スポーツ選手を対象とする姿勢特徴を考慮した複数人物追跡

小林 万葉[1,a)]　小池 和輝[1]　植田 諒大[1]　滝本 裕則[2,b)]

概要：複数物体追跡（Multi-Object Tracking, MOT）は，コンピュータビジョン分野における重要なタスクの 1 つであり，ビデオシーケンスに登場する複数の物体を一意に識別し，その動きを正確に追跡することを目的とする．近年，スポーツ領域においても運動解析や戦術分析への応用を背景として，試合映像中の選手を対象とした MOT の研究が注目されている．しかし，スポーツシーンでは，選手が加減速や急な方向転換を伴う不規則な運動を示す上に，チーム競技では共通したユニフォームを着用することで，選手間でその外観が類似しやすくなる．そのため，位置情報や外観情報に基づく従来の関連付けでは物体の誤割当てが生じやすく，異なる物体同士を時間的に結び付けてしまうことで追跡精度の低下を招く．
本研究では，スポーツ選手を対象とした MOT の高精度化を目的とし，姿勢情報を考慮した追跡手法を提案する．具体的には，試合中に選手が多様な姿勢をとる点に着目し，深層学習に基づく姿勢推定モデルの中間層から抽出される特徴量の類似度を，物体検出結果と過去の Tracklet の関連付けに導入した．これにより，複雑な運動や類似した外観を持つ物体が登場するシナリオでも，安定した物体の関連付けが期待できる．評価実験では，アイスホッケー選手を追跡対象とする VIP-HTD データセットを用い， MOTA や IDF1 をはじめとする主要評価指標により提案手法の有効性を検証した．

1. はじめに

複数物体追跡 (Multi-Object Tracking, MOT) は，コンピュータビジョン分野における重要なタスクの 1 つであり，ビデオシーケンスに登場する複数の物体を一意に識別し，その動きを正確に追跡することを目的とする．

近年，スポーツ領域においても運動解析や戦術分析への応用を背景として，試合中の選手を対象とした MOT の研究が注目されている．しかし，スポーツシーンは歩行者や車両を追跡対象とする一般的な MOT ベンチマークと比べて追跡難易度が高いことが知られている．スポーツシーンにおいては，選手が高速で不規則な動きを示すことが多い．また，チーム競技では各選手が共通したユニフォームを着用することで，その外観が類似しやすくなる．その結果，物体の位置や外観を手掛かりとして追跡を行う従来手法は，その性能を十分に発揮することができない．

本研究では，スポーツ選手を対象とした MOT の高精度化を目的とし，姿勢情報を考慮した追跡手法を提案する．具体的には，試合中に選手が多様な姿勢をとる点に着目し，深層学習に基づく姿勢推定器の中間層から得られる特徴量に基づいて姿勢の類似性を定式化することで，物体の識別性を強化する．これにより，複雑な運動や類似した外観を持つ物体が登場するシナリオにおいても，安定した Association が期待される．

2. 関連研究

MOT を解く代表的なパラダイムは “Tracking-by-Detection” である．このパラダイムでは，まず各フレームに登場する物体を検出し，その後，既存のトラックと検出結果を何らかの指標に基づいて対応付ける．これにより，異なるフレーム間で同じ物体には同じ識別子（ID）が割り当てられ，時間的な物体追跡が実現する．ここで，既存トラックと検出結果の間で同一物体を対応付ける操作を関連付け（Association）と呼ぶ． Association は，トラックの集合を $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}_i\}_{i=1}^{M}$，検出結果の集合を $\mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^{N}$ とすると，両集合の要素を重複なく 1 対 1 で対応付け，コスト関数 $\mathcal{L}_{\text{match}}(\cdot)$ の総和を最小化するインデックス割当 $\hat{P}$ を求める問題に帰着する．これは，式 1 のように線形割当問題として定式化され，ハンガリアン法 [1] などのアルゴリズムによって解かれる．

[1] 岡山県立大学大学院　情報系工学研究科　システム工学専攻
Okayama Prefectural University　Graduate School of Systems Engineering　Course of Advanced Systems Engineering
[2] 岡山県立大学　情報工学部　情報通信工学科
Okayama Prefectural University　Faculty of Computer Science and System Engineering　Department of Communication Engineering
[a)] sk625018@c.oka-pu.ac.jp
[b)] takimoto@c.oka-pu.ac.jp

$$\hat{P} = \underset{P \in \mathcal{P}_{M \times N}}{\operatorname{argmin}} \sum_{i=1}^{M} \sum_{j=1}^{N} \mathcal{L}_{\text{match}}(\hat{\mathbf{x}}_i, \mathbf{y}_j) \qquad (1)$$

SORT[2] は Tracking-by-Detection に基づく代表的な追跡手法である．SORT では，物体の位置とその速度によって状態変数を記述し，Kalman Filter[3] を用いて状態の予測・更新を行う．この際，Association のコストには，IoU(Intersection over Union) が用いられる．IoU は 2 つの図形の重なり度合いを評価する指標であり，SORT では前時刻から状態予測されたトラックと現在の検出結果の Bouding-Box 間で IoU を計算し，物体の位置的類似性に基づいた Association を行う．SORT は，非常にシンプルは枠組みで物体追跡を実現する一方で，複雑に動く物体の追跡を苦手とする．これは，Kalman Filter が前提とする状態空間が，状態の遷移に線形性を仮定することに起因する．物体が不規則な動きを示したり，カメラモーションが加わると，物体の見かけの動きが複雑化し，線形運動を仮定した状態予測は，予測トラックと検出結果の位置的類似性を低下させる．その結果，IoU のみに基づいた Association では ID Switch（フレーム間で同一物体の ID が切り替わる現象）などの誤追跡が生じやすくなる．これに対して，物体の非線形運動や社会性を考慮した高精度な予測器 [4], [5], [6] が提案されているが，いずれも追跡に要する計算コストが増大することから，現在でも Kalman Filter による状態予測が MOT では主流となっている．

DeepSORT[7] では，SORT の課題を克服することを目的に，IoU に加えて物体の外観情報を Association に利用している．具体的には，各検出物体の BBox(Bounding-Box) に対応する画像領域から，深層学習に基づく再同定器によって物体の外観特徴量を抽出し，その特徴量の類似性に基づいて既存トラックと検出結果を対応付ける．これにより，複雑な動きを示す物体に対しても頑健な物体追跡を実現している．しかし，物体の外観類似性に基づく Association は必ずしも MOT に有効であるとは限らない．例えば，外観識別性の高い物体が登場するシーン [8] においては DeepSORT が高い追跡精度を示す一方で，選手が共通のユニフォームを着用するスポーツシーン [9] や，野生動物を追跡対象とするシーン [10] では，その追跡精度が SORT よりも劣ることが報告されている．

近年，外観情報に依存せず，追跡精度の向上を図る手法も提案されている．例えば，OC-SORT[11] では，Occlusion 等によって一時的にダミー更新されたトラックが再び検出結果と対応付けられた際に，過去の検出結果に基づいてトラックの状態変数を再更新することで．Kalman Filter の状態推定誤差を緩和させている．また，姿勢情報を Association に利用した手法も提案されている [12], [13]．これらの手法は，物体の身体構造を姿勢推定器が出力する推定関節点（keypoint）によって捉えることで，外観が類似する物体に対しても識別性を強化している．しかし，スポーツシーンにおいては，選手同士の接触や交錯による Occlusion が頻発し，keypoint の欠落・誤推定が生じやすい．その結果，keypoint の類似度が同一物体で不安定となり，Association の精度が低下する恐れがある．さらに，keypoint ベースの追跡は高精度な姿勢推定器に精度が大きく依存し，対象ドメインに特化させることを目的とした姿勢推定器の訓練が追加で必要となる．これには，keypoint アノテーションが付与されたデータを要することから，コスト面での制約も伴う．

本研究では，姿勢情報の有用性を活かしつつ，keypoint の完全性に強く依存しない表現として，姿勢推定器から得られる中間特徴量の類似性に基づいた Association を提案する．これにより，スポーツ特有の「外観識別性にの低い物体群が複雑に動き，Occlusion が多発する」環境においても安定した物体追跡を目指す．

3. 提案手法

スポーツシーンでは，選手が急な加減速や方向転換を繰り返すため運動の不規則性が高く，さらにユニフォームの共通性により外観の識別性も低い．このため，IoU や外観特徴に依存した従来の Association には限界がある．一方，各選手は試合中に多様な姿勢をとり，隣接フレーム間では姿勢の連続性が保たれやすいことから，姿勢情報は有効な識別手掛かりとなり得る．しかし，keypoint に基づくコスト関数は，Occlusion 等に起因する欠落・誤推定の影響を受けやすく，Association が不安定になる．そこで本研究では，keypoint の完全性に過度に依存しない表現として，姿勢推定器の中間層から得られる特徴量を用い，姿勢の類似性を定式化して Association に組み込む手法を提案する．提案手法のパイプラインを図 1 に示す．

3.1 姿勢特徴量の抽出

ある時刻において，検出結果の集合 $\mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^{N}$ に対して得られる姿勢特徴量の集合を $\mathcal{F} = \{\mathbf{f}_j\}_{j=1}^{N}$ とすると，検出結果 $\mathbf{y}_j$ の姿勢特徴量 $\mathbf{f}_j$ は，次のように定義される．

$$\mathbf{f}_j = \text{GAP}(F_j) \in \mathbb{R}^C \qquad (2)$$

ここで，$F_j \in \mathbb{R}^{C \times H \times W}$ は，$\mathbf{y}_j$ の BBox に対応する画像領域を Top-Down 方式の姿勢推定器に入力した際に，その中間層が出力する特徴マップである．この特徴マップ F_j は，対象物体 $\mathbf{y}_j$ の局所的な間接配置や身体構造を捉えており，外観が類似する物体が多く登場するシナリオでは，比較的頑健な物体表現が得られることが期待される．また，$\mathbf{f}_j \in \mathbb{R}^C$ は，F_j に GAP(Global Average Pooling) を適用し，C 次元のベクトルに変換することで得られる．

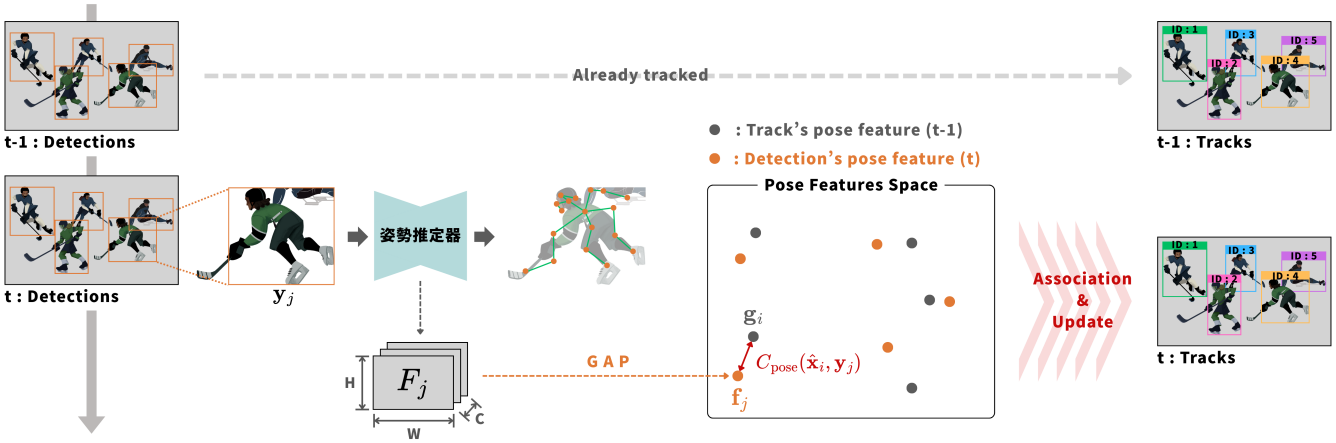

図 1: 提案する姿勢情報に基づいた Association の概要．検出結果 $\{\mathbf{y}_j\}_{j=1}^{M}$ に対して姿勢推定器を用いた特徴抽出を独立に行い，姿勢特徴量 $\{\mathbf{f}_j\}_{j=1}^{M}$ を得る．その後，コスト関数 $C_{\text{pose}}(\cdot)$ によって算出した姿勢類似性を利用して，トラックと検出結果を対応付ける．

3.2 姿勢特徴を考慮した Association

ある時刻において，トラックの集合 $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}_i\}_{i=1}^{M}$ が保持する姿勢情報の集合を $\mathcal{G} = \{\mathbf{g}_i, c_i\}_{i=1}^{M}$ とする．ここで，$\mathbf{g}_i \in \mathbb{R}^C$ はトラック $\hat{\mathbf{x}}_i$ が持つ姿勢特徴量であり，$c_i \in [0,1]$ はその姿勢信頼度を表す．集合 $\mathcal{G}$ は，Association の結果に応じて時間的に更新される（後述）．

本研究では，姿勢特徴を考慮した Association を行うにあたり，トラックと検出結果の姿勢類似度を式 3 で定まる姿勢コスト関数 $C_{\text{pose}}(\cdot)$ を用いて求める．

$$C_{\text{pose}}(\hat{\mathbf{x}}_i, \mathbf{y}_j) = \frac{\mathbf{g}_i^{\mathrm{T}} \mathbf{f}_j}{\|\mathbf{g}_i\|_2 \, \|\mathbf{f}_j\|_2} \in [0,1] \qquad (3)$$

式 3 は $\mathbf{g}_i$ と $\mathbf{f}_j$ のはコサイン類似度を表し，その値が 1 に近づくほどトラック $\hat{\mathbf{x}}_i$ と検出結果 $\mathbf{y}_j$ の姿勢類似性が高いことを意味する．最終的なコスト関数 $\mathcal{L}_{\text{match}}(\cdot)$ は，OC-SORT の既存コスト関数 $C_{\text{ocsort}}(\cdot)$ に姿勢コスト関数 $C_{\text{pose}}(\cdot)$ を加えて式 4 で与える．

$$\mathcal{L}_{\text{match}}(\hat{\mathbf{x}}_i, \mathbf{y}_j) = C_{\text{ocsort}}(\hat{\mathbf{x}}_i, \mathbf{y}_j) + \lambda_p c_i C_{\text{pose}}(\hat{\mathbf{x}}_i, \mathbf{y}_j) \quad (4)$$

ここで，$C_{\text{ocsort}}(\cdot)$ は IoU と速度方向の類似性に基づくコスト関数であり，λ_p は姿勢項の寄与を制御する重みである．また，c_i を掛けることで，トラック $\hat{\mathbf{x}}_i$ が信頼度の低い姿勢特徴量 $\mathbf{g}_i$ を持つ場合には，$\mathbf{g}_i$ から算出される姿勢項の影響を抑える狙いがある．実装上は，$\mathcal{L}_{\text{match}}(\cdot) \leftarrow -\mathcal{L}_{\text{match}}(\cdot)$ として，ハンガリアン法によって式 1 で表される総コストの最小化問題を解く．

3.3 姿勢情報の更新

ある時刻において，式 4 のコスト関数 $\mathcal{L}_{\text{match}}(\cdot)$ を用いた Association の結果，トラック $\hat{\mathbf{x}}_i$ と検出結果 $\mathbf{y}_j$ が対応付けられたとする．このとき，$\hat{\mathbf{x}}_i$ が保持する姿勢情報 $\{\mathbf{g}_i, c_i\}$ は次のように更新する．

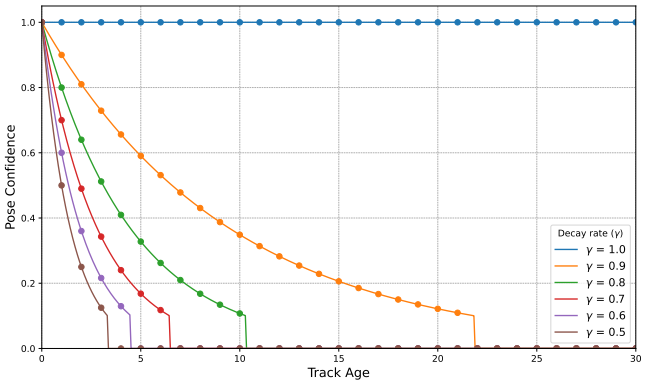

図 2: 姿勢信頼度の減衰挙動．縦軸は姿勢信頼度，横軸はトラックが連続で検出結果と対応付けられなかった時間ステップ数を表す．

$$\mathbf{g}_i \leftarrow \mathbf{f}_j, \quad c_i \leftarrow 1.0 \qquad (5)$$

これは，最新の検出結果が持つ姿勢特徴量を対応付けられたトラックが引継ぎ，信頼度を最大値に戻す操作である．一方，どの検出結果にも対応付けられなかったトラック $\hat{\mathbf{x}}_k$ については，保持している姿勢特徴量 $\mathbf{g}_k$ を維持しつつ，姿勢信頼度 c_k のみを時間減衰させる．

$$\mathbf{g}_k \leftarrow \mathbf{g}_k, \quad c_k \leftarrow \mathbb{1}_{\{\gamma c_k > \tau\}} \gamma c_k \qquad (6)$$

ここで，$\gamma \in (0,1]$ は減衰率，τ は姿勢信頼度の閾値であり，$\mathbb{1}_{\{\gamma c_k > \tau\}}$ は減衰値 γc_k が閾値 τ 以上の場合に 1，それ以外の場合に 0 を出力する指示関数である．式 6 による信頼度の減衰挙動を図 2 に示す．式 5,6 に基づいて集合 $\mathcal{G}$ を更新することで，時間的に物体の姿勢が変化する中で，長期間対応付けられていないトラックの姿勢情報が Association に与える影響を抑制し，追跡精度の低下を防ぐ狙いがある．

表 1: aiueo

Method	MOTA ↑	|FP| ↓	|FN| ↓	|IDs| ↓	|IDF1| ↑	|IDP| ↑	IDR ↑	|FM| ↓
OC-SORT	80.472	2350	2624	29	80.743	81.180	80.311	208
提案手法	**80.495**	**2349**	**2621**	**27**	**82.110**	**82.550**	**81.645**	208

4. 評価実験

4.1 実験条件

本研究では， VIP-HTD(Vision and Image Processing Hokey Tracking Dataset)?を用いて，従来手法である OC-SORT と提案手法の比較評価を行った．この際， OC-SORT と提案手法の両方に ByteTrack[14] で提案された検出信頼度に基づいた二段階の Association を導入した．これは，スポーツシーンにおいてはカメラモーションによる Motion Blur が頻発し，検出信頼度の低い検出結果が得られる傾向にあることを考慮してのことである． VIP-HTD は，試合中のアイスホッケー選手を追跡対象とする MOT データセットであり，計 22 個（訓練:14 個/評価:1 個/テスト:7 個）のシーケンス，計 73,890 枚（訓練:45,396 枚/評価:2,981 枚/テスト:25,513 枚）のフレームで構成されている．

物体検出器には， VIP-HTD の訓練データで事前学習済みの YOLOX[15] を使用した．物体追跡では， YOLOX が出力した検出結果の内，検出信頼値 $c_d \in [0,1]$ が $0.6 \leq c_d \leq 1.0$ のものを第一段階の Association に，$0.1 < c_d < 0.6$ のものを第二段階の Association に利用した．また，姿勢推定器には MS COCO で事前学習済みの HRNet(High-Resolution Network)[16] を用い，物体追跡では HRNet の layer4 が出力する特徴マップから得られた姿勢特徴量 $\mathbf{f} \in \mathbb{R}^{48}$ を Association に利用した．

評価実験では， MOTA(Multi-Object Tracking Accuracy)[17]， IDF1(ID F1-score)[18] を評価指標として用いた． MOTA は $|\mathrm{TP}|, |\mathrm{FP}|, |\mathrm{FN}|, |\mathrm{IDs}|$ から算出される MOT の総合評価指標であり，追跡精度を物体の BBox 単位で評価する．ここで $|\mathrm{TP}|$ は推定 BBox と真値の一致数， $|\mathrm{FP}|$ は真値と一致しなかった推定 BBox の数， $|\mathrm{FP}|$ は推定 BBox に一致しなかった真値の数， $|\mathrm{IDs}|$ は ID Switch の総発生回数を表す．また， IDF1 は IDP と IDR の調和平均として算出される MOT の総合評価指標であり，追跡精度を物体の軌跡単位で評価する．ここで， IDP と IDR は推定軌跡と真値の間で算出される適合率・再現率を表す．

4.2 実験結果および考察

表 1 に， VIP-HTD データセットのテストデータに対する定量評価結果を示す．

図 3 に，定性評価結果を示す．

5. おわりに

本研究では，

まとめと，今後の展望．

参考文献

[1] Kuhn, H. W.: The Hungarian method for the assignment problem, *Naval Research Logistics (NRL)*, Vol. 52 (1955).

[2] Bewley, A., Ge, Z., Ott, L., Ramos, F. and Upcroft, B.: Simple online and realtime tracking, *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468 (online), DOI: 10.1109/ICIP.2016.7533003 (2016).

[3] Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems, *Journal of Basic Engineering*, Vol. 82, No. 1, pp. 35–45 (1960).

[4] Gordon, N. J., Salmond, D. and Smith, A. F. M.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation (1993).

[5] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L. and Savarese, S.: Social LSTM: Human Trajectory Prediction in Crowded Spaces, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–971 (2016).

[6] Saadatnejad, S., Gao, Y., Messaoud, K. and Alahi, A.: Social-Transmotion: Promptable Human Trajectory Prediction (2024).

[7] Wojke, N., Bewley, A. and Paulus, D.: Simple online and realtime tracking with a deep association metric, *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649 (2017).

[8] Milan, A., Leal-Taixe, L., Reid, I., Roth, S. and Schindler, K.: MOT16: A Benchmark for Multi-Object Tracking (2016).

[9] Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K. and Luo, P.: DanceTrack: Multi-Object Tracking in Uniform Appearance and Diverse Motion, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20961–20970 (2022).

[10] Zhang, L., Gao, J., Xiao, Z. and Fan, H.: AnimalTrack: A Benchmark for Multi-Animal Tracking in the Wild (2022).

[11] Cao, J., Pang, J., Weng, X., Khirodkar, R. and Kitani, K.: Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9686–9696 (2023).

[12] Xiu, Y., Li, J., Wang, H., Fang, Y. and Lu, C.: Pose Flow: Efficient Online Pose Tracking, *British Machine Vision Conference* (2018).

[13] Ning, G., Pei, J. and Huang, H.: LightTrack: A Generic Framework for Online Top-Down Human Pose Tracking, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4456–

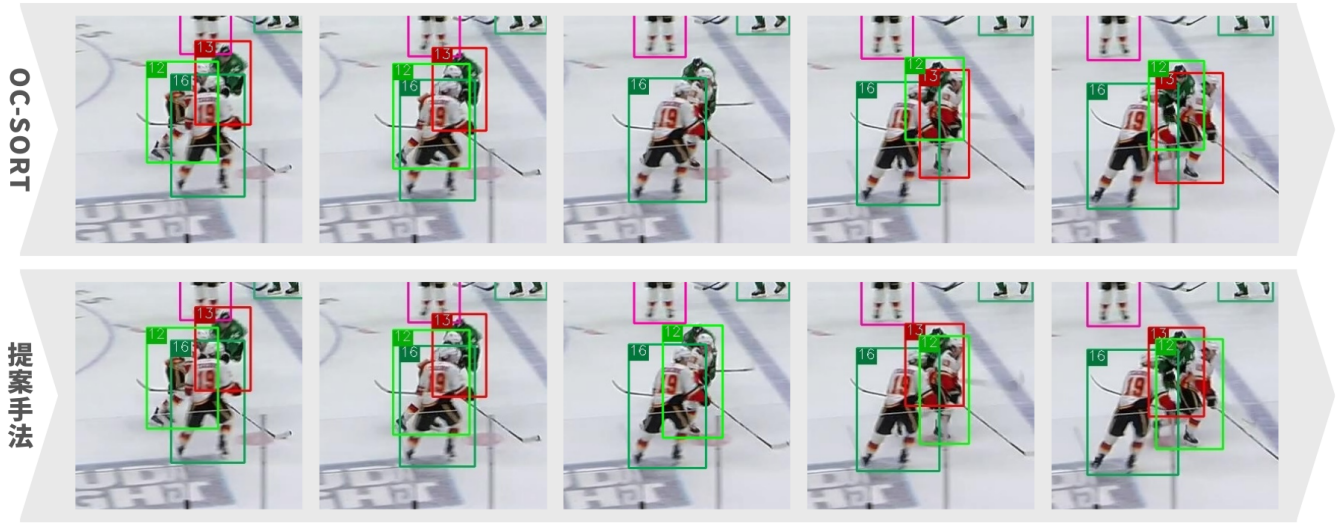

図 3: VIP-HTD の CGY_VS_DAL_003 シーケンスに対する追跡結果の例.

4465 (2020).

[14] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W. and Wang, X.: ByteTrack: Multi-object Tracking by Associating Every Detection Box, *Computer Vision – ECCV 2022* (Avidan, S., Brostow, G., Cissé, M., Farinella, G. M. and Hassner, T., eds.), Cham, Springer Nature Switzerland, pp. 1–21 (2022).

[15] Ge, Z., Liu, S., Wang, F., Li, Z. and Sun, J.: YOLOX: Exceeding YOLO Series in 2021 (2021).

[16] Sun, K., Xiao, B., Liu, D. and Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5686–5696 (2019).

[17] Bernardin, K. and Stiefelhagen, R.: Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics, *EURASIP Journal on Image and Video Processing*, Vol. 2008, pp. 1–10 (2008).

[18] Ristani, E., Solera, F., Zou, R., Cucchiara, R. and Tomasi, C.: Performance Measures and a Data Set for Multi-target, Multi-camera Tracking, *Computer Vision – ECCV 2016 Workshops* (Hua, G. and Jégou, H., eds.), Cham, Springer International Publishing, pp. 17–35 (2016).