Vol.2022-MBL-102 No.27
Vol.2022-UBI-73 No.27
2022/3/8

センサ情報から分析した人の行動認識情報を用いた会話システムの提案

豊坂 祐樹[†1]　大北 剛[†1]

概要：近年, ロボットなどを用いた AI による会話システムは多々存在し, その技術も向上している．しかし, 会話の対象者と事前情報なしに場当たり的に会話するのは依然難しい．そこで, カメラなどのセンサ情報から対象者の行動などを認識し, 情報として組み込めばより幅広いな会話が可能ではないかと考えた．本研究では, カメラ等のセンサから認識した人の動きや人の行動に関わった物の情報を抽出し，会話情報として組み込むシステムを提案する．今回はその足掛かりとして，カメラから得られた映像に対する情報の抽出に焦点を絞り，映像内の情報を人の動作を主とした知識グラフとして抽出できるかどうかの分析と検証を行った．

キーワード：相互作用，物体検知，姿勢推定，知識グラフ

Proposal of conversation system using action recognition information analyzed from sensor information

YUKI TOYOSAKA[†1]　TSUYOSHI OKITA[†1]

Abstract: In recent years, there are many AI-based conversation systems that use robots, etc., and their technologies are improving. However, it is still difficult to have a natural conversation with a person without prior information. Therefore, we thought that a wide range of conversations would be possible if the behavior of the target person is recognized from the sensor information of the camera and incorporated into the system as information. In this research, we propose a system that extracts information on human movements and objects related to human behavior recognized from sensors such as cameras and incorporates them as conversation information. This time, as a stepping stone, we focused on extracting information from camera video and verified whether the information could be extracted as a knowledge graph mainly for human movements.

Keywords: interaction, object detection, pose estimation, knowledge graph

0.1　著者紹介

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．一昔前のロボットは予め決められた単純な動作を繰り返すだけだったが，複雑である程度自立した動作や人間からの言葉に対してロボットが判断して受け答えするなど，ロボットに求められる能力が高くなっている．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．ロボットに目を移してみよう．ロボットとの対話でも音声情報だけではなく，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してそれを対話に生かすケースも存在する．最近では，会話するロボットやソフトウェアが数多く開発されており，その代表歴な例としてソフトバンクの会話ロボットのペッパーや Apple 社の iPhone や iPad に搭載されているバーチャルアシスタントシステムである Siri 等が挙げられる．

しかし，この対話するロボットやシステムにおいて，問

[†1] 現在，九州工業大学
Presently with Kyushu institute of technology

題点がいくつか存在する．1 つ目に, 対話を始めるあるいは終わるタイミングである．通常ならばユーザが対話システムに話しかけることでスタートするが, たまたま遠くから聞こえた音声を拾ったのか（ノイズ）ユーザが会話システムに話しかけてきたのかの判断は意外と難しく，Bohus らの研究 [1] ではカメラに写るユーザーの顔の位置や大きさからシステムと対話しようとしているかを判定している．また，杉山らの研究 [2] ではシステムに対する入力音に対して，応答義務があるかどうかの判定を行うシステムを開発している．

2 つ目に, ユーザの意図をどれだけ正確に把握できるかである．基本的に，同じユーザとの会話をするとしてもユーザのその時の状態により会話の話題は常に変化するが，それをシステム側が認識するのは難易度が高く，ユーザの会話中の音声情報や会話情報を元に学習していくのがスタンダードな方法である．そんな中，もし，カメラなどのセンサを用いてリアルタイムで変化し続けるユーザの動作や状態をその都度，認識できればユーザの意図を把握する一助となり，さらにシステムが応答する際の情報としても幅を持たせることが可能になる．

そこで，本研究ではカメラ等のセンサ情報を相互作用検知による情報抽出を行い，その情報を組み込む手法の提案を行った．前述の既存の研究でもカメラ等のセンサ情報からユーザの顔情報からシステムへと対話しようとしているかの判定などはあったが，我々はセンサ情報から抽出したより詳細な人の動作情報を対話システムに利用することで，これまでの対話システムより幅広い会話の選択肢のを狙い，さらに，センサ情報から得られた動作データを使って，突発的に会話のきっかけとしてシステム (ロボット) 側から対話を始めることもできるのではないかと考えた．先行研究 [3] では動画における何フレーム目に行動が始まったあるいは，終わったがわかるシステムを作成したので，リアルタイムに変化する状況に対応することが可能となり，対話システム側がユーザーに対して話しかけることもできるので，例えば，目の前でユーザが握手していた場合，「今，握手していましたか？」といったロボット側からの質問としての会話を始めるといったことも可能である．

本研究におけるわれわれの貢献は以下の通りである．

- センサ情報から人の行動を認識することによる対話システムの話題情報の拡張
- 検知した人の行動の始まりや終わりの瞬間を用いて対話システム側から会話を始める契機にする可能性の模索

また, 本論文の手法の有用性を以下の手順で示す．

(1) センサから対話システムに使用可能な情報の抽出法について説明
(2) 構築した検出システムが対話システムに使用可能な情報を抽出できているかの有効性のテスト

今回は，対話システムの内部情報に組み込むための属性情報抽出が可能かの段階までを焦点とし，テストではカメラの映像から情報抽出のテストを行い，対話システムに組み込むことが可能な情報を取り出せるかどうかの調査を行った．映像内容は人同士の動作「握手する」と単体での人の動作「物を手で持つ」を組み合わせた内容で例えば，映像内でボトルを手に持った状態からボトルを置いて，他の人間と握手するなど，同一映像内で動作を変化させ，その都度動作を行う人（主体），どんな動作をしたかの行動内容（述部），主体となる人の動作に関わった対象（目的体）の情報が取り出せるかを検証した．その結果，正解率は 84.2%を達成し，情報の抽出ができていることを確認した．

2. 本研究に使用するモデルの概要

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による行動認識を利用してカメラなどのセンサ情報からユーザの行動を認識し，それを対話システムに応用する方法を考案した．本節では，本研究に使用した相互作用を用いた行動認識のモデルと対話システムについての説明を行う．

2.1 相互作用を用いた行動認識モデル

対話システムにカメラなどのセンサを組み込み, センサ情報を対話システムに利用しようとしたとき，逐一流動するリアルタイムでの行動の変化を捉えることが重要となる．従来の行動認識では，対象の動画や静止画において包括的に行動を推定する場合が多い．例えば，ある動画に対してこの動画の人間の行動はランニングをしている等，「いつ, どの瞬間にどんな行動が始まり，終わるのか？」といった時系列的な要素が抽出できないので，会話に使用するための情報としては不足している．そこで，我々は先行研究 [3][4] で人と人，人やモノとの相互作用を用いて行動を認識する手法を提案した．人同士の接触の瞬間（相互作用）を捉えることにより，ただ行動を認識するだけではなく，行動の始まりの瞬間や終わりの瞬間を把握することで，リアルタイムでの変化の情報の抽出に成功した．

相互作用を用いた行動認識システムの概要を図 1 に示す．この行動認識システムは入力された画像内における人や物を検知するための物体検知である centernet[5][6]，姿勢推定である openpose[7] と人や物体の接触を判定するための 3 次元座標を推定する深度マップ [8] を組み合わせており，動画などでも何フレーム目に相互作用が発生したかも認識可能となっているため，時系列の情報も抽出できる．また，出力はその画像の空間で起こるすべての事象を認知したい

Vol.2022-MBL-102 No.27
Vol.2022-UBI-73 No.27
2022/3/8

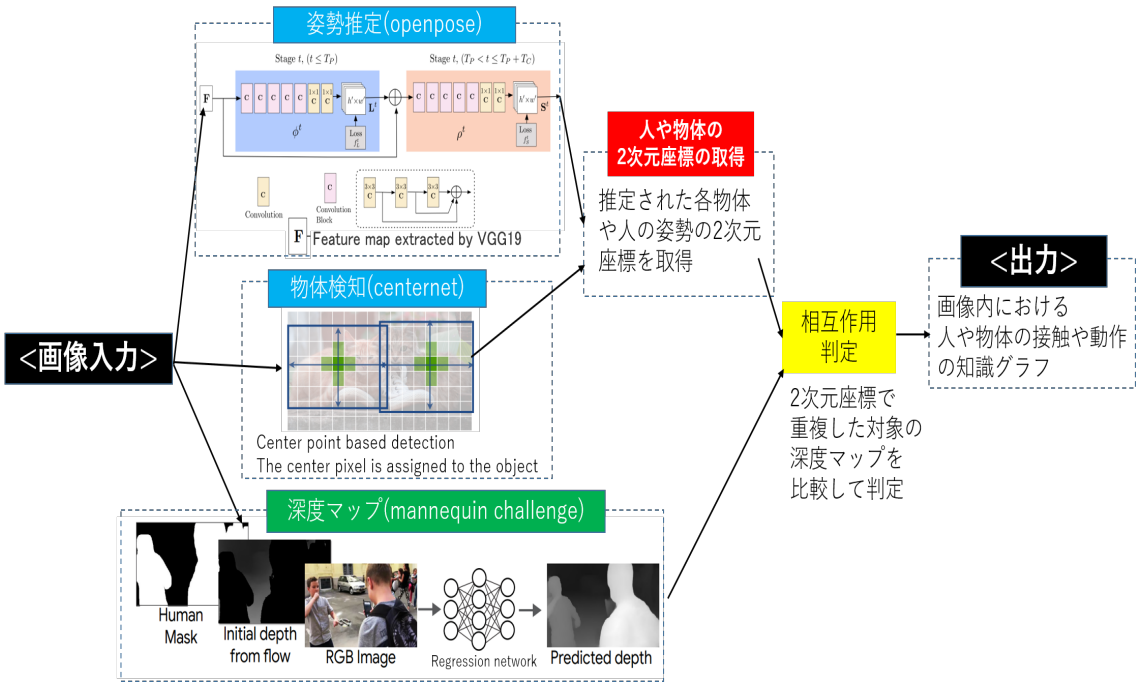

図 1 センサ情報を分析するためのモデル

ので，知識グラフとなっており，知識グラフの具体的な内容は後述の 3.1 節（図 3）に記載する．本研究では，このシステムを使用することでセンサ情報から人の行動を認識する際に，人の動作や人の動作に関わった物体の検知，それらの動作の始まりの瞬間や終わりの瞬間を把握し，会話システムに利用する．

2.2 対話システム

対話システムとは，人間と対話するコンピュータやソフトウェア，ロボットのことを指し，代表歴な例としてソフトバンクの会話ロボットのペッパーや Apple 社の iPhone や iPad に搭載されているバーチャルアシスタントシステムである Siri 等が挙げられる．対話システムには大きく分けて解決すべきタスクが存在するタスク指向型と解決すべきタスクがない非タスク指向型の 2 種類があり，タスク指向型は例えば，音声による条件に合ったレストランの検索システムなど，目的に合致した答えにたどり着くための対話を行うシステムであり，非タスク指向型は雑談会話システムとも呼ばれている．本研究では雑談等の対話を目指していることから非タスク指向型の分類に入るニューラル対話システムを用いることにする．

ニューラル対話システムは Neural Conversation Model(NCM) による発話と応答の対応を学習することで対話を構築するシステムである．RNN をベースとした対話文脈から次の応答を予測するモデルで，RNN を使用して図 2 に示すように「ABC」を「WXYZ」にマッピングするように学習する．このシステムは相手の発話を理解した上で応答を生成しているわけではなく，応答中の単語予測を誤差関数に組み込むことで学習するシステムであり，例として Oriol Vinyals, Quoc V. Le らの研究 [9] などが挙げられる．このニューラル対話システムに対してセンサから得られた人の動作等の認識情報を特徴量として組み込むことで，応答の幅や対話の精度を向上させることを試みる．

3. 提案手法

この節では, カメラセンサから得られた映像からの抽出と対話システムに組み込むための方法について説明する．まず，対話システムに用いるための情報の抽出法について説明し，次にその抽出した情報をどのようにして対話システムに組み込むのかを述べる．

3.1 センサ情報認識器

ここでは，対話システムに組み込むためのセンサ情報認識器について説明する．本研究のセンサ情報認識器は, 部屋という時空間を示して，そこで起こったすべての動作を知識グラフとして記録する (少なくとも 2 次元カメラでキャプチャしたことすべてを記録する) システムである．人と人，または人と物体との相互作用検知は先行研究 [3] の手法を使用するが，基本的に行動認識の方法は相互作用が判定された人や物によって動作も決定するようになっている．例えば，人の手同士の相互作用が判定された場合は，「人同士の握手」の動作として認識され，人の手とコップやナイフ，スプーン等の物との相互作用が判定された場合は．「手でその物体を持つ」動作と認識されるそのことか

Vol.2022-MBL-102 No.27
Vol.2022-UBI-73 No.27
2022/3/8

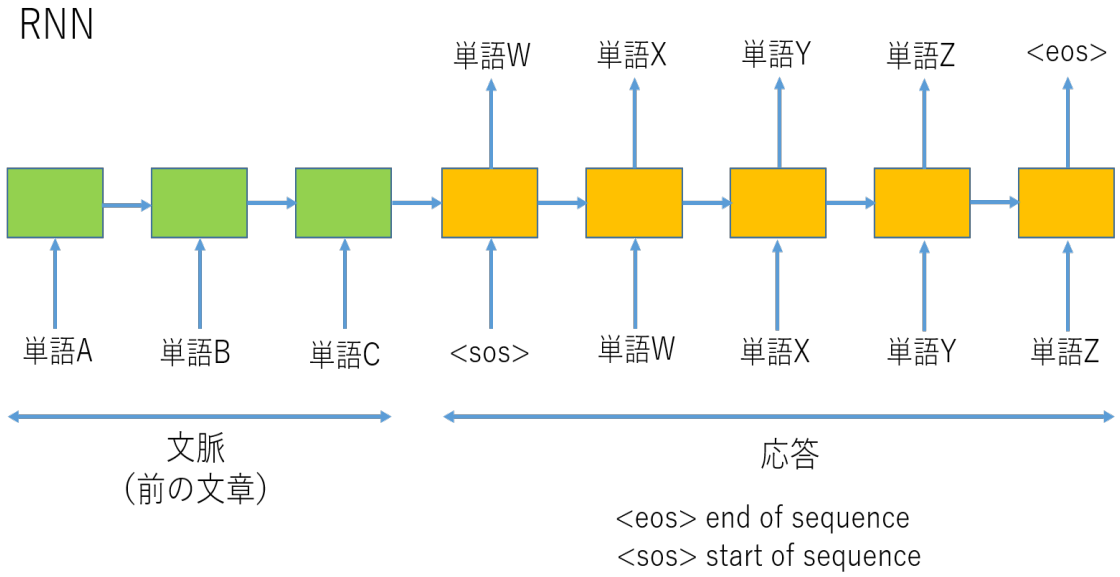

図 2 RNN を用いたニューラル対話システムの発話・応答の学習のフレームワーク

ら，相互作用が判定された場合，抽出できるのは「主体となる人」,「対象が触れた目的体（人やモノ）」,「その時の行動（述部）」,「相互作用が発生した時間（動画なら何フレーム目かの情報，日時が入っている動画ならばその時間）あるいは終わった時間」となり，これらの情報を対話システム用の属性情報として変換する．

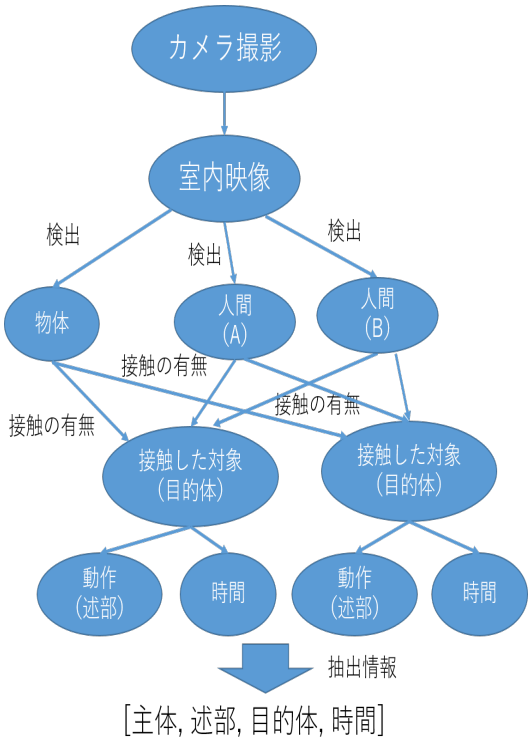

図 3 カメラで撮影された映像をセンサ情報認識器に適用した場合の情報の知識グラフ化（人間 2 人 (A,B) の場合）

図 3 はカメラでの撮影した映像からセンサ認識器を使用した際に生じる情報を知識グラフ化したものである．（部屋内に存在する人間が人（A）, 人（B）の 2 人と仮定した場合の図）その際に，属性情報として設定されるのは主体，述部，目的体，時間の 4 つであり，カメラで撮影された部屋内に存在する人間がそれぞれ「その時刻に何をしているのか？」という情報をその人間と接触した対象（目的体）との相互作用情報を用いて定める．

例えば，図 4 の上図は認識器による知識グラフ化の具体例であり，人の手同士の相互作用が検出された場合，主体「人（A）」と相互作用している対象は「人（B）」，行動は「握手」となる．もし，人の手とコップとの相互作用だった場合は，相互作用している対象は「コップ（物）」，行動は「持つ」となる．時間は相互作用が検知された時刻がそのまま記録される．ただし，時間に関してはそれだけではなく，本システムは動作の始まりと終わりを検知できるようになっているので，動作の始まりの瞬間，または，終わりの瞬間であった場合，それも情報として記録される．これにより，対話システムは撮影された人が何時どんな動作をしたかの詳細な記録を保持することができる．

3.2 対話システムに行動認識を組み込むための手法

相互作用を使用して抽出した情報の対話システムへの組み込む方法について説明する．基本的にセンサ情報はセンサ情報認識器に逐一送られる形となっており，映像内に写された部屋中の情報の抽出が時系列も含めて行われる．センサから得られた情報をセンサ情報認識器によって，対話システムに使用可能な情報に変換し，その情報を対話システムのデータベースに格納するが，特徴量として扱えば対話システムの新たな学習に使用でき，あるいはその情報自体を発話として扱えば，話しかけられなくてもユーザに発

Vol.2022-MBL-102 No.27
Vol.2022-UBI-73 No.27
2022/3/8

図 4 相互作用を用いた対話システムに組み込むための情報抽出の一例

声することが可能となる．例えば，センサが 2 人の人間の手の相互作用を検知した場合，センサ情報認識器によって，「主体，述部，目的体，時間」の情報が抽出，対話システムに組み込んだ際にこの情報自体を一種の発話とみなし，それをもって「あなたは人と握手しているのですか？」といった会話をするきっかけとなる応答を自発的に行うことも可能である．

4. テスト

本節では，カメラ等からの映像情報からの人の動作や動作に関わった物体等における情報抽出を行うシステムのテストを行い，抽出した情報の精度について検証する．その後，テストにおける分析及び考察を行う．

4.1 テスト条件

PC に内蔵されたカメラで撮影した映像を用いて対話システムに組み込むための各情報が抽出できるかどうかをテストする．カメラで撮影した部屋内で起こった全てのこと（動作）を知識グラフとして記録できるかどうかが焦点となるため，カメラで撮影する際には，最初は「手にボトルを持つ」動作を行い，その後，次にボトルをコップに持ち替えて「手にコップを持つ」動作をする，あるいは，「手にボトルを持つ」動作の後，ボトルは置いて，「人と握手する」動作を行うなど，撮影したカメラの映像において動作を何度か変化させてその都度，それらの動作を検出できるか調査した．テスト条件は以下のものとする．

- PC に内蔵されたカメラを用いて撮影した映像
- 映像内容：「手に物体を持つ」や「握手を行う」
- 映像数：19
- 撮影時間：10 から 20 秒
- 評価方法：撮影された映像に対してシステムを適用し，その抽出された情報が正しいかどうか，その正解数によって精度を検証する．

具体的な撮影映像の内容としては，「手でボトルを持った」後，ボトルを置いて，「手でコップを持つ」に切り替える映像，「手でコップを持った」後，ボトルを置いて，「手でボトルを持つ」に切り替える映像，最初から「両手でボトルとコップを持つ」映像，最初に「人と握手をした」後，握手をやめて「ボトルを持つ」映像，最初に「ボトルを持った」後，ボトルを置いて「人と握手をした」映像を用意した．これらの映像に対してテストを行い，対話システムに組み込むための知識グラフとしての情報「主体」「述部」「目的体」が正しく抽出できるかを検証する．例えば，最初に「人と握手をした」後，握手をやめて「ボトルを持つ」映像の場合，最初の動作で「主体：人」「述部：握手する」「目的体：人」，動作が変わった後に「主体：人」「述部：手で持つ」「目的体：ボトル」が検出されれば正解とする．

4.2 テスト結果

テスト結果の一例を図 5 に示す．図 5 の上図は最初にボトルを持ち，後にボトルを置いてコップに持ち替えた映像であり，検出時間はフレーム数で表されている．2 フレーム目では何も持っていないが，101 フレーム目ではボトルを持っており，それを主体「人」，述部「手で持つ」，目的体「ボトル」として検出できているを見ることができ，その後の 371 フレーム目では，ボトルではなくコップに持ち替えているが，その情報は主体「人」，述部「手で持つ」，目的体「コップ」として目的体が変わって検知できているのがわかる．一方，下図では 106 フレーム目に両手にそれぞれコップとボトルを持っており，それぞれに対して主体，述部，目的体が検出されているのがわかる．さらに，161 フレームでは，物体を離した瞬間として，動作の終了（図ではインタラクション終了と表現）が検知されている．この結果から一度に複数の動作検知が可能となっていることと，変化した場合の情報抽出も正しく行われているのがわかる．

各映像における情報を抽出した際の正解率を表 1 に示す．

表 1 各映像における抽出した情報の正解率

	抽出結果
正解率	84.2% (16/19)

Vol.2022-MBL-102 No.27
Vol.2022-UBI-73 No.27
2022/3/8

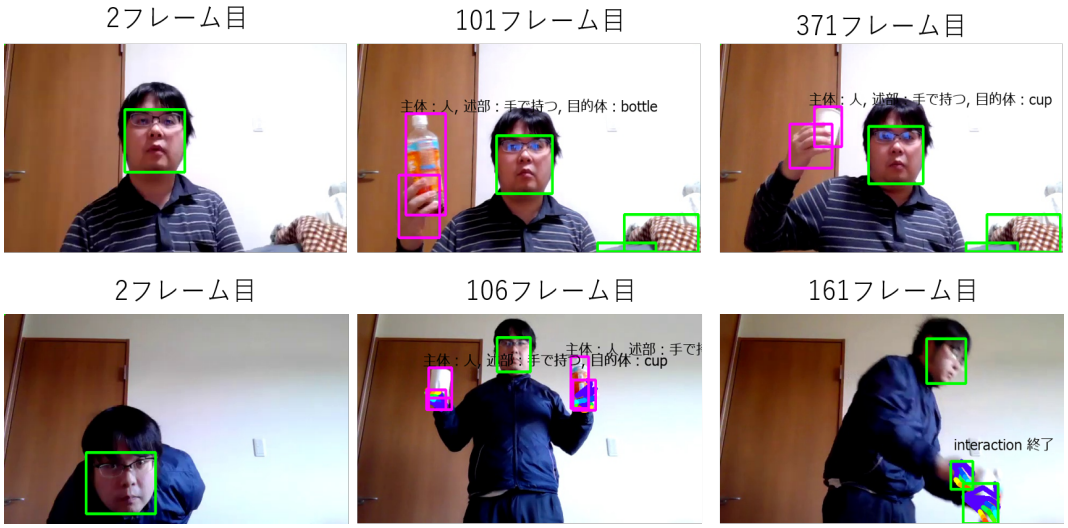

図 5 テストデータに対する結果の一例
(上図：途中でボトルからコップに持ち替えたのを検知，下図：両手でそれぞれ物を持った場合の検知と物を離した瞬間の検知)

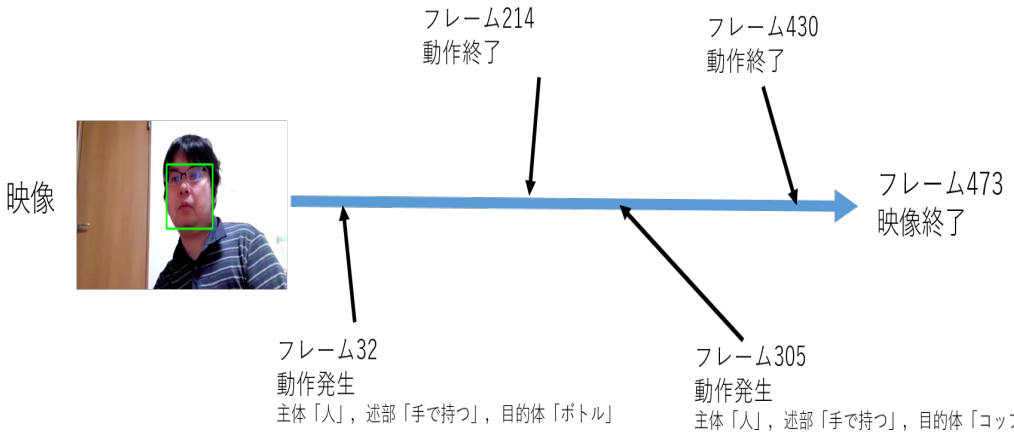

図 6 映像内で検知された動作の始まりと終わりの履歴
(最初にボトルを持ち，その後ボトルを置いてコップに持ち替える映像)

システムを適用して映像を分析した結果，正解率は 84.2%と高い精度になり，たとえ映像内の動画が変化してもその都度，情報を抽出できることができた．図 6 は最初にボトルを持ち，その後ボトルを置いてコップに持ち替える映像に対して，システムを適用して時系列も含めて分析して得られた情報を示した図であるが，最初は何も動作はなく，32 フレーム目から 214 フレーム目までボトルを持ち，305 フレーム目から 430 フレーム目までコップを持つという情報を主体，述部，目的体として抽出できているのをみることができる．また，同時に両手にコップとボトルを持った映像でも右手にコップ，左手にボトルを持っているといったそれぞれの動作の情報として複数の動作情報を同時に取得できることも確認した．

ただし，問題点として動作の始まる瞬間と終わる瞬間の検知の正確性が挙げられる．動作自体は変化したとしてもある程度問題なく検知を行えたが，その動作がいつ始まったか（終わったか）に関しては若干，誤検知が発生するケースが存在した．映像内で手の動きなどで角度が変わった際に人や物体が検知しにくい体勢となった場合，その瞬間だけ動作の検出が途切れてしまい，途切れた瞬間の箇所を動作の終わり，あるいはその後また検知が始まった箇所を動作の始まりといったことが主な原因として挙げられ，改善の必要があることを確認した．

5. おわりに

本研究では，カメラ等のセンサから認識した人の動きや人の行動に関わった物の情報を会話情報として組み込むシステムを提案した．今回はその足掛かりとして，カメラから得られる映像に対して認識システムを用いて会話システムに組み込むための情報の抽出に焦点を絞り，その情報抽

Vol.2022-MBL-102 No.27
Vol.2022-UBI-73 No.27
2022/3/8

出の精度について検証した．その結果，精度は 84.2%を達成し，映像内での途中で変化し続ける動作でも情報をその都度抽出できることを確認した．ただし，今回は情報抽出までしか行っていないので，その情報を実際にニューラル対話システムに特徴量として組み込み応答に使用することが今後の課題として挙げられる．

参考文献

[1] Bohus, Dan, Chit W. Saw, and Eric Horvirz, Directions Robot: In-the-Wild Experiences and Lessons Learned, International Foundation for Autonomous Agents and Multiagent Systems,2014
[2] 杉山, 船越, 中野, 駒谷, 多人数対話におけるロボットの応答義務の推定, 人工知能学会全国大会論文集 29, 1-4, 2015.
[3] 豊坂祐樹, 大北剛, 距離画像推定情報を用いた複数人の行動認識, 第 72 回ユビキタスコンピューティングシステム（UBI）研究発表会,2021 年 12 月.
[4] Yuki Toyosaka, Tsuyoshi Okita, Activity Knowledge Graph Recognition by Eye Gaze: Identification of Distant Object in Eye Sight for Watch Activity, HASCA2021 Workshop(Human Activity Sensing Corpus and Applications), UbiComp/ISWC, Pages 334–339, 2021.
[5] Xingyi Zhou, Dequan Wang, Philipp Krhenbhl, Objects as Points, Computer Vision and Pattern Recognition, Apr 2019.
[6] xingyizhou:centernet(objects as points) . 入手先 ⟨https://github.com/xingyizhou/CenterNet⟩（参照 2020-04-15）
[7] 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.
[8] 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.
[9] Oriol Vinyals, Quoc V. Le, Neural Conversational Model, ICML Deep Learning Workshop, 2015.