基本行動特徴量を用いたオンライン複数人物追跡
Online Multiple Human Tracking using Primitive Action Features

西村仁志 †‡ 田坂和之 ‡ 川西康友 †‡ 村瀬洋 †‡
Hitoshi Nishimura†‡ Kazuyuki Tasaka‡ Yasutomo Kawanishi†‡ Hiroshi Murase†‡
† 株式会社 KDDI 総合研究所 ‡ 名古屋大学
†KDDI Research, Inc. ‡Nagoya University

Abstract: In this paper, we propose a online multiple human tracking method using primitive action features. The proposed method extracts rich action features from a single frame by utilizing a global context such as nearby objects and humans. Online data association is performed using Hungarian algorithm. In the experiments, we verified the effectiveness of the proposed method using the Okutama-Action dataset. Our code is available online (`https://github.com/hitottiez/mht-paf`).

1 はじめに

複数人物追跡は，ロボティクス・サーベイランス・マーケティング等，様々な分野で用いられる基礎的な技術である．従来より複数人物追跡手法 [1] は提案されているが，ドローンによる空撮映像に適用した場合，人物の大きさやアスペクト比が急激に変化するため ID スイッチが多発する．この問題を解決するため，我々は基本行動特徴量を用いた複数人物追跡手法を提案している [2]．MHT-PAF [2] では，近傍の物体や人物のような空間的なコンテキスト情報を考慮することで，単一フレームからでも人物追跡に有用な行動に関する特徴量が抽出可能となっている．しかし，オフラインでの人物追跡しか行っておらず，全フレームから特徴量を抽出した後でなければ人物追跡結果が算出できない．

本論文では，基本行動特徴量を用いてオンラインで複数人物追跡を行う手法を提案する．図 1 に提案手法の概要図を示す．各フレームに対して人物検出を行った後，基本行動特徴量を含めた各種特徴量を抽出する．得られた特徴量を用いて隣接フレーム間で逐次的に人物対応付けを行うことで，オンライン人物追跡が実現される．

2 特徴抽出

人物追跡のための特徴量として，位置特徴量，見え特徴量，そして基本行動特徴量を抽出する．

位置特徴量：フレーム f において人物検出を行うことで，m 番目の位置特徴量 $\mathbf{b}_f^m$ を抽出する [2]．

見え特徴量：フレーム f において，$\mathbf{b}_f^m$ 領域内から見え特徴量 $g(\mathbf{b}_f^m)$ を抽出する [2]．

基本行動特徴量： フレーム f において，各 $\mathbf{b}_f^m$ 領域内から基本行動特徴量 $\mathbf{a}_f^m$ を抽出する．図 2 に基本行動特徴量抽出モデルを示す．モデルは 4-stream のニューラルネットワークで，RGB 画像とオプティカルフロー画像を入力とした 2-stream ネットワーク [3] に基づいている．

各モダリティ（RGB と FLOW）について，局所的切り出し画像と大域的切り出し画像を入力とする．局所的切り出し画像は，バウンディングボックスを長辺に合わせて正方形にリサイズして得られる．大域的切り出し画像は，局所切り出し画像の領域を μ 倍にしてから切り出すことによって得られる．大域的切り出し画像によって，近傍の物体や人物のような空間的なコンテキスト情報が利用できる．

また，ネットワークの出力を複数ラベルとすることで，より追跡に有用な行動情報を抽出できる．損失関数には行動ごとの二値交差エントロピーを用いる．

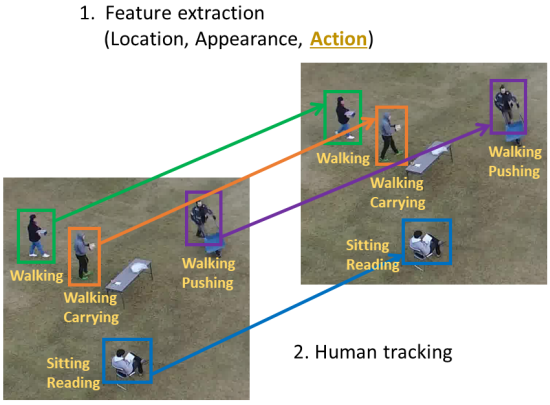

図 1: 提案手法．基本行動特徴量を含めた各種特徴量を抽出した後，隣接フレーム間で逐次的に人物対応付けを行う．

3 人物追跡

各種特徴量を用いて，オンラインで隣接 2 フレーム間の人物対応付けを行う．対応付けは以下の目的関数を最小化することによって行う．

$$\min_{\mathbf{v}_f} \sum_n \sum_m c_{f-1,f}(n,m) \quad s.t.\ \forall n, \forall m, v_f^n \neq v_f^m. \qquad (1)$$

目的関数の最小化は，Hungarian 法によって行う．Hungarian 法による対応付けは，下記の三種類のコスト（$c1, c2, c3$）を用いて三段階で行う．

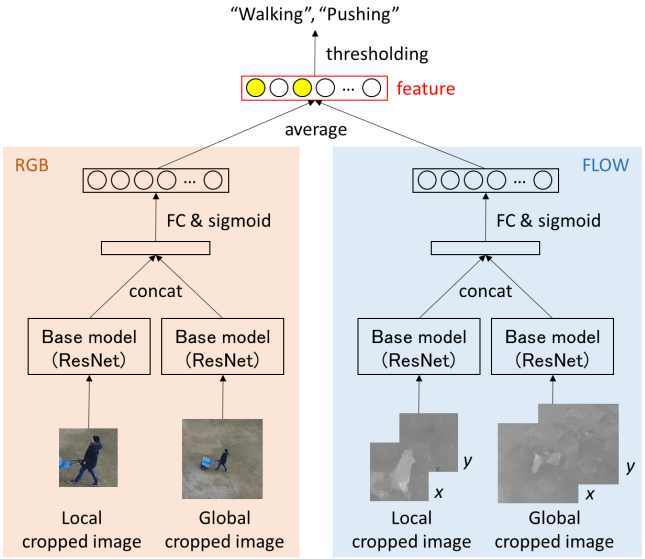

図 2: 基本行動特徴量抽出モデル．

一段階目の対応付けでは，フレーム $f-1$ における n 番目の位置特徴量 $\mathbf{b}^n_{f-1}$ と，フレーム f における m 番目の位置特徴量 $\mathbf{b}^m_f$ との間のコストを以下のように表す．

$$c1_{f-1,f}(n,m) = -s(\mathbf{b}^m_{f-1}) - d(\mathbf{b}^n_{f-1}, \mathbf{b}^m_f) - s(\mathbf{b}^m_f). \quad (2)$$

ここで，$s(\mathbf{b})$ は領域 $\mathbf{b}$ に対する人物検出器の最終層スコアを示す．$d(\mathbf{b}^1, \mathbf{b}^2)$ は二つの人物領域間の重複率（IoU）を示し，積領域/和領域で表される．

次に，一段階目で対応付けられなかった人物を対象に，以下の見え特徴量に基づくコストを用いて対応付けを行う．

$$c2_{f-1,f}(n,m) = 1 - \cos(g(\mathbf{b}^n_{f-1}), g(\mathbf{b}^m_f)). \quad (3)$$

そして，二段階目で対応付けられなかった人物を対象に，以下の基本行動特徴量に基づくコストを用いて対応付けを行う．

$$c3_{f-1,f}(n,m) = 1 - \mathrm{softmax}(t(\mathbf{a}^n_{f-1}) \cdot t(\mathbf{a}^m_f)). \quad (4)$$

ここで，$t(\mathbf{a})$ は行動特徴量抽出モデルの最終層のうち，最大となる行動スコアを示す．こうして，$\mathbf{b}_f$ 中のそれぞれに対する人物 ID $\mathbf{v}_f = (v^1_f, v^2_f, \cdots)$ が得られる．以上の処理を毎フレーム行うことで，オンライン人物追跡が実現される．

4 実験

ドローンからの空撮映像で構成される Okutama-Action データセット [4] を用いて，提案手法の有効性を検証した．データセットは 43 個の動画で構成され，33 個は学習データ，10 個はテストデータとして使用した．映像は 30fps で撮影され，解像度は 4K $(3,840 \times 2,160)$ である．各バウンディングボックスに対して，1 つ以上の行動ラベルが付与されている．

表 1: 人物追跡精度．

	Recall (%) ↑	Precision (%) ↑	ID switch ↓	Fragment ↓
DeepSORT	36.47	70.40	756	2324
提案手法	35.45	70.78	701	2225

表 1 に人物追跡精度を示す．DeepSORT [1] と比較して，提案手法は Recall・Precision を同等に保ちながら，ID スイッチが 55 回削減された．また，追跡が途切れるフラグメントについても 99 回削減された．

5 おわりに

本論文では，基本行動特徴量を用いてオンラインで複数人物追跡を行う手法を提案した．近傍の物体や人物のような空間的なコンテキスト情報を考慮することで，単一フレームからでも人物追跡に有用な行動に関する特徴量が抽出可能となる．隣接フレーム間の人物対応付けには Hungarian 法を用い，オンライン人物追跡を実現した．実験では，Okutama-Action データセットを用いて，Recall・Precision を同等に保ちながら，ID スイッチが 55 回削減されたことを確認した．今後は，他のデータセットを用いて提案手法の有効性を検証する．

参考文献

[1] N. Wojke *et al.*, "Simple online and realtime tracking with a deep association metric," in *Proc. ICIP*, 2017, pp. 3645–3649.

[2] H. Nishimura *et al.*, "Multiple human tracking using multi-cues including primitive action features," in *arXiv:1909.08171*, 2019.

[3] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. ECCV*, 2016, pp. 20–36.

[4] M. Barekatain *et al.*, "Okutama-action: an aerial view video dataset for concurrent human action detection," in *Proc. CVPR Workshops*, 2017, pp. 28–35.

株式会社 *KDDI* 総合研究所
〒*356–8502* 埼玉県ふじみ野市大原 *2* 丁目 *1* 番 *15* 号
Phone: 070-3623-9911
E-mail: ht-nishimura@kddi-research.jp