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

A Study on Online Spatiotemporal Action Detection Based on Human Detection and Action Recognition

西村 仁志†‡　田坂 和之†　川西 康友†‡　村瀬 洋†‡
Hitoshi NISHIMURA†‡ Kazuyuki TASAKA† Yasutomo KAWANISHI†‡ and 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 novel online spatiotemporal action detection method. The proposed method is based on human detection and action recognition, and can detects humans even if their actions change. In the experiment, we showed an example superior to the existing method using UCF101-24 dataset.

1. はじめに

カメラ映像中に写る人物を対象として，「誰が，いつ，どこで，何をしているか」を推定する時空間行動検出は，スポーツ，対話ロボット，サーベイランス等，様々な分野で活用されている．本研究では，リアルタイムでのスポーツ実況やロボット対話実現のため，オンラインでの時空間行動検出を目指す．

あるフレームfにおいて，N人の人物が行っている真の行動を，$T_f = (t_f^1, t_f^2, \cdots, t_f^{N_f})$で表す．ただし，$t_f^n = (b, g, l, a)$とする．$b = (x, y, w, h)$は人物位置を示す．$x, y$はそれぞれフレーム$f$における，左上の点の x 座標， y 座標を示し，w, hはそれぞれ，そのフレーム中での，幅，高さを示す． gは，実世界の人物を一意に表すグローバル人物 ID を示す．lは，人物に対して追跡処理中に仮に付与されたローカル人物 ID を示す．aは，行動クラスを示す．オンライン時空間行動検出とは，時系列の画像フレームが一枚ずつ順番に与えられたときに，順次$T_f = (t_f^1, t_f^2, \cdots)$を推定する問題である．

これまでにも，いくつか時空間行動検出手法は提案されてきた [1, 2]．T-CNN [1]では，オフラインで人物位置bを推定した後，行動クラスaを推定する．ROAD [2]では，オンラインでb, aを同時に推定可能である．しかし両者ともに，ある“行動”に着目して人物位置bを推定するため，人物の行動が時間的に変化すると，別の人物として扱われてしまう．つまり，“人物”に着目したローカル人物 ID が推定できない．そこで本論文では，人物検出と行動認識を統合し，行動が時間的に変化した場合でも，同一人物として認識可能な手法を提案する．

2. 提案手法

提案手法では，人物検出と行動認識を統合して，オンラインで人物行動検出を行う．

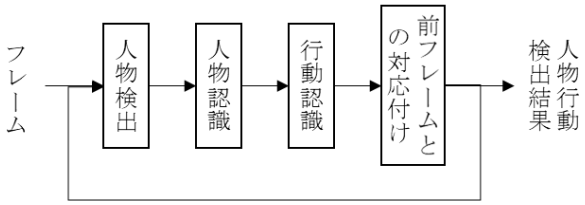


図１．提案手法の流れ．

2.1. 人物検出

フレームfから人物検出を行い，人物位置$b_f = (b_f^1, b_f^2, \cdots)$を推定する．人物検出には，SSD [3]を用い，事前に検出モデルを学習しておく．

2.2. 人物認識

2.1 節で得られた$b_f = (b_f^1, b_f^2, \cdots)$について，それぞれ人物認識を行い，グローバル人物 ID$g_f = (g_f^1, g_f^2, \cdots)$を推定する．具体的には，まずフレーム f から SSD [3]を用いて顔検出を行う．次に，顔検出結果と，2.1 節で得られた人物位置との間で，Hungarian 法によって対応付けを行う．対応付けができなかった顔検出結果は削除する．対応付けができなかった人物位置に対しては，グローバル人物 ID を“Noface”等と付与する．そして，得られた顔領域から特徴量を算出し，あらかじめ収集しておいた各人物の顔特徴量と照合し，その結果をグローバル人物 ID とする．照合には k 近傍法を用いる．

2.3. 行動認識

2.1 節で得られた$b_f = (b_f^1, b_f^2, \cdots)$に該当する領域を，それぞれフレーム$f$から切り出す．各切り出し画像に対して，TSN [4]によって行動認識を行い，行動クラス$a_f = (a_f^1, a_f^2, \cdots)$を推定する．行動認識スコアが一定の閾値よりも小さい場合は，“Unknown”とする．

2.4. 前フレームとの対応付け

2.1～2.3節の結果を基に，フレーム間での人物の対応付けを行う．対応付けは三段階で行う．一段階目の対応付けについて，フレーム$f-1$におけるn個目の人物位置b_{f-1}^{n}と，フレームfにおけるm個目の人物位置b_{f}^{m}との間のコストを以下のように表す．

$$c1_{f-1,f}(n,m) = -s(b_{f-1}^{m}) - d(b_{f-1}^{n}, b_{f}^{m}) - s(b_{f}^{m}).$$

$s(b_f)$は，2.1節の人物検出スコアを示す．$d(b_{f-1}, b_f)$は，bの領域間のIoU (Intersection over Union)を示す．IoUとは二つの人物矩形間の重複率を意味し，積領域/和領域で示される．

次に，一段階目で対応付けられなかった人物を対象に，以下のコストを用いて二段階目の対応付けを行う．

$$c2_{f-1,f}(n,m) = 1 - cos(f(b_{f-1}^{n}), f(b_{f}^{m})).$$

$f(b_f^m)$は，b_f^m領域内で抽出した特徴量を示し，特徴量は，WideResNet [5]によって算出する．

そして，二段階目で対応付けられなかった人物を対象に，以下のコストを用いて三段階目の対応付けを行う．

$$c3_{f-1,f}(n,m) = 1 - \mathrm{softmax}\left(s(a_{f-1}^{m}) \cdot s(a_{f}^{m})\right).$$

$s(a_f)$は，2.3節の行動認識スコアを示す．

上記のコスト$c1, c2, c3$を用いた人物の対応付けは，以下の目的関数を最小化することでによって行う．

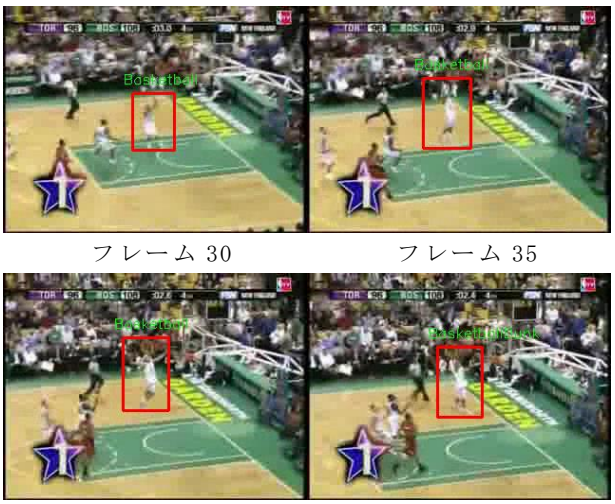
$$\min_{l_f} \sum_{n} \sum_{m} c_{f-1,f}(n,m) \quad s.t. \forall n, \forall m, l_f^n \neq l_f^m.$$

目的関数の最小化は，Hungarian 法によって行う．こうして，b_f中のそれぞれに対するローカル人物 ID $l_f = (l_f^1, l_f^2, \cdots)$が得られる．

3. 実験

UCF101 [6]のサブセットである UCF101-24 を用いて，提案手法の精度評価を行った．学習データには，ランダムに選択した 2290 動画を使用した．人物検出モデルには，VOC データセットで学習された公開モデルを用いた．行動認識モデルは，UCF101-24 の画像から，人物領域のみを切り出したものを入力として学習した．フレーム間の対応付けに用いる特徴量は，WideResNet の公開モデル [5]を用いて抽出した．フレーム間の対応付けの後，過去Mフレーム分の行動認識スコアを行動クラスごとに平均し，平均値が最大となる行動クラスを，そのフレームにおける推定結果とする．Mは実験的に 5 に設定した．

図２に，提案手法による人物行動検出結果の一例を示す．同じ赤い人物領域は，同じローカル人物 ID と推定されたことを示している．初めは行動クラスが “BasketBall”と推定されているが，ダンクシュートの瞬間は “BasketBallDunk”と推定されている．このように行動が時間的に変化した場合でも，同一人物として推定されている．従来手法 [1, 2]では，このように時間的に行動が変化した場合は，別の人物として扱われる．



フレーム 30　　フレーム 35

フレーム 40　　フレーム 45

図２：提案手法による人物行動検出結果の一例．

4. まとめ

本論文では，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を提案した．実験では，提案手法によって，行動が時間的に変化した場合でも，同一人物として認識される例を示した．今後は，テストデータの数を増やして精度評価を行う．また，さらなる精度向上のため，フレーム間の対応付けに人物認識結果の使用を検討する．

文　献

[1] R. Hou, C. Chen, M. Shah: “Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos”, *ICCV*, pp.5823-5832 (2017)

[2] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin: “Online Real-time Multiple Spatiotemporal Action Localisation and Prediction”, *ICCV*, pp. 3657-3666 (2017)

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg: "SSD: Single Shot Multibox Detector", *ECCV*, pp. 21–37 (2016)

[4] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool: "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition", *ECCV*, pp. 20-36 (2016)

[5] N. Wojke, A. Bewley, and D. Paulus: “Simple Online and Realtime Tracking with a Deep Association Metric”, *ICIP*, pp. 3645-3649 (2017)

[6] K. Soomro, A. R. Zamir, and M. Shah: "UCF101: A dataset of 101 human action classes from videos in the wild", Technical report, CRCV-TR-12-01 (2012)

†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

E-mail: ht-nishimura@kddi-research.jp