7ZE-05

手の周辺領域の分析による 人物と物体のインタラクション検出

小西 陸斗 †1　阿部 亨 †2,†3　菅沼 拓夫 †2,†3
†1 東北大学工学部電気情報物理工学科
†2 東北大学大学院情報科学研究科　†3 東北大学サイバーサイエンスセンター

1　はじめに

近年，人物と物体のインタラクション検出は，監視・防犯，人物の行動分析など様々なアプリケーションでの応用が期待されている [1,2]．その中でも，人物が手で物体を動かす動作の検出は特に重要である．

これに対し，人物の手の周辺領域の状態に基づいて人物が手で物体を動かす動作を検出する手法が提案されている．この手法では，従来の手法で必要であった「物体の抽出」や「手と物体との関連の認識」が不要であるため，より多様な状況に適用できる等の利点を有する．しかし，手の周辺領域の状態を獲得する処理や，得られた状態からインタラクションの有無を判定する処理で，各種パラメータと閾値の設定を適切に行う必要があった．

本稿では，手の周辺領域の状態から人物と物体のインタラクション検出を行う Tsukamoto らの [3] の手法に着目し，手の周辺領域の状態を獲得する処理や，得られた状態からインタラクションの有無を判定する処理で必要とされるパラメータと閾値を，機械学習を用いて適切に設定し，安定・高精度なインタラクション検出の実現を図る手法を提案する．

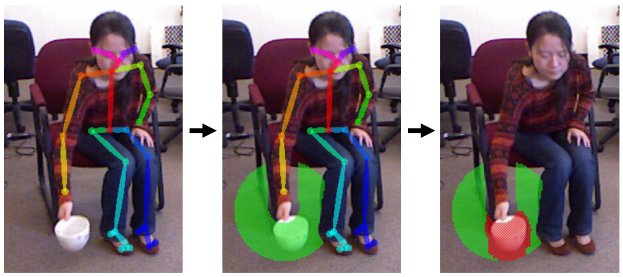
2　関連研究

Tsukamoto らは，画像から抽出した人物の骨格に基づき，前腕領域および手の周辺領域を設定し，周辺領域での動きを分析することで，前腕の動きとの関連性から，人物が手で物体を動かすインタラクションを検出する手法を提案している [3]．

この手法における処理の概要を図 1 に示す．

まず，図 1(a) に示すように，OpenPose [4] を用い各画像で人物の骨格を抽出する．次に，図 1(b) に示すように，骨格および骨格をパラメータ（幅）W で膨張させ求めた人物領域に基づき，矩形として近似した前腕の領域 F と，手を中心として設定した円から人物領域と重なる部分を除去した周辺領域 S を設定する．

次に，前腕領域 F の各画素で求めたオプティカルフローから前腕の動きを推定し，周辺領域 S の各画素 p_i でもオプティカルフロー v_i を求める．推定した前腕の動きから，前腕で物体を動かした場合に p_i で観測されるはずのオプティカルフロー ve_i を予測し，実際に観測された v_i と比較する．観測された v_i と予測された ve_i の差が閾値 α 以下 ($\|v_i - ve_i\|/\|ve_i\| \leq \alpha$) となる p_i（図 1(c) の赤い箇所）は，前腕が動かす物体に対応すると考えられるため，その総数 N を求め，S の総画素数 $|S|$ に対する割合が閾値 β 以上 ($N/|S| \geq \beta$) ならば，S 内に前腕が動かした物体が含まれていると見なし，インタラクションが生じていると判定する．



(a)　(b)　(c)

図 1: Tsukamoto らの手法における処理の概要

以上のように，この手法では，画像から物体を抽出し，手との関連を認識する処理が不要であり，物体が未知である場合や，一部遮蔽された場合でもインタラクションの検出が可能となる．しかし，

(P1)　前腕領域 F，周辺領域 S の設定に必要となる人物領域を決定するためのパラメータ W

(P2)　周辺領域 S 内に前腕が動かした物体が含まれるかを判定するために必要となる閾値 α, β

が必要となり，これらの値が適切に設定されない場合，人物と物体のインタラクションを安定・高精度に検出することが困難となる．

Detecting human-object interactions by analysis of areas surrounding hands
Rikuto Konishi[†1], Toru ABE[†2,†3], and Takuo SUGANUMA[†2,†3]
[†1]Department of Electrical, Information and Physics Engineering, Tohoku University
[†2]Graduate School of Information Sciences, Tohoku University
[†3]Cyberscience Center, Tohoku University

3 提案手法

本稿では，Tsukamoto らの手法に着目し，従来は経験的に設定されていた処理中のパラメータと閾値を機械学習を用いて適切に設定することで，安定・高精度なインタラクション検出の実現を図る．具体的には，前述の (P1)，(P2) に対し，以下の (S1)，(S2) を各々適用する．

(S1) Tsukamoto らの手法では，OpenPose を用い画像から抽出した人物の骨格をパラメータ W で膨張させることで人物領域を決定し，前腕領域 F，周辺領域 S の設定に用いていた．近年，Detectron2 [5] など，人物の骨格だけでなく，人物領域も高精度に検出可能な手法が開発されている．そこで，Detectron2 等により人物領域に抽出を行い，その抽出結果に基づき F を設定することで，前腕の動きのより正確な推定を行う．また，手を中心として設定した円から人物領域を除去する際にも，Detectron2 等による抽出結果を利用することで，S のより適切な設定を行う．

(S2) 周辺領域 S 内の各画素 p_i が前腕で動かす物体に対応するか否かを判定するための閾値 α（p_i で観測されたオプティカルフロー v_i と予測されたオプティカルフロー ve_i の差に対する閾値）は，Support Vector Machine (SVM) [6] 等を用いることで，事前に用意したサンプルから機械学習により適切な値を設定する．また，前腕で動かす物体に対応すると判定された S 内の画素の割合でインタラクションの検出を行う際の閾値 β についても，SVM 等を用いることで，サンプルから機械学習により適切な値を設定する．さらに，前腕で動かす物体に対応する画素の S 内での分布を特徴量として用いることで，インタラクションのより安定・高精度な検出が可能になると考えられる．

4 評価方法

提案手法の評価は以下の方法で行う．

- **使用データ**
  一般公開されている CAD-120 [7] 等のデータセットや，新たに撮影した映像を，パラメータ設定のための学習用データ及び評価用データとして用いる．

- **評価基準**
  実験に使用する全データ（全画像）で，インタラクションの有無，インタラクションが生じている場合はその位置（手の位置）を正解として用意し，インタラクションが生じているフレーム中での位置を正しく検出する精度で評価を行う．また，手で動かしている物体の状況（物体の種類，遮蔽の状態など）で検出精度を比較し，提案手法が有効となる状況の分析を行う．

(a) 入力画像　(b) 人物領域の抽出結果

図 2: Detectron2 による人物領域の抽出例

5 おわりに

本稿では，人物の手の周辺領域の分析によりインタラクションを検出する手法に関し，分析の際に必要となるパラメータを機械学習で適切に設定するための検討を行った．今後，提案手法を実装し，実際の映像を用いた実験により，インタラクション検出の安定性・精度の向上に対する効果を検証する予定である．

参考文献

[1] Chao, Y. et al.: Learning to detect human-object interactions, *IEEE winter Conf. Appl. Comput. Vision.*, pp. 381–389 (2018).

[2] Gkioxari, G. et al.: Detecting and recognizing human-object interactions, *IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 8359–8367 (2018).

[3] Tsukamoto, T. et al.: A method for detecting human-object interaction based on motion distribution around hand, *15th Int. Conf. Comput. Vision Theory Appl.*, pp. 462–469 (2020).

[4] Cao, Z. et al.: OpenPose: Realtime multi-person 2D pose estimation using part affinity fields, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 43, No. 1, pp. 172–186 (2021).

[5] Wu, Y. et al.: Detectron2, `https://github.com/facebookresearch/detectron2` (2019).

[6] Pedregosa, F. et al.: Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, Vol. 12, pp. 2825–2830 (2011).

[7] Koppula, H. S. et al.: Learning human activities and object affordances from RGB-D videos, *Int. J. Rob. Res.*, Vol. 32, No. 8, pp. 951–970 (2013).