情報処理学会論文誌　デジタルプラクティス　Vol.5　No.3　60–69 (July 2024)

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

西岡 拳[1,2,a)]　野口 渉[5,b)]　飯塚 博幸[4,c)]　山本 雅人[3,4,d)]

受付日 2023年10月20日, 採録日 2024年3月15日

概要：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．しかし，監視カメラ映像からの行動記録は飼育員が映像を目視で確認しながら記録をしており，負担が大きいものとなっている．そこで筆者らは，札幌市円山動物園の協力のもと，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．本稿では，行動記録の自動化のベースとなる個体識別を伴ったトラッキング手法の検討とその実践について述べる．提案手法は，物体検出，個体識別，トラッキングのそれぞれの既存手法を組み合わせる手法である．筆者らは，札幌市円山動物園のゾウ舎における 2 頭のゾウを対象に実験をおこない，提案手法の有効性を示し，さらに，実践を通して得られた知見をまとめた．

キーワード：物体検知，個体識別，トラッキング，画像認識，ゾウ，動物園，監視カメラ

Consideration and Practice on Tracking Methods for Zoo Elephants Using Single Surveillance Camera Footage

Ken Nishioka[1,2,a)]　Wataru Noguchi[5,b)]　Hiroyuki Iizuka[4,c)]　Masahito Yamamoto[3,4,d)]

Received: October 20, 2023, Accepted: March 15, 2024

Abstract: Behavioral observation of animals in zoos is indispensable for animal health management and improvement of the breeding environment. However, zookeepers currently record the behavior of animals from surveillance camera images while visually checking the images, which places a heavy burden on the zookeeper. Therefore, we have developed a new method to record the behavior of animals in the zoo keeping area using video surveillance cameras, in cooperation with the Sapporo Maruyama Zoo. We aim to automate the recording of animal behaviors (ethograms) from surveillance camera images of zookeepers. In this paper, we propose a tracking method with identification as a basis for the behavior identification. In particular, we describe a method for tracking the location of individual elephants over a long period of time. Experimental results show the effectiveness of the method on two elephants in the elephant house. We created a dataset by acquiring images from surveillance cameras at the Maruyama Zoo and annotating the positions and individuals of the elephants. We evaluated the dataset for nearly 24 hours and obtained high accuracy with the proposed method.

Keywords: object detection, identification, tracking, image recognition, elephants, zoo, surveillance cameras

[1] 北海道大学 大学院情報科学院
Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060–0814, Japan
[2] 株式会社システム計画研究所
Research Institute of Systems Planning, Inc., Shibuya,Tokyo 150–0031, Japan
[3] 北海道大学 大学院情報科学研究院
Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060–0814, Japan
[4] 北海道大学 人間知・脳・AI 研究教育センター
Center for Human Nature, Artificial Intelligence, and Neuroscience, Hokkaido University, Sapporo, Hokkaido 060–0812, Japan
[5] 北海道大学 数理・データサイエンス教育研究センター
Education and Research Center for Mathematical and Data Science, Hokkaido University, Sapporo, Hokkaido 060–0812, Japan
a) nishioka@ist.hokudai.ac.jp
b) w.noguchi@mdsc.hokudai.ac.jp
c) iizuka@chain.hokudai.ac.jp
d) masahito@ist.hokudai.ac.jp

1. はじめに

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．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．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．

そこで筆者らは，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）の自動化をめざしている．本稿では，行動記録の自動化のベースとなる個体識別を伴ったトラッキング手法について提案する（**図 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．

本稿では，「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 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 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．また，実験では物体検出と個体識別を分けた手法のほうが精度が高いことを示した．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 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．筆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．

本稿の構成を述べる．2 章で本研究に用いるデータセットについて述べ，3 章では関連研究について説明する．4 章では本手法についての詳細を述べ，5 章では実施した実験手法と結果について報告する．6 章に実践から得られた知見について述べたあと，7 章にまとめを述べる．

2. データセット

本章では，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．

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 5 FPS（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

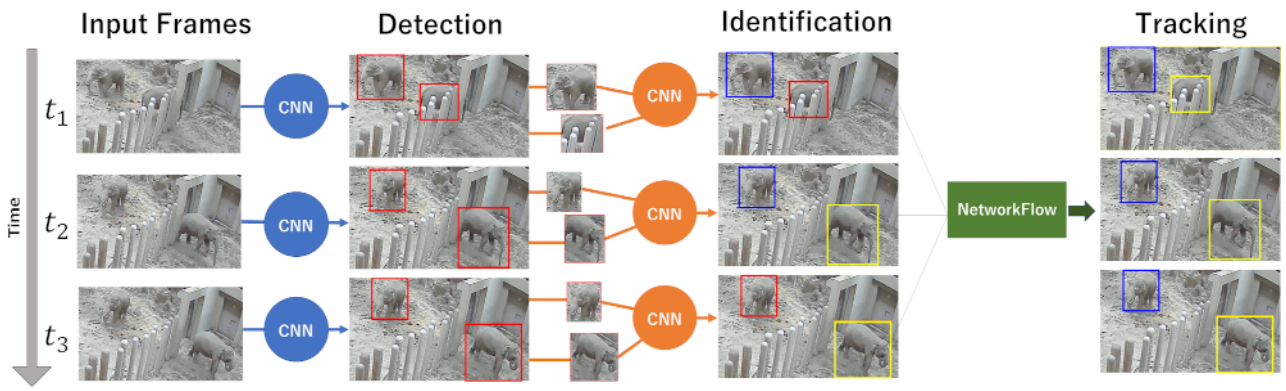

図 1　提案手法の概要図

Fig. 1　Overview of our proposed method.

図 2　監視カメラ映像
Fig. 2　Surveillance camera images.

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 2）.

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，ゾウの繁殖にかか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ないため，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．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できなくなり，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.　関連研究

関連研究としては，物体検出，個体識別，トラッキングがあげられる．本章では，この 3 つのタスクに関連する研究を述べる．

3.1　Detection

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．この物体検出によってトラッキングをおこなうことは，「tracking by detection」とも呼ばれる．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 [1], [2]．本提案手法も，この tracking by detection にもとづいている．

本節では実験に用いたモデルを数点挙げる．Tan らは EfficientNet [3] 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 [4]．このスケーリング手法により，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．これをさらに推し進め，精度を高めたものに YOLOv5 [5] がある．

また，動物に特化したモデルとしては，MegaDetector [6] 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 がベースとなっている．

3.2　Identification

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 のほかに，ResNet [7]，MobileNetV2 [8]，EfficientNet [3] 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，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 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．

次に，動物を対象とした個体識別の研究例を挙げる．Schofield らは，23 体のチンパンジーの顔識別のために VGG [9] を用いており，92.5% の正答率を達成したと述べている [10]．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 [11]．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おり，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 [12]．しかし，顔やキーポイントを利用する場合は，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．本研究では，学習データセットを作成する労力が高いことから，これらの手法は取らなかった．

3.3 Tracking

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 [1]．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ごと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．たとえば，本稿で扱う 5 FPS の動画には，24 時間分であっても 432,000（=5×24×3600）フレームに対してオス・メスの矩形位置を示す必要がある．本稿では，データセットとして 5 カ月間の動画を用いており，すべてのフレームに対してアノテーションを用意するのは現実的ではない．また，ゾウの成長や飼育環境の変化に応じてアノテーションをやりおなす必要があることを考慮し，今回は労力のかからない教師データを用いない方法を採用することとした．

また，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．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ほう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 [2]．行動記録の作成にリアルタイム性は必須ではないため，本稿ではネットワークフローを用いたオフライン手法を用いることにした．手法には主に Jiang ら，および Zhang らによる定式化をベースとしている [13], [14].

次に，動物を対象にしたトラッキングの研究例を述べる．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 による顔のトラッキングをおこなっている [9]．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．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 [15]．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）の特徴量と比較し，軌跡単位で識別 ID を割り当てている．筆者らの提案手法と同様に，idTracker も個体識別を伴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限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．本提案手法は，深層学習とネットワークフローを用いることで，複雑な背景や斜め上方向からのカメラ映像でもトラッキングを可能としており，画像上でのゾウのサイズが大きく変わる，施設の陰にゾウの姿が隠れる，といった状況にも適用できる．

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 [16]．ただし，前述したように，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．

4. 提案手法

前章にまとめたように，物体検知単独では正確なトラッキング結果を出力するのは難しいため，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

本章では，1，2，3 の各処理内容を詳述する．

4.1 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，モデルが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．

4.2 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 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たびたび発生する．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．ほかに，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

図 3　Identification データセット
Fig. 3　Dataset of identification.

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**）．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．

4.3　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ゾウの軌跡に対応する．次に，ネットワークから最小費用流を求め，それを各ゾウの軌跡とする．次に，互いのゾウが近接する場合，軌跡を分割する．最後に，軌跡ごとに識別 ID を割り当てる．以下にそれぞれについて詳述する．

4.3.1　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($1 \leqq n \leqq N$) 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），$K(\geq 2)$ 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 $\boldsymbol{d}_i^n (1 \leqq i \leqq M(n))$ を導入する．ここで，$\boldsymbol{d}_i^n$ は Detection 出力の予測位置 $\boldsymbol{r}_i = (r_i^{\mathrm{top}}, r_i^{\mathrm{left}}, r_i^{\mathrm{bottom}}, r_i^{\mathrm{right}})$，および，Identification の出力判別スコア $p_i \in \mathbb{R}^{K+1}$（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 の場合にあたるため，各フレーム対して 1 個のオクルージョンノードが存在する．

次に，各ノードは時間的に最も近い前後のフレームのすべてのノードとエッジで接続される．また，$\boldsymbol{d}_i^n$，$\boldsymbol{d}_j^m$ を検出ノードとしたとき，エッジ $\boldsymbol{d}_i^n \cdot \boldsymbol{d}_j^m$ に対するコスト関数 $c_{\boldsymbol{d}_i^n \cdot \boldsymbol{d}_j^m} = \mathrm{cost}(\boldsymbol{d}_i^n, \boldsymbol{d}_j^m)$ を以下のように定義する．

$$\mathrm{cost}(\boldsymbol{d}_i^n, \boldsymbol{d}_j^m) = \lambda c_{\mathrm{IoU}}(\boldsymbol{r}_i, \boldsymbol{r}_j) + (1-\lambda) c_{\mathrm{class}}(\boldsymbol{p}_i, \boldsymbol{p}_j), \tag{1}$$

$$c_{\mathrm{IoU}}(\boldsymbol{r}_i, \boldsymbol{r}_j) = 1 - \mathrm{IoU}(\boldsymbol{r}_i, \boldsymbol{r}_j), \tag{2}$$

$$c_{\mathrm{class}}(\boldsymbol{p}_i, \boldsymbol{p}_j) = 1 - \exp(-\mu \cdot KL), \tag{3}$$

$$KL(\boldsymbol{p}_i, \boldsymbol{p}_j) = \sum_{k=0}^{K} p_i^k \log \frac{p_i^k}{p_j^k}, \tag{4}$$

ここで，c_{IoU} は物体の位置の差に対する費用であり，$\mathrm{IoU}(\boldsymbol{r}_i, \boldsymbol{r}_j)$ は $\boldsymbol{r}_i$，$\boldsymbol{r}_j$ 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，$c_{\mathrm{class}}(\boldsymbol{p}_i, \boldsymbol{p}_j)$ 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，$0 \leq c_{\mathrm{class}} \leq 1$ となるようにする．式（1）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．

4.3.2　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．

$$J = \sum_{v,w} c_{vw} x_{vw}, \quad v, w \in V, \tag{5}$$

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 [13]．式から分かるように，フローは線形計画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 4** に示した．この例では，3 フレームにおいて生

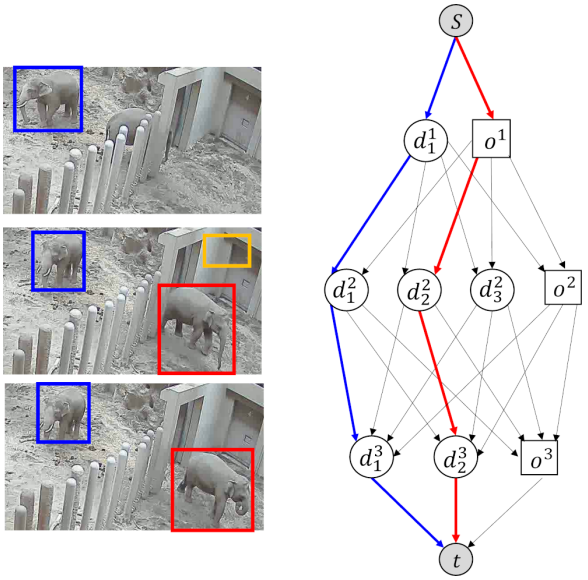

図 4　ネットワークフロー
Fig. 4　Network flow.

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．

4.3.3　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．

4.3.4　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．だだし，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．

5.　実験

本章では，4 章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．

表 1　Detection データセット
Table 1　Detection dataset.

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

5.1　Detection

5.1.1　データセット

実験で用いたカメラ映像は 2 章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 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．

5.1.2　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 [17] 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．

5.1.3　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．

5.1.4　結果

結果を**表 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，より精度が良いほうが好ましいが，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．

5.2　Identification

5.2.1　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** 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 [8]，ResNet50 [7]，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．

5.3　Tracking

本節では，5.1，5.2 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 τ_{div}=0.2 と設定した．また，20 フレーム内で 1 頭もゾウが検出されていない場合は，その時点でネットワークを分け別々に解析をおこなう．

表 2　Detection 結果
Table 2　Detection result.

モデル	AP
EfficientDet-D6	0.779
YOLOv5-l6	**0.810**
MegaDetector v5.0	0.804

表 3　Identification データセット
Table 3　Identification dataset.

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

5.3.1　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 4** 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 [18] 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りうる．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，たとえば，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．

5.3.2　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 [19]，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 MOT Challenge [18] の評価値として採用されている．軌跡が途中で

表 4　トラッキングデータセット
Table 4　Tracking dataset.

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

入れ替わる ID スイッチが評価値に反映される．ただし，HOTA，MOTA，IDF1 は途中で ID スイッチが起きたとき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．

5.3.3　コスト関数

4.3.1 項のコスト関数の，λ=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）cost(v, w)=c_{IoU},
2. （Class Score）cost(v, w)=c_{class},
3. （IoU + Class Score）cost(v, w)=(c_{IoU}+c_{class})/2,

ここで，クラススコアの違いに対する費用 c_{class}（式 (3)）のパラメータ μ を，$0 \leq c_{\mathrm{class}} \leq 1$ となるように μ =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．

5.3.4　結果

結果を**表 5** 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分かった．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分かる．

5.4　Detection と Identification の統合と分離

Detection と Identification の統合するモデル（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 1 章の 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 6** 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．

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（EfficinetNetB2）を学習する際は，表 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 5**，**図 6** 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 5 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 6 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．

5.5.1　精度が低下するケース

本項では，精度が低下するケースを 3 点述べる．まず，各軌跡上にメス・オス判定に有効な画像が少ない場合である．「ゾウが見切れて映る」「カメラと逆方向に向いてい

表 6　統合・分離モデル比較
Table 6　Integration and separation model comparison.

Model	mAP
統合モデル	0.656
分離モデル	**0.664**

表 5　Tracking 結果
Table 5　Tracking result.

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**

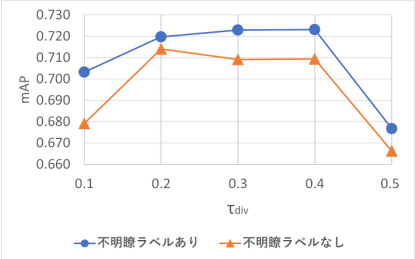

図 5　不明瞭ラベルの効果（軌跡分割基準）

Fig. 5　Effects of unclear labels with trajectories division criteria.

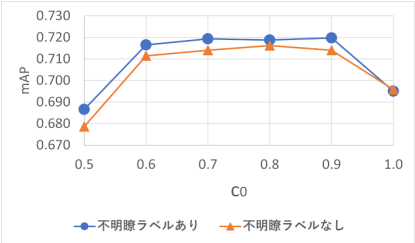

図 6　不明瞭ラベルの効果（オクルージョンコスト）

Fig. 6　Effects of unclear labels with occulusion cost.

る」「カメラから遠い位置にいる」等の場合はメス・オス判別に有効な牙が見えづらく，個体判別率が落ちる．次に，Detection が誤検出する場合である．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．最後に，映像にゾウが 1 頭しか映らない場合である．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．

6.　実践から得られた知見

動物のトラッキングの特性の 1 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 5 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も分かる．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（たとえば，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分かった．これは，表 5 の NF（IoU）が NF（Class Score）と比べて精度が高かったことから分かる．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．

7.　まとめ

筆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 0.820 と十分な精度が得られたうえ，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．

謝辞　本研究は JSPS 科研費 22H03637 の助成を受けたものです．また，本研究遂行にあたっては，札幌市円山動物園の協力をいただきました．この場を借りて感謝申し上げます．

参考文献

[1] Ciaparrone, G., Sánchez, F. L., Tabik, S., Troiano, L., Tagliaferri, R. and Herrera, F.: Deep learning in video multi-object tracking: A survey, *Neurocomputing*, Vol.381, pp.61–88 (2020).
[2] Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W. and Kim, T.-K.: Multiple object tracking: A literature review, *Artificial intelligence*, Vol.293, p.103448 (2021).
[3] Tan, M. and Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks, *International conference on machine learning*, PMLR, pp.6105–6114 (2019).
[4] Tan, M., Pang, R. and Le, Q. V.: Efficientdet: Scalable and efficient object detection, *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, pp.10781–10790 (2020).
[5] Glenn., J.: YOLOv5, (online), available from ⟨https://ultralytics.com/yolov5⟩.
[6] Beery, S., Morris, D. and Yang, S.: Efficient Pipeline for Camera Trap Image Review, arXiv preprint arXiv: 1907.06772 (2019).
[7] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.770–778 (2016).
[8] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks, *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.4510–4520 (2018).
[9] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *International Conference on Learning Representations* (2015).
[10] Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D. and Carvalho, S.: Chimpanzee face recognition from videos in the wild using deep learning, *Science Advances*, Vol.5, No.9, pp.1–9 (online), DOI: 10.1126/sciadv.aaw0736 (2019).
[11] Pang, Y., Yu, W., Zhang, Y., Xuan, C. and Wu, P.: Sheep face recognition and classification based on an improved MobilenetV2 neural network, *International Jour-*

nal of Advanced Robotic Systems, Vol.20, No.1 (online), DOI: 10.1177/17298806231152969 (2023).
[12] Moskvyak, O., Maire, F., Dayoub, F. and Baktashmotlagh, M.: Learning landmark guided embeddings for animal re-identification, *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp.12–19 (2020).
[13] Jiang, H., Fels, S. and Little, J. J.: A linear programming approach for multiple object tracking, *2007 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp.1–8 (2007).
[14] Zhang, L., Li, Y. and Nevatia, R.: Global data association for multi-object tracking using network flows, *2008 IEEE conference on computer vision and pattern recognition*, IEEE, pp.1–8 (2008).
[15] P ´ erez-Escudero, A., Vicente-Page, J., Hinz, R. C., Arganda, S. and De Polavieja, G. G.: IdTracker: Tracking individuals in a group by automatic identification of unmarked animals, *Nature Methods*, Vol.11, No.7, pp.743–748 (online), DOI: 10.1038/nmeth.2994 (2014).
[16] He, Y., Wei, X., Hong, X., Shi, W. and Gong, Y.: Multi-Target Multi-Camera Tracking by Tracklet-to-Target Assignment, *IEEE Transactions on Image Processing*, Vol.29, pp.5191–5205 (online), DOI: 10.1109/TIP.2020.2980070 (2020).
[17] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft coco: Common objects in context, *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014*, Proceedings, Part V 13, Springer, pp.740–755 (2014).
[18] Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S. and Leal-Taixé, L.: MOTChallenge: A Benchmark for Single-camera Multiple Target Tracking, *International Journal of Computer Vision*, pp.1–37 (2020).
[19] Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L. and Leibe, B.: HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking, *International Journal of Computer Vision*, pp.1–31 (2020).

西岡 拳（学生会員）

2017 年北海道大学理学院数学専攻修士課程修了．同年，株式会社システム計画研究所入社，現職．2020 年北海道大学情報科学院情報科学専攻博士課程入学，同大学在学中．

野口 渉（正会員）

2019 年北海道大学大学院情報科学研究科博士後期課程修了．同年同大学人間知・脳・AI 研究教育センター博士研究員．2021 年同大学大学院情報科学研究院博士研究員．2023 年同大学数理・データサイエンス教育研究センター特任助教となり現在に至る．博士（情報科学）．深層学習を用いた認知モデルの研究に従事．

飯塚 博幸（正会員）

2004 年，東京大学総合文化研究科博士課程修了．2005 年，日本学術振興会特別研究員（PD，はこだて未来大学），イギリスサセックス大学客員研究員．2008 年，大阪大学大学院情報科学研究科助教．2013 年，北海道大学大学院情報科学研究科准教授．2024 年より，同大学人間知・脳・AI 研究教育センター准教授．専門は人工生命，複雑系科学，認知科学．博士（学術）．

山本 雅人（正会員）

1996 年北海道大学大学院工学研究科システム情報工学専攻博士後期課程修了．同年日本学術振興会特別研究員（PD）．1997 年北海道大学大学院工学研究科助手．2000 年同大学院工学研究科助教授．同大学院情報科学研究科助教授を経て，2007 年北海道大学大学院情報科学研究科准教授．この間，科学技術振興機構さきがけ研究員，デューク大学客員研究員．博士（工学）．現在は，人工知能技術の実社会応用，ゲーム情報学，スポーツ情報学等の研究に従事．NPO 法人観光情報学会副会長．，人工知能学会，計測自動制御学会，精密工学会等，各会員．