4D2-OS-37d-4

物体の共起性に基づいたマルチモーダル場所領域学習による場所の理解

Understanding of Places by Multimodal Spatial Concepts Learning Based on Coccurrence of Objects

磯部 匠汰 *1 Shota Isobe　谷口 彰 *1 Akira Taniguchi　萩原 良信 *1 Yoshinobu Hagiwara　谷口 忠大 *1 Tadahiro Taniguchi

*1立命館大学
Ritsumeikan University

In order to support people, robots are required to perform tasks such as cleaning up or picking up objects. To achieve these tasks, we consider that robots need to learn the relationships between objects and places. In this paper, we propose a model that can learn the existence probability of objects in each place by multi-modal spatial concept learning based on co-occurrence of objects. In the experiments, we evaluate quantitatively to estimation results of objects from a word expressing the place. Furthermore, the robot actually performs tasks of cleaning up objects for proving the usefulness of the proposed model. We showed that the robot properly learned the relationships between objects and places. In addition, we showed that the robot can perform the task of cleaning up objects using the learning result.

1. はじめに

人のサポートを目的としたロボットは，人とのコミュニケーションを通してタスクを理解し遂行できることが求められている．ロボットが人のサポートをするタスクとして，物を片付けるタスクや物を取りに行くタスクが挙げられる．例えば,「お皿を取って来て」や「これを片付けて」といったタスクが考えられる．前者のタスクでは,「お皿」という単語が示す物体がどの場所にあるかを推定できる必要がある．さらに，お皿が複数の場所に存在している場合，どの場所にあるお皿を取りに行くべきかをロボットが知るために，場所を表す語彙を用いて人とコミュニケーションを取ることでタスクを効率的に行うことができると考える．また，後者のタスクでは，提示された物体から片付ける場所やその場所を表す語彙を推定する必要がある．

場所の学習に関する先行研究として，谷口らは，場所の名前および場所の空間的領域を場所概念と定義し，場所概念に関連する潜在変数を介して，自己位置推定のための生成モデルと発話文の教師なし単語分割を統合したノンパラメトリック場所概念獲得手法 (SpCoA) を提案している [1]．萩原らは，ロボットが自身の観測情報から場所の階層的なカテゴリ分類を行う手法を提案している [2]．石伏らは，Convolutional Neural Network (CNN) の画像認識結果と位置推定手法を統計的に統合したモデルを提案している [3]．しかし，谷口らの研究では，場所概念形成に自己位置情報と単語情報のみを用いているため,「お皿」などの物体と「キッチン」などの場所の関係性を学習していないことから,「お皿を取ってきて」や「これを片付けて」といったタスクを行うことができない．また，萩原らの研究と石伏らの研究も同じく，CNN の物体認識結果や中間層の特徴量を視覚的特徴として場所領域学習に用いており，物体の個数や物体の名称を用いていないことから，物体と場所の関係性を学習できていないため，上記で述べたタスクを行うことができない．

本研究では，場所を表す語彙とそれに対応する空間的な領域を場所領域，画像中から物体の位置や範囲，個数を検出することができる物体検出手法によって得られる物体を Bag-of-Objects (BoO) 表現した情報を物体情報，場所を表す語彙を語彙情報とし，物体情報と語彙情報，自己位置情報から物体と場所の関係性と場所領域を学習するモデルを提案する．実験では，学習させたモデルを用いて，場所を表す語彙情報から物体の推定や指定した物体から場所を表す語彙の推定を行い，定量的に評価することで物体と場所の関係性の妥当性を検証する．また，提案モデルを用いて，実際にロボットが物を片付けに行くタスクを行うことで，本提案モデルの有用性を示す．

連絡先: 磯部匠汰，立命館大学情報理工学研究科，滋賀県草津市野路東 1-1-1，isobe.shota@em.ci.ritsumei.ac.jp

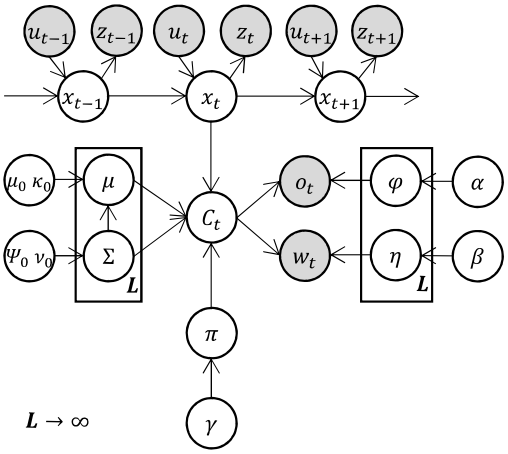

図 1: 提案手法のグラフィカルモデル

2. 提案手法

本研究では，物体検出手法で検出された物体を BoO 表現した情報を物体情報とし，物体情報や自己位置情報，語彙情報を場所領域の学習に用いることで，各場所領域の物体の存在確率を学習する．ここでは，各場所領域の物体の存在確率を物体と場所の関係性とする．本研究のグラフィカルモデルを図 1，グラフィカルモデルの要素を表 1，生成モデルを式 (1)-(10) に示す．GEM (·) は Stick Breaking Process (SBP)[4]，IW (·) は逆ウィシャート分布，N (·) はガウス分布，Mult (·) は多項分布，Dir (·) はディリクレ分布である．事前に Simultaneous

表 1: 提案手法のグラフィカルモデルの要素

x_t	ロボットの自己位置情報
z_t	距離センサの観測情報
u_t	モータの制御情報
o_t	観測画像から得られる物体情報
w_t	語彙情報
C_t	場所領域の index
μ, Σ	ガウス分布のパラメータ
φ, η, γ	多項分布のパラメータ
π	場所領域の index の多項分布
$\mu_0, \kappa_0, \psi_0, \nu_0$	ガウス-ウィシャート事前分布の ハイパーパラメータ
α, β	ディリクレ事前分布の ハイパーパラメータ

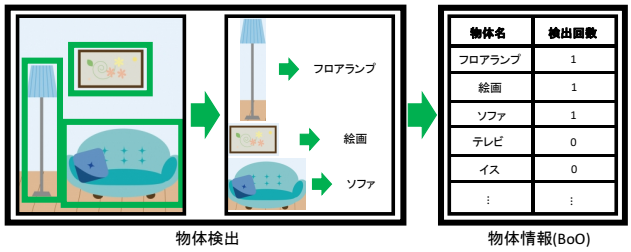

図 2: 物体情報の取得方法

Localization and Mapping (SLAM)[5] により作成した地図を用いて，自己位置推定手法である Monte Carlo Localization (MCL) により自己位置を推定する．また，画像から物体を検出するために，物体検出手法の一つである Faster Region-based CNN (Faster R-CNN)[6] を用いる．物体情報の取得方法を図 2 に示す．時刻 t の時に取得した画像から Faster R-CNN を用いて，画像内にある物体を検出し BoO 表現した物体情報 $\boldsymbol{o_t}$ を取得する．自己位置情報 $\boldsymbol{x_t}$ は，$\boldsymbol{x_t} = (x_t, y_t, \sin\theta_t, \cos\theta_t)$ と定義する．ここで x_t，y_t は，(x, y) の 2 次元座標でのロボットの自己位置の値である．θ_t はロボットの向きであり，x 軸の正の方向を 0°，y 軸の正の方向を 90° として定義する．u_t，z_t は，それぞれロボットの制御情報と距離センサからの観測情報を表す．物体情報 $\boldsymbol{o_t}$ は，$\boldsymbol{o_t} = \left(o_t^1, o_t^2, \cdots, o_t^I\right)$ と定義する．I は，物体検出手法により検出できる物体のカテゴリ数である．また，自己位置情報 x_t に該当する場所の名前を語彙情報 $\boldsymbol{w_t}$ として与える．語彙情報 $\boldsymbol{w_t}$ は，$\boldsymbol{w_t} = \left(w_t^1, w_t^2, \cdots, w_t^M\right)$ と定義する．M は，人が与える場所を表す語彙の総数である．この物体情報と語彙情報，自己位置情報を用いて場所領域を学習させる．場所領域の数は，SBP により確率的に求める．場所領域の学習には，ギブスサンプリングを用いる．ギブスサンプリングによる各パラメータのサンプリングの手順を式 (11)〜(14) に示す．NIW $(\cdot)$ はガウス-逆ウィシャート分布であり，$\psi_{n_l}, \nu_{n_l}, \mu_{n_l}, \kappa_{n_l}$ は更新後のハイパーパラメータである．x_l, o_l, w_l は，それぞれ $C_t = l$ の時の自己位置情報，物体情報，語彙情報のデータの集合である．ギブスサンプリングにより，パラメータ C_t，μ，Σ，φ，η を推定する．

$$\pi \sim \mathrm{GEM}(\gamma) \tag{1}$$

$$C_t \sim p(C_t|x_t, \mu, \Sigma, \pi) \propto \frac{\mathrm{N}(x_t|\mu_{C_t}, \Sigma_{C_t})\,\mathrm{Mult}(C_t|\pi)}{\sum_{c'}\mathrm{N}(x_t|\mu_{c'}, \Sigma_{c'})\,\mathrm{Mult}(c'|\pi)} \tag{2}$$

$$\Sigma \sim \mathrm{IW}(\Sigma|\psi_0, \nu_0) \tag{3}$$

$$\mu \sim \mathrm{N}(\mu|\mu_0, (\Sigma/\kappa_0)) \tag{4}$$

$$\varphi \sim \mathrm{Dir}(\alpha) \tag{5}$$

$$\eta \sim \mathrm{Dir}(\beta) \tag{6}$$

$$o_t \sim \mathrm{Mult}(o_t|\varphi_{C_t}) \tag{7}$$

$$w_t \sim \mathrm{Mult}(w_t|\eta_{C_t}) \tag{8}$$

$$x_t \sim p(x_t|x_{t-1}, u_t) \tag{9}$$

$$z_t \sim p(z_t|x_t) \tag{10}$$

$$C_t \sim p(C_t = l|x_t, \mu, \Sigma, \pi, \varphi, \eta) \propto \mathrm{N}(x_t|\mu_{C_t}, \Sigma_{C_t})\,\mathrm{Mult}(o_t|\varphi_{C_t}) \times \mathrm{Mult}(w_t|\eta_{C_t})\,\mathrm{Mult}(C_t|\pi) \tag{11}$$

$$\mu_l, \Sigma_l \sim \mathrm{N}(x_l|\mu_{C_t}, \Sigma_{C_t})\,\mathrm{NIW}(\mu_l, \Sigma_l|\psi_0, \nu_0, \mu_0, \kappa_0) \propto \mathrm{NIW}(\mu_l, \Sigma_l|\psi_{n_l}, \nu_{n_l}, \mu_{n_l}, \kappa_{n_l}) \tag{12}$$

$$\varphi_l \sim \mathrm{Multi}(o_l|\varphi_l)\,\mathrm{Dir}(\varphi_l|\alpha) \tag{13}$$

$$\eta_l \sim \mathrm{Multi}(w_l|\eta_l)\,\mathrm{Dir}(\eta_l|\beta) \tag{14}$$

3. 実験

3.1 実験目的

実験では，学習させたモデルを用いて，場所を表す語彙情報から物体の推定や指定した物体から場所を表す語彙の推定を行い，定量的に評価することで物体と場所の関係性の妥当性を検証する．また，提案モデルを用いて，実際にロボットが物を片付けに行くタスクを行うことで，本提案モデルの有用性を示す．

3.2 実験条件

本実験では，トヨタの家庭用開発ロボット HSR (Human Support Robot)*1 を用いて実験を行う．立命館大学テクノコンプレックス 2 階の実験室 10 の実験環境とする．実験環境の配置図を図 3 に示す．地図はレーザーレンジセンサを用いて，事前に SLAM により生成しておき，ロボットは地図を持っていることを前提とする．自己位置推定は，Robot Operating System (ROS)*2 の amcl (adaptive MCL) パッケージを用いて行った．語彙情報は,「台所前，冷蔵庫前，ゴミ箱前，キャビネット前，レンジラック前，ダイニング前，本棚前，テレビ前，テーブル前，入り口前，ソファー前」とし，自己位置情報のデータ数の 1 割に語彙情報を割り当てることとする．また，ギブスサンプリングのイテレーション回数を 100 回とし，715 個の位置と画像の訓練データと 72 個の語彙データからパラメータを学習した．Faster R-CNN は，データセット MS-COCO*3 でプレ学習済みの VGG-16 モデルを使用したため，物体情報は 80 次元である．ディリクレ分布のハイパーパラメータはそれぞ

*1 HSR: http://newsroom.toyota.co.jp/en/detail/8709536
*2 ROS: http://www.ros.org/
*3 MS-COCO: http://mscoco.org/

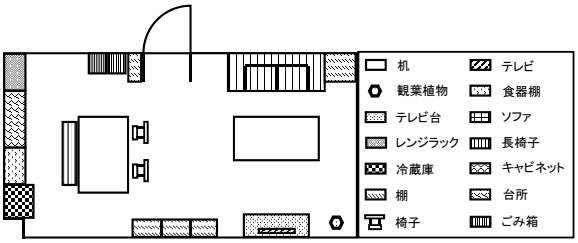

図 3: 実験環境の配置図

れ $\alpha = 0.01$，$\beta_0 = 0.01$，$\gamma_0 = 4$，$\mu_0 = (-1.0, -0.5, 0.0, 0.0)$，$\kappa_0 = 0.03$，$\nu_0 = 0.1$，$\psi_0 = \mathrm{diag}(0.1, 0.1, 0.5, 0.5)$ とする．学習された物体と場所の関係性の妥当性を検証するために，物体検出できる物体 80 個の中から選んだ物体 4 個を正解ラベルとし作成するものとする．正解ラベルは，実験環境を知る人に作成してもらう．

3.3 実験方法

ロボットは MCL により自己位置推定をしながら，コントローラの操作により環境中を移動し，各自己位置での自己位置情報と画像を取得する．自己位置情報のデータ数の 1 割に語彙情報を割り当てる．語彙情報を与える 1 割のデータは，ランダムに決定する．また，取得した画像は Faster R-CNN を用いて，画像内にある物体を検出し BoO 表現した物体情報を取得する．語彙情報や自己位置情報，各自己位置の物体情報を用いて場所領域を学習させる．学習させて得られる場所領域は，地図上に場所領域のガウス分布を描画し確認する．また，学習させたモデルを用いて，場所を表す語彙情報から物体の推定や指定した物体から場所を表す語彙の推定を行い，定量的に評価することで物体と場所の関係性の妥当性を検証する．式 (15)，式 (16) より，場所の名前からその場所にある物体を推定する．式 (15)，式 (16) では，指定した物体を O，得られる場所を表す語彙を W とする．さらに，推定結果と正解ラベルを比較し，物体と場所の関係性の妥当性を検証する．そして，式 (17)，式 (18) より，特定の物体を提示した時，その物体の存在する確率の高い場所の名前を推定する．式 (17)，式 (18) では，指定した場所を表す語彙を W，得られる物体を O とする．最後に，提案モデルを用いて，実際にロボットが物を片付けに行くタスクを行うことで，本提案モデルの有用性を示す．

$$C^* = \arg\max_{C_t} p\left(\varphi_{C_t} | o_t = O\right) \tag{15}$$

$$W = \arg\max_{w_t} p\left(w_t | \eta_{C^*}\right) \tag{16}$$

$$C^* = \arg\max_{C_t} p\left(\eta_{C_t} | w_t = W\right) \tag{17}$$

$$O \sim p\left(o_t | \varphi_{C^*}\right) \tag{18}$$

3.4 実験結果

形成された場所領域と分類された画像の例を図 4 に示す．図 4 では，学習により推定された 11 つの場所領域を示しており，色で識別できるようにしている．矢印が場所領域の中心となる位置と向き，半透明の円が場所領域の共分散行列を表している．各画像は場所領域に割り当てられた画像の例である．以降，各場所領域は，場所領域の index で示す．図 4 より，index3 と index5 は向いている方向によって見える物体が異なるため，

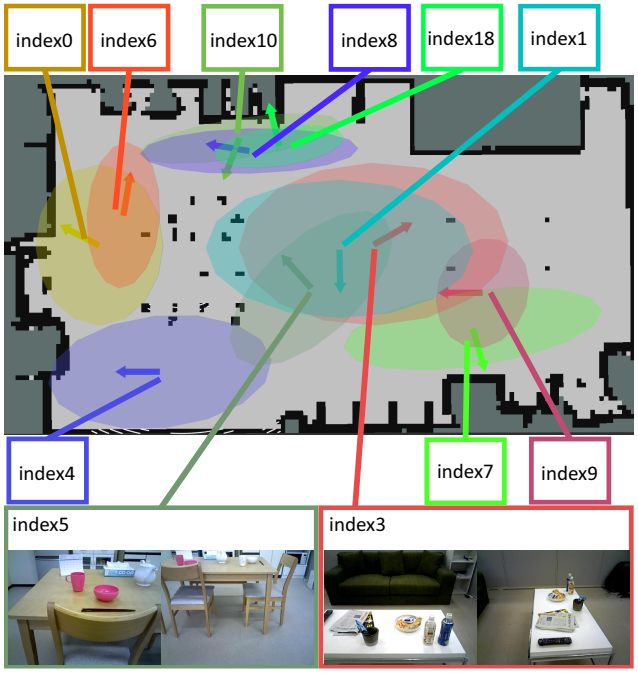

図 4: 形成された場所領域と分類された画像の例

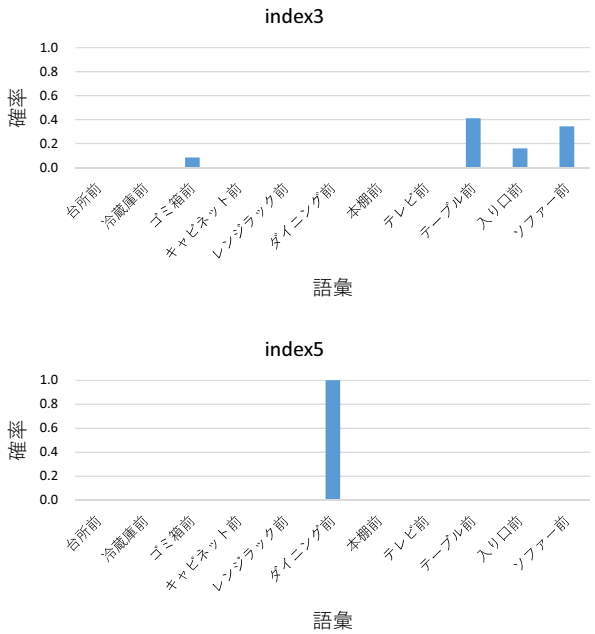

図 5: 各場所領域における語彙の生起確率

別々の場所領域となっていることが分かる．各場所領域における語彙の生起確率を図 5 に示す．図 5 より，index5 はダイニング前である確率が高いことが分かる．また，場所を表す語彙から物体を推定した結果を表 2 に示す．表 2 では，場所を表す語彙をダイニング前，テーブル前，テレビ前とした時の推定結果を示しており，式 (17)，式 (18) により物体を 4 つ推定している．正解ラベルと比較した結果，ダイニング前の場合正解率 1.0，テーブル前の場合正解率 0.8，テレビ前の場合正解率 0.8 となった．以上から，学習された物体と場所の関係性は妥当である．そして，指定した物体から場所領域とその場所

表 2: 場所を表す語彙から物体を推定した結果

語彙情報	推定結果 (物体)	正解ラベル	正解率
ダイニング前	chair cup dining table bowl	cup chair dining table bowl	1.0
テーブル前	book remote bowl bottle	remote bottle cup bowl	0.8
テレビ前	laptop tv keyboard potted plant	keyboard tv mouse potted plant	0.8

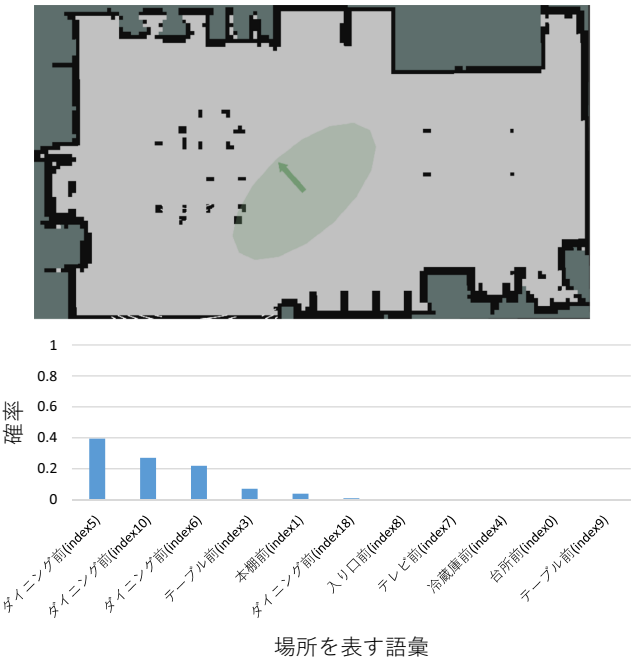

図 6: cup から場所を表す語彙と場所領域を推定した結果

を表す語彙を推定した結果を図 6 と図 7 に示す．図 6 と図 7 は，物体をそれぞれ cup，remote と指定した時の推定結果を示している．図 6 と図 7 より，cup はダイニング前 (index5)，remote はテーブル前 (index3) にある確率が高いことが分かる．また，提案モデルを用いて，実際にロボットが物を片付けに行くタスクを行った結果，タスクを遂行することできたことを確認した．

4. おわりに

本研究では，物体情報や語彙情報，自己位置情報を場所領域の学習に用いることで，場所領域や物体と場所の関係性を学習するモデルを提案した．実験結果から場所を表す語彙から物体の推定や指定した物体から場所を表す語彙の推定ができることを示した．また，場所を表す語彙から物体を推定した結果と正解ラベルとの定量的な評価より，本提案モデルの学習で得られた物体と場所の関係性の妥当性を示した．さらに，ロボットが物を片付けに行くタスクを遂行できたことから，本提案

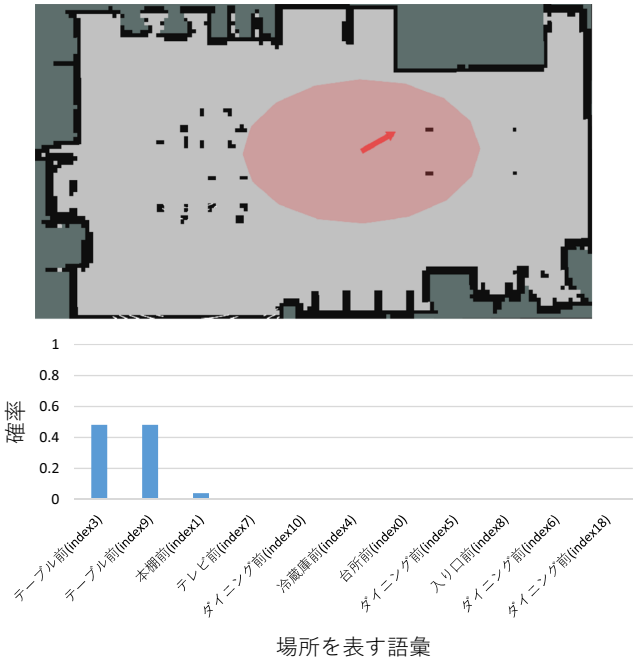

図 7: remote から場所を表す語彙と場所領域を推定した結果

モデルの有用性を示した．今後の課題として，ロボットが新しい環境に行った時，学習した物体と場所の関係性を用いて場所領域や場所を表す語彙を推定できるようにすることが挙げられる．また，今回は MS-COCO データセットのプレ学習済みモデルを使用していたため検出できる物体が 80 個しかないことから，検出できる物体数を増やすために Faster R-CNN を fine-tuning する必要があることが挙げられる．

参考文献

[1] A. Taniguchi, T. Taniguchi, and T. Inamura, “Spatial Concept Acquisition for a Mobile Robot that Integrates Self-Localization and Unsupervised Word Discovery from Spoken Sentences,” IEEE Transactions on Cognitive and Developmental Systems, vol. 8, no. 4, pp. 285-297, 2016.

[2] Y. Hagiwara, M. Inoue, and T. Taniguchi, “Place Concept Learning by hMLDA Based on Position and Vision Information,” in 13th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems, 2016, pp. WebC301.

[3] 石伏智, 谷口彰, 萩原良信, 高野敏明, 谷口忠大. 位置情報と画像の高次特徴量に基づく教師なし場所領域学習. 人工知能学会全国大会論文集, Vol. 30, pp. 1-4, 2016

[4] J. Sethuraman, “A constructive definition of Dirichlet priors,” Statistica Sinica, vol. 4, pp. 639-650, 1994.

[5] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. 確率ロボティクス (mynavi advanced library), 2015

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards Real-Time Object Detection with Region Proposal Networks. In Advances in neural information processing systems, pp. 91-99, 2015