人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-060-02

RoboCup@Homeのヒューマンインタラクションタスクに向けた解法の提案

Proposal for Solution of Human Interaction Task in RoboCup@Home

矢野 優雅 [1*]　松本 生弥 [1] 福田 有輝也 [1] 小野 智寛 [1,2] 田向 権 [1,3]
Yuga Yano[1], Ikuya Matsumoto[1], Yukiya Fukuda[1], Tomohiro Ono[1,2], and Hakaru Tamukoh[1,3]

[1] 九州工業大学大学院生命体工学研究科
[1] Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Japan
[2] 日本学術振興会特別研究員 DC
[2] Research Fellow of the Japan Society for the Promotion of Science
[3] ニューロモルフィック AI ハードウェア研究センター
[3] Research Center for Neuromorphic AI Hardware, Kyushu Institute of Technology, Japan

Abstract: ホームサービスロボットの技術発展を目的として，RoboCup@Home という競技会が開催されている．RoboCup@Home では，実際の家庭環境を模したフィールドを用いてタスクを行うことで，より現実に近い環境でロボットの性能を評価することができる．本研究では，RoboCup@Home のタスクである Find My Mates を解くために，人物認識や音声認識を用いた動的環境でも動作する手法を提案する．また，提案した手法をロボットに実装し，2022 年 7 月にバンコクで行われた RoboCup@Home にて性能評価を行った．競技会では満点を獲得し，提案手法の有効性を示した．競技中の様子は，`https://www.youtube.com/watch?v=ucoP8_j6Kig` にて公開している．

1　序論

RoboCup@Home[1] は，ホームサービスロボットの技術発展を目的に開催されている国際的な競技会である．本競技会では，人間とロボットの協調を目標の一つに掲げており，音声認識や物体認識，ナビゲーションといったテストが動的環境下で行われている．そのため，より現実に近い家庭環境でロボットの性能を評価することができ，多くの注目を集めている．RoboCup@Home には，Open Platform League，Domestic Standard Platform League（DSPL），Social Standard Platform League という 3 つのリーグがある．我々の参加している DSPL では，トヨタ自動車株式会社が開発した Human Support Robot（HSR）[2] を標準機に採用しテストを行っている．図 1 に，HSR の外観と搭載されている主なデバイスを示す．HSR は移動台車やアームに加えて，RGB-D カメラやマイクが搭載されているため，物体認識や音声認識を通して動的環境下においても多様な動作を実現できるロボットである．

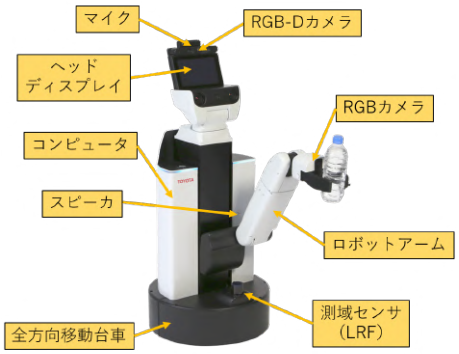
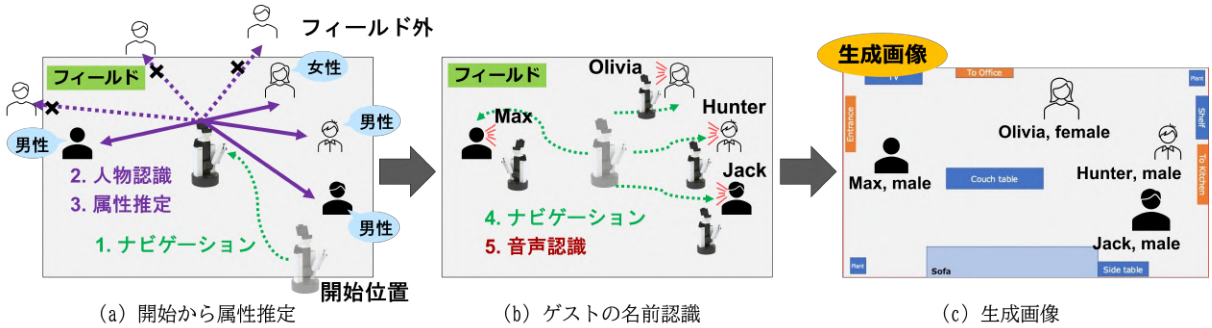


図 1: トヨタ自動車株式会社が開発した HSR

本研究では，特にヒューマンインタラクションの性能をはかる Find My Mates（FMM）というタスクに向けて，その解法を提案する．また，提案した手法を HSR に実装し，2022 年 7 月にバンコクで行われた RoboCup@Home にて性能評価を行った．競技会では満点を獲得し，本手法の有効性を示した．

*連絡先：九州工業大学大学院生命体工学研究科人間知能システム工学専攻
〒 808-0135 福岡県北九州市若松区ひびきの 2-4
E-mail: yano.yuuga158@mail.kyutech.jp



(a) 開始から属性推定　(b) ゲストの名前認識　(c) 生成画像

図 2: 提案した FMM の解法

2 Find My Mates

本節では，RoboCup@Home で行われるタスクの一つである FMM について述べる．FMM では 4 人のゲストと 1 人のホストが登場し，ホストの家にゲスト全員が訪れたという状況を想定している．しかし本タスクでは，ホストはゲストの外見や特徴については何も知らされておらず，名前のみを知らされている．そのため，ロボットが家に訪れたゲストを見つけ出し，顔や特徴，また部屋のどこにいるのかをホストに伝えることが FMM のメインゴールである．FMM の得点表を表 1 に示す．

FMM を遂行するためには，3 次元位置を含む人物認識に加えて，人物の特徴を推定する技術が必要になる．また，ゲストの名前を取得するためには，音声認識の技術も不可欠である．このように，FMM はヒューマンインタラクションのために必要な技術を包括的に評価することができるタスクである．

2.1 登場人物について

RoboCup@Home では，タスクに登場する人物はボランティアから選出され，トライごとに変化する．また，登場人物は自分の本名を使用するのではなく，事前に公開されている名前リストよりランダムに決定される．この名前リストには，アメリカで一般的に使用されている名前から選出した男女 11 個ずつの名前が含まれている．しかし，名前のみで男女の判別ができないように，男女で共通している名前が複数存在する．そのため，性別については別の手段を用いて認識する必要がある．

3 提案手法

本章では，FMM で満点を獲得するための解法と，HSR に実装した機能について述べる．

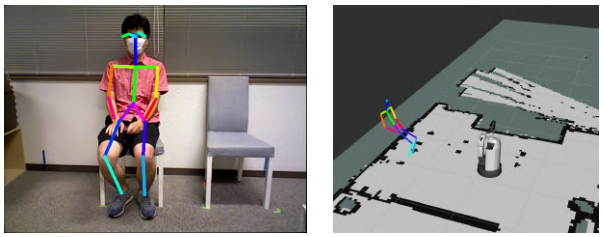
表 1: FMM の得点表

動作項目	回数	点/回数	合計点
メインゴール			
ゲストの位置報告	2	100	200
明示的に位置を報告する	2	50	100
ゲストの特徴報告	2	150	300
ボーナス			
3 人目のゲストも報告	1	150	150
3 人目のゲストの特徴も報告	1	250	250
減点対象			
ゲストから合図をもらう	2	-75	
ゲストの位置を教えてもらう	2	-75	
ゲストからロボットに近づく	2	-150	
合計			1000

3.1 FMM に向けた解法

我々は FMM で満点を獲得するために，図 2 に示す手法を提案する．始めに，ロボットを部屋の中央までナビゲーションさせ，部屋全体を見渡しながら人物認識を用いて 4 人のゲストを見つける．認識には RGB 画像のみを用いるが，Depth 情報も活用することでそれぞれのゲストの位置も同時に算出する．次に，算出したゲストの位置を基に，各ゲストの正面までナビゲーションを行い，音声認識を用いて名前を聞く．更に Class-Specific Residual Attention [5] という属性推定手法を用いて，ゲストの性別を推定する．

最後に，獲得したすべての情報（ゲストの画像，位置，名前，性別）を集約した 1 枚の画像を生成し，HSR



(a) RGB 画像での認識結果　(b) 3 次元の位置推定

図 3: 人物位置推定アルゴリズム

のヘッドディスプレイに表示することでホストに伝える．本手法で生成する画像を図 2 (c) に示す．このように，フィールドと各ゲストの情報をまとめて表示するため，メインゴールとボーナスを同時に達成できる．

3.2　音声認識

近年ではスマートフォンやスマートスピーカーなどの普及により，クラウドを用いた音声認識の研究が盛んである [3, 4]．しかし，RoboCup@Home では会場のネットワークが不安定である場合が想定され，クラウド上での安定した音声認識は困難である．また，ネットワークの課題は一般の家庭環境においても想定されるものであるため，オフラインでの音声認識技術が必要である．そこで本研究では，vosk[6] と呼ばれるオフライン音声認識手法を用いる．

更に，vosk は認識する単語リスト（辞書）を指定することで，辞書にない単語を認識から除くことができる．2.1 節で述べた通り，RoboCup@Home ではタスクに登場する人物の名前リストが公開されている．そのため，名前リストを基に辞書を作成し vosk に適用することで，認識精度向上を図る．

また，RoboCup@Home では音声認識を QR コードによってバイパスすることが認められている．そこで，音声認識に失敗した場合は自動的に QR コードによる認識に切り替え，名前を取得する．

3.2.1　ノイズ除去

RoboCup@Home は実際の家庭環境を模したフィールドで行われるが，実際の家庭環境と異なる点もある．その一つが，周囲の外音（ノイズ）である．RoboCup@Home には多くの観客がおり，また他のリーグも同時に行われているため，実際の家庭環境では見られないようなノイズが発生する．さらに，RoboCup@Home 会場でのノイズの特性は大会ごとに異なり，現地での調整が必要である．そのため，ノイズ除去の強度をパラメータで容易に調整可能なノイズ除去 [7] を音声認識の前段に組み込んでいる．

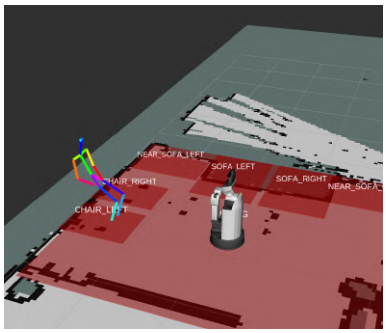


図 4: 図 3 で認識した結果に意味づけを行った結果

3.3　人物認識

本研究では人物認識の手法に Lightweight Human Pose Estimation[8] を用いた．本手法は処理が非常に軽量であり，CPU でも高速に動作する手法である．本手法を用いることで，図 3 (a) に示すように，RGB 画像から人物認識を行うことができる．次に，RGB 画像における認識結果と，Depth 画像を合わせることで，人物の 3 次元位置推定を行う．図 3 (b) に，人物の 3 次元位置推定を行った結果を示す．

3.4　ゲストの位置報告

FMM では，認識した人物の位置をホストに伝えるために，認識した人物が部屋の中のどこにいるのかを識別する必要がある．そこで本研究では，Simultaneous Localization and Mapping を用いて事前に作成したマップに対して意味づけを行う．RoboCup@Home ではフィールドの配置が事前に公開されるため，部屋の内外と椅子などの家具がどの位置にあるのかという情報も含めて意味づけを行う．図 3 に示した 3 次元の人物認識に対して，フィールドの意味づけを行った結果を図 4 に示す．この場合では，ゲストはフィールド内の左側の椅子に座っていると正しく判定している．

4　競技概要

我々は 2022 年 7 月にバンコクで行われた RoboCup@Home に参加し，提案手法の評価と現地環境における音声認識の精度検証を行った．RoboCup@Home では HSR の制御用に PC を 1 台使用することが認められているため，以下のような PC を使用した．CPU：Intel core i7-7820HK, GPU：Geforce RTX 1080，メモリ：32GB, OS：Ubuntu18.04．また，PC と HSR の通信には Robot Operating System [9] を用いている．

実際に使用されたリビングルームの概略図を図 5 に，競技中の様子を図 6 に示す．

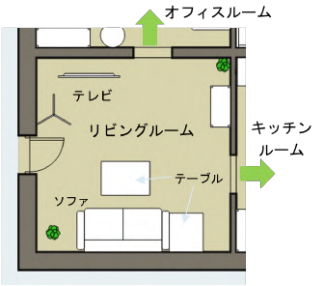
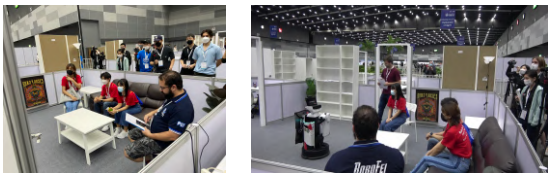


図 5: RoboCup@Home2022 で使用されたフィールド



(a) 1 度目のトライ　(b) 2 度目のトライ

図 6: FMM が行われた実際の会場

5　競技結果

5.1　音声認識の性能評価

始めに，提案手法における音声認識の性能評価を行った．実験に際して，アメリカで一般的に使用されている名前から男女それぞれ 11 個を選出し，本番に近い環境で話者やノイズの強弱を変化させながら 4 度ずつ読み上げ，計 88 個のデータを作成した．表 2 に，ノイズ除去と辞書指定の有無による認識結果を示す．ノイズ除去と辞書指定を行っていないの場合の精度は 13.6%であったが，辞書指定を行うことで 55.7 ポイント向上し 69.3%となった．加えてノイズ除去を行うことによって，精度が 2.3 ポイント向上し 71.6%で最も高くなった．しかし，辞書指定を行っていない場合では，ノイズ除去を行うことで精度が 3.4 ポイント低下した．

5.2　FMM の結果

我々は RoboCup@Home で FMM を 2 度トライし，性能評価を行った．1 度目のトライでは部屋中央へのナビゲーションに失敗し，ゲストから遠い位置に HSR

表 2: 音声認識の精度

辞書指定	ノイズ除去	認識精度 (%)
なし	なし	13.6
	あり	10.2
あり	なし	69.3
	あり	**71.6**

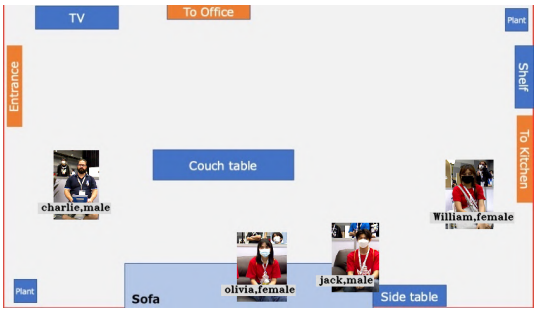


図 7: 2 回目のトライで作成したマップイメージ

が停止してしまった．それでも，人物検出と 3 次元の位置推定は正常に動作したが，各ゲストの顔画像が非常に低い解像度となってしまった．そのため属性推定が正常に動作せず，ゲストの 2 人が男性で 2 人が女性であったが，全員を女性と判定した．また，音声認識では認識結果を得ることが出来ず，QR コードによるバイパスを用いた．結果としては，ヘッドディスプレイに表示した人物画像が不明瞭であったため，人物報告，位置報告の両方が認められず 0 点であった．

2 度目のトライは，1 度目にあったナビゲーションの問題点を修正してからトライした．その結果，ゲストをより近い位置から認識することが出来たため，取得画像が鮮明になり属性推定も間違いなく動作した．しかし，音声認識部においてはゲストの前までナビゲーションを行うことは出来たが，名前を聞き取ることは出来ず，再度 QR コードのバイパスを使用することとなった．2 度目のトライにおいて，フィールド内の状況を説明するために生成した画像を図 7 に示す．今回のトライでは，ゲストは図 6(b) の通りに座っており，生成画像では全員の座っている位置を間違いなく報告できている．更に，性別と名前も正解しているため，結果として満点の 1000 点を獲得した．

6　考察

6.1　ノイズ除去

本研究では，音声認識の精度向上を目的としてノイズ除去を適用した．その結果，検証データ全体では認識精度が向上していることが確認できた．しかし，検証データ 88 個のうち 3 個のデータにおいては，ノイズ除去を行うことでかえって認識に失敗するという結果となった．このことから，ノイズ除去が必ずしも認識精度の向上に繋がるわけではないということが分かった．今後は，音声認識にとってより有効なノイズ除去の手法について検討を進めていく必要がある．

6.2 音声認識

我々はRoboCup@Home2022でFMMを2度実施したが，音声認識が未検出となり結果を得ることができなかった．1つ目の原因として，音声認識時間外に発話されたことが挙げられる．HSRはマイクとスピーカが別デバイスであるため，HSRが発話している間に音声認識を行うと，HSRの音声もマイクに入力されてしまう．また，本研究で実装した音声認識はゲストの発話状態にかかわらず一定時間のみ行うため，認識精度が発話タイミングによって大きく変動してしまう．そこで，HSRのヘッドディスプレイに発話のタイミングを誘導するような表示をしていたが，この表示が発話者に伝わっておらず，認識時間外に発話されることがあった．

2つ目の原因として，発話者の近くまでナビゲーション出来なかったことが挙げられる．バンコクで実際に使用された会場では，ゲストの座っているソファの手前にテーブルがあったため，ゲストの手前まで移動することが出来なかった．そのため，遠い位置からの音声認識となり，マイクに入力される発話者の音声が非常に小さくなってしまった．このことから，音声認識の結果を得ることが困難であったと考えられる．今後は，発話のタイミングに応じて音声認識を開始，終了するような機能を作成する必要がある．また，発話者の音声が小さいことも考慮して，音声強調 [10, 11] の技術を活用する必要がある．

6.3 位置推定

提案手法では，HSRが事前に取得したマップのどこがフィールドで，どこに椅子やソファがあるのかという情報を事前に与える必要がある．RoboCup@Homeのルールでは，事前に部屋の情報が公開されることになっているが，本大会では椅子の位置が何度も変更されたため，対応が困難であった．今後は3次元的な物体認識手法 [12] を用いて，家具の位置変化に頑健なシステムを構築する必要があると考える．

7 結論

本研究では，国際的な競技会であるRoboCup@Homeで行われるFMMに向けての解法を提案し，実機実装を通してその性能評価を行った．競技会では2回目のトライで満点を獲得し，提案手法の有効性を示した．一方で，音声認識やナビゲーションに関しての課題点も見つかったため，今後はこれらの課題を解決するために研究を続けていく必要がある．

参考文献

[1] RoboCup@Home. https://www.robocup.org/domains/3, (Accessed 2022-09-01).

[2] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara and K. Murase, “Development of Human Support Robot as the research platform of a domestic mobile manipulator,” ROBOMECH Journal, Vol. 6, Art. no. 4, (2019).

[3] Google Speech-to-Text. https://cloud.google.com/speech-to-text, (Accessed 2022-09-03).

[4] Amazon Transcribe https://aws.amazon.com/jp/transcribe/, (Accessed 2022-09-03).

[5] K. Zhu and J. Wu, ”Residual Attention: A Simple but Effective Method for Multi-Label Recognition,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 184-193 (2021).

[6] α cephei Vosk Offline speech recognition. https://alphacephei.com/vosk/, (Accessed 2022-09-04).

[7] T. Sainburg, M. Thielk and T. Q Gentner, “Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires,” *Public Library of Science PLoS computational biology*, Vol.16, No.10, pp.e1008228, (2020).

[8] D. Osokin, ”Real-time 2d multi-person pose estimation on cpu: Lightweight openpose.” arXiv preprint arXiv:1811.12004 (2018).

[9] Robot Operating System Wiki. https://wiki.ros.org/, (Accessed 2022-09-01).

[10] J. Serrà, S. Pascual, J. Pons, R. O. Araz and D. Scaini, “Universal Speech Enhancement with Score-based Diffusion,” arXiv (2022).

[11] S. Welker, J. Richter, and T. Gerkmann, ”Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain”, ISCA Interspeech, (2022).

[12] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang and X. Zhou, “OnePose: One-Shot Object Pose Estimation without CAD Models,” Conference on Computer Vision and Pattern Recognition(CVPR), (2022).