手と物体の相互作用の認識とその知識グラフの生成

豊坂 祐樹[1]　大北 剛[1]

概要：知識グラフの形での行動述語表現としてのビッグデータは, スマートシティ, スマートホーム, ロボットの計画などにおいて中心的な役割を担うことが考えられる. これは, 行動述語より, これに付随した行動の主体や被主体が上位アプリケーションには大きな情報源となるからである. 本論文では, 手と手に関係する物体の相互関係を自動的に認識して, その関係を動作述語, 動作主体, 被主体という知識グラフとして記述するシステムを深層学習を用いて構築した. 特に, 人間の手の形状を足掛かりとして, 手と物体との近接性の判定アルゴリズムに手の形状を考慮するアルゴリズムとして構築した. Stair Lab の一部のデータセットを用いて検証し, 83%の正解率を達成した.

Perception and Knowledge Graph Generation: Interaction between Hand and Object

YUKI TOYOSAKA[1]　TSUYOSHI OKITA[1]

1. はじめに

スマートホームのように, IoT センサデバイスを用いて, 家の中をセンシングし, この情報を有用に用いる形の人工知能アプリケーションはスケールを変えて, さまざまな形で登場している. 漁場において魚群を探知するアプリケーション, 病院での手指消毒を探知するアプリケーション, 介護施設において転倒を探知するアプリケーション, 都市における人の流れを探知するアプリケーションなどさまざまなものが出て来た. 一般にセンサをベースとする行動認識とビジョンをベースとする行動認識の 2 通りの行動認識が知られている. いずれにおいても, 人や物体の動きを探知するという基本的な技術を若干応用した程度で実現可能な技術であった. 行動認識した情報を上位のアプリケーションから使用する場合, 現状の枠組みではさまざまな点で問題が出て来る. 仕様自体が不十分と考えられる問題である.

たとえば, ロボットがスマートホームを動き周り, 家の情報を得る場合, SLAM [3] のように 3 次元地図を構築するアプリケーションは可能である. しかし, ロボットが家の中でいくつかのタスクを実行するような計画を建てる場合, 行動認識の技術を用いて新たにセンシングした情報を, どのように計画に組込む情報とするかにおいて, 述語情報だけでは情報が非常に限られる. 1 つ目, 述語情報には動作主体と被主体の状態が記録されないため, たとえば,「花を贈呈する」などの動作を検知しても, 誰に贈呈したかの記録が必要となる. 2 つ目, 行動認識したいイベントは常に主体がロボットであるとは限らない. 述語動作を起こす主体の情報は行動認識では得られず, 上位アプリケーションで計画を練る場合には仕様不足となる. これを可能とするためには, 行動認識として術語情報を得る場面において, 動作主体, 被主体, 物体との関係など人と人の関係, 人と物体の関係もセンシングできる必要がある.

動作述語, 動作主体, 被主体などの関連する情報を認識するためには, まず複数物体の認識が必要とされる. 環境センサなどを用いてセンサベースの行動認識で複数物体の認識を行う場合 [5], 信号のオーバーラップなどが起こるため, 行動認識を行うこと自体が困難な状況であることを示した. Mairittha らは 3 軸加速度などのセンサをベースとする行動認識の場合, スマートフォンを用いて, 動作を入力し, 動作主体と被主体を入力する形の, 手動で記録する形で実現した. 一般にセンサをベースとする行動認識では主体の行動のみを記録し, 周辺との関係を記録することは非常に難しいがそれを克服したやり方である. 羽鳥ら [4] はビジョンベースの行動認識を用いて, 物体のピッキングタスクを実現している. ここでは複数物体の認識を行い, ターゲット

[1] 九州工業大学

とする被主体を選択して, ロボットを用いて動作を働き掛けている. ここでは, 自動の動作主体は一人称で変更されず, また動作述語もピックアップするという一意な非常に単純な例ではあるが, 動作主体, 動作述語, 被主体という組を認識している. 画像キャプションづけ [7] は, 画像から直接, 画像内の物体を記述する. しかし, 複数物体のインタラクションを目的としてはいないものの, 複数物体のインタラクションを記述するには困難である. 単体ではなく, 複数の物体という設定では, 自由度が大き過ぎることと, 未知の画像に対して遮蔽などの条件を推定することが困難であるからである.

以上, われわれの背景を述べたが, このような一般的なシステムを構築することは非常に難しそうである. 第 1 ステップで物体検知を用いて画像内に物体検出し, 第 2 ステップで検知された物体間に関係性があるかどうかを検証し, 第 3 ステップで関係性があれば知識グラフ的な動作述語, 動作主体, 被主体という表現を記述することが可能であろうと思われる. しかし, 2 物体間の関係を一般的に求めること, ひいては, 2 物体間の関係がないことを示すことは困難な場合が多い. そこで, 本論文においてはスコープを狭め, 手と近接物体とのインタラクションに焦点を絞った. なお, 今回スコープを狭めたが, 同じアルゴリズムでかなり広い行動を知識グラフに一般化することは可能と思う.

本論文におけるわれわれの貢献は以下の通りである.

- 物体検知システムと手を骨格レベルで検知できる姿勢検知システムを組合せることで幅広い複数物体の検知を行い, 物体の近接性の判定アルゴリズムにより, 概念的な動作主体と被主体の間のインタラクションを検知するシステムを構築したこと,
- インタラクションによる行動認識の足掛かりとして人間の手が物体に影響を与える際に頻繁に使用されることに注目し, 人間の手と物体との近接性の判定アルゴリズムに手の形状を考慮するアルゴリズムを構築したこと,
- インタラクションに関わった手と物体に対する動作主体と被主体の間のインタラクションを動作述語, 動作主体, 被主体という知識グラフへと変換する技術を構築したこと

である.

本論文の構成は以下の通りである. 第 2 節において物体検知として用いた Centernet[9] と Openpose[1] の概要について説明する. 第 3 節では, インタラクションの判定方法とインタラクションを用いた人の行動と物体を情報として得る方法, そして得た情報を文章として抽出する方法について述べる. 3.3 節においては, Stair lab で公開されている動画 [6] を静止画にした画像データをテストデータとして, 人の手とボトルのインタラクションの判定のテストを行う. また, その結果, 人が飲み物を飲む行動に対する画像に対して, われわれのシステムは手と手に持つボトルのインタラクションを判定することができ, この判定をベースとして「手にボトルを持つ」という知識グラフを構成できることを確認する. 第 4 節において, 結論と今後の課題について述べる.

2. 物体検知システムと姿勢推定システム

本節では, 人と物体のインタラクションを判定するためのシステムに使用する Centernet[9] と Openpose[1] について説明する. また, 本論文で述べるような認識レベルの話題が, 人工知能のアプリケーションから用いられる文献の代表として松尾 [10] の論文をレビューする.

2.1 物体検知システム

Centernet はヒートマップによるキーポイント検出ベースの物体検出手法である. 物体を boundingbox の中心点としてモデル化することで物体を検出するキーポイント推定によって中心点を探索するとともに, タスクに合わせて bounding box の大きさや 3D location, orientation, ポーズなどを回帰で推定する. ガウシアンカーネルを用いてヒートマップを生成し, そのヒートマップの特徴と bounding box（矩形領域）のサイズなどを学習する. Centernet のメリットはアルゴリズムがシンプルで計算が早く, 精度も他の手法とも遜色ない. そのうえ, 推定可能なオブジェクトの特性の幅が広く, 応用性がある.

2.2 姿勢推定システム

Openpose はリアルタイムの複数人の 2D ポーズ推定が可能となる人間の骨格抽出ソフトである. Openpose で提案された方法では, ノンパラメトリック表現（パートアフィニティフィールド（PAF）と呼ぶ）を使用して, 画像の身体部分と個人を関連付ける方法を学習している. このボトムアップシステムは, 画像の人物数に関係なく, 高精度とリアルタイムのパフォーマンスを実現している. 我々の研究では, より詳細な行動認識に必要となる体の部位（手・顔等）を検出するために Openpose を用いる.

Openpose のメリットは動画・静止画問わず簡単に画面に映る複数人の人間の骨格標本の抽出が可能な点と単純な全身の骨格だけではなく, 顔や手などの詳細な人体パーツの検出（手のひら・甲から顔の輪郭のパーツまで）も可能な点である.

2.3 認識と人工知能アプリケーションとの関係

センサなどで認識した情報を上位レベルで用いるアイデアはさまざまな論文で触れられている. このような中, 松尾はこれが古典的なシンボルグラウンディング問題と関係があることを説く [10]. 同時に深層学習による生成モデルで世界をシミュレートする方向へ向かうはずだとも説く.

この論文は深層強化学習などを用いて, ロボットなどにおける身体性を獲得していく上位部分に焦点を絞るが, 人工知能としての認識の果たす人工知能としてさまざまなのアプリケーションが本論文で述べる認識を土台として展開されることの重要性を解く側面ももつ. ただ, この論文では, 本論文で示すような現状の認識技術で欠けている部分については記述されていない.

3. 人と物体の相互作用の認識

複数の人や物体を検知し, 検知した物体から動作述語, 動作主体, 被主体などの関連する情報を認識しようとしたとき, セマンティックセグメンテーションなどの手法を用いてある程度汎用性のある認識システムを実現しようとしたならば, 膨大な種類のラベル付けとそれぞれのラベルごとに大量のデータが必要となる. なぜならば, 「立つ」, 「歩く」など人のみで完結する行動とは違い, 人が「何かの物体に触れる・持つ」や「料理をする」,「何かに乗る」などの人と物体が互いに影響を与える行動を画像から把握しようとした場合, 「物体を手に持っている」という行動一つとっても「皿を持っている」と「ボトルを持っている」では情報が異なる. そのため, 区別するためには「皿を持っている」と「ボトルを持っている」というラベルをそれぞれ付けて別のラベルデータとして分ける必要があり, セマンティックセグメンテーションでは「皿」「ボトル」「手」の個別の物体検知は可能でも, それらが影響し合う（インタラクション）行動を把握して文章として抽出するのは難しい. また, 画面内に複数の人間が存在するときにそれぞれの行動を検知するためには, 一枚の画像そのものに人の行動をラベル付けするようなデータではなく, 画像内のそれぞれの人物の行動にラベル付けするアノテーションデータが必要となる. 大量のアノテーションデータの作成には多大な労力がかかることから明確な問題点である.

それらを踏まえ, 本研究では人と物体の検知で詳細な人の行動や物体を把握し, 動作述語, 動作主体, 被主体などの情報取得が可能なアルゴリズムの構築を目指した. それを実現するのが, 人と物体のインタラクションの検知である. ただ人や物体を検知するだけではなく, 画面内に存在する人と物体とのインタラクション（相互影響）を判定することができれば,「より詳細な人間の行動の把握」や「その行動を文章として抽出する」など様々な応用が可能となる. セマンティックセグメンテーションとは異なるのは, たとえば「手でボトルを持っている」という行動に対して,「手でボトルを持っている」というラベル付けした画像データを学習させるのではなく, 人の手とボトルはそれぞれ単体として検出されるが両者にインタラクションがあると判定された場合, インタラクションにより「手でボトルを持っている」行動と判断される. さらに, 手に持たれているボトルと顔などの別のインタラクションが判定された場合, 顔とボトルのインタラクションにより, 「手で持っているボトルを飲んでいる」という複数のインタラクションによる詳細な行動の把握を行うことも可能である. また, 検出されたインタラクション自体（インタラクションがあるか否か）をラベルとして, トレーニングデータとして扱うこともできる. インタラクションによる行動認識のシステムには Centernet と Openpose を用いることで, インタラクションの判定に必要な複数の人間の詳細な体の部位と画面内に存在する複数の物体を一度に検知可能なシステムを構築した. この節では, 人と物体とのインタラクションを判定する手法とインタラクションによる行動認識（行動の文章化）, 並びに構築したインタラクションの判定システムに対する精度の検証について説明する.

3.1 人と物体のインタラクション判定方法

ここではまず人と物体のインタラクション判定を行うシステムの構築方法について説明する. システムの構築には 2 節で記載した Ceternet と Openpose を用いた. 図 1 における上図は Stair lab で公開されている動画 [6] を静止画にした画像データであり, 下図は上図に対して Centernet と Openpose によってそれぞれで検知された物体をキーポイントとして抽出し, それらのキーポイントから検知された物体を 1 枚の画像にまとめたものとなる. Openpose で詳細な骨格標本としての体の部位を検知し（図 1 では全体の骨格と手の詳細を検出している）, Centernet ではボトルや紙などの物体の検出を行った. これにより画像内の人の具体的な手の領域（位置）と物体の領域が抽出できる.

次に人の手と物体とのインタラクションの判定方法について説明する. インタラクション判定の概要を図 2 に示す. インタラクションの判定は人の手の領域と重複がある領域を持つ物体があればその物体は手とのインタラクションがあると判定される. 図 2 では中央部のボトルが手の領域との重複が認められたためインタラクションがあると判定され, 他の検知されたボトル, 紙, 本などの物体は手の領域との重複が認められないため, インタラクションはないと判定されている. Centernet と Openpose を併用して用いる利点は, 図 2 の画像内に存在する人は一人のみのケースだが, もし画像内に複数人いたとしてもそれぞれの人に対して手と物体のインタラクション判定が可能であるため, 画像内の複数の人に対してそれぞれ独立して物体とのインタラクション判定ができる点である.

3.2 インタラクションを用いたチェイン構造から知識グラフの生成

インタラクションがあると判定された後, どのように知識グラフを生成するかについて述べる. 本研究では行動認識を知識グラフとして表す. 行動は「A で（は）B を C する」といった主語 A, 目的語 B, 述語 C で構成される文章で

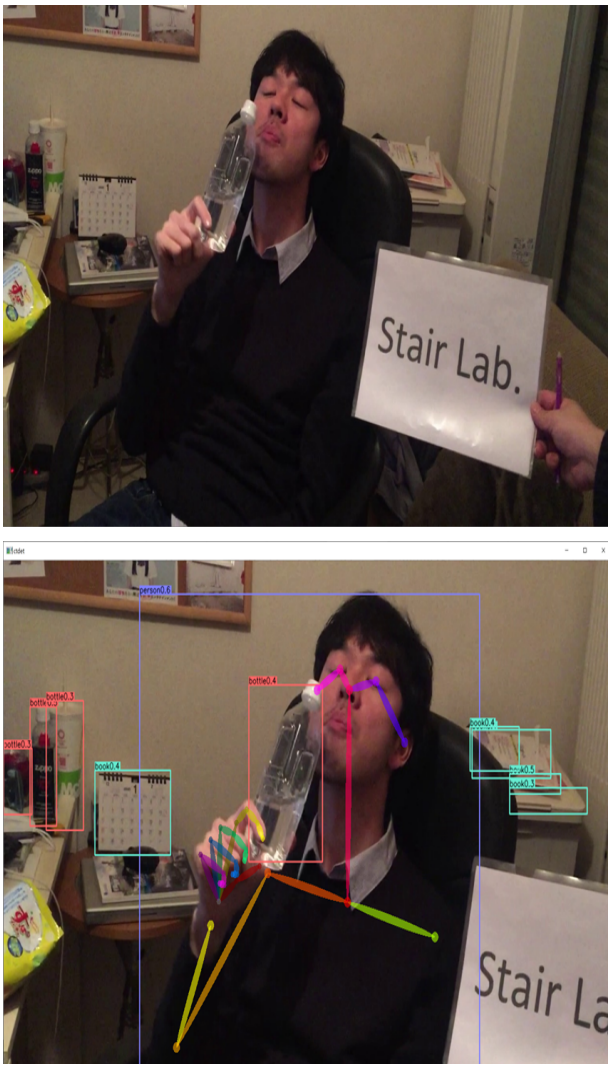

図 1　Centernet と Openpose による人と物体の検知

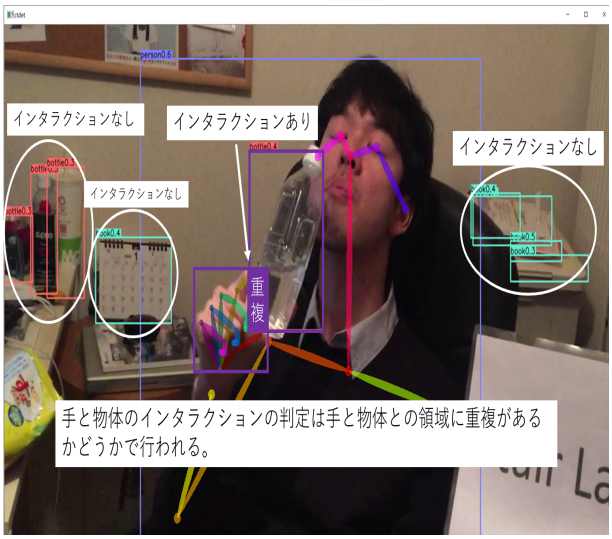

図 2　人と物体のインタラクションの判定方法

表されるが, この A,B,C に対してそれぞれインタラクション判定で得られた情報を当てはめることで文章化する. 例えば, 図 2 のように手とボトルのインタラクションがあると判定された場合, 主語 A に「手」, 目的語 B に「ボトル」を入れると文章は「手でボトルを C する」となる. 問題は C に何を入れるかということだが, ここで体の部位毎に可能な基本アクションを予め設定しておく. 「手」ならば「触れる・持つ（握るも含む）」, 足ならば「履く・乗る・蹴る」, 顔ならば「見る・聞く・食べる・飲む」と体の部位ごとにできることはかなり限定されるため, その中から述語 C を選ぶ.「手でボトルを C する」ならば, 手の基本アクションは「触れる・持つ」なので, 図 2 の人の行動はひとまず「手でボトルを持つ」という行動として認識される.

このように基本的に手と物体のインタラクションの場合, どのような行動の目的があろうとも大抵は「手で触れる・持つ」という動作を伴ったものである（例えば,「飲み物を飲む」という動作も細かく書くと「手で持っている（触れている）ボトルで飲み物を飲む」）. そのため, 基本的に手とのインタラクションがあると判定された物体に対して,「インタラクションを判定された物体を手で触る・持つ」を行動の基本として定義すれば, どんな物体であったとしても物体検知さえできるものならばまずは物体の情報を伴った初歩的な行動認識は可能となる.

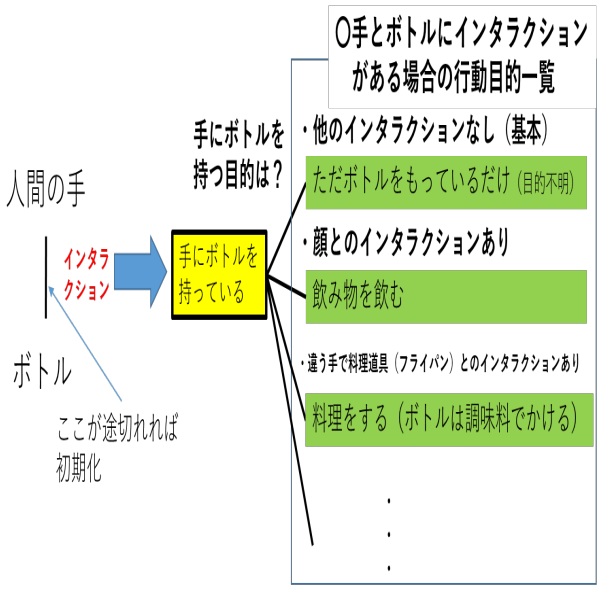

図 3　インタラクションによる行動の限定化

次に, 複数のインタラクションにおける行動認識について考える. 具体例として, 前述の「手にボトルを持っている」という動作に別のインタラクションが加わったケースについて考えた時の図を図 3 に示す. 図 3 のように「手にボトルを持っている」目的は, 一番可能性が高いのは「飲み物を飲む」ためであり, 他にはボトルは実は調味料で「料理をする」等, いくつかのパタンが考えられる（「ただ持っているだけ」というパタンもある）.「手にボトルを持っている」というインタラクションがある時点で行動はかなり

限定化されるが, それらをどうやって見分けて行動認識を行うかは図 3 で記述しているように手以外における別の体の部位とのインタラクションによって判断する. 図 3 の行動目的一覧に, 顔とボトルとのインタラクションがあれば「飲み物を飲む」など, 手以外のインタラクションがある場合の対応の一例が記載されており, このように手とボトルのインタラクションがある状態でインタラクションがある物体と手以外のインタラクションが判定されば場合, 図 4 にあるように「手にボトルを持っている」から「手に持っているボトルで飲み物をのむ」に変化する. 行動の文章化に関しては, 「(最初のインタラクションによって得られた文章) で (新たなインタラクションによって得られた目的) をする」といった初期のインタラクションによって構成された文章の後に新たなインタラクションで得られた文章を繋げる (チェイン構造) ことにより, 幅広い行動認識の文章を作成することが可能となる.

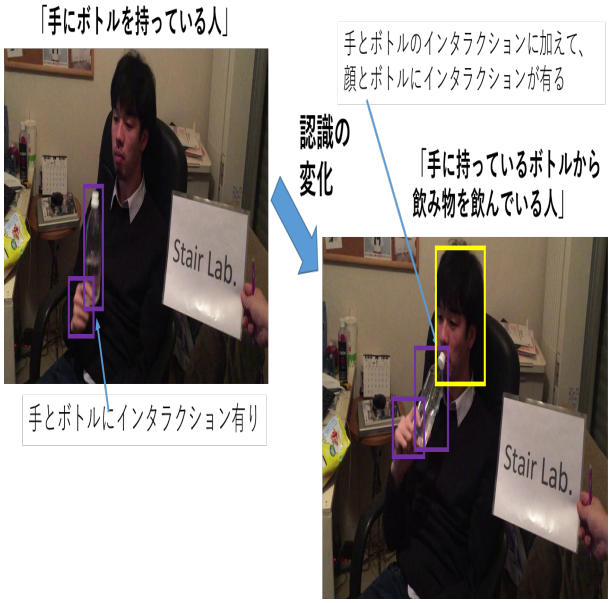

図 4　複数のインタラクションによる行動認識

ボトルを例として複数のインタラクションによるチェイン構造による行動の文章化について説明したが, ボトル以外でもシステムで検知可能な物体ならばボトルの時と同じように「(手で物体を持っている) + (行動の目的) をする」といった文章化が可能である (図 5). 行動目的の条件分岐として,「人がボトルを手に持っている」ならば「飲む・持っている」,「人が皿を持っている」ならば「皿を洗う・皿を運ぶ・料理をしている」等が考えられ, これらの情報を上手く繋げ合わせることができれば動作述語, 動作主体, 被主体などの情報もかなり詳細に得られるのではないかと期待している. ただし, 現時点のシステムでは「手で物体を持っている」という単体のインタラクションの判定と文章化までの実装に留まっており, 複数のインタラクションによる行動目的の設定も単体のインタラクションと同様に体の各部位で定められた基本アクションを組み合わせて文章を結合させることである程度の自動生成は可能だが, それだけで十分なのか検討する必要がある.

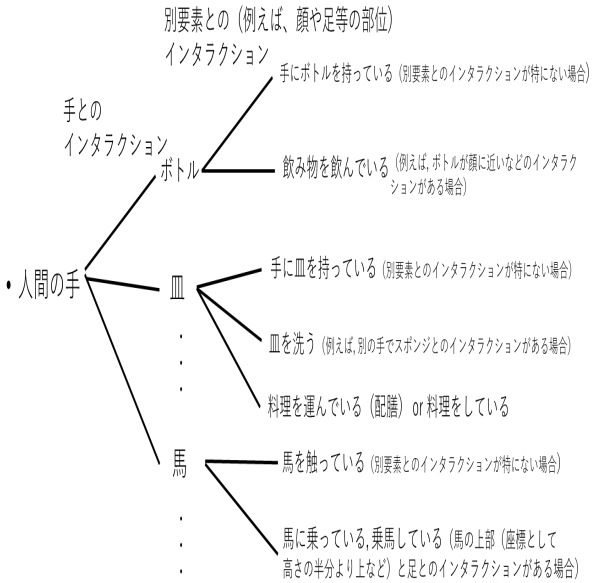

図 5　インタラクションによる条件分岐

3.3　テストデータによるインタラクションの判定精度の検証

ここでは, Centernet と Openpose を用いて作成したインタラクションを判定するシステムの精度について調査した. 画像内の人の手とボトルにおけるインタラクション判定と行動の文章化についてテストデータを用いてシステムの精度の検証を行った. テストを行う際の条件を表 1 に記載する. テストデータには図の例に使用されている画像データと同じく Stair lab で公開されている人の行動の一つ「drinking」の動画の内,「手にボトルを持っている」と目視で判断できる動画を静止画にした画像データ 60 枚を用いることとする. 今回は物体検知の対象を手とボトル, 文章化の内容は手とボトルのインタラクションに関するものだけに絞った条件下でテストを行った. 評価にはインタラクション判定の正解数 (正解率) を用いる. システムに用いている Centernet は公開されている事前学習したモデル [8] とその学習済みデータ ctdet_coco_dla_2x.pth を使用し,Openpose は公開されている事前学習したモデル [2] を使用した.

表 1　インタラクションのテスト条件

データ枚数	60 枚
テスト内容	「手でボトルを持っている」画像に対する手とボトルのインタラクション判定
評価基準	判定の正解数 (率)

テスト結果の一例を図 6 に示す. 図 6 では検知された手とボトルが四角形の領域で囲まれているが, 手とのインタラクションがあると判定されたボトルは紫色の領域となり, インタラクションがないボトルは緑色の領域になるように設定している. 図 6 に記載している 2 枚の図は手とボトルのインタラクションを正確に判断できていることを視覚的に見ることができる. また, 画像内に記載されている「手にボトルを持っている人がいる」から検知された物体名を主語, 述語, 目的語に当てはめて文章化することも成功していると判断できる. 全テストの結果,60 枚の画像の内 50 枚は手とボトルのインタラクションを検出に成功した. 正解率は約 83%であり, 一部の画像では検出できなかったものの大半の画像でインタラクションの検知に成功した. 一部の画像でインタラクションを検知できなかった原因は, 急角度から見るボトルを検出できなかったケース, 手で掴むことでボトルの形状が把握し辛くなったケースなどが挙げられる.

この結果からある程度の精度でインタラクションの判定ができたことを踏まえると, 検知された領域を annotation データ, 文章を「ラベル」とすれば人と物体のインタラクションそのものを新たなトレーニングデータとして収集することも可能となる.

図 6 実験結果の一例

4. おわりに

本論文では, Centernet と Openpose を用いて人と物体のインタラクションにおける行動認識を行うためのシステムの提案を行った. その結果, ごく初歩的な段階ではあるがインタラクションの判定とそれに伴う文章化を行うことができることを確認した. この結果から, 物体の中心点を予測することで比較的計算コストが軽く, 精度も他の手法と見劣りしない物体検知である Centernet と複数の人間の詳細な骨格抽出が可能な Openpose を組み合わせることで, 人の手と物体の検出のインタラクションの判定による詳細な行動認識の抽出ができる可能性を見出した. しかし, 問題点もいくつか存在する. まず, 現時点でのインタラクションの判定方法が単純な 2D における領域の重複の有無なので, 遠近法などによる重なって見えるだけの物体を誤って検知してしまう問題がある. また, 実際に実験で判定したインタラクションは手と物体という一つのインタラクションのみだったが,3.2 節で記述したような複数のインタラクション行動目的の設定に関しても実装する必要がある. さらに, Openpose で検知した手の形はかなり細かい形状までわかるので, 将来的にはその形から触れているのか, あるいは持っているのかの区別がつくアルゴリズムの作成が課題となる. 今回は検討しなかったが, 顔における基本アクションである「食べる・飲む・聞く・見る」のうち「見る」という動作は行動認識において他の動作とは異なり, 物体と接していないにも関わらず影響を与えうるという点で非常に重要となるので非接触型のインタラクション判定方法についても考慮する必要がある.

参考文献

[1] 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.

[2] CMU-Perceptual-Computing-Lab:openpose 入手先 ⟨https://github.com/CMU-Perceptual-Computing-Lab/openpose⟩ (参照 2020-04-15)

[3] H. Durrant-Whyte, T. Bailey, "Simultaneous localization and mapping: part I". IEEE Robotics & Automation Magazine. 13 (2): 99110. CiteSeerX 10.1.1.135.9810. doi:10.1109/mra.2006.1638022. ISSN 1070-9932. 2006.

[4] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, Jethro Tan. Interactively picking real-world objects with unconstrained spoken language instructions, Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2018.

[5] Tsuyoshi Okita, Sozo Inoue, Recognition of Multiple Overlapping Activities Using Compositional CNN-LSTM Model, Proceedings of the 2017 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers Adjunct, 2017.

[6] Stair lab:A Large-Scale Video Dataset of Everyday Human Actions 入手先 ⟨https://actions.stair.center/⟩ (参照 2020-04-13)

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, Show, Attend and Tell: Neural Image

Caption Generation with Visual Attention, Proceedings of the International Conference on Machine Learning, 2015.
[8] xingyizhou:centernet(objects as points) 入手先 ⟨https://github.com/xingyizhou/CenterNet⟩ (参照 2020-04-15)
[9] Xingyi Zhou, Dequan Wang, Philipp Krhenbhl, Objects as Points, Computer Vision and Pattern Recognition, Apr 2019.
[10] 松尾 豊, AI の未解決問題と身体性, シンボルグラウンディングへ, 人工知能学会, 2016