実世界のマルチモーダル情報に基づく指示語を含んだ言語指示の外部照応解析

大山瑛[1]　長谷川翔一[1]　中川光[1]　谷口彰[1]　萩原良信[1]　谷口忠大[1]
[1]立命館大学
{oyama.akira,hasegawa.shoichi,nakagawa.hikaru}@em.ci.ritsumei.ac.jp
{a.taniguchi,yhagiwara,taniguchi}@em.ci.ritsumei.ac.jp

概要

ロボットが家庭環境で生活支援を提供するには，指示語を含む曖昧な言語指示から対象を特定する外部照応解析が重要になる．内部照応解析と異なり外部照応解析はテキスト外の情報を求めるため実世界の情報を総合的に用いて曖昧性を解消することが重要になる．本研究では，実世界のマルチモーダル情報を用いた外部照応により，指示語を含む言語指示における曖昧性の解消を目指す．具体的には，物体カテゴリ，指示語，指差しの 3 つの情報とロボットが事前に環境を探索して得た物体知識を用いて外部照応を行う．家庭環境を模したフィールドにおいて，ユーザが指示した物体を特定するタスクを実施し，複数の条件において物体特定の精度を評価した．この結果，提案手法は複数のモダリティ情報を用いた外部照応により，ベースライン手法よりも高い精度で物体が特定できる事が明らかになった．

1　はじめに

「あれ取って」などのユーザの曖昧な言語指示を現場環境の情報に基づいて理解し，具体的なタスクを達成する事は，生活支援ロボットにおいて重要な課題である [1]．従来，自然言語処理の分野において照応解析の研究が実施されているが，その多くは言語情報のみを文脈として扱っている [2–4]．照応解析とは代名詞や指示語のような照応詞の指示対象を推定することであり，内部照応と外部照応の 2 種類がある．内部照応は文章中の照応詞の前後の文章からその照応詞にあたる語を予測することである．一方，外部照応はその場の環境のものから照応詞に対応するものを予測することである．答えや手掛かりが文章中にある問題の場合は内部照応により解決できるが，「あれ取って」などの指示語を含んだ

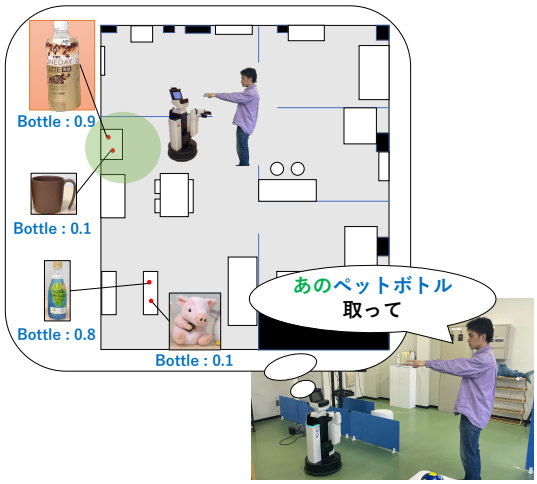


図 1　本研究の概要図．ユーザからの曖昧な言語指示と環境内を事前に探索して得た情報からユーザが指示した物体を特定する．緑色の円はユーザが指示した物体が存在する確率が高い領域を表す．この場合は赤い四角で囲まれたコーヒーを指示物体として特定した．

言語指示を理解するには，言語以外の情報を参照する外部照応を解く必要がある [5]．また，指示語を含んだ言語指示では，指差しなどのジェスチャーも重要な情報となり，ジェスチャー情報に基づいて対象を推定する研究が報告されている [6–9]．Chen らは，画像と言語の大規模データセットから対象を特定する深層学習モデルを提案したが，画像内にユーザと対象物体が含まれる制約がある [9]．Hu らは，物体と言語，指差し情報から空間における対象の位置を推定する手法を提案したが，指示語の情報は推定に用いられていない [7]．

本研究では，実世界のマルチモーダル情報を外的な文脈として用いて，指示語を含んだ言語指示の外部照応解析の実現を目指す． 具体的には，物体カテゴリ，指示語，指差しの 3 つの情報を用いた外部照応解析により，指示語の指示対象における曖昧性を解消する手法を提案する．図 1 に本研究の概要図を示す．ここで，物体カテゴリ情報は言語情報内の物

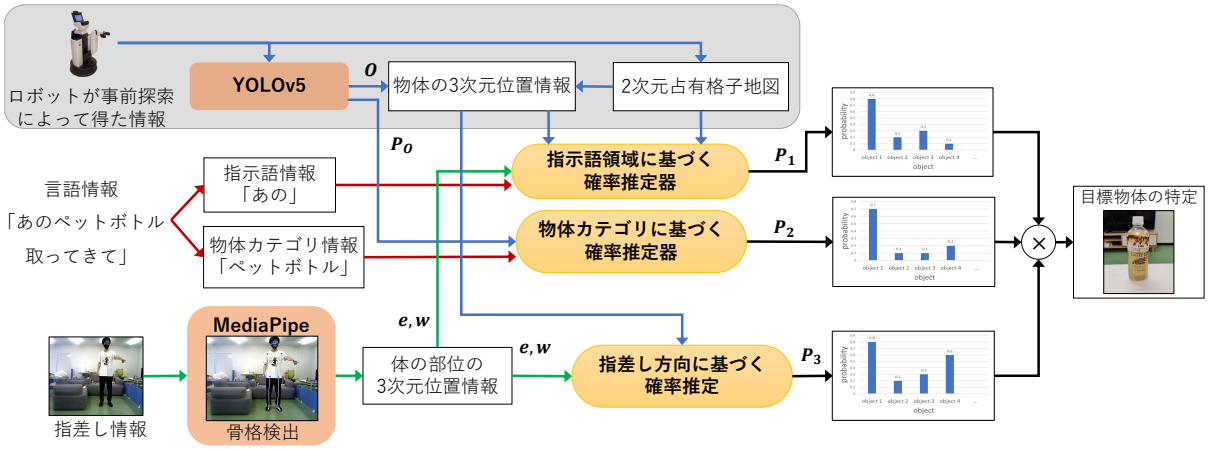


図 2　提案手法の概要図．矢印の色はロボットのセンサ情報の流れを表す．青色矢印はロボットが家庭環境の事前探索により収集した情報，赤色矢印は言語情報，緑色矢印は視覚情報である．$\boldsymbol{O}$ は物体の 3 次元位置，$\boldsymbol{P_O}$ は各物体の信頼度スコア，$\boldsymbol{e}$ はユーザの目の 3 次元位置，$\boldsymbol{w}$ はユーザの手首の 3 次元位置，$\boldsymbol{P_1}, \boldsymbol{P_2}, \boldsymbol{P_3}$ は各推定器から出力される確率である．ロボットは環境で事前に観測した情報と指示語情報，物体カテゴリ情報，指差し情報を用いて 3 種類の推定器からそれぞれ確率を得て，それらを乗算することで目標物体を推定する．

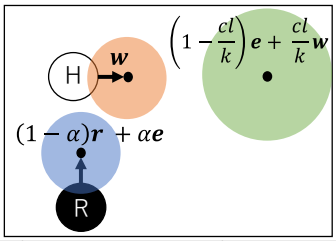
表 1　指示語領域の予測に用いられる変数の定義

$\boldsymbol{e} = (x_e, y_e, z_e)$	ユーザの目の 3 次元位置
$\boldsymbol{w} = (x_w, y_w, z_w)$	ユーザの手首の 3 次元位置
$\boldsymbol{r} = (x_r, y_r, z_r)$	ロボットの足元の 3 次元位置
α	ユーザとロボット間の内分点を決定する実数値
β	ガウス分布の共分散を決定する実数値
l [m]	ユーザとロボット間のユークリッド距離
k [m]	ユーザの目と手首間のユークリッド距離
c	ア系列の平均と分散を決定する実数値
E_3	3×3 の単位行列

体カテゴリを表す語，指示語情報は言語情報内の指示語を表す語とする．例えば，図 1 のように “あのペットボトル取って” という発話がされた時，“ペットボトル” が物体カテゴリ情報，“あの” が指示語情報に該当する．また，指差し情報はユーザが指をさしている画像とする．実験では，指示語を含む言語命令から目標物体を特定するタスクを複数の条件で実施し，物体特定の精度をベースライン手法 [7,9] と比較し，提案手法の特性と課題を明らかにする．

2　提案手法

本研究では指示語情報，物体カテゴリ情報，指差し情報の 3 種類の外部照応情報からユーザがロボットに指示した物体を特定する手法を提案する．提案手法の概要図を図 2 に示す．提案手法ではロボットの事前探索で収集した情報と外部照応情報を 3 種類の推定器に入力する．各推定器は，全ての候補物体についてその物体が指示対象である確率を推定する．ここで各推定器が出力した候補物体毎の確率を対象確率と呼ぶこととする．



指示語	平均	分散共分散行列
コ系列	$\boldsymbol{w}$	$\frac{\alpha}{\beta}l\boldsymbol{E_3}$
ソ系列	$(1-\alpha)\boldsymbol{r}+\alpha\boldsymbol{e}$	$\frac{\alpha}{\beta}l\boldsymbol{E_3}$
ア系列	$\left(1-\frac{cl}{k}\right)\boldsymbol{e}+\frac{cl}{k}\boldsymbol{w}$	$\frac{\alpha}{\beta}cl\boldsymbol{E_3}$

図 3　ガウス分布による指示語領域の表現の例．H の白丸はユーザ，R の黒丸はロボットであり，矢印はそれぞれの向きを表す．赤い領域がコ系列，青い領域がソ系列，緑色の領域がア系列の指示語領域を表す．

2.1　指示語領域に基づく確率推定器

指示語領域に基づく確率推定器は指示語情報，ロボットが事前に環境内を探索して得た 2 次元占有格子地図と候補物体の 3 次元位置，指差し情報から得たユーザの体の部位の 3 次元位置の 4 つの情報を入力として対象確率を出力する．対象確率を出力するために，図 3 のように指示語の系列毎にユーザが指示した物体が存在する可能性の高い領域を 2 次元占有格子地図上で予測する．ここで，物体が存在する可能性の高い領域を指示語領域と呼ぶ．また，候補物体の 3 次元位置とユーザの体の部位の 3 次元位置は地図の座標系における位置情報である．

各指示語系列にはそれぞれ異なる性質があり，コ系列は話し手から近い物体や人を参照する場合，ソ系列は聞き手から近い物体や人を参照する場合，ア系列は話し手からも聞き手からも遠い物体や人を参照する場合にそれぞれ用いられる．本研究では各系列の性質を用いて 3 次元ガウス分布によって指示語領域を生成する．指示語領域を生成する際にユーザやロボットの位置，指差しの方向なども考慮する必要がある．指差し情報からユーザの体の座標を予測する方法については 2.3 節で述べる．指示語領域に基づく確率推定器は指示語領域を生成する 3 次元ガウス分布に各候補物体の 3 次元位置情報を入力し，得られた確率を対象確率として出力する．指示語領域の予測に必要な変数を表 1 に示す．各系列の指示語領域を形成する 3 次元ガウス分布の平均と分散共分散行列の定義を図 3 の下の表に示す．各パラメータの設計指針は以下である．（詳細は付録 A 参照）

- コ系列の平均：ユーザに近い手首の座標
- ソ系列の平均：ロボットに近いユーザとロボットの内分点の座標
- ア系列の平均：ユーザから遠い指差しベクトルの延長線上の座標
- 各系列の分散：ユーザまたはロボットの位置と指示語領域の平均の座標との距離に比例する値

また，ユーザの手首や目の位置情報を得ることができなかった場合は，両肩の中央の座標などの代わりとなる位置情報を推定器に与える．

2.2 物体カテゴリに基づく確率推定器

物体カテゴリに基づく確率推定器は，物体カテゴリ情報とロボットが事前に探索して得た候補物体のカテゴリ確率を入力とし，対象確率を出力する．対象確率を予測するために Objects365 [10] で事前学習済みの You Only Look Once version 5 (YOLOv5) [11] を物体検出器として用いる．YOLOv5 を用いて事前に候補物体全てを検出し，各候補物体が属する物体カテゴリの信頼度スコアを計算する．各候補物体の物体カテゴリ情報に対応する信頼度スコアを正規化したものを対象確率とする．

2.3 指差し方向に基づく確率推定器

指差し方向に基づく確率推定器は，指差し情報から骨格検出をすることで得たユーザの体の部位の 3 次元位置とロボットが事前に環境内を探索して得た候補物体の 3 次元位置を入力とし，対象確率を出力する．骨格検出には MediaPipe [12] を用いる．対象確率を予測するために，ユーザの目を始点とし，手首を終点とする指差しベクトルとユーザの目から各候補物体へのベクトルの 2 種類のベクトルを使用し，内積の定義式から 2 つのベクトルのなす角 θ を得る．ここで得た θ を用いて，確率変数に角度を用いる 2 次元フォンミーゼス分布から確率を得る．指差し方向に基づく確率推定器はフォンミーゼス分布によって得た確率を対象確率として出力する．

3 実験

実験では物体カテゴリ情報，指示語情報，指差し情報の 3 種類の外部照応情報を用いて，ユーザが指示した物体をどの程度の精度で特定することができるか検証した．また外部照応情報の内，どの情報がどの程度外部照応に寄与するかを検証した．

3.1 実験条件

実験は家庭環境を模した実環境で行った．本実験の環境の家具配置，物体配置，指差しを行う際の人とロボットの位置を予め設定し，実験を行った．実験環境の詳細な図は付録 B に記述した．また，ロボットには Human Support Robot (HSR) [13] を用いた．ロボットはユーザからの言語による指示を正確に認識でき，指示文から物体カテゴリ情報と指示語情報を正確に抽出することができると仮定する．

実験では 4 カテゴリの物体を使用し，“Bottle”, “Book”, “Stuffed Toy”, “Cup” を 1 カテゴリ 5 物体，計 20 物体をユーザが指示する物体とした．これらのカテゴリは Objects365 [10] に基づいて，筆者が一般的な家庭環境にあると考えられる物体を選択した．

3.2 検証方法

実験では 2 種類の検証を行い，提案手法の有効性を確認した．1 つ目は，ベースライン手法と提案手法を比較した．実験データには，外部照応情報が全て含まれた場合，3 種類の情報の内いずれかの情報が欠けた場合，指差し情報を誤って得た場合の 5 種類を用いた．外部照応情報が全て含まれた場合は図 4 において各地点で 4 データ，計 40 データ用意し，物体カテゴリ情報が欠けた場合と指示語情報が欠けた場合は 40 データからそれぞれの情報を削除したものを用いた．また，指差し情報が欠けた場合と指差し情報を誤って得た場合は新たに 10 データ

表 2 ベースライン手法と提案手法における指示物体の特定精度の比較．() 内は分数表示．太字は条件での最大値．

	指示における観測条件				
手法	物体カテゴリ欠損	指示語欠損	指差し欠損	指差し誤認識	欠損無し
VGPN [7]	0.28 (11/40)	**1.00 (40/40)**	0.30 (3/10)	**0.40 (4/10)**	**1.00 (40/40)**
YouRefIt [9]	0.18 (7/40)	0.20 (8/40)	0.10 (1/10)	0.00 (0/10)	0.20 (8/40)
提案手法	**0.53 (21/40)**	0.85 (34/40)	**0.90 (9/10)**	**0.40 (4/10)**	**1.00 (40/40)**

表 3 提案手法におけるアブレーションスタディの実験結果．チェックマークは確率推定器有りを示す．

確率推定器の有無			
物体カテゴリ	指示語領域	指差し方向	精度
	✓	✓	0.53
✓		✓	0.85
✓	✓		1.00
✓	✓	✓	1.00

ずつ用意した．指差し情報を誤って得た場合はユーザの体をユーザ自身が隠してしまう自己オクルージョンが発生した際などで得られる．2 つ目は，提案手法の中でどの情報が外部照応にどの程度寄与するかを検証するために，アブレーションスタディを行った．3 種類の推定器の内，1 種類の推定器を削除した場合の 3 パターン，全ての推定器を用いたパターンの計 4 パターンを比較した．

実験ではベースライン手法を 2 種類設定した．1 つ目は Hu らの研究 [7] を参考に，ユーザの指差しと物体カテゴリ情報を用いた手法を設定した．基本的には，Hu らの手法における目標領域を推測する手順と同じである．異なる点は，言語情報内に物体カテゴリ情報がない時は対象とする全ての物体で候補集合を作る点と，ユーザの指差しを観測できなかった場合は候補集合の中から目標物体をランダムに選択する点である．2 つ目は Chen らの研究 [9] のモデルを用いた．このモデルにユーザが指差しを行っている画像，指示語情報と物体カテゴリ情報を含んだ文を入力し，得られた矩形領域から目標物体の予測が成功しているかを判断した．

評価指標は，ユーザが指示した物体をロボットが正しく特定できたかの精度である．精度は $\frac{1}{N}\sum_{i=1}^{N} S_i$ で計算される．N は試行回数，S_i は i 番目の試行が成功すれば 1，失敗すれば 0 が入力される．

4 実験結果

表 2 にベースライン手法と提案手法の比較を示す．まず Voice-Guided Pointing Robot Navigation for Humans (VGPN) [7] に基づくベースライン手法と提案手法を比較した．外部照応情報が全て含まれた場合は同じ場所に同じカテゴリの物体が 2 つ以上ないため，両手法とも精度が 1.00 となったが，物体カテゴリ情報が欠損した場合と指差し情報が欠損した場合に精度に約 2〜3 倍の差が見られた．また，指示語を考慮する提案手法において指示語情報が欠損した場合でも 0.85 という結果を示した．これらの結果より，提案手法は 3 種類の外部照応情報の内 1 つが欠損した場合でも互いに補い，目標物体を予測できたと推察される．次に，YouRefIt [9] で用いられたフレームワークと提案手法を比較した．YouRefIt で用いられたフレームワークは画像内に目標物体がなければ正しく予測ができないため全体的に精度が低くなり，提案手法が上回る結果となった．

次に表 3 にアブレーションスタディの結果を示す．物体カテゴリに基づく確率推定器を除いた場合，同じ場所に物体が複数配置される条件下では目標物体の特定は困難と考えられるが，0.53 という精度を示した．この結果から，残り 2 つの確率推定器により目標物体の候補を絞り込めたと考えられる．また，指差し方向に基づく確率推定器を除いた時は精度が 1.00 となった．これは指示語領域に基づく確率推定器が指差し方向を考慮し，目標物体が配置された領域を推定できたと考えられる．指差し方向に基づく確率推定器は同じ場所に同カテゴリの物体が 2 つ以上ある場合などに有効であると考えられる．

5 まとめ

本稿ではマルチモーダル情報を外的な文脈として用いて外部照応解析を行うためのモデルを提案した．物体カテゴリ情報，指示語情報，指差し情報を用い，外部照応を行うことで外部照応にこれらの情報が有効であることを示した．

今後の展望として，指示語領域を形成する 3 次元ガウス分布のデータからのパラメータ推定や，自然な文からの指示語や物体カテゴリに対応する語の認識に取り組む予定である．

謝辞

本研究は【Moonshot R&D – MILLENNIA Program】課題番号 JPMJMS2011，JSPS 科研費 JP22K12212 の助成を受けたものです．

参考文献

[1] Tadahiro Taniguchi, Daichi Mochihashi, Takayuki Nagai, Satoru Uchida, Naoya Inoue, Ichiro Kobayashi, Tomoaki Nakamura, Yoshinobu Hagiwara, Naoto Iwahashi, and Tetsunari Inamura. Survey on frontiers of language and robotics. **Advanced Robotics**, Vol. 33, No. 15-16, pp. 700–730, 2019.

[2] 村田真樹, 黒橋禎夫, 長尾真. 表層表現を手がかりとした日本語名詞句の指示性と数の推定. 自然言語処理, Vol. 3, No. 4, pp. 31–48, 1996.

[3] 関和広, 藤井敦, 石川徹也. 確率モデルを用いた日本語ゼロ代名詞の照応解析. 自然言語処理, Vol. 9, No. 3, pp. 63–85, 2002.

[4] 飯田龍, 乾健太郎, 松本裕治. 文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定. 情報処理学会論文誌, Vol. 45, No. 3, pp. 906–918, 2004.

[5] 杉山治, 神田崇行, 今井倫太, 石黒浩, 萩田紀博, 安西祐一郎. コミュニケーションロボットのための指さしと指示語を用いた 3 段階注意誘導モデル. 日本ロボット学会誌, Vol. 24, No. 8, pp. 964–975, 2006.

[6] Dadhichi Shukla, Ozgur Erkent, and Justus Piater. Probabilistic Detection of Pointing Directions for Human-Robot Interaction. In **International Conference on Digital Image Computing: Techniques and Applications (DICTA)**, pp. 1–8, 2015.

[7] Jun Hu, Zhongyu Jiang, Xionghao Ding, Taijiang Mu, and Peter Hall. VGPN: Voice-Guided Pointing Robot Navigation for Humans. In **IEEE International Conference on Robotics and Biomimetics (ROBIO)**, pp. 1107–1112, 2018.

[8] Qi Wu, Cheng-Ju Wu, Yixin Zhu, and Jungseock Joo. Communicative Learning with Natural Gestures for Embodied Navigation Agents with Human-in-the-Scene. In **IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)**, pp. 4095–4102, 2021.

[9] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. YouRefIt: Embodied Reference Understanding with Language and Gesture. In **IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 1385–1395, 2021.

[10] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A Large-scale, High-quality Dataset for Object Detection. In **IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 8430–8439, 2019.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 779–788, 2016.

[12] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A Framework for Building Perception Pipelines. **arXiv preprint arXiv:1906.08172**, 2019.

[13] Takashi Yamamoto, Koji Terada, Akiyoshi Ochiai, Fuminori Saito, Yoshiaki Asahara, and Kazuto Murase. Development of Human Support Robot as the research platform of a domestic mobile manipulator. **ROBOMECH journal**, Vol. 6, No. 1, pp. 1–15, 2019.

A　各指示語系列の領域を形成するガウス分布の平均と分散の設定

A.1　コ系列指示語領域のガウス分布の平均

コ系列の指示語は話し手から近い物体を参照する際に用いられるので，ユーザの手首の座標をガウス分布の平均とした．ガウス分布の平均を人の体の中心や足元の座標にせずに手首の座標を平均としたのは，ユーザの指示する方向を考慮するためである．

A.2　ソ系列指示語領域のガウス分布の平均

ソ系列の指示語は聞き手から近い物体を参照する際に用いられるので，ロボットの足元とユーザの目をつなぐ線分を $\alpha : (1-\alpha)$ に内分する点の座標をガウス分布の平均とした．ロボットはユーザの体の部位の 3 次元座標を検出するために，ユーザの方向を向いているので，この内分点はロボットの向いている方向にあり，平均がロボットの背後の位置になることはない．しかし，他人に物を取ってもらうことを考えたとき，頼んだものがその人の背後にあることよりもその人が見えていることの方が多いと考えられるので上記のようにガウス分布の平均を設定した．

A.3　ア系列指示語領域のガウス分布の平均

ア系列の指示語は話し手からも聞き手からも遠い物体を参照する際に用いられるので，ユーザの指をさした方向にベクトルを cl だけ伸ばした先の点をガウス分布の平均とした．他人に物を取ってもらう場面を考えたとき，頼まれた人から遠すぎる物体はその人に頼まず，もっと近くにいる人に頼むなどの別の方法をとることが考えられる．この「遠すぎる」の程度を定義しているのが c である．また，ア系列の平均は l に比例し，これは話し手と聞き手の距離が遠いほど，聞き手からより遠い物体を指示する可能性が高くなるという仮定の下このように定義した．

A.4　各系列指示語領域のガウス分布の分散共分散行列

各系列の分散共分散行列は $\frac{\alpha}{\beta} l \boldsymbol{E_3}$ が共通箇所となっており，ア系列の分散共分散行列にのみ係数 c が掛けられている．β は分散が小さくなりすぎないように調整するためのパラメータ，$\boldsymbol{E_3}$ は 3×3 の単位行列である．

まず，ソ系列の分散を図 3 にある表で定義した理由を述べる．分散は目標の物体がユーザまたはロボットから離れているほど大きくなると考えられるので，目標物体がある確率に影響を与える α に比例するように定義した．ソ系列の指示語が観測された際の目標物体はロボットの近くにある確率が高いということはユーザと目標物体はおよそ l の距離だけ離れていると考えられるので，分散は l にも比例するように定義した．

コ系列の場合は目標物体はユーザの近くにある確率が高いのでロボットと目標物体はおよそ l 離れていると考えられる．コ系列の指示語が観測された際もソ系列の指示語が観測された際もユーザまたはロボットと目標物体まではおよそ l だけ離れているので，コ系列のガウス分布の分散もソ系列の分散と同じものを用いることとした．

また，ア系列の場合はユーザからもロボットからも目標物体はおよそ cl だけ離れていると考えられるので，cl に比例するように定義した．

B　実験環境

本研究で使用した実験環境を図 4 に示す．

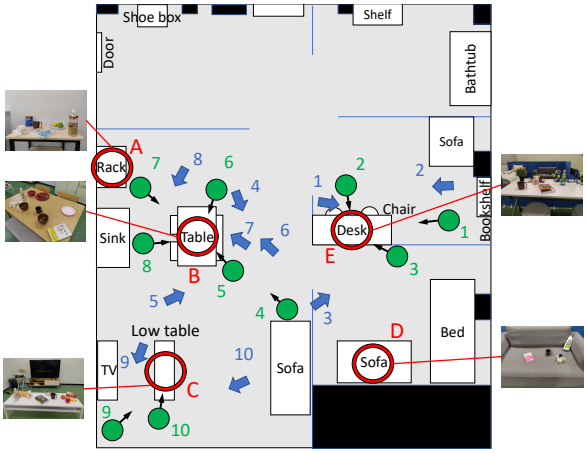


図 4　家庭環境を模した実験環境の家具配置，物体配置位置，指差しを行う人とロボットの立ち位置．緑丸がユーザの位置，青矢印がロボットの位置と方向，赤丸は物体が配置されている領域を表す．緑色と青色の同じ数字はロボットに外部照応を行わせたときのシナリオを表している．例えば，緑色の 1 番と青色の 1 番で E の領域内の物体を指差ししたシナリオがある．