仮想空間における人間とオブジェクトとのインタラクションに着目した行動認識の研究

佐藤秀輔[†1]　角薫[†2]

概要：本研究では，仮想空間において人と物体との相互作用に着目することにより言語を出力する行動認識システムを開発する．従来の行動認識は，人と物体とのインタラクションを検出する場合，人を主体として動作認識を行うことが多い．人を主体とした動作認識センサでは，物体によって人体の一部が隠れる等インタラクションの検知が困難になる場合がある．本研究の手法では，インタラクションしている物体の移動を検知し，物体の動きのみを認識する．そのため，物体の動きとそれに対応する言葉のデータベースを作成し，物体の動きの特徴を学習し言語として出力を行う機械学習モデルを用いる．それにより，オブジェクトの 3 次元での動きのデータを 1 次元の特徴ベクトルとして表し行動分類を行う．本研究では，物体を用いた動作を対象とし，仮想空間上の物理演算された物体に対してのインタラクションの結果を用いて，物体の動きと言語を結びつける．

1. 注意

本研究は，仮想空間において，人の行動を言葉として認識する際，物体主体で人の物体との相互作用を認識する手法を提案する．行動認識の手法としては，画像を用いて画像内の人がどのような行動をしているのかを予測し，画像の説明を生成するアプローチ，動画やストリーミング映像などから人とインタラクションしている物体を推測し行動認識をするアプローチ，また，深度センサを用いて人の骨格を取得し，その骨格の動きからどのような行動をしているのか推測をするというアプローチなどがある．これらは，人を主体として行動を認識する手法であることが共通点である．しかし，人が主体となった行動認識の場合，例えば人がものを持っていることで体の一部分が隠れ，体を認識できなくなった場合には行動認識できない．

そこで本研究では，人主体での行動認識ではなく，物体主体での行動認識の手法について，これまでの研究との比較や，提案するシステムについて議論する．

人と物体とのインタラクションを検知する研究において様々な取り組みが存在する．画像から人がどのような行動をしているのか推測する手法[1][2]では，人がどの部分にいるのか画像認識を行い，人に近い物体とのインタラクションを機械学習のモデルを用いて判断する手法である．これは，画像のある動作の一部分を切り取って判別を行うため，動作が連続的なである場合には認識が難しい．画像内に写っている人と画像内の物体などを認識して人と物体の空間的な成約を考慮することで，推定の精度をより高めた研究[1]がある．

動画などの連続した情報から人と物とのインタラクションを推測するという手法[3][4][5][6][7]もある．これらの研究では，常に人と物体を認識し続け，それらの位置関係などをもとに機械学習を行い，人とインタラクションのある物体を認識し，それがどのような行動であったかの説明を生成する．また，人の行動を分解して，ある時点の行動から次の行動は何をするのかなどの行動予測に用いられる[3]．別の研究では，RGB-D カメラなどを用いて人の骨格を認識してその骨格の動きを学習し，どのような行動であったのかを認識するという手法もある[4]．

行動と言語を結びつける研究では，ロボットの動きを文章に表し，自然言語にてロボットへの動きを制御する研究[8]，人の行動を認識して文章に表し，自然言語でどのような動作かを入力すると，3D モデルの動作に反映されるという研究[9][10]がある．これらの手法では，入力と出力が同じとなる機械学習のモデルを作成し，言語と動きのそれぞれを学習させている．学習後に，機械学習のモデル同士を変換式により相互変換させている．

行動認識の際に人体が隠れているからといって認識ができないということは行動認識において難しい問題であると考え，本研究では人が物体に隠れた際にも行動認識できるように検討した．人と物体とのインタラクションにおいて物体に人体が隠れるということはよくあることである．また，「ドアを開ける」といったインタラクションでは，人やその人物の状況によって物体とのインタラクションの仕方が変わる動作が存在する．人の骨格に着目したシステムの場合，それがうまく認識が行えないという問題がある．本手法では，このように物体で体が隠れてしまう場合や，インタラクションの仕方によって認識が難しいという課題に着目し，物体中心の行動認識アプローチに関する手法とデータベースの構築に関して議論する．

† 公立はこだて未来大学
1 g2119018@fun.ac.jp
2 kaoru.sumi@acm.org

2. オブジェクトとのインタラクションに着目した行動認識

本システムは物体の動きをもとに人とのインタラクションを検知し，行動認識を行う．本システムは仮想空間上でデータを集められるように Oculus Quest を用いて仮想空間上の物体に対しての行動を認識，記録を行う．仮想空間を用いてデータを集めるメリットとして，あまり多くの特別な機材が必要でなく Oculus Quest などの VR 機器を利用すればよいという点，物体の認識やそれらのデータの書き出しなどが容易であるという点で VR 機器を利用してシステムの構築を行う．開発環境として，Unity を使用し，仮想環境の作成を行い，その仮想空間内での動きを認識して言語化するスクリプトを作成する．図 1 は本研究の提案システムでの推定の流れである．また，実際に使用した場合は図 2 のような形で推定結果が表示される．

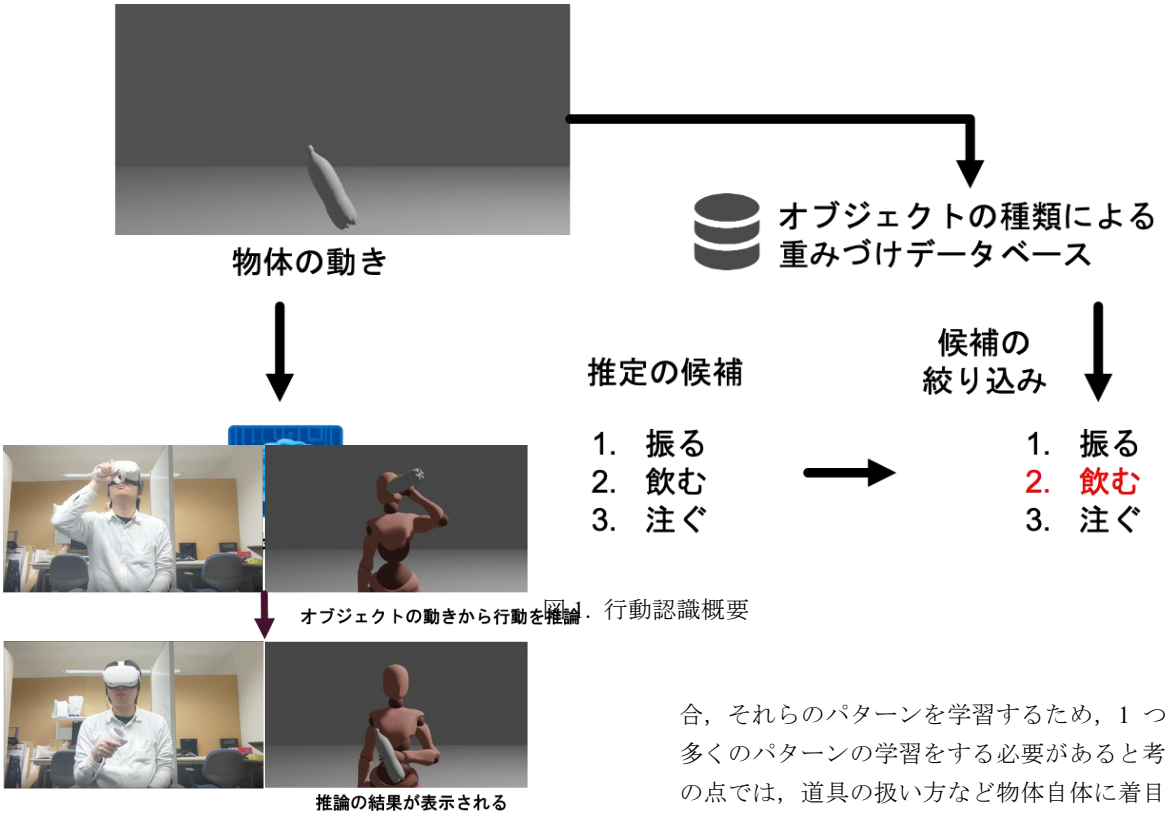

図 2. 実際の使用シーン

物体の動きから，どのような行動であるのか推定を行い候補を出力し，その候補の中から物体に対する重みづけのデータベースより候補の絞り込みをすることでどのような動作であったのか推定を行う．

本システムで扱う物体のデータとして，物体の加速度を利用する．物体の加速度というのは，仮想空間上での移動量よりその物体の加速度を計算によって取得を行う．このような方法で物体のデータを取得し，それらのデータに対して機械学習を用いて認識しその行動がなんであったのか出力する．

本手法では，人の骨格推定を行わず，物体の状態やその物体が何かという点に着目すればよいため，既存の機械学習を用いた研究などよりも学習にかかるコストを低減した状態で行動の推定ができると考えている．また，人によって動作が変わる動作の認識においても機械学習を用いる場合，それらのパターンを学習するため，1 つの動作につき多くのパターンの学習をする必要があると考えられる．その点では，道具の扱い方など物体自体に着目した場合多くのパターンがなく，学習も簡単で認識の精度が上がると考えている．現在は，簡易的な物体に対しての動作を考えているため既存手法との比較は行えていないが，既存手法との認識精度の差の比較などを行いシステムの比較することが必要だと考えられる．既存の研究との違う点として物体の動きを認識してどのような動きであるのか認識を行うため，物体を認識し続ける必要がある．

本手法の活用する場面として，3 次元空間を認識してその中で人の物体とのインタラクションを認識する際により軽量な認識手段として扱えるのではないかと考えている．その理由として，物体の動きを確認する必要があるが，人の骨格を認識して推定を行う場合は各関節などを認識してそれらすべての動きを学習する必要があるため使用するパラメータが多いということが挙げられる．物体の動きであれば，関節などほど多くのパラメータがなく，1 物体につきあまり多くのパラメータを利用する必要がないため，推定が簡単であると考える．

推定の上位 3 件をシステムは出力する．その後，動作を受けた物体の認識を行い，その情報をもとにその結果を絞り込む．絞り込みには重み付けを行ったデータベースを使用

する．データベースは，物体の特徴のデータと，その特徴に対してどのような動作が行われることが多いのかというデータにより重みづけされる．

本研究のシステムは，従来の行動認識と違い物体の加速度のみで人の行動を認識することができるため，仮想空間上ユーザがフルボディトラッキングを行っておらず，人の体の動きを取得できないというような状況でもどのような行動をしているのかを認識することが可能になる．例えば，仮想空間上でどのユーザがどのような行動をとっているのかなどが分かることで仮想空間上でも比較的少ない計算量でユーザの行動を知ることができると考えられる．

また本研究のシステムでは，現実世界の物体に加速度センサを付け，その情報を通信することでどのような動作をされているのか認識することができると考えている．仮想空間上の物体が物理演算で動いており，動きの特徴を抽出することで現実でも活用できると考えられる．

3. まとめ

仮想空間上の物体の状態に着目した行動認識システムについて検討した．物体のみに着目しても，人のみの行動認識と同様な認識できるのではないかと考えている．また，すべての認識を人のみの認識から変更することはできないかもしれないが，特定の行動の認識などは人のみの認識より高く認識を行い利用できることが考えられる．

参考文献

[1]A. Gupta, A. Kembhavi and L. S. Davis, "Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 10, pp. 1775-1789, Oct. 2009.

[2]G. Gkioxari, R. B. Girshick, P. Doll?r, K. He, "Detecting and Recognizing Human-Object Interactions", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8359-8367, 2017.

[3]J. Ji, R. Krishna, L. Fei-Fei, J. Carlos Niebles, "Action Genome: Actions as Composition of Spatio-temporal Scene Graphs" in arXiv, 2019

[4]S. Lei, Y. Zhang, J. Cheng, H. Lu. "Skeleton-Based Action Recognition With Directed Graph Neural Networks." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019

[5]I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-8.

[6]W. He, P. Sören, Y. Ersin, K. Vladimir, S. Ozan, S. Srinath, G. Leonidas. "Learning a Generative Model for Multi‐Step Human‐Object Interactions from Videos", Computer Graphics Forum. 38. 367-378. 10.1111/cgf.13644, 2019

[7]A. Pablo, C. Javier, M. Ana, "Finding Regions of Interest from Multimodal Human-Robot Interactions", 73-77. 10.21437/GLU.2017-15, 2017

[8]T. Yamada, H. Matsunaga and T. Ogata, "Paired Recurrent Autoencoders for Bidirectional Translation Between Robot Actions and Linguistic Descriptions," in IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 3441-3448, Oct. 2018.

[9]W. Takano, Y. Yamada, Y. Nakamura, "Linking human motions and objects to language for synthesizing action sentences" in Autonomous Robots, Volume 43, pp 913-925, April 2019.

[10]M. Plappert, C. Mandery, T. Asfour, "Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks", Robotics and Autonomous Systems, 2018.