

RoboCup@Home タスク：Find My Mates に向けた解法の提案とロボット実機での性能評価

Solving RoboCup@Home Task: Find My Mates and Evaluation Using Domestic Standard Robot

矢野 優雅^{1*} 松本 生弥¹ 福田 有輝也¹ 小野 智寛¹ 田向 権^{1,2}
Yuga Yano¹, Ikuya Matsumoto¹, Yukiya Fukuda¹, Tomohiro Ono¹, and Hakaru Tamukoh^{1,2}

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 ニューロモルフィック AI ハードウェア研究センター

² 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 のタスクの一つである Find My Mates に向けて、満点を取得するための手法を提案する。また、提案した手法をロボットに実装し、RoboCup@Home にて現地実験を行った。現地実験では満点を取得し、提案手法の有効性を示した。現地実験の様子は、https://www.youtube.com/watch?v=Kh_eAm3_ZVw に公開している。

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、非常に注目を集めている。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。私たちの参加している DSPL では、トヨタ社が開発した 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。

本研究では、特にヒューマンインターフェースの性能をはかる Find My Mates (FMM) というタスクに向け

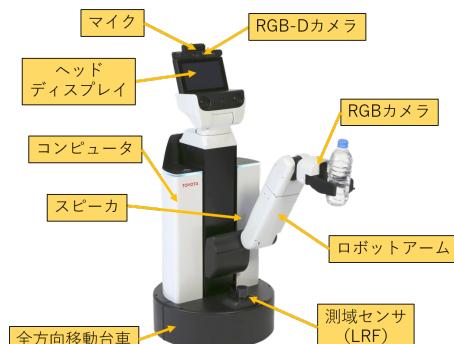


図 1: トヨタ社が開発した HSR

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。現地実験では満点を取得し、本手法の有効性を示した。

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 の得点表を表に

*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

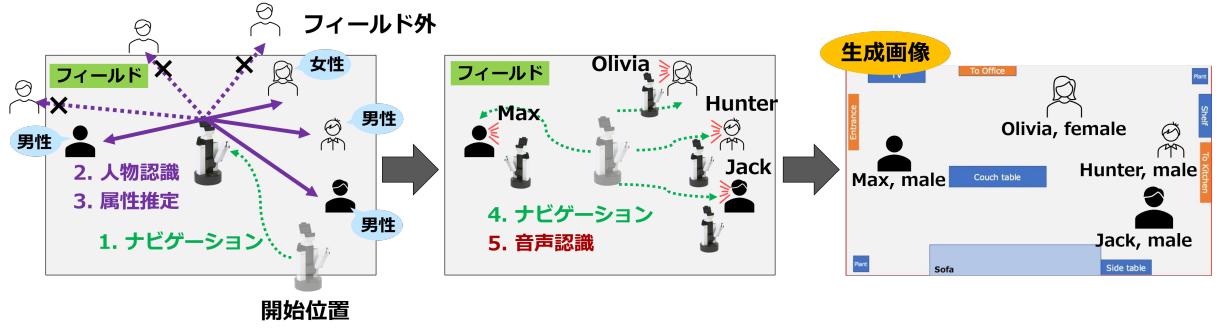


図 2: FMM の解法

示す。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、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 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。

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
ボーナス得点			
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

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。

3 提案手法

本章では、FMM で満点を取得する解法と、HSR に実装した機能について述べる。

3.1 FMM に向けた解法

私たちは、FMM で満点を取得するために、図 2 に示す手法を提案する。初めに、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行う。人物の認識には RGB 画像を用いるが、Depth 画像も用いることで、認識した人が部屋のどこにいるのかも同時に算出する。次に、算出した人物の位置情報を基に、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。更に属性推定の手法を用いて、ゲストの性別を推定する。

最後に、取得したすべての情報（人物の画像、位置、名前、性別）を集約した 1 枚の画像を生成し、HSR 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 の右側に示す。この生成画像ではフィールドを簡易的な画像で表現し、その上にゲストを示しているため、明示的に位置を報告することができている。また、4 人同時に報告が行えるため、メインゴールとボーナスを同時に獲得できる。

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、安定したクラウド上での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[5] と呼ばれるオフライン音声認識手法を用いる。

本研究では、音声認識を次のように実装している。まず、HSR のヘッド部に搭載されているマイクを用いて、音声認識を開始したタイミングから一定時間録音を行う。次に、録音した音声を PC 側へ Robot Operating System (ROS) [6] を介して送信する。しかし、ROS には音声ファイルをそのまま扱うメッセージ型がないため、本研究では音声ファイルを numpy 形式に変換して送信する。PC 側では、受信した numpy の配列から音声ファイルを再構築し、音声認識を行う。

3.2.1 辞書設定

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、本研究では名前リストを基にした辞書を作成し、vosk に適用させることで音声認識の精度向上を図る。辞書を設定していない場合では、名前を話してもまったく違う単語として認識されることがほとんどであったが、辞書設定をすることで認識率は飛躍的に向上した。

3.3 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）が大きいことである。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起きないような大きなノイズが発生する。本研究では、音声認識の精度を高めるために、ノイズ除去 [7] を音声認識の前段に組み込んでいる。

3.4 音声認識の性能評価

RoboCup@Home で使用される名前は、アメリカで一般的に用いられる名前からランダムに決定される。そこで、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名前を 11 個選出し、本研究で作成した音声認識の精度を検証した。検証はノイズの大きな環境で

行い、話者とノイズの環境を変化させながらそれぞれ 4 度ずつ読み上げて検証を行った。

表 2 に、ノイズ除去を行った場合と辞書指定を行った場合における認識結果を示す。辞書指定を行い、ノイズ除去を行った手法が最も精度が高くなっていることが確認できる。

表 2: 音声認識の精度

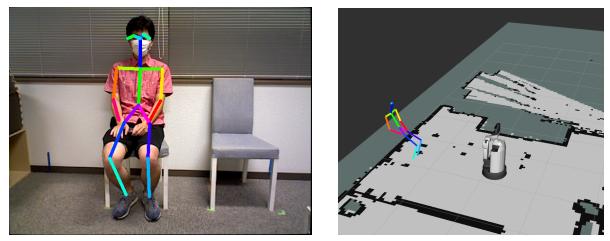
辞書指定	ノイズ除去	認識精度 (%)
なし	なし	13.6
	あり	10.2
あり	なし	69.3
	あり	71.6

3.5 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、HSR より取得した 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。

3.6 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必要があるため、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、事前に rtabmap[9] を用いて作成したマップに対して json ファイルを用いて意味づけを行う。ここで、フィールドの情報も事前に公開されるため、部屋の内外の情報と椅子などの家具がどこにあるのかといった情報も含めて意味づけを行う。図 (b) 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



(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

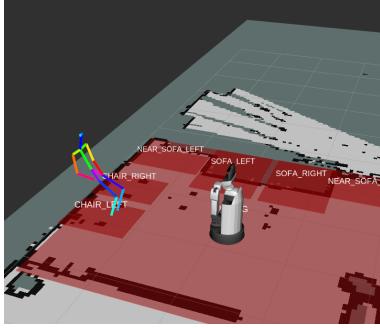


図 4: 図 3 で認識した結果にエリア判定を付加した結果

の椅子に座っており、それを正しく判定できている。

4 現地実験概要

提案手法を HSR に実装し、2022 年 7 月にバンコクで行われた RoboCup@Home で性能評価を行った。図 5 に、実際に使用されたフィールドを示す。4 つのルームがある中で、FMM はリビングルームにて実施された。現地の実際の写真を図 6 に示す。

また、本研究で使用した PC 環境は、CPU : Intel core, GPU : Geforce RTX 1080, メモリ : 64GB, OS : Ubuntu18.04 である。

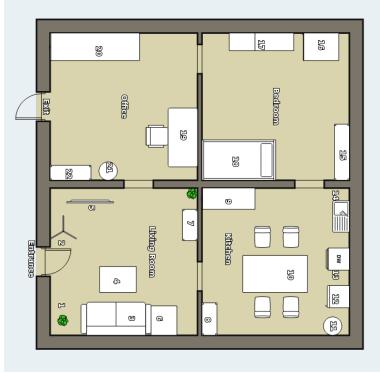


図 5: バンコクで開催された RoboCup@Home2022 で使用されたフィールド



図 6: FMM が行われた実際の会場

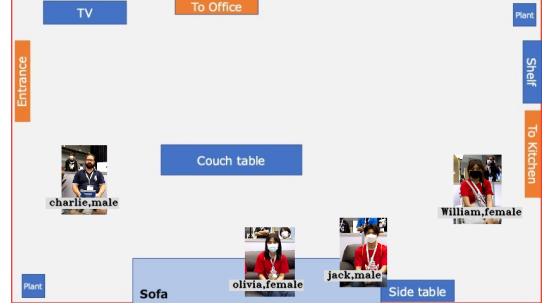


図 7: 2 回目のトライで作成したマップイメージ

5 現地実験結果

私たちは 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 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、鮮明な画像を得ることができ、属性推定も間違なく動作した。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、また 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 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 点を取得した。

6 考察

6.1 音声認識について

今回の現地実験では、FMM を 2 回実施したが、いずれも音声認識の結果を得ることは出来なかった。1 回目の原因として、音声認識時間外に発話されたことが挙げられる。まず、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、発話のタイミングが音声認識の結果に大きく影響してしまう。そこで、HSR のヘッドディスプレイに認識中を示すような GUI を作成していたが、この GUI 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で移動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と考える。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と考える。

6.2 位置推定について

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。今後はロバスト性の高い3次元的な物体認識手法[12, 13]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。現地実験で、2回目のトライで満点を取得し、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] Yamamoto, T., Terada, K., Ochiai, A., Saito, F., Asahara, Y., and Murase, K. “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 音声をテキストに自動的に変換する. <https://aws.amazon.com/jp/transcribe/>, (Accessed 2022-09-03).
- [5] a cephei Vosk Offline speech recognition. <https://alphacepheli.com/vosk/>, (Accessed 2022-09-04).
- [6] Robot Operating System Wiki. <https://wiki.ros.org/>, (Accessed 2022-09-01).
- [7] Sainburg, T., Thiels, M., and Gentner, T. Q., “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] Osokin, D. ”Real-time 2d multi-person pose estimation on cpu: Lightweight openpose.” arXiv preprint arXiv:1811.12004 (2018).
- [9] M. Labb  and F. Michaud, “RTAB-Map as an Open-Source Lidar and Visual SLAM Library for Large-Scale and Long-Term Online Operation,” in *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, 2019.
- [10] Serr , J. and Pascual, S. and Pons, J. and Araz, R. O. and Scaini, D. “Universal Speech Enhancement with Score-based Diffusion,” arXiv (2022).
- [11] Welker, S., Richter, J., and Gerkmann, T. “Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain”, ISCA Interspeech, (2022).
- [12] Garrick Brazil and Julian Straub and Nikhila Ravi and Justin Johnson and Georgia Gkioxari, “Omni3D: A Large Benchmark and Model for 3D Object Detection in the Wild,” arXiv:2207.10660, (2022).
- [13] Sun, Jiaming and Wang, Zihao and Zhang, Siyu and He, Xingyi and Zhao, Hongcheng and Zhang, Guofeng and Zhou, Xiaowei, “OnePose: One-Shot Object Pose Estimation without CAD Models,” Conference on Computer Vision and Pattern Recognition(CVPR), (2022).