H-021

混雑した状況での人物と物体のインタラクション検出に関する研究
Detecting human-object interactions in crowded environments

三橋 優人† 阿部 亨‡† 菅沼 拓夫‡†
Yuto Mitsuhashi Toru Abe Takuo Suganuma

1. はじめに

従来，映像や画像から人物や物体を検出する手法に関し様々な研究が行われてきた．近年は，状況の監視やモニタリング等への応用を目指し，人物と物体の検出を組み合わせて行動認識を行う研究が特に盛んに進められている．しかし，環境内に多数の人物が存在する混雑した状況では，各人物と各物体の関連を特定し，両者の様々なインタラクションを検出することは依然として困難な課題である．

この問題の解決を目指し，本稿では，入力された映像を階層的に解析し，個々の人物について得られた特定部位（前腕）の位置情報を利用することで，混雑した状況でも人物と物体のインタラクション（人物が物体を動かしている状態）を安定に検出するための手法を提案する．

2. 関連研究

人物と物体のインタラクション検出に関する従来の研究では，人物と物体をそれぞれ独立に検出した後で，両者を結び付け相互の関連を特定するものが多い．そのため，多数の人物が隣接しているような状況では，人物と物体の関連を特定することが困難となる．例えば，人物と物体の位置関係から両者を結び付ける手法[1]では，複数の人物が近接しているような場合，ある 1 つの物体に対するそれらの人物の位置関係は類似したものとなり，人物と物体の関連を一意に特定することは難しい．

また，人物とは独立に物体の検出を行う従来手法では，形状モデル等を用い特定の物体の検出するものが多い[2]．このような手法では，検出対象である物体を予め限定し，想定した対象物体の形状モデル等を事前に作成しておく必要がある．従って，インタラクションを検出可能な物体も限定されることになり，人物と任意の物体の様々なインタラクションを検出することは困難となる．

3. 提案手法

提案手法では，入力された画像を階層的に解析し，個々の人物について得られた前腕の状態に基づき人物と物体のインタラクションの検出（人物が物体を動かしている／いないの判定）を行う．これにより，人物と物体の位置関係だけでは両者の関連を一意に特定することが難しい場面でも，個々の人物の前腕の状態を用いてインタラクションの安定した検出を図る．また，インタラクションの対象となる物体が存在している可能性が高い箇所の範囲（前腕の近傍）を推定し，その範囲内で，前腕との関連が高い動きを示している物体を探索することにより，対象を特定の物体に限定しないインタラクション検出の実現を図る．

本提案手法の処理は 4 つの Stage で構成される．Stage 1 で個々の人物領域の抽出・頭部位置の決定を行い，Stage 2 で前腕領域の特定を行い，Stage 3 で物体の候補箇所を推定し，Stage 4 でインタラクションを検出する．提案手法の処理の流れを図 1 に示し，以下で各 Stage の詳細を述べる．

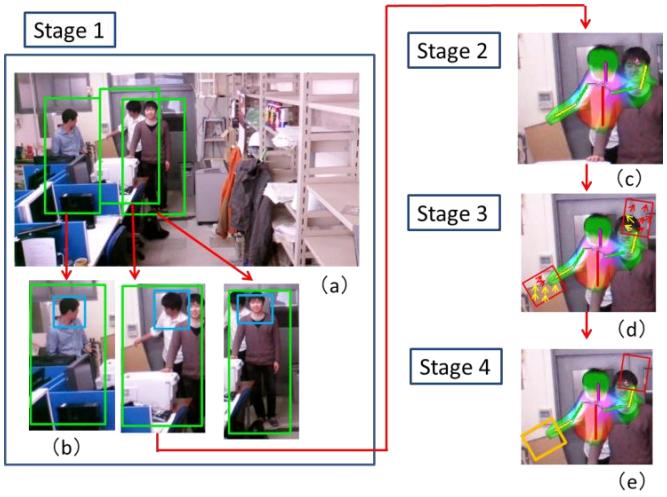


図 1．提案手法の処理の流れ： (a) 個々の人物領域の抽出，(b) 頭部位置の決定，(c) 腕領域の特定，(d) 物体候補箇所の推定，(e) インタラクションの検出

Stage1　個々の人物領域の抽出・頭部位置の決定

入力された時系列画像から，輝度勾配方向を複数のヒストグラムで表現した HOG 特徴量[3]を用いて個々の人物領域を矩形として抽出する．さらに，抽出した各人物領域から，輝度勾配の強度をスカラーで表現した Haar-like 特徴量[4]を用いて人物の頭部の位置を決定する．

Stage 2　前腕領域の特定

抽出された個々の人物領域に対し，人体構造モデルを用いて前腕領域の特定を行う．ここで用いる人体構造モデル[5]は，人体の各部位（左右の前腕・上腕，胴体）の位置関係と部位毎の色の均一性を確率で表現したものである．人体構造モデルを各人物領域へ当てはめることで人体の各部位を求め，その結果に基づき前腕領域を特定する．

以上のように，提案手法では，入力された画像を階層的に解析し，個々の人物の前腕の状態を獲得する．人物と物体とのインタラクションにおいて人物の前腕は中心的な役割を担うため，前腕の状態は，人物と物体の関連を混雑した状況で特定する際の効果的な情報となる．

Stage 3　物体候補箇所の推定

提案手法では，特定の対象物体を画像中で探索するのではなく，前腕近傍で物体らしい箇所（物体候補箇所）の推定を行う．推定された物体候補箇所をインタラクションの検出に用いることで，対象を限定しないインタラクション検出の実現を図る．

まず，前腕領域 F の先端位置，方向，面積から矩形の近

† 東北大学大学院情報科学研究科
Graduate School of Information Sciences, Tohoku University
‡ 東北大学サイバーサイエンスセンター
Cyberscience Center, Tohoku University

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

傍領域 N を設定し，F 内の各画素 p と N 内の各画素 q で動きベクトル（オプティカルフロー）$v(p)$，$v(q)$を計算する（図２）．各 q で $v(q)$と全 $v(p)$の相関の総和 $S(q)$を式(1)より求める．

$$S(q) = \Sigma_{p \text{ in } F} \mathrm{NCR}(v(p), v(q)) \qquad (1)$$

ここで，NCR($v(p)$, $v(q)$)は，$v(p)$と $v(q)$の正規化相互相関（normalized cross correlation）を表す．$S(q)$が閾値 Ts 以上の場合，q は前腕領域の動きと相関が高いと見做し，その人物と関連する（インタラクションの対象である）物体の候補箇所 q_{object} であると判定する．前腕領域に対し設定された近傍領域と推定された物体候補箇所（Ts=0.5）の例を図 3 に示す．

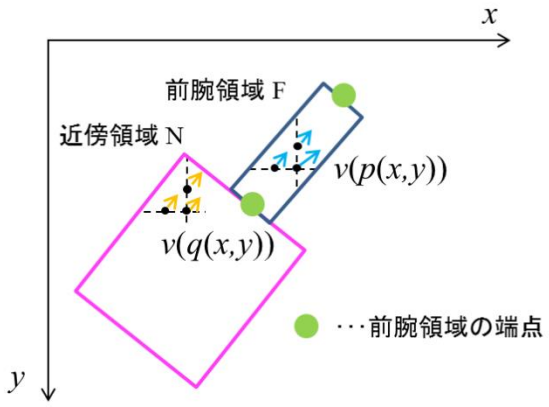
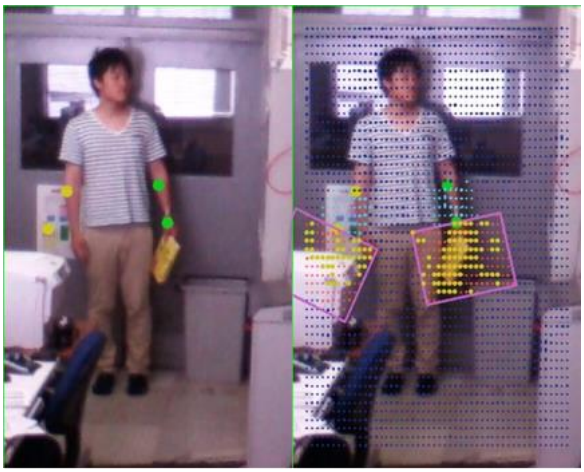


図２．領域内の動きベクトル



（a）　（b）

図３．(a) 前腕領域の先端・終端位置（●の箇所），(b) 近傍領域，前腕領域の動きベクトル（水色），近傍領域の動きベクトル（桃色），物体候補箇所（黄色）

Stage 4　インタラクションの判定

Stage 3 で物体候補箇所と判定した画素 q_{object} の集合から特徴量を抽出し，事前に学習した分類器に入力することでインタラクションの判定を行う．特徴量には，位置に関する特徴量（前腕領域 F の先端と物体候補箇所の重心間の距離，方向），動きベクトルに関する特徴量（$S(q_{\text{object}})$の平均，$\Sigma_{p \text{ in } F} |v(p)-v(q_{\text{object}})|$の平均），物体候補箇所の面積（$q_{\text{object}}$ の画素数）を用いる．なお，画像中の人物の大きさの影響を防ぐためにスケールに依存する特徴量に関しては人物領域の大きさで正規化を行う．分類器には，サンプル数の不足から生じる汎化誤差が比較的少ない SVM (Suppot Vector Machine)[6]を用いる．分類器は，時系列画像から抽出した特徴量をラベルつきで与え事前に学習を行う．特した特徴量をラベルつきで与える．画像中の人物の大きさの影響を防ぐためにスケール依存を受ける特徴量に関しては人物領域の大きさで正規化を行う．

4. 実験

時系列画像に提案手法を適用し，人物と物体のインタラクションを検出する実験を行った．

時系列画像は，USB カメラで撮影した 25 シーケンス（720×1280 画素，30ftp，全 5154 フレーム）を実験に使用した．この中で，インタラクション有りの状態は 4209 フレーム，無しの状態は 945 フレームである．実験では，時系列画像を 10 セットに分割し，1 セットをテストデータ，残り 9 セットを学習データに用いた交差判定法を適用した．なお，人体構造モデルの人物領域への当てはめには 2D articulated human pose estimation software v1.21 [7]を用い，インタラクション検出のための分類器は LIBSVM [8]を用いて作成した．

時系列画像の各フレームに対しインタラクションの有無を判定した結果，正解率は 82.29%であった（交差判定法による結果の平均）．この結果は，人物と物体のインタラクション検出に対する提案手法の有効性を示すものである．

5. おわりに

本稿では，入力された映像を階層的に解析し，人物の前腕領域と近傍領域の特徴を利用することで，混雑した状況でも人物と物体のインタラクション（人物が物体を動している状態）を安定に検出する手法を提案した．今後は，インタラクション検出の精度向上を図るため，前腕領域や物体候補箇所を高精度に特定する手法，インタラクション検出に用いる抽出特徴量について検討を進める予定である．

参考文献

[1] Y.L. Tian, et al., "Robust detection of abandoned and removed objects in complex surveillance videos," IEEE Trans. Syst., Man, Cybern. C, Vol.41, No.5, pp.1094-6977 (2011).
[2] A. Prest, et al., "Explicit modeling of human-object interactions in realistic videos," IEEE Trans. Pattern Anal. Mach. Intell., Vol.35, No.4, pp.835-848 (2013).
[3] N. Dalal, et al., "Histograms of oriented gradients for human detection," CVPR'05, Vol.1, pp.886-893 (2005).
[4] P.I. Wilson, et al., "Facial feature detection using Haar classifiers," J. Comput. Sci. Coll., Vol.21, No.4, pp.127-133 (2006).
[5] M. Eichner, et al., "Human pose estimation and search in (almost) unconstrained still images," ETH Zurich, D-ITET, BIWI, Tech. Report, No.272 (2010).
[6] F. Li, et al., "A Bayesian hierarchical model for learning natural scene categories," CVPR'05, Vol.2, pp.524-531 (2005).
[7] http://groups.inf.ed.ac.uk/calvin/articulated_human_pose_estimation_code/
[8] http://www.csie.ntu.edu.tw/~cjlin/libsvm./

Copyright © 2014 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.