

Find My Mates に向けた解法の提案と実機での性能評価

Solution of Find My Mates and evaluation on Domestic Standard Robot

矢野 優雅^{1*} 福田 有輝也¹ 小野 智寛¹ 田向 権^{1,2}
Yuga Yano¹, 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、2022 年 7 月にバンコクで行われた 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。

1 序論

1.1 RoboCup@Home

RoboCup@Home 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動的環境におけるテストが行われている。そのため、より現実環境を想定した性能評価をすることができ、非常に注目を集めているリーグとなっている。RoboCup@Home には、Open Platform, Domestic Standard Platform (DSPL), Social Standard Platform 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おり、認識を通して多様なヒューマンインタラクションを行うことができる。

本研究では、ヒューマンインタラクションの性能をはかる Find My Mates というテストに向けて、その解法を提案するとともに、HSR への実機実装を行い

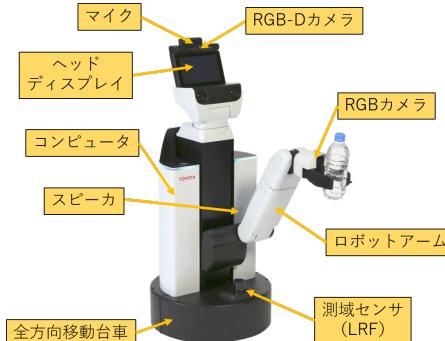


図 1: トヨタ社が開発した HSR

RoboCup@Home での性能評価を行う。

1.2 Find My Mates

本章では、RoboCup@Home で行われる Find My Mates (FMM) というタスクについて述べる。FMM では、4人のゲストが1人のホストを訪れたという状況を想定している。FMM は、1人のホストの家に訪れた4人のゲストをロボットが探し、その場所、名前に加えて人物の特徴をホストに報告するというタスクである。そのため、人物を3次元的に認識する技術と、それぞれのゲストの特徴を抽出する属性推定の技術が必要に

*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.yuuga15@mail.kyutech.jp

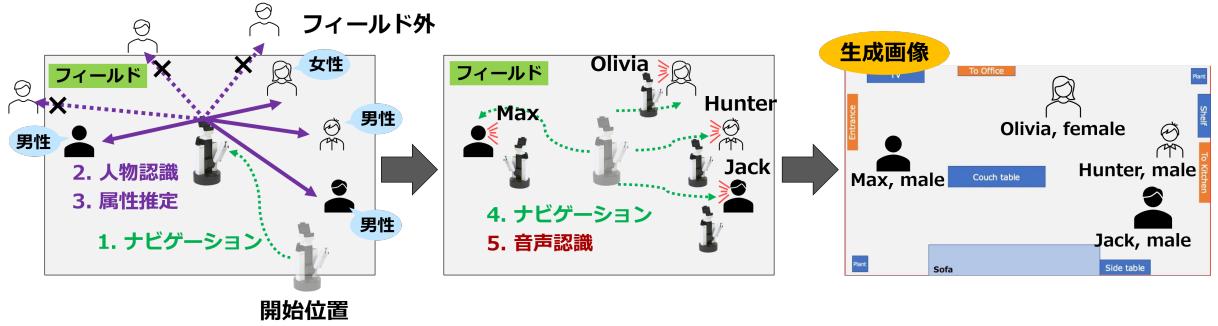


図 2: FMM の解法

なる。更に、ロボットは事前にゲストの名前を知らされていないため、音声認識を通してゲストの名前を知る必要がある。

2 関連研究

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

私達は FMM で満点を取得するために、次のような解法を提案する。提案する手法の概要を図 2 に示す。初めに、ロボットを部屋の中央までナビゲーションを行い、部屋全体を見渡しながら人物認識を行う。ここで、Depth 画像も用いることで、認識した人が map 上のどこにいるのかを算出する。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、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 枚の画像を作成し、ヘッドディスプレイに表示することでホストに伝える。

3.2 音声認識

近年ではスマートフォンなどの普及により、Siri などのクラウドを用いた音声認識の精度が非常に高くなっている。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の音声認識が困難である。また、ネットワークの課題は一般的家庭環境においても想定されるものであり、オフラインでの音声認識技術を利用すること

は非常に有効である。そこで本研究では、vosk[5] と呼ばれるオフラインの手法を用いて音声認識を行う。

本研究では、音声認識を図のように実装している。HSR のヘッド部に搭載されているマイクを用いて、一定時間録音する。次に、録音した音声を ROS を介して PC に送信する。ここで、ROS には音声ファイルをそのまま送信できるメッセージ型がないため、音声ファイルを numpy の配列に読み直して送信を行う。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

RoboCup@Home では、タスクに登場する人物は本名を使用するのではなく、事前に公開している名前リストから毎回ランダムに決定され名前を割り当てる。この名前リストには、男性用と女性の用の名前が約 10 個ずつ用意されている。ただし、名前だけで性別の区別ができないようにするために、男性と女性で共通している名前も存在する。今回は、この vosk を使用する際に名前リストをもとにした辞書を作成し、名前の音声認識精度向上を行う。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、周囲のノイズが大きいことである。RoboCup@Home の他にも、サッカーリーグやレスキューリーグが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 [6] を音声認識の前段に組み込み、精度を高めている。

表 1: 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

3.4 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、本研究で作成した音声認識の性能を評価するため、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、認識精度を検証した。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 1 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。

3.5 人物認識

本研究では人物認識の手法に Lightweight Human Pose Estimation[7] を用いた。本手法は処理が非常に軽量であり、CPU でも高速に動作できる手法である。今回使用した PC では、x fps で動作しており、HSR のカメラ周期と同等な速度である為、リアルタイム動作を実現できていると言える。本研究では、人物を 3 次元的に認識する必要があるため、RGB 画像から人物認識を行った後、認識位置の Depth 画像を参照することで、3 次元的な位置を推定する。図 3 に、人物の 3 次元的な位置推定を行った結果を示す。

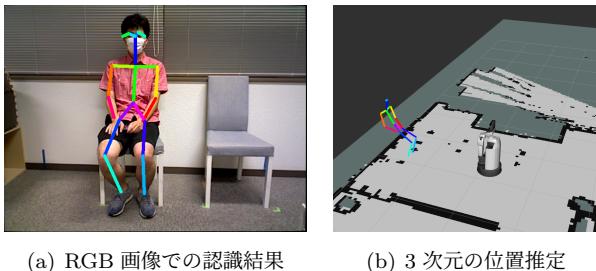


図 3: 人物位置推定アルゴリズム

3.6 口けーション報告

FMM では、認識した人物をホストに伝える必要があるが、この際に考慮すべきこととして、認識した人

が本当に部屋の中にいる人なのか、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、事前に作成しているマップに対して json ファイルを用いて意味づけを行い、フィールド内判定を行う。また RoboCup@Home では、事前に部屋の形が公開されるため部屋の中のどこに椅子があるのかという意味づけも行う。人物のフィールド内判定と、位置推定の結果を図 4 に示す。この場合では、ゲストはフィールド内

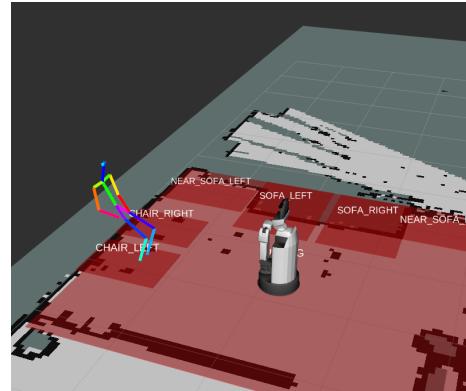


図 4: 図 3 で認識した結果にエリア判定を付加した結果

の左側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

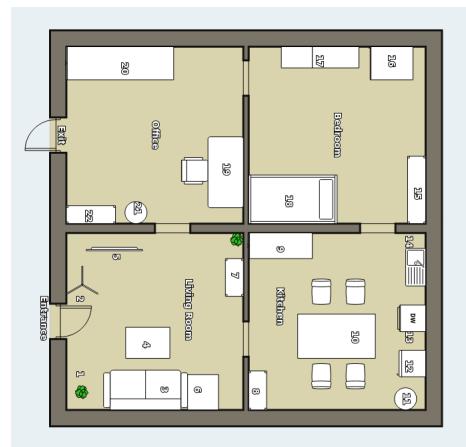


図 5: バンコクで開催された RoboCup@Home2022 で使用されたフィールド

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。



図 6: FMM が行われた実際の会場

5 現地実験

5.1 実験概要

RoboCup@Home の DSPL では, HSR に PC を接続して使用することが認められている。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。

5.2 実験結果

1 度目のトライでは, ナビゲーションの目的地がゲストから遠い位置であったため, 遠距離から認識することとなり, 不鮮明・不明瞭な画像を取得することになった。人物検出と 3 次元の位置推定は正常に動作したが, 画像が不鮮明であったため, 属性推定に失敗した。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。結果は, 属性推定の結果があつてないこと, 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, 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目的地が遠いという問題点を修正してから行った。その結果, 人物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ていたが, 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。

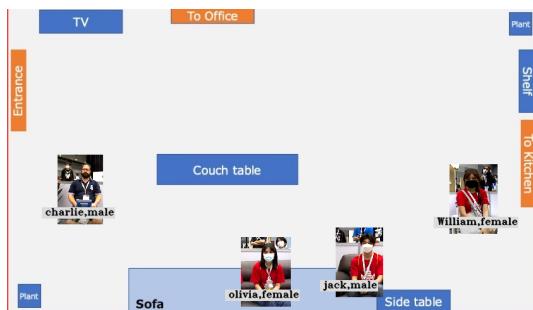


図 7: 2 回目のトライで作成したマップイメージ

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 つの原因として, 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 の音声も認識してしまう恐れがある。また, 音声認識は発話の有無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, 遠い位置からの認識となり, 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 [8, 9] の技術を活用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する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] Author, A., Author, B.: JSAI SIGs Conference Paper Format Sample, *International Journal of Examples*, Vol. 19, No. 4, pp. 1–2 (2007)
- [4] 第一著者, 第二著者: 人工知能学会研究会原稿フォーマットサンプル, *International Journal of Examples*, Vol. 19, No. 4, pp. 1–2 (2007)
- [5] <https://alphacepheli.com/vosk/>
- [6] Sainburg, T., Thielk, 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.
- [7] Osokin, D. ”Real-time 2d multi-person pose estimation on cpu: Lightweight openpose.” arXiv preprint arXiv:1811.12004 (2018).
- [8] Serrà, J. and Pascual, S. and Pons, J. and Araz, R. O. and Scaini, D. “Universal Speech Enhancement with Score-based Diffusion,” arXiv (2022).
- [9] Welker, S., Richter, J., and Gerkmann, T. ”Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain”, ISCA Interspeech, (2022).