一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2023-77(2024-03)

映像中の微小な鳥に対する検出領域絞込みと検出履歴を考慮した追跡

リュウ　テイイ†-†† 川西　康友††-† 駒水　孝裕†　井手　一郎†

† 名古屋大学大学院情報学研究科
†† 理化学研究所ガーディアンロボットプロジェクト
E-mail: †tingweiryu@outlook.com, ††yasutomo.kawanishi@riken.jp, †††taka-coma@acm.org,
††††ide@i.nagoya-u.ac.jp

あらまし　本研究では，広角映像中に小さく映った鳥の追跡を目的とする．画像の大きさに対して追跡対象の大きさが十分に小さい場合（小物体追跡），物体の検出と対応付けにおいて，それぞれ問題が生じる．これらの問題点に対して，本研究では，検出において検出候補領域の絞り込み（Adaptive Slicing Aided Hyper Inference；Adaptive SAHI）を，対応付けにおいて検出履歴を考慮した類似度判断基準（Detection History-aware Similarity Criterion；DHSC）を，各々提案したことで，鳥の追跡速度と精度の向上を実現した．NUBird2022 データセットを用いた実験により，提案手法の有効性を検証し，MOTA，IDF_1 と検出時間の明らかな改善を示した．

キーワード　小物体追跡，検出候補領域，類似度

1. ま え が き

鳥の移動行動，特に木々の間での移動は，採食や繁殖といった生態学的に重要な活動の理解において不可欠である．そのため，鳥を詳細に追跡する技術が求められている．鳥を追跡するためには，様々なセンサが候補に挙げられるが，広範囲にわたる移動を 1 台の装置で観測できるため，広角カメラの利用が適している．しかし，広角映像中の鳥は同距離から普通のカメラで撮影した鳥よりも見えが相対的に小さいため，追跡が困難である．本研究では，このような映像中に小さく映った鳥の追跡を目的とする．

物体追跡では，Tracking-by-Detection という各画像からの物体検出とフレーム間での検出物体の対応付けの 2 つの処理を繰り返す方法が主流である．本研究のように，画像の大きさに対して追跡対象の大きさが十分に小さい場合（小物体追跡），物体の検出と対応付けにおいて，それぞれ問題が生じる．まず，物体検出については，Slicing Aided Hyper Inference（SAHI）[1] のような小物体検出手法は，画像を小領域（検出候補領域）に分割し，各検出候補領域に対して検出処理を適用するが，画像中で実際に鳥を含む領域はごく一部であり，多くの領域で無駄な計算が行なわれる．一方，対応付けについては，時系列データ（以前のフレームの検出領域）を用いて，現フレームにおける物体領域を予測し，過去の検出領域と照合する手法が一般的である．その際に，領域の重なりの割合（Intersection-of-Union；IoU）と外見特徴を基準とすることが一般的であるが，鳥の見えは小さく，動きが速いため，隣接フレームであっても検出した領域同士が重ならないことが多い．また，見えが小さな鳥に対して，個体を識別できるほどの外見特徴を抽出することは容易ではない．さらに，木の間を鳥が移動すると，遮蔽が頻発するため，時系列予測を信頼できないことがある．

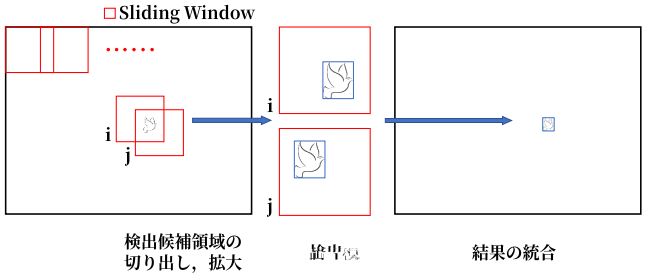


図 1: SAHI による小物体検出処理．

これらの問題点をふまえて，本研究では，検出に対して Adaptive Slicing Aided Hyper Inference（Adaptive SAHI）という検出候補領域の絞り込みを，対応付けに対して Detection History-aware Similarity Criterion（DHSC）という検出履歴を考慮した類似度判断基準を，各々提案することで，鳥の追跡速度と精度の向上を実現する．

本研究の貢献は以下の通りである．

- Adaptive SAHI という領域の絞込みを使用した検出法を提案し，検出候補領域を直前フレームで検出した領域周辺に絞ることにより，高速化するとともに，誤検出を低減する．
- DHSC という検出履歴を考慮する類似度判断基準を提案し，予測が信頼できない時でも安定して追跡できるようにする．

2. 関 連 研 究

2.1 小物体検出

一般物体検出とは，指定された種類の物体について，画像中の位置を推定するタスクである．近年盛んな深層学習により，画像中の見えが大きい物体については高精度で検出することができるようになった．しかし，ニューラルネットワークの制約上，入力画像を特定の大きさ（一般に元の画像よりも小さい）

に拡縮して処理する必要がある．つまり，高解像度な画像を入力すると，画像中の物体は拡縮後にとても小さくなる．この問題は，小物体検出タスクにおいて致命的である．そこで小物体検出に特化した SAHI [1] が提案されている．SAHI では図 1 に示すように，重複を許して小さな Sliding Window で画像全体を走査し，検出候補領域を得る．次に，各検出候補領域を拡大して任意の物体検出器に入力する．最後に，全ての検出結果を元の画像の座標系で統合する．この手法は確かに小物体検出には有効であるが，拡大した検出候補領域ごとに物体検出を行なうため，Sliding Window の数に応じて計算量が大きくなる．また，画像中の物体数が少ない場合には，実際に物体を含む領域はごく一部であり，多くの領域で無駄な計算が行なわれる．この問題に対して，音声情報も利用し，音源定位により，物体検出候補領域を絞込む手法が提案されている [2]．しかし，この手法では音声情報の取得が必要なだけでなく，常に高品質な音声情報を取得できるとは限らないため，安定して利用できない問題点がある．

2.2 物体追跡

物体追跡とは，複数フレームに写った同一の物体に対して一貫した ID を付与するタスクである．深層学習の発展により，物体検出の精度が向上したことから，TbD による物体追跡の精度が向上している．TbD では，各フレームの物体を照合して類似度が高いものを対応付ける（同じ ID を付与する）．TbD の代表的な手法である Simple Online and Realtime Tracking（SORT）[3] では，Kalman フィルタ [4] を用いて物体の状態（位置と速度）を推定し，物体検出器からの検出結果と照合して対応付ける．単純な運動モデルを用いる場合でも高速高精度な追跡を実現している．しかし，位置や速度に基づいているために，遮蔽が発生し検出できない事があると，対応付けに失敗する弱点がある．これに対して，DeepSORT [5] は，SORT に物体の外見特徴に基づく類似度を導入し，遮蔽による物体の消失からの再出現時に対応付けを誤る問題に対処している．さらに，StrongSORT [6] は，DeepSORT の各モジュールを更新し，画像特徴量の改善，Gaussian 平滑補間による軌跡の修復などを加え，物体追跡性能を向上させている．また，Kalman フィルタは，遮蔽や非線型的な移動が生じて追跡が途切れた場合も予測を続けるので，新しい検出による修正がないため，物体の状態の推定誤差が蓄積し，追跡が回復した後に再度追跡を失敗しやすい問題がある．Observation-Centric SORT（OC-SORT）[7] は，追跡が回復した後に，以前の検出結果を利用して，物体状態の再更新をすることで，長時間の安定した追跡を実現している．

遮蔽と非線型的な移動における Kalman フィルタの問題に関して，DeepSORT と StrongSORT は外見特徴を利用し，中断した追跡の回復を実現している．しかし，小物体追跡の場合，物体の解像度が非常に低いため，個体を識別できる外見特徴の抽出が難しい．また，OC-SORT は追跡が復帰した後の維持に注目しているが，どのように復帰するかは依然問題として残っている．

2.3 類似度

追跡における対応付けのためには，これまで追跡してきた領域と，現フレームで検出した領域との類似度を算出する必要がある．類似度の尺度には領域の重なりの割合を表す Intersection-of-Union（IoU）と，外見特徴の余弦尺度がよく利用される．

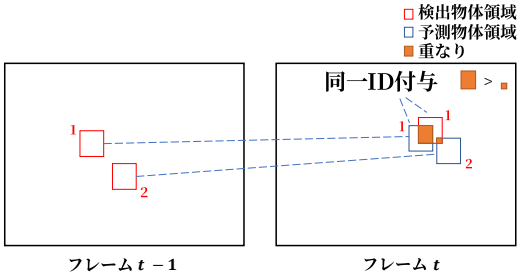
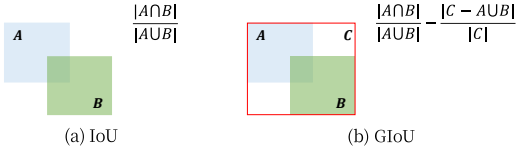
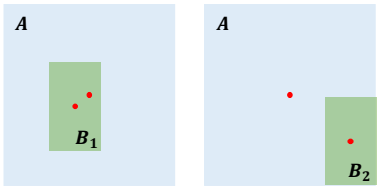


図 2: IoU による対応付け．



(a) IoU　　(b) GIoU

図 3: IoU と GIoU の計算．



$\mathrm{IoU}(A,B_1) = \mathrm{GIoU}(A,B_1) = \mathrm{IoU}(A,B_2) = \mathrm{GIoU}(A,B_2)$

図 4: IoU と GIoU が 2 つの領域の類似度を区別できない例．

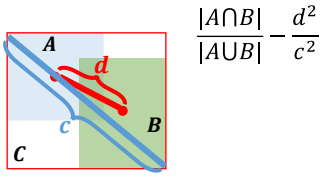


図 5: DIoU の計算．

IoU による対応付けは，図 2 のように，時系列フィルタによる各予測領域に対して，現フレームから検出されたそれぞれの検出領域との重なりの割合を計算し，最も高い検出領域に対してその追跡対象の ID を付与する．IoU は位置だけではなく，大きさの類似度も反映できるが，物体領域が全く重ならない時，その類似度を反映できない．鳥のような対象の場合，動きが速いため，物体領域が重ならない状況が多発する．その場合，すべての IoU が 0 となり，どの領域が類似しているか判断できない．

これに対して，Generalized IoU（GIoU）[8] は，IoU の概念を拡張することによって，物体領域が全く重なっていなくても類似度を計算できるようにしたものである．図 3(b) のように，領域　と　の重なりの割合だけではなく，領域　と　を囲む最小の領域　に着目し，領域　のうち，領域　と　以外の部分が領域　に占める割合も計算に入れる．

一方，Distance-IoU（DIoU）[9] は，図 4 に示すような状況に

図 6: 提案手法の処理手順

おいて，IoU と GIoU が 2 つの領域の類似度を区別できない問題を指摘し，図 5 のように，2 つの領域の中心点距離 *3* と 2 つの領域を囲む最小の領域の対角線の長さ *2* の割合を考慮することで，この問題を解決している．

3. 提 案 手 法

提案手法の処理手順を図 6 に示す．処理は検出候補領域の絞り込みによる検出と，検出履歴を考慮した類似度判断基準による対応付けの 2 つの部分から構成される．物体検出においては，SAHI [1] に従って，入力画像を分割して生成した検出候補領域群をそのまま利用するのではなく，追跡中の物体の予測位置に基づくサンプリングと一様サンプリングにより，検出候補領域の効率的な絞り込みをする．一方，対応付けにおいては，位置の類似度による照合と外見の類似度による 2 段階照合を行なう．位置の類似度 DIoU [9] の計算には，時系列フィルタにより予測した位置だけでなく，直近で検出した物体の位置も利用し，検出履歴を考慮した対応付けを行なう．この対応付けに漏れたものに対してのみ，外見特徴による対応付けをする．

3.1 検出候補領域の絞り込みによる高速な物体検出

SAHI では，対象物体を含まない多くの候補領域に対しても検出処理を適用するため，検出効率が悪いだけでなく，対象物体が存在しない可能性が高い領域に検出処理を適用することは，誤検出の増加を招きかねない．物体追跡は連続したフレームに対する処理であるため，直前フレームで検出した位置付近に追跡対象物体が存在する可能性が高い．提案手法では，検出候補領域をその周辺に絞ることで，高速化するとともに，誤検出を低減する．一方，追跡結果を利用した絞り込みだけでは新規物体の出現に対処できない．そのため，絞り込んだ領域以外の検出候補領域に対する非復元抽出の一様サンプリングによって得られる検出候補領域に対しても検出処理を適用する．これにより，新しい物体の発見だけでなく，遮蔽により追跡を中断した物体の再発見も期待できる．この検出候補領域のサンプリング手法を Adaptive SAHI と名付ける．

SAHI が合計 # 個（すべての検出候補領域）の検出候補領域に対して検出処理を適用するのに対して，提案手法では合計 ″ 個（″ ≪ #）の検出候補領域を選択して適用する．そのうち，追跡結果に基づいて物体が存在しそうな領域周辺の　個（　Ÿ ″）をサンプリングし，残り ″ −　個を画像全体から一様サンプリングする．物体が存在しそうな領域の判断は，物体追跡結果に基づく物体の存在確率分布による．時系列フィルタとして Kalman フィルタを用いる場合には，予測された物体位置の分布を，そのまま物体の存在確率分布として用いる．

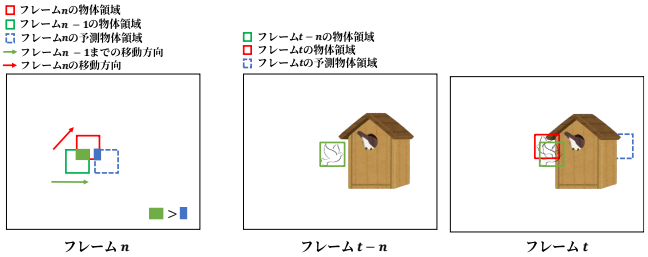
予測分布からのサンプリングによる検出の処理手順は以下の通りである．

（1） 画像全体に対して，SAHI と同様に Sliding window により物体検出候補領域の集合を得る．

（2） 追跡中の各物体に対して，個々の物体の存在確率分布の平均値に相当する点を囲む複数の物体検出候補領域を無作為に　個サンプリングする．

（3） 画像全体から ″ −　個の検出候補領域を一様分布に従ってサンプリングする．

（4） サンプリングした各検出候補領域に対して物体検出を



(a) 非線型的な移動(転向)　　(b) 出入による遮蔽

図 7: 2 つの類似度計算対象の比較.

適用し，物体領域を検出する．

（5） 各検出結果を元の画像の座標系で統合する．

3.2 検出履歴を考慮した類似度判断基準による高精度な物体追跡

物体追跡における対応付けは，それまで追跡してきた物体領域と，現フレームで検出した物体領域とを照合し，その類似度に応じて対応付けを行なうことにより，各物体に対して一貫した ID を付与することである．提案手法では，高精度な対応付けを実現するために，位置に基づく照合と，外見特徴に基づく照合の 2 段階照合で対応付けを行なう．

2.2 節で述べたように，類似度計算の対象について，Kalman フィルタの予測が信頼できない場合がある．対応付けは検出結果に基づく処理なので，最も信頼できるのは検出結果である．そのため，提案手法はそれまでの検出履歴も類似度計算の対象にする．それまでの検出履歴と予測結果の両方と現在の検出結果との類似度を計算し，その重み付き和を計算することで，予測が信頼できない時でも有効な類似度計算となり，精度向上が期待できる．Kalman フィルタの予測が信頼できない 2 つの状況を図 7 に示す．まず，図 7(a) は非線型的な移動である．鳥が移動する際に，急転向する場合があり，非線型的な移動となる．このような場合，予測位置よりも，前フレームでの検出位置の方が現在の検出位置に近い．一方，図 7(b) は，鳥の巣箱などの出入による遮蔽の状況である．フレーム = で鳥が遮蔽物に入った場合，Kalman フィルタは鳥が消失する前の移動方向・移動量に基づいて予測し続けるため，フレーム = + : では図の右上の青い破線枠の場所にいると予測し，対応付けを誤る．実際は出入による遮蔽であるため，物体が再び出現する時は消失する時と同じ場所にいるので，直近の検出履歴とのなら対応付けができる．

鳥の見えは小さく，また動きが速いため， 2.3 節で紹介した DIoU を 1 段階目の照合に用いる．DIoU により，物体領域の大きさと中心点の距離に着目した類似評価をする．1 段階目の照合によって対応付けできなかった対象に対してのみ，2 段階目の照合をする．2 段階目の照合では，外見特徴同士の余弦類似度を用いる．つまり，物体の外見特徴を抽出して，特徴ベクトルがなす角の大きさで類似度を評価する．

ある直近 = フレーム（= ≥ 0）で検出されていない追跡対象 $-^{i} = (-^{i}_{t-n}, -^{i}_{t-n-1}, \cdots)$ について，現フレームで検出したある物体の位置 $\widehat{-}^{j}_{t}$ との類似度を計算する．時刻 t における追跡対象 $-^{i}_{t}$ の予測位置を $-^{i}_{t|t-n}$ とすると，

$$5_1(-^{i}, \widehat{-}^{j}_{t}) = U''(-^{i}_{t|t-n}, \widehat{-}^{j}_{t}) + (1-U)''(-^{i}_{t-n}, \widehat{-}^{j}_{t}) \quad (1)$$

と書ける．ただし，" は 2 つの物体領域の類似度を表す関数，U は重み付け和の重みである．関数 " には，DIoU を用いる．この DIoU の重み付き和を用いた類似度を Detection History-aware DIoU（DH-DIoU）と名付ける．

外見特徴に基づく照合について，追跡対象の直近の検出領域 $-^{i}_{t-n}$ に対応する外見特徴を $\mathbf{f}^{i}_{t-n}$，現フレームで検出した物体領域 $\widehat{-}^{j}_{t}$ に対応する外見特徴を $\widehat{\mathbf{f}}^{j}_{t}$ とする．ここで， を特徴抽出器，時刻 t における観測画像を $_t$ としたときに $\widehat{\mathbf{f}}^{j} = (_t, \widehat{-}^{j}_{t})$，$\mathbf{f}^{i}_{t-n} = (_{t-n}, -^{i}_{t-n})$ である．

適用し，データセットを増強した．具体的には，両セットの画像に対して，重複を許して小さな Sliding Window で画像全体を走査して，鳥を含む領域を切り出したうえで，拡縮して元のセットに追加した．切り出し方として，鳥を含む画像は，少なくとも 1 羽の鳥が存在する範囲を切り出して学習セットあるいは検証セットに入れて増強した．鳥を含まない背景だけの画像は，全ての切り出しを学習セットに入れて増強した．Sliding Window の大きさは 256×128 画素，重複率は 0.25，拡縮後の解像度は 640×320 画素である．分割と切り出しの結果は以下の通りである．

- 検出器用学習セット：27,395 枚の画像，そのうち，元の画像は 9,095 枚（解像度 4-096 × 2-048 画素），増強画像は 18,300 枚（解像度 640×320 画素）である．
- 検出器用検証セット：9,157 枚の画像，そのうち，元の画像は 3,039 枚（解像度 4-096 × 2-048 画素），増強画像は 6,118 枚（解像度 640×320 画素）である．
- 評価セット：
  - シーン 1：874 枚（フレーム番号 1,690~2,563），解像度は 4-096 × 2-048 画素である．鳥が連続して交叉移動するシーンを含む．
  - シーン 2：1,451 枚（フレーム番号 10,490~11,940），解像度は 4-096 × 2-048 画素である．巣の入り口で 2 羽の鳥が交互に出入りするシーンを含む．

4.2.3 評 価 指 標

本実験に用いる評価指標は以下の通りである．

- False Positive（FP）：誤検出の数
- False Negative（FN）：見落しの数
- IDentity switches（IDsw）[13]：追跡 ID 入替わりの回数
- Multi-Object Tracking Accuracy（MOTA）[14]：次式で定義される複数物体追跡精度．

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDsw}_t)}{\sum_t \mathrm{GT}_t} \tag{3}$$

ただし，時刻 t において，FN_t は見落しの数，FP_t は誤検出数，IDsw_t は追跡 ID 入替わりの回数，GT_t は正解物体領域数である．

- IDF_1 Score（IDF_1）：次式で定義される ID に対する IDF_1 Score．

$$\mathrm{IDF}_1 = \frac{2\mathrm{IDTP}}{2\mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}} \tag{4}$$

ただし，IDTP，IDFP，IDFN は正確な ID の数（フレーム数），誤った ID の数（フレーム数），見落した ID の数（フレーム数）をそれぞれ表す．

- Detection Time（DT）：全フレームの検出に要する時間［秒］．

4.3 実 験 結 果

評価セットの 2 シーンに対する実験結果を表 2 に示す．追跡結果の実例を図 8 に示す．検出候補領域の絞り込みは検出時間を大幅に削減するだけでなく，MOTA ではシーン 1 で 134.6 ポイント，シーン 2 で 134.2 ポイント向上した．一方，IDF_1 ではシーン 1 で 4.3 ポイント，シーン 2 で 23.1 ポイント向上した．

表 1: 評価セットの 2 シーンにおけるベースライン手法と提案手法の性能比較．

(a) シーン 1

手法	MOTA↑	IDF1↑	FP↓	FN↓	IDsw↓	DT[秒]↓
ベースライン	−119.6	20.2	3,049.0	697.0	91.0	3,563.6
提案	**15.0**	**24.5**	**776.0**	**700.0**	**9.0**	**135.0**

(b) シーン 2

手法	MOTA↑	IDF1↑	FP↓	FN↓	IDsw↓	DT[秒]↓
ベースライン	−64.5	22.3	3,435.0	644.0	150.0	5,898.7
提案	**69.7**	**45.4**	**264.0**	**479.0**	**36.0**	**179.3**

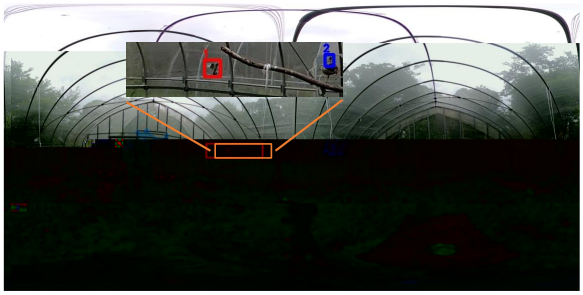
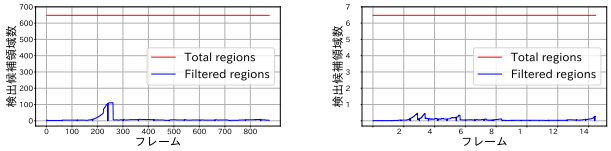


図 8: 追跡結果の実例．



(a) シーン 1　(b) シーン 2

図 9: 検出候補領域数の変化．

検出時間ではシーン 1 で 3,428.6 秒，シーン 2 で 5,719.4 秒短縮した．MOTA が大幅に向上した原因は，FP と IDsw が大きく減少したためである．FP はシーン 1 で 2,273 個，シーン 2 で 3,171 個減少した．そして IDsw はシーン 1 で 82 回，シーン 2 で 114 回減少した．

SAHI は鳥を含まない領域に多数の誤検出が発生し，MOTA は 0 以下になった．提案手法による領域の絞り込みは，検出時間を大幅に短縮しただけではなく，誤検出の発生も大きく削減した．図 9 は評価セットの 2 シーンにおける検出候補領域数の変化を示す．提案手法の検出履歴を考慮した類似度判断基準は，以前の検出結果も用いることで，IDF_1 の向上を実現した．そして外見特徴ではなく，DIoU [9] を 1 段階目の照合に用いることで，IDsw を大きく削減した．

4.4 考　　察

鳥の非線型的な移動と遮蔽が発生した状況を同時に含んでいる複雑なシーンであるシーン 2 について考察を加える．

前節で提案手法の有効性を確認したが，提案手法のうち追跡精度の改善に寄与した部分は不明であった．そこで，照合方法と指標の様々な組み合わせについて実験した．ここでは 2 段階

表 2: 照合方法と指標の組み合わせに関する実験結果.

1 段階照合	2 段階照合	MOTA↑	IDF_1↑	IDsw↓
外見特徴	—	81.8	30.7	176
IoU	—	98.9	33.5	11
DIoU	—	**99.0**	40.6	8
DH-DIoU	—	**99.0**	41.4	6
外見特徴	IoU	89.3	37.5	99
外見特徴	DIoU	90.0	46.4	150
外見特徴	DH-IoU	86.6	42.7	120
外見特徴	DH-DIoU	96.7	**51.2**	51
DIoU	外見特徴	**99.0**	40.7	8
DH-DIoU	外見特徴	**99.0**	50.5	**5**

照合以外，各指標独自の性能を確認するため，1 段階照合のみの手法も比較した．また，物体検出の失敗の影響を排除して追跡部分のみを評価するため，検出領域はデータセットで真値として与えられたものを用いた．実験結果を表 2 に示す．ここで，DH-DIoU は提案する Detection History-aware DIoU を表す．

まず，各 2 段階照合手法は，対応する 1 段階照合手法と比べて，すべて同等以上の精度を示したことから，2 段階照合が精度向上に役立つことを確認した．1 段階照合手法では，外見特徴を用いた場合が最も低精度になり，また 2 段階照合手法でも 1 段階目で外見特徴を用いると多くの ID 入替わりが発生した．その原因は，本シーンでは鳥をうまく識別できる外見特徴の抽出が難しかったためと考えられる．次に，DH-DIoU は 2.2 節で述べたフレーム内で物体領域が重ならない問題と，Kalman フィルタによる予測が信頼できない場合がある問題を改善できていることがわかる．2 行目と 3 行目，5 行目と 6 行目の差に着目すると，IoU を DIoU に切り替えることで，IDF_1 がそれぞれ 7.1 ポイントと 8.9 ポイント向上したことがわかる．また，3 行目と 4 行目，6 行目と 8 行目の差に着目すると，DIoU を DH-DIoU に切り替えることで，IDF_1 がそれぞれ 0.8 ポイントと 4.8 ポイント向上したことがわかる．外見特徴 → DH-DIoU の組み合わせは最も高い IDF_1 を得たが，総合的に評価すると，DH-DIoU → 外見特徴の方が良かった．そのため，提案手法である，DH-DIoU を 1 段階目，外見特徴を 2 段階目の照合に用いることの有効性を確認できた．

5. むすび

本研究では，広角映像中の見えが小さな鳥の追跡のために，Tracking-by-Detection のアプローチにおける検出・追跡の精度と速度を向上させることを目的とする．小物体検出について，SAHI のように画像を小領域に分割して処理する手法において，物体を含む領域が画像全体のごく一部である場合に，多くの領域で無駄な計算が行なわれる問題がある．また対応付けについて，追跡から予測した物体領域と現フレームで検出した物体領域の IoU が重ならない問題と，非線型的な移動と遮蔽によって予測が信頼できない状況が生じる問題がある．これらの問題点をふまえて，本研究では，検出に対して検出候補領域の絞り込みを，対応付けに対して検出履歴を考慮した類似度判断基準を，各々提案したことで，鳥の追跡速度と精度の向上を実現した．

NUBird2022 というデータセットに評価実験を行ない，提案手法の有効性を確認した．今後の課題として，提案手法を他モデルへの応用，マルチモーダルな追跡や未知クラスの検出と追跡への拡張を検討している．

謝　辞

本研究は JSPS 科研費 JP20H00475, JP21H03519 の助成を受けた．

文　献

[1] F.C. Akyon, S.O. Altinuc, and A. Temizel, “Slicing aided hyper inference and fine-tuning for small object detection,” Proceedings of the 2022 IEEE International Conference on Image Processing, pp.966–970, Oct. 2022.

[2] Y. Kawanishi, I. Ide, B. Chu, C. Matsuhira, M.A. Kastner, T. Komamizu, and D. Deguchi, “Detection of birds in a 3D environment referring to audio-visual information,” Proceedings of the 18th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp.1–7, Nov. 2022.

[3] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” Proceedings of the 2016 IEEE International Conference on Image Processing, pp.3464–3468, Sept. 2016.

[4] R.E. Kalman, “A new approach to linear filtering and prediction problems,” Journal of Basic Engineering, vol.82, no.1, pp.35–45, March 1960.

[5] N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” Proceedings of the 2017 IEEE International Conference on Image Processing, pp.3645–3649, Sept. 2017.

[6] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, “StrongSORT: Make DeepSORT great again,” IEEE Transactions on Multimedia, vol.25, pp.8725–8737, Jan. 2023.

[7] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, “Observation-Centric SORT: Rethinking SORT for robust multi-object tracking,” Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9686–9696, June 2023.

[8] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.658–666, June 2019.

[9] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-IoU loss: Faster and better learning for bounding box regression,” Proceedings of the 34th AAAI Conference on Artificial Intelligence, pp.12993–13000, June 2020.

[10] H.W. Kuhn, “The Hungarian method for the assignment problem,” Naval Research Logistics Quarterly, vol.2, no.1–2, pp.83–97, March 1955.

[11] G. Jocher, “Ultralytics YOLOv5,” 2020. https://github.com/ultralytics/yolov5/ (Accessed Jan. 29, 2024)

[12] S. Sumitani, R. Suzuki, T. Arita, K. Nakadai, and H.G. Okuno, “Non-invasive monitoring of the spatio-temporal dynamics of vocalizations among songbirds in a semi free-flight environment using robot audition techniques,” Birds, vol.2, no.2, pp.158–172, April 2021.

[13] Y. Li, C. Huang, and R. Nevatia, “Learning to associate: Hybrid-Boosted multi-target tracker for crowded scene,” Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.2953–2960, June 2009.

[14] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: The Clear MOT metrics,” EURASIP Journal on Image and Video Processing, vol.2008, no.246309, pp.1–10, May 2008.