3D1-OS-37a-2

ノンパラメトリックベイズモデルとSLAMの統合による地図と場所概念の逐次学習

Online Learning of Spatial Concepts and an Environmental Map by Integrating Nonparametric Bayesian Model and SLAM

谷口 彰[*1] Akira Taniguchi　萩原 良信[*1] Yoshinobu Hagiwara　谷口 忠大[*1] Tadahiro Taniguchi　稲邑 哲也[*2*3] Tetsunari Inamura

[*1]立命館大学 Ritsumeikan University　[*2]国立情報学研究所 National Institute of Informatics　[*3]総合研究大学院大学 The Graduate University for Advanced Studies

Robots operating in human-living environments are required to adaptively learn and use the spatial concepts and lexicon related to places for human-robot speech interaction. We propose a nonparametric Bayesian generative model SpCoSLAM integrating SpCoA and FastSLAM. In additon, we propose an online learning algorithm of SpCoSLAM. In the experiments, we tested online learning of spatial concepts and environmental maps in a novel environment of which the robot did not have a map.

1. はじめに

人間と共存し様々な環境で動作するロボットは，適応的に空間の概念や語彙を学習，活用することが求められる．また場所や空間に関する概念は，物体に関する概念と比べて指示対象の範囲が明確ではなく，物体の配置や人により異なる．そのため，事前に概念を人手で設計することは難しく，ロボットが自らの経験を元に自律的に概念を形成する必要がある．本研究は，未知な環境においても環境および人とのインタラクションから場所に関する概念を逐次的に獲得することのできる手法の開発を目的とする．

本研究の概要図を図 1 に示す．本研究では，移動ロボットが事前知識のない状態から，空間の形状や場所の領域，場所の呼び名を逐次的に学習する手法の構築を目指す．これまで我々は，移動ロボットの自己位置推定と語彙獲得を統合した自己位置と語彙の同時推定モデル [谷口 14] やノンパラメトリックベイズ場所概念獲得モデル (SpCoA) [Taniguchi 16] を提案した．また，位置情報と画像情報に基づく場所の領域推定モデル [Ishibushi 15] の提案も行っている．これらの手法で用いられているアルゴリズムはバッチ学習であり，時間的な環境の変化や場所の呼び名の変化に逐次的に対応することはできない．また，SLAM により生成した地図を既知とした上で場所概念の獲得を行うため，空間形状の変化や事前に地図生成を行っていない未知な環境に対しては場所概念の獲得ができないという問題があった．そこで本研究では，ノンパラメトリックベイズ生成モデルの枠組みで地図生成および位置・音声言語・画像を統合した場所概念を逐次的に学習できるオンラインアルゴリズムの構築を行う．

Simultaneous Localization And Mapping (SLAM) は，自己位置推定と地図生成を同時に行う問題である．特に FastSLAM [Grisetti 05] では，Rao-Blackwellized Particle Filters (RBPFs) [Doucet 00] によるオンラインアルゴリズムによって効率的な自己位置推定と地図生成を実現している．本研究では，FastSLAM アルゴリズムを場所概念獲得の確率的生成モデルに対し適用することを行う．SpCoA は Particle Filter に基づいた自己位置推定手法である Monte Carlo Localization (MCL) に語彙獲得を統合したモデルであるため，SLAM への拡張は自然に行うことができると考える．これにより，ロボットは環境の地図を生成すると同時に場所に関する語彙を逐次的に獲得することが可能となると考える．

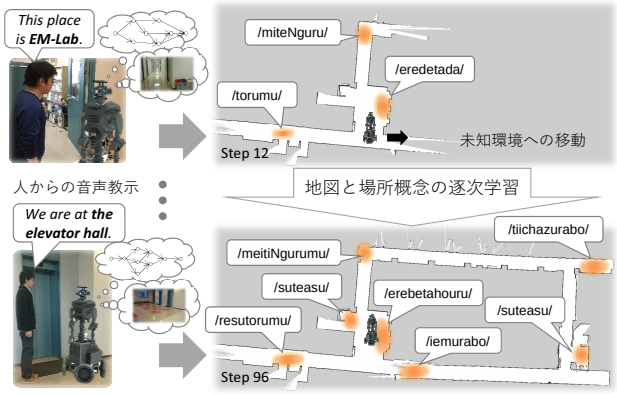

図 1: 地図と場所概念の逐次学習の概要図

2. オンライン場所概念獲得

SpCoSLAM のグラフィカルモデルを図 2 に，各変数の説明を表 1 に示す．提案手法 SpCoSLAM は SpCoA と FastSLAM をベイズ生成モデルの枠組みで統合したものであり，未知環境からの逐次的な場所概念獲得と地図生成を可能とする．また提案手法は，場所のカテゴリゼーションと語彙獲得，および地図生成を一つのモデルとして表現することにより，互いの情報の不確実性を相互補完することが可能となる．

オンラインで教師なし語彙獲得を行うためには，(i) 音声認識誤りへの逐次的な対処と，(ii) 発話文の単語分割の問題への対処が必要である．従来手法である SpCoA では，音声認識ラティスの教師なし形態素解析手法 latticelm [Neubig 12] を用いることで，音声認識のゆらぎと誤りを抑えた発話文の単語分割を行っていた．本研究ではこれをオンライン推定アルゴリズムとするために，[Araki 12] が NPYLM において行っているのと同様の方法で latticelm を疑似的にオンライン化する．また，逐次的に言語モデルを更新することで音声認識の精度向上を図る．

連絡先: 谷口彰，立命館大学大学院 情報理工学研究科，a.taniguchi@em.ci.ritsumei.ac.jp

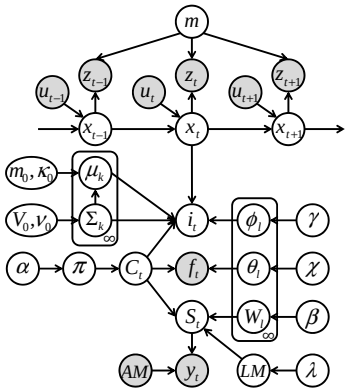

図 2: SpCoSLAM のグラフィカルモデル

表 1: SpCoSLAM のグラフィカルモデルの変数

m	環境地図
x_t	ロボットの自己位置
z_t	距離の観測値
u_t	制御値
y_t	音声情報
S_t	単語列 (単語分割結果)
f_t	画像特徴
C_t	場所概念の index
i_t	位置分布の index
μ_k, Σ_k	位置分布 (平均ベクトル，共分散行列)
π	場所概念の index の多項分布
ϕ_l	位置分布の index の多項分布
θ_l	画像特徴の多項分布
W_l	場所の名前の多項分布
LM	言語モデル (単語辞書)
AM	音響モデル
α,β,γ,χ,λ m_0,κ_0,V_0,ν_0	ハイパーパラメータ

2.1 オンライン学習アルゴリズム

提案手法のオンライン推定アルゴリズムは，RBPFs に基づく FastSLAM の定式化に，場所概念のパラメータを推定するための逐次更新式を導入することにより導出できる．本研究では，Grid-based FastSLAM 2.0 [Grisetti 05] のアルゴリズムを仮定する．SpCoSLAM の 1 ステップにおける手順は以下の通りである．手順 2-6 はパーティクル r ごとに実行される．

1. 前回のステップにおける言語モデル LM_{t-1} を用いて音声認識を行い，Weighted Finite-State Transducer (WFST) 形式で音声認識結果 $\mathcal{L}_{1:t}$ を得る．
2. WFST 形式の音声認識結果 $\mathcal{L}_{1:t}$ を用いて，latticelm による教師なし単語分割を行い，単語列 $S_{1:t}^{[r]}$ を得る．
3. FastSLAM の動作モデルと計測モデルを実行し，自己位置 $x_{0:t}^{[r]}$ と自己位置に関する観測尤度 $\omega_z^{[r]}$ を得る．
4. 式 (4) により，場所概念に関する潜在変数 $C_t^{[r]}$，$i_t^{[r]}$ のサンプリングを行う．
5. 式 (5) により，画像特徴 $f_{1:t}$ と単語列 $S_{1:t}^{[r]}$ の周辺尤度 $\omega_f^{[r]} \cdot \omega_S^{[r]}$ と $\omega_z^{[r]}$ との積を重点重み $\omega_t^{[r]}$ として得る．
6. 自己位置 $x_{0:t}^{[r]}$ と観測値 $z_{1:t}$ から地図 m の更新を行う．
7. 最大重み $\omega_t^{[\mathrm{best}]}$ のパーティクルの単語列 $S_{1:t}^{[\mathrm{best}]}$ を言語モデル LM_t に登録する．
8. 重み $\omega_t^{[r]}$ に従ってパーティクルをリサンプリングする．

SpCoSLAM の定式化とオンライン学習アルゴリズムの導出について述べる．場所概念に関するパラメータ集合を $\Theta = \{\mathbf{W}, \mu, \mathbf{\Sigma}, \theta, \phi, \pi\}$ とし，ハイパーパラメータの集合を $\mathbf{h} = \{\alpha, \beta, \gamma, \chi, \lambda, m_0, \kappa_0, V_0, \nu_0\}$ とする．また，潜在変数の集合を $\mathbf{C}_{1:t} = \{i_{1:t}, C_{1:t}, S_{1:t}\}$ と表記する．SpCoSLAM において推定する同時事後分布を式 (1) に示す．

$$\begin{aligned} &p(x_{0:t}, \mathbf{C}_{1:t}, LM, \Theta, m \mid u_{1:t}, z_{1:t}, y_{1:t}, f_{1:t}, AM, \mathbf{h}) \\ &= p(LM \mid S_{1:t}, \lambda) p(m \mid x_{0:t}, z_{1:t}) p(\Theta \mid x_{0:t}, \mathbf{C}_{1:t}, f_{1:t}, \mathbf{h}) \\ &\quad \cdot \underbrace{p(x_{0:t}, \mathbf{C}_{1:t} \mid u_{1:t}, z_{1:t}, y_{1:t}, f_{1:t}, AM, \mathbf{h})}_{\text{Particle Filter}} \end{aligned} \tag{1}$$

Particle Filter のアルゴリズムは Sampling Importance Resampling (SIR) に基づく．これより後の数式はパーティクル $[r]$ ごとに計算されるが、パーティクル番号を表す添え字は省略して記する．時刻 t における提案分布を式 (2) に示す．

$$\begin{aligned} q_t \quad = \quad & p(x_t \mid x_{t-1}, z_t, m_{t-1}, u_t) p(S_t \mid S_{1:t-1}, y_{1:t}, AM, \lambda) \\ & \cdot p(i_t, C_t \mid x_{0:t}, i_{1:t-1}, C_{1:t-1}, S_{1:t}, f_{1:t}, \mathbf{h}) \end{aligned} \tag{2}$$

このとき，$p(x_t \mid x_{t-1}, z_t, m_{t-1}, u_t)$ は FastSLAM の提案分布と同様である．

また，$p(S_t \mid S_{1:t-1}, y_{1:t}, AM, \lambda)$ は前時刻の言語モデル LM_{t-1} を用いた WFST 音声認識とその音声認識結果 $\mathcal{L}_{1:t}$ を用いた教師なし単語分割によって近似する．SR() は音声認識を表す．

$$\begin{aligned} &p(S_t \mid S_{1:t-1}, y_{1:t}, AM, \lambda) \\ &\approx latticelm(S_{1:t} \mid \mathcal{L}_{1:t}, \lambda) \mathrm{SR}(\mathcal{L}_{1:t} \mid y_{1:t}, AM, LM_{t-1}) \end{aligned} \tag{3}$$

$p(i_t, C_t \mid x_{0:t}, i_{1:t-1}, C_{1:t-1}, S_{1:t}, f_{1:t}, \mathbf{h})$ はパラメータ集合 Θ の周辺分布である．この分布は周辺化ギブスサンプリングと同様の式で計算することができる．このとき，$n_t^{(l)}$ は時刻 $t-1$ までのデータのうち，l 番目の場所概念に割り当てられたデータの個数である．$n_t^{(l,k)}$ は時刻 $t-1$ までのデータのうち，l 番目の場所概念かつ k 番目の位置分布に割り当てられたデータの個数である．$n_t^{(l,g)}$ は時刻 $t-1$ までのデータのうち，l 番目の単語分布の g 番目の単語 s_g の数を表す．G は単語の種類数 (単語分布の次元数) であり，$F_t = \sum_E f_{t,e}$ とする．$n_t^{(l,e)}$ は時刻 $t-1$ までのデータのうち，l 番目の視覚特徴の多項分布の e 次元目の視覚特徴の数を表す．E は視覚特徴の次元数である．発話文中の単語数は B_t である．St() は多変量スチューデントの t 分布である．事後ハイパーパラメータの計算については [Murphy 12] を参照のこと．

$$\begin{aligned} &p(i_t = k, C_t = l \mid x_{0:t}, i_{1:t-1}, C_{1:t-1}, S_{1:t}, f_{1:t}, \mathbf{h}) \\ &\propto \begin{cases} \prod_{B_t} \frac{n_t^{(l,g)}+\beta}{\sum_{g'=1}^{G}(n_t^{(l,g')}+\beta)} \prod_E \left(\frac{n_t^{(l,e)}+\chi}{\sum_{e'=1}^{E}(n_t^{(l,e')}+\chi)} \right)^{f_{t,e}} \\ \cdot \mathrm{St}(x_t \mid m_k, \frac{V_q^{-1}(\kappa_k+1)}{\kappa_k(\nu_k-d+1)}, \nu_k - d + 1) \cdot \frac{n_t^{(l,k)}}{n_t^{(l)}+\gamma} \cdot n_t^{(l)} \\ \qquad (n_t^{(l,k)} > 0) \\ \prod_{B_t} \frac{n_t^{(l,g)}+\beta}{\sum_{g'=1}^{G}(n_t^{(l,g')}+\beta)} \prod_E \left(\frac{n_t^{(l,e)}+\chi}{\sum_{e'=1}^{E}(n_t^{(l,e')}+\chi)} \right)^{f_{t,e}} \\ \cdot \mathrm{St}(x_t \mid m_0, \frac{V_0^{-1}(\kappa_0+1)}{\kappa_0(\nu_0-d+1)}, \nu_0 - d + 1) \cdot \frac{\gamma}{n_t^{(l)}+\gamma} \cdot n_t^{(l)} \\ \qquad (n_t^{(l)} > 0 \ \cap \ k \text{ is new}) \\ \frac{1}{G^{B_t}} \frac{1}{E^{F_t}} \cdot \mathrm{St}(x_t \mid m_0, \frac{V_0^{-1}(\kappa_0+1)}{\kappa_0(\nu_0-d+1)}, \nu_0 - d + 1) \\ \cdot \frac{\gamma}{n_t^{(l)}+\gamma} \cdot \alpha \\ \qquad (l \text{ and } k \text{ are new}) \end{cases} \end{aligned} \tag{4}$$

最終的に，目標分布/提案分布より重点重み ω_t は以下のようになる．これは，z_t, f_t, S_t に関する周辺尤度と前時刻の重み ω_{t-1} との積で表される．

$$\omega_t \approx p(z_t \mid m_{t-1}, x_{t-1}, u_t)p(f_t \mid C_{1:t-1}, f_{1:t-1}, \mathbf{h}) \cdot \frac{p(S_t \mid S_{1:t-1}, C_{1:t-1}, \alpha, \beta)}{p(S_t \mid S_{1:t-1}, \beta)} \cdot \omega_{t-1}. \quad (5)$$

3. 実験

本実験では，未知環境からの場所概念のオンライン学習を行う．また，場所に関する語彙獲得に関する評価を行う．本実験では，以下の 4 つの手法を比較した．

(A) SpCoSLAM (提案手法)
(B) Online SpCoA based on RBPFs
(C) Online SpCoA
(D) SpCoA (バッチ学習) [Taniguchi 16]

手法 (A)，(B)，(C) では，Chinese Restaurant Process (CPR)，手法 (D) では，Stick-Breaking Process (SBP) によりノンパラメトリックベイズモデルを表現した．(D) の場所概念の上限数は $L = 100$，位置分布の上限数 $K = 100$ とした．(D) のギブスサンプリングのイテレーションは 100 回行った．手法 (B)，(C)，(D) は，画像特徴を使用せず，また言語モデルの更新を行わない．

3.1 オンライン学習

提案手法の実装は，FastSLAM が実装された Robot Operating System (ROS) 上の gmapping のパッケージを拡張する形で行った．音声認識には Julius dictation-kit-v4.3.1-linux (GMM-HMM decoding) を使用した．Julius の単語辞書には日本語音節のみを登録した．教師なし単語分割には latticelm [Neubig 12] を使用した．画像特徴抽出器として Convolutional Neural Network (CNN) のツールである Caffe を使用した．CNN のモデルはプレトレーニングされた Places205-AlexNet [Zhou 14] を使用した．本実験では，Robotics Data Set Repository (Radish) [Howard 03] のオープンデータセットを用いた．データセットには音声データは含まれていないため，10 種類の言い回しを含む発話文の音声データを用意した．発話は合計 50 回であり，教示した場所の数は 10 種類，場所の名前は 9 種類である．パーティクルの数は 30 個とした．各ハイパーパラメータの値は，$\alpha = 20$, $\gamma = 10$, $\beta = 0.2$, $\chi = 0.2$, $m_0 = [0, 0]^{\mathrm{T}}$, $\kappa_0 = 0.001$, $V_0 = \mathrm{diag}(2, 2)$, $\nu_0 = 3$ とした．

図 3 に地図と位置分布が逐次的に推定される様子を示す．図の上部は，各ステップにおいて位置分布ごとに割り当てられた画像と推定された $p(S_t \mid i_t, \Theta_t, LM_t)$ の確率値が上位三つの単語の例を示す．結果として，地図と場所概念と語彙のオンライン学習が行えていることがわかる．

3.2 逐次語彙獲得における単語分割

場所の名前を表現する音素列が適切に分割されているかどうかについて確認する．図 4 に分割された単語数の推移を示す．さらに，正確な音素列を形態素で分割した場合とフレーズで分割した場合の単語数の推移も示した．形態素分割では，日本語形態素解析システムである MeCab を使用した．フレーズ分割では，場所の名前を表す音節列とそれ以外の音節列に分割した．どちらの分割法においても、場所の名前は一単語とした．表 2 に発話文の分割結果の例を示す．手法 (A) はフレーズ分割に非常に近い値を示した．手法 (B)，(C)，(D) は形態素で分割した場合よりも多く分割され，Over-segmentation が起き

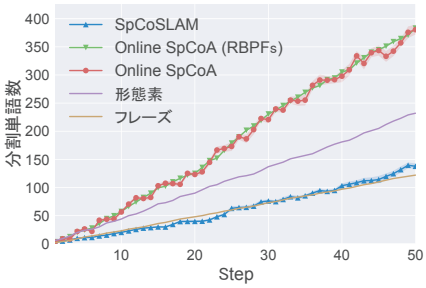

図 4: 分割された単語数の推移

表 2: 発話文の単語分割結果の例

形態素	**いきどまり** | に | き | まし | た
フレーズ	**いきどまり** | にきました
(A)	**ああえりきどまり** | にけいわすた
(B), (C), (D)	**ぴきど** | **ま** | **え** | に | き | ま | しゃ
形態素	**きょういんけんきゅうしつ** | わ | こちら | です
フレーズ	**きょういんけんきゅうしつ** | わこちらです
(A)	**きょいいんいんてんきゅしつ** | わっごちがです
(B), (C), (D)	**きょお** | **い** | **ん** | **てん** | **きゅ** | **し** | **す** | わ | こ | ち | が | です
形態素	この | ばしょ| の | なまえ | わ | **きゅうけいじょ**
フレーズ	このばしょのなまえわ | **きゅうけいじょ**
(A)	うこのましょなまえわあ | **きゅうっきりじょ**
(B), (C), (D)	この | ばしょ| の | なま | え | わ | **きゅう** | **け** | **い** | **じょ**

ていることがわかる．結果として，手法 (A) は言語モデルを逐次的に更新することによって，Over-segmentation の問題を改善した．

3.3 音声による場所の認識

ロボットがユーザからの音声発話 y_t を聞いたとき，その発話文内に含まれる場所の名前が指し示す位置 x_t^* を推定する．ユーザは “** に いって” と発話する．音声発話が与えられたときの位置推定を以下に示す．

$$x_t^* = \underset{x_t}{\mathrm{argmax}}\, p(x_t \mid y_t, \Theta, AM, LM) \quad (6)$$

しかし，式 (6) を直接計算するのは困難であるため，以下の二つの手順で推定を行う．式 (8) は音声認識の式であり，N-best の音声認識結果 $S_t^{1:N}$ を得る．式 (8) は単語が与えられたときの位置推定の式であり，潜在変数 C_t，i_t を周辺化することで求めることができる．本実験では $N = 10$ とした．

$$S_t^{1:N} \sim \mathrm{SR}(S_t \mid y_t, AM, LM) \quad (7)$$

$$x_t^* = \underset{x_t}{\mathrm{argmax}} \sum_{n=1}^{N} p(x_t \mid S_t^n, \Theta) \quad (8)$$

また，可能なすべての位置座標について式 (8) を計算するのは困難であるため，各位置分布に対して 10 個のサンプル点を候補点として用意する．本実験では，推定された位置が同じ場所として発話した位置座標を囲む長方形の領域内に入っていれば正解とする．これを各場所の名前に対して行い，場所認識率 n_c/n_u を計算する．n_u は発話の回数であり，n_c は正解数である．

図 3: 位置分布の学習の推移と位置分布に割り当てられた画像および推定された上位三つの単語の例

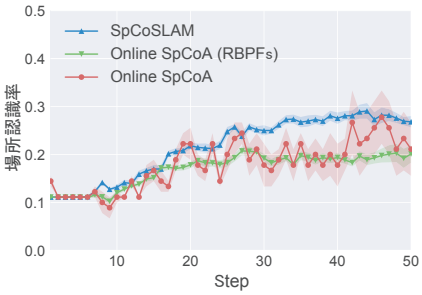

図 5: 場所認識率の推移

図 5 に 10 試行の学習結果に対する場所認識率の平均値の推移を示す．また手法 (D) の場所認識率は 0.500 であった．すべての手法で評価値は上昇傾向にあった．手法 (A) は全体として他のオンライン手法より高い評価値を示した．結果として，手法 (A) は逐次的に生成された地図上に，より適切に単語と場所の関係性を学習できたと言える．

4. おわりに

本稿では，場所概念獲得と地図生成を逐次的に行う手法について述べた．提案手法は，RBPFs に基づく FastSLAM のオンラインアルゴリズムに場所概念獲得を統合したものである．実験では，事前に語彙や地図を持たないロボットによる新規な環境上でのオンライン学習を行った．実験結果は，提案手法により地図と場所概念の逐次的な学習が可能であることを示した．また，音声からの場所の認識において提案手法が有効であることを示した．今後はさらなる精度向上のために，忘却 [Araki 12] や活性化 [Canini 09] などを提案手法に取り入れる予定である．

参考文献

[Araki 12] Araki, T., Nakamura, T., Nagai, T., Nagasaka, S., Taniguchi, T., and Iwahashi, N.: Online learning of concepts and words using multimodal LDA and hierarchical Pitman-Yor Language Model, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1623–1630IEEE (2012)

[Canini 09] Canini, K. R., Shi, L., and Griffiths, T. L.: Online Inference of Topics with Latent Dirichlet Allocation., in *AISTATS*, Vol. 9, pp. 65–72 (2009)

[Doucet 00] Doucet, A., De Freitas, N., Murphy, K., and Russell, S.: Rao-Blackwellised particle filtering for dynamic Bayesian networks, in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pp. 176–183Morgan Kaufmann Publishers Inc. (2000)

[Grisetti 05] Grisetti, G., Stachniss, C., and Burgard, W.: Improving Grid-based SLAM with Rao-Blackwellized Particle Filters by Adaptive Proposals and Selective Resampling, in *IEEE International Conference on Robotics and Automation (ICRA)* (2005)

[Howard 03] Howard, A. and Roy, N.: The Robotics Data Set Repository (Radish) (2003)

[Ishibushi 15] Ishibushi, S., Taniguchi, A., Takano, T., Hagiwara, Y., and Taniguchi, T.: Statistical localization exploiting convolutional neural network for an autonomous vehicle, in *Industrial Electronics Society, IECON 2015-41st Annual Conference of the IEEE*, pp. 1369–1375IEEE (2015)

[Murphy 12] Murphy, K. P.: *Machine learning: a probabilistic perspective* (2012)

[Neubig 12] Neubig, G., Mimura, M., and Kawahara, T.: Bayesian learning of a language model from continuous speech, *IEICE TRANSACTIONS on Information and Systems*, Vol. 95, No. 2, pp. 614–625 (2012)

[Taniguchi 16] Taniguchi, A., Taniguchi, T., and Inamura, T.: 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)

[Zhou 14] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A.: Learning Deep Features for Scene Recognition using Places Database, in *Advances in Neural Information Processing Systems 27 (NIPS)*, pp. 487–495 (2014)

[谷口 14] 谷口彰, 吉崎陽紀, 稲邑哲也, 谷口忠大：自己位置と場所概念の同時推定に関する研究, システム制御情報学会論文誌, Vol. 27, pp. 166–177 (2014)