Vol.2021-CVIM-224 No.1
2021/1/21

複数ウェアラブルカメラ映像の人物識別とセグメンテーション

坂下 晴哉[1,a)]　武村 紀子[1,b)]　白井 詩沙香[1,c)]　Alizadeh Mehrasa[1,d)]　長原 一[1,e)]

概要：近年，IoT 機器を用いた人同士のコミュニケーションやグループ活動の分析が注目されている．特に，ウェアラブルカメラは，視野情報にカメラ装着者の意図や関心がよく反映されるため，多くの研究で使用されている．ウェアラブルカメラを装着した人同士のグループ活動では，それぞれの一人称カメラ映像間における人物同定が重要であるが，従来研究では十分な検討が行われていない．そこで，本研究では，複数一人称ウェアラブルカメラ映像間の人物のセグメンテーションおよび識別のためのフレームワークを提案する．また，グループ活動における一人称視点映像のデータセットを構築し，評価実験に用いることで，本提案手法の有効性を実証した．

1. はじめに

近年, IoT 機器を用いた人同士のコミュニケーションやグループ活動の分析が注目を集めており, 様々なデータが記録されてきた. 特に人の顔の表情, 視線, 体の形状, 動作など, 人同士のコミュニケーションやグループ活動の分析に有用な要素を捉えるためにカメラがよく用いられている. 更に多くの研究では, 1 人称視点映像を記録するために, ユーザーの頭に装着されたウェアラブルカメラを使用している. これはウェアラブルカメラの視野が着用者の頭の動きに応じて変化し, 1 人称映像に対象となる人間または物体が映るため, これらの動きからユーザーの意図や関心を検出できるためである. 人同士のグループ活動を分析する時に, たとえば, ある人が別の人に注意を向けている, すべての人が同じ人に関心を持っている, または人同士がお互いを見つめているといった, 各人物の意図や関心を表現し分析するために, それぞれの人がウェアラブルデバイスを着用し, 記録された 1 人称視点映像を分析する事は非常に有望なアプローチである. したがって, 複数ウェアラブルカメラ映像の人物識別とセグメンテーションは, 人同士のコミュニケーションやグループ活動の分析のための重要なプロセスである.

一般に複数のカメラに映る人物を追跡するためのタスクとして, マルチカメラトラッキング [1], [2] や人物再同定 [3], [4], [5], [6] といった問題が広く研究されてきた. しかし, ほとんどの手法が監視カメラのような固定されたカメラで撮影された人物の全身が映った状況を想定している. これは, よりカメラ同士が近接し, 人物のオクルージョンが発生するようなウェアラブルカメラを想定する本タスクとは大きく異なる問題設定である.

一方で物体認識と追跡のタスクは近年非常に盛んに研究されているコンピュータビジョンのタスクの一つである. しかし, これらに対するアプローチは単一動画を対象としており, 複数のウェアラブルカメラにおける人物の追跡と同定のために直接適応することはできない.

これらに対して, Xu ら [7] は, ウェアラブルカメラを含む異なるカメラ間における, 人物セグメンテーション情報の伝搬を基にした追跡のための手法を提案した. しかし, 彼らの手法はカメラが常にシーンを共有し, カメラに映る人数が一定でかつ既知であることを想定している. 各カメラの視野が人物の頭の動きに応じて絶えず変化し, カメラの映像には常に人物が新たに映る, もしくは消える可能性があるため, 彼らの仮定はウェアラブルカメラを使用する実際の状況においては現実的なものではないと考えられる.

本研究では複数ウェアラブルカメラ映像の人物同定とセグメンテーションのための新しいフレームワークを提案する. 図 1 に示すように, 提案フレームワークはグループメンバーとの会話などの人間の相互作用の分析への応用が期待されるものである. 提案フレームワークは, ローカル追跡モジュールとグローバル同定モジュールで構成されている. ローカル追跡モジュールは MaskTrack R-CNN [8] を土台

[1] 大阪大学
Osaka University
[a)] sakashita.haruya@ist.osaka-u.ac.jp
[b)] takemura@ids.osaka-u.ac.jp
[c)] shirai@ime.cmc.osaka-u.ac.jp
[d)] mehrasa@cmc.osaka-u.ac.jp
[e)] nagahara@ids.osaka-u.ac.jp

図 1 複数のウェアラブルカメラにおける人物同定とセグメンテーションの例．各行は各カメラで撮影された映像を表している．また，同じ色の領域は同じ人物である事を示している．

として用い，単一の動画における人物の検出と追跡を行う．グローバル同定モジュールは，複数の動画に対するローカル追跡モジュールで抽出したそれぞれの人物の外見の特徴量を用いて，異なる動画間での人物の対応付けを行う．また，3 つの新しい評価指標として，IoU_{multi} と AP_{multi}，AR_{multi} を提案する．IoU_{multi} は，複数の動画における IoU を評価するための指標であり，AP_{multi} と AR_{multi} は，複数の動画の平均適合率と平均再現率を評価するための指標である．提案手法と既存手法で実験と評価を行い，提案方法の有効性が示された．

2. 関連研究

2.1 人物同定

マルチカメラトラッキング

一般に複数のカメラで人物を検出，同定および追跡するタスクはマルチカメラトラッキングと呼ばれ，人物監視や群集分析への応用が期待されている．Wen ら [1] は，3 次元の地理的情報と人物動作の連続性制約を考慮した時空間グラフを提案した．Ristani と Tomasi [2] は，人物の外観情報と軌跡を利用した相関クラスタリングによって人物を同定した．しかし，これらの手法は監視カメラなどの静的カメラを想定しているため，複数のウェアラブルカメラ間における人物追跡と同定には適応できない．

人物再同定

動画に登場する人物を同定するタスクは，人物再同定と呼ばれ，コンピュータビジョンの重要なタスクの 1 つとして広く研究されている．このタスクでは，カメラから得た動画像に登場する人物が入力として与えられ，これを同定することを目標としている．Li ら [3] は，人物の部位ごとの特徴の重要性の差に着目し，パッチベースの同定モデルを提案した．Qian ら [4] は，マルチスケールの人物特徴量を用いることで，カメラと人物の位置による外見と解像度の違いに対応した．しかし，これらの手法は，特徴量を用いて現在のフレーム内の人物を同定するだけであり，動画の時間的情報は用いていない．

一方で，動画を対象とした人物同定を目標とした研究もされている．Zheng ら [5] は，オプティカルフローと体の形状情報を用いて人物の動きを捉えることで，異なるカメラ間で人物を同定した．Liang ら [6] は，信頼度指標によって人体の各関節に重みを付け，それを動き情報と組み合わせた手法を提案した．しかし，これらの手法では，各単一ビデオでの人物追跡が完全になされていることを前提としており，人物追跡とセグメンテーションの問題は考慮されていない．

複数のウェアラブルカメラ

本タスクと同様に，複数のウェアラブルカメラを想定した研究がされている．Fan ら [9] は，一人称と三人称のビデオ間で人物を対応づけし，カメラの着用者を特定する方法を提案した．Ardeshir ら [10] は，グラフマッチングを用いて，ウェアラブルカメラを装着した歩行者の 1 人称視点映像と天井定点カメラに映る人物との間の対応づけを行った．Xu ら [7] は，複数のウェアラブルカメラで撮影された人物をセグメンテーションおよび同定する手法を提案し，両方のタスクに対する手法の有効性を実証した．しかし，彼らの設定では，常に異なるウェアラブルカメラで常にシーンが共有され，同定とセグメンテーションは，グローバル個人 ID の一貫性を考慮せずに別々に評価されている．

2.2 物体追跡とセグメンテーション

物体追跡とセグメンテーションは，物体の位置とシーン内の物体形状を推定するための，コンピュータビジョンにおける重要なタスクである．Caelles ら [11] は，単一ビデオの 1 つのフレームで指定された物体を追跡するタスクを提案し，初期フレームで指定された単一の物体領域情報を用いてセグメンテーションモデルを再学習することで，複数のフレームで指定した物体のみを同定しセグメンテーションするモデルを構築した．Yang ら [8] は，Mask R-CNN [12] を単一画像から単一動画に拡張することで，ビデオインスタンスセグメンテーションとそのタスクに対する初めてのアプローチを提案した．彼らのモデルは，外部メモリに保存した過去の物体特徴を現在の物体特徴と照合することで，単一動画における一意な物体セグメンテーションおよび追跡を可能にした．しかし，ほとんどの物体追跡手法では，複数カメラを対象としておらず，単一動画を想定している．

3. 複数ウェアラブルカメラ映像の人物識別とセグメンテーションのための提案フレームワーク

本研究では，複数のカメラで人物を同定して追跡するために，単一動画で人物を追跡するためのローカル追跡モジュールと，複数の動画間で人物を照合するためのグロー

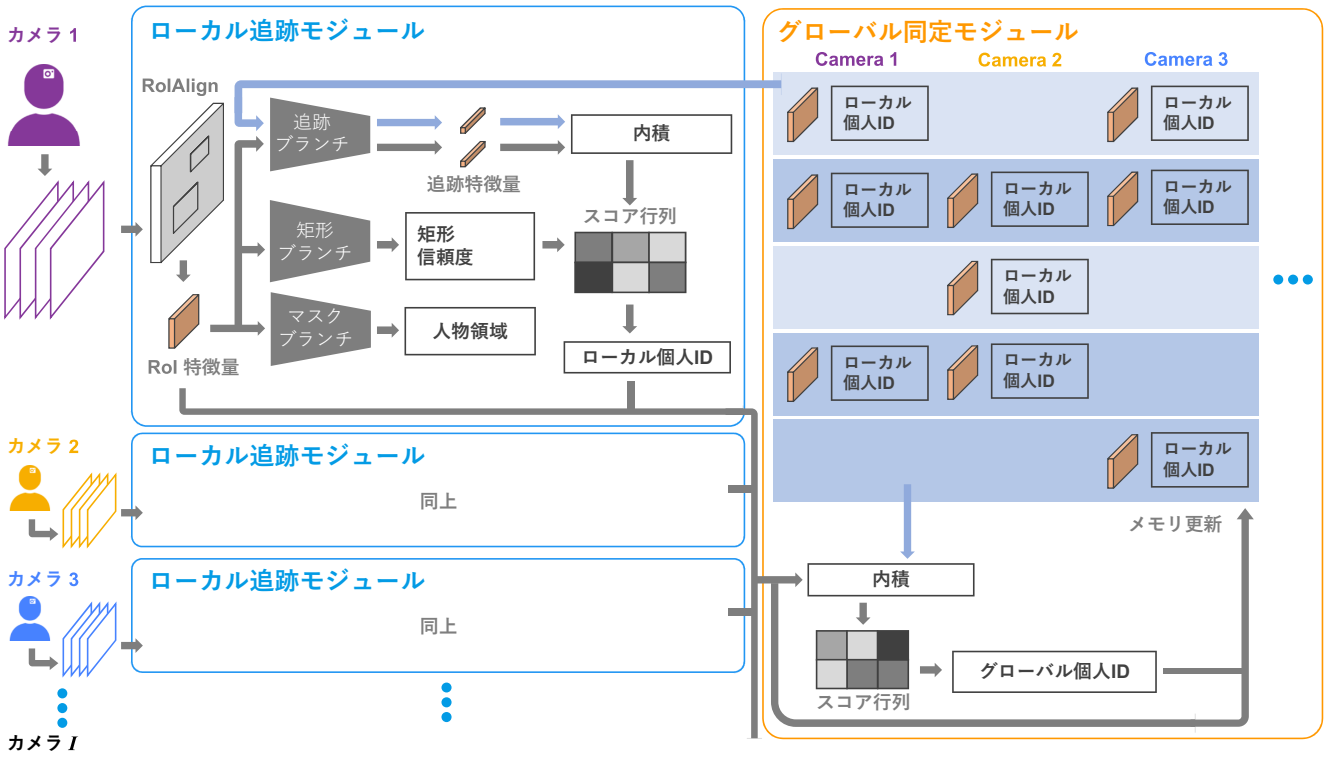

図 2 提案フレームワークの概要

バル同定モジュールから構成されたフレームワークを提案する．図 2 に提案手法の概要を示す．ローカル追跡モジュールは各カメラそれぞれに 1 つずつに用意され，単一の動画内の人物を一意に検出，セグメンテーション，追跡し，異なるカメラ間における同定のためにグローバル同定モジュールに入力するための特徴量を抽出する．グローバル同定モジュールは，ローカル追跡モジュールから入力された人物の特徴量を用いて，複数の動画に映る人物を同定する．

3.1 ローカル追跡モジュール

図 2 に示すように，カメラごとに 1 つのローカル追跡モジュールが割り当てられ，それらは並行して実行される．このローカル追跡モジュールは，単一動画において人物を検出し，追跡するために，MaskTrack R-CNN に基づいて構築されている．まず，RPN [13] が動画の各フレーム画像を入力として，フレーム画像に映る人物のバウンディングボックスを人物の候補として検出する．この時，人物の候補として検出されたバウンディングボックスから RoIAlign [12] を通して固定長ベクトルで表される特徴量が抽出される．本フレームワークではこの特徴量を RoI 特徴量と呼ぶ．次に，追跡ブランチ，バウンディングボックスブランチ，マスクブランチの 3 つのブランチがそれぞれ RoI 特徴量を入力として追跡と回帰的な矩形修正と物体領域検出を行う．追跡ブランチ [8] は，追跡のために RoI 特徴量から 1024 次元の特徴量を抽出する．本フレームワークではこれを追跡特徴量と呼ぶ．ローカル追跡モジュールは，現在のフレーム画像から得た追跡特徴量と，グローバル同定モジュールに保存されている過去の特徴量を使用して，単一動画で人物追跡を行う．バウンディングボックスブランチは，RoI 特徴量を入力として物体クラス分類とバウンディングボックス回帰を行う．マスクブランチは，RoI 特徴量を入力として，バウンディングボックス内の人物の形状のセグメンテーションを行う．そして，ローカル追跡モジュールは，RoI 機能と単一動画で追跡した個人 ID をグローバル同定モジュールに渡す．

追跡ブランチは，RoI 特徴量を入力として，2 つの全結合層を通して，単一動画での追跡のために追跡特徴量を抽出する．また，過去のフレームにおける人物の特徴量も同様にして，グローバル同定モジュールに保存されている過去の RoI 特徴量から追跡特徴量を抽出する．検出された人物が同一人物だと同定する確率は，現在および過去の追跡特徴量を使用して計算される．N 個の個人 ID$n(n = 0, 1..., N)$ がすでに同定されているとする．このとき，検出された候補人物 j を個人 IDn に割り当てる確率 $P_j(n)$ は，次のように計算される．

$$P_j(n) = \begin{cases} \frac{e^{\mathbf{f}_j^\top \mathbf{f}_n}}{1+\sum_{k=1}^{N} e^{\mathbf{f}_j^\top \mathbf{f}_n}} & (n \in [1, N]) \\ \frac{1}{1+\sum_{k=1}^{N} e^{\mathbf{f}_j^\top \mathbf{f}_n}} & (n = 0), \end{cases} \quad (1)$$

ここで，$\mathbf{f}_j$ と $\mathbf{f}_n$ は，追跡ブランチから抽出された追跡特徴量である．

バウンディングボックスブランチでは，まず全結合層に RoI 特徴量を入力することで，1 次元ベクトルの特徴量を

得る. 次に物体クラス分類のための全結合層とバウンディングボックス回帰のための全結合層に入力することで, 各レイヤーから 1 つずつの合計 2 つの出力を得る. これらによって, ローカル追跡モジュールは, 動画に映る人物に対応するバウンディングボックスを推定できる. またバウンディングボックスとその信頼度は, 単一の動画で人物追跡を行うための同定スコアで使用される.

マスクブランチは畳み込み層に RoI 特徴量を入力することで, 空間的な人物情報を示す特徴量を抽出する. 次に, 逆畳み込み層でこの特徴量をアップサンプリングすることにより, ピクセルレベルの人物領域を推定することができる.

単一動画における人物追跡のための有用な手がかりとして, 追跡ブランチで計算された同定確率 P, バウンディングボックスブランチで取得されたバウンディングボックスとその信頼度である b と g を用いる. 検出された候補人物 j を個人 IDn に割り当てることの追跡スコア $u_j(n)$ を次のように定義する.

$$u_j(n) = \log P_j(n) + \alpha \log g_i + \beta \,\mathrm{IoU}(b_j, b_n). \qquad (2)$$

$P_j(n)$ は, 式 1 によって計算された追跡確率である. $\mathrm{IoU}(b_j, b_n)$ は, b_j と b_n の IoU を示す. ここで IoU は, 2 つの領域の重なりを 2 つの領域の合計で割ったものとして定義される指標である [14]. α と β は, 2 つの要素に重みを付けるためのハイパーパラメータである. 次に, すべての検出された候補人物と過去の人物のペアで追跡スコア $u_j(n)$ を計算して, スコア行列を得る.

MaskTrack R-CNN では, スコア行列を用いて追跡スコアが最も高い候補を同一人物であると同定している. しかし, 特定の個人 ID に対して複数候補人物の追跡スコアが高い場合, 追跡スコアが低い個人 ID には候補人物が割り当てられないという問題がある. したがって, 競合することなく可能な限り個人 ID に候補人物を割り当てるために, ハンガリアンアルゴリズム [15] を用いてスコア行列のすべての可能な組み合わせの中でスコアが最大パターンを採用した. さらに, 人物らしくない検出を除くために, クラス分類における人間クラスの信頼度がしきい値 η_{class} を下回る候補人物を破棄する.

3.2 グローバル同定モジュール

グローバル同定モジュールは, ローカル追跡モジュールによって抽出された RoI 特徴量に基づいて, 複数の動画で検出された人物の対応付けを行い, 同定する. 複数の動画全体で M 個のグローバル個人 IDm $(m = 0, 1..., M)$ と N_i 個のローカル個人 IDn_i $(n_i = 0, 1..., N_i$ があるとする. このとき, 単一ビデオに映るローカルに検出された人物には, これまでにいずれかのグローバル個人 ID が割り当てられている場合には特定のグローバル個人 ID を対応付ける. 他のビデオに映ったことのない新しい人物である場合には新しいグローバル個人 ID を対応付ける. 本研究では, 複数の動画で人物を同定するために, 人物の外見の特徴と単一ビデオでローカルに追跡されるローカル個人 ID の一貫性を用いた効果的な手法を提案する.

複数の動画における人物同定では, カメラの違いに対応するために, 人物の外見情報を利用することが効果的だと考えられる. したがって, 検出された人物の外見の類似度は, ローカル追跡モジュールで RoIAlign [12] によって抽出された RoI 特徴量を用いて計算する. さらに, ローカル追跡モジュールでの単一ビデオにおける追跡の一貫性を考慮するために, 人物が過去のローカル個人 ID と一致するかどうかによって人物の外見の類似度に重みを付ける. ここで, カメラ i 内でローカルに検出された人物 n_i をグローバル個人 IDm に割り当てるための同定スコア $s(n, m)$ は, 次のように定義する.

$$\begin{aligned} s(n_i, m) = &\omega(n_i, n'_m)\times \\ &(S_G(n_i, m) + S_L(n_i, m)), \end{aligned} \qquad (3)$$

$$where \quad \omega(n_i, n'_m) = \begin{cases} p & (n_i = n\prime_m) \\ 1-p & (n_i \neq n\prime_m). \end{cases}$$

n'_m は, 以前に検出されたグローバル人物 ID が m である人物うち, ローカルに検出された人物 n が映る動画と同じ動画から検出された人物のローカル個人 ID である. $\omega(n_i, n'_m)$ は, ローカル個人 ID の一貫性を考慮した重み関数であり, p は, ローカル個人 ID の一貫性を示すハイパーパラメーターである. $S_L(n, m)$ は, 以前に検出されたグローバル人物 ID が m である人物うち, ローカルに検出された人物 n が映る動画と同じ動画である人物との類似性である. これに対して $S_G(n, m)$ は, ローカルに検出された人物 n が映る動画と異なる動画である人物との類似性である. $S_G(n, m)$ および $S_L(n, m)$ は次のように定義する.

$$S_G(n_i, m) = \mathrm{avg}\{\mathbf{f}_{n_i}^{\top} \mathbf{f}_{m_{i'}}\} \qquad (i \neq i'), \qquad (4)$$

$$S_L(n_i, m) = \mathbf{f}_{n_i}^{\top} \mathbf{f}_{m_{i'}} \qquad (i = i'). \qquad (5)$$

$\mathbf{f}_{n_i}$ は, カメラ i に割り当てられたローカル追跡モジュールで抽出した, ローカルに検出される人物 n の特徴量である. $\mathbf{f}_{m'_i}$ は, グローバルな個人 IDm を持つ以前に検出された人物の内, カメラ i' で撮影した動画から検出された人物の特徴量である.

それぞれの動画ごとに, すべてのローカルに検出された人物とグローバル個人 ID のペアに対して計算した同定スコアからすべての現実的に可能なパターンの尤度を計算し, ハンガリアンアルゴリズムを用いて最尤パターンを採用する. なお, 特定の動画にのみ映る人物の同定に対応するために, 同定スコアがしきい値 η_G 未満の人物には, 新しいグローバル個人 ID が割り当てる. 新しいグローバル個人 ID が割り当てられると, その ID がメモリに新しく挿入され,

対応する人物の特徴量とローカル個人 ID が保存される．また, 既存のグローバル個人 ID に保存されている以前の特徴量とローカル個人 ID は, 現時点で検出された対応するローカル個人 ID とその特徴量によって上書きされる．

3.3 実装

ローカル追跡モジュールは文献 [8] の設定に基づいており, それらの公開されている実装およびコードを使用した．式 2 のハイパーパラメータ α と β は, 文献 [8] で用いられている値である 1 と 2 とした．また, ネットワークのバックボーンとして ResNet-50-FPN [12] が用いた．これに対して, MSCOCO [16] と YouTubeVIS [8] で学習された事前学習済みモデルを適用した．式 3 のパラメータ p と, しきい値パラメータ η_G, および η_{class} については 4.5 節で説明する．

4. 評価実験

提案手法の有効性を検討するために評価実験を行った．複数の動画全体における追跡と同定を包括的に評価するには, 手法の特性の分析を目的として, 単一動画における時間的一貫性を考慮した追跡精度と, 複数の動画における時間的一貫性と空間的一貫性の両方を考慮した追跡と同定精度の両方を分析するべきである．そこで, 本研究では単一動画における追跡に加えて, 複数動画における追跡を評価するための新しい評価指標を提案する．また, 評価のために関連する手法を本タスクに適応させ, 提案手法と比較した．手法の傾向と特性および精度の分析のために, 既存データセットである IUShareView データセット [7] に加えて, 新たに独自のデータセットを構築して 2 つの異なるデータセットで手法を評価した．IU ShareView データセットには, 屋内の 2 台のカメラで撮影された短い動画セットと評価のためにラベル付けされた個人 ID と人物領域が含まれているため, 限られた環境でしか評価ができない．そこで, 3 台から 5 台のウェアラブルカメラで撮影したより複雑な環境における評価を行うための新しいデータセットを作成した．

4.1 評価指標

本研究では単一動画と複数動画の同定とセグメンテーションの精度を評価するために, 4 つの評価指標を用いた．まず初めに, 単一動画の同定とセグメンテーションを評価する指標である IoU_{single} [8] について説明する．

$$\mathrm{IoU}_{single} = \frac{\sum_{t=1}^{T} |A_t \cap B_t|}{\sum_{t=1}^{T} |A_t \cup B_t|}, \qquad (6)$$

ここで, A_t は, 時刻 t におけるモデルの予測した人物領域であり, B_t は真値の人物領域である．なお, カメラや人物の動きによって人物が消える, もしくは再出現することがあるため, フレームに人物領域が存在しない場合を考慮する必要がある．よって, フレーム t に人物が存在しない場合, A_t と B_t を空白で詰めることとする．IoU_{single} は, 単一の動画の時空間領域を通じて, 予測された人物領域と真値の間の重なりを評価し, 単一動画のセグメンテーションにおける個人の空間的および時間的一貫性を評価することができる．

次に IoU_{multi} を新しく定義する．これは, IoU_{single} を単一動画から複数動画に拡張し, 複数のカメラにおける動画に対するセグメンテーションを評価することができる．

$$\mathrm{IoU}_{multi} = \frac{\sum_{i}^{I} \sum_{t=1}^{T} |A_{(t,i)} \cap B_{(t,i)}|}{\sum_{i}^{I} \sum_{t=1}^{T} |A_{(t,i)} \cup B_{(t,i)}|}, \qquad (7)$$

ここで, i は動画 ID, $A_{(t,i)}$ は動画 i の時間 t における予測された人物領域, $B_{(t,i)}$ は真値の人物領域である．IoU_{multi} は, 複数の動画全体における予測された人物領域と真値の間の重なりを評価し, グローバルな個人 ID の一貫性, および複数の動画における空間的および時間的な一貫性を評価できる．

さらに, 物体検出の標準的な評価指標である平均適合率 (AP) と平均再現率 (AR) [16] を複数の動画に対して定義することで, 複数の動画全体における検出と同定の精度を定量的に評価する．AP は, 信頼スコアに基づいて適合率（横軸）と再現率（横軸）でプロットした曲線の下部の領域として定義される．たとえば, $\mathrm{AP}_{single(25)}$ は, IoU_{single} に対して 25%のしきい値を設けた時の AP を表す．AR は, 固定数の検出された人物領域が与えられた時の再現率（横軸）と IoU（横軸）でプロットした曲線の下部の領域として定義され, 複数の IoU しきい値を平均することによって計算される．MS COCO の評価方法 [16] と同様に, 50%から 95%までの 5%刻みで合計 10 個の IoU しきい値を用いる．たとえば, $\mathrm{AR}_{multi(1)}$ は, 複数の動画セットごとに 1 人の検出された人物領域を与えた時の AR を表す．

4.2 比較手法

本研究では, 複数の動画において人物領域をセグメンテーションして同定するタスクを対象とし, 複数の動画間で個人 ID の一貫性を維持し, 各ビデオの時空間の一貫性を維持する手法を提案している．私たちの知る限り, この点を直接評価した既存研究は存在しないが, Xu ら [7] による手法は, 提案手法に最も類似する研究である．しかし, 彼らは IoU_{single} を使用して, 単一のビデオでのみセグメンテーション精度を評価し, 複数動画に対してフレームごとに同定を行い評価したため, 複数動画に対するグローバル個人 ID の一貫性は考慮されておらず評価されていない．そのため, 本研究では 2 種類の評価を適応した．具体的には, IoU_{single} [7] による単一動画に対する評価と, IoU_{multi} を用いたグローバル個人 ID の一貫性を考慮した複数動画に対する評価を行った．なお, AP_{multi} と AR_{multi} は, 彼らの

手法に複数の動画間におけるグローバル個人 ID の一貫性がないため, 計算できない.

比較のために, 公開されているオリジナルの実装コード [7] を用いて, IoU_{single} の結果を再現した. また, 4.1 節で説明している IoU_{multi} の定義は Xu ら [7] の手法が想定している評価と異なるため, IoU_{multi} を用いて文献 [7] の結果と比較する場合, フレームごとの同定アルゴリズムによって算出された個人 ID のすべての可能な組み合わせの候補から IoU_{multi} が最高となる時のスコアを選択した.

比較手法では, 初期フレームの人物領域が入力としてラベル付けされ, 単一ビデオにおいて人物領域を伝播するため, 人物が動画にフレームインおよびフレームアウトしないことを前提としている. しかし, 提案するデータセットには頻繁なフレームインとフレームアウトが含まれているため, 提案方法と直接比較することができない. したがって, 提案するデータセットでは比較手法に対して初回ラベリングおよび複数回ラベリングの 2 つの設定を想定した. 初回ラベリングでは, 文献 [7] の設定と同じように人物が動画に初めて登場するときに真値の人物領域をラベル付けし, 動画全体でこれを伝播する. 一方で, マルチタイムラベリングでは, 人物が再びフレームインした時に人物領域を再ラベル付けする.

4.3 独自のデータセットにおける評価

本研究では, 既存の IU ShareView データセット [7] に含まれていない環境で評価するためのデータセットを作成した. この独自のデータセットは, 既存のデータセットとは異なり, グループ内の屋内と屋外の会話シーンで構成されており, すべてのシーンは 3〜5 台のウェアラブルカメラで撮影された約 3〜6 分の 6fps の動画で構成されている. また, 評価のために 30 秒のシーケンスを 5 つ選択した. 選択したシーケンスは, 既存のデータセットよりも長時間であり, フレームインまたはフレームアウトの人物が含まれているため, より厳しい環境であるといえる. このシーケンスに対して 1 秒ごとに 1 フレームを切り出し, 人物領域と個人 ID のをラベル付けした. 結果として合計で 1257 個の人物領域と個人 ID がデータセットに含まれている.

表 1 に, 提案手法と既存手法の結果を示す. 提案フレームワークのパラメーター η_G, η_{class} および p は, 3800, 0.7, 0.6 に設定している. 独自のデータセットを用いて, 2 台, 3 台, 4 台, 5 台のカメラを入力とした 4 つの設定で同定精度を評価した. 各設定では, データセットに含まれるカメラ間の取りうるすべての組み合わせを用いており, 提案手法はすべての指標で既存手法に比べて優れた精度を達成した.

まず 2 台のカメラを入力とした時の同定結果において, 提案手法はすべての評価指標で既存手法より同定精度が向上した. これは, 既存手法は時間方向への人物領域の伝搬に基づいているため, 長時間にわたる視野の大きな動作やカメラに映る人物のオクルージョンに対応していないためであると考えられる. よって, 既存手法におけるこれらの問題が, 人物のセグメンテーションと同定の両方に悪影響を及ぼしたと考えられる. 一方で, カメラの数が増えると, 提案フレームワークはすべての評価指標に対して識別精度が低下した. なお, 既存手法は 2 台のカメラ間の人物識別を想定しており, 3 台以上のカメラを同時に使用してセグメンテーションと識別を行うことはできない. カメラの台数の増加に伴いフレームワークの識別精度が低下していることから, 依然として複数カメラにおける人物同定は難しいタスクであり, カメラの台数が増える程より困難なタスクとなることが確認できる.

図 3 に提案手法の定性的結果の例を示す. 図中 (a) は同定に成功した結果を示しており, (b) は同定に失敗した結果を示している. この結果から, 衣服といった人物の外見情報が, 人物同定にとって重要であると考えられる. したがって, 似た衣服を着ていることによって外見が似た人物が複数人カメラに映ると, 同定が困難になる事がわかる. なお, (a) における赤い人物領域で表された人物が, 上段の 3 番目と 4 番目のフレームでフレームアウトしており, 同様に (b) における紫の人物領域で表された人物領域が, 下段の 4 番目のフレームでフレームアウトしている. これらのフレームはいずれも後のフレームで再登場し, 正しく再同定されている, このことから, 提案フレームワークは, 時間的一貫性を考慮した人物同定によって, 人物のフレームインとフレームアウトに対応している事が分かる.

4.4 IU ShareView データセットにおける評価

既存データセットである IU ShareView データセット [7] は, テストセットとして, 2 台のウェアラブルカメラで撮影された 3 つの動画セットから選択された 1 4 秒の 9 つのシーケンスで構成されている. 全ての動画は屋内で撮影され, テストセットに人物のフレームインやフレームアウトは含まれていない. テストセットの各フレームには手動で 580 個の人物領域と個人 ID がラベル付けされている.

表 2 は, IU ShareView [7] データセットにおける提案手法と既存手法の結果を示す. フレームワークのパラメーター η_G, η_{class}, および p は, 3800, 0.7, 0.6 に設定している. 提案手法は既存手法と比較して, 事前情報の参照やカメラに映る人数の制限がないにもかかわらず, 動画全体の時間的かつ空間的一貫性を考慮することで, IoU_{single} と IoU_{multi} の両方でより良い識別精度を達成した.

4.5 パラメータの検討

提案手法には, ハイパーパラメータ η_G, η_{class} および p があり, これらのパラメータに関して最適な値の検証を行った. なお, それぞれのパラメータに関する分析には独自のデータセットを用いた.

表 1 提案方法と既存手法の比較. この表では, AP_m と AR_m は AP_{multi} と AR_{multi} を表している. 最良の結果は太字で強調している.

	Num Videos	$AP_{m(25)}$	$AP_{m(50)}$	$AR_{m(1)}$	$AR_{m(10)}$	IoU_{single}	IoU_{multi}
既存手法 (初回ラベリング)	2 Cameras	N/A	N/A	N/A	N/A	9.5	10.2
既存手法 (複数回ラベリング)	2 Cameras	N/A	N/A	N/A	N/A	12.0	12.4
提案手法	2 Cameras	**51.6**	**17.4**	**7.1**	**20.8**	**62.4**	**47.4**
既存手法	3–5 Cameras	N/A	N/A	N/A	N/A	N/A	N/A
	3 Cameras	31.4	6.6	3.0	10.3	54.2	38.0
提案手法	4 Cameras	21.9	2.1	0.7	4.9	48.2	34.0
	5 Cameras	16.4	0.0	0.0	3.6	47.7	31.2

表 2 IU ShareView 評価セットに対する提案方法と既存手法の比較

Method	$AP_{m(25)}$	$AP_{m(50)}$	$AR_{m(1)}$	$AR_{m(10)}$	IoU_{single}	IoU_{multi}
既存手法 (初回ラベリング)	N/A	N/A	N/A	N/A	62.2	51.4
提案手法	81.3	35.5	15.8	41.8	**77.1**	**55.6**

表 3 しきい値 η_G の検討. η_{class} および p は 0.7 と 0.6 に設定している.

η_G	$AP_{single(25)}$	$AP_{multi(25)}$	IoU_{single}	IoU_{multi}
2600	**62.7**	50.3	**62.7**	42.1
3800	60.1	**51.6**	62.4	**47.4**
5000	35.5	27.9	54.5	43.6

表 4 しきい値 η_{class} の検討. η_G および p は, 3800 と 0.6 に設定している.

η_{class}	$AP_{single(25)}$	$AP_{multi(25)}$	IoU_{single}	IoU_{multi}
0.999	40.6	30.8	51.9	35.0
0.7	**60.1**	**51.6**	**62.4**	**47.4**
0.0	45.9	37.9	58.8	45.1

表 5 重みパラメータ p の検討. η_G および η_{class} は, 3800 と 0.7 に設定している.

p	$AP_{single(25)}$	$AP_{multi(25)}$	IoU_{single}	IoU_{multi}
0.5	34.2	31.7	53.8	41.5
0.6	**60.1**	**51.6**	**62.4**	**47.4**
0.7	58.6	45.5	60.0	44.9

表 3 に, 人物同定の際に検出される人物が複数の動画間で異なるかどうかを判断するためのしきい値である η_G の評価結果を示す. η_G を大きくする程, IoU_{single} と $AP_{single(25)}$ が低下している. これはしきい値を大きくするほど同一人物が異なる人物だと判断されるためだと考えられる. 一方で IoU_{multi} と $AP_{multi(25)}$ に関しては, しきい値が小さすぎると, 異なる人物が同一人物だと判断されるため, 低い値を取ることがわかる. しかし, しきい値が大きすぎると, IoU_{single} と $AP_{single(25)}$ の低下に合わせて IoU_{multi} と $AP_{multi(25)}$ も低下する. これは複数の動画間の識別精度の向上には, 単一動画内の追跡精度が十分高い必要があるためであると考えられる. 表 3 の結果のバランスをとって, 本実験ではしきい値 η_G に 3800 を設定した.

表 4 に, ローカル追跡モジュールにおいて, 人物らしくない検出候補を除外するためのしきい値 η_{class} の評価結果を示す. しきい値を増やしすぎると, より人間らしい検出のみを採用するあまり, 正しく検出された候補も除外されてしまうため, すべての評価指標で値が低下した. 一方で, しきい値が大きすぎると, 人間らしくない検出を採用してしまうため, すべての評価指標で値が低下した. 表 4 の結果のバランスをとって, 本実験ではしきい値 η_{class} に 0.7 を設定した.

表 5 は, 人物の類似性に対するローカル個人 ID の一貫性を考慮するための重みパラメーター p の評価結果を示す. パラメータを高くしすぎると, 単一動画内におけるローカル個人 ID の一貫性をより強く考慮するが, グローバルな人物の整合性を低下させてしまい, またその整合性を単一動画内のローカルな同定にフィードバックしなくなるため, すべての評価指標で値が低下した. 一方で, パラメータが低すぎると, 単一動画内におけるローカル個人 ID の一貫性が失われるため, IoU_{single} と $AP_{single(25)}$ が低下した. 更に, 異なる動画間のグローバルな人物同定には単一動画内のローカルな人物同定が正しく行われる必要があるため, IoU_{multi} と $AP_{multi(25)}$ も低下している. 表 5 の結果のバランスをとって, 本実験ではパラメータ p に 0.6 を設定した.

5. 結論

近年カメラから人同士のコミュニケーションやグループ活動を分析することが大きな注目を集めている. 特に複数の動画にわたって人物を同定することは, アプリケーションの面で重要な処理である. 本研究では複数のウェアラブルカメラにおける人物のセグメンテーションと同定のための新しいフレームワークを提案した. 提案手法は, ローカル追跡モジュールおよびグローバル同定モジュールの 2 種類のモジュールで構成されている. ローカル追跡モジュールは, 単一動画で人物をセグメンテーションし, 追跡するためのモジュールであり, グローバルマッチングモジュールは, 複数の動画間で人物を同定するためのモジュールであ

図 3 提案手法の定性的結果の例. (a) と (b) はどちらも, 2 台のカメラで同定した結果である. (a) は正しい予測の場合であり, (b) は失敗した場合の例である. 同じ色の領域が同じ人物だと予測していることを表している.

る. 提案手法は, 複数の動画において人物を同定するために, 複数の動画間で時空間的な一貫性を保証している. また, 既存手法と異なり, シーン内の人数などの事前知識を必要としない. また, 本タスクを評価するために, 3 から 5 台のウェアラブルカメラで撮影された屋内と屋外のシーンから成る独自のデータセットを作成した. 独自データセットと既存データセットを用いて評価実験を行い, 既存手法と比較して提案手法の有効性を確認した.

参考文献

[1] Wen, L., Lei, Z., Chang, M.-C., Qi, H. and Lyu, S.: Multi-camera multi-target tracking with space-time-view hyper-graph, *International Journal of Computer Vision*, Vol. 122, No. 2, pp. 313–333 (2017).

[2] Ristani, E. and Tomasi, C.: Features for multi-target multi-camera tracking and re-identification, *Proc. Computer Vision and Pattern Recognition (CVPR)*, pp. 6036–6046 (2018).

[3] Li, W., Zhao, R., Xiao, T. and Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification, *Proc. Computer Vision and Pattern Recognition (CVPR)*, pp. 152–159 (2014).

[4] Qian, X., Fu, Y., Jiang, Y.-G., Xiang, T. and Xue, X.: Multi-scale deep learning architectures for person re-identification, *Proc. International Conference on Computer Vision (ICCV)*, pp. 5399–5408 (2017).

[5] Zheng, K., Fan, X., Lin, Y., Guo, H., Yu, H., Guo, D. and Wang, S.: Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras, *Proc. International Conference on Computer Vision (ICCV)*, pp. 2858–2866 (2017).

[6] Liang, G., Lan, X., Zheng, K., Wang, S. and Zheng, N.: Cross-view person identification by matching human poses estimated with confidence on each body joint, *AAAI Conference on Artificial Intelligence* (2018).

[7] Xu, M., Fan, C., Wang, Y., Ryoo, M. S. and Crandall, D. J.: Joint person segmentation and identification in synchronized first-and third-person videos, *Proc. European Conference on Computer Visionn (ECCV)*, pp. 637–652 (2018).

[8] Yang, L., Fan, Y. and Xu, N.: Video instance segmentation, *Proc. International Conference on Computer Vision (ICCV)*, pp. 5188–5197 (2019).

[9] Fan, C., Lee, J., Xu, M., Kumar Singh, K., Jae Lee, Y., Crandall, D. J. and Ryoo, M. S.: Identifying first-person camera wearers in third-person videos, *Proc. Computer Vision and Pattern Recognition (CVPR)*, pp. 5125–5133 (2017).

[10] Ardeshir, S. and Borji, A.: Egocentric meets top-view, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 41, No. 6, pp. 1353–1366 (2018).

[11] Caelles, S., Maninis, K.-K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D. and Van Gool, L.: One-shot video object segmentation, *Proc. Computer Vision and Pattern Recognition (CVPR)*, pp. 221–230 (2017).

[12] He, K., Gkioxari, G., Dollár, P. and Girshick, R.: Mask r-cnn, *Proc. International Conference on Computer Vision (ICCV)*, pp. 2961–2969 (2017).

[13] Ren, S., He, K., Girshick, R. and Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks, *Proc. Neural Information Processing Systems (NIPS)*, pp. 91–99 (2015).

[14] Everingham, M., Gool, V., Williams, C. K., Winn, J. and Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge, *International Journal of Computer Vision*, p. 303–338 (2010).

[15] Kuhn, H. W.: The Hungarian method for the assignment problem, *Naval research logistics quarterly*, Vol. 2, No. 1-2, pp. 83–97 (1955).

[16] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft coco: Common objects in context, *Proc. European Conference on Computer Vision (ECCV)*, pp. 740–755 (2014).