

ジェスチャと音声を用いた家庭内ロボットのためのマルチモーダルインタラクションシステム

Multimodal Interaction System using Gesture Recognition and Voice Recognition for a home robot

下川原(佐藤) 英理 (首都大) 〇甄 継国 (首都大)
正 山口 亨 (首都大)

Eri sato-Shimokawara, Tokyo Metropolitan University, eri@sd.tmu.ac.jp
Jiguo Zhen, Tokyo Metropolitan University, zhen-jiguo@sd.tmu.ac.jp
Toru Yamaguchi, Tokyo Metropolitan University, yamachan@tmu.ac.jp

This paper proposes a system that the robot in the system shares information of objects by using Multimodal Interaction between human and robot with recognizing human's action and voice. Multimodal Interaction works by combining indicating and one word utterance which are the results of gesture recognition and voice recognition. Proposed system obtains an object's place as three dimensional coordinates by indicating motion, and also obtains the object's name, color, and shape by one word utterance. Proposed system manages the objects by storing, deleting, and searching information of the objects. In this study, examinations showed the proposed system is effective.

Key Words: Multimodal Interaction, gesture recognition, voice recognition

1. 緒言

近年、ロボット技術の発展に伴い、一般家庭や公共の場でもロボットのサービス提供が考えられるようになってきている。しかし、このような過程や公共の場で利用されるロボットは様々なサービスの提供を求められるため、多機能化・複雑化する傾向がある。そこで、日常生活でロボットによるサービス提供を行うに当たって、直観的に使用できるようなインタフェースを用いることが重要であると考え、人と人とのインタラクションに着目した。

人と人とのインタラクションでは、指差し・傾き・首振りなどの手振り身振りといった動作からなるノンバーバルコミュニケーションと、発話や効果音などの音声からなるバーバルコミュニケーションが利用されている。Satoら[1]は指差し動作を用いたロボットの操作を提案している。Niitsumaら[2]は空間へ記憶を残す空間メモリのインタフェースとして指差しを用いている。またジェスチャに加えて音声を用いたロボットとの対話型インタフェースの研究も進んでいる[3][4]。

本研究ではジェスチャ動作のうち、指差し動作に着目し、動作認識による「指差し」と音声認識による「単語発話」とを合わせたマルチモーダルなインタラクションシステムを開発した。さらにこのシステムを利用した、対象物の位置や名前の情報を人とロボットとが共有するシステムを提案する。

2. システム構成

提案するシステムの構成を図1に示す。提案システムでは、人とロボットとがインタラクションを行うにあたって、指差しと単語発話を利用する。指差しの動作認識にはステレオカメラを、単語発話の音声認識にはマイクを用いる。また、人へ問い合わせや確認を行うために、インタラクションロボットを用いる。インタラクションロボットには、図1に示す東芝製インタラクションロボット「ApriPoco」を用いる。ApriPocoは東芝が研究・開発を進めているインタラクションロボットであり、頭部・腕部・腰部の7自由度を持っている。また、頭部には2つのカメラを持ち、内部に組み込まれている制御用PCでそれらを統括して動作する[5]。また、音声によるインタラクションのため、音声認識にはJuliusを、音声

合成にはLaLaVoiceを用いる。ユーザの動作認識にはApriPoco頭部の2つのカメラによるステレオ計測を用いる。

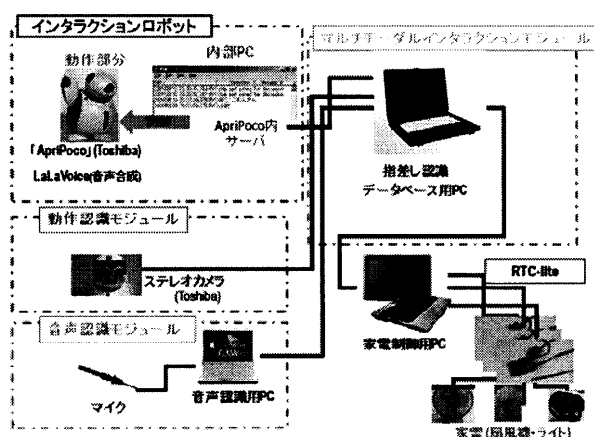


Fig. 1 System outline

2.1 動作認識モジュール

動作認識モジュールでは動作の1つとして「指差し」を認識し、指差し位置の座標を出力する。動作認識の方法として、ステレオカメラから取得した画像を用いた特徴点認識を利用する。本研究では、指差し位置の算出を動作認識モジュールで行う。画像の取得には東芝製ロボットApriPocoの頭部にとりつけられたステレオカメラを利用する。動作認識モジュールで行う特徴点認識にはOpenCVを用いた。SURF (Speeded Up Robust Feature) による局所特徴点抽出を行い、背景差分処理を行う。特徴点はステレオカメラの中心を原点とした三次元座標で得られる。カメラから見て右方向をx軸の正方向、上方向をy軸の正方向、前方がz軸の正方向である。

カメラからz方向に500[mm]間隔で領域を分割し、最も特徴点の数が多い領域を人のいる領域とする。人のいる領域のなかで最も高い点を頭とし、頭から最も遠い点を手とした。得られた頭と手の座標の延長線と、机や壁にあたる面との交点

を指差し座標とする。ただし、手を上げただけで指差しと判断しないように、指差し座標を一定時間取得し、その前後の座標を比較した加速度が一定以下であれば指差しをしていると判断する。指差しをしていると判断した場合は、最新の指差し座標を認識結果としてモジュールから出力する。

本研究で実装した動作認識の指差し位置の精度を測るため、図 2 に示す環境で扇風機とライトを指差してもらい、その実際の位置と指差し座標とのずれを測定した。

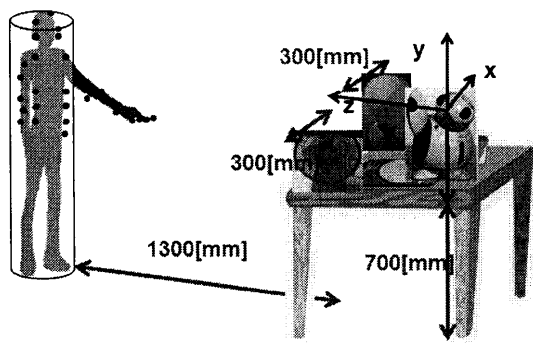


Fig. 2 Measurement Environment of Pointing Precision

測定における指差し位置の結果を図 3 に示す。なお、指差しを行った場合に指差しであると判断する、指差しの認識率は100%であった。指差し位置の座標と実際においた位置とのずれから誤差を求めた。その結果、誤差の平均が291[mm]、最大の誤差が417[mm]であった。この最大誤差を対象物の範囲とし、指差し座標が対象物の範囲よりも外側にあるときは移動、内側にあるときはその対象物を指差ししていると判断する。ただし、精度の測定環境におけるライト・扇風機の位置が[cm]単位が有効桁数であるため、420[mm]を提案システムでの対象物の範囲とした。

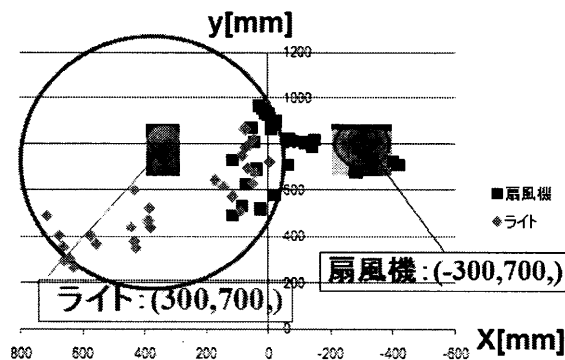


Fig. 3 Result of Pointing Precision

2.2 音声認識モジュール

音声認識モジュールでは「単語発話」を認識し、認識した単語を出力する。動くロボットの近くでは、人は危険性を感じる。人とロボットとがインタラクションを行う時には、一定の距離を保つ必要がある。人同士であっても、親密さによって距離を変えるが、一般的に友人程度の親しい人との距離

にあたる親密距離が 0.45[m]から 1.2[m]程度、初対面の人と話す社会距離では 1.2[m]から 3.5[m]程度と言われる[6]。3.5[m]以上離れている場合、一方的な講演などを行うような距離となるため、インタラクションを行うには不適切である。本研究では、ロボットとの距離が友人でないながらも親しみを覚える程度の距離として 1.3[m]の距離で人とロボットとのインタラクションを行うこととする。この距離であれば、お互いが動くことによって接触する危険がないため、人が安全性を感じられる距離としても十分である。ただし、1[m]以上離れた位置では、長い会話を行うのには向いておらず、一言一言で会話を行うのが望ましい。そこで、本研究では音声認識精度向上のため、単語による発話の認識を用いる。

本研究では、単語発話の認識を音声認識モジュールで行う。音声の取得には ApriPoco の前に設置した外部マイクを用いる。

音声認識モジュールで行う単語発話の認識には、オープンソースの音声認識エンジンである Julius を用いる。Julius は、RT コンポーネントに対応しているオープンソースソフトウェア OpenHRI の一部として組み込まれているものを利用した。音声認識モジュールはマイクから取得した音声と、登録されている単語の音韻とを比較し、発話単語を取得する。本研究では登録単語に対しての利用するコードとして、3 文字の英字に変換し、単語のコードとして取り扱う。

2.3 マルチモーダルインタラクションモジュール

マルチモーダルインタラクションモジュールでは、指差しと単語発話を融合することで人の指示を認識する。また、マルチモーダルなインタラクションを行うためには、ロボット側からも動作・音声を用いて人にアクションを行う必要がある。

動作認識モジュールによる指差しの認識手法では、対象物を指し続けている場合、連続的に座標が入力される。そのため、発話を、指示を認識するタイミングのトリガーとして利用する。そこで、まず指差しと発話の時間的な関係についての実験を行った。

実験では被験者 5 名に対して、2 つの対象物を順番に指差ししながら発話してもらった。実験の結果は図 4 に示すように、発話と関係している指差しのタイミングは、発話を基準として前 0.8[秒]、後 0.3[秒]であることがわかった。この知見から、音声認識モジュールの入力の前 0.8[秒]から入力後 0.3[秒]の動作認識モジュールを指差しの座標として利用することとした。その間に 2 回以上指差し座標が取得された場合はそれらの座標の平均座標を指差し座標として扱う。

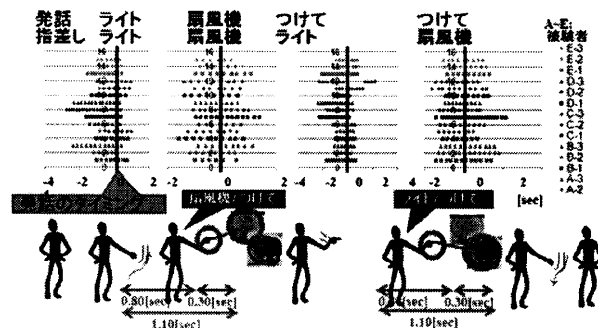


Fig. 4 Relation of Timing between pointing and voice

マルチモーダルインタラクションモジュールの構成と処理の流れについて、図 5 に示す。マルチモーダルインタラクションモジュールではまず、先に述べたように、動作認識による指差しと音声認識による発話を受け取って、それぞれの認識の結果を合わせて出力する。そして環境情報データベースを参照し、指差しと発話単語の内容と統合して人の指示を判断する。

本稿では、環境情報との統合のために、基本機能として実装した、登録・削除・検索について述べる。

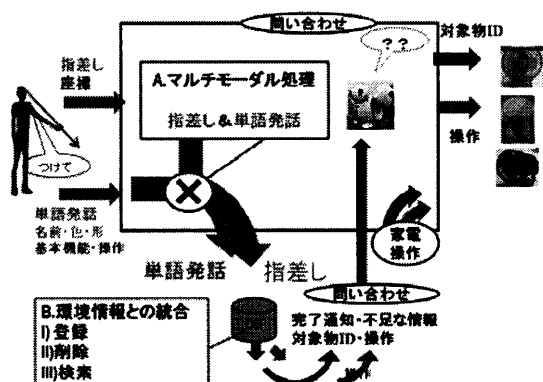


Fig. 5 Relationship of timing between pointing and voice

3. 対象物の情報共有

どこに何があるか、という情報は家庭内ロボットがサービスを行う上で重要な情報であると考えられる。IC タグや屋内 GPS を利用した位置管理が研究されているが、タグリーダの無い場所や電波の届かない場所では利用できない。そこで、本研究では、人とのインタラクションから物の位置を登録・削除するシステムを構築した。また、ロボットがどのような情報を持っているか確認するため、検索機能も構築した。今回の対象物はネットワークから制御可能な家電とした。

情報共有部において、対象物である家電に関する情報の管理と制御を行う。対象物の情報共有部は、データベースを用いて対象物の情報を管理する。

3.1 登録・削除・検索機能

対象物の情報共有のために、対象物の情報を管理するための基本機能として登録・削除・検索という 3 つの機能を実装した。これらの機能はそれぞれ「登録」・「削除」・「検索」という単語発話が入力されることで実行を開始する。

登録では、指差し座標と単語発話の名前を利用する。得られた名前を対象物を、指差し座標を位置として登録する。削除では、指差し座標と単語発話によって対象物を特定し、その対象物の情報をデータベース上から削除する。検索では、指差し座標を利用し、指差しした位置に対象物があるのかどうか検索を行う。そして、検索結果をロボットに発話させる。図 6 に対象物の情報共有のための各基本機能のイメージを示す。

3.2 ロボットとのインタラクション

まず、対象物の登録について述べる。登録を行うには、最初に単語発話として「登録」と発話する。この時点では対象物の名前が不明なため、何を登録して欲しいかが不明である。提案システムはロボットを利用して登録したい対象物の名前を人に問い合わせる。名前を把握した後、データベースを検索し、同じ名前を対象物がない場合は登録を行う。しかし、

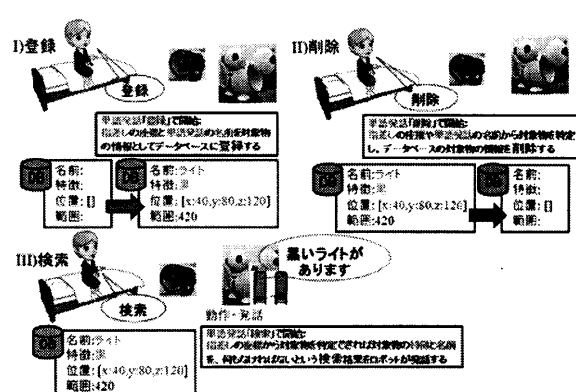


Fig. 6 Basic function of information share for subjects

既にデータベースに同じ名前を対象物がある場合は対象物の特徴を問い合わせ、対象物の絞り込みを行う。登録では、指差し座標を対象物の座標として登録する。

削除は、単語発話の「削除」に応じて処理を開始し、登録と同様、対象物を特定するまで名前と特徴の問い合わせを行う。そして特定された対象物の情報をデータベース上から削除する。なお、削除すべき対象物が存在しない場合は、削除するものが何もない旨を確認し、処理を終える。

検索は、データベース上に対象物の位置を確認するために利用する。処理の開始には「検索」の単語発話を用いる。指差し座標から、データベース上に格納されている対象物を特定する。対象物が特定できない場合には何も無いということ、特定できた場合には特定した対象物の名前を発話する。ただし、指差しの座標から対象物が複数あると判断された場合は、指差し先にある対象物の名前を発話する。また、データベース上に同じ名前を対象物が複数ある場合には、その特徴を含めて発話する。

4. 実験

3 節で述べた対象物の情報共有システムを用いて実験を行った。

4.1 対象物の登録

対象物の登録を行う実験として、部屋において初めて提案システムを用いた家電の操作を行う状況を想定する。対象物の登録の実験の様子を、図 7 に示す。初期の状態であるため、データベース上には何も登録されてない。また、システムには扇風機とライトが接続されている。初めにユーザはロボットに対して「登録」と発話を行った。この時点では、登録したい対象物が「何」であるか不明なので、システムが基本機能の登録を実行するために、対象物の名前が必要となる。そこで、ロボットによってユーザに何をつけるのか、という意味を持つ問い合わせとして「サーバーに名前が必要です」と発話を行うことで、ユーザに対象物の名前の発話を促した。ユーザが指差ししながら「扇風機」と答えたため、登録したい対象物の特定ができる状態となった。そこでシステムは対象物の扇風機を指差しの位置に登録した。登録後に、ユーザは操作を行うため、指差ししながら「つけて」と発話した。登録した扇風機が範囲内にあるため、システムはユーザが扇風機をつけようとしていると判断し、扇風機をつける。このとき、ロボットによって扇風機をつける旨を発話し、ユーザにシステムの動作と意図を伝える。同じように、ライトについても登録し、操作を実行した。

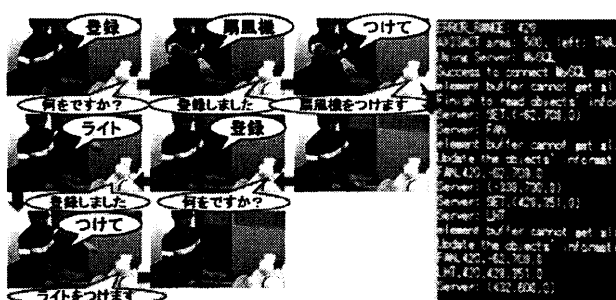


Fig. 7 Experiment scene -registration-

4.2 対象物の削除

4.1 節で行った実験で登録した白いライトと扇風機の情報を引き継いでいる。この実験では、引っ越しにより、扇風機が不要になり、新しく別の色の黒いライトが有用になった状況を想定する。

対象物の情報の削除の実験の実際の様子を図 8 に示す。引っ越しした時の状態であるため、データベース上には 4.1 節の情報が残っている。また、システムには色が異なる 2 つのライトが接続されている。扇風機が不要になったため、情報のみが残る状態である。ユーザはもともとあった白いライトを指差しながら「つけて」と発話を行った。ところが、もともと扇風機のあった範囲と、白いライトの範囲とが重なっている位置を指差してしまった。そのため、ロボットは扇風機とライトのどちらをつけるかと問い合わせを行った。ユーザがライトと答えるのに応じてライトをつけた。ユーザはこのインタラクションを通して扇風機の情報が残っていることを把握したが、扇風機を既に使うことはない状態で、扇風機の情報は不要である。そこでユーザは「削除」と指示し、問い合わせに対して「扇風機」と答えることで扇風機の情報を削除した。また、新規に黒いライトをつないでおり、黒いライトを利用するために登録を行った。データベースには既に白いライトが存在しているため、同じ名前の対象物が複数ある場合の登録として、システムは特徴の問い合わせを行った。黒という特徴によって区別が可能になったため、対象物の登録ができた。また、その後に黒いライトと白いライトがある状態でも、指差しによる座標によって対象を指示することで、発話は指示のみで名前を発話することなくライトを利用することができることも示した。

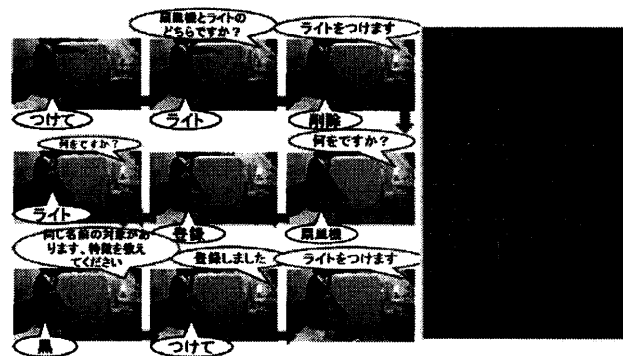


Fig. 8 Experiment scene -deletion-

4.3 対象物の検索

対象物の検索の実験では、4.2 節で行った引っ越しした時の実験の後として、黒と白のライトの情報を引き継いでいると想定した。

対象物の検索の実験では、以下の図 9 に示す流れで実験を行った。引っ越し後に扇風機の削除と黒いライトの登録を行った後の状態でデータベース上の情報を引き継いでいる。引っ越し後しばらくして、他の人も使ったという状況を想定する。使ったのは自分だけでないため、ロボットが記憶している対象物の情報の位置が変わっている可能性を考えて、机の周辺を検索した。検索結果として、ユーザは何もないところと、白いライトがあるところが把握できた。また、白いライトがあるとわかった位置で、そのまま指差しながら「つけて」と指示し、実際に操作を行った。

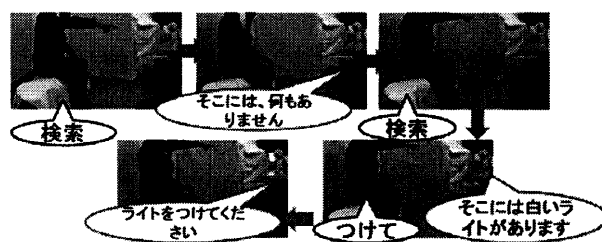


Fig.9 Experiment scene -search-

5. 結言

本研究では、人とロボットとがインタラクションにより対象物の情報共有を行うシステムの提案した。提案システムにおいて利用するマルチモーダルなインタラクションの手法について、動作認識・音声認識およびそれらを合わせて利用する手法について述べ、その精度の検証をおこなった。また、マルチモーダルインタラクションを用いて対象物の情報を共有するために、対象物のデータについて定義し、登録・削除・検索という 3 つの機能と家電の操作について述べた。

一般家庭におけるサービス提供として、家電の操作を例に、登録・削除・検索という機能に沿って 3 つの実験を行い、人とロボットとが対象物の情報を共有するシステムの有効性を示した。

文 献

- [1] Eri Sato, Toru Yamaguchi, and Fumio Harashima, "Natural Interface Using Pointing Behavior for Human-Robot Gestural Interaction", IEEE Transactions on Industrial Electronics, Vol.54, No.2, 2007
- [2] Mihoko Niitsuma, Hiroshi Hashimoto, and Hideki Hashimoto, "Interaction with spatial memory by using human indicator." In Proceedings of the SICE Annual Conference, pp.2330-2335, 2005.
- [3] Richard A. Bolt, "put-that-there": voice and gesture at the graphics interface. Computer Graphics (ACM), 14(3):262-270, 1980.
- [4] Shaolin Qu, Joyce Y. Chai, "Beyond Attention: The Role of Deictic Gesture in Intention Recognition in Multimodal Conversational Interfaces" in International Conference on Intelligent User Interfaces (IUI) 2008 pp 237-246
- [5] 柴野友美, 村上和雅, 藤本泰成, 山口亨, "ロボットを用いた発話行為と同調動作による心身活性化支援", 第 10 回 (社) 計測自動制御学会システムインテグレーション部門講演会 (SI2009), pp.1193-1194, (2009)
- [6] Edward T. Hall: "The hidden dimension", 日高敏隆, 佐藤信行 (訳): "かくれた次元", PP160-181, みすず書房