



Title	監視カメラ映像による飼育動物の個体識別をとまなうトラッキングと行動検出に関する研究
Author(s)	西岡, 拳
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第16173号
Issue Date	2024-09-25
DOI	https://doi.org/10.14943/doctoral.k16173
Doc URL	https://hdl.handle.net/2115/93401
Type	doctoral thesis
File Information	Ken_Nishioka.pdf



令和6年度 博士論文

監視カメラ映像による飼育動物の
個体識別をともなうトラッキングと行動検出に関する研究

北海道大学 大学院情報科学院
情報科学専攻 情報理工学コース 自律系工学研究室

西岡 拳

2024 年 9 月

Doctoral Dissertation

A Study on Tracking with Identification and Behavior Detection of Zoo Animals
Using Surveillance Camera Footage

Ken Nishioka

Laboratory of Autonomous Systems Engineering
Course of Computer Science and Information Technology
Division of Information Science and Technology
Graduate School of Information Science and Technology

Hokkaido University

September, 2024

目次

第 1 章	序論	1
1.1	研究背景	1
1.2	研究目的	2
1.3	本論文の構成	3
第 2 章	自動エソグラム生成システム	4
2.1	エソグラム	4
2.2	センサーを用いた自動エソグラム生成に関する既存研究	9
2.3	映像を用いた自動エソグラム生成に関する既存研究	10
第 3 章	個体識別をともなうトラッキング	20
3.1	はじめに	20
3.2	データセット	22
3.3	ネットワークフローを用いたトラッキング	24
3.4	提案手法	29
3.5	実験	35
3.6	考察	47
3.7	応用-地図へのマッピング	49
3.8	まとめ	50
第 4 章	TAL による食事行動検出	52
4.1	はじめに	52
4.2	典型的な行動検出手法	53

4.3	データセット	61
4.4	手法	65
4.5	実験 - Temporal action classification	69
4.6	実験 - Temporal action localization	72
4.7	考察	77
4.8	本章のまとめ	78
第 5 章 結論		80
研究業績		84
参考文献		85

2.1.1	動物園に求められるシステムの概要図	8
2.2.1	センサー類とその欠点	9
2.3.1	re-identification	12
2.3.2	TAL の推定イメージ	18
3.1.1	提案手法の概要図	21
3.2.1	監視カメラ映像	22
3.2.2	雌雄の違い	23
3.2.3	夜間映像と遮蔽物	25
3.3.1	例：軌跡の割り当て	28
3.4.1	Identification データセット	30
3.4.2	Tracking 処理フロー	32
3.4.3	ネットワークフロー	34
3.5.1	MOTA	43
3.5.2	IDF1	43
3.5.3	HOTA	43
3.5.4	不明瞭ラベルの効果 (軌跡分割基準)	46
3.5.5	不明瞭ラベルの効果 (オクルージョンコスト)	46
3.6.1	時刻ごとの HOTA の結果	48
3.6.2	時刻ごとの軌跡の正答率	48
3.7.1	トラッキング結果と地図へのマッピング	50
3.7.2	位置の 2D ヒストグラム	50

4.2.1	TSN 前処理	55
4.2.2	BMN の概要図	58
4.2.3	BMN の学習方法	60
4.3.1	様々な食事行動	63
4.3.2	食事行動のアノテーション例	64
4.3.3	行動検出データセットのアノテーション作業	65
4.3.4	データセットにおける食事行動の頻度と累積比率	66
4.4.1	食事行動検出手法の概要図	68
4.4.2	特徴抽出と TAL の入力	69
4.5.1	TAC モデル	71
4.6.1	オプティカルフロー出力	73
4.6.2	行動検出結果の PR 曲線	75
4.6.3	TAC と TAL による予測結果の比較	75
4.6.4	カメラと逆向きでの食事行動検出結果	76

表目次

3.5.1	Detection データセット	36
3.5.2	Detection 結果	37
3.5.3	Identification データセット	38
3.5.4	Re-ID 結果	39
3.5.5	Re-ID EfficientNet B2 混同行列結果	39
3.5.6	トラッキングデータセット	40
3.5.7	Tracking 結果	45
3.5.8	統合・分離モデル比較	46
4.3.1	検出行動の候補	62
4.5.1	TAC データセット	70
4.5.2	TAC 実験結果	72
4.6.1	TAL データセット	73
4.6.2	Ablation Study 結果	77

第 1 章

序論

1.1 研究背景

動物園には「種の保存」「教育・環境教育」「調査・研究」「レクリエーション」という 4 つの役割がある [87]. 「種の保存」という観点からは, 数世代にわたった繁殖も求められている. そのため, 動物園における飼育管理は個体の健康を維持し, できるだけ長命となるように飼育することが求められる. 特に, 動物園における飼育動物の行動観察は, 動物の健康管理と飼育環境の改善のために欠かせない [76][88]. 飼育員は日頃から飼育動物を観察し, 食事量は適切か, 他の個体と喧嘩しないかなどをチェックしている. また, 直接観察するだけでなく, 飼育場の監視カメラ映像を見て異常行動や繁殖行動の有無, 発育状態, 睡眠時間などを記録している. 監視カメラは複数台設置されており, 24 時間稼働している. その映像を飼育員が早送り再生し, 目視で確認しながら行動の記録をおこなっている [48]. 仕事内容が幅広く人員が限られる動物園にとって, これは大きな負担となっている. また, 「調査・研究」の観点から, 野生動物の生態の解明のため, 比較的観察が容易な動物園の動物が研究対象になることがある. 動物行動学として対象の動物を研究する場合, 特定の行動の頻度や継続時間を記録することが基本的なデータ収集方法である [14]. 直接動物を観察するには, 研究対象の行動を見逃す可能性があるため, 定点カメラ映像を用いることもある. ただし, 定点カメラを用いた場合でも, 映像を目視確認する作業は残る.

監視カメラ映像を用いた行動記録による負担の解消や作業効率化のため, 接触センサー (加速度計, GPS, RFID など) を対象の動物に取り付け, 得られたデータから行

動を推定することが考えられる。実際に一定の成果が得られている研究も多いが、接触センサーは動物へのストレスになる場合もあるため、接触センサーよりも非接触センサーの方が動物福祉の観点から望ましい。一方、非接触センサー（カメラやマイクなど）は、対象動物が飼育エリアのパーティションや餌台に隠れることがあるなど課題が多く、使用を避けられる傾向にある。そのため、カメラ映像を用いたエソグラム生成に関する既存研究は理想的な撮影条件の動画や実験動物（ラットやマウス等）の動画でしか検証されていなかった。

そのため、本研究では札幌市円山動物園（以下、円山動物園）の協力のもと、動物園飼育場の監視カメラ映像からの動物の行動記録（エソグラム）作成の自動化をめざしている。

1.2 研究目的

本研究は、監視カメラ映像を用いた個体識別をともなうトラッキングおよび行動検出に向けたアルゴリズムの構築を目的とする。ここで「個体識別をともなうトラッキング」とは、対象動物の検出および各個体に与えられた個体 ID（識別 ID）の判別をおこない、各個体の軌跡（各フレームで対象個体の位置を特定し、時系列データとしたもの）を得ることを意味している。複数物体のトラッキング（Multi-Object Tracking; MOT）のタスクでは、各対象の軌跡の入れ替わり（ID スイッチ）の少なさのみを評価し、その対象がどの個体なのかは評価に入れない。一方、個体をともなうトラッキングとは、ID スイッチだけでなく、識別 ID の判別率も評価に用いる。動物園飼育動物の管理には個体ごとの健康状態を把握する必要があるため、個体識別をともなうトラッキングとして考えるのが適切である。本研究では、「個体識別をともなうトラッキング」におけるワークフローを確立することを目的とし、環境の変化やカメラから姿が確認しづらい状況などといった実運用で起こりうる状況に対する頑健性をもつ手法を構築する。

また、監視カメラ映像から動物の行動を確認する場合、カメラの向きやオクルージョンによって対象個体が確認しづらいケースも多い。そのため、行動の判別にはある程度まとまった時間を用いることが自然だと考えられる。本研究では、数十秒から数分の映像を用いることによる、カメラの向きやオクルージョンに頑健な行動検出手法の

構築を目的とする。

1.3 本論文の構成

本論文の構成について述べる。本論文は全 5 章から構成される。1 章では、序章として本研究の背景と目的を述べた。

2 章では、エソグラム自動生成システムが必要とされる理由とその関連研究について述べる。

3 章では、個体識別をともなうトラッキングについての研究を述べる。本章では、個体識別をともなうトラッキングを実現するため「物体検出」「個体識別」「トラッキング」の 3 タスクの既存手法を組み合わせたモデルを提案する。実験では、札幌市円山動物園のゾウ舎における 2 頭のゾウを対象に手法の有効性を示す。提案手法は 24 時間分のカメラ映像で評価し、環境の変化やカメラから姿が確認しづらい状況などといった実運用で起こりうる状況に対する頑健性をもつことを示す。

4 章では、食事行動検出についての研究を述べる。本研究では、監視カメラの映像のみから動物園のゾウの食事行動を自動的に検出する手法を構築する。本研究では、Temporal Action Localization(TAL) のモデルを用いる。TAL とは、動画から対象の行動がいつ現れたのかを予測するタスクであり、動画分類モデルよりも比較的長い動画の入力が想定されている。3 章同様に円山動物園の 2 頭のゾウの動画からデータセットを作成し、学習、評価をおこなうことで、TAL の有効性を示す。また、食事行動に限定し行動検出をおこなうことで、単一フレームからでは判別ができないような行動が検出できるか確かめる。

最後に、5 章において本論文の結論を述べる。本論文が取り組んだ問題と提案手法についてまとめ、その後今後の展望について述べる。

第2章

自動エソグラム生成システム

本章では、エソグラムとそれに関連した研究について説明する。まず、エソグラムが動物行動学および動物飼育管理の立場から必要とされる背景について述べる。つぎに、(カメラ以外の) センサーを用いた研究およびカメラ映像を用いた既存研究についてそれぞれ述べる。最後に、実例として札幌市円山動物園に求められるシステムについて述べる。

2.1 エソグラム

「エソグラム」の定義は文献によって違いはあるが、本研究ではエソグラムを「観察動物の個体ごとの行動記録」とし、この行動記録には行動の種類だけでなく、その行動がなされた時刻と場所も含める。エソグラムを作成する目的は、動物行動学としての知見の蓄積や、動物の体調変化や新しい環境への適応度を定量的に測るためである。以下では、これらについて詳細を述べる。

2.1.1 動物行動学研究からの要請

ここでは、まず動物行動学研究の目的を、Niko Tinbargen が述べた「4つの問い」を用いて説明する。つぎに、動物園の動物を対象にする研究が、動物行動学全体の貢献につながることを説明する。最後に、その目的を達成するために、研究方法としてエソグラムの作成が必要となることを述べる。

■**動物行動学研究の目的** 動物行動学とは「なぜそのような行動をするのか」という問いに答える研究分野であるが、Niko Tinbergen はその問いを「メカニズム（直接的な原因，因果関係）」「適応（生存価，繁殖）」「発達」「系統（進化的歴史）」4 つに分類している [14] [66] [89]．まず「メカニズム」は，その行動にはどのような体の機能が関係しているのか，という問いである．例えば，「どのようにして孵化後のウミガメは海に向かって進むのか」という問いであり，その答えは「海に反射した月の光がウミガメの網膜に映り，脳で光の方向を特定し，脳から手足の筋肉に指令が送られ，光の方向に進む」となる．つぎに「適応」は，その行動が生存率または繁殖率を高めるのにどのように役立っているか，という問いである．例えば，先のウミガメの例を用いると，「なぜ孵化直後のウミガメは海に向かって進むのか」という問いに対して，その答えは「鳥などの天敵に食べられないようにするため」となる．つぎに「発達」は，その行動は生涯でどのように獲得されていったのか，という問いである．例えば，「鳥はどのようにしてさえずりを学ぶのか」という問いに，その答えは「幼鳥期に母鳥のさえずりを聴いて学ぶ」となる [79]．最後に「系統」は，その行動は歴史上どのように進化したのか，という問いである．例えば，カエルには卵を産み子が成長するまで育てる種と，卵を産んだらその場を離れて子の成長に関与しない種がいるが，育てるカエルと卵の大きさには相関があることが分かっている．そこで，「まず大きな卵を産むカエルが現れて，そのあとに子の世話をするようになったのか，もしくは，世話をすカエルが現れ，そのあとに大きな卵を産むようになったのか」という問いが生まれる．その答えは，現在では「大きな卵を産むカエルが現れ，そのあとに世話をするようになった」と考えられている [14]．動物行動学者は，これらの問いに応じた適切なデータ収集の方法および解析手法を選択する必要がある．

■**研究方法** 動物行動学の研究方法として，観察または実験によるデータ収集がおこなわれる．さらにデータの解析には統計処理や戦略モデル [89] が用いられることもある．観察によってデータを収集するためには，直接対象を観察するだけでなく，カメラ映像や GPS などのセンサーも用いることがある．ただし，これらの機器を用いても，すべての行動を調べ挙げるのは難しいため，仮説に基づいて特定の行動データを集中的に取得する．研究例として，タンザニアのトムソングゼル在行动研究を紹介す

る [20]. トムソンガゼルは、ハイエナなどの捕食者の前で「ストットティング」と呼ばれる 4 本の脚を伸ばしながら空中に跳ぶ行動をおこなう. FitzGibbon と Fanshawe は、その行動によってガゼル自らの健康状態を捕食者に示すことができ、結果として捕食者から逃れることができると仮説を立てた. 仮説の有効性は、2 年間にわたり国立公園で直接観察をおこない、日中のストットティングの頻度や継続時間のデータを収集し、統計による分析をおこなうことにより証明した. 結果、より多くのストットティングをおこなうガゼルは捕食者から逃れる可能性が高いことが示された.

観察によって仮説を示すには、適切なデータ収集方法が必要である. Grafen と Hails は、実験計画の 3 原則を示し、その原則に基づいた統計的解析手法をまとめた [14] [24]. その原則は、「反復 (独立試行で得られたデータあること)」「無作為化 (交絡変数を統制すること)」「ブロック化 (背景因子を制御すること)」であり、行動観察にも当てはまる. 仮説を正しく検証するためには、場合によっては年単位のデータ収集が必要である.

最近では、その 3 原則を守りながら効率的な行動観察のためのアプリケーションも登場している. 例えば、ZooMonitor [74] は動物の行動を記録するためのウェブアプリケーションである. スマートフォンやタブレットで行動観察記録を作成し、サーバーに保存して分析レポートとして閲覧できる.

■動物園における行動学研究 動物行動学研究の目的を果たすため、上記の 3 原則に基づきながら野生生物を観察するのは現実的に難しいことが多い. そこで、動物園などの飼育管理下に置かれた動物で観察データを収集することが考えられる. 家畜とは異なり、動物園では野性を残すように飼育されているため、動物園での観察も有益であると考えられている [14] [88]. ただし、動物園では野生動物とは異なる環境で飼育されているため、野生との環境の違いが結果に影響することがないように考慮する必要がある. 例えば、動物のコミュニケーション行動について研究する場合、行動の発現には個体数や環境が影響することがある [14] [88].

動物園の動物を対象にした研究例として、Martin による母親と分離されたチンパンジーの行動調査を挙げる [45]. Martin は、イギリスの 5 箇所の動物園において 43 頭のチンパンジーを対象に調査を行い、チンパンジーの遊びや毛づくろいを記録した. これらの行動は社会性を反映していると考えられる. 母親と分離されたチンパンジー

は、母親と共に飼育されたチンパンジーよりも社会性が劣ると予想されていたが、母親との分離による影響はほとんど無いことが示された。観察データは複数の動物園で収集されており、それぞれの個体は生育条件が異なる。それゆえ、導き出された結果は、動物園や野生などの環境には左右されない普遍的なものであり、チンパンジーの社会的行動の解明に貢献していると考えられている [14]。

ただし、たとえ動物園においてデータを収集できるとしても、24 時間の監視カメラ映像などを見て、検証のために十分な量の行動パターンを記録するのは膨大な時間がかかる場合がある。そのため、研究者支援のための自動エソグラム生成システムが求められる。

2.1.2 動物飼育管理からの要請

ここでは、個体管理や健康維持のための具体的な手段について説明する。また、動物園従事者が意識すべきである動物福祉について述べる。最後に、実際に求められる自動エソグラムシステムについて述べる。

■**個体管理** 「種の保存」「調査・研究」という役割から、動物園での個体管理は必須である。個体管理には、種ごとの個体識別が必要である。外見から個体差が判別できる場合、カメラに写真を撮る、図に書くなどの方法で管理をしておく。外見から判断できない場合、足環や耳標、マイクロチップなどを用いる。個体情報はデータベース化され、性別、繁殖年月日、出生地といった情報とともに整理される [88]。

■**健康維持** 動物園の役割には、動物の健康が前提にある。動物園の動物は野生性を残しているため、外見上で異常が見られないことも多い。ただし、日常に十分な観察し、健康状態を把握することで異常に気付くこともできる。また、日常の観察に加えて定期的な健康診断をおこなうこともある。そこでは、糞便検査や血液検査、レントゲン検査などがおこなわれ、検査結果はデータベースに蓄積されていき、病気の診断にも役立てられる。

また、発情期や妊娠の兆候があれば、飼育環境を変えることも考えられる。精確な診断には血や糞尿を分析し生殖関連ホルモンの測定をする必要があるが、特定の鳴き声や行動からそのような兆候がわかることもある。そのため、日常の観察により個体

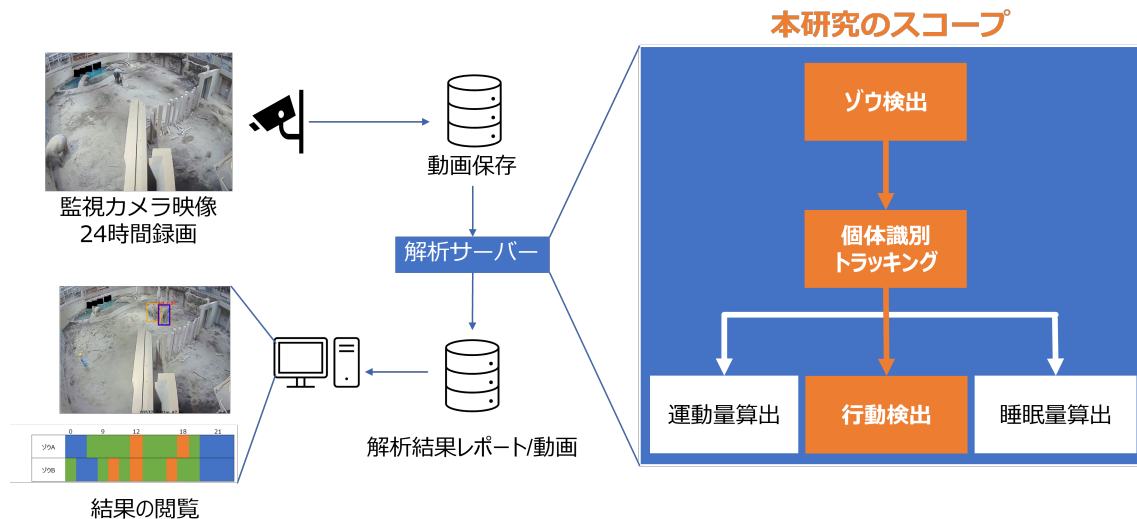


図 2.1.1 動物園に求められるシステムの概要図

の変化に気付くことが求められる [88].

■**動物福祉** 動物福祉とは、「動物の立場に立って個々の動物の生活の質を高め、それを維持するように人間が飼育していくこと」を指す [88]. 欧米諸国では 1950 年代ごろから動物福祉に関する問題提起がおこり、その影響はニューヨークのブロンクス動物園をはじめとした野生に近い状態で飼育する動物園の登場にもつながっている. 近年、日本でも動物福祉の意識が高まりつつあり、2022 年には札幌市で動物園における動物福祉の条例が整定されている（札幌市動物園条例令和 4 年 6 月 6 日条例第 30 号）. 動物園としても動物福祉を考慮した対応が求められている.

■**求められるシステム** 札幌市円山動物園に実際に求められているシステムを説明する (図 2.1.1). 実際の円山動物園で使われている監視カメラ映像を NAS に保存していき、順次、解析サーバーでゾウの検出と、個体識別、トラッキングをおこなって、その結果から各個体の運動量や睡眠量を算出する. また、指定した行動パターンを検出する、ということも行う. 解析の結果は、後で都合のよいタイミングで、動物園従事者が確認できるようにして、動物の飼育に生かせるようなものがよい.

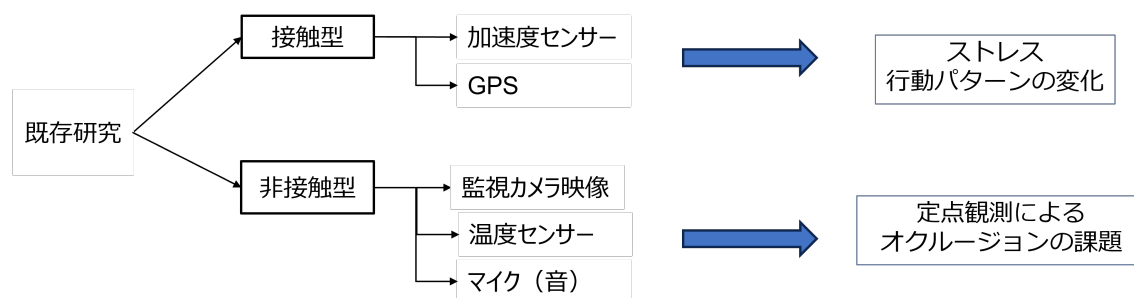


図 2.2.1 センサー類とその欠点

2.2 センサーを用いた自動エソグラム生成に関する既存研究

ここでは入力データの形式に着目し、センサー類を用いた自動エソグラム生成に関する既存研究を紹介する。センサー類は接触型センサーと非接触型センサーに分かれる (図 2.2.1)。

2.2.1 接触センサー

接触センサー（加速度センサー、GPS、RFID など）を用いる場合、センサーを対象動物個体に取り付けることで、その個体の位置のトラッキングや行動判別をおこなうことが可能である。動物園で行われた接触センサーを用いた研究の多くは有効性が示されている [16][76]。例えば、加速度センサーは、軽量であり、夜行性の動物や姿が隠れやすい動物などの直接観察が難しい種にも用いることが可能である。また、動物の運動パターンを定期的に監視し、その個体の痛みや不快感の兆候を検出するのに役立つ。以下、センサーを用いた研究例を数点紹介する。Hokkanen らは、子牛の首に加速度センサーを取り付け、睡眠時間や睡眠の回数などを正確に推定したと報告している [27]。また、Holdgate らは、アジアゾウとアフリカゾウの足に加速度センサーを取り付け、寝転がる頻度と時間を計測した [28]。彼らは、寝転がる行動がゾウの睡眠の代わりとなり、さらにその睡眠パターンを解析すれば動物の管理に役立つと述べている。一方、接触センサーを用いる場合、動物の体型や動き方、行動パターンによって適切なデータの取得が難しいこともある [76]。また、接触センサーを動物の体に取り

り付けた場合、その動物の行動パターンが変化してしまったという事例も報告されている [14]。加えて、動物園の場合には展示や動物福祉といった観点も考慮する必要がある [88]。

2.2.2 非接触センサー

非接触センサー（カメラやマイクなど）については、対象動物が飼育エリアのパーティションや餌台に隠れることがあるといったオクルージョンの課題がある。ただし、動物の状態の把握に役立つことが知られている [76]。Siebert らは、ヤギが社会から孤立するほど低い鳴き声が多いことを発見した。著者らは低い鳴き声が自己を落ち着かせるメカニズムとして機能する可能性があるとして述べている [57]。また、接触センサーは動物へのストレスになる場合もあるため、接触センサーよりも非接触センサーの方が動物福祉の観点から望ましい。以下では、非接触センサーとしてのカメラを用いた既存研究について詳しく述べる。

2.3 映像を用いた自動エソグラム生成に関する既存研究

関連した研究としては、物体検出、個体識別、トラッキングおよび行動検出が挙げられる。本節では、この各タスクに対しての既存研究について述べる。

2.3.1 物体検出

近年、深層学習を用いた高精度な物体検出モデルが非常に多く提案されている。深層学習による物体検出モデルの性能が、深層学習を用いないモデルよりも高いため現在の主流となっている [13][44]。

Tan らは EfficientNet [64] をベースにしたモデルである EfficientDet を提案し、入力解像度、ネットワークの深さ、ネットワークの幅をスケールする方法について考案、検証している [65]。このスケールング手法により、同様の構造をもつモデルを速度・精度のトレードオフを考慮して使用することができる。EfficientDet の処理フローを説明する。まず、画像はバックボーンネットワークである EfficientNet で特徴抽出される。この際、プーリング層で空間解像度を下げながら畳み込むことで、複数スケール

ルの特徴を得る．さらに、この特徴を BiFPN と呼ばれる構造により、複数スケール間の特徴ベクトルを混ぜ合わせていく．BiFPN は FPN (Feature pyramid networks) 構造 [39] と呼ばれる手法をベースにしており、CNN を用いた多くのモデルでこの構造が取り入れられている [8] [22] [40] [82]．最後に、特徴ベクトルはクラスを出力する層と矩形位置を回帰する層に渡される．

EfficientDet を改良し、精度を高めたものに YOLOv5 [23] があげられる．ただし、YOLOv5 のモデルの詳細は論文として示されておらず、コードを直接確認する必要がある．

■**動物の物体検出** 動物に特化したモデルとしては、MegaDetector [4] がある．これは主に自動撮影カメラ（カメラトラップ）映像から野生動物を検出するモデルとして作成されている．MegaDetector は生物学者が観察に注力できるようにすることを目的にしている．トラップカメラ画像に映る「動物」、「人」、「車」を検出するが、個体の識別はおこなわない．MegaDetector は複数のバージョンがあるが、2023 年 9 月時点の最新モデルである v5 は YOLOv5[23] がベースとなっている．画像解像度は 1280 x 1280 サイズであり、学習データは 複数の公開データセットと非公開データから構成されている．公開データセットとしては、Snapshot Serengeti [3] [19], Caltech Camera Traps [62], iNaturalist [69], COCO [41] などが用いられている．

2.3.2 個体識別

個体識別では、目的に応じて手法を考える必要がある．動物園の飼育動物など、各個体に識別 ID が定まっている場合は判別問題として解くことが考えられる．一方で、野生動物などの観察の際は、事前に識別 ID が定まっているわけではない．この場合、再照合 (Re-Identification) として問題をとらえることが自然である．ここでは、(1) 判別問題として解く、(2) Re-Identification として解く、の 2 つに分けて関連研究を述べる．

■**判別問題** 個体識別をおこなう場合、個体の識別 ID を判定する CNN モデルを用いることが考えられる．CNN は、VGG の他に、ResNet [25], MobileNetV2 [53], EfficientNet [64] 等現在まで数多くのモデルが提案されている．ResNet は、スキップ

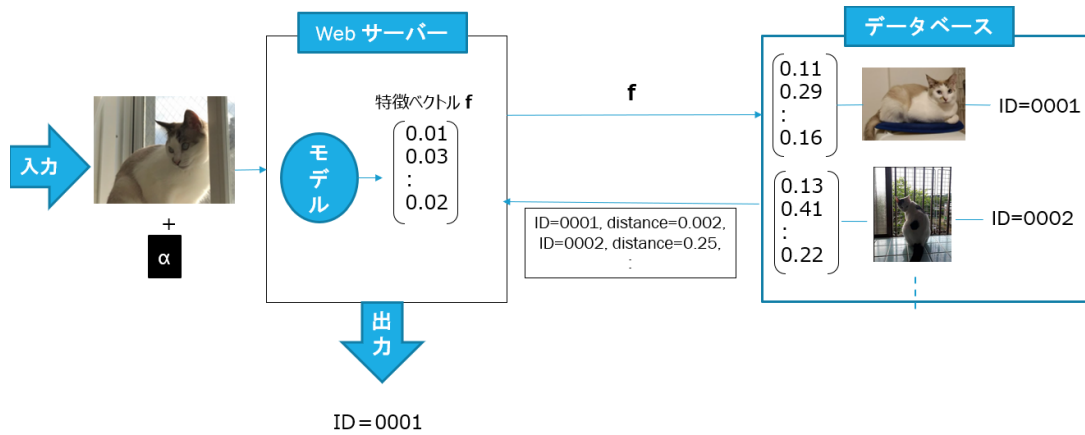


図 2.3.1 re-identification

接続と呼ばれる手法を導入することで、VGG と比較して非常に深いネットワークの学習を可能としており、結果として高い性能を達成できる。ResNet は 2016 年の登場以来、様々なタスクで広く利用されている。一方で、計算量やメモリ消費量が多く、画像の解像度や求める精度によっては多大な計算リソースを必要とする。EfficientNet は、2.3.1 節で述べたように、モデルのスケーリングによって効率的に処理することが可能である。一方で、ResNet と比較すると EfficientNet の適用例は少ない。MobileNet は、Depthwise Separable Convolution[12] と呼ばれる畳み込み演算の手法を導入し、計算効率を向上させている。エッジデバイスなどの計算リソースが制約された環境でも動作ができる一方で、ResNet や EfficientNet よりも精度面で劣る。

■Re-Identification Re-Identification の処理の流れを説明する (図 2.3.1)。識別 ID、識別 ID に紐づいた画像、および、その画像から得られる特徴ベクトルで構成されるデータベースを用意する。特徴抽出の際に用いる学習モデルは、距離計量学習によって得ることができる。距離計量学習とは、ここでは各識別 ID 間の距離を学習する手法を指す。距離計量学習の手法としては、CNN を Contrastive Loss[70]、Triplet Loss [55] によって学習するものが代表的な手法として挙げられる。ある個体が検出された際は、学習モデルによって特徴ベクトルに変換し、データベース内の特徴ベクトルとの距離を算出する。登録済みの識別 ID との距離が近い場合、その識別 ID、画像、および、特徴ベクトルを紐づけ、そうでないときは未登録 ID として新規にデータベースに追加する [81]。

■動物の個体識別 ここでは、画像のみを用いた個体識別の既存研究について述べる。

Schofield らは、23 体のチンパンジーの顔識別のために VGG [58] を用いており、92.5% の正答率を達成したと述べている [54]。また、Pang らは動画を入力とした羊の顔識別に取り組んでおり、これには MobileNetV2 をベースにしたモデルを用いている [50]。また、Moskvyak らは、目、耳、尾などのキーポイントを入力に加えることにより、個体識別の精度が大幅に向上したと報告している。合成データセットと水中で撮影されたマンタ（大型のエイ）の画像に対して実験をおこない、手法の有効性を示した [47]。Moskvyak らの研究は、動物の Re-Identification に関する研究にあたる。画像に加え、ランドマークの位置を CNN に与えることで少ないデータセットに対応でき、さらに、ランドマーク位置のヒートマップを再構築する学習で精度を向上させることが可能となったと報告している。論文では、以下のように動物の個体識別の難しさについて述べられている。

- 動物は「人の顔」に相当する箇所が少ない。そのような箇所があったとしても画像では見えにくい場合がある。
- データ量が少ない場合、深層学習モデルでは個体識別のカギとなる箇所（キーポイント）を自動で見つけるようにならない。
- 人の場合、顔の向きで体の大まかな方向がわかるため、向きを揃えることでバリエーションを減らすことができるが、動物の場合はキーポイントから体の向きを推定するのが難しい。

また、体の部位の形状に着目する研究もある。Weideman らは、画像から体の一部の輪郭を FCNN を用いて抽出し、曲線の曲率を用いた特徴量を算出している [75]。手法はクジラの尾びれとアフリカゾウの耳で有効性を示している。

顔やキーポイントを利用することは有効である反面、物体検出用のデータセットとは別に、顔またはキーポイントをアノテーションしたデータセットを用意し、学習モデルを作成する必要がある。

2.3.3 トラッキング

ここでは、複数物体の追跡 (Multi Object Tracking; MOT) に焦点を当て、既存研究を紹介する。以降、単に MOT を「トラッキング」と呼ぶ。トラッキングは物体検出の出力と組み合わせる「Tracking by Detection」による手法の精度が高いことから、近年の主流になっている [44]。なお、トラッキングは教師ありの手法と教師なしの手法がある [13]。教師ありの場合は精度がよい傾向にある半面、教師データとして、連続したフレームに対する個体毎の矩形位置情報が必要であり、そのアノテーション作業には多大な労力を必要とする。なお、異なる方向から撮影した映像が利用可能な場合、より正確にトラッキングできると考えられる [26]。ただし、本研究では施設の制約から単一カメラの映像のみを用いる。

■**オンライン手法** トラッキングはオンライン (またはリアルタイム) とオフライン (またはバッチ処理) の手法に分かれる。ここではオンライン手法について述べる。SORT は、カルマンフィルタとハンガリー法を組み合わせたシンプルな追跡手法を採用している [6]。IoU Tracker [7] は、画像情報を使用せずに高速なトラッキングを実現しており、隣接フレームとの IoU を比較して軌跡の割り当てをおこなっている。IoU Tracker は、従来のトラッキング手法と比較して、計算コストを大幅に削減しながら高い性能を達成している。また、DeepSORT [77] は ImageNet 等で事前学習済みモデルの中間出力を利用することで、見た目を重視したマッチングが可能である。最近では、ByteTracker [86] や Transformer [71] ベースの TransTrack [61] 等続々と性能がよい手法が提案されている。ただし、基本的に人や車など、見た目の区別が着きやすいデータで実験されていることが多い。動物のような個体差が見た目で判断しづらい場合、最近の手法では必ずしも精度が上がるわけではないという報告もある [85]。

■**オフライン手法** オフライン手法の場合には、オンライン手法と異なり、過去フレームだけでなく未来のフレームも使えるという利点がある。そのため、精度はオフライン手法の方がオンライン手法よりもよい傾向にある。オフライン手法では、ネットワークフローとして定式化をおこない、最小費用流問題を解くことによって各トラッキングの軌跡を求めることが多い [29][30][44][33][84]。ネットワークフローを用いた

トラッキングについては 3.3 節で詳しく述べる。

2.3.4 個体識別をともなうトラッキング

本研究のような個体識別をともなうトラッキングに関する研究は多くはない。ここでは、2 点の既存研究について紹介する。

idTracker は、実験動物の行動分析を目的としたトラッキング手法であり、ゼブラフィッシュ、ショウジョウバエ、クロナガアリおよびハツカネズミを対象に手法の有効性が示されている [51]。idTracker は、それぞれのフレームから二値化に基づく個体の領域抽出をおこなった後、他個体とオーバーラップしていない個体に対して軌跡を求める。次に、各軌跡上の個体の画像と、事前に収集した各個体画像（参照画像）の特微量と比較し、軌跡単位で識別 ID を割り当てている。著者らの提案手法と同様に、idTracker も個体識別をともなったトラッキング手法である。ただし、idTracker の適用範囲は、背景が一様で変化が少ない環境に限られており、カメラの角度も対象を真上から撮影し、対象の画像上のサイズがあまり変化しない状況が想定されている。また、オクルージョンは対象物同士の重なりに関わり対策が取られており、何かの陰に隠れるという状況は考えられていない。

Schofield らは、23 体のチンパンジーの個体識別のために、フレームごとの顔検出に続いて Kanade-Lucas-Tomasi (KLT) tracker[42] による顔のトラッキングをおこなっている [58]。ただし、Schofield らの手法はチンパンジーの顔がカメラから確認できる状況に限定されている。

2.3.5 行動検出 - Temporal Action Classification

行動検出に関わるタスクとして、短時間の動画のクラスラベルを予測する Temporal Action Classification (TAC) と、比較的長い動画に対して検出対象の行動の開始時刻と終了時刻を予測する Temporal Action Localization (TAL) が挙げられる。本節では、TAC について述べる。

■データセット TAC の代表的なベンチマークデータセットとしては、HMDB-51 [35], UCF101 [60], Kinetics [34] が挙げられる。これらは約 10 秒 [30 フレーム/秒

(FPS)] 程度の動画ファイルである。UCF101 は行動ラベル数が 101, 計 13,320 の動画ファイルで構成される。HMDB-51 は映画や Web 動画などの幅広い動画データを用いており, 51 ラベル 6,766 の動画ファイルで構成される。また, Kinetics はさらに大規模な動画データセットであり, 400 クラス 306,245 の動画ファイルで構成される。

■**代表手法** 畳み込みニューラルネットワーク (CNN) は画像分類タスクで高い性能を示しているが, 多くの行動分類手法も CNN を用いている。Tran らは 3 次元 CNN (3D CNN) を用いた手法を提案し, 3D CNN が時空間特徴を学習できることを示した [67]。画像分類の場合, 大量の画像データによる事前学習が有効であると示されているため, 動画判別の場合も事前学習をおこなうことが当然考えられる。ただし, 大量の動画データを用意する必要があるため, 3D CNN はモデルの学習は難しいとされている。そこで, Carreira と Zisserman は I3D [11] と呼ばれる, 画像分類のために事前学習された 2D CNN を 3D CNN に変換する手法を提案し, その高い性能を示した。しかし, 3D CNN は計算コストが高い。Wang らは, 3D CNN の代わりに 2D CNN で抽出された特徴を用いる Temporal Segment Network (TSN) と呼ばれる手法を提案した [73]。TSN は時系列をパーティションに分割し, 各パーティションから特徴量を抽出し, 各特徴量のコンセンサス (平均化など) によって分類を行う。また, モデル構造を工夫するだけでなく, 複数モーダル of データを用いる手法も考えられておる。I3D と TSN は, オプティカルフローを用いた学習も組み込むことが可能である。これらは 2 ストリーム方式と呼ばれ, 事前に計算されたオプティカルフローと RGB 画像を同時に学習に用いる。

■**最近の手法** 最近の研究では, TSN や I3D の欠点を補うモデルが提案されている。例えば, TSM [36] は TSN の欠点を補うものであり, TSN では取れない時間情報を時間軸方向に特徴ベクトルをスワッピングすることで精度をあげている。また, X3D [18] は I3D の計算コストを削減し, 効率的な学習と推定を実現した。上述した手法は, 主に CNN を用いる手法だが, vision Transformer [17] を適用する手法もいくつか提案されている [56]。これらの手法は精度を大幅に向上させると報告されている。

■**長期間の動画に対する研究** また, 長期間の動画を対象にした研究も提案されている。LFB (Long-term feature banks) は, 動画全体から特徴抽出する仕組みである。

[78]. LFB では、事前学習した 3D CNN を用いて、スライディングウィンドウしながら動画全体から特徴抽出をおこない、複数のタスク（分類、物体検出など）に適用し精度が向上することを示した. Context R-CNN は LFB をベースに静止カメラ (static camera) で撮影された動画を対象に、高精度な物体検出モデルを提案した [5]. LFB と異なり、Context R-CNN の場合は、R-CNN[52] の RoIPool 後の特徴ベクトルを用いている. 訓練データとテストデータは同じカメラで撮影し、ラベルの無いフレームも利用している. カメラのデータセットにおいて、シングルフレームのモデルと比較し、mAP が 17.9% 向上したと報告し、1 ヶ月のコンテキスト情報が精度向上につながることを示した.

2.3.6 行動検出 - Temporal Action Localization

TAL は、TAC が対象とする動画よりも比較的長時間の動画を入力として、検出対象の行動が起きた開始時刻と終了時刻を予測するタスクである.

■**データセット** TAL のベンチマークとなる代表的なデータセットとして、THUMOS 14 [31] と ActivityNet v1.3 [80] が挙げられる. これらのデータセットの検出対象の行動ラベルは、複数のクラスとなることが多い. 例えば、THUMOS 14 には「SoccerPenalty」クラスが含まれており、検出対象となる動作はボールが蹴られた瞬間からゴールに入った瞬間までとなっている.

■**代表手法** 典型的な TAL のプロセスは以下の通りである. まず、事前に学習された行動分類モデルを使って、短期的な特徴量を抽出する. この特徴抽出のモデルとして TSN と I3D がよく用いられる. 次に、これらの特徴量を時間軸に沿って結合し、長期的なコンテキストを持った特徴量 (特徴列) とする. 次にこの特徴列から、別のネットワークを用いて、クラスを限定しない行動の開始時刻と終了時刻の組を出力する. この組は「プロポーサル」と呼ばれる. その後、プロポーサルごとに行動ラベルが予測される. 行動ラベルの予測には、まず、プロポーサル内のセグメントからフレームをサンプリングし、行動分類モデルに入力する. 次に、分類モデルのスコアの加重平均あるいはソフトマックスのスコアの平均を用いて行動ラベルの判定をおこなう. また、分類モデルのスコアによって事前学習された別のモデル（たとえば、ニューラルネッ

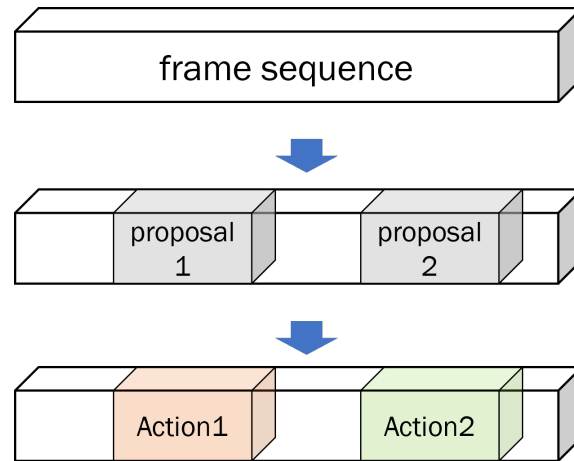


図 2.3.2 TAL の推定イメージ

トワークや線形分類) の出力を用いる手法もある [68] [72].

典型的な TAL アーキテクチャとして, Boundary-Matching Network (BMN) [37] が挙げられる. BMN は行動の開始時刻と終了時刻のスコアと, 各プロポーサルの信頼度スコアを出力する. BMN は, 開始時刻と行動継続時間を軸とする 2 次元の密なマップとして予測することで, 信頼性の高いスコアを出力することができる. TAL の研究では, BMN を改良するためのモデルがいくつか提案されている. オリジナルの BMN は時間特徴ベクトルに CNN を用いているが, CNN は局所的な特徴を重要視するため, 大域的な特徴が反映されない傾向にある.

■最近の手法 長時間の特徴量を得るために, グラフ畳み込みネットワーク (GCN) や Transformer を用いる手法が提案されている [68]. ただし, これらの方法は BMN よりも精度が高いが, 計算コストが高く, 比較的多くの学習データを必要とする.

2.3.7 動物の行動検出

動物を対象にした行動検出の研究例は少ないが, ここでは 2 点の既存研究を挙げる.

Yin らは EffientNet-LSTM を提案し, 家畜牛の行動認識に関してその有効性を示した [82]. 実験は主に行動分類モデルを用いており, 5 つのクラスラベル (飲む, 立つ, 横たわる, 歩く, 食べる) が用いられている. Yin らはオプティカルフローを用いていないが, これは動画がハンドカメラで撮影されたものであるためオプティカルフローは精度向上に寄与しない, と理由を述べている. また, スライディングウィン

ドウを用いた分類により，トリミングされていない動画に対しても実験をおこない，日々の牛の健康チェックに使えるほどの精度を記録したと述べている．しかし，短期間の動画でしかクラスラベルを予測できないため，誤判定するケースが報告されている．例えば，「食べる」を「歩く」と誤判定した事例がある．Yin らはこの原因を採餌中に頭を上げる動作は歩いている状態とよく似ているためであると報告している．

DeepEthogram [10] は，実験動物の動画からエソグラムを推論する行動認識モデルである．RGB フレームに加えてオプティカルフローを用いる．ただし，この研究で用いられた動画は，背景がほとんど変化しないといった理想的な条件で取得されたものである．

第 3 章

個体識別をともなうトラッキング

3.1 はじめに

本章では、行動記録の自動化のベースとなる個体識別をともなうトラッキング手法について提案する (図 3.1.1). 実環境での運用に耐えうる手法を構築することを目的とし、特に、環境の変化やカメラから姿が確認しづらい状況などといった実運用で起こりうる状況に対する頑健性をもつ手法を提案する。特に、アジアゾウ（以下、単にゾウ）を対象にして各動物個体の位置を長時間にわたり追跡する手法について述べる。

本研究では、「物体検出」「個体識別」「トラッキング」の 3 つのタスクの既存手法を組み合わせたモデルを構築する。それぞれのタスクを Detection, Identification, Tracking と表記する。以下に手法の概要について説明する。

まず、深層学習における物体検出モデルによって動画フレームごとに対象動物の位置を推定する (図 3.1.1 Detection). 飼育場ではエサ箱や遊具の定期的な変更や床砂がたびたびゾウによって掘り返されるが、深層学習を用いることでこのような背景の流動性にも対応することができる。次に、推定された位置で画像を切り抜き、別の深層学習モデルによって各個体の判別をおこなう (図 3.1.1 Identification). ここでは各個体の識別 ID を予測する分類モデルを用いる。最後に最小費用流を用いたトラッキング手法によって各個体の軌跡（各フレームでゾウの位置を特定し、時系列データとしたもの）を求める (図 3.1.1 Tracking).

実験では円山動物園のゾウ舎における 2 頭のゾウを対象に手法の有効性を示す。著者らは円山動物園の監視カメラから映像を取得し、ゾウの位置と個体のアノテーション

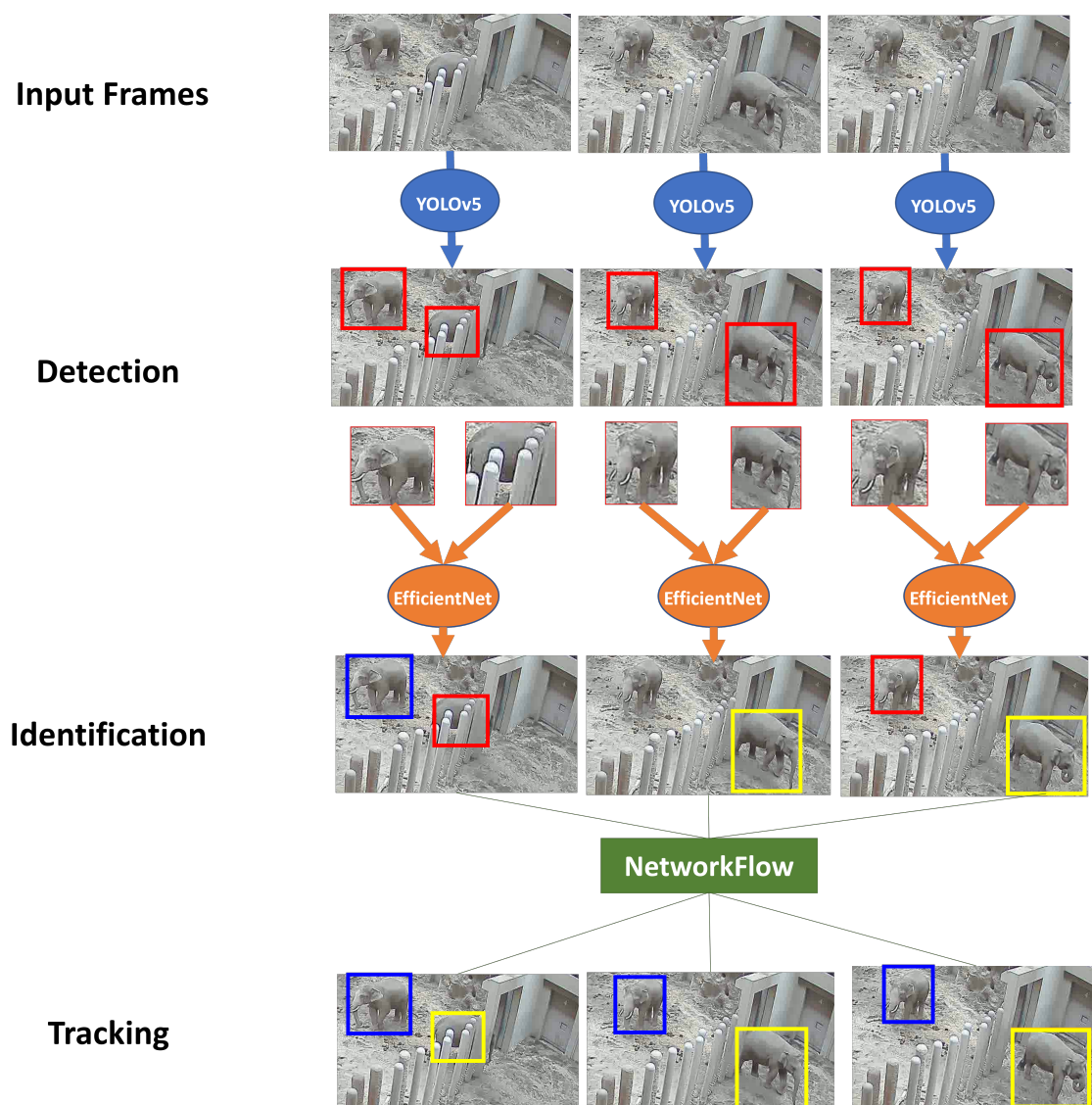


図 3.1.1 提案手法の概要図

ンをおこないデータセットを作成した。また、提案手法は 24 時間分のカメラ映像で評価している。

本章の構成を述べる。3.2 節で本研究に用いるデータセットについて述べる。3.3 節では、提案手法に用いられるネットワークフローを用いたトラッキング手法について説明する。3.4 節では本手法についての詳細を述べ、3.5 節では実施した実験手法と結果について報告し、3.6 節に考察を述べる。3.7 節では、得られたトラッキング結果の応用として、地図へのマッピングとその活用例について簡単に述べた。最後に、3.8 節にまとめを述べる。



図 3.2.1 監視カメラ映像

3.2 データセット

本節では、本研究で用いる動画データセットの概要と特性を説明する。

3.2.1 監視カメラ映像

本研究で扱う動画データは、円山動物園のゾウの監視カメラ映像である。映像は園内の屋内飼育場で録画されており、手動でのカメラ操作時を除き 24 時間連続で録画されている。録画データは 1 時間ごとに 5FPS(frame-per-second) の動画ファイルとして取得できるが、通常、ゾウの飼育エリアごとに 1 台の監視カメラが設置されており、複数のカメラによって獣舎全体をカバーしている。そのため、施設の制約上、ゾウの観察には単一の監視カメラ映像のみしか用いることができない。また、昼間（およそ午前 5 時から午後 6 時）はカラー映像として記録されるが、夜間（およそ午後 6 時か



図 3.2.2 雌雄の違い

ら午前 5 時) は自動で暗視モードに切り替わり、グレースケールの映像として記録される。また、カメラ位置と撮影角度は常に固定であり、本研究で用いる映像は屋内飼育場を上から見下ろす形で撮影されたものである (図 3.2.1)。

3.2.2 対象物

円山動物園には 5 体のゾウ (メス 4 頭, オス 1 頭) が飼育されている。それぞれのゾウは諸般の事情から同一の飼育場で飼育するか, 別々の飼育場で分けて飼育されるかが決められている。本稿では, 特定のメスとオスが 1 頭ずつ映る動画を対象にする。これは, ゾウの繁殖に関わる行動分析の飼育員からのニーズが強いためである。オスには牙があるがメスには牙が無いいため, 牙の有無を確認することでオス・メスの判別が可能である (図 3.2.2)。本研究に用いる監視カメラ映像からでも牙は確認できる。しかし, ゾウの正面がカメラと反対側に向いてしまうと牙の位置が映像から確認出来なくなり, 1 フレーム単位ではオス・メスの判別は難しくなる。また, ゾウがプールに潜る, 飼育場の仕切りに隠れるといった場合にも同様の理由でオス・メスの判別は難しい。本提案手法では, トラッキングによって複数フレームから軌跡を判別するため, 1 フレーム単位では判別が難しい状況下でも正しい推定は可能である。

3.2.3 解析の障害となりうるもの

ここでは, 映像からトラッキングを実現するうえで障害となりうるものをまとめる。

背景の流動性

定期的にゾウに与えるエサ箱や遊具が変わる。学習時の動画データには無かったものが映り、誤検出される可能性がある。また、床砂はゾウによって掘り返されることもあり、見た目が変わる場合も多い。

対象物と床砂の色の類似性

床砂は映像上ではゾウの肌の色と同じように見える。この場合、ゾウと背景の境界があいまいとなり、検出精度が低くなる原因となる。

対象物のサイズの変動

ゾウの映像上のサイズが変わる。作成した学習データでは、ゾウを囲む矩形サイズは、最大高さ 381 ピクセル、最大幅 565 ピクセル、最小高さ 36 ピクセル、最大幅 19 ピクセルとなった。

夜と昼の違い

24 時間録画のため、夜間は自動で暗視用の映像に変換される (図 3.2.3)。このとき、映像はグレースケール映像であり、色情報が使えず背景をゾウと誤検知する可能性が高まる。また、夜間はシャッター速度が遅くなるため、ブレも大きくなりやすい。

個体差

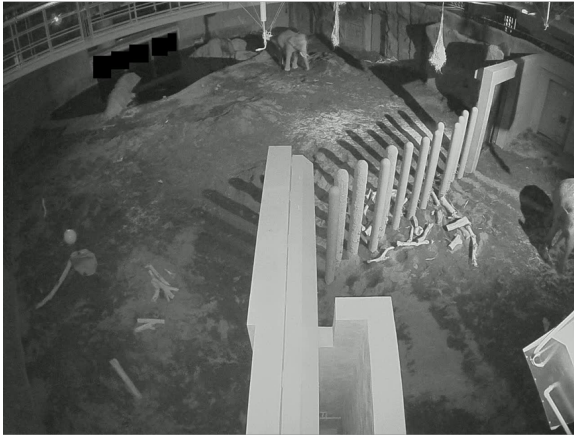
個体は見た目で判別するのは難しい。オスとメスは牙の有無で判別できるが、逆に言うと、ゾウの正面がカメラと反対側に向いている場合など牙が見えない場面では判別が難しい。

遮蔽物

カメラに向かって左奥にはプールがあり、ゾウが潜ってしまうと、当然カメラからは見える範囲が少なくなる。また、カメラの手前側には仕切りがあり、ゾウがその仕切りに隠れる場合もある (図 3.2.3)。

3.3 ネットワークフローを用いたトラッキング

本節では、まず一般的な最小費用流問題について説明する。つぎに、トラッキングの軌跡を求める問題（以下、トラッキングの割り当て問題）を定式化し、トラッキン



夜間映像



遮蔽物

図 3.2.3 夜間映像と遮蔽物

グの問題が非重複条件をもつ最小費用流問題に帰着できることを示す.

3.3.1 最小費用流問題

最小費用流問題とはネットワーク最適化問題の1つであり, 次のような入出力および制約条件をもつ.

- 入力

$G = (V, E)$: $V = \{V_i\}_{i=0}^N$ をノード, $E = \{E_{ij}\}_{i,j=0}^N$ をエッジとした有向グラフ

$s = V_0$: 入口 (source) を表すノード

$t = V_N$: 出口 (sink) を表すノード

$\{b_i\}_{i=0}^N$: 需要量 (または, 供給量)

$\{u_{ij}\}_{i,j=0}^N$: 容量

$\{c_{ij}\}_{i,j=0}^N$: 費用

- 出力

$\{x_{ij}\}_{i,j=0}^N$: 制約条件を満たす流量 (フロー) であって, 目的関数 (総費用) が最小になるもの.

- 目的関数

以下の目的関数を最小化する

$$J = \sum_{i,j} c_{ij} x_{ij}, \quad 0 \leq i, j \leq N.$$

• 制約条件

1. 容量条件:

$$0 \leq x_{ij} \leq u_{i,j}, \quad 0 \leq i, j \leq N,$$

2. 流量保存条件:

$$\sum_{e \in \delta^-(j)} x_e - \sum_{e' \in \delta^+(j)} x_{e'} = b_j, \quad 1 \leq j \leq N-1.$$

ただし、 $\delta^-(j)$ は終点 j をもつエッジの集合、 $\delta^+(k)$ は始点 k をもつエッジの集合とする。

3. 非重複条件: 通常、最小費用流量問題の制約条件は上記 1, 2 の条件のみであるが、トラッキングの定式化にはこの条件を加える [33].

$$\sum_{e \in \delta^-(i)} x_e \leq 1, \quad \sum_{e' \in \delta^+(i)} x_{e'} \leq 1, \quad i \in V - \{s, t\}$$

これは、各ノードの始点および終点に接続するエッジがそれぞれ多くとも 1 本のみであることを表す。

3.3.2 トラッキングの割り当て問題の定式化

本節では、Zhang らの論文 [84] を参考に、確率モデルによるトラッキングのネットワークフローとしての定式化について述べる。

$\mathcal{D} = \{\mathbf{d}_i\}$ を物体検出および個体識別の出力とする。すなわち、予測位置 $\mathbf{r}_i$ およびクラススコア $\mathbf{p}_i$ を用いて $\mathbf{d}_i = (\mathbf{r}_i, \mathbf{p}_i)$ とする。また、(単独の) 軌跡 T_k を、 $T_k = \{\mathbf{d}_{k_0}, \mathbf{d}_{k_1}, \dots, \mathbf{d}_{k_{l_k}}\}$ とおく。本節で求めるのは、トラッキングにおいて複数の軌跡の集合 $\mathcal{T} = \{T_k\}$ である。ここで、最適な軌跡の割り当て $\mathcal{T}^*$ は、以下のように

事後確率が最大となるものと仮定する.

$$\begin{aligned}
 \mathcal{T}^* &= \operatorname{argmax}_{\mathcal{T}} P(\mathcal{T}|\mathcal{D}) \\
 &= \operatorname{argmax}_{\mathcal{T}} P(\mathcal{D}|\mathcal{T})P(\mathcal{T}) \\
 &= \operatorname{argmax}_{\mathcal{T}} \prod_i P(\mathbf{d}_i|\mathcal{T})P(\mathcal{T}).
 \end{aligned} \tag{3.1}$$

ただし, $\mathcal{T}$ の空間が大きすぎるため式 3.1 を直接求めるのは難しい. そこで, 互いの軌跡は共通部分をもたないと仮定する. すなわち,

$$\mathcal{T}_k \cap \mathcal{T}_l = \phi, k \neq l \tag{3.2}$$

とする. このとき, 式 3.1 は以下のようになる.

$$\mathcal{T}^* = \operatorname{argmax}_{\mathcal{T}} \prod_i P(\mathbf{d}_i|\mathcal{T}) \prod_{\mathcal{T}_k \in \mathcal{T}} P(\mathcal{T}_k). \tag{3.3}$$

ここで, $P(\mathbf{d}_i|\mathcal{T})$ は検出 $\mathbf{d}_i$ の信頼度を表すと見なせる. また, マルコフ性を仮定すると $P(\mathcal{T}_k)$ は,

$$P(\mathcal{T}_k) = P(\mathbf{d}_{k_0})P(\mathbf{d}_{k_1}|\mathbf{d}_{k_0})P(\mathbf{d}_{k_2}|\mathbf{d}_{k_1}) \cdots P(\mathbf{d}_{k_{l_k}}|\mathbf{d}_{k_{l_k-1}})P(\mathbf{d}_{k_{l_k}}) \tag{3.4}$$

となる. 式 3.3, 3.4 の対数を取り, 和で表現すると,

$$\mathcal{T}^* = \operatorname{argmin}_{\mathcal{T}} \sum_i -\log P(\mathbf{d}_i|\mathcal{T}) + \sum_k \sum_{l_k} -\log P(\mathbf{d}_{k_{l_k}}|\mathbf{d}_{k_{l_k-1}}). \tag{3.5}$$

ただし, $P(\mathbf{d}_{k_{l_k+1}}|\mathbf{d}_{k_{l_k}}) = P(\mathbf{d}_{k_{l_k}}), P(\mathbf{d}_{k_{l_0}}|\mathbf{d}_{k_{l_0-1}}) = P(\mathbf{d}_{k_0})$ とおいた.

3.3.3 最小費用流問題への帰着

本節ではトラッキングの割り当て問題が非重複条件をもつ最小費用流問題に帰着できることを示す.

$$\begin{aligned}
 x_i^{\mathcal{T}} &= \begin{cases} 1 & \exists \mathcal{T}_k \in \mathcal{T}, \mathbf{d}_i \in \mathcal{T}_k \\ 0 & \text{その他} \end{cases} \\
 x_{i,j}^{\mathcal{T}} &= \begin{cases} 1 & \exists \mathcal{T}_k \in \mathcal{T}, \mathcal{T}_k \text{の中で } \mathbf{d}_i \text{ と } \mathbf{d}_j \text{ が隣接している} \\ 0 & \text{その他} \end{cases}
 \end{aligned}$$

とおくと, $\{x_i^{\mathcal{T}}\}, \{x_{i,j}^{\mathcal{T}}\}$ は軌跡 $\mathcal{T}$ を漏れなく表す. 言い換えると, $\{x_i^{\mathcal{T}}\}, \{x_{i,j}^{\mathcal{T}}\}$ から検出 $\mathcal{D}$ の割り当て方法 $\mathcal{T}$ を復元することが可能である. 図 3.3.1 に例を示す. こ

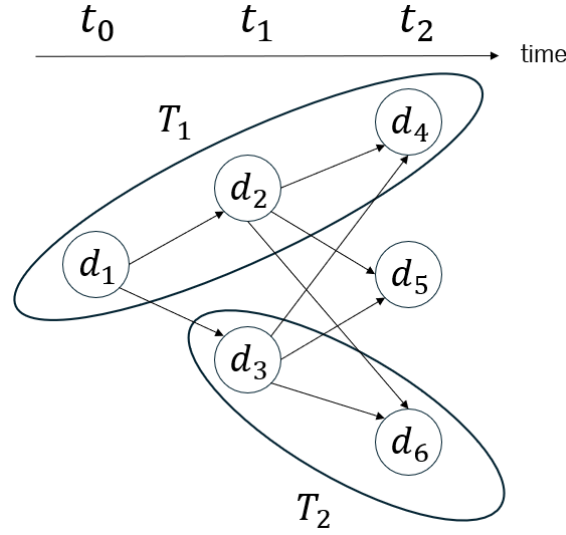


図 3.3.1 例：軌跡の割り当て

の例では, $\mathcal{D} = \{d_1, d_2, \dots, d_6\}$, $T_1 = \{d_1, d_2, d_4\}$, $T_2 = \{d_3, d_6\}$, $\mathcal{T} = T_1 \cup T_2$ である. また, d_5 はどの軌跡にも割り当てられていない. このとき, 添え字 $\mathcal{T}$ を省略すると, $\{x_i\}$ は $x_1 = x_2 = x_3 = x_4 = x_6 = 1$, $x_5 = 0$ であり, $\{x_{i,j}\}$ は $x_{1,2} = x_{2,4} = x_{1,3} = x_{3,6} = 1$ で, それ以外は 0 である.

したがって,

$$\begin{aligned} c_i &= -\log P(d_i | \mathcal{T}) \\ c_{i,j} &= -\log P(d_j | d_i) \end{aligned} \quad (3.6)$$

とおくと, 式 3.5 は 0-1 線形計画問題

$$\mathbf{x}^{\mathcal{T}*} = \operatorname{argmin}_{\mathbf{x}^{\mathcal{T}}} \sum_i c_i x_i^{\mathcal{T}} + \sum_{i,j} c_{i,j} x_{i,j}^{\mathcal{T}} \quad (3.7)$$

と同値である. ただし, $\mathbf{x}^{\mathcal{T}} = \{x_i^{\mathcal{T}}\} \cup \{x_{i,j}^{\mathcal{T}}\}$ とおいた.

さらに条件 3.2 より,

$$x_i^{\mathcal{T}} = \sum_j x_{i,j}^{\mathcal{T}} = \sum_j x_{j,i}^{\mathcal{T}} \quad (3.8)$$

となる. これは $\{x_{i,j}^{\mathcal{T}}\}$ をフローとして見なすと, 3.3.1 節で述べた 3 条件 (容量条件, 流量保存条件, 非重複条件) を満たすため, この問題は最小費用流問題としてみなせる.

3.4 提案手法

本節では提案手法について説明する。

3.4.1 手法概要

物体検知，個体識別，トラッキングを組み合わせることにより最終的な精度を高めることを目的とする．本手法は以下の順に処理を進める．

1. Detection (物体検出)

CNN 物体検出モデルによりゾウの矩形位置を推定する

2. Identification (個体識別)

フレーム画像を矩形位置で切り抜き，CNN 個体識別モデルに入力して各個体の識別 ID を推定する

3. Tracking (トラッキング)

Detection と Identification の結果から各ゾウの軌跡を求める

以下に 1, 2, 3 の各処理内容を詳述する．

3.4.2 Detection

監視カメラ映像に対して，1 フレームずつゾウの位置を検出する．モデルは CNN をベースにした物体検出モデルを採用する．

モデルはゾウの位置を矩形で示したアノテーション付き画像データセットで学習する．検出時には，まず，モデルは信頼度スコアが閾値以上である矩形位置を出力する．次に，出力した矩形位置を用いて検出対象をフレームから切り抜き，切り抜いた画像を Identification モデルに渡す．

3.4.3 Identification

モデルは CNN を用いた画像分類モデルを採用する．本データセットの場合，牙の有無によってオス・メスの判別ができるため，画像上で外見を考慮した判別ができる

Male:オス**Female:メス****Unclear:不明瞭**

図 3.4.1 Identification データセット

CNN は有効だと考えられる．ただし，監視カメラ映像ではゾウの牙の位置が確認できない状況が度々発生する．たとえば，ゾウの正面がカメラと反対側に向いてしまうと，もはや牙の位置が確認できない．他に，ゾウがプールに潜る，飼育場の仕切りに隠れるといった場合もある．このような画像に対してはトラッキングの精度が低下することが予想された．そこで，牙の見えない画像を明示的に区別するため，牙の見えない画像を 1 クラスにまとめ，「不明瞭」としてクラスを分けた．すなわち，モデルは Male（オス），Female（メス），Unclear（不明瞭）の 3 種類の判別をおこなう（図 3.4.1）．ここで「Unclear」は，ゾウがカメラ映像から牙の有無が確認できない場合につけるラベルとするが，ゾウが 2 頭映っていて，どちらも牙が確認できないときは，どちらも Unclear としている．Unclear は，Tracking におけるネットワークの構成と軌跡に対する識別 ID の割り当てに用いる．

3.4.4 Tracking

ここでは Detection と Identification の結果から、トラッキングをおこなう方法について述べる。基本的な流れは以下のとおりである。まず、検出位置および Identification の判別スコアからネットワークを構成する。このネットワーク上のフローは各個体の軌跡に対応する。次に、ネットワークから最小費用流を求め、それを各個体の軌跡とする。次に、個体同士が近接する場合、軌跡を分割する。最後に、軌跡ごとに識別 ID を割り当てる。

図 3.4.2 にオス・メスの 2 体のゾウに本手法を適用した処理イメージを示す。 d_i^n は、各フレームの Detection と Identification の推定結果を表しており、ここでは「検出ノード」と呼ぶ。Tracking では、すべての検出ノードに、メス (Female)、オス (Male) および誤検出のいずれかを割り当てる (図 3.4.2, 1 段目)。メスまたはオスに割り当てられた検出ノードの集合は、それぞれメスの軌跡 (図中の軌跡 a)、オスの軌跡 (図中の軌跡 b) を表す。ネットワーク上の最小費用流によって求められるのは、同一個体を表す軌跡である (図 3.4.2, 2 段目)。この時点では、この軌跡上の検出ノードにオス・メスは割り当てられていない。さらに、個体同士がすれ違う場合、軌跡の入れ替わりが発生する可能性がある。この入れ替わりを抑制するために、互いのゾウが近接する検出ノードを取り除き、各軌跡を 2 分割する (図 3.4.2, 3 段目)。その後、各軌跡に対し、Identification のスコアを利用してクラス判別をおこなう (図 3.4.2, 4 段目)。途中取り除いた検出ノードを割り当てる場合は、それぞれの検出ノードの Identification のスコアで判別する。

以下、「ネットワークの構成」「最小費用流問題の求解」「軌跡の分割」「識別 ID の割り当て」について、それぞれ詳述する。

■**ネットワークの構成** 検出された物体と誤検出を含めたすべての組み合わせを表すネットワークを構成する。 N を総フレーム数、 n ($1 \leq n \leq N$) をフレーム番号、 $M(n)$ を n フレーム目の検出個体数 (すなわち、Detection が出力する矩形の個数)、 K (≥ 2) を検出対象動物の最大個体数とする。各フレームの Detection と Identification の推定結果をもつ検出ノード d_i^n ($1 \leq i \leq M(n)$) を導入する。ここで、 d_i^n は Detection

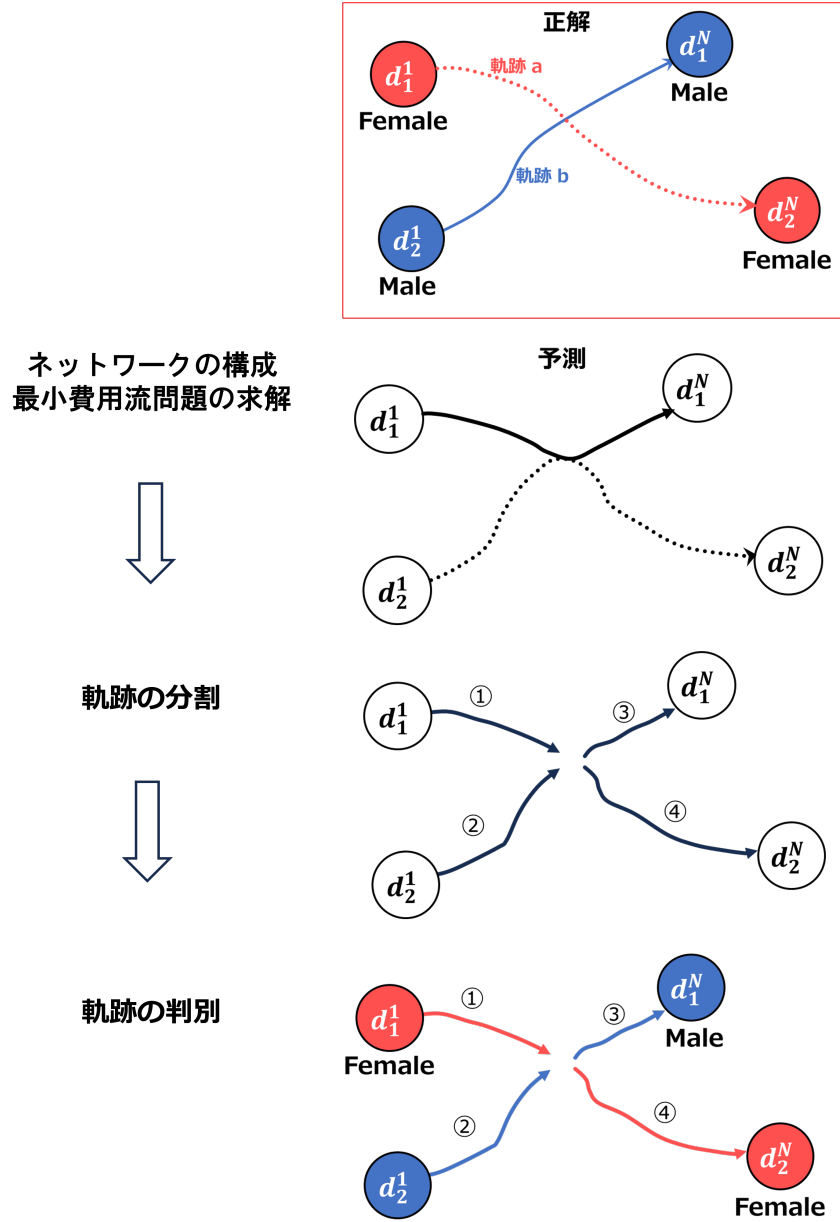


図 3.4.2 Tracking 処理フロー

出力の予測位置 $\mathbf{r}_i = (r_i^{\text{top}}, r_i^{\text{left}}, r_i^{\text{bottom}}, r_i^{\text{right}})$, および, Identification の出力判別スコア $\mathbf{p}_i \in \mathbb{R}^{K+1}$ (個体ごとの推定値 K 次元, および, 不明瞭ラベルに対する推定値 1 次元) で構成されているとする. また, いずれかの個体が検出漏れしていることを表す $K-1$ 個のオクルージョンノード o^n を各フレームに導入する. ただし, 検出ノードが存在しないフレーム (すなわち, 対象物が検出されていないフレーム) は除いておく. よって, 各フレームには $M(n)$ 個の検出ノードと $K-1$ 個のオクルージョン

ノードがある．なお，本稿で扱うデータは $K = 2$ の場合に当たるため，各フレームに対して 1 つのオクルージョンノードが存在する．

次に，各ノードは前後に最も近い時刻のフレームの全てのノードとエッジで接続される．また， $\mathbf{d}_i^n$, $\mathbf{d}_j^m$ を検出ノードとしたとき，エッジ $\mathbf{d}_i^n \cdot \mathbf{d}_j^m$ に対するコスト関数 $c_{\mathbf{d}_i^n \cdot \mathbf{d}_j^m} = \text{cost}(\mathbf{d}_i^n, \mathbf{d}_j^m)$ を以下のように定義する．

$$\text{cost}(\mathbf{d}_i^n, \mathbf{d}_j^m) = \lambda c_{\text{IoU}}(\mathbf{r}_i, \mathbf{r}_j) + (1 - \lambda) c_{\text{class}}(\mathbf{p}_i, \mathbf{p}_j), \quad (3.9)$$

$$c_{\text{IoU}}(\mathbf{r}_i, \mathbf{r}_j) = 1 - \text{IoU}(\mathbf{r}_i, \mathbf{r}_j), \quad (3.10)$$

$$c_{\text{class}}(\mathbf{p}_i, \mathbf{p}_j) = 1 - \exp(-\mu \cdot KL), \quad (3.11)$$

$$KL(\mathbf{p}_i, \mathbf{p}_j) = \sum_{k=0}^K p_i^k \log \frac{p_i^k}{p_j^k}, \quad (3.12)$$

ここで， c_{IoU} は物体の位置の差に対する費用であり， $\text{IoU}(\mathbf{r}_i, \mathbf{r}_j)$ は $\mathbf{r}_i$, $\mathbf{r}_j$ が表す矩形領域の IoU (Intersection Over Union) 値である．ここで IoU とは，2 矩形の共通部分の面積を和集合の面積で割った値のことである．また， $c_{\text{class}}(\mathbf{p}_i, \mathbf{p}_j)$ はクラススコアの差に対する費用であり，同じ軌跡の中でクラスが変わることに対してペナルティを与える． μ はクラス数に応じて設定し， $0 \leq c_{\text{class}} \leq 1$ となるようにする．式 (3.9) の λ は， c_{IoU} と c_{class} の重要度を表すパラメータであり，実験によって決定する．オクルージョンノードに接続する場合は一定値のコスト c_0 を与える．

■最小費用流問題の求解 V を始点 s ，終点 t ，検出ノードおよびオクルージョンノードからなる集合， vw ($v, w \in V$) をエッジ， x_{vw} をフローとする．ネットワーク上のフロー x_{vw} は 0 または 1 の値をとる．ここで， $x_{vw} = 0$ は v, w が異なる識別 ID であること，また， $x_{vw} = 1$ は v, w が同じ識別 ID であることを意味するものとする．フロー x_{vw} は複数の実行可能解が取りうるが，その中で総コストが最小となる最小費用流を求める．すなわち，以下の目的関数 J が最小となるフロー x_{vw} を求める．

$$J = \sum_{v,w} c_{vw} x_{vw}, \quad v, w \in V, \quad (3.13)$$

ただし，容量条件，流量保存条件，非重複条件を満たす必要がある．非重複条件はトラッキングの定式化に必要であり，各ノードの始点および終点となるエッジはそれぞれ多くとも 1 本のみであることを表す [33]．式から分かるように，フローは線形計画

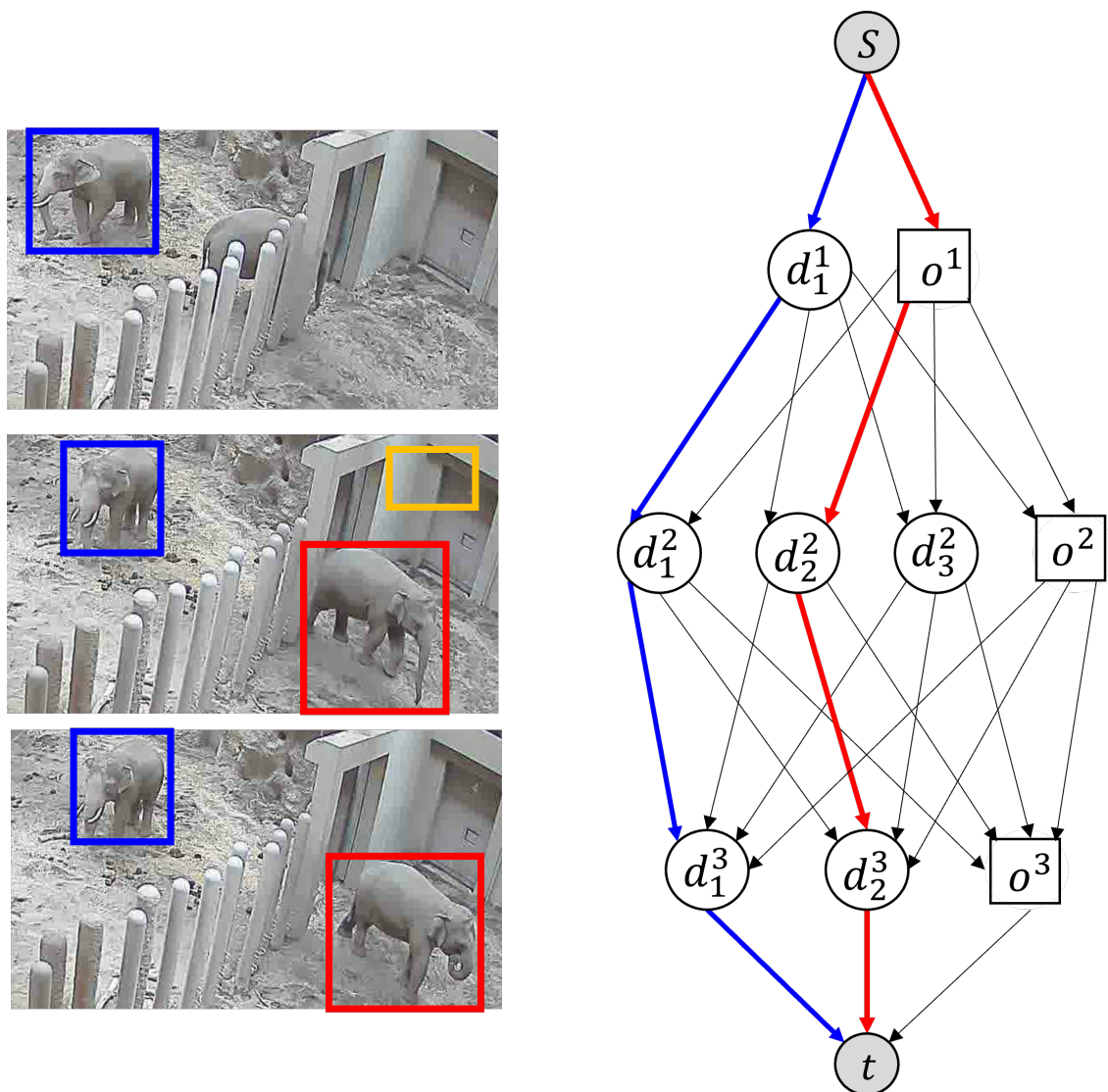


図 3.4.3 ネットワークフロー

問題としても求めることが可能である．実際には，フロー x_{vw} が 0 または 1 であるから，0-1 整数計画問題となるが，変数の数が少ないため短時間で解くことが可能である．

例を図 3.4.3 に示した．この例では，3 フレームにおいて生成されるネットワークとそのフローを示す．丸型のノードが検出ノード，四角型のノードがオクルージョンを表すノードである．図では，パス $s \cdot d_1^1 \cdot d_1^2 \cdot d_1^3 \cdot t$ ，および，パス $s \cdot o^1 \cdot d_2^2 \cdot d_2^3 \cdot t$ が $x_{vw} = 1$ となり，それ以外が $x_{vw} = 0$ となる．

■**軌跡の分割** ID スイッチ（複数の個体の軌跡に割り当てた ID の入れ替わり）はゾウのすれ違い時に発生しやすく、ID スイッチの発生後はそれ以降誤った識別子と対応することになる。そのため、2 頭それぞれに対し検出された矩形の IoU があらかじめ定めた閾値 τ_{div} 以上となる区間の前後で軌跡を分割する。

■**識別 ID の割り当て** ここでは、分割された各軌跡に識別 ID を割り当てる。まず、各軌跡に対して Identification モデルの出力スコアの平均値を求める。ただし、平均値は Unclear と判定されたインスタンスを含めずに計算する。次に、その値が Unclear ラベル以外で最も高い識別 ID を割り当てる。ただし、同一時刻における軌跡は同じクラスが 2 つ以上にならないように識別 ID を割り当てる。加えて、検出矩形の IoU が τ_{div} 以上となる区間の各インスタンスに、Identification モデルの出力スコアが高いクラスから順に識別 ID を割り当てる。

3.5 実験

本節では、3.4 節で述べた手法の有効性を示す実験について述べる。なお本実験では、学習データとして、対象物の矩形位置のアノテーションと識別対象物の画像および個体識別不能な画像（「不明瞭」画像）が必要である。

3.5.1 Detection

■**データセット** 実験で用いたカメラ映像は 3.2 節で述べたとおりであるが、ここでは学習と評価に用いるデータセットについて述べる。表 3.5.1 にデータセットの内訳を示す。学習データは 2022 年 1 月 1 日から 2022 年 5 月 31 日の監視カメラ映像を用いており、1 月 1 日から 3 月 31 日を訓練データ、4 月 1 日から 4 月 30 日をバリデーションデータ、5 月 1 日から 5 月 31 日をテストデータの区間としている。データセットは上記期間の動画から一部のフレームを画像として取り出したものを使用する。アノテーションとして、各ゾウの位置を矩形 (4 点) で指定する。外見のバリエーションが増えるように、動画からフレームを取り出す間隔を 1 分以上あけた。また、睡眠時などのほとんど動きがない場合、その動作の始めのフレームのみをアノテーションし、

それ以降は使用しなかった。昼間の画像（カラー画像）は 70.1%，夜間の画像（グレースケール画像）は 29.1% であった。

表 3.5.1 Detection データセット

	フレーム数	オブジェクト数
training	2270	4032
validation	771	1232
test	740	965

■**学習モデルとパラメータ** 学習モデルは CNN ベースの物体検出モデルを用いて、ファインチューニングをおこなう。検出対象ラベルは「elephant」とし、1 クラスのみを検出するモデルを作成する。モデルは EfficientDet, YOLOv5, MegaDetector(v5.0) を用いて精度を比較する。EfficientDet および YOLOv5 に関しては、COCO [41] で事前学習済みのモデルを用いて、ファインチューニングをおこなう。また、MegaDetector は野生動物を主なターゲットとした物体検出器だが、多様な背景や学習データに含まれない動物の種にも対応できるように訓練されている。モデルはカメラトラップ映像を含めたデータセットで事前学習されており、実験ではこの事前学習済みのモデルを用いてファインチューニングをおこなう。また、画像拡張処理として、左右反転、ランダムクロップ、拡大縮小をランダムに適用した。YOLOv5-x6, MegaDetector はバッチサイズ 2, それ以外のバッチサイズは 4 とした。

■**評価方法** モデルの精度比較にはテストデータに対する $AP(@IoU=0.50:0.95)$ を用いる。これは、COCO の評価指標値であり、IoU 閾値 0.5 から 0.95 まで 0.05 刻みで変えたときの平均 AP (average precision) である。

表 3.5.2 Detection 結果

モデル	入力解像度	AP @IoU=0.50:0.95	AP @IoU=0.50	AP[昼間] @IoU=0.50:0.95	AP[夜間] @IoU=0.50:0.95
EfficientDet-D1	640	0.605	0.868	0.668	0.526
YOLOv5-n	640	0.743	0.917	0.781	0.696
YOLOv5-s	640	0.764	0.971	0.792	0.731
YOLOv5-x	640	0.794	0.966	0.825	0.754
EfficientDet-D5	1280	0.769	0.957	0.804	0.719
EfficientDet-D6	1280	0.779	0.960	0.815	0.734
YOLOv5-m6	1280	0.782	0.961	0.817	0.736
YOLOv5-l6	1280	0.810	0.982	0.839	0.774
YOLOv5-x6	1280	0.805	0.974	0.839	0.759
MegaDetector v5.0	1280	0.633	0.795	0.723	0.534
MegaDetector v5.0 (fine tuning)	1280	0.804	0.971	0.837	0.764

表 3.5.3 Identification データセット

	Male	Female	UnClear	計
training	969	1216	971	3156
validation	331	348	490	1169
test	274	225	418	917

■**結果** 結果を表 3.5.2 に示す．最も良いモデルは，YOLOv5-l6 となり， $AP(@IoU=0.50:0.95)$ が 0.810 となった．Detection の結果は Tracking の精度に直結するため，より精度が良い方が好ましいが，動物の飼育環境によっては学習が不要となる可能性もある．

3.5.2 Identification

■**実験方法と結果** クラスは Male (オス), Female (メス), Unclear (不明瞭) とし，Detection において用いたデータセットの画像から，アノテーションの矩形位置を切り抜いて作成した．内訳は表 3.5.3 のとおりである．オス，メスの判別は牙の有無で判断し，牙の位置が見えない画像には Unclear ラベルを付けた．ただし，ゾウが牙の位置が見える位置にいたとしても，ゾウの画像上のサイズが小さく，オス・メスの判別に迷う画像は取り除いた．加えて，データセット全体において昼間の画像（カラー画像）は 66.8%，夜間の画像（グレースケール画像）は 33.2% であった．学習モデルとしては MobileNetV2 [53], ResNet50 [25], EfficientNet を用いて比較した．EfficientNet に限り入力サイズの異なるモデル (B0-B5) を比較した．すべてのモデルは ImageNet で事前学習されているため，最終層のノード数を 3 に変更して全層ファインチューニングをおこなった．最も良いモデルは EfficientNet-B2 となり，正答率は 0.869 となった．また，表 3.5.5 に EfficientNet-B2 の混同行列を示す．Male の正答率が最も高く，Female と Unclear がそれと比較するとやや低い．これは，Female と Unclear の見た目の違いが小さいことを示していると考えられる．

表 3.5.4 Re-ID 結果

モデル	入力解像度	正答率
MobileNet-v2	160	0.845
Resnet50	224	0.865
EfficientNet-B0	224	0.865
EfficientNet-B1	240	0.852
EfficientNet-B2	260	0.869
EfficientNet-B3	300	0.869
EfficientNet-B4	380	0.856
EfficientNet-B5	456	0.866

表 3.5.5 Re-ID EfficientNet B2 混同行列結果

		予測ラベル			再現率
		Female	Male	UnClear	
真のラベル	Female	195	9	21	0.867
	Male	11	256	7	0.934
	UnClear	51	21	346	0.828
	適合率	0.759	0.900	0.933	

3.5.3 Tracking

本節では, 3.5.1, 3.5.2 節で学習したモデルを適用した結果を用いて Tracking を行う. また, Detection モデルとして YOLOv5-l6, Identification モデルとして EfficientNet B2 を用いた. 軌跡分割時の閾値 τ_{div} は, 本文で言及しない場合 $\tau_{\text{div}} = 0.2$ と設定した.

■**データセット** テストデータとして 2022 年 5 月 1 日 のカメラ映像を用いた. 6 秒おきにゾウの矩形位置および識別 ID (メス・オス) のアノテーションをおこない, 結果として, データセットは表 3.5.6 のようになった. データセットを作成する際は, 元動画からゾウが映っていない時間, 30 分以上の睡眠時間, 肉眼で判別しづらい時間を

除いた。結果として、作成した動画はそれぞれ1分間から48分間の分割されたものとなった。また、データセットは特定の時間帯に偏らないように作成しており、データセット全体において昼間のフレーム（カラー）は55.5%、夜間のフレーム（グレースケール）は44.5%であった。

本データセットでは、多人数トラッキングのコンペティションである MOTChallenge [15] と比べると2点の違いがある。まず、MOTChallenge には個体を識別する ID が存在せず、トラッキング対象の照合をおこなう必要はないという点である。一方、本データセットではトラッキングと同時に事前に定められた識別 ID（オス・メス）との照合をおこなう必要がある。次は、本データセットは長時間のデータセットとなっている点である。ここでは「長時間」を実時間（実際の動画撮影時間）の意味で用いている。本稿で扱うような識別 ID も同時に推定する場合、推定区間内の初期に ID スイッチが生じると、それ以降の識別 ID がすべて入れ替わるということが起きうる。また、推定区間でゾウの総移動量が大きくなるため、ID スイッチが起こるようなゾウの重なりも増える傾向にある。長時間のデータを用いることで、そのような識別 ID の入れ替わりも評価することができる。MOTChallenge には複数のデータセットがあるが、例えば、MOT17 のテストデータの実時間は15秒から85秒であり、合計20本の動画（合計動画時間744秒、17,757フレーム、オブジェクト数564,228）で構成されている。本データセットは、MOT に比べてオブジェクト数とフレーム数は少ないが、合計動画時間が9.2時間となっており長時間のデータであることが特徴である。そのため、本データセットは実際のアプリケーションで想定される状況に近く、研究段階から実用化に移る際の指標となる。

表 3.5.6 トラッキングデータセット

フレーム数	オブジェクト数	動画時間	合計
5525	8665	9.2 時間	

■実装の詳細および実験環境 最小費用流問題は整数線形計画法問題として定式化したあと、Python の線形計画法ライブラリ PuLP[46] を用いて最適解を求めた。PuLP は内部で COIN-OR[21] の CBC モジュールを呼んでいる。これは、混合整数問題の

ソルバーであり、実装は分枝カット法 [49] に基づいている。本実験では、Intel(R) Core(TM) i9-9960X CPU @ 3.10GHz, RAM 64GB 搭載の PC を用いて実験した場合、1 時間の動画 (約 18000 フレーム) で 3 分以内で求めることが可能だった。

■**評価方法** 評価値はオス・メスごとの AP (@IoU=0.50:0.95) の平均値 (以下, mAP と略記) と, HOTA [43], MOTA, IDF1 の値を記録した。HOTA, MOTA, IDF1 は, 多人数トラッキングのコンペティションである MOTChallenge [15] の評価値として採用されており, 軌跡が途中で入れ替わる ID スイッチが評価値に反映される。以下, 図 3.5.1, 3.5.2, 3.5.3 を用いて各評価指標を説明する。gtDets, prDets を, それぞれ Detection の正解データと予測データ (bounding-box), gtIDs, prIDs を識別 ID の正解データと予測データとする。このとき, gtDet と prDet の類似度 S が定義されているとする。まず, MOTA, IDF1, HOTA を計算する際は gtDets と prDets の一対一の対応をおこなう。gtDets は最大でも 1 つの prDets と対応させ, 過検出か未検出の場合はペナルティを課す。ハンガリー法によって類似度を用いたマッチングスコアが最適になるようにマッチングする。

- TP (True Positive) = 正しくマッチしたもの
- FN (False Negative) = 未検出の gtDets の数 (prDets に対応するものがない gtDets の数)
- FP (False Positive) = 過検出の prDets の数 (gtDets に対応するものがない prDets の数)

とする。MOTA は以下のように定義される (図 3.5.1) 。

$$MOTA = 1 - \frac{FN + FP + IDSW}{|gtDet|}.$$

ここで, IDSW は軌跡が入れ替わった回数である。図 3.5.1 の例では, $FN = 1, FP = 5, IDSW = 1, |gtDet| = 11$ より,

$$MOTA = 1 - \frac{1 + 5 + 1}{11} \approx 0.364.$$

となる。MOTA は単純な計算で ID スイッチを加味できる評価値だが, Detection の精度を重視する傾向にある [43]。

つぎに, IDF1 を説明する (図 3.5.2)。まず, ID に着目し,

- IDTP = 正しく ID がマッチしたもの
- IDFN = 未検出の gtIDs の数
- IDFP = 誤判定の prIDs の数

とする。IDF1 は以下のように定義される。

$$\text{IDF1} = \frac{|\text{IDTP}|}{|\text{IDTP}| + 0.5|\text{IDFN}| + 0.5|\text{IDFP}|}$$

IDF1 は ID の割り当てを重視する傾向にある [43]。

最後に HOTA を説明する (図 3.5.3)。まず, TP (Detection が正しいもの) のうち 1 つ (c) に対して以下 3 つの値を計算する:

$$\begin{aligned} \text{TPA}(c) &= \{k \in \text{TPs} | \text{prID}(k) = \text{prID}(c), \text{gtID}(k) = \text{gtID}(c)\}, \\ \text{FNA}(c) &= \{k \in \text{TPs} | \text{prID}(k) \neq \text{prID}(c), \text{gtID}(k) = \text{gtID}(c)\} \\ &\quad \cup \{k \in \text{FNs} | \text{gtID}(k) = \text{gtID}(c)\}, \\ \text{FPA}(c) &= \{k \in \text{TPs} | \text{prID}(k) = \text{prID}(c), \text{gtID}(k) \neq \text{gtID}(c)\} \\ &\quad \cup \{k \in \text{FPs} | \text{gtID}(k) = \text{gtID}(c)\} \end{aligned}$$

さらに, Detection の閾値 α ごとに以下を定義する:

$$\mathcal{A}(c) = \frac{|\text{TPA}(c)|}{|\text{TPA}(c)| + |\text{FNA}(c)| + |\text{FPA}(c)|}$$

,

$$\text{HOTA}_\alpha = \sqrt{\frac{\sum_{c \in \text{TPs}} \mathcal{A}(c)}{\text{TP} + \text{FN} + \text{FP}}}.$$

HOTA は以下で定義される:

$$\text{HOTA} = \int_0^1 \text{HOTA}_\alpha d\alpha$$

積分は実際には, 0.05 から 0.95 まで 0.05 刻みで足し合わされる。HOTA は MOTA と IDF1 の欠点を補うものとして考案されている。HOTA, MOTA, IDF1 は途中で ID スイッチが起きた時は評価値が低下するが, 推論区間の最初から最後までオス・メスが互いに入れ替わっていても評価値は低下しない。行動記録では個体の区別が必須であるため, モデルの良さの指標としては個体の識別性能を表す mAP を最も優先する。

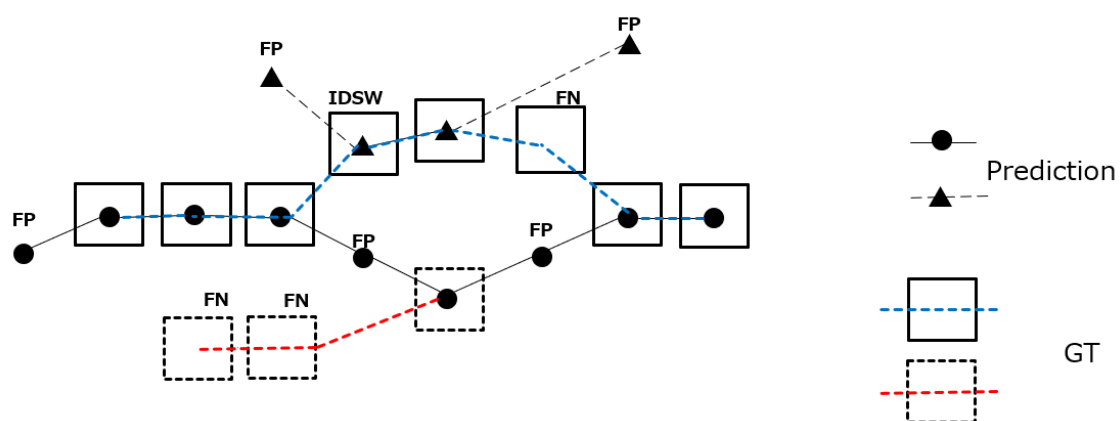


図 3.5.1 MOTA

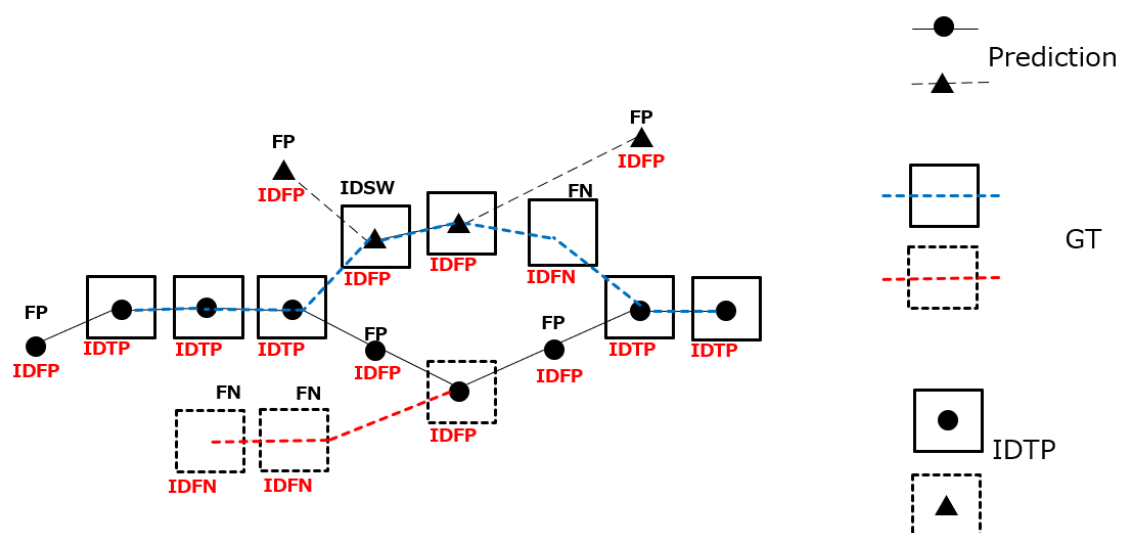


図 3.5.2 IDF1

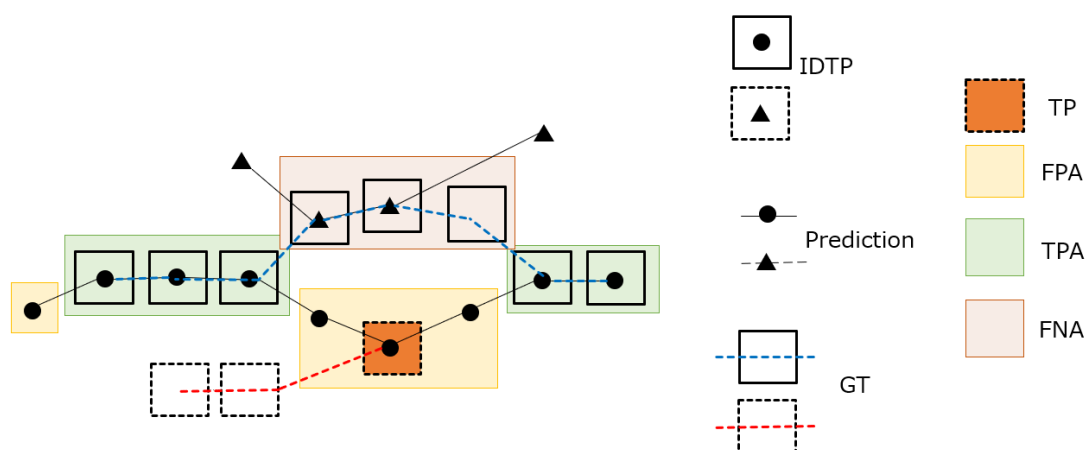


図 3.5.3 HOTA

■**コスト関数** 3.4.4 節で述べたコスト関数の、 $\lambda = 1, 0.5, 0$ の場合を比較する．すなわち、以下の3種類のコスト関数を比較する．

1. (IoU) $\text{cost}(v, w) = c_{\text{IoU}}$,
2. (Class Score) $\text{cost}(v, w) = c_{\text{class}}$,
3. (IoU + Class Score) $\text{cost}(v, w) = (c_{\text{IoU}} + c_{\text{class}})/2$.

ここで、クラススコアの違いに対する費用 c_{class} (式 (3.11)) のパラメータ μ を、 $0 \leq c_{\text{class}} \leq 1$ となるように $\mu = 1.3358$ と定めた．また、オクルージョンノードに接続するコストは $c_0 = 0.9$ とした．

■**結果** 結果を表 3.5.7 に示す．比較のため、Detection と Identification の結果のみを用いて評価をおこなった値 (Baseline) も示した．また、「NF」はネットワークフローによるトラッキングを表し、括弧内はコスト関数の種類を表す．まず mAP に着目すると、NF (IoU) は Baseline よりも大幅に精度が高くなり、mAP は 0.603 から 0.720 に向上した．一方で、NF (Class Score) は Baseline より、mAP は 0.603 から 0.473 に大きく減少した．また、NF (IoU + Class Score) は NF (IoU) よりも mAP は低く 0.691 に留まっている．これから、精度向上には、隣接フレーム間の位置の差が外見の差よりも重要であることがわかった．HOTA, MOTA, IDF1 に着目すると、ネットワークフローを用いたモデルは BeseLine よりも総じて評価値が向上している．これから ID スイッチの抑制には IoU と Class Score のコスト関数のどちらも効果があり、両コスト関数を合わせるとさらに効果的であることがわかる．

3.5.4 Detection と Identification の統合と分離

Detection と Identification を統合するモデル (以下、統合モデル) と分離するモデル (以下、分離モデル) の比較をおこなった．ここで統合モデルとは、Detection モデルが、単に elephant を検出するのではなく、Female, Male, Unclear の3種を直接別々に検出することを指す．分離モデルは Detection と Identification を分けたモデルのことを指す．データセットは 3.5.1 節の Detection データセットの一部を利用して作成したが、統合モデルにおけるデータセットには、各オブジェクトに対して矩形

位置と識別 ID が付加されている。結果として, training 1135 枚, validation 461 枚, test 324 枚となった。

結果を表 3.5.8 に示す。Detection と Identification を分離した場合（分離モデル）は, Detection と Identification を統合した場合（統合モデル）と比較して, わずかながら精度は向上した。

3.5.5 不明瞭ラベルの効果

不明瞭ラベルの効果を確認するため, 不明瞭ラベルを用いない Identification モデルによるトラッキングをおこなった。Identification モデル (EfficientNetB2) を学習する際は, 表 3.5.3 のデータセットから Unclear ラベルを除いて学習・評価をおこなった。Identification モデルの test データに対する精度は 94.2% となった。

結果を図 3.5.4, 図 3.5.5 に示す。不明瞭ラベルの有無によって Identification のスコアが大きく変わるため, 最適なパラメータが異なる可能性がある。よって, 不明瞭ラベルの精度比較に加え, 複数のパラメータ値で精度を比較している。

図 3.5.4 は, 軌跡分割時に基準とする IoU 値 τ_{div} を 0.1 から 0.5 まで 0.1 刻みで増加させたときの mAP を表す。図 3.5.5 は, オクルージョンノードに接続するコスト c_0 を 0.5 から 1.0 まで 0.1 刻みで増加させたときの mAP を表す。「不明瞭ラベルあり」は「不明瞭ラベルなし」と比較すると, 小幅ながら $c_0 = 1.0$ を除くほぼすべての閾値で上回った。

表 3.5.7 Tracking 結果

Model	NF	IoU	Class Score	mAP	HOTA	MOTA	IDF1
BaseLine		-	-	0.603	0.565	0.601	0.687
NF (IoU)	✓	✓		0.720	0.805	0.909	0.917
NF (Class Score)	✓		✓	0.473	0.627	0.817	0.693
NF (IoU+Class Score)	✓	✓	✓	0.691	0.826	0.910	0.948

表 3.5.8 統合・分離モデル比較

Model	mAP
統合モデル	0.656
分離モデル	0.664

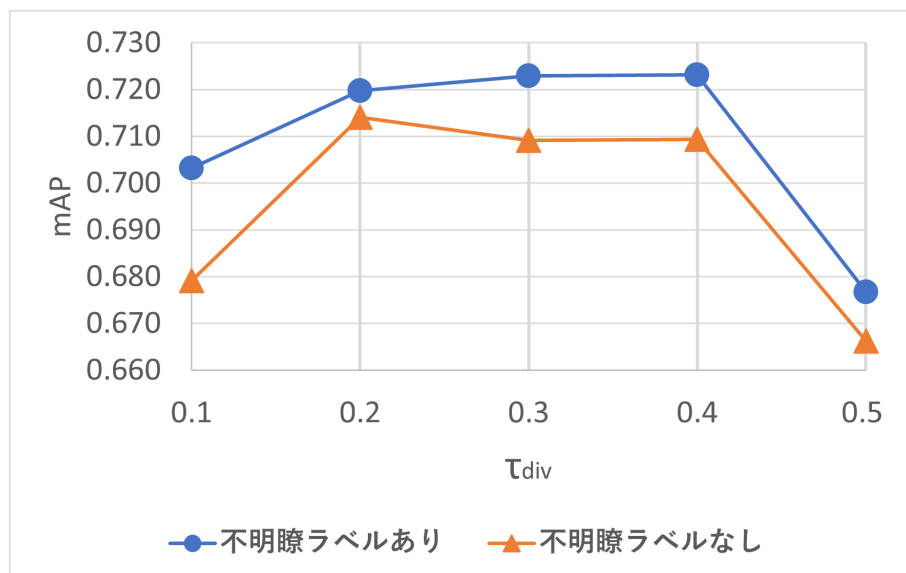


図 3.5.4 不明瞭ラベルの効果 (軌跡分割基準)

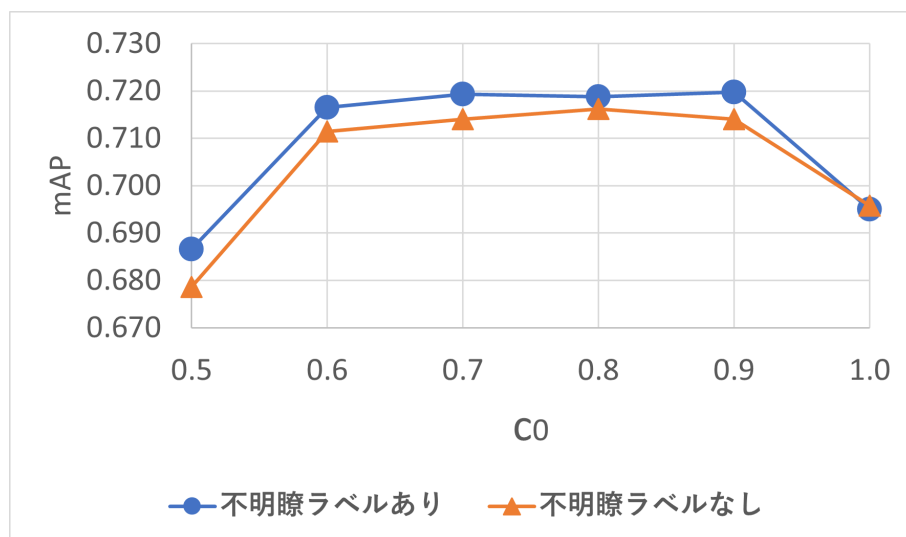


図 3.5.5 不明瞭ラベルの効果 (オクルージョンコスト)

3.6 考察

ここでは、本研究の考察を述べる。

動物のトラッキングの特性の一つとして、各個体の外見が非常に近いことが挙げられる。この場合、単独のフレーム画像から個体の判別ができないことも多い。これは、表 3.5.7 において、「Baseline」の結果が他のモデルよりも著しく精度が低かったことからわかる。そのため、トラッキングの軌跡を求めるには、各個体のフレーム間の相対的な変化（例えば、位置や外見の差）を考慮することが重要になる。本稿の実験により、特に位置情報が重要であるとわかった。これは、表 3.5.7 の NF (IoU) が NF (Class Score) と比べて精度が高かったことからわかる。また、外見を学習モデルに組み込む工夫として「不明瞭」ラベルを付加的に用いたが、それによって各軌跡の判別を実現できることを示した。

■時刻ごとの結果 図 3.6.1 にテストデータにおける時刻ごとの結果を示す。ここで時刻 N の値は、 N 時 00 分から $(N + 1)$ 時 00 分までの HOTA の平均値としている。ここで、昼間（午前 5 時 20 分から午後 6 時 20 分）の平均 HOTA は、0.814、夜間（昼間以外の時間帯）の平均 HOTA は、0.723 となり、夜間の精度が昼間に比べて大きく精度が悪化することが分かる。これは、3.2 章「昼と夜の違い」において述べたように、夜間はグレースケールであることやブレが影響していると考えられる。

また、我々は 2022 年 5 月 1 日動画に対して、クラススコアが 0.5 以上のフレームを出力し、トラッキングが失敗したフレーム数を計測した。ここで、「トラッキングが失敗した」とは「Detection は正しく検出されているが、オス・メスが入れ替わっている」ことを指す。0 時から 24 時の間で Detection によって検出されたフレームは 271,787 フレームであり、この中でトラッキングが失敗したフレームは 12,690 フレーム（約 4.7%）であった。

■精度低下の原因 実際の出力例から誤ってトラッキングされた場合の原因を数点述べる。

(1) 各軌跡上に雌雄判定に有効な画像が少ない

「ゾウがカメラから見切れて映る」「カメラと逆方向に向いている」「カメラから遠い位

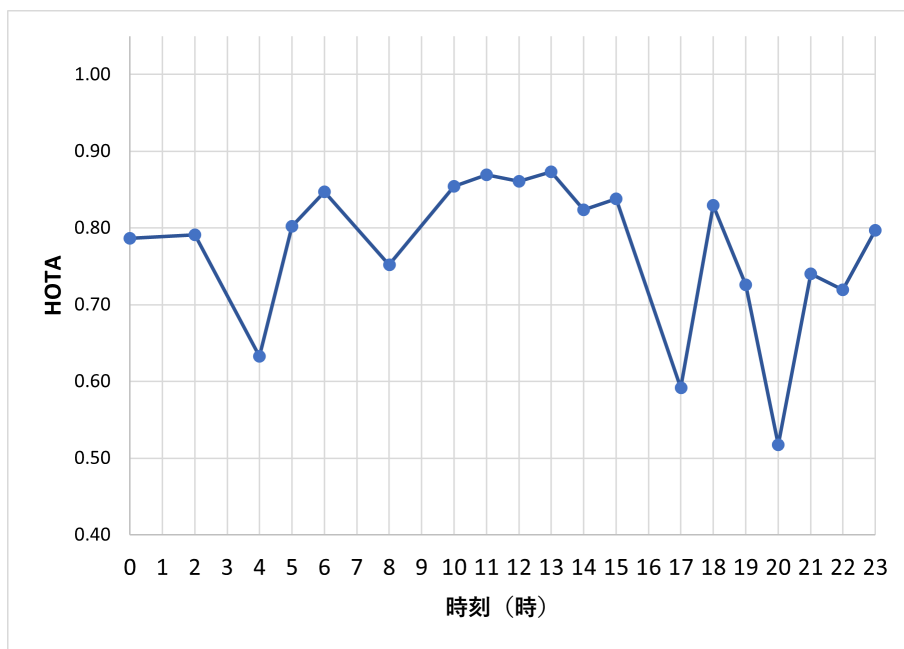


図 3.6.1 時刻ごとの HOTA の結果

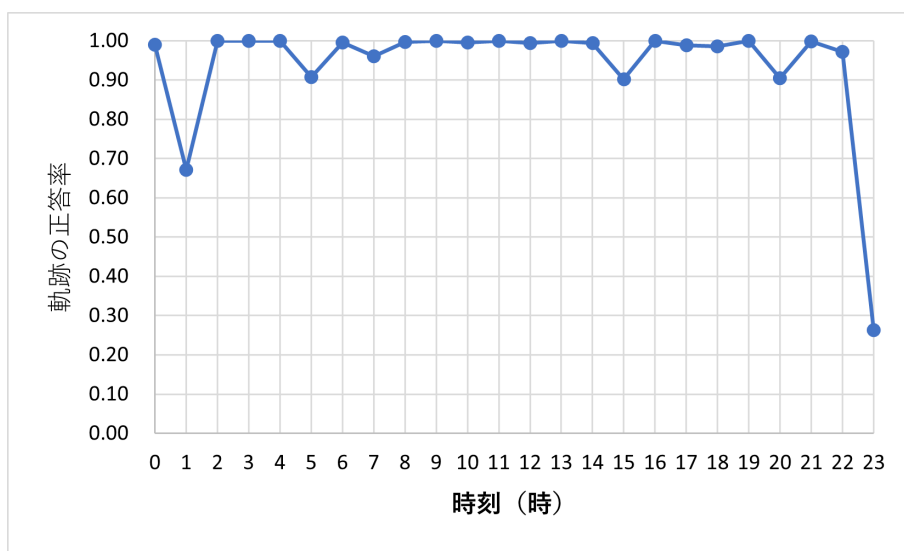


図 3.6.2 時刻ごとの軌跡の正答率

置にいる」等の場合は雌雄の判別に有効な牙が見えづらいことが多く、結果として全体の精度が悪くなる。図 3.6.1 の 17 時は昼間にも関わらずやや精度が悪いが「ゾウがカメラから見切れて映る」時間が大半であった。また、ゾウの睡眠時は「カメラと逆方向に向いている」ことが多く、精度が低下した。ただし、1 時間ごとに映像を区切って推論しているため、判別に有効なフレームが取得できていない場合もあった。この

ため、Detection の検出位置が変化しないフレームをある程度除き、動きのあるフレームまで映像の時間を進めて解析をおこなう、といった対策が考えられる。

(2) Detection が誤検知する

ゾウでない物体（背景の足場や給餌器など）が検出されてしまうと、当然全体の精度に影響する。対策としては、背景の画像を多く追加学習することが考えられる。

(3) 映像にゾウが 1 頭しか映らず排他的に判別できない

本手法では、同一時刻における複数の軌跡には重複しない識別 ID をそれぞれの軌跡に割り当てる。それゆえ、ゾウが 2 頭の場合、一方が判別に難しい外見をしていても、もう一方が容易に判別できれば、正しい識別 ID を割り当てることは可能である。しかし、1 頭しか映っていない場合は、そのような戦略がとれないため比較的精度が悪くなる。対策としては、アノテーションの労力は必要だが、キーポイントを利用した個体識別 ([47]) を用いることで、個体識別の正答率を上げることが考えられる。

3.7 応用-地図へのマッピング

動物の食事や睡眠の場所を知ることができれば、飼育方法の方針の策定や健康チェックに役立てられる。著者は、図 3.7.1 に示すように、飼育場の俯瞰図（地図）に軌跡をマッピングする機能を設けた。図 3.7.1 中の、オレンジがメスのゾウ、青がオスのゾウを表し、矩形はゾウの Bounding box、曲線は各ゾウの軌跡を表している。監視カメラ映像と俯瞰図は縮尺やゆがみの影響から必ずしも一致しないため、以下の手順にしたがって図を作成した。まず、地図上に等間隔に点をプロットする。つぎに、適当な 4 点を選び射影変換によって監視カメラ画像上に対応点をマッピングする。この時点では対応点がずれていてもよい。つぎに、画像編集ソフト等を用いてカメラの歪みを推定し、監視カメラ画像上の点の位置を修正する。修正しきれない部分は手作業で地図と監視カメラ画像上の点を調整する。最後に、対応点をもとに区分的アファイン変換を定義して、監視カメラ画像から地図へのマッピングを実現する。

また、著者は上記の方法により、各個体の地図上の位置をもとに 2 次元ヒストグラムとして表した。図 3.7.2 は、2021 年 11 月 27 日 13 時から 14 時までのヒストグラムであり、明るい箇所がその場にいた時間が長いことを示す。オス・メスともに同じ箇所が明るい、これは吊り上げ式のネットによる干し草がある場所であり、食事のた



図 3.7.1 トラッキング結果と地図へのマッピング

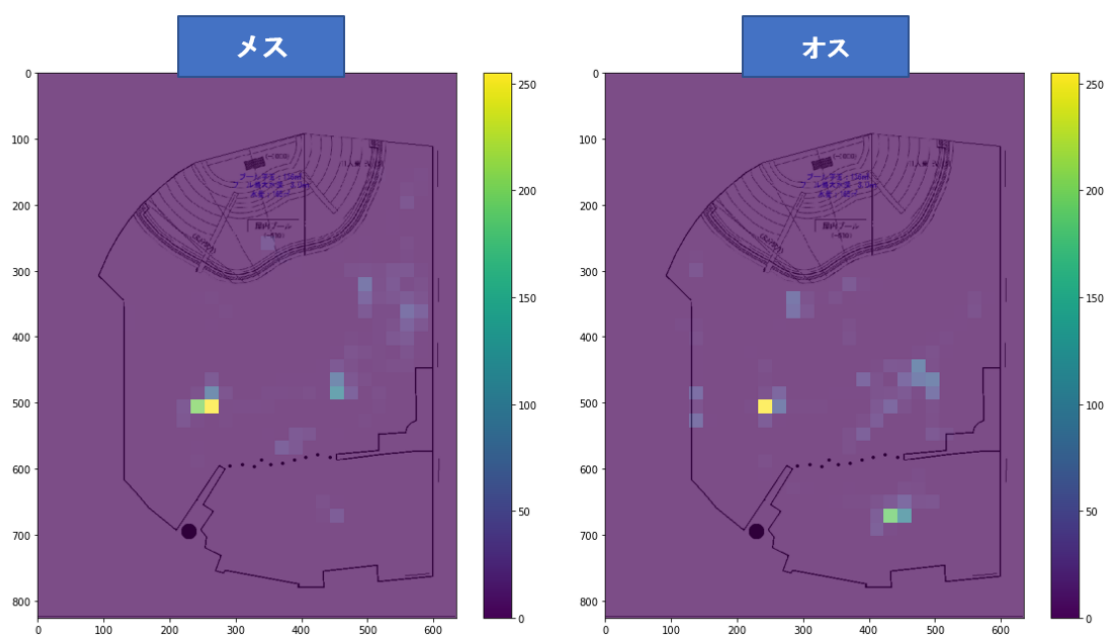


図 3.7.2 位置の 2D ヒストグラム

めに比較的長い時間その場所にいたためである。

3.8 まとめ

著者らは複数の手法を組み合わせ、ゾウを長時間にわたり安定してトラッキングできる手法を提案した。これには、深層学習ベースの物体検出モデル（YOLOv5）と分

類モデル (EfficientNet), また, ネットワークフローを用いている. 監視カメラ映像から Detection, Identification, Tracking 用のデータセットを作成して学習と評価をおこなった. 結果, トラッキング精度は HOTA 0.805 と十分な精度が得られた上, 個体の識別精度は mAP 0.720 となり, 行動記録のベースとなりうる精度を得ることができた.

第 4 章

TAL による食事行動検出

4.1 はじめに

一般に、動物園における行動記録の目的は、動物の生態調査と飼育管理のためである(2.1 節). 行動記録には、対象行動の頻度と持続時間を正確に記録する必要がある. ただし、監視カメラ映像から動物の行動を確認する場合、カメラの向きやオクルージョンによって対象個体が確認しづらいケースも多い. そのような場合でも、人は前後の映像を確認すれば行動の継続性を判断することができる. これは、行動の判別にはある程度まとまった時間を観察する必要があることを意味する.

本章では、ゾウのトラッキング情報をもとにゾウの行動の自動検出手法を提案する. 本手法は、数十秒から数分の映像の特徴量を抽出することで、カメラから確認しづらい状況でも行動検出を可能にしている. 本研究では、主に temporal action localization (TAL) の手法に基づく. TAL とは、比較的長い動画に対して対象の行動がいつ出現したのかを予測する手法の総称である. TAL の手法は人に対しての研究がほとんどであり、動物に対する研究はいままで無かった. 既存の動物の自動行動認識に関する研究では、10 秒程度の短時間の動画に対する手法に基づいており、動物の瞬間的な動きに敏感に反応してしまう. 本研究では、TAL が動物の行動検出にも有効であると示す.

また、本研究では、事前に定めたゾウの食事行動のみに着目し、その検出を試みた. 「歩く」や「寝る」といった基本的な動作を検出することもゾウの健康状態を確認する上で重要であるが、トラッキング情報によりゾウの 1 日の動きを観察できること、また睡眠は他の行動と比較して外見が特徴的であることから、これらの検出は容易であ

と考えた。それに対して、ゾウの行動が「食べる」と判断するのは、ある程度まとまった時間を観察する必要がある。

本章の構成について説明する。まず、4.2 節で典型的な行動検出手法について述べる。つぎに、4.3 節では、本研究で取り扱う行動データセットについて述べ、特定の食事行動に制限した理由についても説明する。4.4 節では提案手法について述べ、4.5 節および 4.6 節で実験について述べる。4.7 節で議論をおこない、4.8 節で本章のまとめを述べる。

4.2 典型的な行動検出手法

本節では、まず TAC の代表的な手法である Inflated 3D ConvNet(I3D) [11] と Temporal Segment Network(TSN) [73] の処理について述べる。つぎに、TAL の代表的な手法である BSN[38] と BMN[37] について述べる。

4.2.1 Inflated 3D ConvNet(I3D)

■**手法の概要** I3D [11] は、3D CNN を用いて動画判別をおこなうが、事前学習のために ImageNet 等で学習済みの 2D CNN の重みを用いる。ベースネットワークとして、バッチ正規化構造をもつ Inception[63] を用いる。

■**モデルの入力と出力** 入力された画像は 224×224 サイズにクロップされる。フレーム数が少ない動画に関しては、あらかじめ定めたフレーム枚数になるようにフレームをコピーして用いる。また、オプティカルフローの画像も同様にして用いる。

■**学習方法** 3D CNN の畳み込み層では $H_l \times W_l \times C_l \times T_l$ 次元のフィルターを用いる。I3D の重みはデータセットによる学習前に 2D CNN のフィルターの $H_l \times W_l \times C_l$ 次元の重みを T_l 回繰り返しコピーすることで、事前学習の代わりとする。また、データ拡張処理として、短辺 256 ピクセルにリサイズしたあと、ランダムに 224×224 サイズの範囲を切り出す。また、水平方向反転をランダムにおこなう。テスト時には、画像中心から 224×224 の範囲を切り出す。

元論文 [11] の実験では、訓練時には 64 枚のフレーム画像を学習に用いて、テスト時

には250枚のフレーム画像を用いて推定をおこなっている。オプティカルフローを用いる場合は同数のフレーム画像を用いる。

4.2.2 Temporal Segment Network (TSN)

■**手法の概要** TSN [73] は、動画からフレーム画像をサンプリングし、CNNによる画像判別モデルを用いることで動画判別を実現する。モデルは3D CNNと比較すると大幅に計算量を落とすことが出来る。一方で、フレームの順序を考慮して推論できないため、例えば、「ドアが開く」「ドアが閉まる」といった判別は原理的に不可能である。

■**モデルの入力と出力** まず動画（フレーム画像の系列）を等間隔にセグメントに分割する（図4.2.1）。つぎに、各セグメントから1枚の画像をランダムサンプリングする。サンプリングされた画像はスニペット (snippet) と呼ばれる。続いて、スニペット内の各画像に対してクロップやリサイズを施す。ここでは、学習時と推論時でやや異なる方法でおこなわれることが多い。続いて、各画像を2次元CNNからなる学習済みモデルに入力し、ソフトマックス前のスコアを得る。つぎに、スコアをクリップ内で平均などで集約した後、ソフトマックスを取る。

また、TSNはRGBフレーム画像だけでなく、事前に計算したオプティカルフロー画像など複数モーダルの画像を合わせて用いることができる。その場合、モデルを各モーダルの画像で学習しておき、推論時にはRGB画像と同様の処理をおこなった後、出力の最後に複数のモーダル間で平均するだけでよい。

■**ネットワーク構造** TSNに用いるモデルは画像判別モデルであればよいが、元論文 [73] では、バッチ正規化構造をもつInception [32] [63] を用いている。Inceptionは、 1×1 の畳み込み層を用いてチャンネル方向の次元を調整することで、VGG [58] などの以前のモデルよりも大幅に計算量を落とし、なおかつ、精度の向上に成功している。

■**学習方法** 画像判別モデルはImageNetで事前学習された重みを用いる。256x340サイズにリサイズしたあと、データ拡張処理として、ランダムで水平方向の反転と画像のクロップがおこなわれる。クロップ時は、画像の4隅および中央から切り出される。また、クロップサイズは、256, 224, 192, 168ピクセルの中から幅・高さとしてそ

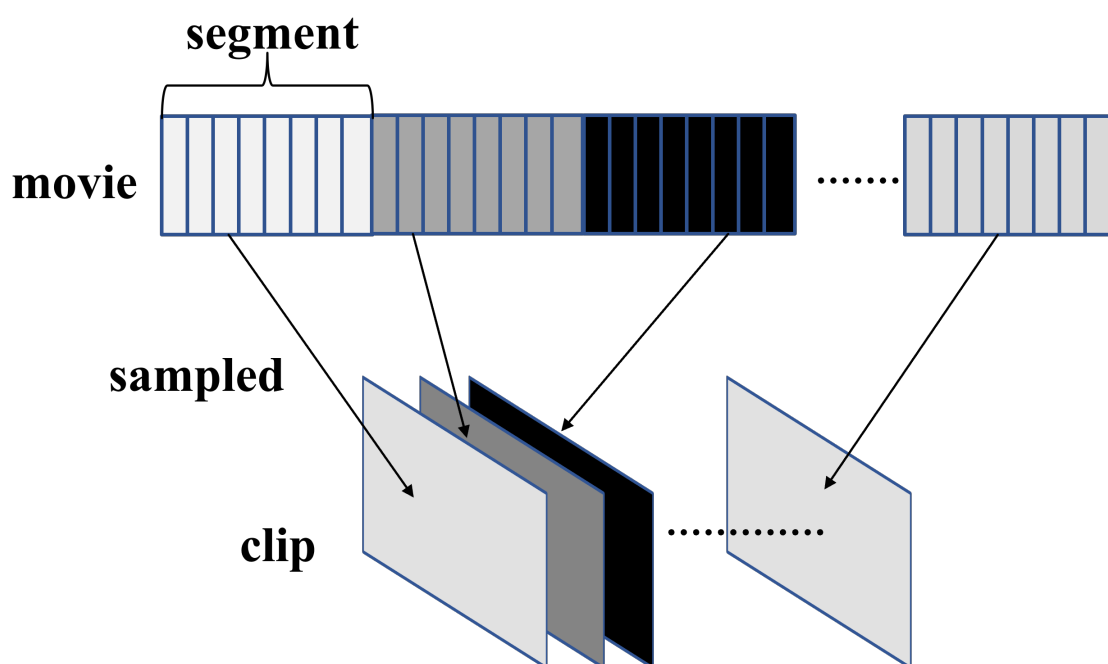


図 4.2.1 TSN 前処理

それぞれランダムで選択される。その後、 224×224 にリサイズされ、モデルに入力される。また、UCF101 のデータセットを用いて学習するには、4 枚の TITANX GPU で 2 時間程度かかる。また、オプティカルフローを用いる場合はさらに 9 時間程度かかる [73]。

4.2.3 Boundary Sensitive Network (BSN)

■**手法の概要** モデルの入力は学習された動画判別モデルの特徴列である。出力はプロポーサルである。処理は動画から学習された動画判別モデルを通して、特徴列を得る。つぎに、特徴列は Temporal evaluation module と呼ばれる 1 次元 CNN からなるモデルに入力される。このモデルは、カーネルサイズ 3、チャンネル数 512 の 1 次元の CNN により畳込みがおこなわれ、最後にシグモイド関数を通して「Actionness (行動中)」「Startingness (開始)」「Endingness (終了)」を表す 3 種の確率を予測する。そのあと、3 種の確率から行動の開始、終了を表すプロポーサルを出力する。また、プロポーサル候補と確率列から開始時刻、終了時刻、その間の周辺の特徴列からサンプリングをおこない、固定長の特徴ベクトルに変換する。これは、Boundary-Sensitive

Proposal (BSP) と呼ばれる．プロポーサルの出力と BSP への変換処理は合わせて Proposal Generation Module と呼ばれる．つぎに，BSP 特徴量のプロポーサルから 2 層の FC 層からなるネットワークで行動のスコアを再計算する．このネットワークは Proposal evaluation module と呼ばれる．最後に，Soft-NMS によってプロポーサルを予測する．推論方法および学習方法については BMN と同様であるため，ここでは省略する．

4.2.4 Boundary Matching Network (BMN)

■**手法の概要** BMN [37] は BSN を元にされた提案手法であり，多くの類似点がある．入力された動画は BSN 同様にまず動画判別モデルによる特徴量抽出がおこなわれる (図 4.2.2)．つぎに，1 次元 CNN からなるモデルによって特徴量の変換がおこなわれる．これは，Base Module と呼ばれる．Base Module の出力は Temporal evaluation module (TEM) と Proposal evaluation module (PEM) に送られる．TEM は，開始時刻と終了時刻の 2 種の確率の系列を出力し，プロポーサル候補を出力する．PEM (Algorithm 1) では，まず，BM 特徴量と呼ばれる開始終了のペアを表す 4 次元の特徴量に変換する (Algorithm 2)．BM 特徴量は 3 次元の CNN と 2 次元の CNN によって畳み込まれ，「行動の開始時刻」と「行動の時間の長さ」を軸にもつ 2 次元信頼度マップを出力する．プロポーサル候補と信頼度マップによって，信頼度付きプロポーサルを得る．最後に，Soft-NMS によって最終的なプロポーサルを出力する．

Algorithm 1 Proposal Evaluation Module (PEM)

Input: D : プロポーサルの最大幅, $S_F \in \mathbf{R}^{C \times T}$: 特徴列, N : サンプル数

Output: $M_C \in \mathbf{R}^{D \times T}$: BM 信頼度マップ

$$M_F = \text{CREATE_BM_FEATURE}(S_F, C, N, D, T) \in \mathbf{R}^{C \times N \times D \times T}$$

$$x_1 = \text{Relu}(\text{Conv3D}(M_F)) \in \mathbf{R}^{C_1 \times D \times T} \quad \triangleright \text{3 次元 CNN による畳み込み}$$

$$x_2 = \text{Relu}(\text{Conv2D}_1(x_1)) \in \mathbf{R}^{C_2 \times D \times T} \quad \triangleright \text{2 次元 CNN による畳み込み}$$

$$x_3 = \text{Relu}(\text{Conv2D}_2(x_2)) \in \mathbf{R}^{C_3 \times D \times T}$$

$$M_C = \text{Sigmoid}(\text{Conv2D}_3(x_3)) \in \mathbf{R}^{2 \times D \times T}$$

Algorithm 2 Boundary-Matching(BM) 特徴量の算出

function CREATE_BM_FEATURE(S_F, C, N, D, T)初期化: $M_F \in \mathbf{R}^{C \times N \times D \times T}$ **for all** $i \in \{1, \dots, D\}, j \in \{1, \dots, T\}$ **do** $w_{i,j} = \text{CREATE_SAMPLING_MASK}(i, j) \in R^{N \times T}$ **for all** $c \in \{1, \dots, C\}, n \in \{1, \dots, N\}$ **do**

$$M_F[c, n, i, j] = \sum_{t=1}^T S_F[c, t] \cdot w_{i,j}[n, t]$$

end for**end for****return** $w_{i,j}$ **end function**

Algorithm 3 サンプリングマスク

function CREATE_SAMPLING_MASK(i, j)初期化: $w_{i,j} \in R^{N \times T}$ $t_s = t_i, d = t_j$ **for all** $n \in \{1, \dots, N\}, t \in \{1, \dots, T\}$ **do**

$$t_n = t_s - 0.25d + 1.5d \cdot \frac{n}{N} \quad (\text{区間 } [t_s - 0.25d, t_s + 0.25d] \text{ の } N \text{ 分割})$$

$$w_{i,j}[n, t] = \begin{cases} 1 - (t_n - \lfloor t_n \rfloor) & \text{if } t = \lfloor t_n \rfloor \\ t_n - \lfloor t_n \rfloor & \text{if } t = \lfloor t_n \rfloor + 1 \\ 0 & \text{if } t = \text{others} \end{cases}$$

end for**return** $w_{i,j}$ **end function**

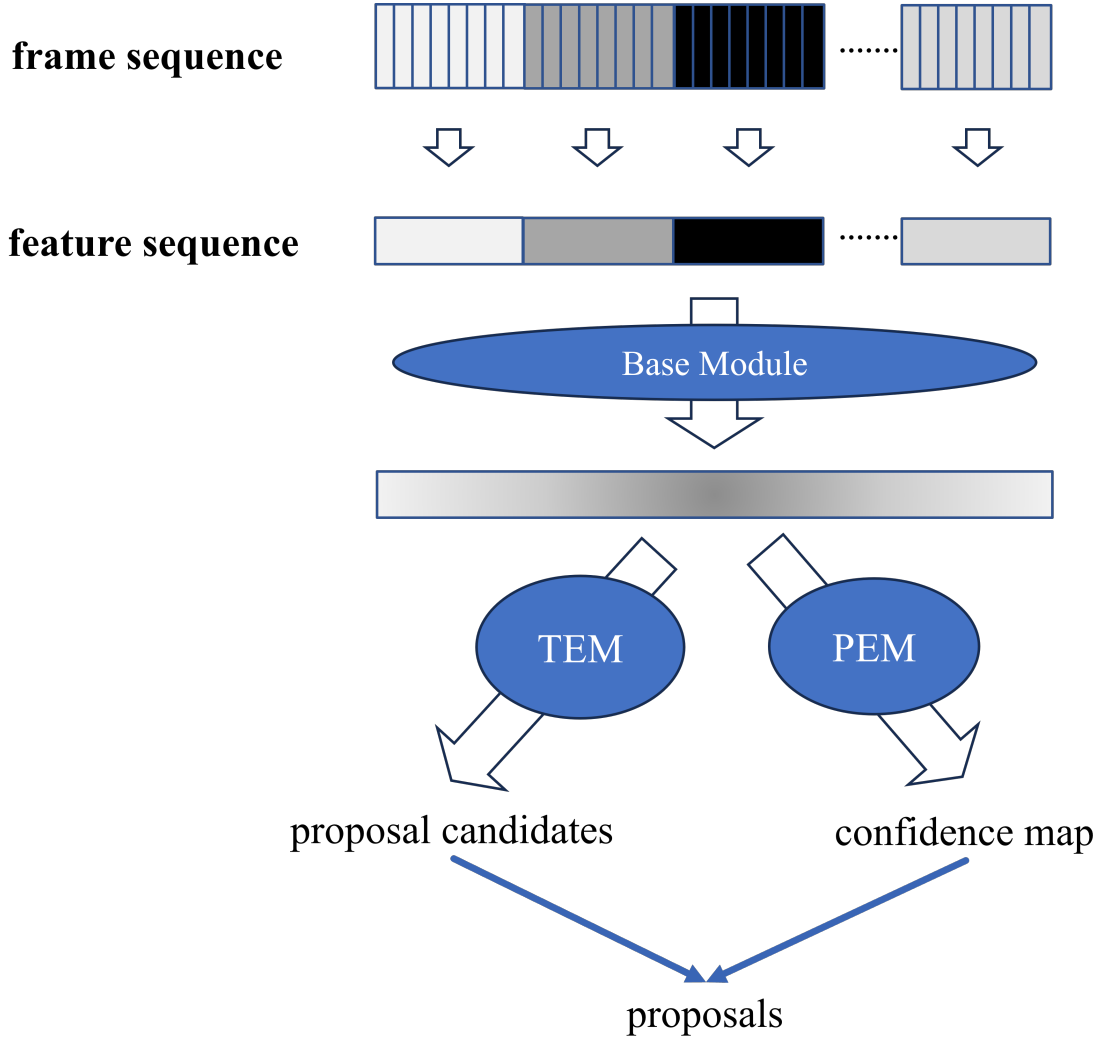


図 4.2.2 BMN の概要図

■**学習方法** ここでは、BMN の学習方法について述べる．学習データとしてまず、動画 X とアノテーション（正解ラベル） $G_X = \{\varphi_g^i = (t_s^i, t_e^i)\}$ ，を用意する．つぎに、フレーム数 T の動画 $V = \{X_t\}_{t=1}^T$ から、長さ l_f の特徴ベクトルの系列（以下、特徴列）、 $F = \{f_{t_n}\}_{n=1}^{l_f}$ を生成する． f_{t_n} は、時刻 $t = t_n$ を中心とした σ 枚のフレーム画像の特徴を表す．次に、あらかじめ定めた長さ l_w の窓を用いて特徴列および対応する正解ラベルを切り出す．この時、50% オーバーラップするように窓をスライドさせながら切り出す．ここまで、特徴ベクトルの集合 $\Omega = \{\omega_n\}_{n=1}^{N_\omega}$ と正解ラベルの組 Ψ_ω が得られる．

TEM は各 ω_n ごとに開始時刻 t_s と終了時刻 t_e の確率の系列 $P_{S,\omega} = \{p_{t_n}^s\}_{n=1}^{l_\omega}$ ，

$P_{E,\omega} = \{p_{t_n}^e\}_{n=1}^{l_\omega}$ を出力する. TEM の学習のために, 以下の方法で開始・終了を表す長さ T のラベルの系列 G_S および G_E を構成する (図 4.2.3). 正解データ $\varphi_g = (t_s, t_e)$ ごとに開始区間 $r_S = [t_s - d_g/10, t_s + d_g/10]$ および終了区間 $r_E = [t_e - d_g/10, t_e + d_g/10]$ を定めておく. ここで, $d_g = t_e - t_s$ である. ここで, 特徴 f_{t_n} の時間中心 $t = t_n$ を用いて $r_{t_n} = [t_n - d_f/2, t_n + d_f/2]$ とする. ここで, $d_f = t_n - t_{n-1}$ である. ここで, r_S, r_{t_n} がオーバーラップしている分を加味して $g_{t_s}^s = |r_S \cap r_{t_n}|/|r_S|$ とし, 正解ラベルを $G_{S,\omega} = \{g_{t_n}^s\}_{n=1}^{l_\omega}$ とする. 同様に, $G_{E,\omega} = \{g_{t_n}^e\}_{n=1}^{l_\omega}$ が構成できる. TEM の損失関数 L_{TEM} は以下のように定義される:

$$L_{TEM} = L_{bl}(P_S, G_S) + L_{bl}(P_E, G_E).$$

関数 L_{bl} は重み付きのクロスエントロピーであり, 以下のように定義される

$$L_{bl}(P, G) = \frac{1}{l_\omega} \sum_{i=1}^{l_\omega} (w^+ \cdot y_i \cdot \log p_i + w^- \cdot (1 - y_i) \cdot \log(1 - p_i))$$

ここで, $y_i = 1[g_i > \theta]$ は r_{t_n} とのオーバーラップ比 θ を超えたときに 1 をとる指示関数である. w^+, w^- は, 系列の長さに応じて正規化する重みであり,

$$w^+ = \frac{l_\omega}{\sum y_i}, \quad w^- = \frac{l_\omega}{l_\omega - \sum y_i}$$

となる.

PEM の出力は $M_C \in R^{2 \times D \times T}$ であるが, これを 2 種の信頼度マップ $M_C = (M_{CC}, M_{RC})$, $M_{CC}, M_{RC} \in R^{D \times T}$ と解釈し, それぞれ異なる損失関数で学習する. M_{CC}, M_{RC} とともに, 同じ正解ラベルマップ $G_C \in R^{D \times T}$ を用いる. それぞれのプロポーサル $\varphi_{i,j} = (t_s, t_e)$ に対して, Ψ_ω 内のすべての IoU 取った時の最大値を $g_{i,j}^c$ とする. $G_C = \{g_{i,j}^c\}_{i,j=1}^{D,l_\omega}$ をラベルとする. PEM の損失関数 L_{PEM} は以下のように定義される:

$$L_{PEM} = L_{bl}(M_{CC}, G_C) + \lambda L_2(M_{CR}, G_C)$$

ここで, 第一項は TEM で用いたクロスエントロピーであり, 第 2 項は L2 関数 (2 乗誤差の和) である.

モデル全体は以下の損失関数で学習される:

$$L = L_{TEM} + L_{PEM}$$

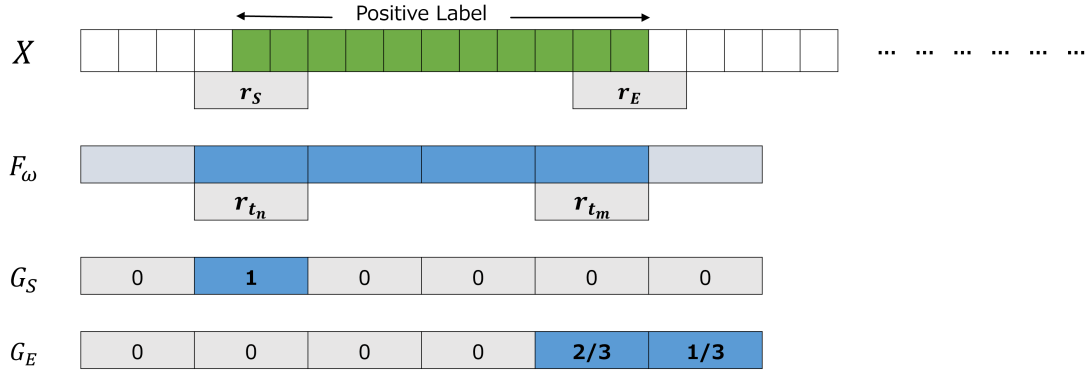


図 4.2.3 BMN の学習方法

4.2.5 推定方法

PEM からの出力により, $P_{S,\omega} = \{p_{t_n}^s\}_{n=1}^{l_\omega}$, $P_{E,\omega} = \{p_{t_n}^e\}_{n=1}^{l_\omega}$ が得られるが, まず, 境界の候補 $B_S = \{t_n | p_{t_n}^s \geq 0.5 \cdot \max P_S, \text{かつ}, p_{t_n}^s \text{ は極大値} \}$ とし, B_E も同様にする. プロポーサルの信頼度 p_f を以下のように定義する:

$$p_f = p_{t_s}^s \cdot p_{t_e}^e \cdot \sqrt{p_{cc} p_{cr}}$$

B_S, B_E から, 事前に定めた最大のプロポーサルの長さを超えない組み合わせを, $\Psi' = \{\varphi_i = (t_s^i, t_e^i, p_f^i)\}$ とし, さらに Ψ' の Soft-NMS をとったものを最終的な出力 Ψ とする.

4.2.6 評価方法

検出率を評価する場合は, Mean Average Presision (mAP) が用いられる. ActivityNet では, IoU の閾値が 0.5, 0.75, 0.95 のときの AP の平均値, IoU の閾値を 0.5 から 0.05 刻みで 0.95 まで変更していき, その AP の平均を評価指標とする [80]. また, THUMOS14 では, 0.3 から 0.7 まで 0.1 刻みで閾値を定めたものを用いる [31].

また, プロポーサル精度のみを評価する場合も多い. その評価指標としては, 複数の IoU 閾値を用いて計算する Average Recall (AR) が用いられることが多い [59]. ActivityNet では, 正解データとプロポーサルの IoU の閾値を 0.5 から 0.05 刻みで

0.95 まで変更していき、その AR の平均を指標とする。また、THUMOS14 では、IoU の閾値を 0.5 から 0.05 刻みで 1.0 までの値を用いている。プロポーサルの数が多いほど、当然再現率 (Recall) は上がる。そのため、平均のプロポーサルの数で再現率を評価した値 AR@AN (Average Recall with Average Number of proposals) も用いられる。また、AR と AN の関係をグラフ化し、そのグラフ下の面積 (Area Under the Curve, AUC) を用いることがある。

4.3 データセット

本節では、実験で用いるデータセットについて、その作成方法と内容について述べる。

4.3.1 行動ラベルの選定

著者は円山動物園の動画データを確認し、表 4.3.1 に示す行動を実際に確認した。この中で、以下の 3 条件を満たす「食事」行動を検出することにした。

1. 健康状態の把握に有効である。
2. 一定数以上のデータが確保できる。
3. 画像だけでは判別が難しく系列情報が必要である。

さらに、著者は、様々は食事方法のうち、“eat_net”を対象とした。主な食事方法は吊るされたネット（以下、ネット）から食べることであるが、タイヤの中の餌や壁の穴に隠された餌など、他の方法で餌を食べることもある (図 4.3.1)。これらの食事方法のなかには、試験的に短期間だけ採用されたものもあり、監視カメラに収められなかったものもある。食事方法のすべてに対応しようとする、「食べている」かどうかの判断は、アノテーターによって判断が異なる可能性がある。そのため、本研究ではネットによる食事に限定した。典型的な食事行動は以下の通りである。

1. ゾウは食事エリア（ネットの下）に近づき、立ち止まる。
2. 地面やネットの上にある干し草を鼻を使って口に運び、数秒間咀嚼する。
3. 次の干し草をつかむ。

種別	クラス名	説明
食事	drink	水を飲む
	eat_net	ネットから食べる
	eat_ground	地面から食べる
	eat_tire	タイヤから食べる
遊び	play_ball	ボールで遊ぶ
	play_water	泳ぐ
	play_with	じゃれ合う
動作	shake	ものを揺らす
	stand	立つ
	walk_back	後進
	walk_forward	前進
生理現象	excretion	排泄
	sleep	睡眠
	copulation	交尾

表 4.3.1 検出行動の候補

4. この行動を数回繰り返した後、餌場を離れる。

4.3.2 アノテーションポリシー

上記を踏まえ、以下のようなアノテーションポリシーを設定した。また、図 4.3.2 にアノテーションの一例を示す。

1. ゾウがネットから干し草を鼻で食べる行動（"eat_net"）にラベルを付ける。
2. 食べる行動が3秒以上続く場合を対象とする。
3. ネット以外の食事行動（"eat_tire" など）にはラベルを付けない。
4. 行動開始は、対象個体が歩行を停止し、食事を開始した時点とする。
5. 行動の終了は、対象個体が食べ終わり、その場を離れ始める時点とする。
6. すべての「食べる」行動には、「鼻を上にする」動作と「噛む」動作の両方が



eat from the tire



eat from the ground



eat from the net

図 4.3.1 様々な食事行動



図 4.3.2 食事行動のアノテーション例

含まれる．この2つの動作のどちらか一方しか観察されない場合，その行動はラベル付けしない．

No.6 により，対象のゾウが地面に落ちている干し草を食べたとしても，対象個体が鼻を上げない限り，その行動は食事とラベリングされない．このポリシーを立てた理由としては，ラベル対象となる状況は，意図的に餌場に近づく場合に限りたためである．「鼻を上げずに地面の干し草を食べる動作」は，ゾウが歩いているときに地面に落ちている干し草をたまたま見つけた場合にも起こる．本研究では，このような行動は「食べる」とみなさないことにした．

実験用のデータセットは，2022 年 2 月 1 日から 3 月 31 日までの監視カメラ映像を用いた．食事行動が見られる場面を重点的に収集したが，それ以外の行動も含んでいる．時間帯は午前 11 時から午後 4 時，または午後 7 時から午後 11 時の間に録画された．2022 年 2 月 1 日から 28 日までのデータを訓練データとして用い，3 月 1 日から 9 日まで日を検証データとして用いた．また，3 月 10 日から 3 月 31 日をテストデータとした．その結果，オスとメスのアノテーションの総時間は約 130 時間であった．このうちオスとメスの食事行動は 61 時間であった．

4.3.3 アノテーション作業の詳細

アノテーションは図 4.3.3 のように，録画動画と Excel を用いて記録をおこなった．Excel シートには雌雄，開始時刻（分，秒），終了時刻（分，秒）の列があり，後からプログラムで読みとることができる．

また，作業は 1 名のアノテーターに依頼をしている．作業時間は約 60 時間程度と見積もり，アノテーターと複数回アノテーション方法について確認した後，実際にアノ

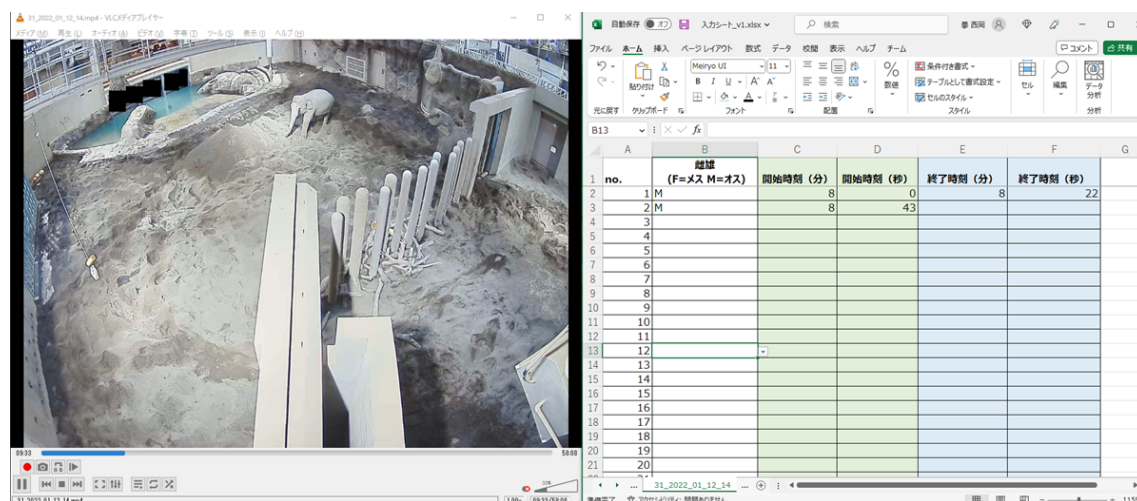


図 4.3.3 行動検出データセットのアノテーション作業

テーション作業を開始してもらった。

4.3.4 データセットの特徴

本節では、著者のデータセットにおいて行動ラベルを推定する際の特徴について、述べる。

1. 「餌を探す」行動と「餌を食べる」行動の間にほとんど差がない。

著者らは、餌を探すような行動を、「食べる」ではなく「背景（その他）」とラベル付けした。「探す」行動と「食べる」行動を区別するためには、比較的長い間その行動を観察する必要がある。

2. 食事時間は大きなばらつきがある。

各食事行動の継続時間は、最小 15 フレームから最大 10565 フレームであった。図 4.3.4 は「食べる」とラベル付けされたフレーム数の頻度とその累積比率を示している。累積比率を示す。「食べる」の 60 % 以上が 1000 フレーム以上である。

4.4 手法

本節では、問題を定式化後に手法について述べる。つぎに、手法の概要について説明した後、処理の順番にしたがって手法の詳細について述べる。

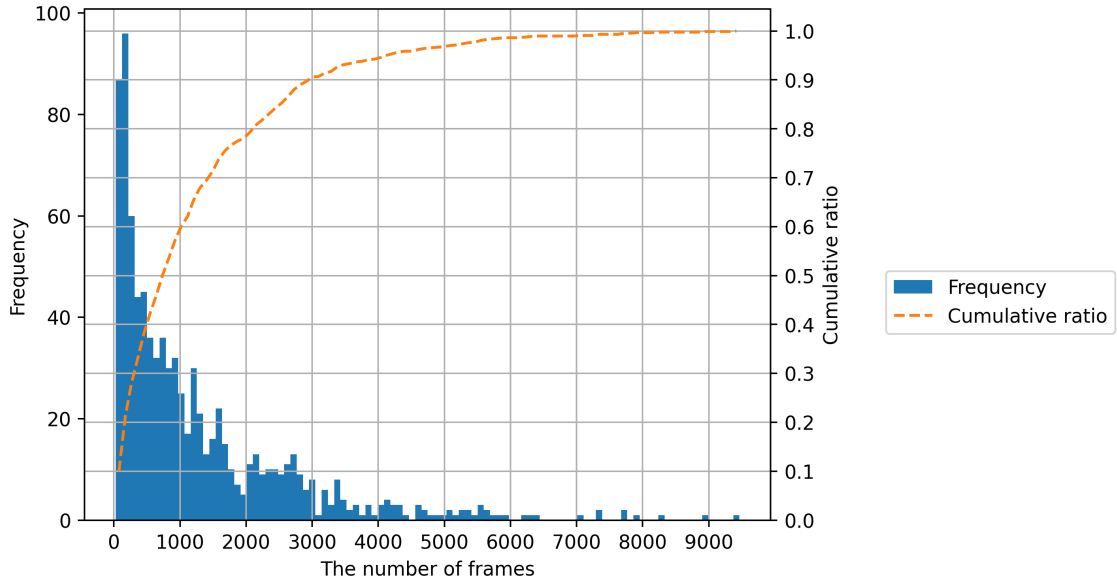


図 4.3.4 データセットにおける食事行動の頻度と累積比率

4.4.1 問題の定式化

本節では、提案手法を説明する前に解くべき問題を定式化する． $V = \{X_n\}_{n=1}^{l_v}$ を動画データとする．ここで、 X_n はフレーム画像、 l_v は動画の総フレーム数である．動画データ V にはトラッキング情報 $T = \{T_n\}_{n=1}^{l_v}$ が付随している．この T_n はフレーム画像における各個体の矩形位置と識別 ID を表す．モデルは動画 V とトラッキング情報 T から、個体ごとのプロポーサル $P = \{(t_{s,n}, t_{e,n}, p_{conf})\}_{n=1}^{N_p}$ を出力する．ここで、 N_p は映像中の対象行動の数、 $t_{s,n}, t_{e,n}, p_{conf}$ は開始時刻と終了時刻、信頼スコアである．

4.4.2 手法の概要

ここでは、処理の流れを概説する（図 4.4.1）．なお、本処理は事前に学習されたモデルを用いてトラッキングをおこなう．まず、動画、ゾウの推定矩形位置、識別 ID が入力される．動画を一定のフレーム数（8 から 64 フレーム）ごとに分割する．次に、各フレーム画像を推定矩形位置で個体ごとに切り出し、固定サイズにリサイズする．つぎに、事前に学習された行動分類モデルを用いて特徴を抽出する．最後に、特徴列

を TAL に入力し、対象動物の行動区間を推定する．本研究では、行動検出を 2 頭のゾウが互いに離れている状況に限定している．本手法は、トラッキングが正確であることを前提としている．ただし、実際には 2 頭のゾウが重なる状況下ではトラッキング精度は低下するため、その行動も認識することができない．この問題は、トラッキング精度を向上させることで解決できる．

4.4.3 前処理

各フレーム画像 X_n に付随するトラッキング情報 T_n は、各ゾウの矩形位置を切り出すために用いる．切り出された画像は一定の大きさにリサイズされる．実験では、画像サイズは 256×256 とし、切り出し方法は以下の 2 種類とした：(1) 個々の象の矩形位置に沿って切り出す (2) 正方形で切り出す．(1) の場合、リサイズによって縦横比が変化し、学習に影響を与える可能性がある．(2) の場合、縦横比は変わらないが、別のゾウが含まれる可能性が高まる．さらに、切り取られた画像に対してパディング処理をおこなった場合とおこなわなかった場合の精度を比較する．パディングは、切り出し後の画像は正方形になるように周囲を黒で埋め、 256×256 サイズにリサイズする．

4.4.4 特徴抽出

図 4.4.2 に特徴抽出手順と TAL の入力を示す． $x = \{x_n\}$ を切り出し後の画像とし、 x を重複しない短い時間セグメント $s_1, \dots, s_L$ に分割する．ここで L はセグメントの数であり、各 s_n は最大フレーム数を持ち、その中心 $\{t_n\}$ における t_n 番目のフレームは $\{1, \dots, l_v\}$ から等間隔で選択される．各セグメントに対して、学習された動画分類モデルを用いて特徴抽出が行われる．その結果、 $F = \{f_{t_n}\}_{n=1}^L$ が得られる．ここで、 $f_{t_n} \in R^C$ はフレーム x_{t_n} の周りの特徴ベクトル、 C は特徴次元である．BMN に関するオリジナルの研究では、特徴ベクトルは分類モデルの出力スコアであった．ここでは、分類モデルの出力スコアに加え、BMN の中間層の出力の精度も比較する．

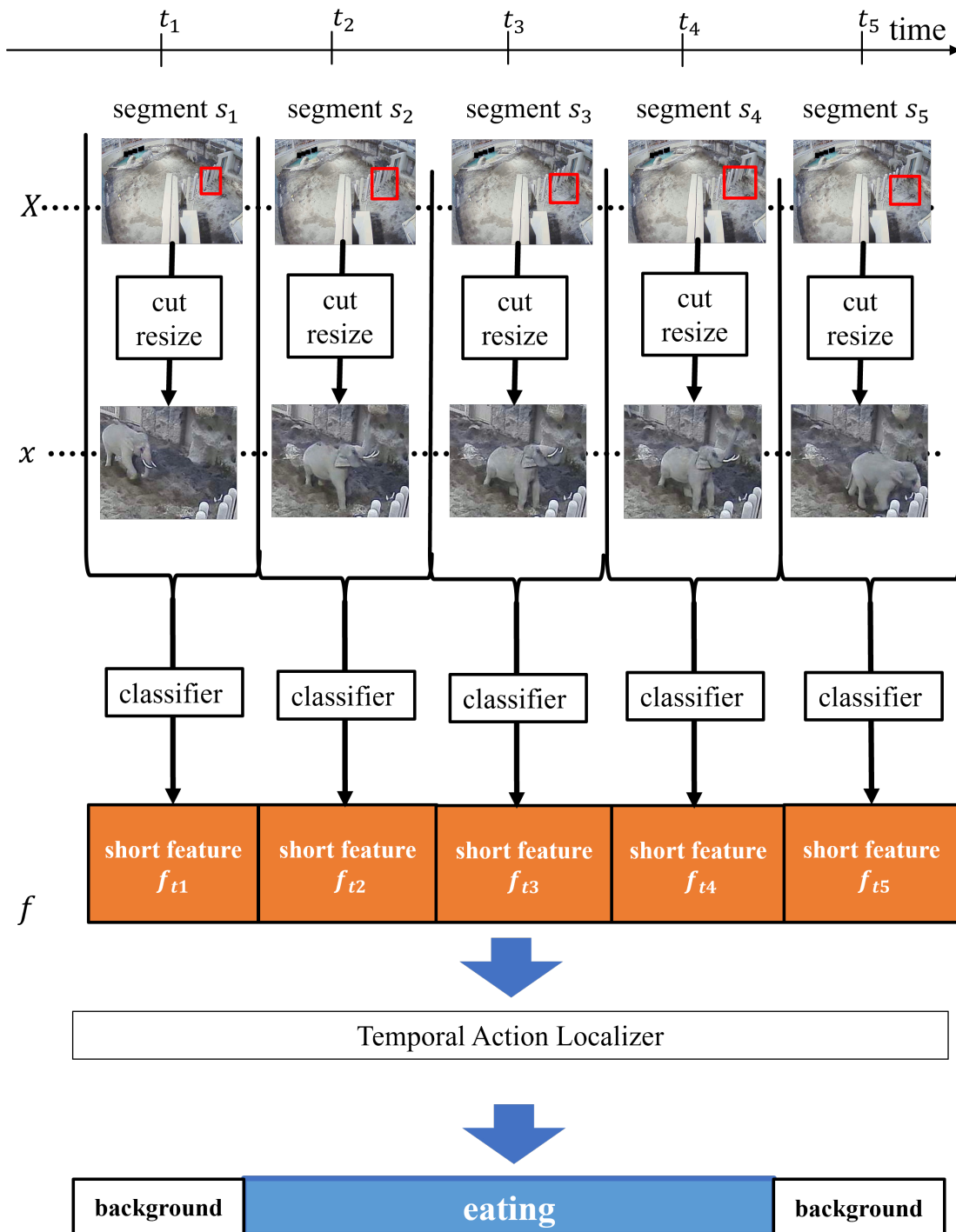


図 4.4.1 食事行動検出手法の概要図

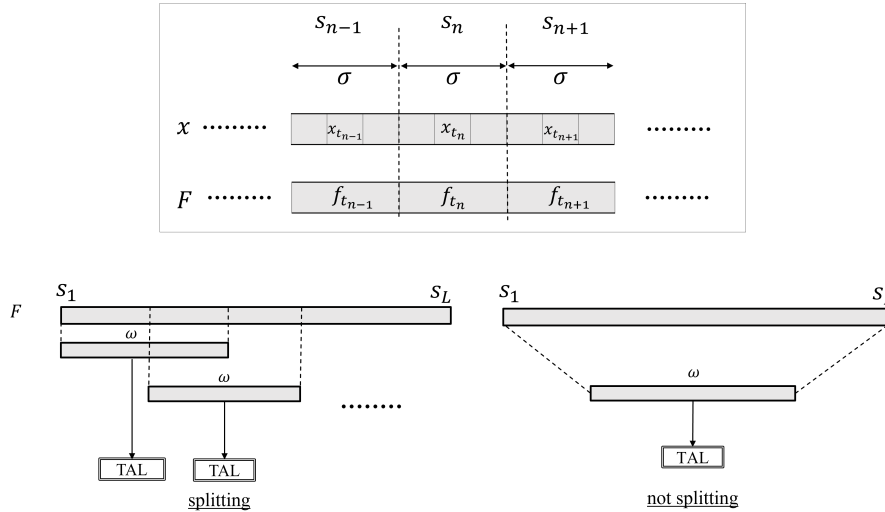


図 4.4.2 特徴抽出と TAL の入力

4.4.5 Temporal action localization

TAL モデルは特徴列 $F = \{f_{t_n}\}_{n=1}^L$ を入力とし、プロポーサル $\mathcal{P}$ を出力する。しかし、モデルに入力される特徴の次元は固定でなければならないため、特徴列 ω の長さ L をあらかじめ定めておき、 $L \leq \omega$ の場合と $\omega < L$ の場合で処理方法を分ける。まず、 $L \leq \omega$ の場合、線形補間を行い、特徴列の長さを $L = \omega$ とする。 $\omega < L$ の場合、「分割する (splitting)」方法 (図 4.4.2 左下) と「分割しない (not splitting)」方法 (図 4.4.2 右下) がある。前者では、特徴列を ω ごとに分割し、後者では特徴列を ω にリサンプリングする。最終的な出力は、モデルの予測プロポーザルに SoftNMS [9] を適用した後に得られる。「分割する」場合は、一定の閾値以上の信頼スコアをもつ予測区間の和集合をとる。

4.5 実験 - Temporal action classification

本節では、まず TAL に用いるための特徴抽出器の学習実験について述べる。次に、特徴抽出器を用いた TAL モデルの学習実験について述べる。最後に複数のパラメータでの結果を比較し、その結果を示す。

Phase	Clips	Period
Train	1816	2022年2月1日～2月28日
Valiation	604	2022年3月1日～3月9日
Test	1100	2022年3月10日～3月31日

表 4.5.1 TAC データセット

4.5.1 分類データセット

クラスラベルは「食べる」と「背景（その他）」とした．データ数はクラス間で等しくなるように調整した．既存の分類モデルの入力は、「クリップ」と呼ばれる約 300 フレームのデータである．これに合わせて，今回のデータセットでは最大フレーム数を 300 とした．表 4.5.1 に作成したデータセットを示す．クリップの総数は 3,530 であり，各クリップには平均 266.5 フレームが含まれる．

4.5.2 モデル

動画分類モデルとして広く用いられている TSN および I3D モデルを用いて結果を比較した．両モデルは，ImageNet で事前学習された Resnet50 をバックホーンネットワークとしている．オプティカルフローを用いる方法と最近提案されたモデルは精度が高くなる可能性はあるが，以下の理由から用いなかった．第一に，オプティカルフローは計算コストが高く，解析時間が増えるためである．第二に，TSN と I3D は TAL のベースネットワークとしてよく用いられているためである．著者らは実験で TSN と I3D を用いて，TAL が行動検出の精度を向上させるかどうかを検証した．図 4.5.1 に，分類モデルである TSN と I3D の前処理を示す．TSN では，まず入力の数フレームがパーティションに分割される．次に，分割されたパーティションから特徴が抽出され，TSN は各パーティションの信頼スコアを出力し，最後に，そのスコアはパーティション間で平均化される．なお，訓練時の計算コストが高いため，パーティションの数は通常，訓練とテストで変更される．訓練時には，各パーティションから 1 フレームがランダムに選択される．テスト時には，すべてのフレームが等間隔になる

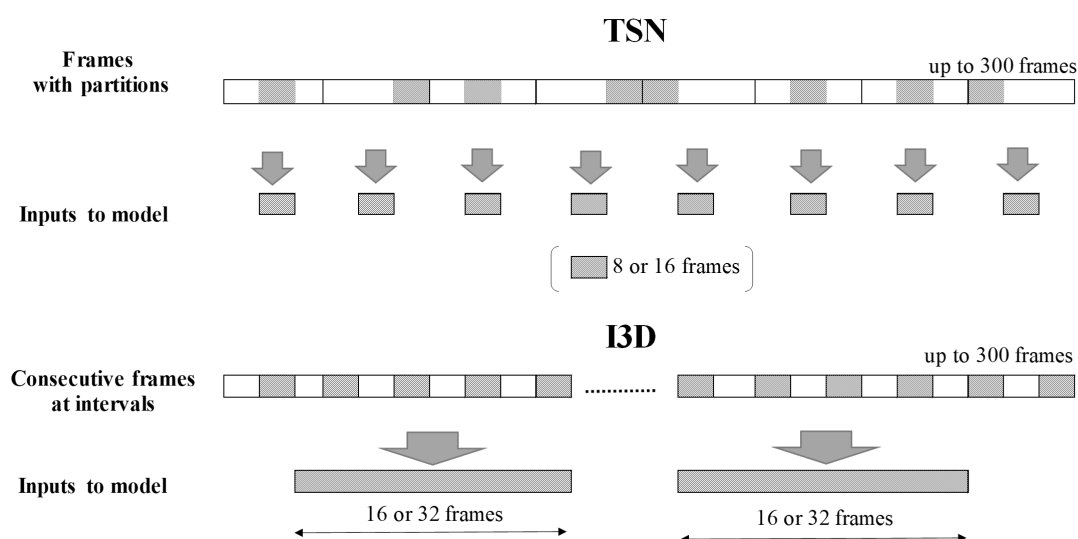


図 4.5.1 TAC モデル

ように、各パーティションから 1 フレームが選択される。I3D の場合、2 フレーム間隔で連続するフレームが入力となる。

4.5.3 結果

表 4.5.2 に結果を示す。「Model」列において、「TSN-x-y」は、TSN モデルが訓練時は x 個のパーティションを用いて学習し、バリデーションおよびテスト時には、 y 個のパーティションを用いて評価したことを示す。また、「I3D-z」は、2 フレーム間隔で連続する z 枚のフレーム画像を入力とする I3D モデルを示す。その結果、「I3D-16, padding, cut out by a raw rectangle」が最も精度が高いことがわかった。テストデータセットには、ゾウがカメラと反対を向いて「食べる」行動および「餌を探す」行動が含まれており、それぞれ「食べる」と「背景」とラベル付けされている。しかし、これらの行動は誤判定が多かった。

No.	Model	Padding	Preprocess	Accuracy
1	TSN-8-30		cut out by a raw rectangle	88.47%
2	TSN-8-30	✓	cut out by a raw rectangle	87.66%
3	TSN-8-30	✓	cut out by a square	88.02%
4	TSN-16-60	✓	cut out by a raw rectangle	87.12%
5	I3D-16	✓	cut out by a raw rectangle	89.19%
6	I3D-16	✓	cut out by a square	88.29%
7	I3D-32	✓	cut out by a raw rectangle	87.03%

表 4.5.2 TAC 実験結果

4.6 実験 - Temporal action localization

4.6.1 モデル

TAL モデルとして BMN[37] を用いた。これは、BMN を動物に適用した例が存在せず、多くの研究者によって再検証されている典型的なモデルを用いて、TAL の精度を検証したかったからである。BMN は、開始時刻と終了時刻を予測するための 1 次元畳み込み層と、開始時刻と終了時刻を評価するための 2 次元および 3 次元畳み込み層をもつ。本実験では、データセット kinetics[11] で学習済みの I3D モデルを特徴抽出に用いた。BMN の元論文 [37] ではオプティカルフローが使われていたが、本実験では用いなかった。図 4.6.1 は TVL1[83] によるオプティカルフローの出力例であり、輝度値は物体の速さ、色は移動方向を表している。どちらのゾウも動いてはいるが、手前のゾウはオプティカルフローとして検出できていることがわかる。ただし、奥のゾウはほとんど検出されなかった。これらから、オプティカルフローが精度向上に寄与する可能性は低いと考えた。フレーム間隔は 16 に設定した。各動画について、I3D は「食べる」または「背景」の信頼スコアの列を出力する。そして、この特徴列は最大長 ω の複数の特徴列に分割される。本節では、 $\omega = 100$ の場合について述べる。学習率や他のパラメータは BMN の原著論文 [37] で述べられている ActivityNet[80] に関する実験で用いられたものにした。



図 4.6.1 オプティカルフロー出力

Phase	Videos	Period	統計量			
			最大	最小	平均	総数
Train	271	2022 年 2 月 1 日～2 月 28 日	4370	15	979	317,292
Validation	117	2022 年 3 月 1 日～3 月 9 日	4109	15	898	124,752
Test	353	2022 年 3 月 10 日～3 月 31 日	4749	15	894	183,262

表 4.6.1 TAL データセット

4.6.2 TAL データセット

表 4.6.1 に TAL のデータセットの詳細を示す．訓練データセットと検証データセットでは，各動画に少なくとも 1 回の食事行動が含まれる．一方，テストデータセットには，食事行動のない動画も含んでいる．また，最小フレーム数は 1 動画 800 フレームとした．したがって，1600 フレーム未満の特徴列は，線形補間によってリサンプリングされない．また，線形補間によって最短の 800 フレームは 2 倍に伸張される．

4.6.3 評価

スライディングウィンドウとの比較では，フレーム単位で正しく分類されたかどうかで評価した．しかし，背景フレームが全体の 90 %を占めているため，平均精度 (AP) を評価に用いた．精度は $p(r)$ と表す．この場合，補間精度 p_{interp} は式 (1) の

ように設定される.

$$p_{interp}(r) = \max_{r' \leq r} p(r') \quad (4.1)$$

AP は式 (2) のように計算される.

$$AP = \int_0^1 p_{interp}(r) dr \quad (4.2)$$

実際には, 積分区間は 101 分割で近似される.

4.6.4 スライディングウィンドウとの比較結果

TAL に基づく方法と, 分類モデルとスライディングウィンドウを併用する方法を比較した. 分類モデルとしては, 「I3D-16, padding, cut out by a raw rectangle」を用いた. 分類モデルは 32 フレームを推定し, 各フレームは 50 % の重なりをもつ.

図 4.6.2 は precision-recall 曲線を表す. 「I3D-16+BMN」と「I3D-16+スライディングウィンドウ」の AP はそれぞれ 0.853 と 0.801 であり, BMN の AP はスライディングウィンドウの AP より 0.052 ポイント高かった. 図 4.6.3 は 2022 年 3 月 14 日 14:00 の結果の一部で, 横軸は 14 時 00 分からの経過時間 (秒), 緑色の部分が「食べる」行動, 灰色の部分がその他の行動を表している. I3D-16 +スライディングウィンドウの場合は, 線が途切れていることがわかる. TAL の場合, 推定結果は明瞭であるが本来複数の食事区間が 1 つの区間に合算される場合もある.

図 4.6.4 は, 「食事」の TAL とスライディングウィンドウの結果である. 緑色の領域は, ゾウが「食べている」と判定された区間である. これらのフレームでは, ゾウは方向を変えながらネットやネット下で干し草を食べている. スライディングウィンドウの場合, ゾウがカメラの反対側を向いているときには検出できないが, TAL では正しく検出された.

4.6.5 Ablation study

本節では, 複数の異なる条件下での比較結果を示す. 比較をおこなう理由は, 既存の TAL の研究は主に人に対してのものであり, 動物ではあまり研究されておらず, 前

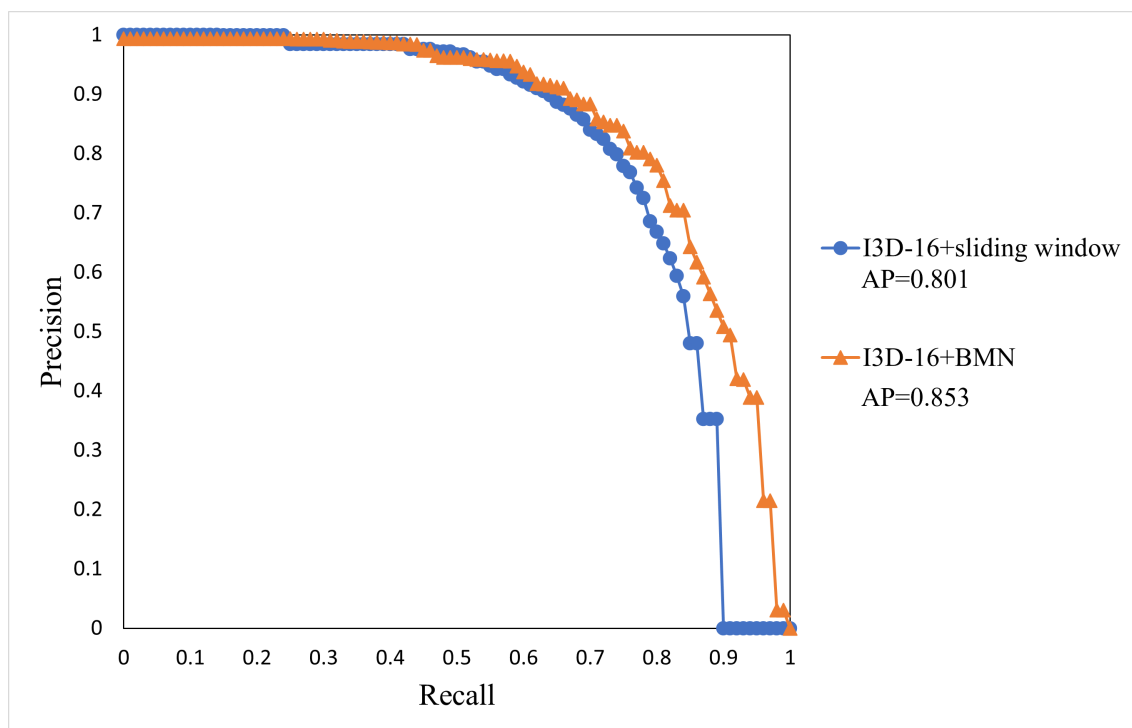


図 4.6.2 行動検出結果の PR 曲線

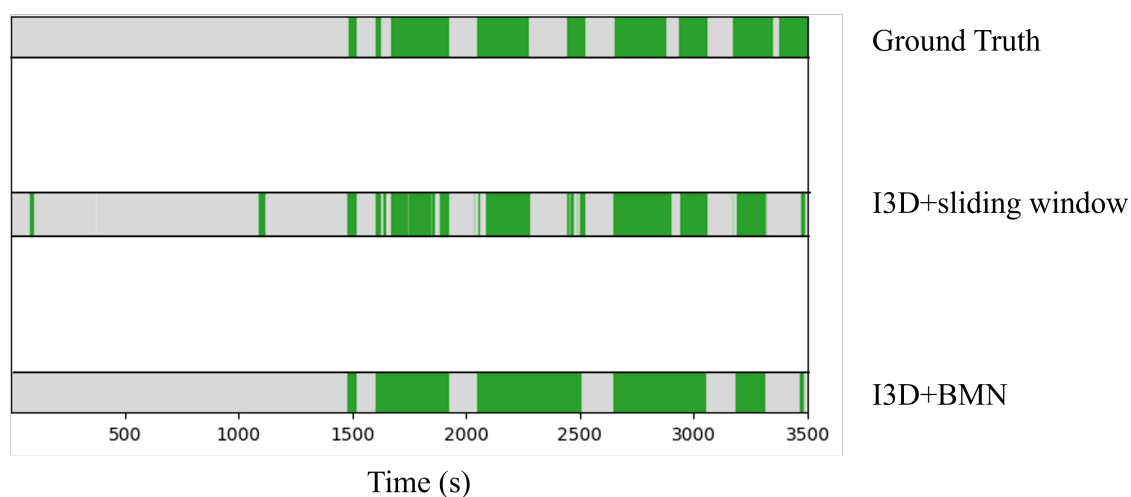


図 4.6.3 TAC と TAL による予測結果の比較

処理やモデルの違いによる効果に関する報告が少ないからである．本節での実験ではフレーム間隔 σ は 16 に固定し，その他のパラメータは 4.6.4 節と同様とした．

表 4.6.2 に結果を示す．最も高い AP (C-2, 0.854) と最低の AP (C-3 および C-4, 0.794) の差は 0.058 であった．この結果は，特徴抽出とモデルの違いで結果が異なる可能性があることを示している．

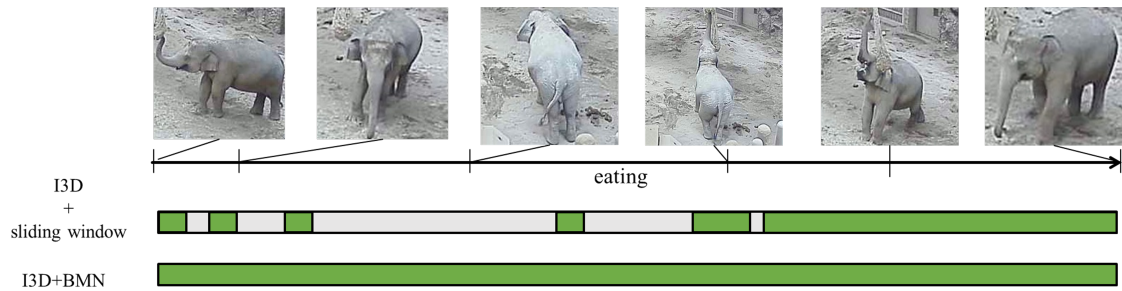


図 4.6.4 カメラと逆向きでの食事行動検出結果

■**特徴抽出器** 特徴抽出器として I3D-16 と TSN-8-30 を比較した。パディングと矩形による切り出しを行った。A-1 と A-2 の結果は、I3D-16 の AP が TSN-8-30 より優れていることを示している。これは、I3D-16 が TSN-8-30 よりも特徴を効果的にとらえられているためと考えられる。これはおそらく、I3D-16 が高密度なサンプリングによって、TSN-8-30 よりも特徴量を効果的に用いているためであると考えられる。

■**分割方法** 4.4.5 節で説明したように、特徴列を分割する方法と分割しない方法がある。I3D-16 を特徴抽出器として用いた場合 (B-1, B-2)、「分割する」方が「分割しない」よりも結果は良好であった。しかし、TSN-8-30 を特徴抽出器として用いた場合 (B-3, B-4) は、その逆の結果となった。これは、特徴抽出器のアーキテクチャによって結果が異なる可能性があることを示している。

「分割しない」場合、特徴列が長ければ長いほど、より正確な推定が期待できる。「分割しない」場合の特徴列の長さは 200 とした。しかし GPU メモリの制限により、それ以上の場合は評価できなかった。解像度を下げることなく、より長い特徴列を用いた場合、モデルはより正確な予測が可能となると考えている。

■**オーバーラップ** フレームを分割する場合、フレーム列のオーバーラップ比を調整することができる。オーバーラップ比が高いほど、より多くのフレームが用いられる。ただし、C-1～C-4 の結果を見るとオーバーラップ比の影響はわずかであることがわかる。

■**特徴抽出層** 特徴タイプとは、分類モデルにおける特徴ベクトルのタイプを指す。特徴抽出器の出力特徴抽出器の出力スコア (表 4.6.2 の「Score」) の他に、中間層の

ID	特徴抽出器	分割	特徴列の長さ	オーバーラップ比	特徴抽出層	AP
A-1	I3D-16	✓	100	0.5	Score	0.853
A-2	TSN-8-30	✓	100	0.5	Score	0.816
B-1	I3D-16	✓	100	0.5	Score	0.853
B-2	I3D-16		200	0	Score	0.843
B-3	TSN-8-30	✓	100	0.5	Score	0.816
B-4	TSN-8-30		200	0	Score	0.824
C-1	I3D-16	✓	100	0.5	Score	0.853
C-2	I3D-16	✓	100	0.9	Score	0.854
C-3	TSN-8-30	✓	100	0.5	Score	0.796
C-4	TSN-8-30	✓	100	0.9	Score	0.796
D-1	I3D-16	✓	100	0.5	Score	0.853
D-2	I3D-16	✓	100	0.5	Middle	0.843
D-3	TSN-8-30	✓	100	0.5	Score	0.816
D-4	TSN-8-30	✓	100	0.5	Middle	0.809

表 4.6.2 Ablation Study 結果

出力（表 4 の「Middle」）も TAL の入力として採用できる．中間層の出力を用いる場合は，Resnet50 の Average Pooling 後の特徴量を用いた．この時，特徴ベクトルの次元は 2048 となる．また，GPU のメモリ使用量の制限により，4 ケースのみで実験を行った．D-1～D-4 の結果は，特徴抽出器のアーキテクチャによって異なる可能性があることを示している．

4.7 考察

本研究の目的は，ゾウの行動検出において，分類とスライディングウィンドウを用いた手法よりも TAL の方が有効であると示すことである．本研究は食事行動検出に限定されたものであったが，本研究の結果は，正確な行動検出が可能であることを示唆するものである．また，TAL は分類とスライディングウィンドウを用いた手法より

も高い AP を達成しただけでなく、ゾウがカメラと反対を向いているときの食事行動も検出することができた。

しかし、不連続な食事行動が 1 つの食事行動として検出される場合があった。食事回数が健康チェックに重要である場合には問題となる可能性もある。より正確に検出するためには、以下のことが考えられる。

1. 長い特徴列を用いた学習

4.6.5 節で述べた方法では、GPU のメモリ使用量が多く情報量を減らす必要があったが、このようなハードウェアの問題が解決できれば、より精度の高い予測ができる可能性がある。

2. 多くの行動ラベルを用いて特徴抽出器を学習させる

TAL の性能は特徴抽出器の性能に依存しているが、本研究では、「食べる」と「背景」という限られた行動ラベルで特徴抽出器を訓練している。そのため、特徴抽出器を訓練するための行動ラベルを増やせば、TAL の性能が向上することが期待される。

また、本研究では、食事行動のみを検出する方法を説明したが、より正確な日常の健康チェックを行うためには、複数の行動を検出する必要がある。これは、生成されたプロポーサルごとに行動ラベルを推定することで実現できる。具体的な方法を説明する。まず、プロポーサル内のセグメントからフレームをサンプリングし、行動分類モデルに入力する。次に、分類モデルのスコアの加重平均あるいはソフトマックスのスコアの平均を用いて行動ラベルの判定をおこなう。また、分類モデルのスコアによって事前学習された別のモデル（たとえば、ニューラルネットワークや線形分類）の出力を用いる手法もある [68] [72]。ただし、これらの研究では人の行動が検出の主な対象であるため、動物の場合はさらなる検証が必要である。

4.8 本章のまとめ

非接触センサー（カメラやマイクなど）は、オクルージョンが多いことや夜間の検出が必要なことから敬遠される傾向にある。しかし、非接触センサーは動物福祉の観点から接触センサーよりも望ましい。著者は、監視カメラの映像のみから動物園のゾ

ウの食事行動を自動的に推定する手法を提案した。本手法は、既存の TAL の方法である BMN をベースとしている。2 ヶ月分の監視カメラ映像フレームのデータセットを作成した。実験では、分類とスライディングウィンドウを用いた手法と TAL を比較した。TAL は分類・スライディングウィンドウよりも高い AP を達成し、より明瞭な行動の境界を生成した。最終的な目標は、ゾウの体調の変化を検出することであるが、そのためには行動の時間、場所、数を測定することが必要である。また、現在のモデルは精度が高いが、結果を検証するためには、実際のゾウの体調の記録と比較する必要がある。これらの課題は今後の研究に委ねたい。

第 5 章

結論

本論文では、エソグラム生成の目的と関連研究 (2 章) および個体識別をともなうトラッキング (3 章) と食事行動検出 (4 章) についての研究を述べた。ここでは、それぞれのまとめと今後の展望について述べる。

■**エソグラム生成** 動物行動学研究者および動物飼育管理者のそれぞれの立場から、エソグラムの自動生成ツールの必要性を述べた。エソグラム生成自動化ツールは、動物行動学研究者が大量のデータを解析するための研究支援として需要があり、また、動物飼育管理者からは、監視カメラからの観察記録業務の効率化ツールとして需要がある。さらに、関連研究としてセンサーを用いた既存研究について述べた。接触センサーは、よい結果を示す研究例が多いものの動物福祉の立場から避けられるべきであることを述べた。また、カメラ映像を用いた既存研究についてそれぞれ述べ、深層学習ベースの手法を中心にまとめた。本論文で述べた動物を対象にした「個体識別をともなうトラッキング」と「行動検出」の既存研究は少ないことを述べた。最後に、実例として札幌市円山動物園に求められるシステムについて述べた。

■**個体識別をともなうトラッキング** 個体識別をともなうトラッキングを実現するため、「物体検出」「個体識別」「トラッキング」の 3 タスクの既存手法を組み合わせたモデルを提案した。実験では、円山動物園のゾウ舎における 2 頭のゾウを対象に手法の有効性を示した。円山動物園の監視カメラ映像から、識別 ID ごとにゾウの位置のアノテーションをおこない学習データセットを作成した。「物体検出」「個体識別」の

実験では、複数モデルを比較し、YOLOv5-l6 と Efficient B2 を用いることを決定した。また、「トラッキング」では、最小費用流における複数のコスト関数やパラメータを比較した。また、提案手法は 24 時間分のカメラ映像で評価した。結果、行動記録のベースとなりうる精度を得ることができた。本研究の新規性は、深層学習を軸とした画像・動画認識手法を組み合わせることで、動物の行動解析に必要な精度のトラッキングが実現可能であると示した点である。「物体検出」「個体識別」「トラッキング」のそれぞれの既存アルゴリズムを単純に適用するだけでは実現が難しい。本研究では、3 タスクの組み合わせ方法を工夫することで、高精度なトラッキング手法を構築している。また、動物に対する画像認識、動画認識の特性の一つとして各個体の外見が非常に近いことが挙げられる。トラッキングの軌跡を求めるには、各個体のフレーム間の相対的な変化 (例えば、位置や外見の差) を考慮することが重要になる。本研究では、実験により、特に位置情報の利用が重要であることを明らかにした。

■**食事行動検出** 本研究では、監視カメラの映像のみから動物園のゾウの食事行動を自動的に検出する手法を構築し、データセットを用いて学習、評価をおこなった。また、食事行動に限定し行動検出をおこなうことで、単一フレームからでは判別ができないような行動が検出できることを確かめた。実験では、2 ヶ月分の監視カメラ映像フレームのデータセットを作成し、動画分類モデルを用いる方法と TAL による方法を比較した。それに加えて、TAL の前処理や特徴抽出方法を複数検討し比較実験をおこなった。TAL を動物に用いた既存研究が無いため、基本的な前処理や特徴抽出方法による複数比較した。結果、用いる方法によって大きく精度が変わりうることが明らかとなった。また、TAL を用いるモデルは、動画分類モデルを用いるよりも高い精度を達成し、より正確な行動検出が可能となることを明らかにし、加えてカメラから確認しづらい状況でも食事行動の検出が可能であることを確認した。

■**今後の展望** ここでは、今後の展望として 2 点述べる。まず、本論文で述べた手法の汎化性を評価することである。本論文では、円山動物園ゾウ舎の 2 頭のゾウの監視カメラ映像でのみ評価をおこなっている。別環境での検証や、動物の種や数に対しての定量的な評価することが必要である。また、睡眠や遊び等の異なった行動の検出を試みることも考えられる。

次に、自動エソグラム作成システムの構築と運用である。実際の動物園管理者の声を反映させるため、システム運用からのフィードバックを得ることが必要であると考えている。1年程度の監視記録を通して実際にゾウの体調変化が捉えられるのか検証をおこなうことも考えられる。

謝辞

本論文の執筆にあたり多くの方々にご協力いただきました。指導教官である北海道大学大学院情報科学研究院 山本雅人教授には、本学の受験から現在まで、研究や種々の学内手続きや相談に真摯に対応していただきました。心より感謝いたします。また、北海道大学 人間知・脳・AI 研究教育センター 飯塚博幸准教授、北海道大学 数理・データサイエンス教育研究センター 野口渉特任助教には、ゼミでのご指摘や論文のご指導をしていただきました。わたしの研究に関して具体的な助言をいただき、論文の完成に至ることができました。深く感謝いたします。また、北海道大学大学院情報科学研究院 野田五十樹教授、北海道大学大学院情報科学研究院 川村秀憲教授、北海道大学大学院情報科学研究院 小野哲雄特任教授には、本論文の副査を引き受けて下さり、論文内容のご指導と共に研究者としての助言をしていただきました。深く感謝いたします。

本研究は、札幌市円山動物園のゾウ舎映像を使用させていただき、飼育員の方々からの意見を参考にさせていただきました。札幌市、および札幌市円山動物園の関係者の皆様方に深謝いたします。

また、わたしの所属研究室である自律系工学研究室のみなさまに感謝いたします。秘書である神裕子氏には、学習データ作成にご協力いただきました。学生の方々の発表や議論は参考になるとともに、よい刺激になりました。

また、現職である株式会社システム計画研究所の関係者の方々には、会社と大学を両立させるための環境を整備していただきました。深く感謝申し上げます。

最後に、わたしを精神的に支えてくれた妻に感謝いたします。妻にはわたしの決断を後押ししてもらい、ここまで応援し続けてくれました。この場をお借りして、感謝の意を表し謝辞とさせていただきます。

研究業績

- [1] Ken Nishioka, Wataru Noguchi, Hiroyuki Izuka, and Masahito Yamamoto. Detecting Eating Behavior of Elephants in a Zoo Using Temporal Action Localization. *Sensors and Materials*, 35(11):3927–3945, 2023.
- [2] Ken Nishioka, Wataru Noguchi, Hiroyuki Izuka, and Masahito Yamamoto. 動物園のゾウを対象にした単一監視カメラ映像によるトラッキング手法の検討と実践. 情報処理学会 デジタルプラクティス, 5, 2024.

参考文献

- [3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In Ferrari et al. [19], pages 472–489.
- [4] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019.
- [5] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long term temporal context for per-camera object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 13072–13082, 2020.
- [6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.
- [7] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2017.
- [8] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.
- [9] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-NMS - Improving Object Detection with One Line of Code. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 5562–5570, 2017.

- [10] James P. Bohnslav, Nivanthika K. Wimalasena, Kelsey J. Clausing, Yu Y. Dai, David A. Yarmolinsky, Tomás Cruz, Adam D. Kashlan, M. Eugenia Chiappe, Lauren L. Orefice, Clifford J. Woolf, and Christopher D. Harvey. DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels. *eLife*, 10, 2021.
- [11] João Carreira and Andrew Zisserman. Quo Vadis, action recognition? A new model and the kinetics dataset. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:4724–4733, 2017.
- [12] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
- [13] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020.
- [14] Marian Stamp Dawkins. *Observing animal behaviour: design and analysis of quantitative data* (邦訳: マリアン・S・ドーキンス 著, 黒沢 令子 訳, 動物行動の観察入門 計画から解析まで, 白揚社 (2015)). Oxford University Press, 2007.
- [15] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, pages 1–37, 2020.
- [16] Alessia Diana, Marina Salas, Zjef Pereboom, Michael Mendl, and Tomas Norton. A systematic review of the use of technology to monitor welfare in zoo animals: is there space for improvement? *Animals*, 11(11):3048, 2021.
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

-
- [18] Christoph Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 200–210, 2020.
 - [19] Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors. *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*. Springer, 2018.
 - [20] C D FitzGibbon and J H Fanshawe. Stotting in Thomson’s gazelles: an honest signal of condition. *Behavioral Ecology and Sociobiology*, 23(2):69–74, 1988.
 - [21] John Forrest, Ted Ralphs, Stefan Vigerske, Haroldo Gambini Santos, John Forrest, Lou Hafer, Bjarni Kristjansson, jpfasano, EdwinStraver, Miles Lubin, Jan-Willem, rlougee, jpgoncal1, Samuel Brito, h-i gassmann, Cristina, Matthew Saltzman, tosttost, Bruno Pitrus, Fumiaki MATSUSHIMA, and to st. coin-or/cbc: Release releases/2.10.11, October 2023.
 - [22] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOx: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
 - [23] Jocher Glenn. YOLOv5, <https://ultralytics.com/yolov5>.
 - [24] Alan Grafen and Rosemary Hails. *Modern statistics for the life sciences*. Oxford University Press, 2002.
 - [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
 - [26] Yuhang He, Xing Wei, Xiaopeng Hong, Weiwei Shi, and Yihong Gong. Multi-Target Multi-Camera Tracking by Tracklet-to-Target Assignment. *IEEE Transactions on Image Processing*, 29:5191–5205, 2020.
 - [27] Ann-Helena Hokkanen, Laura Hänninen, Johannes Tiusanen, and Matti Pastell. Predicting sleep and lying time of calves with a support vector machine classifier using accelerometer data. *Applied Animal Behaviour Science*, 134(1-2):10–15, 2011.

- [28] Matthew R Holdgate, Cheryl L Meehan, Jennifer N Hogan, Lance J Miller, Jeff Rushen, Anne Marie de Passillé, Joseph Soltis, Jeff Andrews, and David J Shepherdson. Recumbence behavior in zoo elephants: determination of patterns and frequency of recumbent rest and associated environmental and social factors. *PLoS One*, 11(7):e0153301, 2016.
- [29] Andrea Hornakova, Roberto Henschel, Bodo Rosenhahn, and Paul Swoboda. Lifted disjoint paths with application in multiple object tracking. In *International Conference on Machine Learning*, pages 4364–4375. PMLR, 2020.
- [30] Andrea Hornakova, Timo Kaiser, Paul Swoboda, Michal Rolinek, Bodo Rosenhahn, and Roberto Henschel. Making higher order mot scalable: An efficient approximate solver for lifted disjoint paths. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6330–6340, 2021.
- [31] Haroon Idrees, Amir R. Zamir, Yu Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS challenge on action recognition for videos “in the wild” . *Computer Vision and Image Understanding*, 155:1–23, 2017.
- [32] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.
- [33] Hao Jiang, Sidney Fels, and James J Little. A linear programming approach for multiple object tracking. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.
- [34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. may 2017.
- [35] H Kuehne, H Jhuang, E Garrote, T Poggio, and T Serre. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563, 2011.

-
- [36] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, pages 7082–7092, nov 2019.
 - [37] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: Boundary-Matching Network for Temporal Action Proposal Generation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, volume 2019-Octob, pages 3888–3897. IEEE, oct 2019.
 - [38] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11208 LNCS, pages 3–21, 2018.
 - [39] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
 - [40] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
 - [41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
 - [42] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI’81: 7th international joint conference on Artificial intelligence*, volume 2, pages 674–679, 1981.
 - [43] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order met-

- ric for evaluating multi-object tracking. *International Journal of Computer Vision*, pages 1–31, 2020.
- [44] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial intelligence*, 293:103448, 2021.
- [45] J E Martin. The effects of rearing conditions on grooming and play behaviour in captive chimpanzees. *Animal Welfare*, 14(2):125–133, 2005.
- [46] Stuart Mitchell, Michael OSullivan, and Iain Dunning. Pulp: a linear programming toolkit for python. *The University of Auckland, Auckland, New Zealand*, 65, 2011.
- [47] Olga Moskvayak, Frederic Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Learning landmark guided embeddings for animal re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 12–19, 2020.
- [48] Yoshida Nobuaki. 動物園における飼育動物映像の多面的な活用に向けた 監視カメラ映像システムの検討. *情報処理学会論文誌*, 2019(8):1–8, 2019.
- [49] Manfred Padberg and Giovanni Rinaldi. A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems. *SIAM review*, 33(1):60–100, 1991.
- [50] Yue Pang, Wenbo Yu, Yongan Zhang, Chuanzhong Xuan, and Pei Wu. Sheep face recognition and classification based on an improved MobilenetV2 neural network. *International Journal of Advanced Robotic Systems*, 20(1), jan 2023.
- [51] Alfonso Pérez-Escudero, Julián Vicente-Page, Robert C. Hinz, Sara Arganda, and Gonzalo G. De Polavieja. IdTracker: Tracking individuals in a group by automatic identification of unmarked animals. *Nature Methods*, 11(7):743–748, 2014.
- [52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

-
- [53] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
 - [54] Daniel Schofield, Arsha Nagrani, Andrew Zisserman, Misato Hayashi, Tetsuro Matsuzawa, Dora Biro, and Susana Carvalho. Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, 5(9):1–9, 2019.
 - [55] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
 - [56] Javier Selva, Anders S. Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B. Moeslund, and Albert Clapes. Video Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–26, 2023.
 - [57] Katrin Siebert, Jan Langbein, Peter-Christian Schön, Armin Tuchscherer, and Birger Puppe. Degree of social isolation affects behavioural and vocal response patterns in dwarf goats (*capra hircus*). *Applied Animal Behaviour Science*, 131(1-2):53–62, 2011.
 - [58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
 - [59] Sorn Sooksatra and Sitapa Watcharapinchai. A Comprehensive Review on Temporal-Action Proposal Generation. *Journal of Imaging*, 8(8), 2022.
 - [60] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402*, 2012.
 - [61] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv: 2012.15460*, 2020.

- [62] AB Swanson, M Kosmala, CJ Lintott, RJ Simpson, A Smith, and C Packer. Data from: Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna, 2015.
- [63] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [64] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [65] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.
- [66] Niko Tinbergen. On aims and methods of ethology. *Zeitschrift für tierpsychologie*, 20(4):410–433, 1963.
- [67] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 4489–4497, 2015.
- [68] Elahe Vahdani and Yingli Tian. Deep Learning-based Action Detection in Untrimmed Videos: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–24, 2022.
- [69] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.
- [70] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The*

-
- Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 135–153. Springer, 2016.
- [71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
 - [72] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. UntrimmedNets for weakly supervised action recognition and detection. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:6402–6411, 2017.
 - [73] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9912 LNCS, pages 20–36, aug 2016.
 - [74] Jason D Wark, Katherine A Cronin, Tony Niemann, Marisa A Shender, Adrienne Horrigan, Alison Kao, and Megan R Ross. Monitoring the behavior and habitat use of animals to enhance welfare using the zoomonitor app. *Anim. Behav. Cogn*, 6(3):158–167, 2019.
 - [75] Hendrik J. Weideman, Charles V. Stewart, Jason R. Parham, Jason Holmberg, Kiirsten Flynn, John Calambokidis, D. Barry Paul, Anka Bedetti, Michelle Henley, Jerenimo Lepirei, and Frank G. Pope. Extracting identifying contours for african elephants and humpback whales using a learned appearance model. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1265–1274, 2020.
 - [76] JC Whitham and LJ Miller. Using technology to monitor and improve zoo animal welfare. *Animal Welfare*, 25(4):395–409, 2016.
 - [77] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and real-time tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

- [78] Chao Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 284–293, 2019.
- [79] Tristram D Wyatt. *Animal behaviour: A very short introduction* (邦訳：基礎からわかる動物行動学，トリストラム・D・ワイアット (著)，沼田英治 (監修)，青山 薫 (訳)，*Newton Press*). Oxford University Press, 2017.
- [80] Yuanjun Xiong, Limin Wang, Zhe Wang, Bowen Zhang, Hang Song, Wei Li, Dahua Lin, Yu Qiao, Luc Van Gool, and Xiaoou Tang. CUHK & ETHZ & SIAT Submission to ActivityNet Challenge 2016. 2016.
- [81] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C.H. Hoi. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, jan 2021.
- [82] Xuqiang Yin, Dihua Wu, Yuying Shang, Bo Jiang, and Huaibo Song. Using an EfficientNet-LSTM for the recognition of single Cow’s motion behaviours in a complicated environment. *Computers and Electronics in Agriculture*, 177:105707, oct 2020.
- [83] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*, pages 214–223. Springer, 2007.
- [84] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.
- [85] Libo Zhang, Junyuan Gao, Zhen Xiao, and Heng Fan. Animaltrack: A benchmark for multi-animal tracking in the wild. *International Journal of Computer Vision*, 131(2):496–513, 2023.
- [86] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan,

Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object Tracking by Associating Every Detection Box. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13682 LNCS:1–21, oct 2022.

- [87] 公益社団法人日本動物園水族館協会 (JAZA). 日本動物園水族館協会の 4 つの役割. <https://www.jaza.jp/about-jaza/four-objectives>.
- [88] 村田 浩一, 成島 悦雄, and 原 久美子. **動物園学入門**. 朝倉書店, 2014.
- [89] 日本生態学会. **行動生態学 (シリーズ 現代の生態学 5)**. 共立出版, 2012.