Vol.2021-HCI-195 No.50
Vol.2021-UBI-72 No.50
2021/12/1

距離画像推定情報を用いた複数人の行動認識

豊坂 祐樹[1]　大北 剛[1]

概要：画像内におけるヒトとヒトが 3 次元的に接触しているか否かを知ることは, 現実空間における実際のヒトとヒトの行動を正確に認識するための手助けとなる可能性がある. 本論文においては, 画像の深度として推定した情報を下に画像内におけるヒトや物体の奥行きを再構築して, 姿勢推定を用いて画像内に認識したヒトとヒトの接触状態を検知するシステムを提案する.

Action recognition of multiple people using estimated information of distance image

YUKI TOYOSAKA[1]　TSUYOSHI OKITA[1]

1. はじめに

近年，複雑な動作や目的指向の会話を可能とするための研究が盛んにおこなわれている．さらに, ロボットに複雑な自律動作を行わせ，人間からの言葉に対して会話させることも多い．一方，行動認識という研究分野でいうと, 従来, 高々 50 程度の行動を分類することが多かったが，近年この要求は高まり，500 種類などの大量の行動を分類する要求されることも多くなった．このため, これら大量の行動を解像度良く認識できる, 複雑な世界認識が可能な人工知能システムが求められている．これらさまざまな行動においては, 人と人がどのように接触するのか, どの程度の時間接触を行うのか, 暖く接触しているのか冷く怒りをもって接触しているのかなどさまざまな要因が関係する．したがって, これらの要因を正しく認識することは, 大量の行動を正確に認識するために必要不可欠な技術である.

ロボットに目を移してみよう．ロボットはカメラなどの各種センサを設置することにより, 周りの環境や対象となる人物，モノなどの外界を常時センシングすることができる．行動計画を立てる際には，これらのセンシングした情報を更新しながら利用する．例えば，介護ロボットの場合, サービスする対象は一人ではなく, 複数人である可能性がある．この場合, 対象の人物と相互作用のある他の人物や物体の情報は, 未来の計画を立てる際に重要となる場合がある．特に，対象の人物と相互作用となる瞬間や逆に相互作用がなくなる瞬間を捉えられなければ, ロボットが動作すべきタイミングを逸することになる．病院では転倒防止のために体が不自由な人がベッドから降りると検知する場合，ベッドとの相互作用がなくなった瞬間を検知できなければ, ロボットの動作計画ができなくなる．このタイミングを正確に捉えることで, ロボットが起こすべき行動を正しく計画できることになる.

研究室環境では, ショート動画に対して行動認識を行うことが多い．少なくとも 2 つの問題点がある．1 つ目に, これらにおける行動ラベルは, 飲む，走る，歩くなどの重複のない行動であることが多い．現実的には, これらは重複した行動であることは往々にして起こる．このため, 重複を考慮した形の行動認識を行なってこそ，複雑な行動を認識可能となるはずである．2 つ目に, 不変の情報を認識する必要があることである．握手という行動において，握手をするタイミングや手が離れるタイミングは, 不変情報である．これらの動作を始点や終点を捉える形で行動をアノテーションすることにより, また，これらのタイミングを検知することにより, 人の行動の詳細の把握，行動の履歴の保存，動作の移り変わりを検出することになる.

そこで, 本研究は人と人 (人とモノ) が接触する瞬間と離れるタイミングと相互作用を考慮した上で行動を認識するシステムを構築した.

[1] 九州工業大学
Kyushu Institute of Technology

Vol.2021-HCI-195 No.50
Vol.2021-UBI-72 No.50
2021/12/1

本研究におけるわれわれの貢献は以下の通りである.

- より複雑なロボット等の行動計画に組み込むための人の動作の発生や終わりを検知可能な手法を提案したこと

また, 本論文の手法の有用性を以下の手順で示す.

(1) 物体検知と姿勢推定，深度マップによる相互作用検出方法の説明
(2) 構築した相互作用の発生する瞬間と終わる瞬間（接触する瞬間と離れる瞬間）の検出システムの有効性のテスト

有効性のテストでは，人と物体による「ボトルやコップを手で持つ」動作と人同士による「握手する」動作の二つに焦点を絞り，対象の接触する瞬間と離れる瞬間の検出が判定可能かどうかを調査した結果, 全体で 80.5%の正解率を達成（41 個の動画中 33 個正解）し, 検出ができることを確認した. (テストデータには Stair lab が公開している日常生活シーンをおさめた動作の動画 [2] の「shaking hands」,「drinking」動画を用いている)

2. 構築したシステムに使用したモデル

本研究では, 接触による人と人, または人と物体の相互作用の検知を実現するために姿勢推定と物体検知, 深度マップを組み合わせるシステムを構築した. そこで本節では, 人と人あるいは物体との接触による相互作用を判定するためのシステムに使用する姿勢推定システムと物体検知システム, 深度マップシステムについて説明するが，姿勢推定と物体検知に関しては，以前の研究 [1] でも使用した姿勢推定システムの openpose[3] と物体検知システム centernet[4] [5] をそれぞれ用いるので，深度マップについて説明する.

2.1 深度マップ

深度マップとは 2 次元画像から 3 次元空間の奥行きを表現するために，色の濃淡によって対象となる物体の距離（奥行き）を表した手法である．基本的に通常の画像は縦横の二次元座標で表現されるため奥行きがどれほどかは考慮されず，画像中の物体が重なっている場合には接触しているのか，あるいは奥行きで前後に存在していて接触していないのかの区別がつかない．そこで，深度マップを用いれば 3 次元的な位置の把握が可能となり，接触判定の精度が向上する.

本研究では google が開発し，公開している深度マップ生成手法 MannequinChallenge（マネキンチャレンジ）[6] を使用した．マネキンチャレンジは 2016 年頃に流行した約 2000 本の「人間は静止しているがカメラは動いている」というマネキンチャレンジのムービーをトレーニングデータとして利用しており，三角測量ベースであるマルチビューステレオ法（Multi View Stereo）を用い，入力データとして「RGB Image」,Mask R-CNN を使用して人間の領域をマスクした「Human mask」,optical flow を使用して計算された深度マップ「Initial depth from flow」の 3 つから深度マップを推定する．この手法では，カメラと対象の両方が自由に動くビデオから深度マップを生成しており，従来の深度マップ生成法である DORN[7] や DeMoN[8] では困難であった複雑な動作を伴う動画に対しても高い精度での深度マップを生成可能としている.

3. 相互作用の検出手法

この節では, 相互作用検出の構築方法について説明する. 本システムは人同士または人とモノの相互作用の検知を行い，動作としては「人が物を触れる・つかむ」動作,「人が物を食べる・飲む」動作,「人同士の握手」が検知可能となっている．まず，人同士の接触による握手の検知方法について説明し，その後，構築したシステムの概要と相互作用が接触する瞬間や離れる瞬間の検知方法について説明する.

3.1 人同士の接触による握手の検知方法

人と物体との相互作用検知は先行研究 [1] の手法を本システムでも使用するが，それに加え，人同士の接触による相互作用判定も追加した．人同士の接触の判定は姿勢推定と深度マップを使用することで実現した.「人同士の握手」に関しては，姿勢推定によって手を検出し，人の手の座標の重複により，人同士の手が接触しているかどうかを判定する.

3.2 構築したシステムの概要と相互作用判定方法

先行研究 [1] では，取得した人や物体の座標からそれぞれの領域が重複しているかどうかで判定したが，2 次元座標での判定だったので奥行きが考慮されず，精度に問題が生じた．そこで，本研究では 2 次元座標判定だったシステムに深度マップを加えて計算することで 2 次元座標＋対象位置の色の濃淡の 3 次元的な判定が可能となった.

構築したシステムの概要を図 1 に示す．まず，入力した画像に対して 3 つのモデルを適用して，画像内の手や顔の詳細な座標まで検知した人の姿勢推定と物体検知，画像の深度マップの作成を行う．次に検知した人や物体の 2 次元座標を統合し，座標が重複している人や物体を探し，重複している人や物体に対して深度マップを用いた相互作用判定を行い，結果を出力する．座標の重複の定義に関しては，画像における 2D 座標に加えて深度マップによる濃淡

Vol.2021-HCI-195 No.50
Vol.2021-UBI-72 No.50
2021/12/1

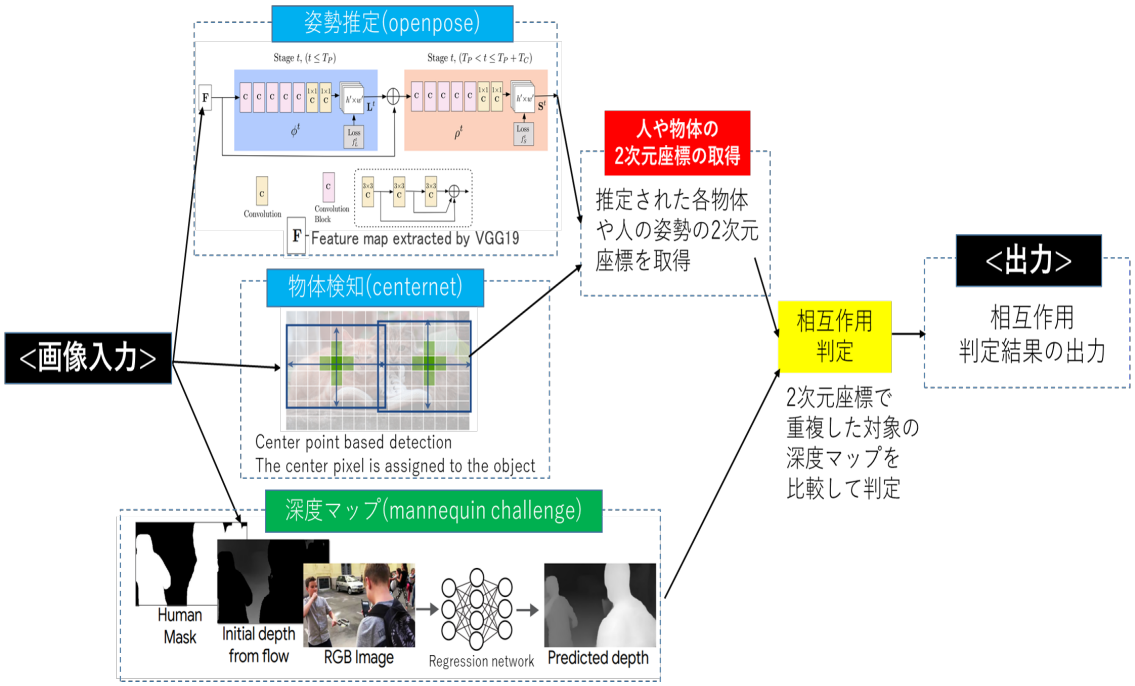

図 1 相互作用検知システムの概要

で表現される色の近さを基準に重複しているかを判定している．

相互作用の判定の具体例を図 2 に示す．図 2 の左上図は姿勢推定と物体検知によって重複判定された人の手が検出されており，左下図は上図の深度マップを表しているが，左上図で検知された重複判定された領域を左下図の深度マップからそれぞれの画素の平均値を算出する．その算出された 2 つの領域の画素の平均値をそれぞれ $[x_0, y_0, z_0], [x_1, y_1, z_1]$ とすると，相互作用の判定式は以下となる．

$$(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2 \qquad (1)$$

(1) 式は 2 つの領域の画素の差分の 2 乗和を表しており，その値が 81 未満の場合，相互作用があると判定される．81 という数値の理由に関しては，色の濃淡が数値として 0 から 255 で表現されるので，2D 座標で接触していると判定された対象同士の色が約± 5 以内である場合，接触していると判定するために誤差の 2 乗和が 81 未満の場合，接触していると定めた．

3.3 接触する瞬間や離れる瞬間の検知方法

相互作用検知はフレーム毎に独立に検出されるが，接触する瞬間と離れる瞬間の判定は時系列に並べた各フレーム毎における判定対象のフレームの前後のフレームの相互作用の状態を用いることで行われる．接触する瞬間と離れる瞬間の判定方法を以下の図 3 に示す．接触する瞬間の条件は，判定されるフレームの対象となる人や物体が前 5 フレームで相互作用なしの状態かつ，判定されるフレームとその後の 4 フレームが相互作用があると判定された場合となり，その時，対象の相互作用は「接触する瞬間である」と判定される．離れる瞬間の条件は，判定されるフレームの対象となる人や物体が前 5 フレームで相互作用ありの状態かつ，判定されるフレームとその後の 4 フレームが相互作用がなしと判定された場合に，対象の相互作用は「離れる瞬間」となる．基本的に，一つの画像から同時に別々の複数の相互作用が検知されることも当然ありうるが，その場合，各フレーム毎に検知されたそれぞれの相互作用の情報が保持され，それぞれに対して判定が行われる．

4. テスト

提案した人と物体の接触する瞬間や離れる瞬間の相互作用検出システムの有効性を確認するためのテストを行い，以下に述べるテストにおける分析及び考察を行う．

4.1 テスト方法

本研究では，人と物体による「ボトルやコップを手で持つ」動作と人同士による「握手する」動作の二つに対して，相互作用判定と各フレーム毎の相互作用の判定を用いた接触する瞬間や離れる瞬間の検出を行い，それらの精度について調査する．テスト条件は以下のものとする．

- テスト動画には stair lab が公開している日常動画 [9] の「drinking」「shaking hands」を用いる
- 動画は，「相互作用が存在し，離れる瞬間または接する瞬間が存在する」動画と「相互作用は存在するが，離れる瞬間または接する瞬間が存在しない」動画の 2 種

Vol.2021-HCI-195 No.50
Vol.2021-UBI-72 No.50
2021/12/1

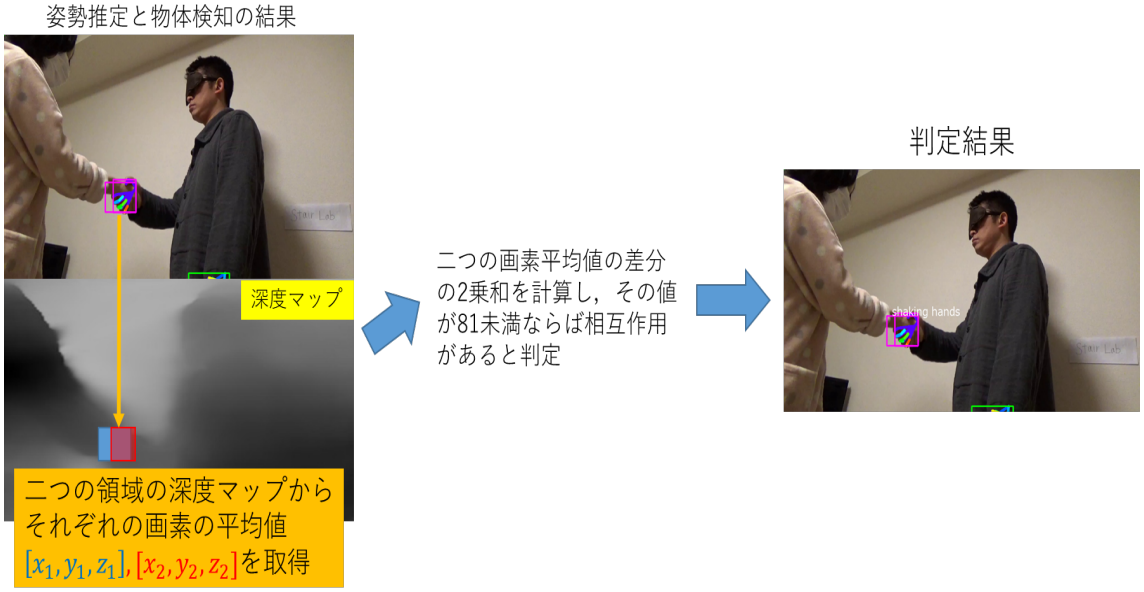

図 2 深度マップによる重複領域の相互作用判定

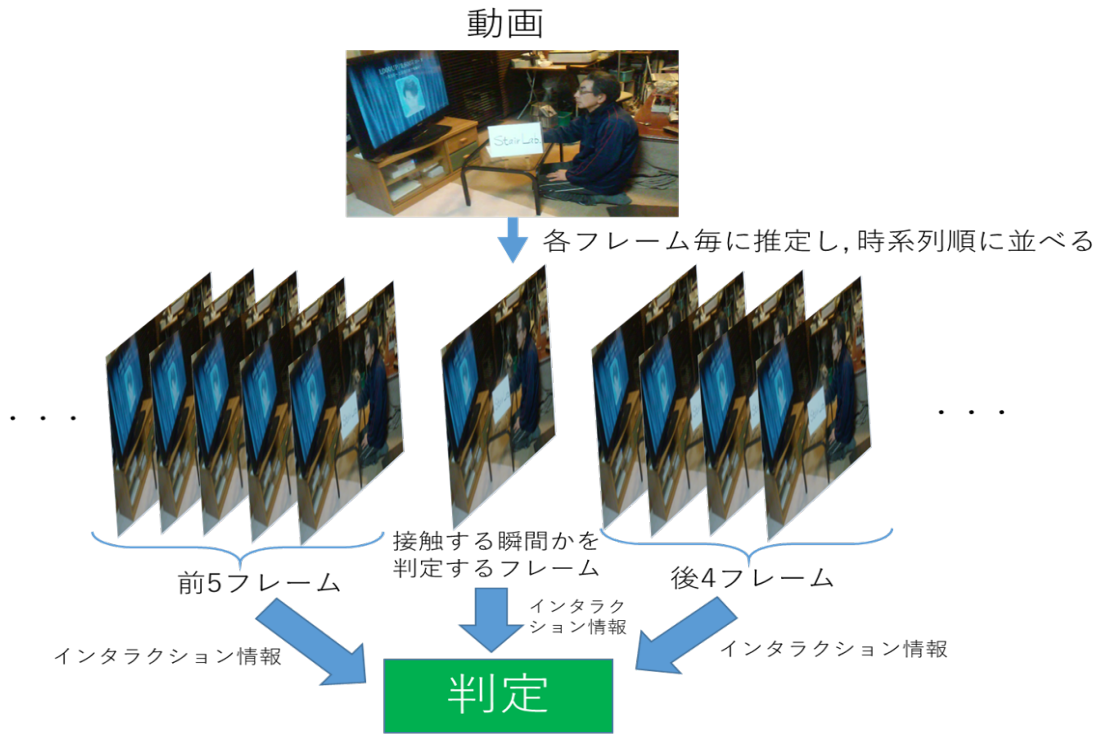

図 3 接触する瞬間や離れる瞬間の検出方法

類を用意する．(前者を positive 動画，後者を negative 動画として扱う)

- 動画数は 41（握手する動画：positive:10,negative:14, 手で物体を持つ動画：positive:7,negative:10）
- 動画の長さは 4 から 7 秒ほどのショート動画
- 評価方法は動画に対してシステムを適用し，接触する瞬間または離れる瞬間が検出できるかを調査し，主にその正解数によって精度を検証する．

条件にある「相互作用は存在するが，離れる瞬間または接する瞬間が存在しない」という negative 動画は例えば「握手する」動作の場合，最初から最後まで手を握っている動画，「手で物体を持つ」動作の場合，最初から最後まで物体を手に持っている動画のことを指し，接する瞬間も離れる瞬間も存在しない動画を選んだ．これらの条件の下でテストを行い，相互作用の検知と接する瞬間と離れる瞬間を検知できるかどうかを検証する．

Vol.2021-HCI-195 No.50
Vol.2021-UBI-72 No.50
2021/12/1

4.2 テスト結果と考察

相互作用の判定と相互作用が接する瞬間と離れる瞬間を検出するシステムのテスト結果の一例を図 4 に示す.

図 4 の上図は左から握手する動画における 25,26,27 フレーム目のデータである．25 フレーム目では相互作用は存在しないが，26 フレーム目に人同士の手が接触していることにより描画される「shaking hands」の文字とともに初めて発生した相互作用の瞬間を示す「インタラクション発生」というコメントが書き込まれており，27 フレーム目でも相互作用は存在するがこちらには「shaking hands」の文字のみとなっていることから相互作用の発生の瞬間を捉えることができているのがわかる．下図も同様に離れる瞬間を捉えたフレームの前後を並べた図となっており，58 フレーム目で手が離れ相互作用が終了したので「インタラクションが終了」と書き込まれているのがわかる.

各動画における相互作用の接する瞬間と離れる瞬間を検出した際の正解率を表 1 に示す.

表 1 相互作用の接する瞬間と離れる瞬間の検出の正解率

		予測結果	
		positive	negative
実	positive	70.6% (12/17)	29.4%(5/17)
画像	negative	12.5%(3/24)	87.5% (21/24)

negative 動画に対する正解率は 87.5%と比較的高い結果となったが，positive 動画は約 70%と 17%程低い結果となった．動画の種別で見ると，握手する動画に関しては negative 動画に対する正解率は 100%，positive 動画は約 80%と高い精度が得られたが，手で物体を持つ動画では，negative 動画に対する正解率は 70%，positive 動画は約 57%とかなり低い結果となった．手で物体を持つ動画の精度が低くなった要因としては，使用した動画が「drinking」動画であるため，飲む動作をするときにグラスやボトルを傾けた際に物体として認識できるフレームと認識できないフレームに分かれ，negative 動画でもその認識が途切れた際に相互作用が離れたと判定されたケースが存在した．握手する動画の精度が比較的高かったのは，握手という動作が手を多少動かす程度であるので動きが少なく，推定が比較的容易であったことが挙げられる.

5. おわりに

本研究では，人と人（人とモノ）との相互作用による接触する瞬間と離れる瞬間を検知することが可能なシステムを構築した．フレーム毎に相互作用の推定を行い，時系列的にどの瞬間から相互作用が発生，または終わるのかを検知することにより，非常に限定的なシチュエーションではあるが，人の動作の始まりと終わりを察知することに成功した．これにより，従来の行動認識のように一つの動画に対して，一つの行動を推定するだけではなく，変化し続ける人間の動作にある程度対応できる可能性を示した．相互作用検知の精度についても，以前の研究では画像から 2 次元座標のみを取得して相互作用を判定していたが，今回は深度マップによる 3 次元的な座標を用いることで，遠近法によって重なっているように見えるだけの誤検知を減らし，精度の向上を図った．しかし，今回のシステムは握手すると手をボトルやカップを持つという非常に限定された動作にしか対応していないので，対象となる動作の範囲を広げることを今後の課題として考慮する必要がある.

参考文献

[1] 豊坂祐樹, 大北剛, 複数の物体と手や顔の相互作用, 第 22 回日本知能情報ファジィ学会九州支部学術講演会,2020 年 11 月.
[2] Stair lab:A Large-Scale Video Dataset of Everyday Human Actions. 入手先 ⟨https://actions.stair.center/⟩（参照 2021-04-01）
[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh, OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, IEEE Transactions on Pattern Analysis and Machine Intelligence, July, 2019.
[4] Xingyi Zhou, Dequan Wang, Philipp Krhenbhl, Objects as Points, Computer Vision and Pattern Recognition, Apr 2019.
[5] xingyizhou:centernet(objects as points) . 入手先 ⟨https://github.com/xingyizhou/CenterNet⟩（参照 2020-04-15）
[6] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, William T. Freeman, Learning the Depths of Moving People by Watching Frozen People, CVPR, 2019.
[7] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In Proc. Computer Vision and Pattern Recognition (CVPR), 2018.
[8] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Doso vitskiy, and T. Brox. DeMoN: Depth and motion network for learning monocular stereo. In Proc. Computer Vision and Pattern Recognition (CVPR), 2017.
[9] STAIR Actions: A Video Dataset of Everyday Home Actions, Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi, arXiv, abs/1804.04326, 2018.

Vol.2021-HCI-195 No.50
Vol.2021-UBI-72 No.50
2021/12/1

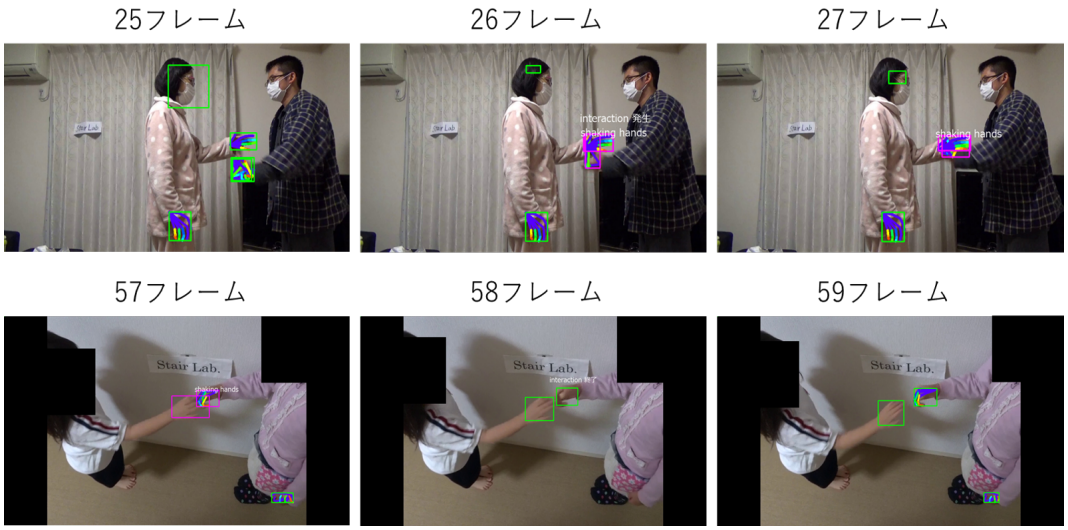

図 4 テストデータに対する結果の一例
(上図：接触する瞬間（26 フレーム目）を検知，下図：離れる瞬間（58 フレーム目）を検知)