広域撮影可能な監視カメラを用いた
人物追跡

西村 仁志

要旨

人々が社会で快適でかつ安心な暮らしを実現するために，これまでにマーケティングや監視のような様々な分野で研究が推進されてきた．そのような分野においては，自動的に人物の行動を認識することが求められており，その際に最も重要な役割を果たすのが人物追跡技術である．本論文では，人物追跡の要素である人物位置推定と人物 ID 推定の誤りに対して，単一人物追跡と複数人物追跡に分けて解決を目指す．単一人物追跡では，パーティクルフィルタの枠組みを用いた確率的な相関フィルタの更新により，人物位置推定のずれを解決する．複数人物追跡では，可能な限り多くの人物を追跡できるように，広域を撮影できる全方位カメラ及びドローン搭載カメラを用いた手法を検討する．局所的パノラマ展開と世界座標表現により，全方位カメラの見えの歪みによって生じる人物 ID 推定の誤りを解決する．また，人物軌跡と複数フレーム行動特徴量（MAF）の交互更新により，ドローン搭載カメラのブレによって生じる人物 ID 推定の誤りを解決する．

第 1 章では，本研究の背景及び目的，本論文の構成について述べる．人物の行動を認識するために重要となる人物追跡は，実応用では人物位置推定と人物 ID 推定に誤りが生じやすく，応用先において致命的な問題となる．本論文ではこれらを単一人物追跡と複数人物追跡に分けて解決することを目指す．単一人物追跡では，見えの変化によって生じる位置推定のずれの解決を目指す．一方，複数人物追跡では，広域を撮影できる全方位カメラ及びドローン搭載カメラを用い，見えの歪みやブレによって生じる人物 ID 推定の誤りの解決を目指す．

第 2 章では，人物追跡に関する関連研究をまとめる．従来，追跡手法は，照合による手法，類似度勾配による手法，時系列フィルタによる手法，検出による手法の順に変遷してきた．まず，単一人物追跡において，検出による手法は遮蔽や変形のような見えの変化に弱いという問題を指摘する．次に，複数人物追跡において，広域を撮影できる全方位カメラ及びドローン搭載カメラを用いる際の問題について各々指摘する．

第 3 章では，見えの変化に頑健な単一人物追跡手法を提案する．従来の検出による追跡手法は，追跡対象とそれ以外の物体を識別する能力は高い．しかし，遮蔽や変形のような見えの変化に対して不安定なため，人物位置推定にずれが生じやすい．そこで，検出による追跡の信頼度が低下した際に，見えの変化に頑健な時系列フィルタと組み合わせる手法を提案する．提案手法では，時系列フィルタであるパーティクルフィルタの観測モデルとして，検出器である相関フィルタによって得られた応答マップを用いる．また，複数の相関フィルタを用意し，パーティクルフィルタの状態変数に追加することで，最適な相関フィルタを選択しながら追跡する．実験では，TB-50 データセットを用いて，提案手法が物体の遮蔽・変形・回転のような見えの変化に対して頑健であることを確認する．

第 4 章では，全方位カメラを用いた見えの歪みに頑健な複数人物追跡手法を提案する．複数人物追跡では，可能な限り多くの人物を追跡するために広域を撮影できることが望ましい．その手段として 360 度の画角をもつ全方位カメラを用いる．しかし，その場合，レンズの歪みによって，人物の見えや位置がフレーム間で非線形に変化してしまう．これに対して通常カメラ向けに設計された従来手法を単純に適用すると，フレーム間の人物対応付けが失敗（ID スイッチが発生）しやすくなる．この問題を解決するため，人物の 3 次元モデルを用いた追跡手法を提案する．提案手法では，1）人物領域のみを局所的に展開してから特徴抽出する，2）距離指標が均一な世界座標系で人物位置を表現する．これらによって，人物の見えや位置の非線形な変化を防ぎ，人物対応付け精度を向上させることができる．実験では，独自に作成した LargeRoom・SmallRoom データセットを用いて，局所的展開と世界座標表現がともに有効であることを確認する．

第 5 章では，ドローン搭載カメラを用いたブレに頑健な複数人物追跡手法を提案する．広域を撮影する別の手段として，移動可能なドローン搭載カメラを用いる．ドローン搭載カメラを用いた場合，ドローンの急な動きによって，人物の見えや位置がフレーム間で急激に変化してしまう．これに対して固定カメラを対象とした従来手法を適用すると，フレーム間の人物対応付けが失敗（ID スイッチが発生）しやすい．この問題に対応するために，人物の行動に関する特徴量を用いた追跡手法を提案する．提案手法では，推定した人物軌跡に基づいて行動特徴量を更新し，再度人物追跡に用いる．実験では，Okutama-Action データセットと Drone-Action データセットを用いて，提案手法により交互更新の反復を繰り返すことで，ID スイッチを防止できることを確認する．

第 6 章では，本論文をまとめ，今後の課題と展望を述べる．上記 3 つの提案手法により，人物追跡における人物位置推定と人物 ID 推定の誤りは解決される．今後は，複数の

カメラからの映像やカメラ以外のセンサからの情報の活用方法を検討する必要がある．また，プライバシー保護も考慮しながら人物追跡技術を利用することで，人々がより快適でかつ安心して暮らせる社会を実現できると考えている．

目次

表目次

図目次

第 1 章

序論

本論文は，広域を撮影可能な監視カメラを用いた人物追跡に関する一連の研究成果についてまとめたものである．本章では，これらの研究の背景，目的，本論文の構成について述べる．

1.1　背景

人々が社会で快適でかつ安心な暮らしを実現するために，これまでに様々な分野で研究が推進されてきた．そのような分野の例として，顧客の需要に応じたサービスを提供するマーケティング分野や，危険な状況の発生を観察する監視分野がある．

マーケティング分野では，“成熟市場においては既存顧客の維持こそが優位性をもたらし，そのためには顧客満足度の向上が重要である” と言われている [5]．近年インターネットショッピングが急速に普及する一方で，実店舗は落ち着いて楽しみながら社会と接することができるという貴重な役割を持つ [6]．実店舗では，来店客は興味がある商品を何度も手に取ったり，会計の待ち時間に不満な態度を取ったりする．そのような行動を認識することができれば，サービスを改善し，顧客満足度の向上につなげることができる．従来は，来店客のそのような行動は，店員が目視で確認し，手作業で分析することが多かったが，このような方法は，人的資源の観点から網羅的に調査したり，また高い認識精度を得ることが難しい．

一方，監視分野では，不審人物の有無を確認し，検知されれば追跡することが重要な任務である [7]．不審人物は，例えば殴ったり蹴ったりしている，あるいは凶器や爆発物を持って徘徊しているといった行動を認識することによって判定できる．近年では，オリン

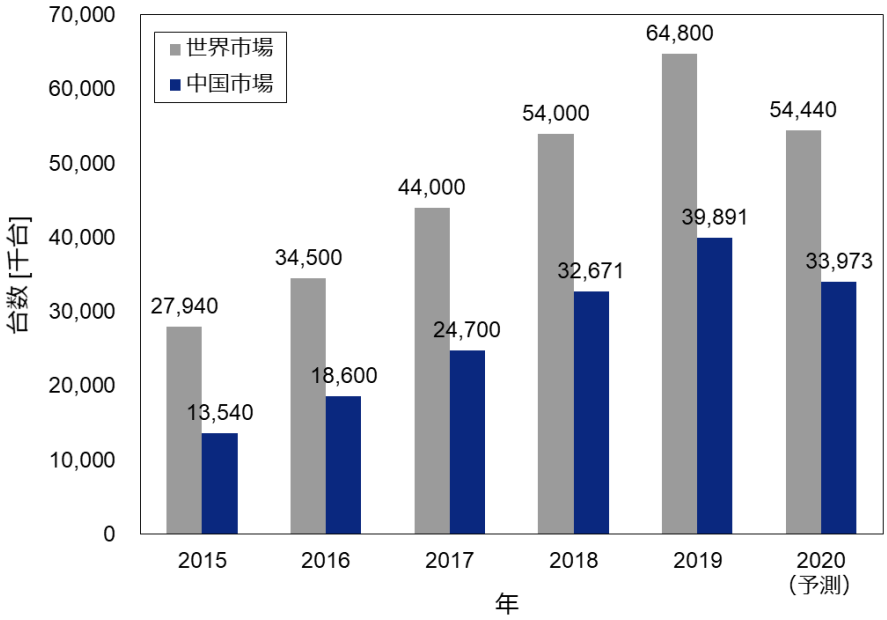


Fig. 1.1: 監視カメラ世界市場規模の推移（文献 [4] に基づいて作成）.

ピックやワールドカップをはじめとした大規模イベントにおいて，犯罪やテロの危険性がますます高まっている．例えばオリンピック開催に要する警備費用は，2000 年の Sydney では 300 億円だったのに対し，2012 年の London では 1,840 億円にまで増加した [8]．従来は，警備員が直接目視で確認し，不審な行動をしている人物の有無を判断することが多かったが，これには人的資源や認識精度の観点からも限界がある．

近年の情報技術の進歩は目覚ましく，センサやコンピュータのようなハードウェアは低価格・高性能化し，固定／無線の通信技術は高速・大容量化し，さらに機械学習アルゴリズムは高精度化した．これによって，マーケティングや監視の分野においても，人間が手作業で行っていた業務の自動化が加速している [7, 9].

人物の行動を捉えるセンサとして，対象人物が端末を保持する必要のある GPS・WiFi や，保持する必要のないカメラ・赤外線センサ・LiDAR センサ等がある．本研究では，対象人物が端末を保持しなくても，行動に関する豊富な情報を抽出できる上，最も市場規模が大きいカメラを用いる．Fig. 1.1 に，監視カメラ世界市場規模の推移を示す．2020 年は新型コロナウイルスの影響により出荷台数の減少が予想されるが，2019 年までは年々出荷台数が増加し，中国を中心に世界市場が拡大していることが分かる．ここで，カメラは実空間における光学情報を時系列画像として取得することができ，各画像はフレームと

(a) 通常のカメラ.

(b) 全方位カメラ.

(c) ドローン搭載カメラ.

Fig. 1.2: 様々なカメラで撮影した画像.

呼ばれる.

可能な限り多くの人物の行動を分析するためには，カメラで広域を撮影できることが望ましい．カメラには撮影できる範囲が定められており，これを角度で表したものを画角と呼ぶ．通常の画角を持つ固定カメラ（Fig. 1.2(a)）では，対象人物が移動するとすぐに撮影範囲から外れて（フレームアウトして）しまう．広域を撮影するためには，カメラの画角を広げる，カメラを移動させる，カメラを複数台設置するといった方法がある．全方位カメラはレンズを工夫して，最大限画角を広げたものである（Fig. 1.2(b)）．その利点は，360 度の画角を持つため，固定して設置するだけで十分広域を撮影できることである．最近の応用例として，日立グループでは，全方位カメラを用いて店舗内の混雑度を推定し，ヒートマップとして可視化するソリューションを提供している [13]．一方，ドローンにカメラを搭載することで，カメラを自由に移動させて広域を撮影することもできる（Fig. 1.2(c)）．その利点は，カメラの設置が難しい場所でも柔軟にシーンを撮影できることである．最近の応用例として，セコム株式会社を中心とするグループは，国内の大規模なイベント会場である花園ラグビー場で，ドローンを用いた不審人物検知の実証実験を成功させた [14]．本研究では，通常の画角を持つ固定カメラだけでなく，全方位カメラやドローン搭載カメラのような様々なカメラを取り扱う．なお，カメラを複数台設置する方法は本研究では取り扱わない．本研究では，そのような方法の基礎となっている 1 台のカメラによる認識に焦点を当てる.

各個人の行動は長時間分析するほど有益で，例えばマーケティング分野ではサービス改

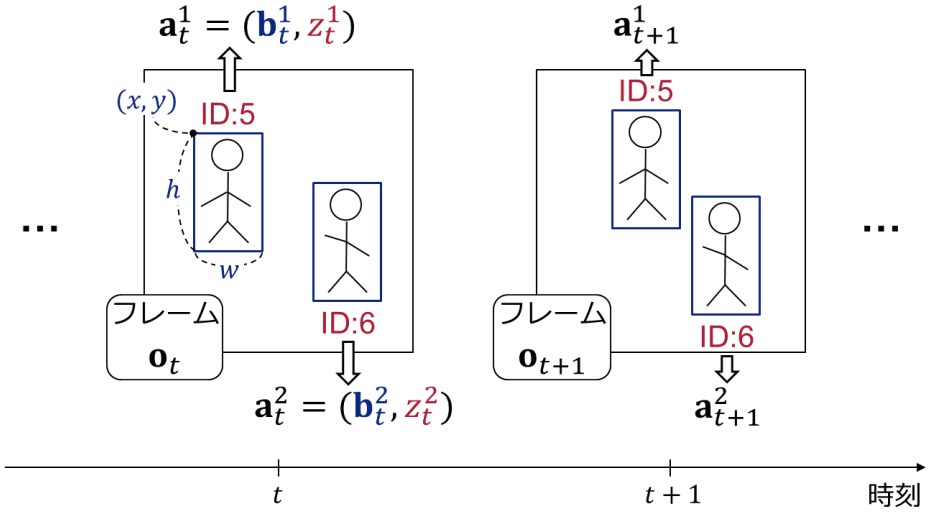


Fig. 1.3: 人物追跡の様子.

善のためのより良い情報が得られ，監視分野では不審人物を判定するためのより良い情報が得られる．そのためには，各人物の位置を他の人物と間違えずに推定し続ける必要がある．これは人物追跡と呼ばれ，人物の行動を認識するための最も重要な役割を果たす技術である．

1.2 目的

本研究では，広域を撮影可能な監視カメラを用いた人物追跡を目的とする．

まず本論文で扱う人物追跡を定式化する．Fig. 1.4 に人物追跡の様子を示す．$B_t = (\mathbf{b}_t^1, \mathbf{b}_t^2, \cdots)$ を，時刻 t におけるフレーム $\mathbf{o}_t$ 中の人物位置とする．ここで，$\mathbf{b}_t^i$ はフレーム $\mathbf{o}_t$ 中の i 番目の人物位置を示す．人物位置は画像座標系において $\mathbf{b} = (x, y, w, h)$ のような矩形で表し，x と y はそれぞれ矩形の左上の x 座標と y 座標を，w と h はそれぞれ矩形の幅と高さを示す．また，フレーム $\mathbf{o}_t$ 中の i 番目の人物位置 $\mathbf{b}_t^i$ に対して，人物 ID z_t^i を付与したものを $\mathbf{a}_t^i = (\mathbf{b}_t^i, z_t^i)$ とし，これらをフレーム $\mathbf{o}_t$ 中で全て集めたものを $A_t = (\mathbf{a}_t^1, \mathbf{a}_t^2, \cdots)$ とする．人物 ID は，フレーム間で人物を対応付けすることによって求まる．人物の対応付けには人物の見えや位置といった特徴量が用いられる．つまり，人物追跡とは，時系列画像（フレーム）$O = \{\mathbf{o}_t \mid t \geq 1\}$ が与えられたとき，$\Omega = \{A_t \mid t \geq 1\}$ を求める問題と定式化できる．よって，人物位置と人物 ID の両方を精度良く求める必要

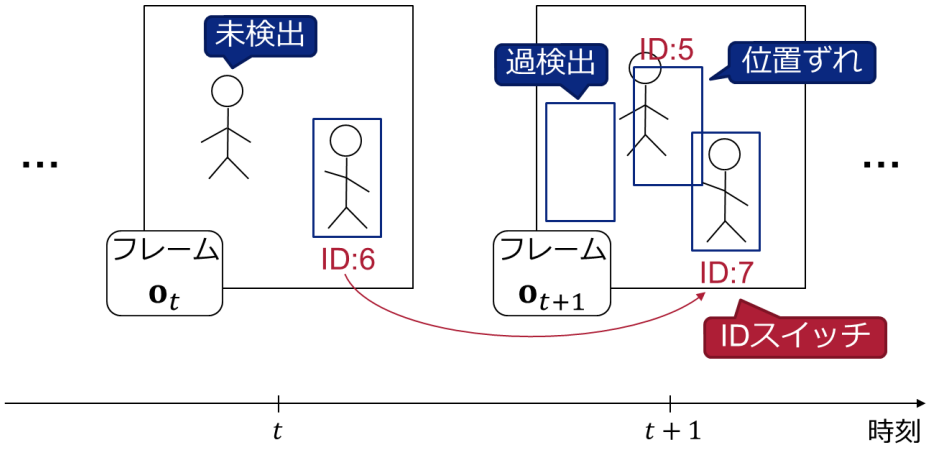


Fig. 1.4: 人物位置推定と人物 ID 推定の誤り．

がある．

しかし，実応用においては，他の物体による遮蔽，対象人物の変形，ブレのような見えの変化が頻発するため，Fig. 1.4 に示すように，人物位置と人物 ID の推定に誤りが生じやすい．本論文では，人物位置推定と人物 ID 推定を研究課題として設定する．

- 人物位置推定：人物位置の推定誤りの種類として，人物がいるにもかかわらず検出されない “未検出”，人物がいない位置に人物が検出される “過検出”，そして人物は検出できたものの位置がずれる “位置ずれ” がある．
- 人物 ID 推定：位置推定は正しくされても，人物 ID の推定を誤ると，異なる人物として追跡されてしまう．これは ID スイッチと呼ばれ，応用先において致命的な問題となる．例えば，マーケティング分野では対象人物の店内行動を別の人物として捉えてしまったり，監視分野では一般の人物を犯人として追跡してしまう恐れがある．

人物追跡の代表的な方法として，時系列フィルタによる手法，検出による手法がある．現在は検出による手法が主流となっているが，通常の画角を持つ固定カメラを想定したのものが多い．そのようなカメラは撮影できる範囲が狭いため，人物が多くなるほど，全ての人物を撮影することは難しい．本論文では，より広域を撮影するため，画角が広い全方位カメラや，カメラ自身が移動可能なドローン搭載カメラを取り扱う．

しかし，全方位カメラを用いた場合は，レンズの歪みによって，人物の見えが位置に

よって大きく異なったり，位置がフレーム間で非線形に変化してしまうという欠点がある．また，ドローン搭載カメラを用いた場合は，ドローンの急な動きによって，人物の見えや位置がフレーム間で急激に変化してしまうといった欠点がある．これらの欠点により，フレーム間の人物対応付け（人物 ID の推定）に誤りが発生する．本論文では，人物に関する豊富な情報に注目してフレーム間の対応付けを行うことで，この問題の解決を目指す．具体的には，全方位カメラのレンズの歪みの問題に対しては人物の形や大きさに関する事前情報を用いることで，また，ドローン搭載カメラの急な動きの問題に対しては人物の行動に関する情報を用いることで，フレーム間の対応付け精度の向上を目指す．

一方，マーケティングにおける会計時や監視における不審者追跡時のように，他の人物や物体が混在する中から，注目した 1 人のみを追跡したい場合もある．その際は人物 ID を推定する必要がなく，人物位置のみが推定対象となる．つまり，前述の定式化は特殊化され，時系列画像 $O = \{\mathbf{o}_t \mid t \geq 1\}$ と $\mathbf{b}_1^1$ が与えられたとき，$\Omega = \{\mathbf{b}_t^1 \mid t \geq 1\}$ を求める問題となる．単一人物追跡は人物位置推定に特化しているため，複数人物追跡とは別で研究が進んでいる．代表的な方法として，照合による手法，勾配による手法，時系列フィルタによる手法，そして検出による手法がある．検出による手法は，追跡対象とそれ以外を識別する能力が機械学習技術により高くなったことにより，現在主流となっているが，時系列情報を考慮していないものがほとんどである．よって，遮蔽や変形のような見えの変化に対して不安定となり，人物位置推定における位置ずれを引き起こす．本論文では，識別能力が高い検出器と，見えの変化に頑健な時系列フィルタを組み合わせた手法を提案することで，この問題の解決を目指す．

なお，本章では，複数人物追跡，単一人物追跡の順に説明をしてきたが，次章以降では，人物位置推定のみを問題とする単一人物追跡，人物位置推定及び人物 ID 推定を問題とする複数人物追跡の順に説明する．ここで，単一人物追跡では通常のカメラ，複数人物追跡では広域撮影可能なカメラを用いる．

1.3　本論文の構成

本論文は Fig. 1.5 に示す通り 6 章で構成されている．第 1 章は序論であり，本研究の背景及び目的，本論文の構成について述べた．第 2 章では関連研究として，人物追跡に関する従来手法について体系的にまとめる．第 3 章では単一人物を対象とし，機械学習による

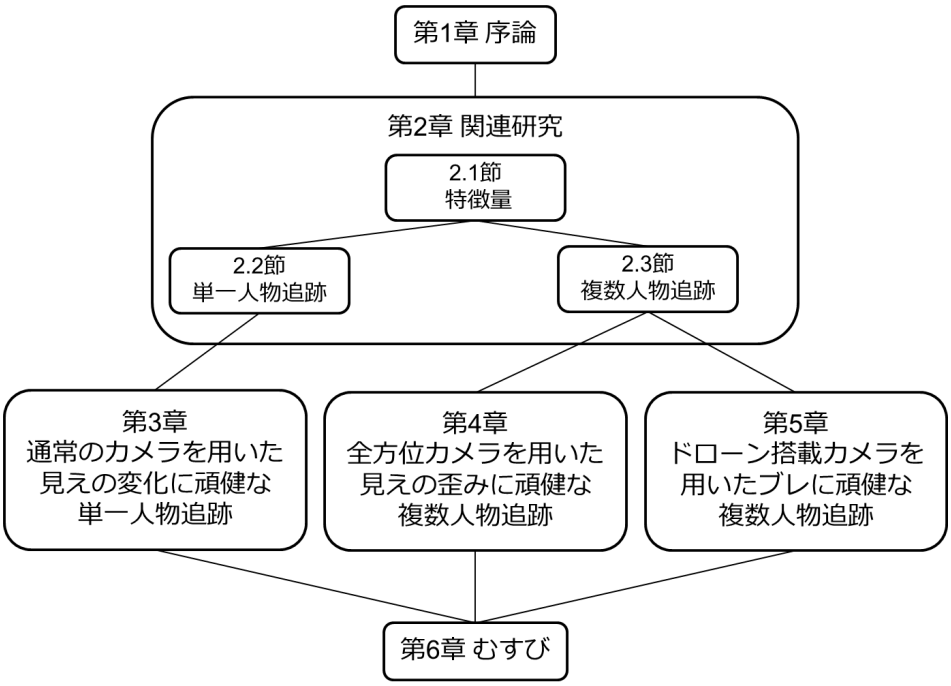


Fig. 1.5: 本論文の構成.

検出器と時系列フィルタを組み合わせた人物追跡手法を提案する．第 4 及び第 5 章では複数人物追跡を対象とする．まず第 4 章では，人物の形や大きさに関する事前情報を用いて，全方位カメラの歪みの影響を低減した人物追跡手法を提案する．次に第 5 章では，人物の行動に関する情報を用いて，ドローン搭載カメラの急な動きの影響を低減した人物追跡手法を提案する．最後に第 6 章で本論文をまとめ，今後の課題と展望を述べる．

第 2 章

関連研究

本章では人物追跡に関する関連研究を体系的にまとめる．人物追跡は人物位置推定と人物 ID 推定で構成され，単一人物追跡では人物位置推定のみが，複数人物追跡では人物位置推定及び人物 ID 推定が解くべき問題となる．本節では，解くべき問題が異なる単一人物追跡と複数人物追跡に分けて関連研究をまとめる．

Fig. 2.1 に，人物追跡に関する従来の研究の変遷を示す．単一人物追跡の研究は，1980 年代以前から現在まで継続して行われてきた．一方，複数人物追跡の研究は，2000 年代に統計的機械学習の技術が発達し，人物検出精度が向上した頃から活発に行われるようになった．そのなかで，人物追跡に用いる特徴量は，大まかには，色，動き，エッジ，局所特徴量，DNN 特徴量の順に変遷してきた．一方，追跡手法は，照合による手法，類似度勾配による手法，時系列フィルタによる手法，そして検出による手法の順に変遷してきた．

本章では，まず 2.1 節で追跡に用いる特徴量についてまとめる．次に 2.2 節で単一人物追跡，2.3 節で複数人物追跡について追跡手法をまとめる．さらに 2.4 節で，第 4 章と関わりが深い全方位カメラを用いた複数人物追跡についてまとめる．最後に 2.5 節で，第 5 章と関わりが深いドローン搭載カメラを用いた複数人物追跡についてまとめる．

2.1　特徴量

本節では，人物追跡に用いる特徴量に関する関連研究をまとめる．

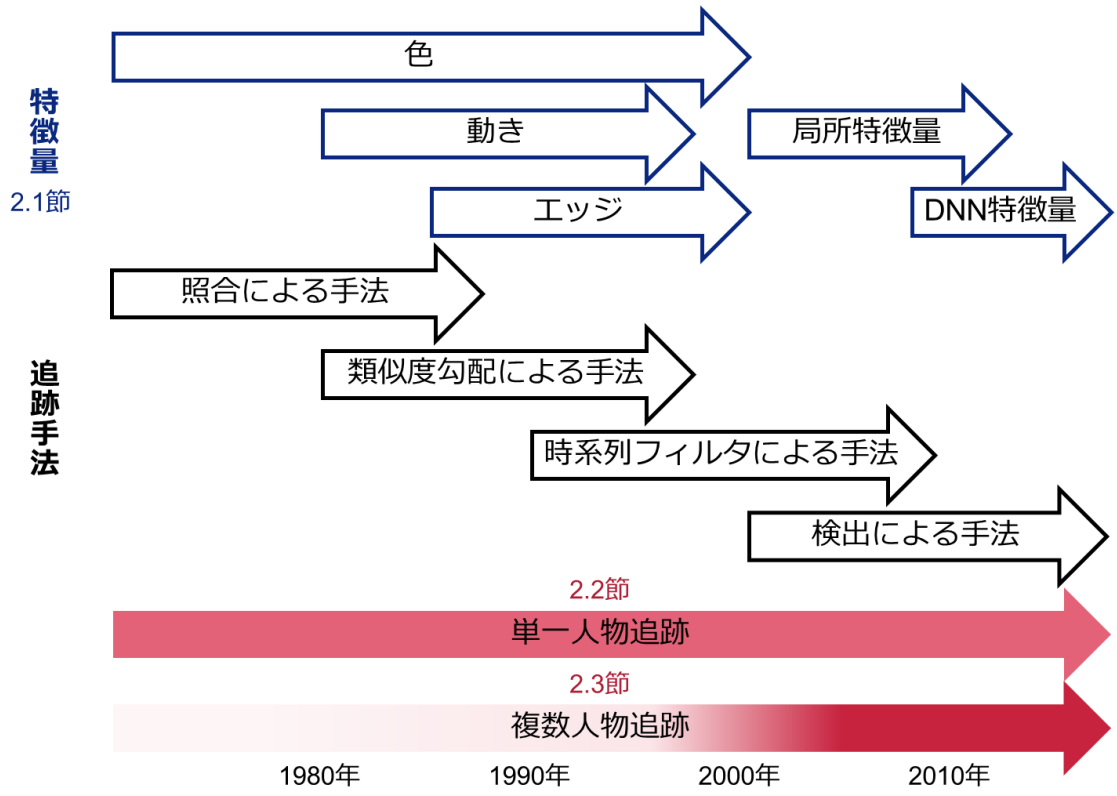


Fig. 2.1: 人物追跡に関する研究の変遷.

2.1.1　色に関する特徴量

カメラで撮影された画像は多数の画素で構成され，一般に各画素は 8 ビット（256 階調）の画素値を持つ．代表的な画像の種類として，濃淡画像と RGB 画像がある．濃淡画像は各画素が明るさを示す 1 つの値を持つのに対し，RGB 画像は赤・緑・青の 3 つの値を持つ．以前は計算量削減のために濃淡画像が用いられることも多かったが，現在では計算機の性能向上により RGB 画像が用いられることがほとんどである．

画素単位の色を用いた追跡手法の代表例としては，2.2.1 項で紹介するテンプレート照合や Circulant Structure of tracking-by-detection with Kernels（CSK）を用いた手法 [15] がある．

しかし，このように画素単位の色を追跡に用いた場合，追跡対象の形状変化に弱いという問題がある．一方，色ヒストグラムは，RGB 各色について領域全体の統計量を取った形状変化に頑健な特徴量である．

色ヒストグラムを用いた追跡手法の代表例としては，アクティブ探索法 [16] や平均値シフト法による手法 [17]，Perera らの手法 [18]，Xing らの手法 [19]，最小費用流（Minimum-Cost Flow ; MCF）による手法 [20] がある．

2.1.2 動きに関する特徴量

動きに関する特徴量の最も基本的なものとして，物体やカメラの動きによって生じる見え方の変化のパターンをフレーム間の画素の変位の 2 次元ベクトルで表現したオプティカルフローがある．代表的な Lucas-Kanade（LK）法 [21] や Horn-Schunck 法 [22] は，隣接フレームにおいて物体の移動は微小であることを仮定した勾配法による手法である．さらに，LK 法では注目画素の近傍で動きは滑らかであるという仮定を置いている．また，Horn-Schunck 法では，オプティカルフローの空間的な滑らかさを拘束条件として，目的関数に加えている．これらの工夫により，これらの手法 [21, 22] は効率的にオプティカルフローを算出することができる．

オプティカルフローを用いた追跡手法の代表例としては，Paragios らの手法 [23]，Support Vector Tracking（SVT） [24]，Median Flow [25] がある．

2.1.3 エッジに関する特徴量

画像中で明るさが急激に変化する部分をエッジと呼ぶ．エッジに関する特徴量は，色に関する特徴量と比較して照明変化に頑健である．エッジを検出するための代表的な手法として，Sobel フィルタ，Laplacian フィルタ，Canny 法がある．Sobel フィルタは 1 次微分と平滑化に基づくフィルタで，画像中のノイズを低減してエッジを検出できる．Laplacian フィルタは 2 次微分に基づくフィルタで，1 次微分に基づく Sobel フィルタよりも鮮明にエッジを検出できる．Canny 法は，Gaussian フィルタで平滑化し，Sobel フィルタで勾配を求め，勾配の極大値を求めた後，2 つのしきい値で処理を行う．これによって，連続で，かつ誤検出が抑制されたエッジを検出することができる．

エッジを用いた追跡手法の代表例としては，Hausdorff 距離を用いた手法 [26]，チャンファマッチングを用いた手法 [27]，CONDENSATION [28] がある．

2.1.4 局所特徴量

局所特徴量は画像の局所領域から抽出される特徴量である．代表的な Scale Invariant Feature Transform（SIFT）[29] や Histograms of Oriented Gradients（HOG）[30] は，局所領域における輝度の勾配方向をヒストグラム化した特徴量で，照明変化や回転のような変化に頑健である．SIFT は特徴点に対して特徴量が記述されるのに対して，HOG はある一定の領域に対して特徴量が記述される．

局所特徴量を用いた追跡手法は，検出による手法であることが多い [31–35].

2.1.5 DNN 特徴量

深層ニューラルネットワーク（Deep Neural Network ; DNN）のうち畳み込みニューラルネットワーク（Convolutional Neural Network ; CNN）は，強力な識別能力を持つ特徴量を抽出できる．AlexNet [36] を起点に，深層学習による画像認識精度は飛躍的に向上した．その後も，Visual Geometry Group による VGG [37] や Residual Network（ResNet） [38] といった深層学習による画像認識モデルが提案されてきた．

ただし，これらのモデルから抽出した特徴量をそのまま用いるのは適切ではない．例えば，2 入力 1 出力の Siamese ネットワーク [39–41] によって，同じ人物は特徴量間の距離が小さく・異なる人物は大きくなるように特徴空間を学習させることで，より良い特徴量を抽出できる．Siamese ネットワークでは，前段のバックボーンネットワークで特徴量を抽出し，後段で特徴量間の距離を算出する．各入力に対応する 2 つのバックボーンモデルは重みを共有し，バックボーンネットワークには AlexNet，VGG，ResNet のようなネットワークを用いることができる．学習の際の誤差関数には，2 つの入力に基づいた contrastive 誤差 [42] の他に，3 つの入力に基づいた triplet 誤差 [43] がよく用いられる．

DNN 特徴量を用いた追跡手法の代表例としては，Deep Learning Tracker（DLT）[44]，Fully Convolutional Network based Tracker（FCNT） [45]，Multi-Domain Network（MDNet） [46]，SiameseFC [47]，SiamRPN [48]，SiamMask [49]，DeepSORT [50]，FairMOT [51]，Joint Detection and Tracking（D&T）[52]，Tracktor [53]，TrackletNet Tracker（TNT） [54]，CenterTrack [55] がある．

2.1.6 行動に関する特徴量

“話している” や “本を読んでいる” というような行動は複数フレームにわたって構成されるため，時間方向の情報が必要となる．行動に関する特徴量を抽出する代表的な手法として，特徴点の軌跡による手法と DNN による手法がある．特徴点の軌跡による手法の代表例として，疎な特徴点を用いる Space-Time Interest Points（STIP） [56] や密な特徴点を用いる Dense Trajectories（DT） [57] がある．その後は DNN による手法が主流となっており，それらは 2 次元畳み込みと 3 次元畳み込みによる手法に大別される．2 次元畳み込みによる手法は，空間方向のみに畳み込みを行い，得られた 2 次元特徴量を組み合わせることで最終的な特徴量を得る [58–60]．一方，3 次元畳み込みによる手法は，空間と時間の両方向に畳み込みを行い，直接特徴量を得る [61, 62]．

また，対象の人物だけでなく，その周辺の人物の行動も含めて特徴表現を行う，行動コンテキスト記述子と呼ばれる手法も提案されている [63]．複数の人物の各行動を組み合わせて，“話している” や “追いかけている” のような集団の行動を表現する手法もある [35, 64]．

行動特徴量を用いた追跡手法の代表例としては，Khamis らの手法 [65]，Choi らの手法 [35]，Li らの手法 [64] がある．

2.2 単一人物追跡

本節では，単一人物追跡の関連研究についてまとめる．単一人物追跡は，初期フレームにおいて指定された 1 つの人物位置を起点として，以降のフレームでその指定された人物の位置を推定し続けるタスクである．

2.2.1 照合による手法

照合による手法は，既に人物位置が分かっているあるフレーム中の人物領域と最も類似する領域の位置を探索する．以下に 2 つの方法を紹介する．

テンプレート照合による手法

テンプレート照合は最も古典的な照合による手法である．フレーム $\mathbf{o}_1$ において与えられた追跡対象人物の位置 $\mathbf{b}_1^1$ を基準とする人物領域をテンプレートとし，次のフレーム中で最も類似する領域の位置を探索する．探索は通常，左上から右下方向に探索するラスタ走査によって行う．

画像間の類似度尺度には様々なものがある．差の 2 乗和（Sum of Squared Difference ; SSD）は最も基本的な類似度尺度であり，

$$R_{\mathrm{SSD}} = \sum_{j=0}^{N-1} \sum_{i=0}^{M-1} (I(i,j) - T(i,j))^2 \tag{2.1}$$

と定義される．ここで，$T(i,j)$ はテンプレートの位置 (i,j) における画素値，$I(i,j)$ は探索中のある照合領域の位置 (i,j) における画素値，M, N はそれぞれテンプレートの横幅と縦幅を示す．他にも，L1 距離を用いることにより，SSD と比べて外れ値の影響を受けにくい差の絶対値和（Sum of Absolute Difference ; SAD）は，

$$R_{\mathrm{SAD}} = \sum_{j=0}^{N-1} \sum_{i=0}^{M-1} |(I(i,j) - T(i,j))| \tag{2.2}$$

と定義される．また，内積を用いることにより照明変化に強い正規化相互相関（Normalized Cross Correlation ; NCC）は，

$$R_{\mathrm{NCC}} = \frac{\sum_{j=0}^{N-1} \sum_{i=0}^{M-1} I(i,j)T(i,j)}{\sqrt{\sum_{j=0}^{N-1} \sum_{i=0}^{M-1} I(i,j)^2} \sqrt{\sum_{j=0}^{N-1} \sum_{i=0}^{M-1} T(i,j)^2}} \tag{2.3}$$

と定義される．

類似度算出に，単純な画素値ではなく，照明変化に頑健なエッジ特徴量を用いた手法もある．集合間の距離を表す Hausdorff 距離を用いて，エッジの差分を重点的に考慮して照合する手法もある [26]．また，チャンファマッチング [66] を用いる手法もある [27]．チャンファマッチングでは，探索対象の画像は距離変換画像に変換して使用する．距離変換画像の画素値はエッジまでの距離になっていることから，類似度の勾配を利用して効率良く探索を行うことができる．

ヒストグラム照合による手法

ヒストグラム照合では，個々の画素値を直接照合するのではなく，照合領域における統計量としてヒストグラムを照合する．ヒストグラムには，形状変化に頑健な色ヒストグラムが用いられることが多い．アクティブ探索法 [16] は色ヒストグラムを用いた高速な照合手法である．基準領域と照合領域との類似度が低ければ，周辺の重複する領域の類似度も低いという考えに基づいて，周辺領域の類似度計算を省略することができることも特長である．

2.2.2 類似度勾配による手法

類似度勾配による手法は，勾配法により類似度が極大となる位置を効率的に探索するため，単純なテンプレート照合よりも高速である．以下では，代表的な Kanade-Lucas-Tomasi（KLT）法による手法と平均値シフト法による手法について紹介する．

Kanade-Lucas-Tomasi（KLT）法による手法

KLT 法 [67] とは，画像から特徴点を抽出し，その特徴点の軌跡を求める手法である．特徴点抽出には，画像 I 内の注目画素 $\mathbf{p}$ を囲む正方形の領域 $S(\mathbf{p})$ を設定し，以下の行列を求める．

$$H = \begin{bmatrix} \sum_{S(\mathbf{p})} I_x^2 & \sum_{S(\mathbf{p})} I_x I_y \\ \sum_{S(\mathbf{p})} I_x I_y & \sum_{S(\mathbf{p})} I_y^2 \end{bmatrix}$$

ここで，I_x, I_y はそれぞれ，画像中のある点における x, y 方向の勾配を示す．この行列の 2 つの固有値がしきい値以上となるような画素 $\mathbf{p}$ を特徴点として抽出する．その後，勾配法に基づく Lucas-Kanade（LK）法 [21] によって特徴点に対するオプティカルフローを求めることで，フレーム間で特徴点の対応付けを行う．

Median Flow [25] では，このようにして対応付けられた特徴点の中央値を用いて矩形を更新することによって追跡を実現している．

平均値シフト法による手法

平均値シフト法とは，カーネル密度推定によって得られた確率密度関数の勾配を求め，極大点を効率良く求める手法である．平均値シフト法を用いて色ヒストグラムの類似度が極大となる位置を効率良く探索することで，高速な追跡が実現される [17]．この際，2

つの正規化色ヒストグラム $\mathbf{p}, \mathbf{q}$ 間の類似度は，以下の Bhattacharyya 係数を用いて算出する．

$$S(\mathbf{p}, \mathbf{q}) = \sum_{u=1}^{U} \sqrt{p_u q_u} \tag{2.4}$$

ここで，$u = (1, 2, \cdots, U)$ はヒストグラムのビンを，p_u, q_u はそれぞれ $\mathbf{p}, \mathbf{q}$ の u 番目のビンの頻度を示す．

平均値シフト法による手法は，探索矩形の大きさが固定で，スケール変化に頑健でない．これに対して，矩形の大きさを可変としたカムシフトによる手法が提案されている [68]．

2.2.3 時系列フィルタによる手法

時系列フィルタによる手法では，人物の位置や大きさなどを状態変数とし，状態変化の Markov 性を仮定した状態変数の時系列推定により人物追跡を行う．具体的には，各フレームにおける状態を，事後確率を最大とする状態として推定する．事後確率は，現在のフレームから観測モデルにより得られる尤度分布と，その直前のフレームから運動モデルにより予測される事前分布に基づいて算出される．この手法では，見えの変化が生じた場合でも直前のフレームからの予測を考慮できるため，頑健な人物追跡が可能である．

時系列フィルタでは，$O_t = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t)$ をフレーム $\mathbf{o}_t$ までの観測，$S_t = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_t)$ をフレーム $\mathbf{o}_t$ までの状態とする．状態 $\mathbf{s}_t$ を $\mathbf{b}_t$ と置けば，$\mathbf{b}_t$ が逐次的に求まることとなり，1.2 節で定式化した問題を解くことができる．

観測モデルと運動モデルをそれぞれ $P(\mathbf{o}_t|\mathbf{s}_t)$，$P(\mathbf{s}_t|\mathbf{s}_{t-1})$ とする．O_t が与えられたとき，状態 $\mathbf{s}_t$ の時間変化に Markov 性を仮定すると，状態 $\mathbf{s}_t$ の事後確率は以下のように Bayes の定理に従って求まる．

$$P(\mathbf{s}_t|O_t) \propto P(\mathbf{o}_t|\mathbf{s}_t) \int P(\mathbf{s}_t|\mathbf{s}_{t-1}) P(\mathbf{s}_{t-1}|O_{t-1}) d\mathbf{s}_{t-1} \tag{2.5}$$

そして，フレーム $\mathbf{o}_t$ における推定状態 $\hat{\mathbf{s}}_t$ は，以下のように事後確率の最大化によって求める．

$$\hat{\mathbf{s}}_t = \arg\max_{\mathbf{s}_t} P(\mathbf{s}_t|O_t) \tag{2.6}$$

以下では代表的な時系列フィルタである Kalman フィルタとパーティクルフィルタによる手法について紹介する．

Kalman フィルタによる手法

Kalman フィルタ [69] による人物追跡手法では，下記の通り，観測モデル $P(\mathbf{s}_{t-1}|O_{t-1})$ と遷移モデル $P(\mathbf{s}_t|\mathbf{s}_{t-1})$ ともに正規分布を仮定し，下記の線形システムで状態推定を行う [70].

$$\mathbf{s}_t = F_t\mathbf{s}_{t-1} + \mathbf{v}_{t-1} \tag{2.7}$$
$$\mathbf{o}_t = H_t\mathbf{s}_{t-1} + \mathbf{w}_t \tag{2.8}$$

ここで，F_t と H_t はそれぞれ遷移行列，観測行列を表す．また，$\mathbf{v}_{t-1}, \mathbf{w}_t$ は，それぞれシステムノイズ，観測ノイズを表し，どちらも正規分布に従って生成される．

パーティクルフィルタによる手法

パーティクルフィルタによる手法では，Monte Carlo 法に基づき，複数のパーティクルによって任意形状の事前／事後状態分布を表現する．具体的には，$\mathbf{s}_t^j$（$j = 1, \ldots, N$）の N 個のパーティクルによって任意形状の状態の分布を表現する．フレーム $\mathbf{o}_t$ における推定状態 $\hat{\mathbf{s}}_t$ は，N 個のパーティクルを用いて以下のように求める．

$$\hat{\mathbf{s}}_t = \underset{j=1,\ldots,N}{\arg\max} P(\mathbf{s}_t^j|O_t) \tag{2.9}$$

CONDENSATION [28] はパーティクルフィルタを用いた人物追跡の古典的な手法であり，人物の位置を状態変数とし，エッジを尤度計算に用いて状態推定を行う．また，Incremental learning for robust Visual Tracking（IVT） [71] は低次元の部分空間における特徴量を利用したパーティクルフィルタによって人物追跡を行っている．さらに Visual Tracker Sampler（VTS） [72] は矩形の位置や大きさだけでなく識別器も複数選択することで状況に応じた追跡を可能にしている．

しかし，これらの手法では，尤度分布を算出するための観測モデルを構築する際に学習の要素が含まれないため，追跡対象とそれ以外との識別能力は低い．この問題に対して，識別能力を高めるために複数の識別器を観測モデルに用いる手法もいくつか提案されている [73, 74].

2.2.4 検出による手法

統計的機械学習の発展により，近年では機械学習により構築した人物検出器を用いた追跡手法が主流となっている．以下では，識別器，相関フィルタ，そして DNN による追跡

手法について説明する．

識別器による手法

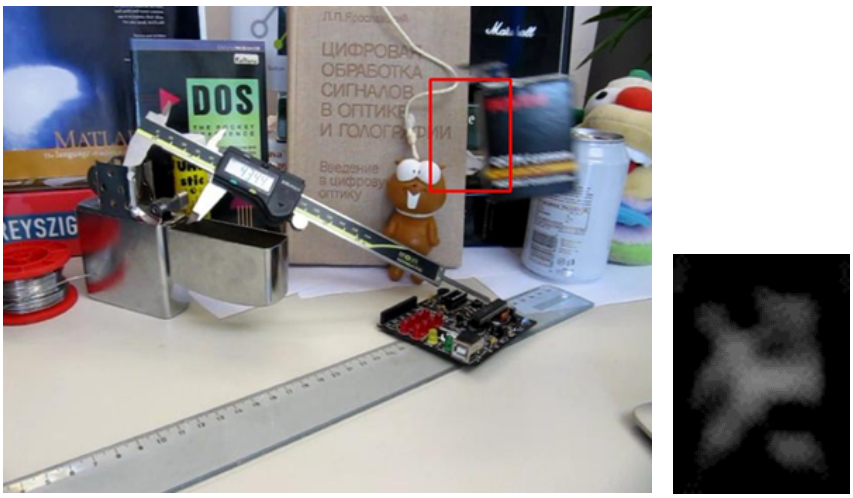
識別器による手法は，候補領域中の様々な位置に対して識別器を適用し，最も出力値が大きくなる位置を検出結果とする．識別器の学習方法として，クラス間のマージンが最大となるように学習する Support Vector Machine（SVM） [75] と呼ばれる手法がある．SVM を追跡に利用した手法である Support Vector Tracking（SVT） [24] により，検出による手法が流行する端緒となった．そこでは，オプティカルフローによる追跡において SVM のスコアを利用し，識別性能を向上させた．Struck [76] は，対象人物の見かけを対象人物か否かの 2 値で SVM を学習するのではなく，人物周辺領域に対する対象人物の占有率も考慮して学習することで，追跡の失敗を防いでいる．また，単純で弱い識別器を逐次的に学習し，識別器の精度を向上させる Boosting と呼ばれる手法がある．Boosting の要素を追跡に取り入れ，随時識別器を更新することで，長時間追跡を可能にする手法も提案されている [77, 78]．

相関フィルタによる手法

相関フィルタは，入力画像に畳み込むことで，それと相関が高い位置を示す応答マップを出力できる．各フレームの応答マップ中で最大値をとる位置を検出することで追跡を実現できる．応答マップ中の値は識別器によって算出され，識別器のパラメータは検出結果によってフレームごとに更新する．相関フィルタによる追跡手法は，追跡精度の高さに加えて，周波数領域で高速に応答マップを出力することができるため，広く用いられている．

Minimum Output Sum of Squared Error（MOSSE） [79] は人物追跡に相関フィルタを用いた先駆的研究であり，これ以降多くの相関フィルタによる追跡手法が提案されてきた．MOSSE では線形空間で識別を行うが，Circulant Structure of tracking-by-detection with Kernels（CSK）を用いた手法 [15] ではより高精度な人物追跡を実現するため，カーネルトリックを用いて非線形空間で効率的に識別を行う．また，CSK が画素値を用いるのに対して，Kernelized Correlation Filter（KCF） [31] は HOG 特徴量を用いる．

ここでは，第 3 章で提案する手法の基になる KCF の詳細を述べる．$\boldsymbol{\alpha}$ を直前のフレームにおいて更新した相関フィルタのパラメータ，$\mathbf{y}$ を直前のフレームにおいて保存した



(a) 入力フレーム.　　(b) 応答マップ.

Fig. 2.2: 追跡が失敗する際の応答マップの例.

HOG 特徴量，そして，$\mathbf{x}$ を現フレームにおいて得られた HOG 特徴量とする．$\boldsymbol{\alpha}$, $\mathbf{x}$, $\mathbf{y}$ は全てベクトルである．これらを用いて，応答マップ R を以下のようにして得る．

$$R = \mathcal{F}^{-1}(\widehat{\mathbf{k}^{\mathbf{yx}}} \circ \widehat{\boldsymbol{\alpha}}) \tag{2.10}$$

ここで，R は行列，$\widehat{\cdot}$ はベクトルの離散 Fourier 変換，$\circ$ は Hadamard 積，$\mathcal{F}^{-1}$ は逆離散 Fourier 変換を示す．$\mathbf{k}^{\mathbf{yx}}$ は $\mathbf{y}$ と $\mathbf{x}$ との間のカーネル相関ベクトルを表し，各要素は $k_i^{\mathbf{yx}} = \kappa(\mathbf{x}, P^{i-1}\mathbf{y})$ で表される．κ はカーネル関数，P はベクトルを巡回シフトさせる置換行列を表す．応答マップ R において値が最大となる位置を人物検出位置 (x, y) とする．矩形の大きさ (w, h) は，HOG 特徴量の次元数を一定にするために，初期フレームで与える固定値とする．

相関フィルタは各フレームで独立に応答マップから追跡対象を検出する．しかし，見えの変化が生じた場合，応答マップを正しく得られずに追跡が失敗することがある．このような場合の入力フレームと応答マップの例を Fig. 2.2 に示す．応答マップには Fig. 2.2(b) のように値の小さな極値が複数存在し，明確な極値を読み取れない．このような複数の極値をもつ応答マップ中で，最大となる位置をフレームごとに選択するだけでは追跡が失敗しやすいため，相関フィルタを用いた追跡手法は見えの変化に頑健でない．

DNN による手法

深層ニューラルネットワーク（Deep Neural Network ; DNN）による追跡手法は，Deep Learning Tracker（DLT） [44] が提案された頃から数多く提案されるようになった．基本的には，候補領域画像を入力とし，人物の位置が出力されるネットワークを用いる．Fully Convolutional Network based Tracker（FCNT） [45] は，VGG [37] によって特徴マップを得た後，“人物” のようなクラス情報を表現するネットワークと，クラス非依存の見えを表現するネットワークに分岐することで，高精度な追跡を可能としている．Multi-Domain Network（MDNet） [46] は，ネットワークを各ドメイン特有の情報を持つ複数のネットワークに分岐させることで，ドメインを認識しながら高精度な追跡を実現している．

一方，最新の研究では，Siamese ネットワークによる手法が主流となっている．そこでは，追跡対象画像と候補領域画像を入力とし，それぞれ特徴マップを得る．その後，候補領域画像からの特徴マップに対して，追跡対象画像からの特徴マップを畳み込むことで相互相関が求まり，得られたマップの中で最大値となる位置を検出位置とする．SiameseFC [47] は，全層畳み込みの Siamese ネットワークを用いている．SiamRPN [48] は，人物を表しているか否かを分類するブランチと，位置補正を行うブランチにネットワークを分岐させることで，より正確な人物検出を実現している．SiamMask [49] は，さらに人物の領域を推定するブランチを追加することで，正確な人物検出を実現している．

2.2.5　単一人物追跡手法の課題

本節で述べた単一人物追跡手法は，対象が複数人になった場合には適用が難しい．これは，単一人物追跡手法は初めに指定した人物を追跡し続けるため，追跡対象が撮影範囲から外れたり，逆に新たな人物が撮影範囲に入ってきた場合，追跡対象人物を判定するのが難しいからである．これに対して，次節では対象が複数人の場合でも追跡できる手法についてまとめる．

2.3 複数人物追跡

本節では，人物位置推定及び人物 ID 推定で構成される複数人物追跡の関連研究についてまとめる．

2.3.1 時系列フィルタによる手法

2.2.3 項で述べた時系列フィルタによる手法は単一人物追跡のための手法であり，そのままでは複数人を対象とする場合に人物の対応付けができない．そこで，Kalman フィルタやパーティクルフィルタを複数人物追跡向けに改良した手法が提案されている．Kalman フィルタによる手法としては，Multiple Hypothesis Tracking（MHT） [80] や Joint Probabilistic Data Association（JPDA）フィルタ [81] が代表的である．その後，パーティクルフィルタによる手法が提案されている [82–85]．

2.3.2 検出による手法

人物検出は深層学習によって著しく進歩し（例：Faster R-CNN [86], You Only Look Once（YOLO） [87], Single Shot multibox Detector（SSD） [88]），精度は大幅に向上した．これに伴い，現在は検出による追跡手法が主流となっている．これらの手法は，検出器で人物を検出し，何らかの特徴量を用いてその結果を対応付けることによって複数人物追跡を実現する．Fig. 2.3 に検出による追跡手法に基づく複数人物追跡の様子を示す．検出による追跡手法の多くは，対応付けに見え特徴量と位置特徴量（矩形）を用いる．以下では，2 部グラフ照合による手法，最小費用流ネットワークによる手法，検出と対応付けを統合した手法，行動特徴量を用いた手法について説明する．なお，これらの手法はオンライン手法とオフライン手法に大別される [89]．オンライン手法では実時間処理するために注目フレームのみしか使用しないのに対し，オフライン手法では事後に処理することにより未来のフレームも使用することができる．

2 部グラフ照合による手法

2 部グラフ照合とは，グラフにおいて 2 つの独立なノード集合間で照合を行う手法である．オンライン手法ではそれまでの軌跡と新たな人物検出結果との照合を行い，オフラ

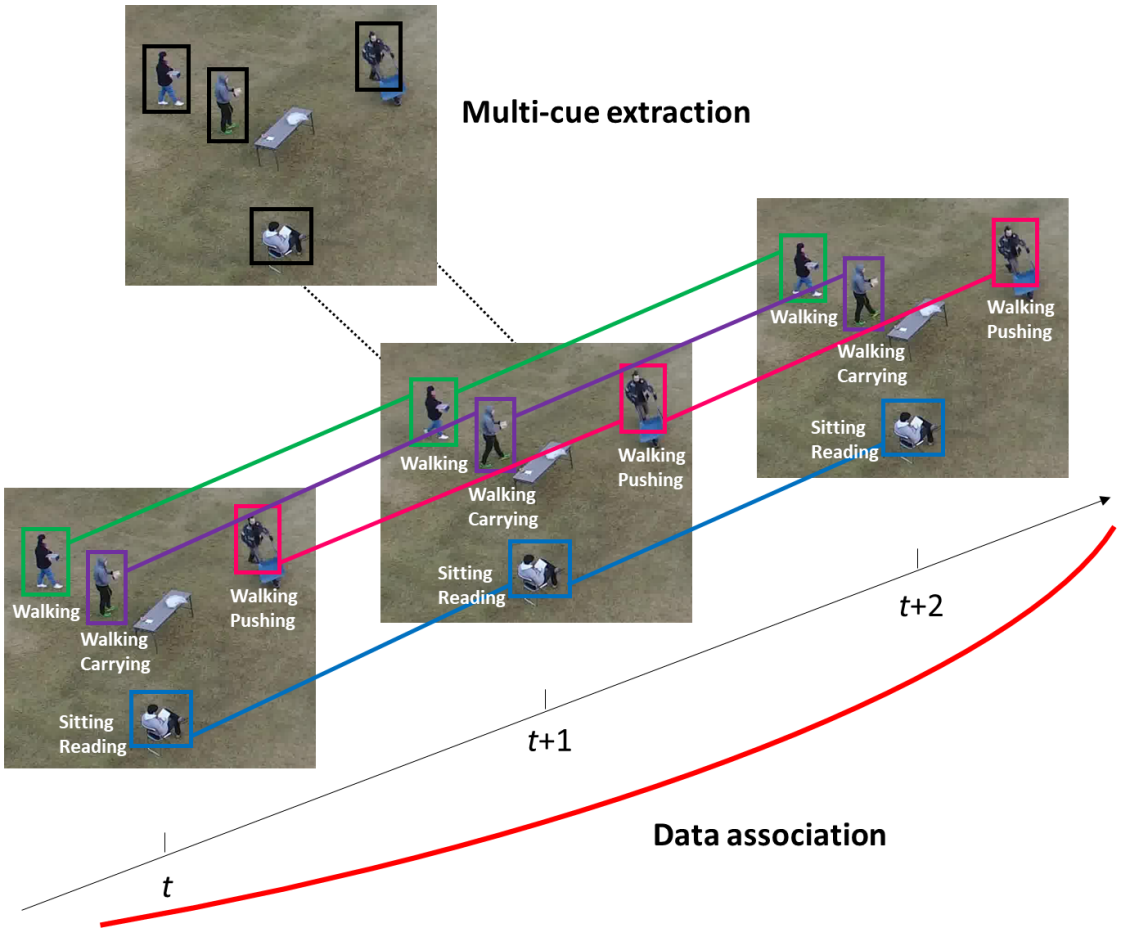


Fig. 2.3: 検出による手法に基づく複数人物追跡.

イン手法では 2 つの軌跡間の照合を行う．照合は，2 部グラフ貪欲割り当てアルゴリズム [32, 90] や Hungary 法 [18, 19, 33, 50, 51, 91] によって組み合わせ最適化問題を解くことで実現する．

ここでは，Hungary 法による手法の 1 つであり，第 4 章で提案する手法の基になる DeepSORT [50] について紹介する．DeepSORT はフレームごとに逐次的に処理を行うオンライン手法のため，1.2 節で示した定式化において，A_{t-1} と $\mathbf{o}_t$ から A_t を推定する問題と考えられる．Fig. 2.4 に DeepSORT による追跡の例を示す．

まず，各フレーム $\mathbf{o}_t$ において，SSD のような人物検出器によって人物を検出し，これを位置特徴量 $\mathbf{b}$ とする．ここでは，フレーム $\mathbf{o}_t$ で 3 人の人物が検出されている．

次に，位置特徴量に対応する見え特徴量 $\mathbf{x}$ を抽出する．特徴抽出器は Siamese ネットワーク [39] を用いて学習する．Siamese ネットワークのバックボーンネットワークには

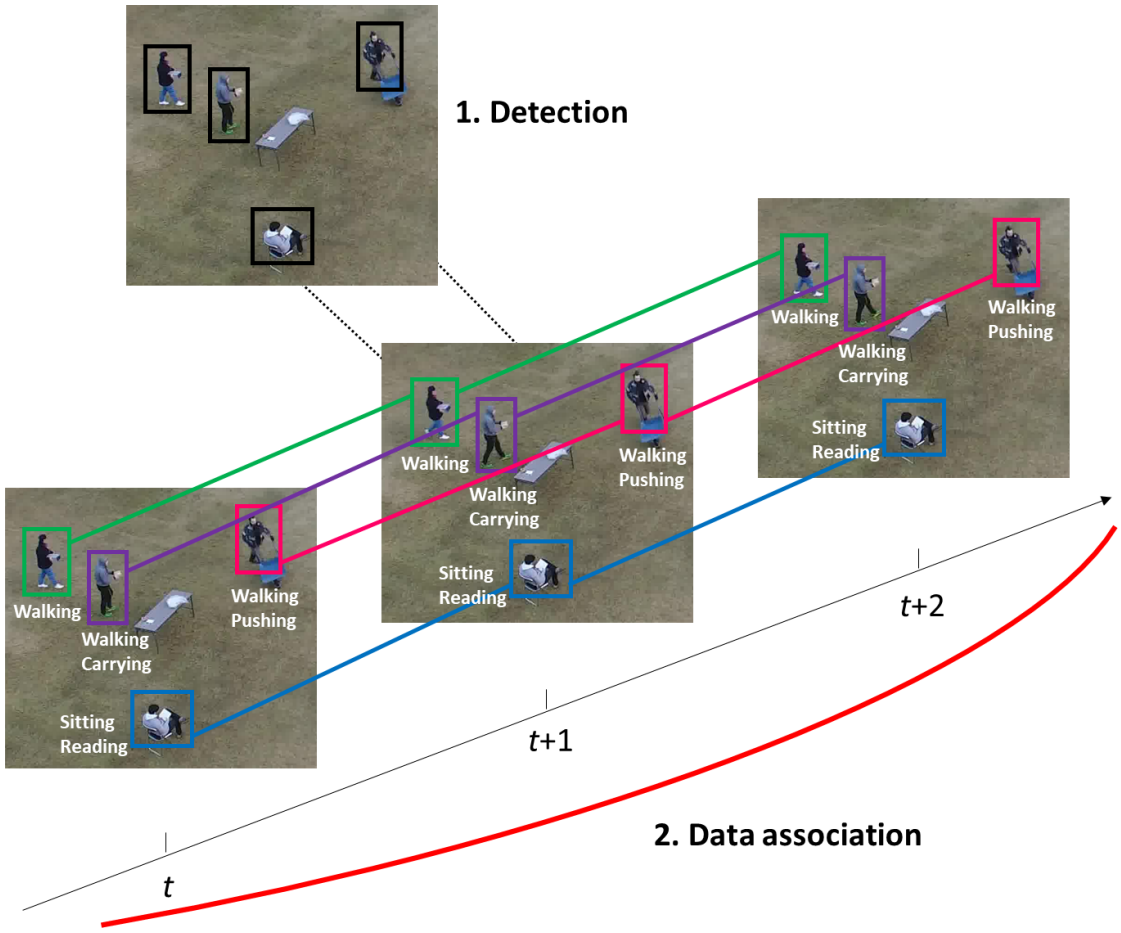


Fig. 2.4: DeepSORT による人物追跡の例.

ResNet [38] を用いる.

そして，フレーム $\mathbf{o}_{t-1}$ までに追跡された矩形とフレーム $\mathbf{o}_t$ で検出された矩形との対応付けを行うために，コスト行列 $C(A_{t-1}, B_t) \in \mathbb{R}^{N_a \times N_b}$ を算出する．$C(A_{t-1}, B_t)$ は，フレーム $\mathbf{o}_{t-1}$ までに追跡された矩形 $\mathbf{a}_{t-1}^i$ とフレーム $\mathbf{o}_t$ で検出された矩形 $\mathbf{b}_t^j$ との間のコスト $c(\mathbf{a}_{t-1}^i, \mathbf{b}_t^j)$ を要素として持つ．Fig. 2.4 の例では，2×3 のコスト行列が算出されている．対応付けは Hungary 法によって効率的に求める．$c(\mathbf{a}_{t-1}^i, \mathbf{b}_t^j)$ は，見え特徴量と位置特徴量についてそれぞれ個別に算出する．つまり，見え特徴量に基づく C^{feat} と位置特徴量に基づく C^{pos} の 2 種類のコスト行列が得られる．対応付けは 2 段階で行う．まず 1 段階目では，C^{feat} に関する割り当て問題を解く．2 段階目では，1 段階目で対応付けられなかった追跡結果に対してのみ，C^{pos} を用いて割り当て問題を解く．もし $c(\mathbf{a}_{t-1}^i, \mathbf{b}_t^j) > \varepsilon$ であれば，$c(\mathbf{a}_{t-1}^i, \mathbf{b}_t^j) = \infty$ とする．ε はあらかじめ定めたパラメータで，

見え特徴量には $\varepsilon_{\mathrm{feat}}$，位置特徴量には $\varepsilon_{\mathrm{pos}}$ と別々に設定する．Fig. 2.4 の例では，a_{t-1}^1 と b_t^1，a_{t-1}^2 と b_t^2 が対応付けられている．b_t^3 はどの人物にも対応付けられず，新たな人物として追跡開始する．

最小費用流ネットワークによる手法

費用流ネットワークは，各辺に対して費用と容量が定義されている有向グラフである．追跡では，各ノードを人物とみなし，各辺の容量は 1，流量は 0 か 1 の 2 値とする．そこで，始点から終点までに流れる全費用が最小となるように，各辺の流量を求める [20, 34, 35, 92]．

ここでは，第 5 章で提案する手法の基になる MCF（Minimum-Cost Flow）による手法 [20] について説明する．便宜上，検出された矩形 $\mathbf{b}_t^j$, $\forall j$, $\forall t$ の番号を並べ替え，通番 i を用いて $\mathbf{y}_i$ と表す．これらの矩形の中で，辺の流量 1 として対応付けられたもの同士を人物軌跡とし，k 番目の人物軌跡を $Y_k = (\mathbf{y}_{k_1}, \mathbf{y}_{k_2}, \cdots, \mathbf{y}_{k_{l_k}})$ として表す．全ての軌跡 $\{Y_k\}$ を求めることは，1.2 節の定式化における $\{A_t\}$ を求めることと等価となる．

まず，対応付けのための特徴抽出について述べる．各 $\mathbf{b}_i$ について，位置特徴量（$\mathbf{x}_i^{\mathrm{loc}}$），見え特徴量（$\mathbf{x}_i^{\mathrm{app}}$），つまり $\mathbf{x}_i = (\mathbf{x}_i^{\mathrm{loc}}, \mathbf{x}_i^{\mathrm{app}})$ を抽出する．位置特徴量は，人物位置 $\mathbf{x}_i^{\mathrm{loc}}$ とその信頼度 x_{sco} を検出器によって推定し用いる．見え特徴量は人物の見えを捉えるもので，Siamese ネットワークのような何らかの特徴抽出器によって抽出する．

Fig. 2.5 に費用流ネットワークの例を示す．人物検出結果 $\mathbf{y}_i$ に対応する 2 つのノード u_i, v_i に関する辺に対して，以下のような費用と流量を持つ．

- 辺 (u_i, v_i)：費用 $c(u_i, v_i) = c_{\mathrm{obsv}}(i)$ と流量 $f(u_i, v_i) = f_{\mathrm{obsv}}(i)$ を持つ．
- 辺 (s, u_i)：費用 $c(s, u_i) = c_{\mathrm{entr}}(i)$ と流量 $f(s, u_i) = f_{\mathrm{entr}}(i)$ を持つ．
- 辺 (v_i, z)：費用 $c(v_i, z) = c_{\mathrm{exit}}(i)$ と流量 $f(v_i, z) = f_{\mathrm{exit}}(i)$ を持つ．

また，$\mathbf{y}_i$ から $\mathbf{y}_j$ への遷移に対して，

- 辺 (v_i, u_j)：費用 $c(v_i, u_j) = c_{\mathrm{tran}}(i, j)$ と流量 $f(v_i, u_j) = f_{\mathrm{tran}}(i, j)$ を持つ．

最小費用流問題は，次式のように全ての辺の流量 F を推定することと等価となる．

$$F = \{(f_{\mathrm{entr}}(i), f_{\mathrm{obsv}}(i), f_{\mathrm{tran}}(i, j), f_{\mathrm{exit}}(i)) \mid \forall i, \forall j, i \neq j, t_i \neq t_j\} \tag{2.11}$$

ここで，$f_{\mathrm{entr}}(i)$, $f_{\mathrm{obsv}}(i)$, $f_{\mathrm{tran}}(i, j)$, $f_{\mathrm{exit}}(i) \in \{0, 1\}$ である．

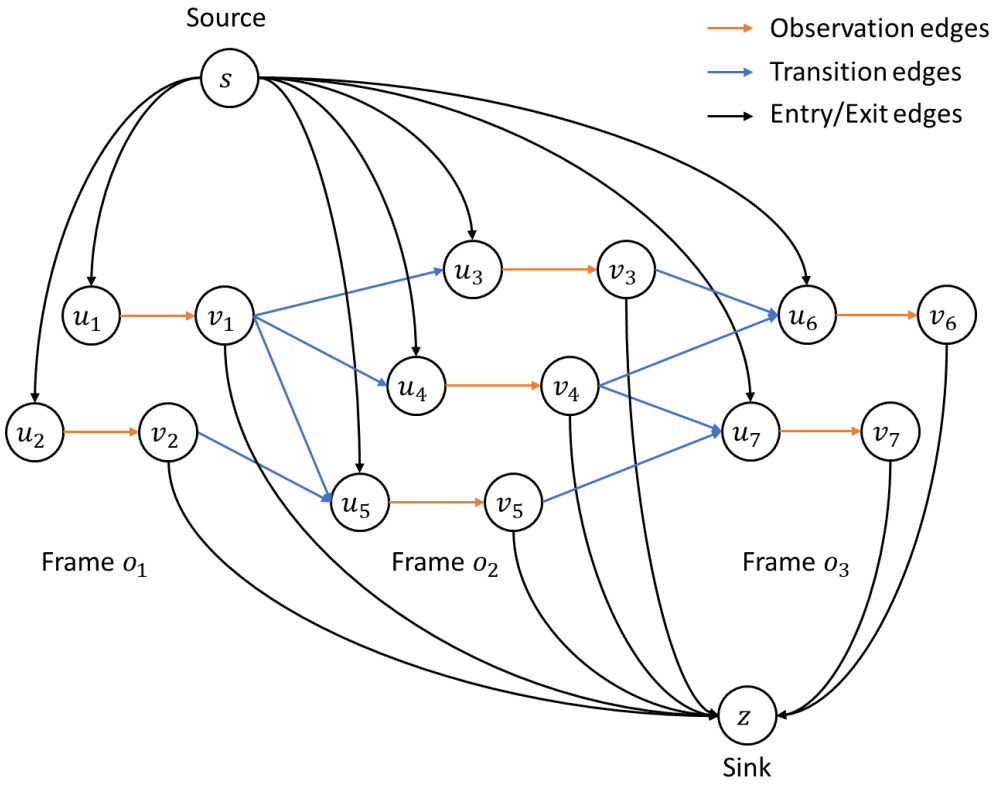


Fig. 2.5: 費用流ネットワークの例．3 フレームに 7 つの人物検出結果が存在する．

i 番目の人物検出結果の観測コストを示す $c_{\text{obsv}}(i)$ は，ロジット関数に基づいている．ロジット関数の変数 p は，次式のように位置特徴量のスコア（x_{sco}）を変数としたロジスティック関数によって算出される．

$$c_{\text{obsv}}(i) = b - \log \frac{p}{1-p} \tag{2.12}$$

$$p = \frac{1}{1 + \exp(\alpha + \beta \cdot x_{\text{sco}}(i))} \tag{2.13}$$

ここで，b は事前に定められたバイアス，α, β はロジスティック関数のパラメータを示す．$c_{\text{obsv}} \in (-\infty, +\infty)$ である．

i 番目の人物検出結果と j 番目の人物検出結果の遷移コストを示す $c_{\text{tran}}(i,j)$ は，ロジスティック関数に基づいている．ロジスティック関数の変数 q は，次式のように非線形関数 g によって算出される．

$$c_{\text{tran}}(i,j) = -\log \frac{1}{1 + \exp(q)} \tag{2.14}$$

$$q = g(c_{\mathrm{iou}}(i,j), c_{\mathrm{app}}(i,j)) \tag{2.15}$$

ここで，$c_{\mathrm{iou}}, c_{\mathrm{app}}$ は，それぞれ位置特徴量間の IoU（Intersection over the Union）スコアと見え特徴量間の余弦距離を示す．g は複数の決定木によって表現され，学習段階ではそのパラメータは勾配ブースティングアルゴリズムによって推定される [93]．$c_{\mathrm{tran}} \in (0, +\infty)$ である．

i 番目の人物検出結果の軌跡の開始コストを示す $c_{\mathrm{entr}}(i)$ は出現コストである．同様に，i 番目の人物検出結果の軌跡の終了コストを示す $c_{\mathrm{exit}}(i)$ は消滅コストである．

ここで，単一の人物は単一の軌跡にしか属さないので，次式のように Y は互いに重複しないという制約を課すことができる．

$$Y_k \cap Y_l = \emptyset, \forall k \neq l \tag{2.16}$$

F は，この非重複制約の下で，次式の目的関数を最小化することによって推定される [20]．

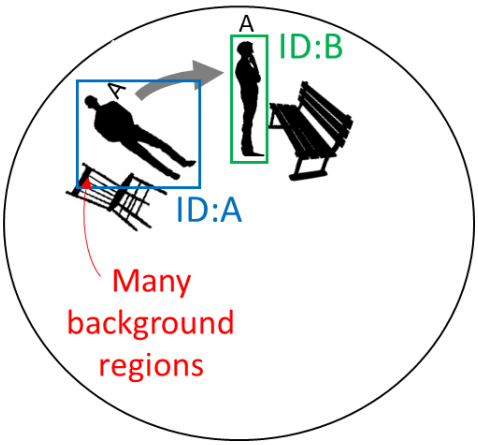
$$\begin{aligned} F^* = \arg\min_F & \sum_i c_{\mathrm{entr}}(i) f_{\mathrm{entr}}(i) + \sum_i c_{\mathrm{obsv}}(i) f_{\mathrm{obsv}}(i) \\ & + \sum_i \sum_j c_{\mathrm{tran}}(i,j) f_{\mathrm{tran}}(i,j) + \sum_i c_{\mathrm{exit}}(i) f_{\mathrm{exit}}(i) \\ \text{s.t. } & f_{\mathrm{entr}}(i) + \sum_j f_{\mathrm{tran}}(j,i) = f_{\mathrm{obsv}}(i) = f_{\mathrm{exit}}(i) + \sum_j f_{\mathrm{tran}}(i,j), \forall i \end{aligned} \tag{2.17}$$

目的関数は Scaling push-relabel 法 [94] によって最小化される．

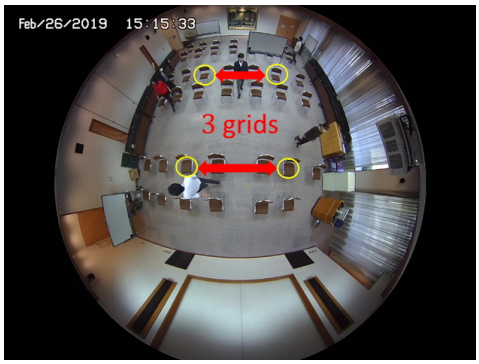
上述の MCF による手法のような手法では，ブレが生じた際に人物の見え特徴量や位置が急激に変化するため，各種コストを適切に算出できず，対応付けに失敗することが多い．

検出と対応付けを統合した手法

本節ではこれまでに検出による手法について述べた．それらの手法では検出とその結果の対応付けという 2 段構成であったが，最新の研究では，検出と追跡を 1 つのニューラルネットワークで実現する手法が提案されている．Joint Detection and Tracking（D&T）[52] では，全層畳み込みで特徴量を抽出し，フレーム間で特徴量の相関を算出することによって，フレーム間の対応付けを行う．また，Tracktor [53] では，既存の検出器を基にして，追加学習なしで次フレームにおける位置を検出できる．CenterTrack [55] では，点単位のヒートマップによって動きを予測することで，フレーム間での人物の移動量が大きい場合でも対応付けを行うことができる．TrackletNet Tracker（TNT） [54] では，グラフモデルとクラスタリングを組み合わせて対応付けを行う．



(a) 矩形内の背景領域が，画像内の位置によって異なる．



(b) 実世界における距離は等しい（3 マス分）が，位置によって画像上での距離が異なる．

Fig. 2.6: 全方位画像は大きな歪みを持つため，2 つの理由によって ID スイッチが生じる．

行動特徴量を用いた手法

本節でこれまでに述べた手法では見え特徴量や位置特徴量がよく用いられるが，例えばブレが生じた場合には，それらが急激に変化し，対応付けに失敗することが多い．それを解決するため，行動特徴量を人物追跡に用いる手法が提案されている．Khamis らは，行動コンテキストを特徴量として，人物追跡と行動認識を効率的に統合するフローモデルを提案している [65]．Choi らは，階層的グラフィカルモデルに基づき，人物追跡と集団行動認識を統合した枠組みを提案している [35]．Li らは，個人・インタラクション・集団単位での行動認識を，人物追跡と統合する手法を提案している [64]．しかし，これらの手法は行動特徴量を単一のフレームのみから抽出しており，他のフレームを用いて更新することはしていない．そのため，ブレが生じた際に行動特徴量が不安定となり，ID スイッチが起こりやすい．

2.4　全方位カメラを用いた複数人物追跡

本節では，全方位カメラを用いた人物追跡に関する関連研究をまとめる．

2.3.2 項で紹介したような通常のカメラを想定した手法の多くは，検出による追跡手法に基づいている．しかし，それらの手法を全方位画像に単純に適用することは難しい．以下にその理由を述べる．全方位画像には大きな歪みが存在するのに対して，多くの検出手

法 [86–88] は人物領域を画像座標に平行な矩形として推定する．その結果，次の 2 つの理由によって ID スイッチが生じる．

1. 人物が斜めの角度から撮影された場合，矩形内に占める背景領域の割合が大きくなってしまい，特徴量に悪影響を及ぼす（Fig. 2.6(a)）．
2. 画像内の位置によって距離指標が不均一であるため，隣接フレーム間で人物の位置が非線形に変化する（Fig. 2.6(b)）．

一方，全方位カメラに特化した手法も数多く提案されている．本節では，全方位画像に対して事前にパノラマ展開を行わない手法と，パノラマ展開を行う手法に分けて説明する．

2.4.1　事前にパノラマ展開を行わない手法

事前にパノラマ展開せずに全方位画像のまま人物追跡を行う手法もいくつか提案されている [95–97]．Chen らは，Markov Random Fields（MRF）によって人物追跡を行う手法を提案している [95]．Zhang らは，前景領域と 3 次元人物モデルの照合によって人物領域を抽出する手法を提案している．ここで，前景領域は背景差分によって推定される．Rameau らは，球体を想定した状態ベクトルによるパーティクルフィルタを用いて人物追跡を行う手法を提案している [96]．

しかし，これらの手法は検出による追跡手法に基づいていないため，高精度な深層学習による人物検出手法との相性が悪い．

2.4.2　事前にパノラマ展開を行う手法

事前に全方位画像をパノラマ展開した後に人物追跡を行う手法は数多く提案されている．パノラマ展開画像では人物の角度は正規化される．Gächter は，時間変化と背景変化の検出手法を提案している [98]．Cielniak らは，人物検出後に Kalman フィルタを適用する手法を提案している [99]．Liu らは，背景モデルに基づいて人物検出を行い，貪欲法によって人物対応付けを行う手法を提案している [100]．Kobilarov らは，Joint Probabilistic Data Association（JPDA）フィルタ [81] をもとに人物対応付けを行う手法を提案している [101]．Song らは，全方位画像の外側のみを展開し，パーティクルフィ

ルタによって人物追跡を行う手法を提案している [102]．Kawasaki らは，静的な背景抽出と動的な背景抽出を組み合わせた人物追跡手法を提案している [103]．これらの手法は，展開後の画像において水平方向に人物が移動した場合には，隣接フレーム間における矩形の位置の非線形な変化を低減することができる．

しかし，パノラマ展開前の画像の中心付近にいる人物は展開後に大きく歪むため，矩形内に占める背景領域の割合が大きい．よって，本節の冒頭に上げた 2 つの問題は解決されていない．

2.5　ドローン搭載カメラを用いた複数人物追跡

本節では，ドローン搭載カメラを用いた人物追跡に関する関連研究をまとめる．ドローン自体の歴史は長いが，空撮用のドローンが身近になったのは 2010 年代と最近である．これは，ジャイロスコープや加速度・画像センサの小型化・高精度化，電池の高性能化によるものが大きい [104]．そのため，ドローン搭載カメラに特化した人物追跡手法はまだ少なく，2.3 節で述べたような通常のカメラを想定した既存手法を用いたものがほとんどである．以下ではそれらの手法について紹介する．

Isard らは，CONDENSATION [28] による追跡を基に，小型で追従可能なドローンを提案している [105]．Comaniciu らは，平均値シフト法 [17] による追跡を行っている [106]．Kim らも同様に平均値シフト法による追跡を基にして，GPS 情報が得られない場合でも飛行できるシステムを提案している [107]．Rodriguez らは，低価格なドローンを用いて，テンプレート照合による追跡を行うシステムを提案している [108]．Bradski らは，カムシフトによる追跡 [68] を基に，楕円形状も追跡できる手法を提案している [109]．

これらは単一人物が対象であるのに対し，複数人物を対象とした手法も提案されている．Zhang らは，TrackletNet Tracker（TNT） [54] によって追跡を行い，そこから 3 次元位置推定を行う手法を提案している [110]．Yang らは，DeepSORT [50] によって行った追跡に基づいて，注視領域を利用した行動認識を行う手法を提案している [111]．

近年は，ドローン映像を対象とした VisDrone と呼ばれる競技会も開催されている [112]．しかし，そこでは，ほとんどの場合は 2.3 節で述べた通常のカメラ向けの手法が適用されており，ドローン映像に特化した手法が少ないことが指摘されている [112]．

第 3 章

通常のカメラを用いた見えの変化に頑健な単一人物追跡

本章では，本論文で対象とする単一人物追跡と複数人物追跡のうち，単一人物追跡に焦点を当て，見えの変化に頑健な追跡手法を提案する．ここでは，通常のカメラ及び広域撮影可能なカメラで共通して行う位置推定を、通常カメラを用いて説明する．Fig. 3.1 に，通常の画角を持つ固定カメラで撮影した画像を示す．従来の検出による追跡手法は，追跡対象とそれ以外の物体を識別する能力は高いものの，遮蔽や変形のような見えの変化に対して不安定となり，位置推定にずれが生じやすい．そこで，本章では，検出による追跡の信頼度が低下した際に，見えの変化に頑健な時系列フィルタと組み合わせた追跡手法を提案する．提案手法では，時系列フィルタの一種であるパーティクルフィルタの観測モデルとして，検出器である相関フィルタによって得られる応答マップを用いる．また，複数の相関フィルタを用意し，パーティクルフィルタの状態変数に追加することで，最適な相関フィルタを選択しながら追跡を行う．評価実験では，50 系列からなる TB-50 データセット [1] を用いて，提案手法が物体の変形・遮蔽・回転のような見えの変化に対して頑健であることを確認する．

3.1　はじめに

単一人物追跡は様々な分野で応用される基礎技術で，例えばマーケティングにおける会計時や，監視における不審者追跡時に用いられる．2.2.4 項で述べた通り，検出による手法は，各フレームで独立に応答マップから追跡対象を検出するため，見えの変化に頑健で

Fig. 3.1: 通常のカメラで撮影した画像の例（文献 [1] より転載）.

ない．一方，2.2.3 項で述べた通り，時系列フィルタによる手法は，尤度分布を算出するための観測モデルに機械学習による識別器を用いていないため，追跡対象とそれ以外の物体との識別能力は低い．

本章では，識別能力が高い複数の相関フィルタと，見えの変化に頑健な時系列フィルタを組み合わせた人物追跡手法を提案する．複数の相関フィルタを用いて複数の応答マップを算出し，それらを確率的に切り替えてパーティクルフィルタの観測モデルとして利用することで，両手法を組み合わせた追跡を実現する．提案手法では，最適な相関フィルタを確率的に選択するため，利用する相関フィルタの情報をパーティクルフィルタの状態変数に追加する．

以後の本章の構成は次の通りである．まず 3.2 節で提案手法について述べる．次に 3.3 節で評価実験について述べる．最後に 3.4 節で本章をまとめる．

3.2　相関フィルタと時系列フィルタによる人物追跡

提案手法では，相関フィルタによる追跡の信頼度が低下した際，パーティクルフィルタに基づく追跡に切り替える．パーティクルフィルタの観測モデルには，相関フィルタによって得られる応答マップを用いることで，両フィルタの同時利用を実現する．しかし，単純に両フィルタを組み合わせるだけでは次の問題が生じる．相関フィルタは，直前のフレームで検出した矩形領域に基づいて識別器のパラメータを更新する．そのため，見えの変化が生じ，推定した矩形領域が追跡対象からずれた場合，誤った矩形領域を用いてパラメータを更新してしまう．その結果，適切でないパラメータを用いた追跡が行われるよう

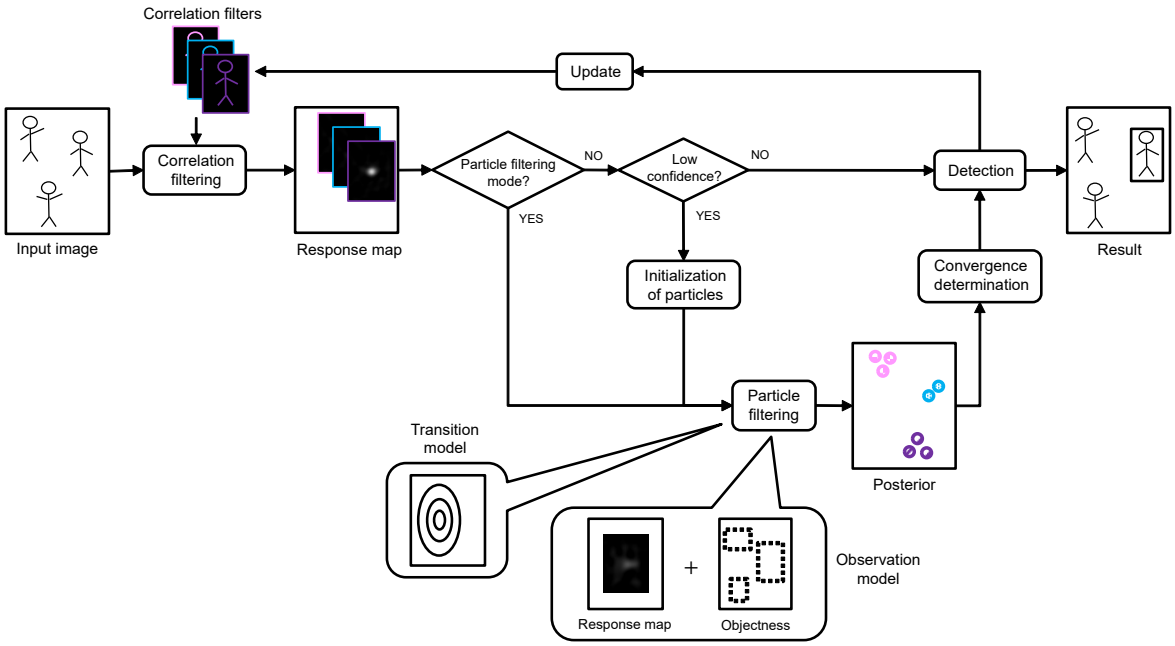


Fig. 3.2: 提案手法全体の処理手順.

になり，追跡対象でない人物を追跡し続ける恐れがある．そこで提案手法では，複数の相関フィルタを用意し，観測モデルに用いる相関フィルタをパーティクルフィルタの状態変数として確率的に推定しながら追跡を行う．

以下で提案手法の詳細について述べる．まず 3.2.1 項では提案手法全体の処理手順について述べ，それ以降の項では提案手法の各構成要素について述べる．

3.2.1　処理手順

Fig. 3.2 に提案手法全体の処理手順を示す．まず，フレームに対して相関フィルタを適用し，応答マップを算出する．次に，応答マップを用いて信頼度判定を行う．信頼度が高ければ，応答マップの値が最大となる位置を人物位置とし，通常の相関フィルタによる追跡を続ける．

一方，信頼度が低ければ，パーティクルフィルタによる手法で追跡を行う．複数の相関フィルタのうち利用するものを状態変数に加え，フレーム $\mathbf{o}_t$ における状態を $\mathbf{s}_t = (\mathbf{b}_t, f_t)$ とする．$\mathbf{b}_t = (x_t, y_t, w_t, h_t)$ であり，フレーム $\mathbf{o}_t$ における人物の矩形を示す．f_t は相関フィルタの番号を示し，$f_t \in \mathbb{N}$ とする．時刻 t における分布を $\mathbf{s}_t^j$（$j = 1, \ldots, N$）の N 個のパーティクルによって表現する．そしてパーティクルの初期化・予測・フィルタリン

グを行い，事後確率を算出する．

パーティクルの初期化では，相関フィルタによって算出した応答マップに対して混合正規分布の当てはめを行い，各正規分布からパーティクルのサンプリングを行う．サンプリングした正規分布に基づき，相関フィルタの番号を決定する．そして，その正規分布から矩形領域の位置と大きさをサンプリングする．フィルタリングの際の観測モデルには，相関フィルタによって得られた応答マップと物体らしさを表す値を利用する．そして事後確率が最大となる位置を人物位置として出力する．

最後に，同一の相関フィルタ番号を持つパーティクルに基づいて，対応する各相関フィルタを更新する．パーティクルが十分に収束した場合にはパーティクルフィルタによる追跡は終了し，通常の相関フィルタによる追跡に戻る．そうでなければ引き続きパーティクルフィルタによる追跡を続ける．

以降の節で，追跡信頼度の判定（3.2.2 項），パーティクルの初期化（3.2.3 項），運動モデル（3.2.4 項），観測モデル（3.2.5 項），パーティクルの収束判定（3.2.6 項），相関フィルタの保存と更新（3.2.7 項）の順に各々の詳細を述べる．

3.2.2　追跡信頼度の判定

応答マップ $R(x, y)$ を用いて，相関フィルタによる追跡に対する信頼度判定を行う．$\max_{x,y} R(x, y) > \varepsilon_1$ の場合は，そのまま通常の相関フィルタによる追跡を行う．ここで，ε_1 はあらかじめ定められたしきい値である．それ以外の場合はパーティクルフィルタによる追跡を開始する．

3.2.3　パーティクルの初期化

各パーティクルの状態は，確率分布に従ってサンプリングすることによって初期化する．パーティクルの初期化アルゴリズムを Algorithm 1 に示す．なおパーティクルの初期化は，パーティクルフィルタによる追跡を開始したフレームでのみ行う．

パーティクルの初期化手順としては，初めに応答マップに基づいて初期状態の分布を混合正規分布で近似した後，そのパーティクルで用いる相関フィルタ番号 f_t をサンプリングし，それに基づいて領域 $\mathbf{b}_t$ をサンプリングする．

Algorithm 1 パーティクルの初期化

· 応答マップに混合正規分布を当てはめる（式 (3.1)）.

for $j = 1$ to N **do**

· Bernoulli 分布に従って，j 番目のパーティクルの u（過去フレームの相関フィルタを使用するかどうかを示す 2 値）をサンプリングする（式 (3.2)）.

· カテゴリカル分布 P_J に従って，j 番目のパーティクルの相関フィルタ番号 f をサンプリングする.

· 混合正規分布中の，相関フィルタ番号 f に対応する正規分布に従って，j 番目のパーティクルの矩形 $\mathbf{b}_t$ をサンプリングする（式 (3.1)）.

if $u = 1$ **then**

· カテゴリカル分布 P_K に従ってサンプリングしたもので，j 番目のパーティクルの相関フィルタ番号 f を置き換える.

end if

end for

混合正規分布による初期状態分布の近似

応答マップ $R(x, y)$ 中の極大値を算出し，それらの位置の集合を得る．極大値をとる位置の集合を応答マップの値の降順に並べ替え，上から M 個を $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M$ とする．$\mathbf{c}_m = (x_m, y_m)$ は応答マップ中の座標を示す.

混合正規分布は以下のように表される.

$$\begin{aligned} & P(\mathbf{b}_t; \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M) \\ &= \sum_{m=1}^{M} p_m \mathcal{N}(\mathbf{b}_t; (\mathbf{c}_m, w_{t-1}, h_{t-1}), S) \end{aligned} \tag{3.1}$$

ここで，$\mathcal{N}$ は正規分布，S は分散共分散行列を示す．w_{t-1}, h_{t-1} は，それぞれフレーム $\mathbf{o}_{t-1}$ における矩形の幅と高さを示す．ここまでの操作によって混合正規分布が決定される.

f_t のサンプリング

相関フィルタ番号 f_t はカテゴリカル分布 P_J に従ってサンプリングする．$P_J = (p_1, p_2, \ldots, p_M)$ と表し，p_m はフレーム $\mathbf{o}_t$ における m 番目の相関フィルタ番号 f_t^m を使用する確率を表す．ここで，$\Sigma_m p_m = 1$ を満たす．なお，m 番目の相関フィルタ番号は，

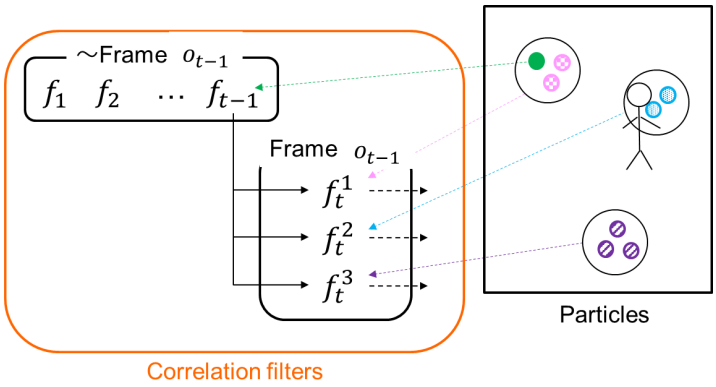


Fig. 3.3: パーティクルへの相関フィルタの割り当て．

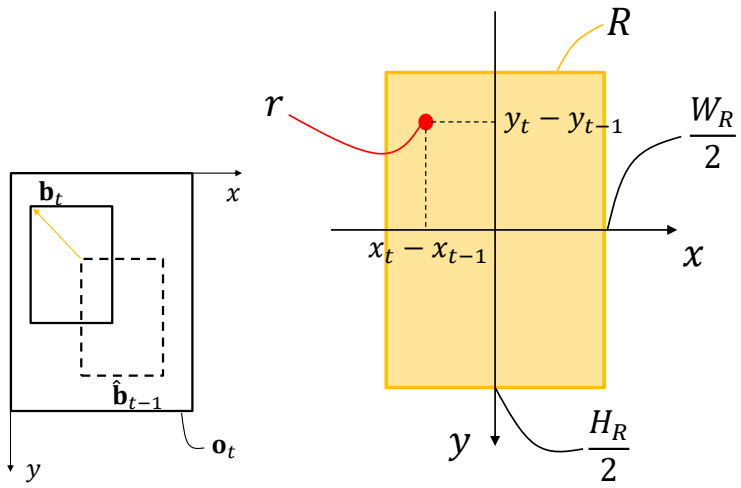
混合分布における m 番目の極大位置と対応している．

しかし，単一人物追跡では，初期フレームのみしか正解位置が与えられないため，時間が経過するほど追跡に失敗する可能性が高くなる．そのため，過去フレームの方が追跡が成功している可能性が高く，その際学習した相関フィルタも利用するのが有効であると考えられる．過去フレームの相関フィルタを使用するか否かを表す 2 値のフラグ $u \in \{0, 1\}$ は，以下の Bernoulli 分布に従う．

$$P(u; \mu) = \mu^u (1 - \mu)^{1-u} \tag{3.2}$$

ここで，母数 μ（$0 \leq \mu \leq 1$）はあらかじめ定められたパラメータである．μ が大きくなればなるほど f が過去フレームの相関フィルタ番号となる確率が高くなる．この Bernoulli 分布に従い，各パーティクルについて u をサンプリングし，$u = 1$ であれば f_t を過去フレームの相関フィルタで置き換える．その場合はカテゴリカル分布 P_K に従って f_t をサンプリングする．$P_K = (p_1, p_2, \ldots, p_{t-1})$ は，フレーム $\mathbf{o}_{t-1}$ までの相関フィルタ番号 $f_1, f_2, \ldots, f_{t-1}$ を使用する確率を示すカテゴリカル分布である．

Fig. 3.3 に，各パーティクルへ相関フィルタが割り当てられる様子を示す．小さな丸記号は，1 つのパーティクルを示す．大きな丸記号の内部にあるパーティクルは全て応答マップ中の同一極大点に対応する正規分布からサンプリングされたことを示す．これらのパーティクルには，同じ相関フィルタ番号 f で表される相関フィルタ番号が割り当てられる．



(a) 状態.　(b) 応答マップ.

Fig. 3.4: 状態 $\mathbf{s}_t$ における応答値 r.

$\mathbf{b}_t$ のサンプリング

前手順においてカテゴリカル分布 P_J に従ってサンプリングされた f_t 番目の正規分布に従って，各パーティクルの矩形 $\mathbf{b}_t$ をサンプリングする．これによって，追跡対象の位置を中心としてパーティクルを配置できる．

3.2.4　運動モデル

各パーティクルについて，運動モデルに従って状態遷移を行う．$\mathbf{b}_{t-1}$ から $\mathbf{b}_t$ への遷移はランダムウォークを仮定し，以下の正規分布に従って行う．

$$P(\mathbf{b}_t|\mathbf{b}_{t-1}) = \mathcal{N}(\mathbf{b}_t; \mathbf{b}_{t-1}, S) \tag{3.3}$$

S は式 (3.1) と同じものを用いる．なお状態遷移の段階では，$\mathbf{b}$ は変化させるが，相関フィルタ番号 f は変化させない．

3.2.5　観測モデル

各パーティクルの尤度 $L(\mathbf{o}_t, \mathbf{s}_t)$ は次式のように算出される．

$$L(\mathbf{o}_t, \mathbf{s}_t) = \alpha \cdot r(\mathbf{o}_t, \mathbf{s}_t) + (1-\alpha) \cdot v(\mathbf{o}_t, \mathbf{s}_t) \tag{3.4}$$

ここで，r は応答マップの値，v は物体らしさの値を示し，α（$0 < \alpha < 1$）は事前に定められた実数値を示す．

観測モデルに相関フィルタから得られた応答マップの値 r を用いることによって，相関フィルタとパーティクルフィルタを統合する．観測 $\mathbf{o}_t$ に f 番目の相関フィルタを適用して得られた応答マップを $R_f^{\mathbf{o}_t}$ とする．応答マップにおける座標系では，応答マップの中心を原点として，前フレームからの追跡対象の移動量を相対位置で示す．よって，状態 $\mathbf{s}_t$ における応答値 r は，応答マップ中の座標 $(x_t - x_{t-1}, y_t - y_{t-1})$ における応答値を参照することによって得られる（Fig. 3.4）．式で表すと次のようにして得られる．

$$r(\mathbf{o}_t, \mathbf{s}_t) = R_f^{\mathbf{o}_t}(x_t - x_{t-1}, y_t - y_{t-1}) \tag{3.5}$$

ここで，$\mathbf{s}_t = \mathbf{b}_t$ であり，$\mathbf{b}_t = (x_t, y_t, w_t, h_t) \in \mathbb{R}^4$ を表す．応答マップ R の横幅と縦幅はそれぞれ W_R，H_R とする．R は 2.2.4 項で述べた高精度かつ高速な Kernelized Correlation Filter（KCF） [31] によって算出する．

また，追跡対象を物体らしい領域に限定するため，物体らしさを表す値 v も尤度計算に用いる．ここで言う “物体” とは，あらゆるカテゴリに属する一般的な物体を示す．物体らしさを用いることにより，検出領域の中に物体全体が含まれるようにする効果が期待できる．状態 $\mathbf{s}_t$ が持つ矩形領域 $\mathbf{b}_t$ が示す領域の物体らしさ v は，$\mathbf{b}_t$ と物体らしい領域との間の重複率をもとに算出する．

まず，物体らしさが高い複数の領域を BInarized Normed Gradients（BING） [113] によって算出する．観測 $\mathbf{o}_t$ のうち，k 番目の物体らしい領域を $\boldsymbol{z}_k^{\mathbf{o}_t}$ とする．次に，物体らしい領域 $\mathbf{p}_k^{\mathbf{o}_t}$ と $\mathbf{b}_t$ との間の重複率 $\mathrm{IoU}(\mathbf{p}_k^{\mathbf{o}_t}, \mathbf{b}_t)$ [79] を算出する．$\mathrm{IoU}(\mathbf{p}_k^{\mathbf{o}_t}, \mathbf{b}_t)$ は，$\mathbf{p}_k^{\mathbf{o}_t}$ と $\mathbf{b}_t$ 同士の和領域の面積に対する共通領域の面積の比で表される．これらをもとに v を以下のように算出する．

$$v(\mathbf{o}_t, \mathbf{s}_t) = \max_k \mathrm{IoU}(\mathbf{p}_k^{\mathbf{o}_t} \mathbf{b}_t) \tag{3.6}$$

3.2.6　パーティクルの収束判定

事後分布の分散は，各パーティクルの尤度を重みとしたときの加重分散によって測る．$\mathbf{s}$ の構成要素に着目すると，$\mathbf{b}$ は座標を表し，f は相関フィルタの番号を表すため，スケールが異なる．そこで，各次元のスケールを正規化して距離計算を行うため，Mahalanobis

距離を用いて加重分散 V を算出する．

$$V = \sum_{j=1}^{N} L_j (\mathbf{s}^j - \bar{\mathbf{s}})^T C^{-1} (\mathbf{s}^j - \bar{\mathbf{s}}) \tag{3.7}$$

ここで，L_j は j 番目のパーティクルの尤度，C は分散共分散行列を示す．パーティクルの加重平均 $\bar{s}$ は以下の通り算出する．

$$\bar{\mathbf{s}} = \sum_{j=1}^{N} L_j \mathbf{s}^j \tag{3.8}$$

ここで， $\mathbf{s}^j$ は j 番目のパーティクルの状態を示す．

$V/(w_1 h_1) < \varepsilon_2$ のとき，パーティクルは十分に収束したとして，パーティクルフィルタによる追跡を終了する．ε_2 はあらかじめ定められたしきい値を示す．それ以後は，全てのパーティクルの中で，最も尤度が高いパーティクルが持つ f 番目の相関フィルタを用いて追跡を行う．

3.2.7 相関フィルタの保存と更新

各フレームにおいて相関フィルタを保存し，$f_1, f_2, \ldots, f_{t-1}$ とする．相関フィルタのみにより追跡を行っている場合はそのままその相関フィルタを保存する．複数の相関フィルタとパーティクルフィルタにより追跡を行っている場合はそのフレームで使用した相関フィルタのうち，最も信頼度が高い相関フィルタを保存する．保存する相関フィルタ数の上限は U フレーム分とし，それ以上になった場合は最も古い相関フィルタを削除する．相関フィルタの信頼度は 3.2.2 項と同様にして算出する．また，各相関フィルタは独立に更新を行い， KCF と同様にして各フレームで更新を行う．

3.3 実験

提案手法の有効性を検証するため，単一人物追跡実験を行った．

3.3.1 データセット

実験には，人物追跡の精度評価に最も広く用いられている TB-50 データセット [1] を用いた．

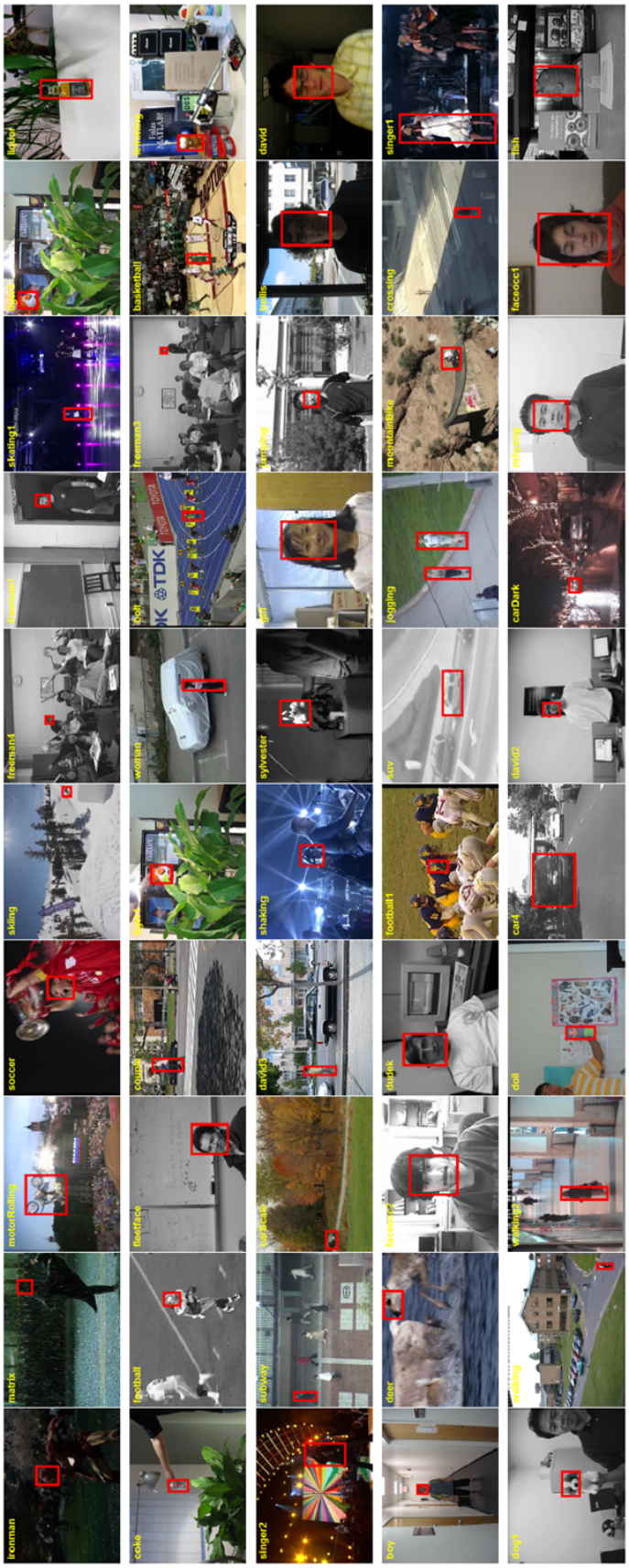


Fig. 3.5: TB-50 データセット [1] に含まれる 50 系列（文献 [1] より転載）．

Table 3.1: TB-50 データセット [1] における見えの変化の種類.

Background Clutters (BC)	乱雑な背景	Motion Blur (MB)	ブレ
Deformation (DEF)	変形	Occlusion (OCC)	遮蔽
Fast Motion (FM)	高速移動	Out-of-Plane Rotation (OPR)	平面外回転
In-Plane Rotation (IPR)	平面内回転	Out-of-View (OV)	画像外移動
Illumination Variation (IV)	照明変化	Scale Variation (SV)	スケール変化
Low Resolution (LR)	低解像度		

Table 3.2: 提案手法の基本パラメータ.

追跡信頼度の判定しきい値	$\varepsilon_1 = 0.2$
パーティクルフィルタによる追跡を行わないフレーム数	$H = 50$
保存する相関フィルタ数の上限	$U = 20$
相関フィルタ番号のカテゴリカル分布	$P_J = 1/M$
過去フレームの相関フィルタ番号のカテゴリカル分布	$P_K = 1/(t-1)$
パーティクル数	$N = 200$
分散共分散行列 S の係数	$a = 10/3$
過去フレームの相関フィルタ番号が選ばれる確率	$\mu = 0.5$

このデータセットは Fig. 3.5 に示す 50 系列で構成されている．人物を対象とした系列は 34 系列であり，追跡対象は全身あるいは顔の場合がある．人物以外にも，車のような人工物やパンダのような動物を対象とした系列も含まれている．実験には人物を対象にしたもの以外も含めた全 50 系列を用いた．撮影場所，被写体との距離，照明状況等の撮影条件は様々である．動画像の解像度は 128×96 画素–768×576 画素，系列長は 71–$1,918$ フレームの範囲である．また，各系列は Table 3.1 に示す 11 種類の見えの変化の種類別に分類される．なお，1 系列に対して複数の見えの変化の種類が付与されている場合もある．

3.3.2 実験条件

あるフレームにおいて，正解と推定結果の間の重複率がしきい値よりも大きい場合は，そのフレームにおいて追跡は成功したと判断する．重複率には，矩形同士の和領域の面

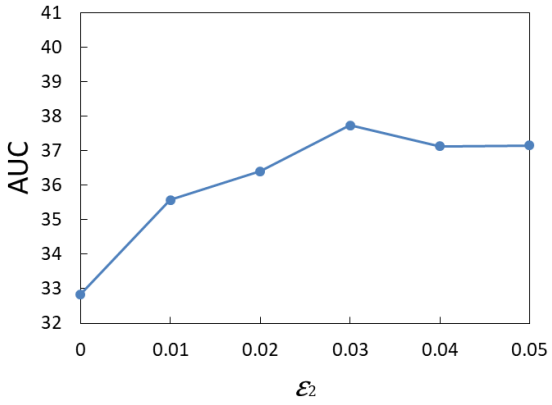


Fig. 3.6: パーティクル収束判定のしきい値 ε_2 を変化させた場合の AUC スコア [%] ($\alpha = 1.0$, $M = 1$).

積に対する共通領域の面積の比を表す Intersection over Union（IoU）を用いる．これを全系列の全フレームに対して実施して平均したものを，平均追跡成功率とする．さらに，IoU のしきい値を 0.0 から 1.0 まで，0.1 刻みで 11 段階に変化させて平均追跡成功率の平均を求めた，Area Under the Curve （AUC）スコア [1] も使用する．

パーティクルの初期化・運動モデルに用いる分散共分散行列 S は，対角成分が $(\frac{w_1}{a}, \frac{h_1}{a}, 0, 0)$ となるような対角行列とした．ここで，a（$a > 0$）はあらかじめ定められた実数である．つまり，式 (3.3) により，w, h は初期フレームから変わらず，常に一定値となるようにした．提案手法の基本パラメータは，Table 3.2 に示した値を実験的に設定した．その他のパラメータ ε_2, α, M については，以下で順次実験的に設定する．なお，提案手法では，パーティクルフィルタによる追跡が終了した後の H フレームは，パーティクルフィルタによる追跡は行わないようにした．これは，パーティクルフィルタによる追跡の後は，応答マップの分散が小さくなるまで 1 つの相関フィルタで対象人物を学習した方が追跡が安定するためである．

3.3.3　人物追跡の評価

本節では人物追跡の評価を行う．まず，提案手法のパラメータを設定した後，他手法との比較を行う．次に，追跡信頼度が低下しているフレーム区間における提案手法の有効性を確認する．さらに，見えの変化の種類ごとに提案手法の有効性を確認する．

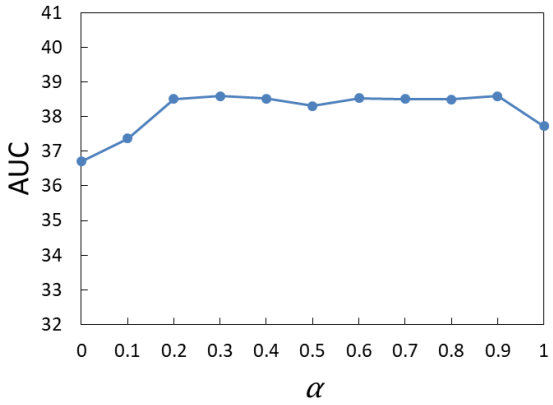


Fig. 3.7: 物体らしさと応答マップ間の重み比率 α を変化させた場合の AUC スコア [%] ($\varepsilon_2 = 0.03$, $M = 1$).

パーティクル収束判定のしきい値

パーティクル収束判定のパラメータ ε_2 を変化させた場合に人物追跡精度を評価した．その他のパラメータは，$\alpha = 1.0$, $M = 1$ とした．全系列，全フレームを対象にした AUC スコアを Fig 3.6 に示す．$\varepsilon_2 = 0.03$ のときに AUC スコアは 37.73% と最も高くなった．また，$\varepsilon_2 = 0$ とした場合，AUC スコアは 32.84% となった．常にパーティクルフィルタを用いた場合は，AUC スコアは 2.31% となった．これより，パーティクルがある程度収束した段階で KCF のみの追跡に戻した方が良いことが分かる．なお，$\varepsilon_2 > 0.05$ とすると AUC スコアはほぼ変化しなかった．

物体らしさと応答マップ間の重み比率

物体らしさと応答マップの重み比率 α を変化させた場合の人物追跡精度の評価を行った．その他のパラメータは，$\varepsilon_2 = 0.03$, $M = 1$ とした．全系列，全フレームを対象にした AUC スコアを Fig. 3.7 に示す．物体らしさを全く用いなかった場合（$\alpha = 1.0$）と比較して，物体らしさを用いた場合（$\alpha = 0.2$–0.9）の方が，AUC スコアが高いことが分かる．特に，$\alpha = 0.3$ の場合に AUC スコアが最大となり，値は 38.60 であった．これより，物体らしさを観測モデルに利用することが有効であることが確認された．ただし，$\alpha = 0.0, 0.1$ の場合は $\alpha = 1.0$ の場合よりも精度が低い．これは，追跡対象ではなく，物体らしさの値が大きい他の物体を追跡してしまうことが増えたためと考えられる．よっ

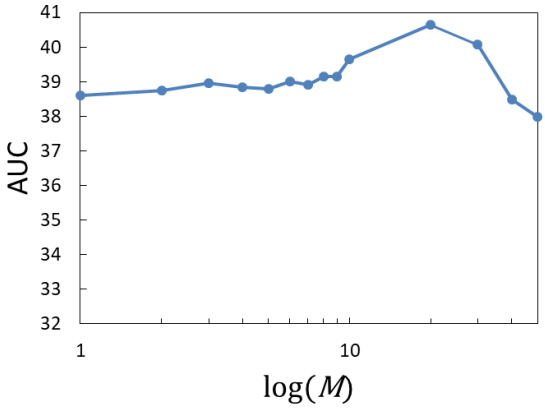


Fig. 3.8: 相関フィルタ数 M を変化させた場合の AUC スコア [%]（$\varepsilon_2 = 0.03, \alpha = 0.3$）.

Table 3.3: 比較に用いた従来手法の一覧.

手法	基にしている手法
CSK [15]	Correlation filter
KCF [31]	Correlation filter
VTS [72]	Particle filter
Struck [76]	SVM
MIL [78]	Others
TLD [114]	Others
SCM [115]	Others

て，物体らしさの比率は大きくしすぎない必要がある．

相関フィルタ数

相関フィルタ数 M を変化させた場合の人物追跡精度の評価を行った．その他のパラメータは，$\varepsilon_2 = 0.03, \alpha = 0.3$ とした．全系列，全フレームを対象にした AUC スコアを Fig. 3.8 に示す．単一の相関フィルタを用いた場合 （$M = 1$）よりも，複数の相関フィルタを用いた場合 （$2 \geq M \geq 30$）の方が AUC スコアが高く，提案手法の有効性が示された．特に，$M = 20$ の場合に AUC スコアが最大となり，値は 40.64 であった．一方，$M = 40, 50$ の場合は，単一の相関フィルタを用いた場合 （$M = 1$）よりも AUC スコアが低くなった．これは，3.2.3 項で述べたパーティクルの初期化において，追跡対象とは

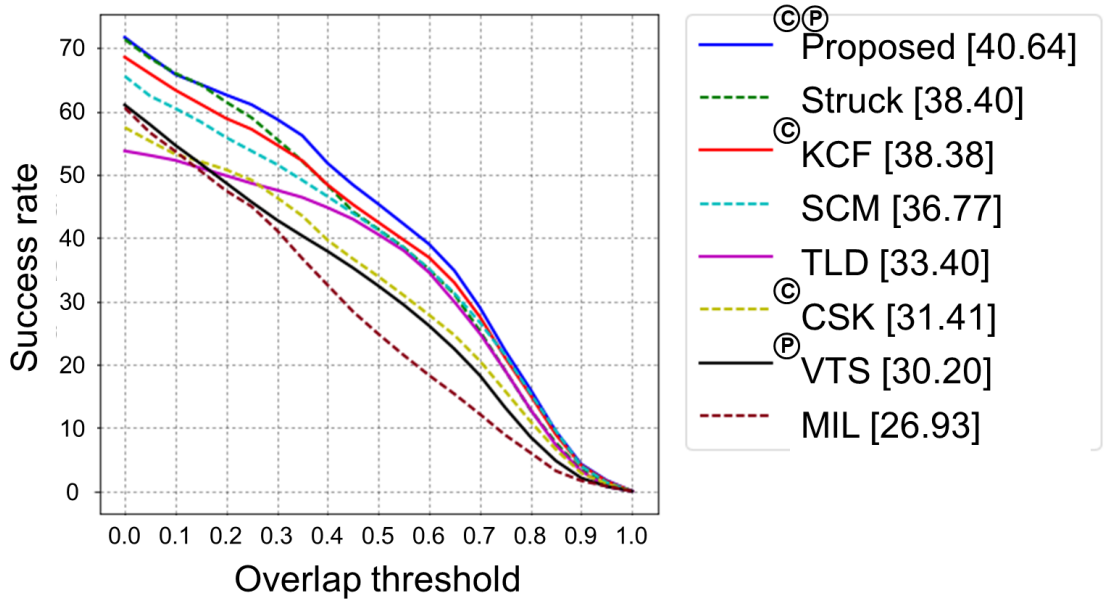


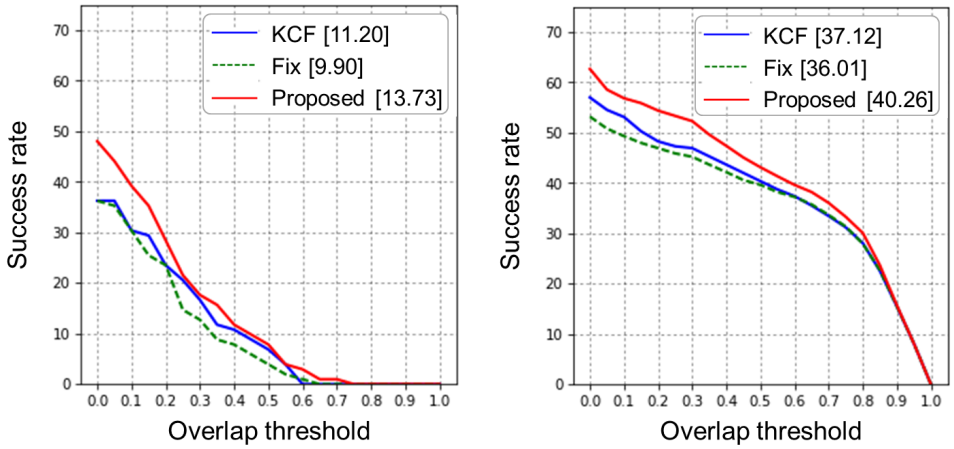
Fig. 3.9: 平均追跡成功率 [%]．丸付き C は相関フィルタによる手法，丸付き P はパーティクルフィルタによる手法を示す．[ ] 内の値は AUC スコア [%] を示す（$\varepsilon_2 = 0.03$, $\alpha = 0.3$, $M = 20$）．

異なる位置に多くのパーティクルが配置され，結果として追跡が失敗することが増えたためだと考えられる．よって，相関フィルタ数は多くしすぎない必要がある．

他手法との比較

提案手法と state-of-the-art の従来手法との精度比較を行った．本節のこれまでの実験結果に基づいて，提案手法のパラメータは，$\varepsilon_2 = 0.03$, $\alpha = 0.3$, $M = 20$ とした．比較に用いた従来手法の一覧を Table 3.3 に示す．これらの手法のうち，KCF [31] は独自らの実装*，それ以外はベンチマーク用実装†を使用して，追跡精度を求めた．正解と推定結果の間の重複率 IoU のしきい値を変化させたときの各手法の平均追跡成功率を Fig. 3.9 に示す．ベースライン手法の KCF では AUC スコアが 38.38% であったのに対し，提案手法では 40.64% に向上した．また，提案手法は他の比較手法よりも AUC スコアが高く，提案手法の有効性が示された．

* https://github.com/joaofaro/KCFcpp/
† https://github.com/jwlim/tracker_benchmark/



(a) 追跡信頼度が低下しているフレーム区間．(b) パーティクルが収束し，KCF による追跡に戻った後の 10 フレーム．

Fig. 3.10: フレーム区間ごとの平均追跡成功率 [%]．凡例は追跡信頼度が下がったフレーム区間における追跡手法を示し，KCF は KCF で追跡した場合，Fix は追跡を停止した場合，Proposed は相関フィルタの確率的選択に基づいて追跡した場合を示す．[ ] 内の値は AUC スコア [%] を示す．

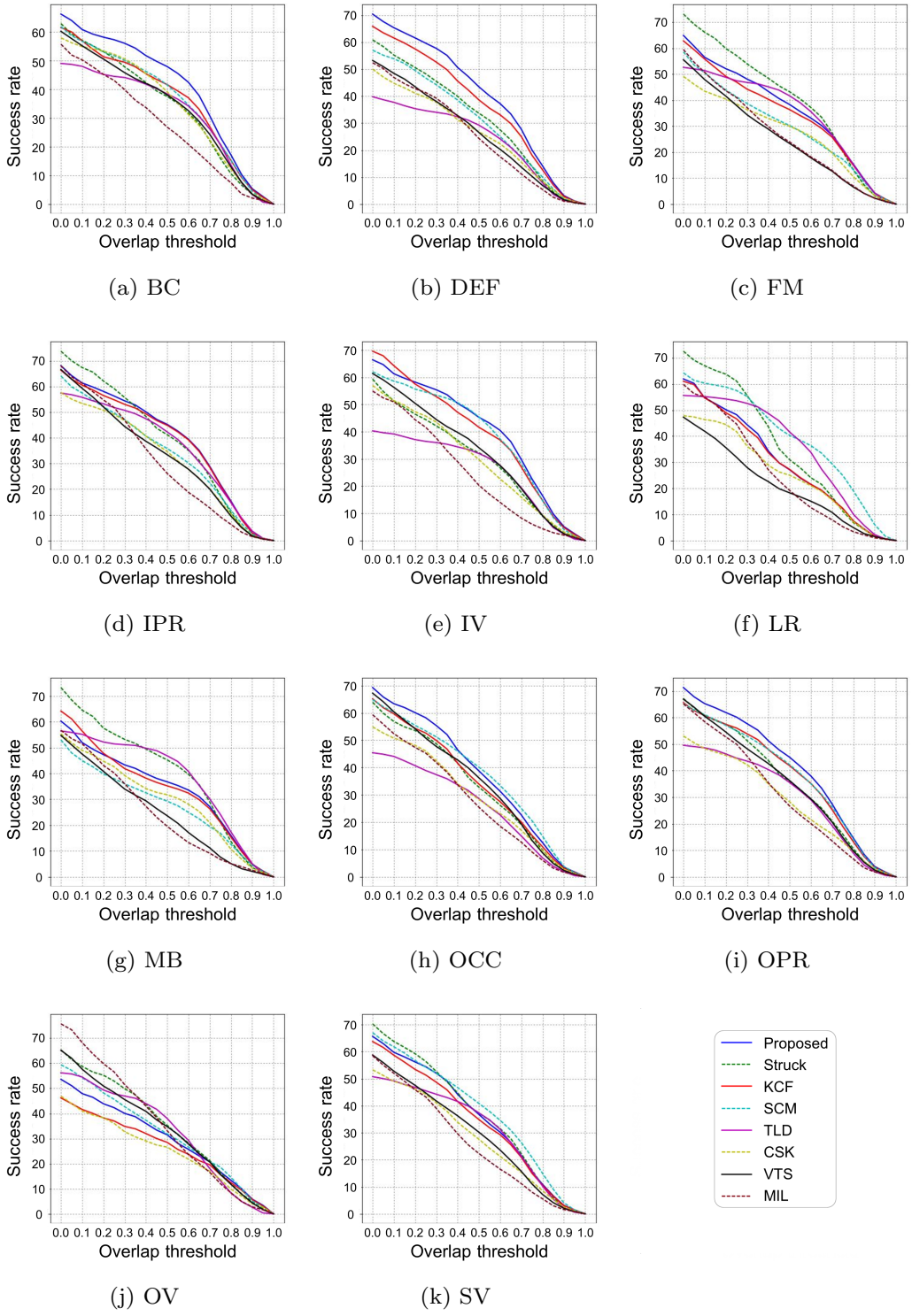
フレーム区間ごとの精度比較

提案手法のうち，相関フィルタの確率的選択に基づく人物追跡の有効性を確認するため，フレーム区間ごとの人物追跡精度の評価を行った．Fig. 3.10(a) は，追跡信頼度が低下しているフレーム区間における平均追跡成功率を示す．Fig. 3.10(b) は，パーティクルが収束し，KCF による追跡に戻った後の 10 フレームにおける平均追跡成功率を示す．

追跡信頼度が低下しているフレームにおいて，提案手法による相関フィルタの確率的選択に基づく追跡が有効であることが確認された（Fig. 3.10(a)）．また，追跡信頼度が低下しているフレームにおいて相関フィルタの確率的選択に基づく追跡を行うと，その後 KCF による追跡に戻った後の 10 フレームで他手法よりも高精度であることが確認された（Fig. 3.10(b)）．

見えの変化の種類別の評価

Table 3.1 に示した 11 種類の見えの変化の種類ごとに，提案手法の得手不得手を調べた．結果を Fig. 3.11 に示す．ほぼ全ての見えの変化において，ベースライン手法の



(a) BC (b) DEF (c) FM

(d) IPR (e) IV (f) LR

(g) MB (h) OCC (i) OPR

(j) OV (k) SV

Fig. 3.11: 見えの変化の種類別の平均追跡成功率 [%].

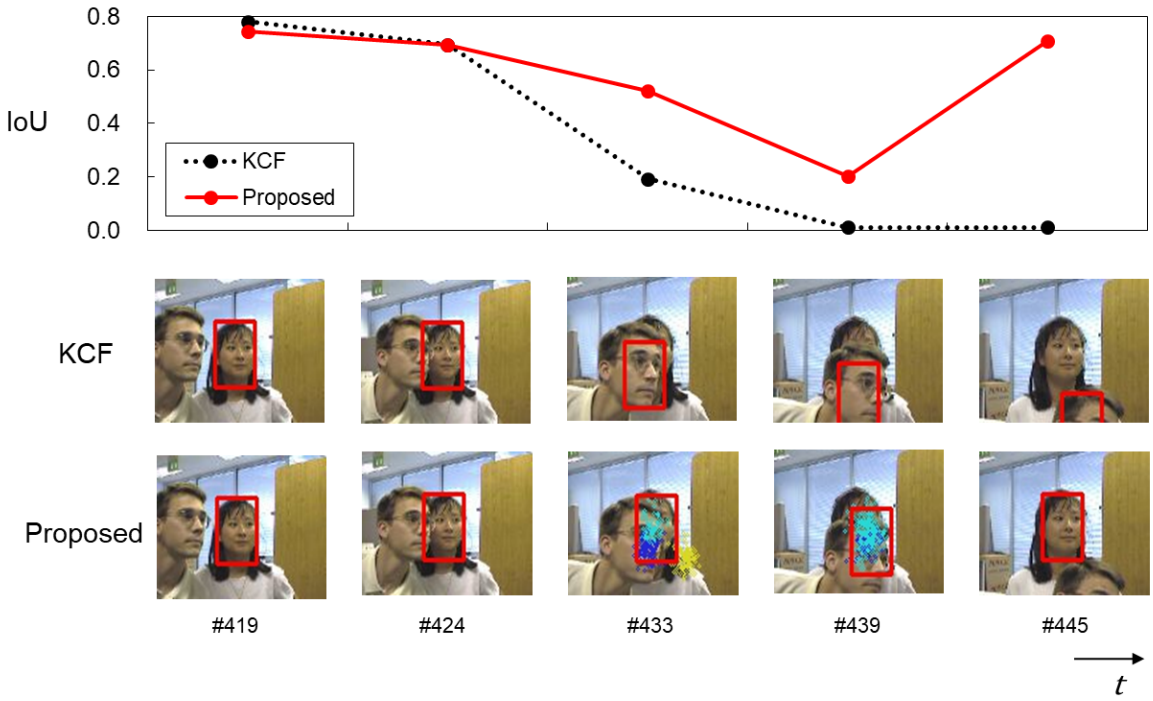


Fig. 3.12: 追跡結果の例（系列名 “女性（Girl）”）.

KCF と比較して，提案手法は追跡成功率が向上している．特に，乱雑な背景（BC），変形（DEF），遮蔽（OCC），平面外回転（OPR）に対して有効であることが分かる．

Fig. 3.12 に，遮蔽が生じた際，従来手法では追跡失敗したが提案手法が成功した例を示す．この例は，系列名 “女性（Girl）” を用いて追跡を行った結果である．KCF では，433 番目のフレームより誤って他の人物（男性）を追跡し続けている．一方，提案手法では，433 番目のフレームで追跡の信頼度が低下した際，応答マップの極大値 3 つの周りを中心にパーティクルフィルタによる追跡が開始される．439 番目のフレームでは，パーティクルフィルタにより正しい人物を追跡できている．その後 445 番目のフレームでは，パーティクルフィルタを用いず，相関フィルタのみを用いた追跡に戻っている．

逆に，照明変化（IV）とブレ（MB）では，重複率 IoU のしきい値が 0.0 から 0.2 の場合にのみ，ベースライン手法の KCF と比較して提案手法は追跡成功率が低下している．これは，提案手法によって相関フィルタを確率的に更新する前段階で，そもそも追跡対象の特徴量を正しく抽出できていない可能性がある．これに対処するためには，より高品質の見え特徴量を抽出するような別の枠組みが必要だと考えられる．

Table 3.4: 過去の追跡器をサンプリングする割合 μ を変化させた場合の AUC スコア [%].

μ	0.00	0.25	0.50	0.75	1.00
AUC	39.97	39.28	**40.64**	38.95	38.84

過去の追跡器をサンプリングする割合

過去の追跡器をサンプリングする割合 μ については，$\mu = 0.50$ のときは $\mu = 0.00$ のときよりも AUC スコアが高くなった．しかし，$\mu = 0.25, 0.75, 1.00$ のときは $\mu = 0.00$ のときよりも AUC スコアが低くなった．これは，確率分布を $P_K = 1/(t-1)$ の一様分布にしていることが原因であると考えられる．例えば，何らかの方法で追跡器自体の評価値を算出し，それに基づいた確率分布を使用する等の対策が必要である．

3.3.4 処理時間

全系列，全フレームを対象にして，1 フレームあたりの平均処理時間を計測した．CPU（i7-6700K, 4.00 GHz）のみでも 35.0 msec/frame と，ほぼ実時間での追跡を実現できた．なお，KCF では 12.7 msec/frame であった．

3.4 まとめ

本章では，識別能力が高い相関フィルタと見えの変化に頑健な時系列フィルタを組み合わせた，高精度な実時間人物追跡手法を提案した．この手法では，見えの変化が生じた際には複数の相関フィルタで追跡を行い，パーティクルフィルタの枠組みを用いて確率的に相関フィルタの更新を行う．このようにすることで，誤った矩形領域を用いた相関フィルタの更新を低減する．TB-50 データセットを用いた評価実験では，提案手法が物体の遮蔽・変形・回転のような見えの変化に対して頑健であることを確認した．追跡精度については，ベースライン手法が 38.38% であったのに対し，提案手法では 40.64% に向上し，提案手法の有効性が確認された．残された課題は，照明変化やブレのある画像に対しても追跡精度を向上させることである．

本章においては，単一人物追跡に焦点を当て，人物位置推定のずれに対して解決を目指した．この問題の解決に貢献したのは，パーティクルフィルタの枠組みを用いた確率的な

相関フィルタの更新であった．以降の第 4 章及び第 5 章では複数人物追跡に焦点を当て，人物 ID 推定の誤りに対して解決を目指す．特に次章では，より多くの人物を追跡するため，広域を撮影できる全方位カメラを用いた複数人物追跡手法を提案する．

第 4 章

全方位カメラを用いた見えの歪みに頑健な複数人物追跡

第 3 章では単一の人物を追跡対象とする単一人物追跡に焦点を当てた．本章及び続く第 5 章では複数の人物を追跡対象とする複数人物追跡に焦点を当てる．複数人物追跡では，可能な限り多くの人物を追跡するために広域を撮影できることが望ましい．本章では，360 度の画角を持つ全方位カメラを用いた見えの歪みに頑健な追跡手法を提案する．全方位カメラで撮影した画像の例を Fig. 4.1 に示す．全方位カメラを用いた場合，レンズの歪みによって，人物の見えが位置によって大きく異なったり，位置がフレーム間で非線形に変化する．このような画像に通常カメラ向けの従来手法をそのまま適用すると，フレーム間の人物対応付けが失敗しやすい．この問題に対応するため，本章では人物の 3 次元モデルを用いた追跡手法を提案する．提案手法では，1）人物領域のみを局所的にパノラマ展開してから見え特徴量を抽出する，2）距離指標が均一な世界座標系での位置を位置特徴量とする．評価実験では，独自に作成した LargeRoom・SmallRoom データセットを用いて，人物の見えや位置がフレーム間で大きく変化する場合でも，提案手法により正しく追跡できることを確認する．

4.1　はじめに

複数人物追跡は幅広い分野で応用される重要技術で，例えばマーケティングにおける顧客の興味商品推定時や，監視における平常時の広域確認に用いられる．複数人物追跡では，より多くの人物を同時に追跡するために広域を撮影できることが望ましく，本章では

Fig. 4.1: 天井に固定した全方位カメラで撮影した画像の例.

天井に固定した 360 度の画角を持つ全方位カメラ 1 台を用いる．2.4 節で述べた通り，全方位画像に通常カメラ向けの追跡手法を適用した場合，以下の 2 つの要素によって ID スイッチが多発する．1）人物が画像座標系上の x 軸と垂直または並行でない角度で撮影された場合，矩形内に占める背景領域の割合が大きくなってしまい，見え特徴量に悪影響を及ぼす（Fig. 2.6(a)）．2）画像内で距離指標が不均一であるため，隣接フレーム間で人物の位置が非線形に変化する（Fig. 2.6(b)）．一方，全方位画像をパノラマ展開した後に通常の画像向けの人物追跡手法を適用することもできる．しかし，パノラマ展開前の画像の中心付近にいる人物は展開後に，頭部から足元に向かって画像の横軸方向に大きく引き延ばされてしまうため，矩形内に占める背景領域の割合は大きくなってしまう．

本章では，以上の 2 つの問題に対応するために設計した対応付け特徴量を用いて，複数人物を追跡する手法を提案する．この特徴量には以下の 2 つの性質がある．1）見え特徴量抽出のために，人物領域のみを局所的にパノラマ展開する．2）位置特徴量として，距離指標が均一な世界座標系で位置を表現する．このようにして設計した特徴量は，任意の

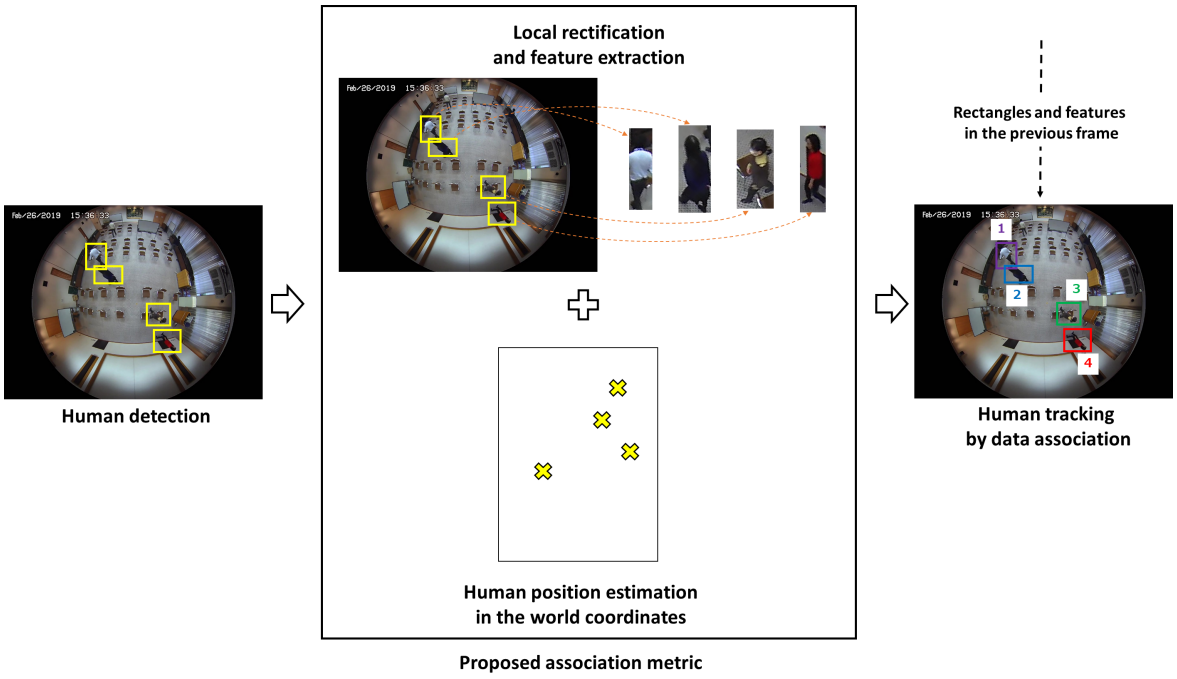


Fig. 4.2: 提案手法の処理手順．まず，対象人物を画像座標系上で検出する．次に，その領域を局所的に展開し，見え特徴量を抽出する．また，対象人物の位置を世界座標系上で推定し，これを位置特徴量とする．そして，得られた見え特徴量と位置特徴量を用いて人物対応付けを行い，フレーム間で追跡する．

state-of-the-art の追跡器で利用可能で，その追跡精度を向上させることができると期待できる．

本章の以後の構成は次の通りである．まず 4.2 節で全方位画像の歪みの影響を低減可能な対応付け特徴量を用いた追跡手法を提案する．次に 4.3 節で評価実験について述べる．最後に 4.4 節で本章をまとめる．

4.2　全方位画像の歪みに頑健な特徴量を用いた複数人物追跡

4.2.1　処理手順

提案手法は検出による追跡手法の一種である．提案手法の処理手順を Fig. 4.2 に示す．まず画像座標系上で人物検出を行った後，2 つの対応付け特徴量を抽出する（4.2.2 項）．次にそれらの特徴量を用いて人物対応付けをすることによって，フレーム間で複数人物追

Algorithm 2 フレーム $\mathbf{o}_t$ における提案手法のアルゴリズム．

Input: : $\mathbf{o}_t$, $A_{t-1} = (\mathbf{a}_{t-1}^1, \mathbf{a}_{t-1}^2, \cdots, \mathbf{a}_{t-1}^{N_a})$
Output: : $A_t = (\mathbf{a}_t^1, \mathbf{a}_t^2 \cdots)$
  Calculate $B_t = (\mathbf{b}_t^1, \mathbf{b}_t^2, \cdots, \mathbf{b}_t^{N_b})$
  for $j = 1$ to N_b **do**
    Estimate $\mathbf{r}_t^j$ and $\mathbf{q}_t^j$ using the estimator.
    Extract $\mathbf{x}_t^j$ using the feature extractor.
  end for
  for $i = 1$ to N_a **do**
    for $j = 1$ to N_b **do**
      Calculate $c^{\text{feat}}(\mathbf{a}_{t-1}^i, \mathbf{b}_t^j)$ using $\mathbf{x}_{t-1}^i$ and $\mathbf{x}_t^j$.
      Calculate $c^{\text{pos}}(\mathbf{a}_{t-1}^i, \mathbf{b}_t^j)$ using $\mathbf{r}_{t-1}^i$ and $\mathbf{r}_t^j$.
    end for
  end for
  Apply the Hungarian algorithm to C^{feat} and C^{pos} for estimating A_t.
  Create new tracklets and delete tracklets.

跡を行う（2.3.2 項）．提案手法が従来手法と異なる点は，4.2.2 項で述べる対応付け特徴量である．提案手法の追跡アルゴリズムを Algorithm 2 に示し，以下ではその詳細について説明する．

4.2.2　全方位画像の歪みに頑健な特徴量

人物検出器により検出された矩形（本章ではこれを通常矩形と呼ぶ）に対して，局所的パノラマ展開（4.2.2.2）と世界座標系上での人物位置推定（4.2.2.3）を行う．そのために事前に，局所的パノラマ展開のための “回転矩形” と，世界座標系上での位置の推定器を学習しておく（4.2.2.1）．回転矩形は $\mathbf{r} = (x_r, y_r, w_r, h_r, \phi)$ で表す．x_r と y_r は，それぞれ矩形の中心点の x 座標と y 座標を表す．w_r と h_r は，それぞれ矩形の幅と高さを表す．ϕ は，x_r と y_r は回転中心とし，時計回りを正とした回転角を示し，水平方向の正の方向を $\phi = 0$ とする．ϕ の定義域は $0 \leq \phi < 2\pi$ である．なお，通常矩形及び回転矩形は画像座標系上で表現する．

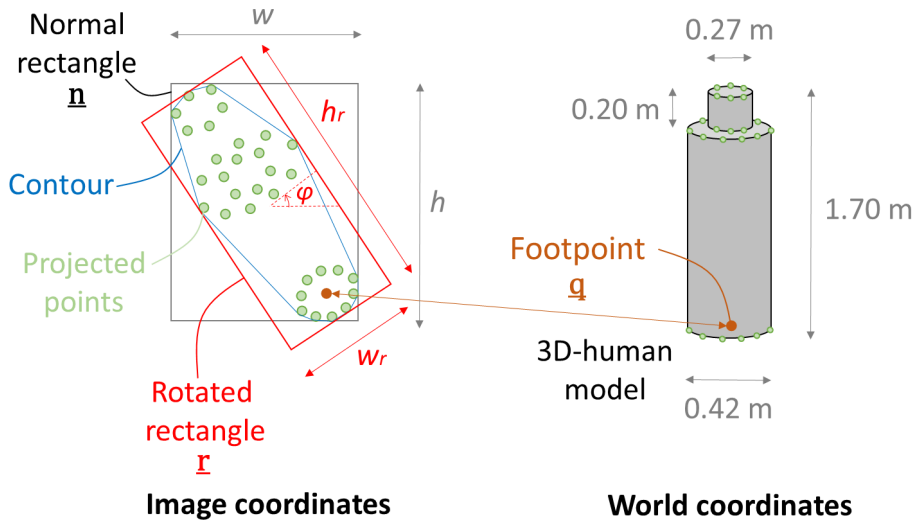


Fig. 4.3: 世界座標系上に 3 次元人物モデルを仮想的に配置し，点群で構成される人物輪郭を画像座標系上に変換する．その後，画像座標系上で通常矩形と回転矩形を求める．

4.2.2.1　回転矩形・位置推定器の学習

提案手法では，推定器の入力は，人物検出器により検出された通常矩形 $\mathbf{b} = (x, y, w, h)$ である．一方，推定器の出力は，回転矩形 $\mathbf{r} = (x_r, y_r, w_r, h_r, \phi)$ と世界座標系における位置 $\mathbf{q} = (q_1, q_2, 0)$ である．

通常矩形 $\mathbf{b}$ は，そのまま位置特徴量として使用することもできる．しかし，人物の形や大きさはある程度決まっており，これは見え特徴量や位置特徴量を抽出する際に大きな情報となる．提案手法では，Fig. 4.3 の右図に示すような 2 つの円筒で構成される 3 次元人物モデル [97, 116, 117] を使用する．

まず，3 次元人物モデルを使用するための準備として，画像座標系と世界座標系との間の変換について説明する．$\mathbf{p} = (p_1, p_2)$ は，画像座標系における位置を表し，p_1 と p_2 はそれぞれ x 座標と y 座標を表す．また，$\mathbf{q} = (q_1, q_2, q_3)$ は世界座標系における位置を表し，q_1, q_2, q_3 はそれぞれ x, y, z 座標を表す．次式のように，カメラ射影行列 $M \in \mathbb{R}^{3\times4}$ を用いて世界座標系から画像座標系への変換を行う [118]．

$$\lambda \mathbf{p}^{\mathsf{T}} = M \mathbf{q}^{\mathsf{T}} \tag{4.1}$$

世界座標系における仮想的な 3 次元人物モデルと，画像座標系における人物輪郭を

Fig. 4.3 に示す．画像座標系と世界座標系との間で，人物領域は足元位置を経由して，対応付けられる．画像座標系における各足元位置 $\mathbf{p} = (p_1, p_2)$ ごとに，以下の手順を反復する．

- $\mathbf{p}$ は，式 (4.1) によって，世界座標系における位置 $\mathbf{q} = (q_1, q_2, q_3)$ に変換され，足元位置は $\mathbf{q} = (q_1, q_2, 0)$ として算出される．
- $\mathbf{q}$ に従って，3 次元人物モデルを世界座標系に仮想的に配置する．
- 配置された 3 次元人物モデルを用いて，複数の点で構成される人物輪郭を画像座標系上で算出する．世界座標系における各頂点は，式 (4.1) によって画像座標系に変換される．
- 通常矩形 $\mathbf{b} = (x, y, w, h)$ と回転矩形 $\mathbf{r} = (x_r, y_r, w_r, h_r, \phi)$ は，人物輪郭を用いて画像座標系上で算出される．両矩形は，面積最小基準による外接矩形として算出される．
- クエリベクトル $\mathbf{b}$ と回転矩形 $\mathbf{r}$ との対応関係を登録する．また，クエリベクトル $\mathbf{b}$ と足元位置 $\mathbf{q}$ との対応関係も登録する．

4.2.2.2　画像座標系上における局所的パノラマ展開と見え特徴量の抽出

局所的パノラマ展開は，回転矩形を推定し，それを回転することによって行う．具体的には，回転矩形 $\mathbf{r} = (x_r, y_r, w_r, h_r, \phi)$ は，クエリベクトル $\mathbf{b}$ を推定器に入力することによって得る．ただし，4.2.2.1 の手順によって登録されたクエリベクトル群は，あり得る $\mathbf{b}$ を網羅しているわけではない．そのため，登録されたクエリベクトル群の中から，入力されたクエリベクトル $\mathbf{b}$ と最も近いものを最近傍探索によって得る．最近傍探索の高速化のために Kd-tree [119] を用いる．

回転矩形 $\mathbf{r} = (x_r, y_r, w_r, h_r, \phi)$ は，足元位置が常に下の位置となるように（$\phi = 0$）回転させる．回転は次式の回転行列によって行う．

$$R = \begin{pmatrix} \alpha & -\beta & (1-\alpha)x_r - \beta y_r \\ \beta & \alpha & \beta x_r - (1-\alpha)y_r \end{pmatrix} \tag{4.2}$$

$$\alpha = \cos(-a) \tag{4.3}$$

$$\beta = \sin(-a) \tag{4.4}$$

ここで，x_r と y_r は回転中心を示す．

Table 4.1: LargeRoom データセットの詳細.

Sequence ID	1	2	3	4	5	6	7	8	9	10
Camera ID	1	2	1	2	1	2	1	2	1	2
Number of humans	6	6	6	6	6	6	6	6	6	6
Sequence length [sec]	180	180	180	180	180	180	180	180	180	180

Table 4.2: SmallRoom データセットの詳細.

Sequence ID	1	2	3	4	5	6	7	8
Camera ID	1	2	1	2	1	2	1	2
Number of humans	4	4	10	10	3	3	3	3
Sequence length [sec]	182	182	155	155	268	268	105	105

そして，矩形 $\mathbf{b}$ に対応する見え特徴量 $\mathbf{x}$ を算出する．人物領域は局所的に展開されるため，背景領域の削減と人物向きの正規化によって特徴量の質が向上する．見え特徴抽出器には 2.3.2 項と同様 Siamese ネットワークを用いる．

4.2.2.3 世界座標系上における人物位置推定

世界座標系における足元位置 $\mathbf{q} = (q_1, q_2, 0)$ は，クエリベクトル $\mathbf{b}$ を推定器に入力することによって得られる．

4.3 実験

提案手法の有効性と効率を評価するため，複数人物追跡実験を行った．

4.3.1 データセット

提案手法が有効である部屋の広さを確かめるため，異なる面積の部屋でデータセットを 2 種類作成した．全方位カメラには，Panasonic WV-SF438*の魚眼カメラを用いた．撮影した画像の解像度は $1,280 \times 960$ 画素で，時系列画像のフレームレートは 15 fps とし

* https://security.panasonic.com/products/wv-sf438/

た．カメラパラメータは OCamCalib[†]を用いて算出した．

LargeRoom データセットでは，面積 128 m^2（8 m × 16 m）の部屋において 6 人の人物が回遊する状況を撮影した．一方，SmallRoom データセットでは，面積は 36 m^2（4 m × 9 m）の部屋において 3–10 人の人物が回遊する状況を撮影した．データセットの詳細を Table 4.1 及び Table 4.2 に示す．

4.3.2　実験条件

人物検出器（SSD）では，Zhreshold らの実装[‡]のデフォルトハイパパラメータを用いた．人物検出器は，様々な部屋（SmallRoom を含む）で撮影された 220,874 枚の画像を用いて学習した．人物対応付けのパラメータは，2 つのパラメータ $\varepsilon_{\text{feat}} \in \{200, 300, 400\}$ と $\varepsilon_{\text{pos}} \in \{0.3, 0.5, 0.7, 0.9\}$ を変化させ，検証用データを用いて $\varepsilon_{\text{feat}} = 300$, $\varepsilon_{\text{pos}} = 0.7$ に設定した．

CPU には Intel Core i7-7700K 4.20GHz，メモリには 32GB RAM，そして GPU には NVIDIA GeForce Titan X Pascal を用いた．

評価指標として，Multiple Object Tracking Accuracy（MOTA）指標 [120] を用いた．MOTA は追跡精度の評価に最も広く用いられている指標で，3 つの誤差指標を組み合わせた総合的な指標であり，次式のように定義されている．

$$\text{MOTA} = 1 - \frac{\text{FN} + \text{IDs} + \text{FP}}{\text{DET}} \tag{4.5}$$

ここで，FN，IDs，FP，DET は，それぞれ，未検出，ID スイッチ，過検出，検出の総数を示す．MOTA の値域は $(-\infty, 100]$ である．MOTA は，ID スイッチ回数が正規化されたものとして捉えることもできる．矩形の正解データは通常矩形で記述されているため，提案手法も通常矩形 $\mathbf{b} = (x, y, w, h)$ を出力するようにした．正解データは 1 fps 分を作成したため，撮影した全フレームのうち，正解データとしてアノテーションされた分のみを評価した．

Table 4.3: 追跡の評価結果（MOTA）.

	Feature	Position	1 fps		15 fps	
			LargeRoom	SmallRoom	LargeRoom	SmallRoom
Baseline	No rectification	— (Without using)	8.4	51.0	−3.5	48.7
SORT [91]	— (Without using)	Rectangle in image	−10.5	24.5	−1.4	49.9
DeepSORT [50]	No rectification	Rectangle in image	−4.3	30.9	6.0	**52.3**
Proposed	Local rectification	— (Without using)	10.2	51.8	−3.4	48.5
	— (Without using)	Point in world	8.1	53.1	−1.1	49.7
	Local rectification	Point in world	**11.7**	**53.3**	**6.6**	52.0

4.3.3 人物追跡の評価

提案する対応付け特徴量を構成する見え特徴量と位置特徴量それぞれ単体と，それらの組み合わせについて評価した．また，提案手法が既存手法の問題を解決することについて確認した．評価データは，フレームレートは 15 fps の時系列画像を用いた場合と，それを 1 fps に間引いて使用した場合に分けて評価した．評価指標には，全系列に対する MOTA の値を用いた．見え特徴量は，局所的パノラマ展開（Local rectification）の有無を比較評価した．また，位置特徴量は，画像座標系における矩形（Rectangle in image coordinates）と世界座標系における点（Point in world coordinates）を比較評価した．追跡の評価結果を Table 4.3 に示す．

4.3.3.1 局所的パノラマ展開の評価

位置特徴量は使用せず，局所的パノラマ展開の有無で追跡精度を比較した．まず LargeRoom データセットに対する評価結果を示す．1 fps では MOTA は 1.8 ポイント向上した（8.4 対 10.2）．一方，15 fps では MOTA はほぼ同等であった（−3.5 対 −3.4）．次に SmallRoom データセットに対する評価結果を示す．1 fps では MOTA は 0.8 ポイント向上した（51.0 対 51.8）．一方，15 fps では MOTA はほぼ同等であった（48.7 対 48.5）．以上より，局所的パノラマ展開は低フレームレートの場合に特に有効であることが分かっ

† https://sites.google.com/site/scarabotix/ocamcalib-toolbox/
‡ https://github.com/zhreshold/mxnet-ssd/

Table 4.4: 提案手法による MOTA の向上幅.

局所的パノラマ展開	展開人物が垂直または平行である場合	+0.011
	人物が斜めである場合	+0.031
世界座標表現	人物が原点の近くにいる場合	+17.703
	人物が周縁部にいる場合	+1.757

た．また，通常のフレームレートの場合でも，局所的パノラマ展開の有無によらず同程度に有効であることが分かった．さらに，1 fps では，局所的パノラマ展開は SmallRoom データセットよりも LargeRoom データセットに対して効果を発揮することも分かった（LargeRoom：+1.8 対 SmallRoom：+0.8）．

4.3.3.2　世界座標表現の評価

見え特徴量は使用せず，画像座標系における矩形を用いた場合と世界座標系における点を用いた場合で追跡精度を比較した．まず LargeRoom データセットに対する評価結果を示す．1 fps では MOTA は 18.6 ポイント向上した（−10.5 対 8.1）．一方，15 fps では MOTA はほぼ同等であった（−1.4 対 −1.1）．次に SmallRoom の結果を示す．1 fps では MOTA は 28.6 ポイント向上した（24.5 対 53.1）．一方，15 fps では MOTA はほぼ同等であった（49.9 対 49.7）．以上より，世界座標表現は低フレームレートの場合に特に有効であることが分かった．また，通常のフレームレートの場合でも，世界座標表現は画像座標表現と同程度に有効であることが分かった．さらに，1 fps では，世界座標表現は LargeRoom データセットよりも SmallRoom データセットに対して効果を発揮することも分かった（LargeRoom：+18.6 対 SmallRoom：+28.6）．

4.3.3.3　傾向分析

提案手法が特に有効である場合について分析する．分析のために，画像座標 (X, Y) を極座標 (θ, r) に変換した．この際に画像座標の中心点 $(640, 480)$ を極座標の原点とした．θ [rad] は原点とのなす角を示し，r [pixel] は原点との距離を示す．ここでは，部屋の面積が広く，傾向がより顕著に表れると考えられる LargeRoom データセットを分析対象とした．

局所的パノラマ展開は θ がどのような値の場合に有効であるか，世界座標表現は

r がどのような値の場合に有効であるかを分析した．局所的パノラマ展開については，人物がおおむね垂直または平行（$-9/8\pi < \theta \leq -7/8\pi$, $-5/8\pi < \theta \leq -3/8\pi$, $-1/8\pi < \theta \leq 1/8\pi$, $3/8\pi < \theta \leq 5/8\pi$）である場合と，人物が斜め（$-7/8\pi < \theta \leq -5/8\pi$, $-3/8\pi < \theta \leq -1/8\pi$, $1/8\pi < \theta \leq 3/8\pi$, $5/8\pi < \theta \leq 7/8\pi$）である場合とで，MOTA の向上を比較した．MOTA の向上は，局所的パノラマ展開がない場合を基準にして算出した．一方，世界座標表現については，人物が原点の近くにいる（$0 < r \leq 200$）場合の MOTA と，人物が周縁部にいる（$300 < r \leq 500$）場合とで，MOTA の向上を比較した．MOTA の向上は，画像座標系における矩形を用いた場合を基準にして算出した．Table 4.4 に分析結果を示す．

まず，局所的パノラマ展開について分析する．人物が垂直または平行である場合，MOTA は 0.011 ポイント向上した．一方，人物が斜めである場合，MOTA は 0.031 ポイント向上した．これより，局所的パノラマ展開は，人物が垂直または平行である場合よりも，人物が斜めである場合の方が有効であることが分かった．これは局所的パノラマ展開により背景が除去されることによるものと考えられる．

次に，世界座標表現について分析する．人物が原点の近くにいる場合に MOTA は 17.703 ポイント向上した．一方，人物が周縁部にいる場合に MOTA は 1.757 ポイント向上した．これより，世界座標表現は，人物が周縁部にいる場合よりも人物が原点の近くにいる場合の方が有効であることが分かった．

4.3.3.4 局所的パノラマ展開と世界座標表現

局所的パノラマ展開と世界座標表現の組み合わせの評価を，1 fps の動画像を用いて行った．なお以下では，局所的パノラマ展開を評価する際は位置特徴量を用いず，同様に，世界座標表現を評価する際は見え特徴量を用いずに精度を算出した．

LargeRoom データセットに対しては，局所的パノラマ展開は世界座標表現よりも有効であった（10.2 対 8.1）．これは，4.3.3.3 で示した通り，局所的パノラマ展開は人物が画像の周縁部にいる場合に有効だからである．局所的パノラマ展開と世界座標表現を組み合わせた場合，MOTA は 1.5 ポイント向上した．これは，4.3.3.3 で示した通り，世界座標表現は人物が原点の近くにいる場合に有効だからである．Fig. 4.4 に LargeRoom データセットを用いた追跡結果の例を示す．いくつかのフレームにおいて，提案手法により ID スイッチを防げたことが分かる．

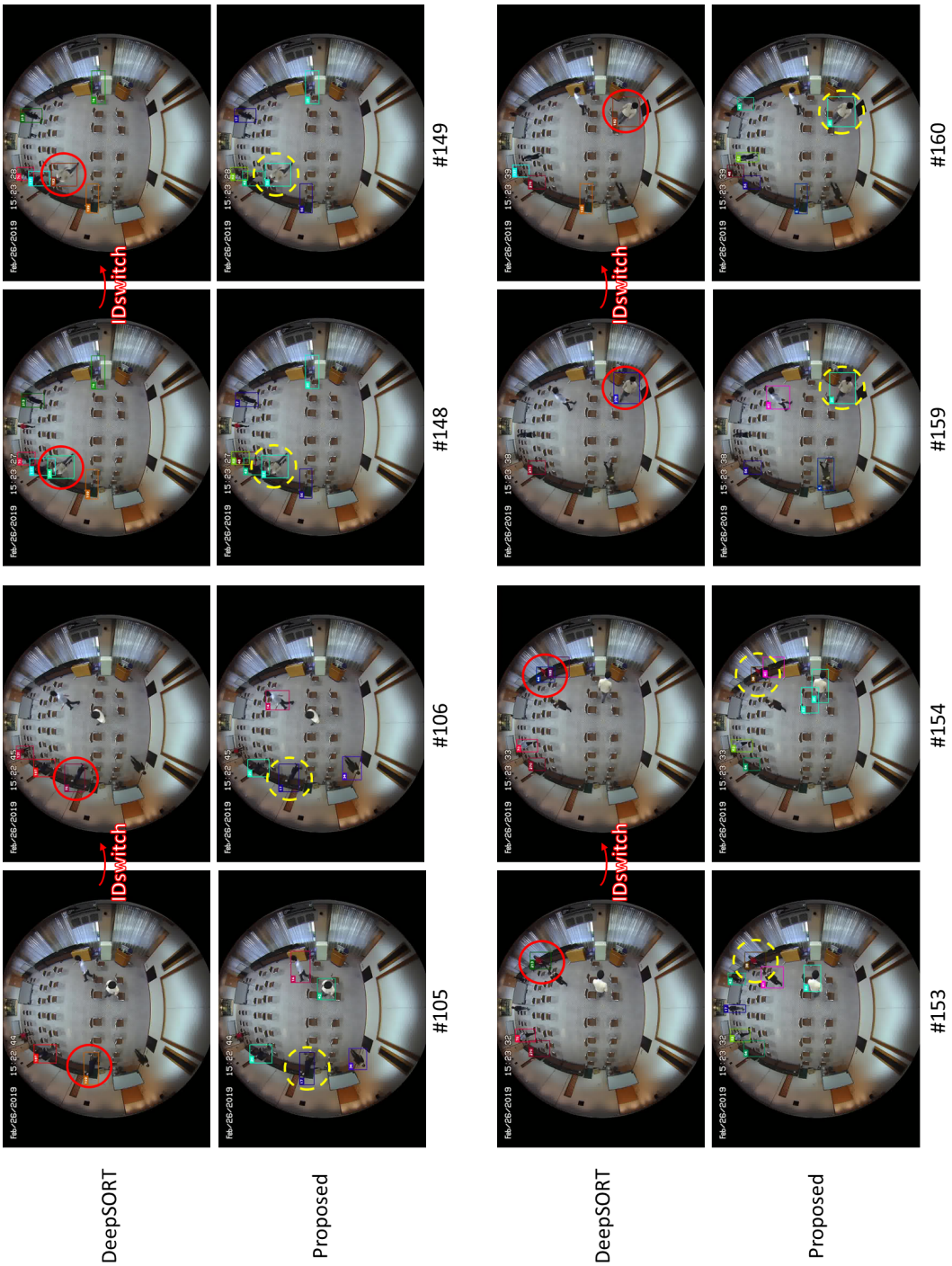


Fig. 4.4: 追跡結果の例．#(number) はフレーム番号を示す．実線の円は ID スイッチ，点線の円は ID スイッチが防止された場合を示す．

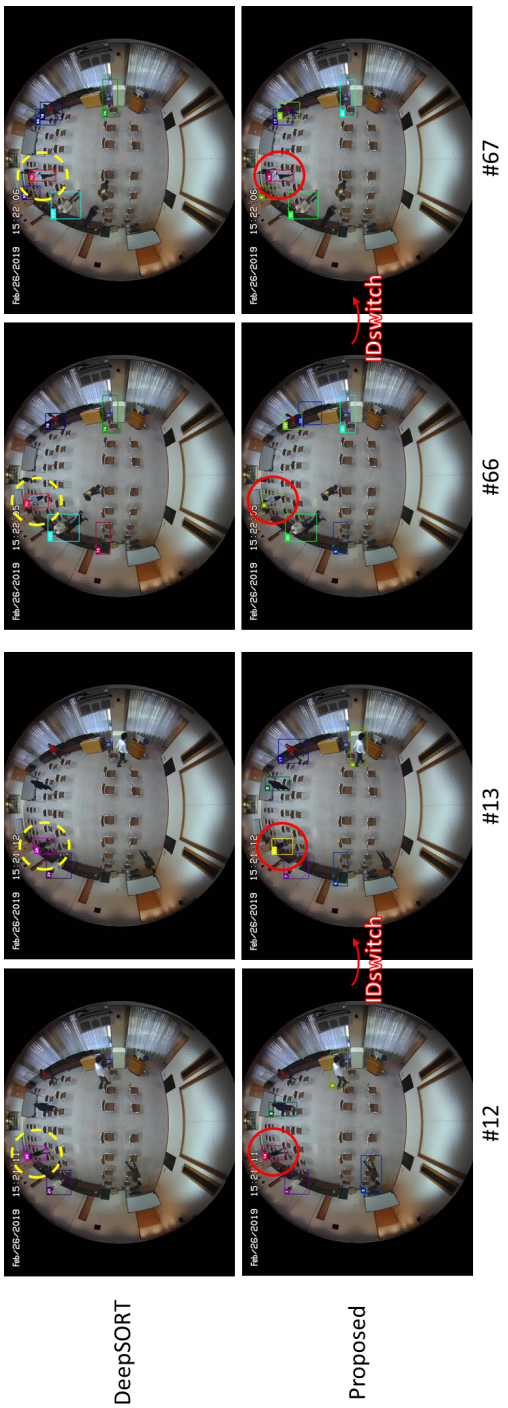


Fig. 4.5: 提案手法で ID スイッチが生じた追跡結果の例．#(number) はフレーム番号を示す．実線の円は ID スイッチ，点線の円は正しく追跡された場合を示す．

一方，Fig. 4.5 に，提案手法で ID スイッチが生じた追跡結果の例を示す．4.2.2.1 で述べた通り，提案手法は画像座標系と世界座標系との間の変換を基にしている．これらの例で人物が写る画像の周縁部では，1 画素が表現する世界座標系での位置の曖昧性が大きくなる．それによって局所的パノラマ展開・世界座標表現の精度が低下したことが，ID スイッチの原因ではないかと推測される．

一方，SmallRoom データセットに対しては，世界座標表現は局所的パノラマ展開よりも有効であった（51.8 対 53.1）．これは，4.3.3.3 で示した通り，世界座標表現は人物が原点の近くにいる場合に有効だからである．しかし，局所的パノラマ展開と世界座標表現を組み合わせた場合，MOTA は 0.2 ポイントのみの向上であった．人物が原点近くにいる場合は背景領域が小さいため，局所的パノラマ展開の効果が限定的であったと考えられる．

以上より，提案手法は低フレームレート（1 fps）でかつ広い部屋（LargeRoom）に適用する場合に特に有効であると考えられる．

4.3.3.5　フレームレートに関する考察

一般的には高フレームレートの方が低フレームレートよりも追跡精度が高いが，Table 4.3 の実験結果を見ると，提案手法では低フレームレートの場合の方が精度が高い．この理由について以下で考察を行う．実験では，高フレームレート（15 fps）の場合も低フレームレート（1 fps）の場合も，正解データは同じ 1 fps 分のみを用いた．よって，高フレームレートで ID スイッチが生じる機会は，低フレームレートの 15 倍であり，ID スイッチが多発したのではないかと考えられる．ただし，15 fps よりもさらに高フレームレートにしていくと，ID スイッチが生じる機会が増加する効果よりも，高フレームレートにすることによる追跡問題の簡単化の効果が上回り，低フレームレート（1 fps）の場合よりも追跡精度が高くなると予想される．

4.3.4　処理時間

回転矩形と世界座標系における人物位置の推定に要する処理時間を計測した．実験には LargeRoom データセットと SmallRoom データセットを用い，それぞれ系列 1 の全フレームの平均処理時間を計測した．Table 4.5 に計測結果を示す．LargeRoom データセット・SmallRoom データセットのいずれにおいても，回転矩形と世界座標系における人

Table 4.5: 1 フレームに要する処理時間 [msec].

	LargeRoom	SmallRoom
Human detection	18.8	19.9
Rotated rectangle estimation	0.2	0.2
Human position estimation	0.1	0.2
Feature extraction	64.4	46.5
Data association	2.1	1.0
Total	85.5	67.8

物位置の推定に要する時間は非常に少ない．合計の処理時間は，LargeRoom データセット・SmallRoom データセットのいずれにおいても 10 fps 以上であった．一方，4.3.3.4 では，提案手法は入力動画像が低フレームレート（1 fps）の場合でも有効であることを示した．よって，提案手法により，高精度な追跡が実時間処理で実現できることが分かった．

さらなる高速化のためには，全体の大部分を占めている人物検出や特徴抽出の処理時間の削減が必要である．そのための方法として，より高速な人物検出器・特徴抽出器を用いる，あるいは入力画像を縮小する等が考えられる．

4.4 まとめ

本章では，全方位画像の歪みの影響を低減する対応付け特徴量を用いることによって，広域中に複数人物が存在する環境において高精度に人物追跡を行う手法を提案した．提案手法には次の 2 つの特徴がある．1）局所的パノラマ展開によって背景領域の影響を低減する．2）距離空間が均一である世界座標系で人物の位置を表現する．独自に作成したデータセットを用いた評価実験では，局所的パノラマ展開と世界座標表現がともに有効であることを確認し，特に低フレームレート（1 fps）の場合に有効であることが分かった．追跡精度については，提案手法により，LargeRoom データセットで MOTA が 3.3 ポイント，SmallRoom データセットで MOTA が 2.3 ポイント向上することを確認した．また，提案対応付け特徴量の算出には，フレーム内の 1 人物あたり，0.43 msec しか要さず，実時間処理可能なことも確認した．今後，背景領域の削減や向きの正規化だけでなく，人物の見え特徴量自体の正規化が必要である．

本章においては，複数人物追跡に焦点を当て，より多くの人物を追跡するため，広域を撮影できる全方位カメラを用いた手法を検討した．その際に発生する人物 ID 推定の誤りに対して解決を目指した．この問題の解決に貢献したのは，局所的パノラマ展開と世界座標表現であった．次章では，広域を撮影する別の手段として，移動して撮影できるドローン搭載カメラを用いた複数人物追跡手法を提案する．

第 5 章

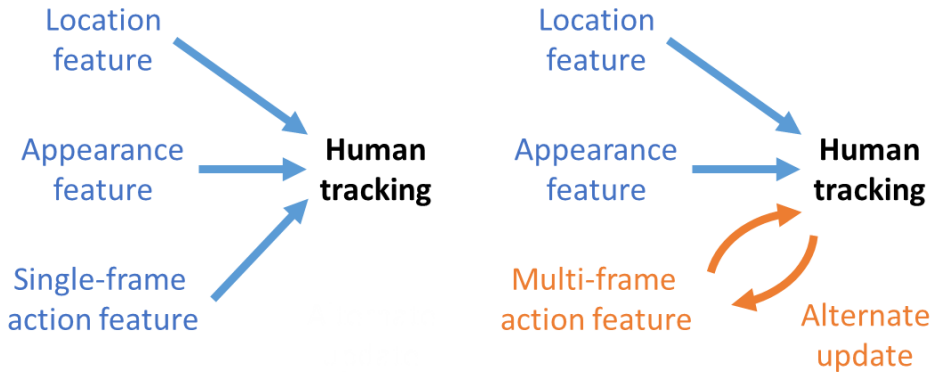
ドローン搭載カメラを用いたブレに頑健な複数人物追跡

本章でも第 4 章と同様に，複数の人物を追跡対象とする複数人物追跡に焦点を当てる．第 4 章では可能な限り多くの人物を追跡するために 360 度の画角を持つ全方位カメラを用いた追跡手法を提案したが，本章では移動可能なドローン搭載カメラを用いたブレに頑健な追跡手法を提案する．ドローン搭載カメラで撮影した画像の例を Fig. 5.1 に示す．ドローン搭載カメラを用いた場合，ドローンの急な動きによって，人物の見えや位置がフレーム間で急激に変化してしまう．そこに固定カメラを対象とした従来手法を適用すると，フレーム間の人物対応付けが失敗しやすい．この問題に対応するために，本章では人物の行動に関する特徴量を用いることでブレに頑健な追跡手法を提案する．提案手法では，推定した人物軌跡に基づいて行動特徴量を更新し，再度人物追跡に用いる．評価実験では，Okutama-Action データセット [2] 及び Drone-Action データセット [3] を用いて，ドローンの急な動きによって人物の見えや位置が急激に変化した場合でも，提案手法により正しく追跡できることを確認する．

5.1　はじめに

複数人物追跡は幅広い分野で応用される重要技術で，例えばマーケティングにおける顧客の興味商品推定時や，監視における平常時の広域確認に用いられる．冒頭で述べたように，複数人物追跡では可能な限り多くの人物を追跡するために広域を撮影できることが望ましいため，本章では移動可能なドローン搭載カメラを用いる．しかし，ドローンの急な

Fig. 5.1: ドローン搭載カメラで撮影した画像の例（文献 [2] より転載）.



(a) 従来手法　　(b) 提案手法

Fig. 5.2: 従来手法と提案手法の違い.

動きによってブレが生じ，人物の見えや位置がフレーム間で急激に変化してしまう．そこに 2.3.2 項で述べた検出による手法を単純に適用すると，フレーム間の人物対応付けが失敗しやすい．ブレに頑健な行動特徴量を用いる手法 [35, 64, 65] も存在するものの，特徴量を単一のフレームのみから抽出している．本章ではこのような特徴量を Single-frame Action Feature（SAF）と呼ぶ．SAF は他のフレームの情報を考慮しておらず，ドロー

ンの急な動きによってブレが生じた場合に不安定であるため，依然としてフレーム間の人物対応付けが失敗しやすい．これに対して本章では，複数フレームから抽出する行動特徴量（Multi-frame Action Feature ; MAF）を提案する．しかし，MAF と人物追跡は相互に依存関係にある．つまり，人物追跡のための MAF を抽出するためには，人物追跡（軌跡推定）があらかじめ完了している必要がある．

本章では，人物軌跡と複数フレーム行動特徴量を交互に更新する複数人物追跡手法（Multiple Human Tracking using MAF ; MHT-MAF）を提案する．提案手法では，一度人物追跡が完了した後，推定した軌跡に基づいて MAF を抽出し，再度人物追跡に用いる．そして，このような軌跡と MAF の更新を数回反復する．Fig. 5.2 に，SAF を用いた従来手法と MAF を用いた提案手法の違いを示す．従来手法では人物追跡に一度のみ行動特徴量を用いるのに対し（Fig. 5.2(a)），提案手法では人物追跡結果（軌跡）と行動特徴量の更新を反復する（Fig. 5.2(b)）．このようにして安定化した MAF によって，ブレが生じた場合でも ID スイッチを防止することができる．なお，提案手法は，一般にオンライン手法よりも精度が高いオフライン手法として設計している．

本章の以後の構成は次の通りである．まず 5.2 節で MAF の基となる SAF を用いた複数人物追跡手法について述べる．次に 5.3 節では人物軌跡と MAF を交互に更新する手法を提案する．さらに 5.4 節では評価実験について報告する．最後に 5.5 節で本章をまとめる．

5.2　単一フレーム行動特徴量（SAF）を用いた複数人物追跡

2.3.2 項で紹介した Minimum-Cost Flow（MCF）による手法 [20] では，人物の見え特徴量や位置特徴量を用いてフレーム間の対応付けをすることで，複数人物追跡を行っている．ブレが生じた際には，それらは急激に変化し，ID スイッチを生じやすい．このような ID スイッチを防止するために，従来手法と同様，単一フレーム行動特徴量（Single-frame Action Feature ; SAF）を対応付けに使用することが考えられる．

5.2.1　単一フレーム行動特徴量（SAF）

画像から，人物検出器によって得られた矩形 $\mathbf{x}_i^{\mathrm{loc}}$ の領域を切り出し，各人物領域から SAF $\mathbf{x}_i^{\mathrm{saf}}$ を抽出する．Fig. 5.3 に SAF 抽出モデルを示す．ネットワークは 4 ストリー

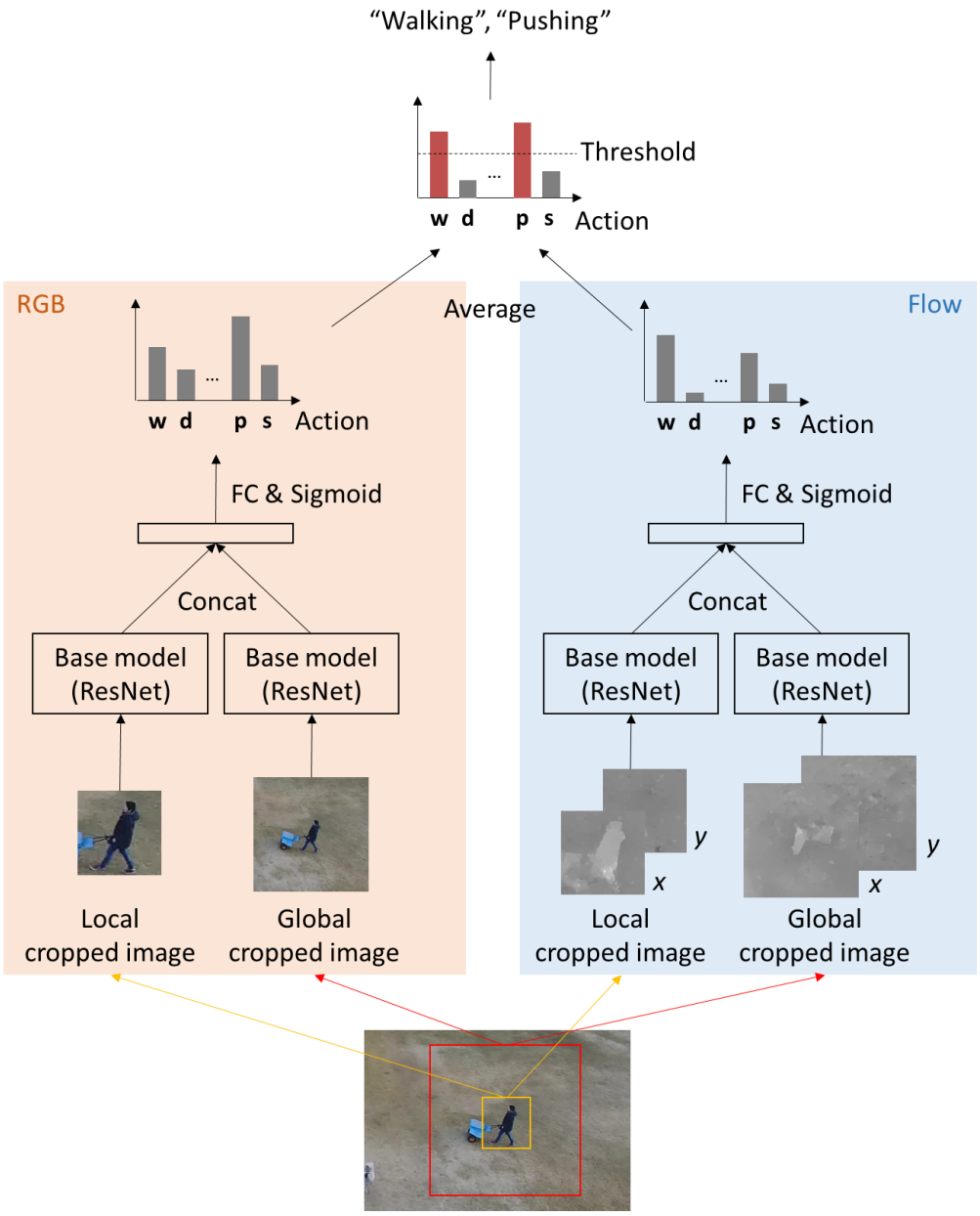


Fig. 5.3: 単一フレーム行動特徴量（SAF）の抽出モデル．

ムニューラルネットワークである．このネットワークは空間と時間の 2 つのモダリティを持ち，各モダリティはさらに 2 ストリームネットワーク [58, 59] によって 2 つのモダリティを持つ．空間のネットワークには RGB 画像を入力し，時間のネットワークにはオプティカルフロー画像を入力する．オプティカルフローの算出には高速かつ高精度な TV-L1 [121] を使用する．水平方向と垂直方向のフローはそれぞれ別々に用いる．各スト

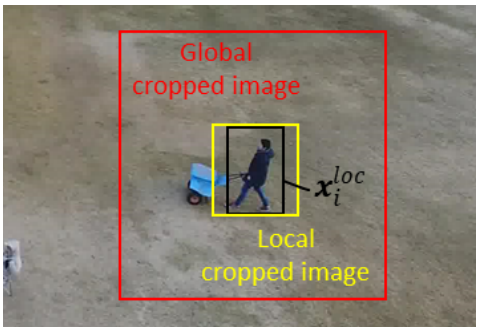


Fig. 5.4: 局所切り出し画像と大域切り出し画像.

リームのバックボーンモデルには ResNet101 [38] を用いる.

各モダリティについて，局所切り出し画像と大域切り出し画像の 2 種類の画像を入力として用いる．局所切り出し画像と大域切り出し画像の例を Fig. 5.4 に示す．まず局所切り出し画像は，$\mathbf{x}_i^{\mathrm{loc}}$ を中心として，$\mathbf{x}_i^{\mathrm{loc}}$ の長辺と同じ長さの正方形として切り出す．次に大域切り出し画像は，$\mathbf{x}_i^{\mathrm{loc}}$ を中心として，局所切り出し画像を拡大して切り出す．拡大率 μ はパラメータとして事前に設定する．大域切り出し画像によって，周りの物体や人物のような空間的なコンテキストを考慮できる．

空間と時間各々のモダリティにおいて，全結合層の後にシグモイド関数を適用したものを特徴量とする．そして，それらを平均することによって SAF $\mathbf{x}_i^{\mathrm{saf}}$ を抽出する．特徴量は確率ベクトルで表現され，Fig. 5.3 の例では，確率があらかじめ定めたしきい値を超えた行動が “Walking” と “Pushing” となっている．

ここまで述べてきたネットワークは，このように複数の行動ラベルを認識できるように学習する．これにより，ネットワークから識別的な行動特徴量が抽出できるようになる．誤差関数は，次式に示すような各クラスの 2 値交差エントロピー誤差とする．

$$E = -\sum_{n=1}^{N}(d_n \log y_n + (1-d_n)\log(1-y_n)) \tag{5.1}$$

ここで，d_n は n 番目の学習データの教師信号，y_n は n 番目の学習データの推定出力，N は学習データの総数を示す.

5.2.2　SAF を含んだ遷移モデル

MCF による手法 [20] では遷移コスト $c_{\mathrm{tran}}(i,j)$ は位置特徴量と見え特徴量を用いて算出するが，ここでは遷移コストに SAF が含まれるように変更する．具体的には，SAF 間の余弦距離を示す $c_{\mathrm{saf}}(i,j)$ を非線形関数 g の変数として追加する．

$$r = g(c_{\mathrm{iou}}(i,j), c_{\mathrm{app}}(i,j), c_{\mathrm{saf}}(i,j)) \tag{5.2}$$

これにより，人物の行動を考慮した対応付けが可能となる．

5.3　軌跡と複数フレーム行動特徴量（MAF）の交互更新

5.2 節で紹介した SAF は他のフレームの情報を考慮していないため，ドローンの急な動きによってブレが生じた場合，特に不安定である．本章では，式 (2.17) で求めた人物追跡結果 F^* を用いて MAF（$\mathbf{x}^{\mathrm{maf}}$）を抽出し，再度人物追跡に利用する．さらに，両者を数回反復して交互に更新する．この交互更新が提案手法の特徴である．

Fig. 5.5 に人物対応付けと MAF 抽出の交互実行の例を示す．この例は，1 人の人物が全フレームにわたって同じ行動 $\mathbf{a}$ をしている際，ブレが生じたという状況である．各人物の行動特徴量（SAF ／ MAF）は，行動 $\mathbf{a}$, $\mathbf{b}$ の確率分布で表される．ブレが生じた際には位置と見え特徴量が大きく変化する．以下，この例に基づいて処理手順を説明する．

- 反復 1 回目：各フレームから SAF を抽出し，人物対応付けを行う．ブレが生じた際には SAF が不安定なため（行動 $\mathbf{b}$ の確率が行動 $\mathbf{a}$ の確率よりも高い），対応付けに失敗している．
- 反復 2 回目：反復 1 回目で得られた軌跡を用いて MAF 抽出を行う．Fig. 5.6 に MAF 抽出の様子を示す．MAF 抽出はスライディングウィンドウにより行うため，フレーム $\mathbf{o}_t$ における MAF $\mathbf{x}(t)^{\mathrm{maf}}$ は次式のように抽出される．

$$\mathbf{x}(t)^{\mathrm{maf}} = \frac{1}{\lambda} \sum_{t=\lceil t-\lambda/2 \rceil}^{\lfloor t+\lambda/2 \rfloor} \mathbf{x}(t)^{\mathrm{saf}} \tag{5.3}$$

ここで，$\lceil\ \rceil$ と $\lfloor\ \rfloor$ は天井関数と床関数を示し，λ はウィンドウ長を示す．ブレによって SAF が不安定な場合でも，平均処理を行うことによって MAF は安定化する（行動 $\mathbf{a}$ の確率が行動 $\mathbf{b}$ の確率よりも高い）．その結果，対応付けが成功する

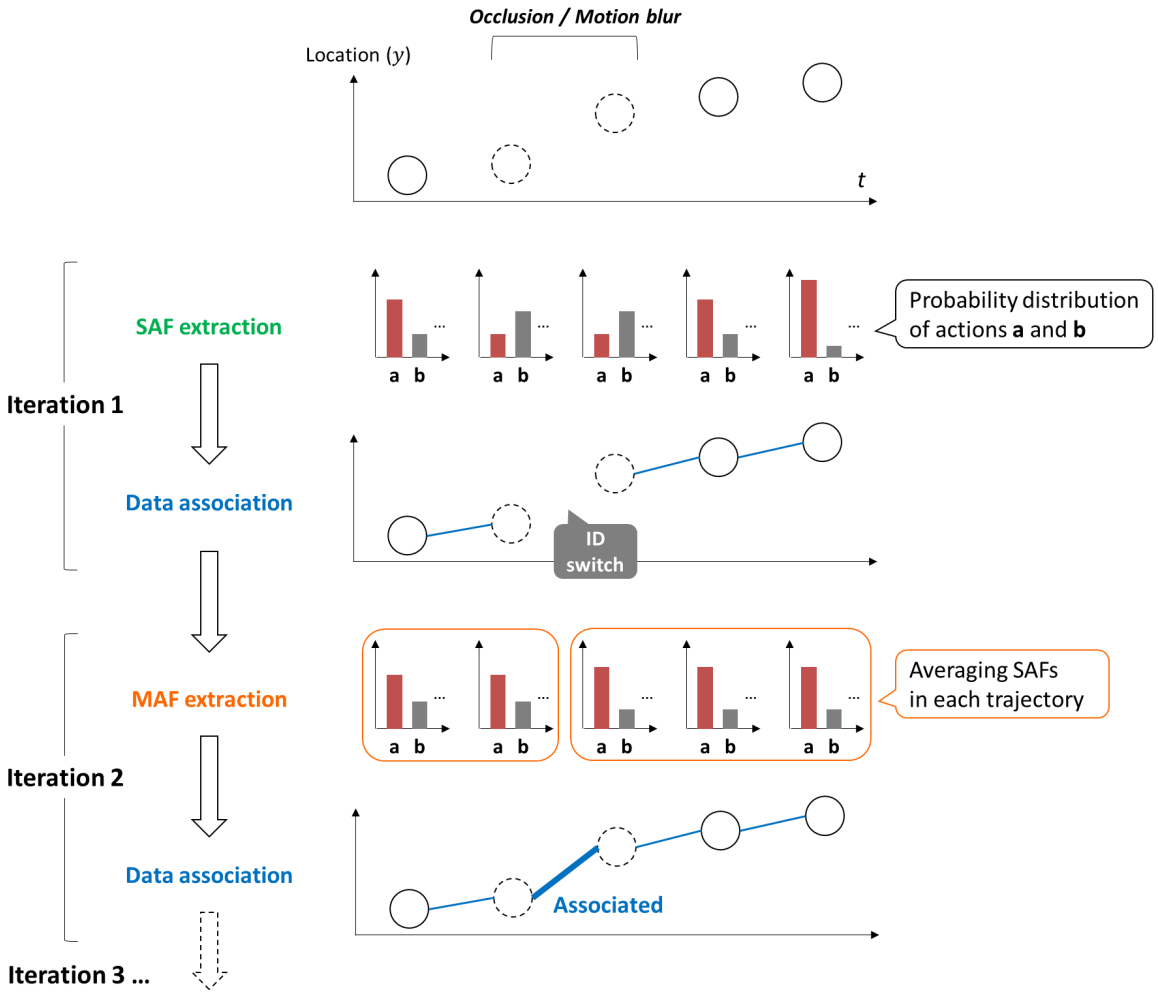


Fig. 5.5: 人物対応付けと MAF 抽出の交互更新の例．1 人の人物が全フレームにわたって同じ行動 **a** をしている際にブレが生じた状況である．点線の円は，ブレによって人物の見え特徴量が変化した様子を示す．各人物の行動特徴量（SAF ／ MAF）は，行動 **a**, **b** の確率分布で表される．所定の回数交互更新を反復し，ID スイッチを防止することができる．

（行動 **a** と行動 **a** が対応付く）．ステップ 3 以降は，ステップ 2 と同様の処理を所定の回数反復する．

5.4　実験

提案手法の有効性を検証するため，複数人物追跡実験を行った．

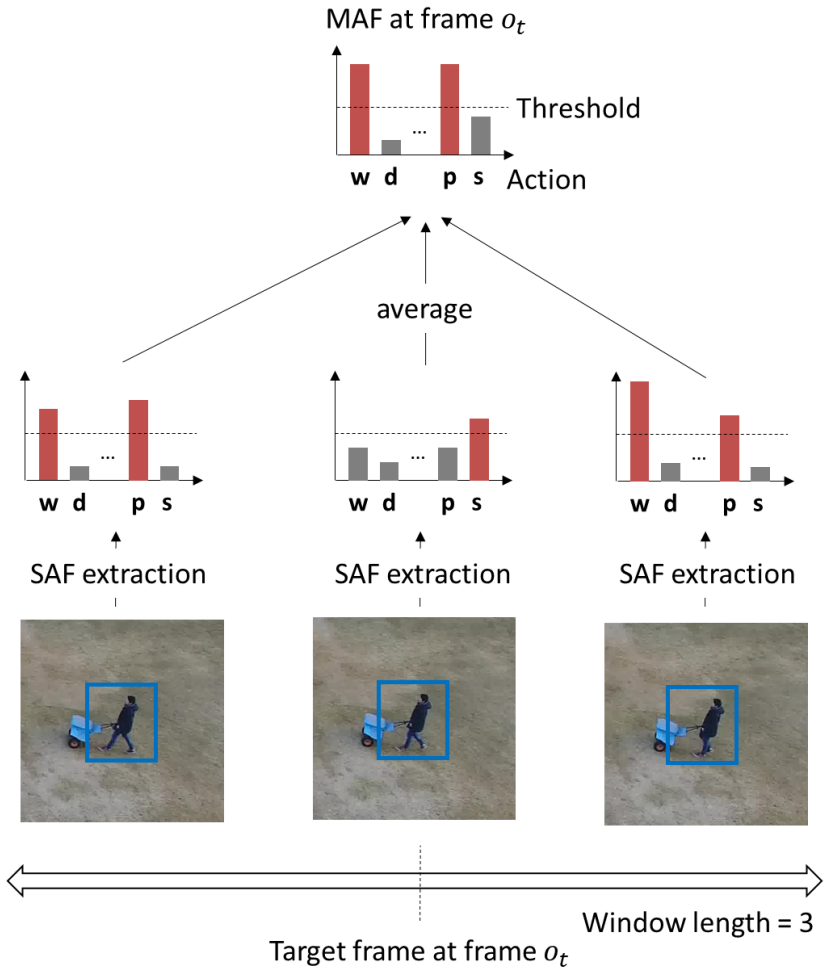


Fig. 5.6: 複数フレーム行動特徴量（MAF）抽出.

5.4.1　データセット

実験には，Okutama-Action データセット [2] と Drone-Action データセット [3] を用いた．主な評価は Okutama-Action データセットを用いて行った．Drone-Action データセットを用いた実験では，Okutama-Action データセットを用いて学習したパラメータを使用し，他のデータセットにおける提案手法の有効性を確かめた．

Okutama-Action データセットは，上空からのドローンで撮影された動画像で構成される人物行動検出用データセットである．このデータセットは，人物の大きさや縦横比の急激な変化，カメラの急激な動き，複数ラベルからなる行動の動的な遷移を含むため，非常

Table 5.1: Okutama-Action データセット [2] における行動ラベル.

人物-人物インタラクション	Handshaking
	Hugging
人物-物体インタラクション	Reading
	Drinking
	Pushing/Pulling
	Carrying
	Calling
インタラクションなし	Running
	Walking
	Lying
	Sitting
	Standing

Table 5.2: Drone-Action データセット [3] における行動ラベル.

Clapping	Kicking
Hitting with bottle	Waving hands
Hitting with stick	Walking (front/back/side view)
Stabbing	Jogging (front/back/side view)
Punching	Running (front/back/side view)

に難易度が高い．データセットは 43 本の動画像で構成され，33 本の学習データと 10 本の評価データに分かれている．動画像は 4K（$3,840 \times 2,160$ 画素）の解像度で 30 fps で撮影されており，データセット内の全画像数は $77,365$ 枚である．各画像には 0–9 人の人物が写っている．2 つのドローンで，高度 10–45 メートルから，水平線を 0 度として地上向きに 45–90 度のカメラ角度で撮影されている．各人物矩形には 1 つ以上の行動ラベルが付与されている．行動ラベルは Table 5.1 に示す 12 種類である．複数の行動ラベルが付与されている場合，1 つの行動はインタラクションなしのカテゴリに含まれるもので，それ以外の行動は人物-人物インタラクションか人物-物体インタラクションのカテゴリのもので構成される．

一方，Drone-Action データセット [3] は行動認識用のデータセットで，低高度で低速

飛行のドローンから撮影した動画像で構成されている．データセットは 80 本の動画像で構成されている．動画像は，HD（$1,920 \times 1,080$ 画素）の解像度で 25 fps で撮影されており，データセット内の全画像数は $9,813$ 枚である．各画像には 1 人の人物が写っている．各人物矩形には 1 つの行動ラベルが付与されている．行動ラベルは Table 5.2 に示す 10 種類である．

5.4.2　実験条件

人物検出モデル（Single Shot multibox Detector ; SSD）を Okutama-Action データセットを用いて学習した．この際のバッチサイズは 8，繰り返し回数は $6,000$，学習率は 10^{-4} とした．SSD の入力サイズは 512×512 とし，バックボーンモデルには VGG16 [37] を用いた．人物検出結果は，従来手法と提案手法で同じものを使用した．見え特徴量の抽出モデル（WideResNet [122]）は MARS データセット [123] を用いて学習した．また，SAF 抽出モデルも Okutama-Action データセットを用いて学習した．この際のバッチサイズは 16，繰り返し回数は $5,000$，学習率は 10^{-4} とした．また，ドロップアウトの割合は 0.7 とした．さらに，データ拡張として無作為の切り出しと水平／垂直切り出しを行った．大域切り出し画像の拡大率を $\mu = 3$，MAF 抽出の際のスライディングウィンドウ長を $\lambda = 15$ と実験的に設定した．さらに，観測コストモデルと遷移コストモデルも Okutama-Action データセットを用いて学習した．人物対応付けのパラメータは，実験的に $c_{\mathrm{entr}}(i) = 10$, $c_{\mathrm{exit}}(i) = 10$, $b = -0.5$ と設定した．

5.4.3　人物追跡の評価

まず，Okutama-Action データセットを用いて人物追跡の評価を行った．推定した矩形の正誤判定の際の基準として，正解の矩形と推定した矩形との間の IoU のしきい値を 0.5 とした．評価指標には，再現率（Recall），適合率（Precision），ID スイッチの回数（IDs），Multiple Object Tracking Accuracy（MOTA），Multiple Object Tracking Precision（MOTP）を用いた [120]．MOTA は未検出，ID スイッチ，過検出の 3 つの指標を組み合わせた指標で，複数人物追跡分野において最も広く用いられている．

比較手法として，オンライン手法で，位置特徴量と見え特徴量を用いた DeepSORT [50] を用意した．また，オフライン手法として，同じく位置特徴量と見え特徴量を用いた MCF

Table 5.3: Okutama-Action データセット [2] での人物追跡性能.

		Recall [%] ↑	Precision [%] ↑	IDs ↓	MOTA ↑	MOTP ↑
Online	DeepSORT [50]	29.10	70.81	584	17.29	32.85
Offline	MCF [20]	32.42	73.93	597	21.43	32.99
	MHT-SAF	32.13	74.10	558	21.20	**33.01**
	MHT-MAF	**32.48**	**74.24**	**528**	**21.57**	**33.01**

による手法 [20] を用意した．提案手法については，位置特徴量・見え特徴量・SAF を用いるが交互更新はしない MHT-SAF（反復回数 : 1）と，交互更新をする MHT-MAF（反復回数 : 6）を用意した．

Table 5.3 に各手法に対する評価結果を示す．MCF による手法と比較して，提案手法（MHT-MAF）の再現率及び適合率はほぼ同等である．ID スイッチの回数は 69 減少し，MOTA は 0.14 ポイント向上した．なお，MOTP は ID スイッチの回数を考慮しないため，ほぼ同等となった．

Fig. 5.7 に，ブレによって ID スイッチが生じたものの，最終的には MHT-MAF によって防止できた例を示す．MCF による手法では，位置特徴量と見え特徴量が不安定なため，ID スイッチがフレーム 1,584 で生じている．MHT-SAF（反復 1 回目）では，“carrying” と “walking” によってフレーム 1,583, 1,584, 1,585 が対応付けられているが，フレーム 1,582, 1,583 は対応付けられていない．MHT-MAF（反復 2 回目）では，フレーム 1,582, 1,583 は MAF 抽出によって “carrying” と推定されており，フレーム 1,582, 1,583, 1,584, 1,585 が対応付けられている．そして，MHT-MAF（反復 6 回目）では，MAF 抽出を反復することによって，フレーム 1,582, 1,583, 1,584, 1,585 で安定して “carrying” と推定されている．

次に，人物追跡と MAF 抽出の交互更新の評価を行った．Fig. 5.8(a) に各反復回における再現率と適合率を示す．反復 1 回目では，再現率が約 74%，適合率が約 32% であった．その後，両者はほぼ変化しなかった．再現率と適合率は人物検出精度に大きく依存するため，交互更新の効果はほぼなかったと推測される．Fig. 5.8(b) に各反復回における ID スイッチの回数と MOTA を示す．反復 1 回目から 6 回目の間に ID スイッチ回数が 30 回減少した．しかし，反復 10 回目ではむしろ増加した．MOTA も ID スイッチ数と同様の

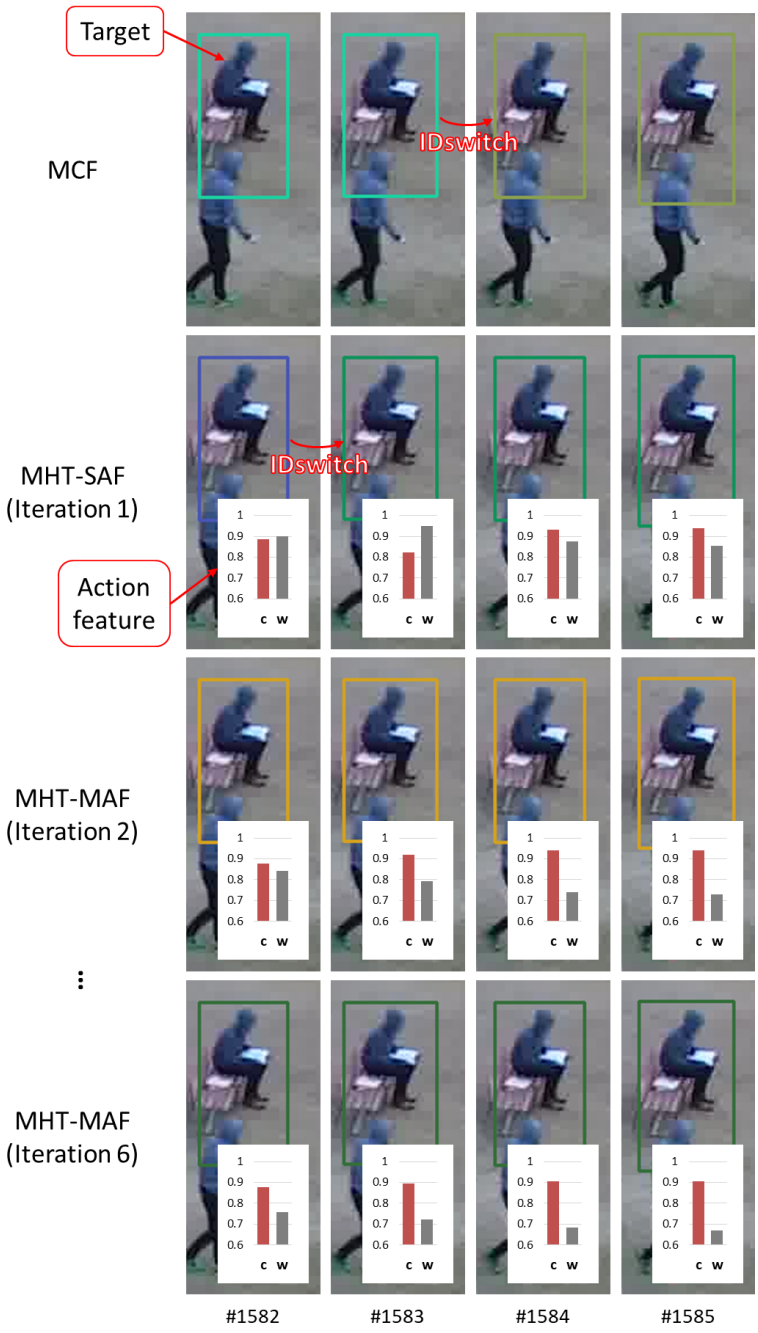
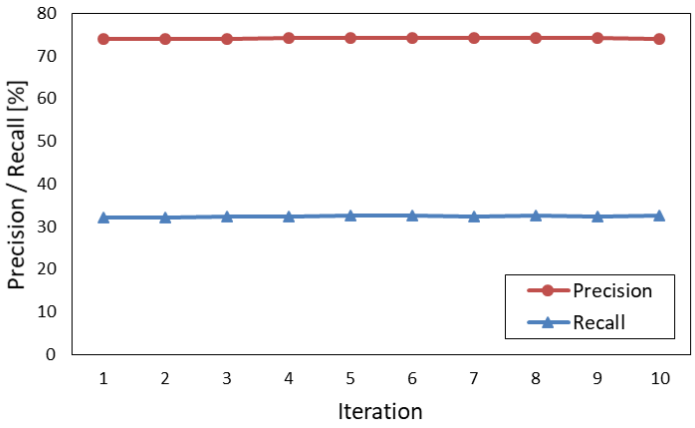
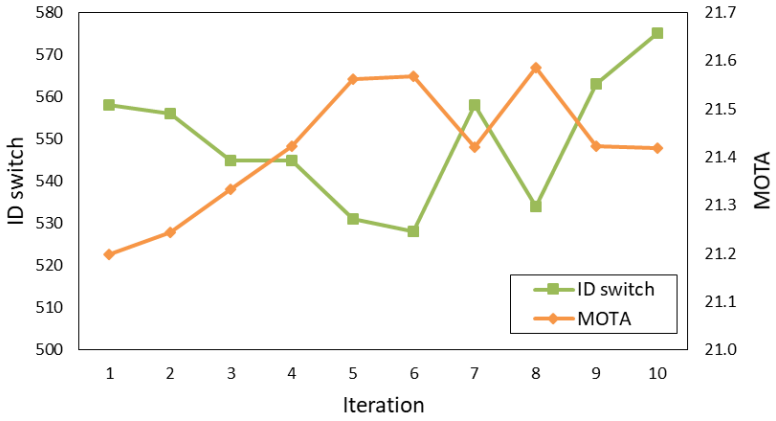


Fig. 5.7: ブレによって ID スイッチが生じたが，最終的には MHT-MAF によって防げた例．この例は動画像名 “1.2.10” の 4 フレーム分である．#(number) はフレーム番号を示す．上部の人物が正解の追跡対象である．推定された矩形各々に対して行動特徴量が記載されており，“c” は “carrying” を，“w” は “walking” を示す．なお，MCF では行動特徴量を用いていないため記載していない．



(a) 再現率（Recall）と適合率（Precision）.



(b) ID スイッチ回数と MOTA.

Fig. 5.8: 各反復回における人物追跡評価指標の値.

傾向を示した．この結果から，適切な反復回数はおおよそ 5, 6 回であることが分かった．

反復 10 回目に ID スイッチが頻繁に生じる原因を分析した．動画像 “1.2.10” から，Fig. 5.9 に示すような 2 人の人物が立っている 5 フレーム（フレーム 1,002〜1,006）の区間を抜き出した．この区間以前のフレームでは，2 人の人物は各フレームにおいて 2 つの軌跡（ID : 0 と ID : 23）として追跡できていた．しかし，抜き出した区間では 2 人の人物間で互いに遮蔽が生じており，両者を含む 1 つの矩形として推定してしまっている．そのため，2 つの軌跡が不安定に推定されている．精度評価においては，これらの 5 つの推定矩形に対応した正解矩形は，全て左の人物であった．上記が ID スイッチが頻繁に生じた主な原因であった．このように遮蔽が生じた際の未検出を防ぐことは，今後の研究にお

Fig. 5.9: 2 人の人物が交互に 1 つの矩形として検出され，ID スイッチが頻発した例．

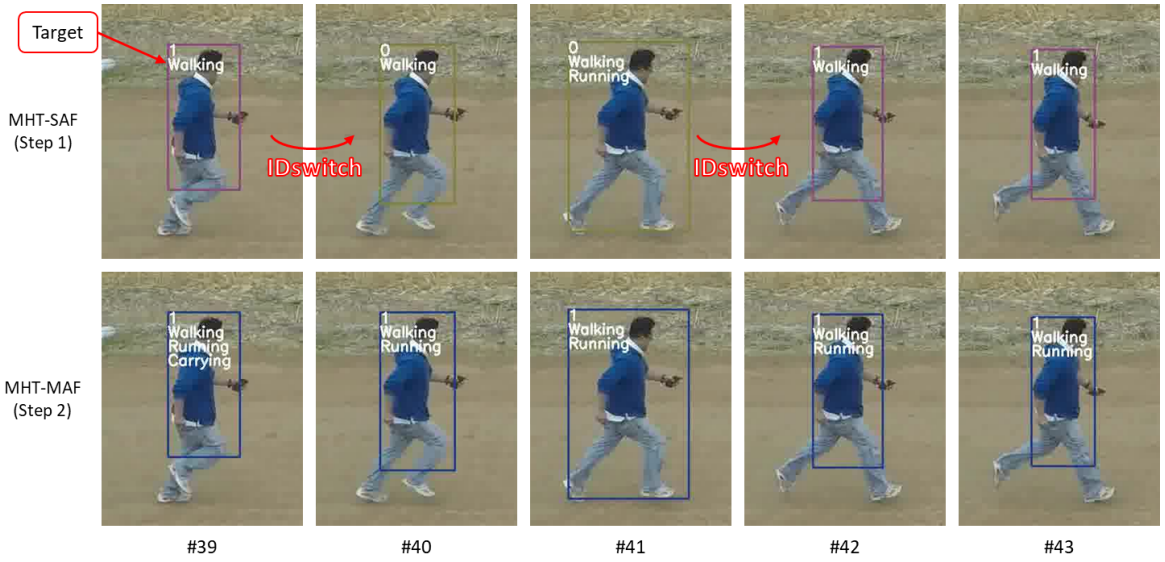


Fig. 5.10: Drone-Action データセット [3] での人物追跡結果の例．

ける重要な課題である．

最後に，Drone-Action データセットを用いて人物追跡の評価を行った．提案手法のパラメータは，5.4.2 項で示したような Okutama-Action データセットを用いて学習したパラメータを流用した．Fig. 5.10 に追跡結果の例を示す．例として，典型的な行動を含む “S8_running_toRight_sideView_HD” の動画像を選択した．MHT-SAF（反復 1 回目）ではフレーム 40 と 42 で ID スイッチが生じたが，MHT-MAF（反復 2 回目）ではこれらの ID スイッチが防止された．“Running” は 5 つ全てのフレームで一貫して推定されており，これによって正しい人物対応付けができたと考えられる．Table 5.4 に Drone-Action データセットを用いた人物追跡の性能を示す．MHT-SAF と比較して，MHT-MAF の再現率及び適合率はほぼ同等である．一方，ID スイッチの回数は 4 回減少し，MOTA は 2.58 ポイント向上した．以上より，対象とするデータセットを用いてパラメータを学習し

Table 5.4: Drone-Action データセット [3] での人物追跡性能.

	Recall [%] ↑	Precision [%] ↑	IDs ↓	MOTA ↑	MOTP ↑
MHT-SAF	89.68	83.23	47	41.29	33.98
MHT-MAF	89.68	83.23	**43**	**43.87**	33.98

なくても，提案手法が有効であることが確認できた.

5.4.4 行動認識の評価

次に，MAF 抽出（複数フレーム行動認識）と大域切り出し画像の評価を行うために，人物追跡結果には正解を与えて行動認識のみの評価を行った．SAF ／ MAF においてしきい値 $\epsilon = 0.4$ を超えた行動を 1，それ以外を 0 とし，フレーム単位で評価した．SAF は単一フレームによって，MAF はスライディングウィンドウ（$\lambda = 15$ フレーム）によって特徴抽出し，行動認識精度を算出した．また，SAF ／ MAF それぞれについて，局所切り出し画像のみ，あるいはさらに大域切り出し画像も入力した場合に分けて，行動認識精度を算出した．Table 5.5 に行動認識精度を示す.

複数フレーム行動認識

SAF と MAF を比較する．局所切り出し画像では，単一フレーム行動認識よりも複数フレーム行動認識の方が精度が高い．また，局所 + 大域切り出し画像でも，単一フレーム行動認識よりも複数フレーム行動認識の方が精度が高い．したがって，複数フレーム行動認識，つまり MAF は有効であることが示された.

Table 5.5: 行動認識精度 [%].

	Human to human interactions		Human to object interactions				
	Handshaking	Hugging	Reading	Drinking	Pushing/Pulling	Carrying	Calling
Single frame (local)	7.78	21.47	57.26	0.00	55.95	53.57	15.08
Single frame (local+global)	17.11	24.60	61.27	0.00	64.31	74.13	**17.82**
Multi frames (local)	9.64	17.68	56.75	0.00	60.88	56.94	13.97
Multi frames (local+global)	**18.07**	**24.81**	**61.37**	0.00	**68.28**	**78.20**	16.76

	No-interaction					Average
	Running	Walking	Lying	Sitting	Standing	
	43.63	79.97	24.19	76.31	79.37	42.88
	41.18	85.17	14.33	75.41	75.89	45.94
	46.72	87.19	**26.56**	**81.94**	**82.83**	45.09
	43.55	**90.73**	15.40	78.04	78.40	**47.80**

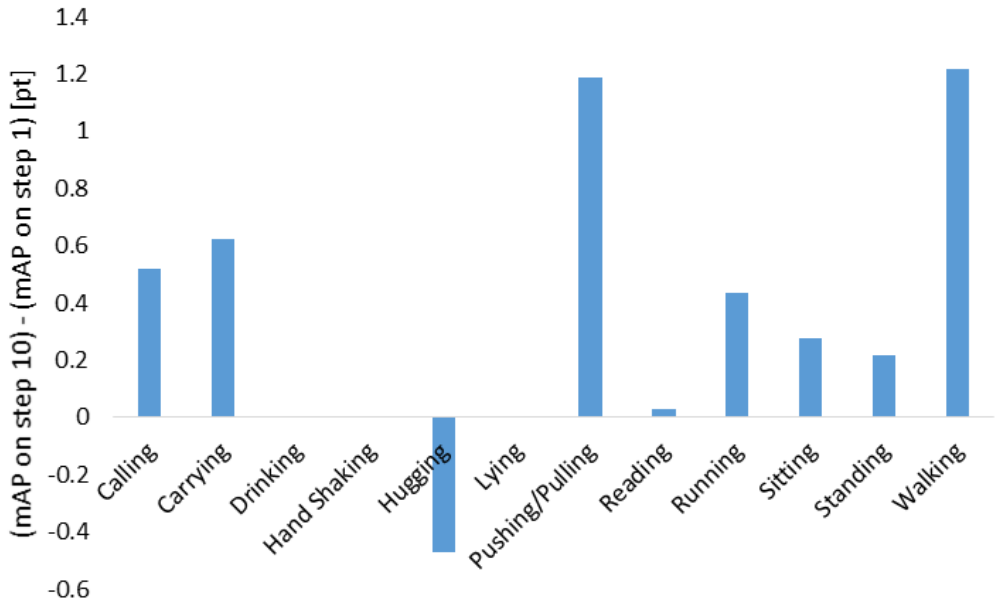


Fig. 5.11: ステップ数 10 のときとステップ数 1 のときの mAP の差.

大域切り出し画像

SAF（単一フレーム行動認識）において，局所切り出し画像と局所 + 大域切り出し画像を比較する．局所 + 大域切り出し画像の場合の精度は，局所切り出し画像の場合よりも高い．人物-人物インタラクションと人物-物体インタラクションの行動では，大域切り出し画像も含めた方が有効である．このような行動の認識には，周りの人物や物体のような大域的なコンテキストが必要となるからであると考えられる．一方，インタラクションなしの行動では，局所切り出し画像の方が有効である．このような行動の認識には対象の人物のみしか必要ないからであると考えられる．

5.4.5 最適な反復回数に関する議論

反復回数は，認識したい行動の種類に基づいて決定するのが良い可能性がある．5.4.4 項で示した通り，行動認識精度は複数のフレームを用いることで向上させることができる．よって，反復回数が増えるにつれて行動認識精度が向上した場合，人物追跡精度も向上すると言える．そのため，評価指標には行動認識に関する mean Average Precision（mAP）を用い，反復回数によるその変化を分析した．Fig. 5.11 に，各行動について，反復

10 回目の mAP と反復 1 回目の mAP の変化を示す．“Pushing/Pulling” と “Walking” の mAP は約 1.2 ポイント増加した．Table 5.5 に示した通り，これら 2 つの行動は単一フレームでも高い認識精度を示している．しかし，“Hugging” の mAP は約 0.5 ポイント減少した．Fig. 5.9 に示した通り，複数の人物が密接している場合は，ID スイッチが頻発する．従って，単一フレームで認識しやすい行動を対象とする場合は反復回数を増やし，複数の人物が密接している場合は反復回数を減らすのが有効だと考えられる．

5.5　まとめ

本章では，広域中に複数人物が存在する環境において，人物軌跡と複数フレーム行動特徴量（Multi-frame Action Feature ; MAF）を交互に更新する人物追跡手法（Multiple Human Tracking using MAF ; MHT-MAF）を提案した．この手法により，ドローンの急な動きによってブレが生じた場合でも，安定化した MAF によって ID スイッチを防止することができる．空撮画像で構成された Okutama-Action データセット [2] 及び Drone-Action データセット [3] を用いた評価実験では，提案手法により交互更新の反復を繰り返すことで，ID スイッチが防止できることを確認した．追跡精度については，再現率と適合率をほぼ同等に保ちながら，ID スイッチ数が 69 回削減され，MOTA は 0.14 ポイント向上することを確認した．また，MAF 抽出と大域画像切り出しの有効性を確認した．今後は，遮蔽が生じた際の人物の未検出を防ぐ手法を検討する必要がある．

本章においては，複数人物追跡に焦点を当て，より多くの人物を追跡するため，移動して広域を撮影できるドローン搭載カメラを用いた手法を検討した．その際に発生する人物 ID 推定の誤りに対して解決を目指した．この問題の解決に貢献したのは，人物軌跡と複数フレーム行動特徴量（MAF）の交互更新であった．次章では，本論文でこれまでに述べた内容をまとめる．

第 6 章

むすび

本章では，本論文の内容をまとめた上で，今後の課題と展望について述べる．

6.1　本論文のまとめ

人々が社会で快適でかつ安心な暮らしを実現するために，これまでにマーケティングや監視のような様々な分野で研究が推進されてきた．そのような分野においては，自動的に人物の行動を認識することが求められており，その際に最も重要な役割を果たすのが人物追跡技術である．人物追跡の要素である人物位置推定と人物 ID 推定の誤りに対して，単一人物追跡と複数人物追跡に分けて解決を目指した．単一人物追跡では，パーティクルフィルタの枠組みを用いた確率的な相関フィルタの更新により，人物位置推定のずれを解決した．複数人物追跡では，可能な限り多くの人物を追跡できるように，広域を撮影できる全方位カメラ及びドローン搭載カメラを用いた手法を検討した．局所的パノラマ展開と世界座標表現により，全方位カメラの見えの歪みによって生じる人物 ID 推定の誤りを解決した．また，人物軌跡と複数フレーム行動特徴量（MAF）の交互更新により，ドローン搭載カメラのブレによって生じる人物 ID 推定の誤りを解決した．

第 2 章では人物追跡に関する関連研究をまとめた．追跡手法は，照合による手法，類似度勾配による手法，時系列フィルタによる手法，検出による手法の順に変遷してきた．まず，単一人物追跡において，検出による手法は遮蔽や変形のような見えの変化に弱いという問題を指摘した．次に，複数人物追跡において，広域を撮影できる全方位カメラ及びドローン搭載カメラを用いる際の問題について各々指摘した．

第 3 章では見えの変化に頑健な単一人物追跡手法を提案した．従来の検出による追跡手

法は，追跡対象とそれ以外の物体を識別する能力は高い．しかし，遮蔽や変形のような見えの変化に対して不安定なため，人物位置推定にずれが生じやすい．そこで，検出による追跡の信頼度が低下した際に，見えの変化に頑健な時系列フィルタと組み合わせる手法を提案した．提案手法では，時系列フィルタであるパーティクルフィルタの観測モデルとして，検出器である相関フィルタによって得られた応答マップを用いる．また，複数の相関フィルタを用意し，パーティクルフィルタの状態変数に追加することで，最適な相関フィルタを選択しながら追跡する．実験では，TB-50 データセット [1] を用いて，提案手法は物体の遮蔽・変形・回転のような見えの変化に対して頑健であることを確認した．追跡精度については，AUC スコアが 2.26 ポイント向上することを確認した．

第 4 章では全方位カメラを用いた見えの歪みに頑健な複数人物追跡手法を提案した．複数人物追跡では，可能な限り多くの人物を追跡するために広域を撮影できることが望ましい．その手段として 360 度の画角を持つ全方位カメラを用いた．しかし，全方位カメラを用いた場合，レンズの歪みによって，人物の見えや位置がフレーム間で非線形に変化してしまう．これに対して通常カメラ向けに設計された従来手法を単純に適用すると，フレーム間の人物対応付けが失敗（ID スイッチが発生）しやすくなる．この問題を解決するため，人物の 3 次元モデルを用いた追跡手法を提案した．提案手法では，1）人物領域のみを局所的に展開してから特徴抽出する，2）距離指標が均一な世界座標系で人物位置を表現する．これらによって，人物の見えや位置の非線形な変化を防ぎ，人物対応付け精度を向上させることができる．実験では，独自に作成した LargeRoom・SmallRoom データセットを用いて，局所的展開と世界座標表現がともに有効であることを確認した．追跡精度については，Multiple Object Tracking Accuracy（MOTA）が，LargeRoom データセットで 3.3 ポイント，SmallRoom データセットで 2.3 ポイント向上することを確認した．

第 5 章ではドローン搭載カメラを用いたブレに頑健な複数人物追跡手法を提案した．広域を撮影する別の手段として，移動可能なドローン搭載カメラを用いた．ドローン搭載カメラを用いた場合，ドローンの急な動きによって，人物の見えや位置がフレーム間で急激に変化してしまう．これに対して固定カメラを対象とした従来手法を適用すると，フレーム間の人物対応付けが失敗（ID スイッチが発生）しやすい．この問題に対応するために，人物の行動に関する特徴量を用いた追跡手法を提案した．提案手法では，推定した人物軌跡に基づいて行動特徴量を更新し，再度人物追跡に用いる．実験では，Okutama-Action

データセット [2] と Drone-Action データセット [3] を用いて，提案手法により交互更新の反復を繰り返すことで，ID スイッチが防止できることを確認した．追跡精度については，MOTA が，Okutama-Action データセットで 0.14 ポイント，Drone-Action データセットで 2.58 ポイント向上することを確認した．

以上のように，本論文では人物位置推定と人物 ID 推定の誤りの解決を目指した．単一人物追跡では，パーティクルフィルタの枠組みを用いた確率的な相関フィルタの更新により，人物位置推定のずれを解決した．複数人物追跡では，局所的パノラマ展開と世界座標表現により，全方位カメラの見えの歪みによって生じる人物 ID 推定の誤りを解決し，人物軌跡と複数フレーム行動特徴量（MAF）の交互更新により，ドローン搭載カメラのブレによって生じる人物 ID 推定の誤りを解決した．

6.2 今後の課題と展望

本論文で提案した手法を用いても，人物位置推定精度及び人物 ID 推定精度は 100% には及ばない．機械学習の各段階において，いかにセンサ情報を得るか，いかに正解情報を得るか，そしていかに学習をするかが重要であり，今後の課題である．以下に各々に分けて課題の詳細を述べる．

センサ情報

本論文では 1 台のカメラのみを用いた手法を検討したが，設置・運用コストに問題がなければ，複数のカメラを用いた方が追跡をしやすくなる場合があると考えられる．例えば第 3 章で扱った遮蔽の問題は，被写体の反対側から撮影できれば簡単に解決できる可能性がある．ただし，複数カメラを使用する場合は，カメラ間での人物の対応付けが必要となる．カメラ間の対応付けとフレーム間の対応付けに矛盾がないように最適化問題を解くことが重要な課題となると考えられる．

また，本論文ではカメラのみを使用した手法を検討したが，赤外線センサ・LiDAR センサ，あるいは GPS・WiFi を利用できる状況であれば，それらの手段で推定した人物位置も利用することができる．その際には，信頼できる人物位置推定手段の選定が重要になる．

正解情報

本論文では，単一人物追跡においては初期フレームのみ人物の正解位置を与え，複数人物追跡においては正解位置は全く与えていない．しかし，実応用を考えると，監視者とのインタラクションによって，いくつかのフレームで正解位置を与えられる可能性がある．例えば，追跡する様子をタブレット端末で映しておき，余裕があるときに人物の位置をタップするようなインタラクションが考えられる．このようにできれば，与えられた正解位置を制約にして最適化問題を解きやすくなったり，半教師あり学習の形で追跡器（検出器）のパラメータを更新できたりする可能性がある．

学習方法

本論文では，単一人物追跡と複数人物追跡に分けて取り組んだ．従来研究においても両者を分けた取り組みがほとんどであるが，両者を効果的に融合できると良い．例えば，基本は複数人物追跡の枠組みで動かしながらも，単一人物追跡手法のように検出器のパラメータを更新する機能を加えるのが有効だと思われる．その際には，更新のタイミングが重要になると考えられる．

一方，監視カメラによる人物追跡・行動認識技術の普及に伴い，プライバシーを侵害する監視社会への懸念も高まっている．プライバシーを保護するためには，利用目的，撮影範囲，取得・加工データの内容，第三者提供の有無等を撮影時に公表することが重要になる．また，個人情報の取り扱いに関するルールを定めた個人情報保護法を遵守し，高い倫理観を持って技術を利用することが必要になる．

最後に，本研究の成果が人物追跡に関する研究の発展に貢献するとともに，マーケティングや監視を初めとした様々な分野で応用されることを願う．それが，人々が快適でかつ安心して暮らせる社会を実現するための一助になればと考えている．

謝辞

初めに，本論文の主査である，名古屋大学大学院情報学研究科 村瀬洋教授に深く感謝致します．社会人博士として研究室に受け入れて下さり，研究活動全般において多大なるご支援を賜りました．先生からご指導頂けることを大変名誉に感じながら，日々研究を進めて参りました．

本論文の副査である，名古屋大学大学院情報学研究科 井手一郎教授，同間瀬健二教授，同出口大輔准教授に感謝致します．ご多忙にもかかわらず，大きな点から詳細に至るまで何度も本論文を確認して下さり，貴重なご助言を頂きました．

日々多大なるご指導を賜りました，名古屋大学大学院情報学研究科 川西康友講師に感謝致します．いつもたたき台の原稿に対して大きく質を高めるご指導をして下さり，頭が上がりません．また，自由な発想で楽しく研究することの大切さを学ばせていただきました．

業務を行いながら本研究を行う機会を与えて下さいました，株式会社 KDDI 総合研究所会長 中島康之博士，同所長 中村元博士，同総務部門長 柳原広昌博士，同メディア ICT 部門長 内藤整博士，同フロンティア研究室長 田坂和之博士，同メディア認識グループリーダー 小森田賢史氏に心より感謝致します

入社時よりご指導頂きました，大阪工業大学情報科学部 酒澤茂之教授，KDDI 株式会社課長補佐 巻渕直哉氏，同課長補佐 永井有希氏，株式会社 KDDI 総合研究所マネージャー 小林達也氏に感謝致します．また，日々の研究生活を支えて下さったメディア認識グループの皆様に感謝致します．

修士課程在籍中にご指導賜り，その後研究職を志すきっかけを作って下さった，神戸大学大学院システム情報学研究科 有木康雄名誉教授，武庫川女子大学生活環境学部 榎並直子講師に感謝致します．

最後に，今日に至るまで生活を支えて下さった，父，母，義父，義母に感謝致します．

また，妻でもあり，研究者でもある，東京電機大学システムデザイン工学部 小篠裕子助教に心より感謝致します．神戸大学在学時から多大なご支援を頂きました．そして，いつも元気を分けてくれる二人の息子達に感謝致します．

参考文献

[1] Y. Wu, J. Lim, and M.-H. Yang, “Object tracking benchmark,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.37, no.9, pp.1834–1848, 2015.

[2] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, “Okutama-Action: An aerial view video dataset for concurrent human action detection,” Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp.28–35, 2017.

[3] A.G. Perera, Y.W. Law, and J. Chahl, “Drone-Action: An outdoor recorded drone video dataset for action recognition,” Drones, vol.3, no.4, pp.82_1–82_16, 2019.

[4] 矢野経済研究所，2020 年度版監視カメラ市場予測と次世代戦略—ビジュアル・コミュニケーション調査シリーズ—，矢野経済研究所，2020.

[5] C. Fornell and B. Wernerfelt, “Defensive marketing strategy by customer complaint management: A theoretical analysis,” Journal of Marketing Research, vol.24, no.4, pp.337–346, 1987.

[6] S.I. Rick, B. Pereira, and K.A. Burson, “The benefits of retail therapy: Making purchase decisions reduces residual sadness,” Journal of Consumer Psychology, vol.24, no.3, pp.373–380, 2014.

[7] 阿部幸司，“空の監視サービス，”電子情報通信学会通信ソサイエティマガジン B-Plus，vol.10，no.3，pp.150–155，2016.

[8] 宝木和夫，“［特集］オリンピックのための情報処理：3. オリンピックのセキュリ

ティ，” 情報処理，vol.55，no.11，pp.1196–1203，2014.

[9] 新美潤一郎，星野崇宏，“顧客行動の多様性変数を利用した購買行動の予測，” 人工知能学会論文誌，vol.32，no.2，pp.B-G63_1–9，2017.

[10] S. Van Spek, J. Van Schaick, P. De Bois, and R. De Haan, “Sensing human activity: GPS tracking,” Sensors, vol.9, no.4, pp.3033–3055, 2009.

[11] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann, “Tracking human mobility using WiFi signals,” PLOS ONE, vol.10, no.7, p.e0130824, 2015.

[12] R.T. Collins, A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson, “A system for video surveillance and monitoring: VSAM final report,” Technical Report of Carnegie Mellon University, no.CMU-RI-TR-00-12, 2000.

[13] 株式会社日立産業制御ソリューションズ，“動線分析ソリューション，” https://info.hitachi-ics.co.jp/product/pss/solution/flow_line.html，2020.

[14] セコム株式会社，“［報道資料］国内初，5G を活用したスタジアム警備の実証実験に成功，” https://www.secom.co.jp/corporate/release/2019/nr_20190819.html，2019.

[15] J.F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting the circulant structure of tracking-by-detection with kernels,” Proceedings of the 12th European Conference on Computer Vision (ECCV) Part 4, Lecture Notes in Computer Science, vol.7575, pp.702–715, 2012.

[16] 村瀬洋，V.V. Vinod，“局所色情報を用いた高速物体探索—アクティブ探索法—，” 電子情報通信学会論文誌（D），vol.J81-D，no.9，pp.2035–2042，1998.

[17] D. Comaniciu, V. Ramesh, and P. Meer, “Real-time tracking of non-rigid objects using mean shift,” Proceedings of the 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp.142–149, 2000.

[18] A.A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, “Multi-object tracking through simultaneous long occlusions and split-merge conditions,” Pro-

ceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp.666–673, 2006.

[19] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp.1200–1207, 2009.

[20] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2008.

[21] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI), pp.121–130, 1981.

[22] B.K. Horn and B.G. Schunck, "Determining optical flow," Techniques and Applications of Image Understanding, vol.281, pp.319–331, 1981.

[23] N. Paragios and R. Deriche, "Geodesic active regions for motion estimation and tracking," Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV), pp.688–694, 1999.

[24] S. Avidan, "Support vector tracking," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.26, no.8, pp.1064–1072, 2004.

[25] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," Proceedings of the 20th International Conference on Pattern Recognition (ICPR), pp.2756–2759, 2010.

[26] D.P. Huttenlocher, J.J. Noh, and W.J. Rucklidge, "Tracking non-rigid objects in complex scenes," Proceedings of the 4th IEEE International Conference on Computer Vision (ICCV), pp.93–101, 1993.

[27] D.M. Gavrila and L. Davis, "3-D model-based tracking of human upper body

movement: A multi-view approach," Proceedings of the 1995 International Symposium on Computer Vision (ISCV), pp.253–258, 1995.

[28] M. Isard and A. Blake, "CONDENSATION: Conditional density propagation for visual tracking," International Journal of Computer Vision (IJCV), vol.29, no.1, pp.5–28, 1998.

[29] D.G. Lowe, "Object recognition from local scale-invariant features," Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV), pp.1150–1157, 1999.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp.886–893, 2005.

[31] J.F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.37, no.3, pp.583–596, 2014.

[32] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1815–1821, 2012.

[33] Z. Qin and C.R. Shelton, "Improving multi-target tracking via social grouping," Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1972–1978, 2012.

[34] H. Pirsiavash, D. Ramanan, and C.C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1201–1208, 2011.

[35] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," Proceedings of the 12th European Conference on Computer Vision (ECCV) Part 4, Lecture Notes in Computer Science, vol.7575,

pp.215–230, 2012.

[36] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems (NIPS), vol.25, pp.1097–1105, 2012.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Computing Research Repository arXiv Preprint arXiv:1409.1556, 2014.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770–778, 2016.

[39] E. Ahmed, M. Jones, and T.K. Marks, "An improved deep learning architecture for person re-identification," Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3908–3916, 2015.

[40] R.R. Varior, M. Haloi, and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identification," Proceedings of the 14th European Conference on Computer Vision (ECCV) Part 8, Lecture Notes in Computer Science, vol.9912, pp.791–808, 2016.

[41] R.R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A Siamese long short-term memory architecture for human re-identification," Proceedings of the 14th European Conference on Computer Vision (ECCV) Part 7, Lecture Notes in Computer Science, vol.9911, pp.135–153, 2016.

[42] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp.1735–1742, 2006.

[43] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," Proceedings of the 2015 International Workshop on Similarity-Based Pattern Recognition, pp.84–92, 2015.

[44] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," Advances in Neural Information Processing Systems (NIPS), vol.26, pp.809–817, 2013.

[45] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," Proceedings of the 17th IEEE International Conference on Computer Vision (ICCV), pp.3119–3127, 2015.

[46] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4293–4302, 2016.

[47] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, and P.H. Torr, "Fully-convolutional Siamese networks for object tracking," Proceedings of the 14th European Conference on Computer Vision (ECCV) Part 2, Lecture Notes in Computer Science, vol.9914, pp.850–865, 2016.

[48] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.8971–8980, 2018.

[49] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P.H. Torr, "Fast online object tracking and segmentation: A unifying approach," Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1328–1338, 2019.

[50] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," Proceedings of the 24th IEEE International Conference on Image Processing (ICIP), pp.3645–3649, 2017.

[51] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," Computing Research Repository arXiv Preprint arXiv:2004.01888, 2020.

[52] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to

detect," Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), pp.3038–3046, 2017.

[53] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," Proceedings of the 17th IEEE International Conference on Computer Vision (ICCV), pp.941–951, 2019.

[54] G. Wang, Y. Wang, H. Zhang, R. Gu, and J.-N. Hwang, "Exploit the connectivity: Multi-object tracking with TrackletNet," Proceedings of the 27th ACM International Conference on Multimedia (ACMMM), pp.482–490, 2019.

[55] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," Computing Research Repository arXiv Preprint arXiv:2004.01177, 2020.

[56] I. Laptev, "On space-time interest points," International Journal of Computer Vision (IJCV), vol.64, no.2-3, pp.107–123, 2005.

[57] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3169–3176, 2011.

[58] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," Advances in Neural Information Processing Systems (NIPS), vol.27, pp.568–576, 2014.

[59] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," Proceedings of the 14th European Conference on Computer Vision (ECCV) Part 8, Lecture Notes in Computer Science, vol.9912, pp.20–36, 2016.

[60] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, "Action tubelet detector for spatio-temporal action localization," Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), pp.4405–4413, 2017.

[61] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," Proceedings of the 15th

IEEE International Conference on Computer Vision (ICCV), pp.4489–4497, 2015.

[62] R. Hou, C. Chen, and M. Shah, "Tube convolutional neural network (T-CNN) for action detection in videos," Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), pp.5822–5831, 2017.

[63] T. Lan, Y. Wang, G. Mori, and S.N. Robinovitch, "Retrieving actions in group contexts," Proceedings of the 11th European Conference on Computer Vision (ECCV) Part 1, Lecture Notes in Computer Science, vol.6553, pp.181–194, 2010.

[64] W. Li, M.-C. Chang, and S. Lyu, "Who did what at where and when: Simultaneous multi-person tracking and activity recognition," Computing Research Repository arXiv Preprint arXiv:1807.01253, 2018.

[65] S. Khamis, V.I. Morariu, and L.S. Davis, "A flow model for joint action recognition and identity maintenance," Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1218–1225, 2012.

[66] H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf, "Parametric correspondence and Chamfer matching: Two new techniques for image matching," Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI), pp.659–663, 1977.

[67] C. Tomasi and T. Kanade, "Detection and tracking of point features," Technical Report of Carnegie Mellon University, no.CMU-CS-91-132, 1991.

[68] G.R. Bradski, "Computer vision face tracking for use in a perceptual user interface," Intel Technology Journal, vol.2, no.2, pp.12–21, 1998.

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

[70] T.J. Broida and R. Chellappa, "Estimation of object motion parameters from noisy images," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.8, no.1, pp.90–99, 1986.

[71] D.A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," International Journal of Computer Vision (IJCV), vol.77, no.1-3, pp.125–141, 2008.

[72] J. Kwon and K.M. Lee, "Tracking by sampling trackers," Proceedings of the 13th IEEE International Conference on Computer Vision (ICCV), pp.1195–1202, 2011.

[73] T. Zhang, S. Liu, C. Xu, B. Liu, and M.-H. Yang, "Correlation particle filter for visual tracking," IEEE Transactions on Image Processing (IP), vol.27, no.6, pp.2676–2687, 2017.

[74] T. Zhang, C. Xu, and M.-H. Yang, "Learning multi-task correlation particle filters for visual tracking," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.41, no.2, pp.365–378, 2018.

[75] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol.20, no.3, pp.273–297, 1995.

[76] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S.L. Hicks, and P.H. Torr, "Struck: Structured output tracking with kernels," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.38, no.10, pp.2096–2109, 2015.

[77] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," Proceedings of the 10th European Conference on Computer Vision (ECCV) Part 1, Lecture Notes in Computer Science, vol.5302, pp.234–247, 2008.

[78] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp.983–990, 2009.

[79] D.S. Bolme, J.R. Beveridge, B.A. Draper, and Y.M. Lui, "Visual object tracking using adaptive correlation filters," Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2544–2550, 2010.

[80] D. Reid, “An algorithm for tracking multiple targets,” IEEE Transactions on Automatic Control (AC), vol.24, no.6, pp.843–854, 1979.

[81] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, “Sonar tracking of multiple targets using joint probabilistic data association,” IEEE Journal of Oceanic Engineering (JOE), vol.8, no.3, pp.173–184, 1983.

[82] Z. Khan, T. Balch, and F. Dellaert, “An MCMC-based particle filter for tracking multiple interacting targets,” Proceedings of the 8th European Conference on Computer Vision (ECCV) Part 4, Lecture Notes in Computer Science, vol.3024, pp.279–290, 2004.

[83] C. Yang, R. Duraiswami, and L. Davis, “Fast multiple object tracking via a hierarchical particle filter,” Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV), pp.212–219, 2005.

[84] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, “Online multiperson tracking-by-detection from a single, uncalibrated camera,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.33, no.9, pp.1820–1833, 2010.

[85] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank, and Z. Zhang, “Single and multiple object tracking using log-Euclidean Riemannian subspace and block-division appearance model,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.34, no.12, pp.2420–2440, 2012.

[86] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” Advances in Neural Information Processing Systems (NIPS), vol.28, pp.91–99, 2015.

[87] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.779–788, 2016.

[88] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg, “SSD: Single shot multibox detector,” Proceedings of the 14th European

Conference on Computer Vision (ECCV) Part 1, Lecture Notes in Computer Science, vol.9905, pp.21–37, 2016.

[89] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao, and T.-K. Kim, "Multiple object tracking: A literature review," Computing Research Repository arXiv Preprint arXiv:1409.7618, 2014.

[90] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," International Journal of Computer Vision (IJCV), vol.75, no.2, pp.247–266, 2007.

[91] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," Proceedings of the 23rd IEEE International Conference on Image Processing (ICIP), pp.3464–3468, 2016.

[92] A.A. Butt and R.T. Collins, "Multi-target tracking by Lagrangian relaxation to min-cost network flow," Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1846–1853, 2013.

[93] J.H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol.29, no.5, pp.1189–1232, 2001.

[94] A.V. Goldberg, "An efficient implementation of a scaling minimum-cost flow algorithm," Journal of Algorithms, vol.22, no.1, pp.1–29, 1997.

[95] X. Chen and J. Yang, "Towards monitoring human activities using an omnidirectional camera," Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI), pp.423–428, 2002.

[96] F. Rameau, D. Sidibé, C. Demonceaux, and D. Fofi, "Visual tracking with omnidirectional cameras: An efficient approach," Electronics Letters, vol.47, no.21, pp.1183–1184, 2011.

[97] Z. Zhang, P.L. Venetianer, and A.J. Lipton, "A robust human detection and tracking system using a human-model-based camera calibration," Proceedings of the 8th International Workshop on Visual Surveillance (VS), HAL-Inria,

no.inria-00325644, 2008.

[98] S. Gächter and T. Pajdla, “Motion detection as an application for the omnidirectional camera,” Research Reports of CMP, Czech Technical University in Prague, Omnidirectional Visual System (7), pp.5–13, 2001.

[99] G. Cielniak, M. Miladinovic, D. Hammarin, L. Goranson, A. Lilienthal, and T. Duckett, “Appearance-based tracking of persons with an omnidirectional vision sensor,” Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, vol.7, pp.84–89, 2003.

[100] H. Liu, W. Pi, and H. Zha, “Motion detection for multiple moving targets by using an omnidirectional camera,” Proceedings of the 2003 IEEE International Conference on Robotics, Intelligent Systems and Signal Processing (RISSP), vol.1, pp.422–426, 2003.

[101] M. Kobilarov, G. Sukhatme, J. Hyams, and P. Batavia, “People tracking and following with mobile robot using an omnidirectional camera and a laser,” Proceedings of the 2006 IEEE International Conference on Robotics and Automation (ICRA), pp.557–562, 2006.

[102] C.-J. Song, C.-M. Huang, and L.-C. Fu, “Human tracking by importance sampling particle filtering on omnidirectional camera platform,” IFAC Proceedings Volumes, vol.41, no.2, pp.6496–6501, 2008.

[103] A. Kawasaki, D.H. Hung, and H. Saito, “Human trajectory tracking using a single omnidirectional camera,” Proceedings of the 2014 Irish Machine Vision and Image Processing (IMVIP) Conference, pp.157–162, 2014.

[104] 久保大輔，“無人航空機システム（ドローン）の歴史と技術発展,” 計測と制御，vol.56，no.1，pp.12–17，2017.

[105] C. Teuliere, L. Eck, and E. Marchand, “Chasing a moving target from a flying UAV,” Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.4929–4934, 2011.

[106] B. Jeon, K. Baek, C. Kim, and H. Bang, "Mode changing tracker for ground target tracking on aerial images from unmanned aerial vehicles," Proceedings of the 10th International Conference on Control, Automation and Systems (ICCAS), pp.1849–1853, 2013.

[107] Y. Kim, W. Jung, and H. Bang, "Visual target tracking and relative navigation for unmanned aerial vehicles in a GPS-denied environment," International Journal of Aeronautical and Space Sciences, vol.15, no.3, pp.258–266, 2014.

[108] J. Rodriguez, C. Castiblanco, I. Mondragon, and J. Colorado, "Low-cost quadrotor applied for visual detection of landmine-like objects," Proceedings of the 2014 International Conference on Unmanned Aircraft Systems (ICUAS), pp.83–88, 2014.

[109] S. Zhao, Z. Hu, M. Yin, K.Z. Ang, P. Liu, F. Wang, X. Dong, F. Lin, B.M. Chen, and T.H. Lee, "A robust real-time vision system for autonomous cargo transfer by an unmanned helicopter," IEEE Transactions on Industrial Electronics, vol.62, no.2, pp.1210–1219, 2014.

[110] H. Zhang, G. Wang, Z. Lei, and J.-N. Hwang, "Eye in the sky: Drone-based object tracking and 3D localization," Proceedings of the 27th ACM International Conference on Multimedia (ACMMM), pp.899–907, 2019.

[111] F. Yang, S. Sakti, Y. Wu, and S. Nakamura, "A framework for knowing who is doing what in aerial surveillance videos," IEEE Access, vol.7, pp.93315–93325, 2019.

[112] P. Zhu, D. Du, L. Wen, X. Bian, H. Ling, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, L. Bo, H. Shi, R. Zhu, B. Dong, D. Reddy Pailla, F. Ni, G. Gao, G. Liu, H. Xiong, J. Ge, J. Zhou, J. Hu, L. Sun, L. Chen, M. Lauer, Q. Liu, S. Saketh Chennamsetty, T. Sun, T. Wu, V. Alex Kollerathu, W. Tian, W. Qin, X. Chen, X. Zhao, Y. Lian, Y. Wu, Y. Li, Y. Li, Y. Wang, Y. Song, Y. Yao, Y. Zhang, Z. Pi, Z. Chen, Z. Xu, Z. Xiao, Z. Luo, and Z. Liu, "Visdrone-VID2019: The vision meets drone object detection in video challenge results,"

Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

[113] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3286–3293, 2014.

[114] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.34, no.7, pp.1409–1422, 2011.

[115] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1838–1845, 2012.

[116] Y. Nagai, D. Kamisaka, N. Makibuchi, J. Xu, and S. Sakazawa, "3D person tracking in world coordinates and attribute estimation with PDR," Proceedings of the 23rd ACM International Conference on Multimedia (ACMMM), pp.1139–1142, 2015.

[117] L. Chen, W. Wang, G. Panin, and A. Knoll, "Hierarchical grid-based multi-people tracking-by-detection with global optimization," IEEE Transactions on Image Processing (IP), vol.24, no.11, pp.4197–4212, 2015.

[118] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.5695–5701, 2006.

[119] J.L. Bentley, "Multidimensional binary search trees used for associative searching," Communications of the ACM, vol.18, no.9, pp.509–517, 1975.

[120] 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.1–12, 2008.

[121] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L1 optical flow," Proceedings of the 29th DAGM Symposium, Lecture Notes in Computer Science, vol.4713, pp.214–223, 2007.

[122] S. Zagoruyko and N. Komodakis, "Wide residual networks," Proceedings of the 27th British Machine Vision Conference (BMVC), no.87, pp.1–12, 2016.

[123] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "MARS: A video benchmark for large-scale person re-identification," Proceedings of the 14th European Conference on Computer Vision (ECCV) Part 6, Lecture Notes in Computer Science, vol.9910, pp.868–884, 2016.

研究業績

本論文に関連する研究業績

査読付き学術雑誌

1. Hitoshi Nishimura, Kazuyuki Tasaka, Yasutomo Kawanishi, and Hiroshi Murase, “Multiple human tracking with alternately updating trajectories and multi-frame action features,” ITE Transactions on Media Technology and Applications (MTA), vol.8, no.4, pp.269–279, October 2020.
2. Hitoshi Nishimura, Naoya Makibuchi, Kazuyuki Tasaka, Yasutomo Kawanishi, and Hiroshi Murase, “Multiple human tracking using an omnidirectional camera with local rectification and world coordinates representation,” IEICE Transactions on Information and Systems, vol.E103-D, no.6, pp.1265–1275, June 2020.
3. 西村仁志，田坂和之，川西康友，村瀬洋，“複数の相関フィルタを用いた見えの変化に頑健な物体追跡，” 映像情報メディア学会誌，vol.73，no.5，pp.1004–1012，August 2019.

査読付き国際会議

1. Hitoshi Nishimura, Yuki Nagai, Kazuyuki Tasaka, and Hiromasa Yanagihara, “Object tracking by branched correlation filters and particle filter,” Proceedings of the 4th IAPR Asian Conference on Pattern Recognition (ACPR), vol.1, pp.79–84, November 2017.

研究会・シンポジウム等

1. 西村仁志，田坂和之，川西康友，村瀬洋，“基本行動特徴量を用いたオンライン複数人物追跡,” 2019 年度画像符号化シンポジウム／映像メディア処理シンポジウム（PCSJ/IMPS），no.P-4-15，November 2019.
2. 西村仁志，田坂和之，川西康友，村瀬洋，“人物検出と行動認識を統合したオンライン時空間行動検出手法の検討,” 2018 年映像情報メディア学会冬季大会，no.21C-1，December 2018.
3. 西村仁志，永井有希，小林達也，酒澤茂之，“相関フィルタを用いた確率的状態推定による長時間物体追跡,” 第 20 回画像の認識・理解シンポジウム（MIRU），no.PS2-7，August 2017.
4. 西村仁志，永井有希，小林達也，酒澤茂之，“カーネル化相関フィルタを用いた確率的運動予測に基づく物体追跡法,” 2016 年映像情報メディア学会冬季大会，no.14C-4，December 2016.

その他

1. 西村仁志，田坂和之，川西康友，村瀬洋，“複数の相関フィルタを用いた見えの変化に頑健な物体追跡,” 映像情報メディア学会誌　研究ハイライト，vol.74，no.6，pp.976–985，November 2020.
2. Hitoshi Nishimura, Kazuyuki Tasaka, Yasutomo Kawanishi, and Hiroshi Murase, “Multiple human tracking using multi-cues including primitive action features,” Computing Research Repository arXiv Preprint arXiv:1909.08171, September 2019.

その他の研究業績

査読付き学術雑誌

1. Houari Sabirin, Hitoshi Nishimura, and Sei Naito, “Synchronized tracking in multiple omnidirectional cameras with overlapping view,” IEICE Transactions

on Information and Systems, vol.E102-D, no.11, pp.2221–2229, November 2019.
2. 西村仁志，小篠裕子，有木康雄，中野幹生，“一般物体認識に基づく音声で指示された物体の選択法，” 電子情報通信学会論文誌（D），vol.J98-D，no.9，pp.1265–1276，September 2015.

査読付き国際会議

1. Hitoshi Nishimura, Yuko Ozasa, Yasuo Ariki, and Mikio Nakano, “Selection of an object requested by speech based on generic object recognition,” Proceedings of the 16th ACM International Conference on Multimodal Interaction (ICMI) Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction (MMRWHRI), pp.23–24, November 2014.
2. Hitoshi Nishimura, Yuko Ozasa, Yasuo Ariki, and Mikio Nakano, “Selection of unknown objects specified by speech using models constructed from Web images,” Proceedings of the 22nd IAPR International Conference on Pattern Recognition (ICPR), pp.477–482, August 2014.
3. Hitoshi Nishimura, Yuko Ozasa, Yasuo Ariki, and Mikio Nakano, “Object recognition by integrated information using Web images,” Proceedings of the 2nd Asian Conference on Pattern Recognition (ACPR), pp.657–661, November 2013.

研究会・シンポジウム等

1. Jianfeng Xu，Kazumasa Oniki，Hitoshi Nishimura，and Kazuyuki Tasaka，“A study on detection of kicking motions in multi-view 4K soccer videos，” 2018 年度画像符号化シンポジウム／映像メディア処理シンポジウム（PCSJ/IMPS），no.P-3-01，November 2018.
2. 西村仁志，小篠裕子，有木康雄，中野幹生，“Web 画像を用いた一般物体認識と指示発話の音声認識を統合した物体選択法，” 第 17 回画像の認識・理解シンポジウム（MIRU），no.SS2-36，July 2014.

3. Hitoshi Nishimura, Yuko Ozasa, Yasuo Ariki, and Mikio Nakano, “Object recognition by integrated information using speech and Web images,” 第 16 回画像の認識・理解シンポジウム（MIRU），no.SS5-25，July 2013.
4. 西村仁志，小篠裕子，有木康雄，中野幹生，“Web 画像を用いたマルチモーダル情報による物体認識，” 2013 年電子情報通信学会総合大会，no.D-12-16，March 2013.

受賞

1. 丹羽高柳賞論文賞，映像情報メディア学会，2019.