2I3-4

自己位置と画像の高次特徴量に基づく教師なし場所領域学習

Unsupervised Learning for Spatial Concepts
by Using Self-position and High-level Features

石伏 智 *1 Satoshi Ishibushi
谷口 彰 *1 Akira Taniguchi
萩原良信 *1 Yosinobu Hagiwara
高野 敏明 *1 Toshiaki Takano
谷口 忠大 *1 Tadahiro Taniguchi

立命館大学 *1
Ritsumeikan University

In this paper, we propose a novel model to form place concept by using stochasitically integrating positional and visual information. Many conventional self-localization method can estimate robot's self-position statistically. However, for human-communication, the robot has to estimate not only its self-position but also place concepts such as "kitchen" and "living room". Our proposed method learns place concepts by using visual information and positional information. In our proposed method, as visual information and, robots use object recognition result obtained by Convolutional Neural Network. Moreover, as positional information, the robot uses the position estimated by Monte-Carlo Localization (MCL) and proposed method integrates their information statistically. The experiment to learn place concepts was performed in a real environment by using a robot integrated our proposed method. The performance of the learned place concept was evaluated by comparing with human place-concepts.

1. はじめに

従来の多くの位置推定手法はオドメトリからの情報や距離センサや RGB センサからの観測情報，さらに，環境の情報を表す地図を用いることで確率的に自己位置を推定することができる [Thrun 05]．しかし，家庭内の環境においてロボットが人間と円滑にコミュニケーションを取り生活するためにはロボット自身の位置だけではなく「キッチン」や「リビング」といった場所を表す空間的な領域（場所領域）を学習する必要がある．さらに，ロボットが位置や視覚，音声などの情報をマルチモーダルに利用して場所領域を学習することにより，ロボットは人間からの「リビングに行って」や「本棚から本をとってきて」などの命令を遂行することができると考える．そこで本研究ではロボットの自己位置情報と画像情報から場所領域を学習するものとし，Convolutional Neural Network (CNN) による物体認識結果と確率的な位置推定手法である Monte-Carlo Localization を統計的に統合したモデルを提案する．本稿では提案手法によって学習された場所領域と人間が認識する場所領域との比較を行い，提案手法によって学習された場所領域がどの程度，人間が認識しているものと近いかの検証を行う．さらに，視覚情報によってクラスタリングされた結果との比較を行い，位置と視覚情報によって場所領域を学習することの有用性を検証する．

2. 提案手法

本研究では場所領域を推定するための生成モデルを提案する．Fig. 1 は本研究で提案する場所領域とロボットの自己位置の同時推定を行うグラフィカルモデルである．まず，本稿では t 時刻に推定した自己位置を $\boldsymbol{x_t}$ とし，以下のように定義する．

$$\boldsymbol{x_t} = (x_t, y_t, \sin\theta_t, \cos\theta_t). \tag{1}$$

ここで x_t, y_t は x, y の 2 次元座標で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．また，グラフィカルモデル上の 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．場所領域はあらかじめ推定した地図中に複数個存在する．本稿では場所領域のインデックスを C_t とし，$C_t \in \boldsymbol{C} = \{1, 2, \cdots, L\}$ と定義する．L は推定した地図中に存在する場所領域の個数である．また，本稿では t 時刻に観測した画像から得られる特徴ベクトルを f_t とし，$f_t = (f_t^1, f_t^2, \cdots, f_t^I)^{\mathrm{T}}$ と定義する．I は特徴量のベクトルの次元数である．これらのパラメータにより，ロボットの自己位置 $\boldsymbol{x_t}$ に関する信念 $p(\boldsymbol{x_{0:t}}|z_{1:t}, u_{1:t}, f_{1:t})$ が以下のように得られる．

$$\begin{aligned} p(\boldsymbol{x_{0:t}}|z_{1:t}, u_{1:t}, &f_{1:t}) \\ &\propto p(z_t|\boldsymbol{x_t})p(f_t|\boldsymbol{x_t})p(\boldsymbol{x_t}|\boldsymbol{x_{t-1}}, u_t) \\ &\times p(\boldsymbol{x_{0:t-1}}|z_{1:t-1}, f_{1:t-1}, u_{1:t-1}). \end{aligned} \tag{2}$$

事後確率 $p(f_t|\boldsymbol{x_t})$ は自己位置 $\boldsymbol{x_t}$ の条件下で特徴ベクトル f_t を観測する確率であり，以下のように得られる．

$$p(f_t|\boldsymbol{x_t}) \propto \sum_{C_t} p(f_t|\varphi_{C_t})p(\boldsymbol{x_t}|\mu_{C_t}, \Sigma_{C_t})p(C_t). \tag{3}$$

ここで式 (2) はベイズの定理により，式 (3) は周辺化によって導かれる．式 (3) に含まれる確率 $p(\boldsymbol{x_t}|\mu_{C_t}, \Sigma_{C_t}), p(f_t|\varphi_{C_t})$ はそれぞれ多次元ガウス分布と多項分布によって得る．多次元ガウス分布の平均ベクトルと共分散行列をそれぞれ μ, Σ とし，多項分布のパラメータを φ とする．ディリクレ事前分布のハイパーパラメータは α とする．多次元ガウス分布と多項分布はそれぞれ場所領域の個数と同じが図だけ存在し，インデックス C_t の場所領域の各分布のパラメータを $\mu_{C_t}, \Sigma_{C_t}, \varphi_{C_t}$ とする．確率 $p(f_t|\varphi_{C_t})$ は以下の式のように多項分布で得る．

$$p(f_t|\varphi_{C_t}) = \mathrm{Mult}(f_t|\varphi, K) \tag{4}$$

ここで K は $K = \sum_{i=1}^{I} f_t^i$ である．また，確率 $p(\boldsymbol{x_t}|\mu_{C_t}, \Sigma_{C_t})$

連絡先: 石伏 智, 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 立命館大学情報理工学部,ishibushi@em.ci.ritsumei.ac.jp

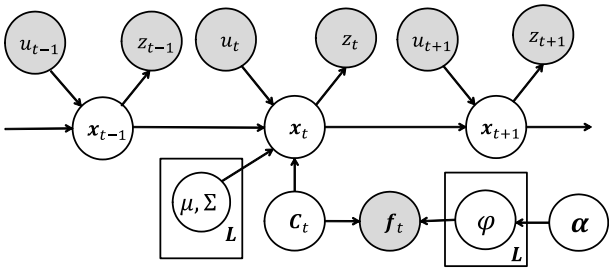

図 1: Graphical model of the proposed method

は以下のように多次元ガウス分布で得る．

$$p(\boldsymbol{x_t}|\mu_{C_t}, \Sigma_{C_t}) = \frac{1}{\sqrt{2\pi}^m |\Sigma_{C_t}|^{1/2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{x_t} - \mu_{C_t})^T {\Sigma_{C_t}}^{-1}(\boldsymbol{x_t} - \mu_{C_t})\right\}. \quad (5)$$

ここで自己位置 x_t の次元数を m とし，本研究では $m = 4$ である．さらに 式 (3) の場所領域 C_t の事前分布 $p(C_t)$ は一様な分布から得られるものと仮定し，$p(C_t) = \frac{1}{L}$ とする．

2.1 CNN の物体認識結果の利用

本研究では CNN によって観測した画像の物体認識結果を高次特徴量とみなし，特徴ベクトル f_t として用いる．CNN は Deep learning と呼ばれる 3 層以上の構造をもつニューラルネットワークの一つである．CNN は畳込み層とプーリング層よばれる層をもつ構造により，大きさの変化や位置のズレに対して不変な特徴が得られることが知られており，物体認識などの研究で用いられている [Krizhevsky 12]．また，最後のプーリング層にすべて結合した層を配置し，出力層を学習した物体のカテゴリ数と同じにすることで，t 時刻に入力された画像 V_t を物体カテゴリ ${O_t}^i$ に分類する確率 $P({O_t}^i|V_t)$ を得ることができる．

$$P({O_t}^i|{V_t}^i) = \frac{e^{h^i}}{\sum_{k=1}^{I} e^{h^k}}. \quad (6)$$

式 (6) のように $P({O_t}^i|{V_t}^i)$ はソフトマックス関数で計算する．ここで h_i は i 番目のユニットからの出力である．本研究では以下の式のように，CNN から得られた確率 $P({O_t}^i|{V_t}^i)$ を 10^s 倍し，小数点以下を切り捨てた値を特徴ベクトルの要素 ${f_t}^i$ とする．

$${f_t}^i = P({O_t}^i|{V_t}^i). \quad (7)$$

Fig.2 は CNN によって物体認識した結果の例と CNN の全体概要である．Fig.2 中の単語と数字は物体のクラスラベルと入力画像をその物体のクラスに分類する確率である．Fig.2 では最も確率が高い物体のクラスは「turnstile (改札口)」である．

Fig.2 の画像に写っている物体は空気清浄機である．ところが，CNN によって物体認識した結果では「turnstile（改札口)」,「postbag(ポスト)」といった別の物体のクラスに分類される確率が高くなっている．しかし，このような結果が得られた場合でも，ロボットは自身が存在している場所領域では「tunstile(改札口)」,「postbag(ポスト)」といった要素の値が大きい特徴ベクトルを高い確率で観測することを学習できる．そのため，ロボットが「tunstile(改札口)」,「postbag(ポスト)」といった要素の値が大きい特徴ベクトルを観測したとき，ロボットは自身の位置が空気清浄機の近くである推定することができる．

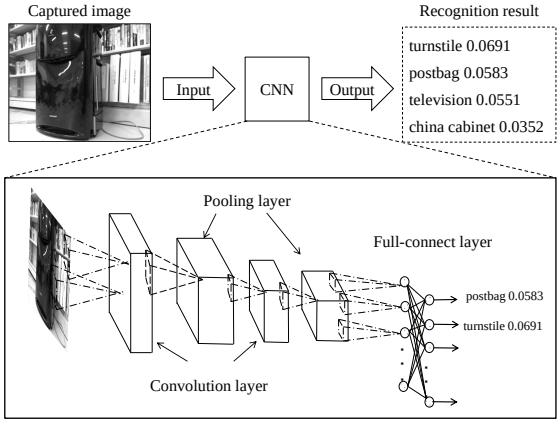

図 2: CNN の構造と物体認識結果の例

提案手法では各場所領域で観測した特徴ベクトルの要素を Bag-of-Features(BoF) 表現で数え上げる．数え上げた特徴ベクトルとその要素を $\boldsymbol{m} = (m_1, \cdots, m_I)^T$ とする．多項分布のパラメータ φ_l をディリクレ事前分布と BoF によって数え上げた結果からディリクレ事後分布を計算して，サンプリングによって得る．

$$\boldsymbol{\varphi}_l \sim \mathrm{Dir}(\boldsymbol{\varphi}_l|\boldsymbol{\alpha} + \boldsymbol{m}) = \frac{\Gamma(\alpha_0 + K)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_I + m_I)} \prod_{i=1}^{I} \varphi_{l,i}^{\alpha_i + m_i - 1} \quad (8)$$

また，$\boldsymbol{\alpha}$ はディリクレ分布のハイパーパラメータであり $(\alpha_1, \cdots \alpha_I)^T$ を表す．

2.2 場所領域の学習

本研究での場所領域の学習は多項分布とガウス分布の各パラメータ μ_c, Σ_c, φ_c を推定することを意味する．提案手法では多くの自己位置データとそこで観測された特徴ベクトルのデータを集め，ギブスサンプリングにより各パラメータを推定する．Algorithm 1 に本研究で用いるギブスサンプリングのアルゴリズムを記述する．ここで集められたデータ数の数を N としている．

3. 実験

本研究では提案手法により学習された場所領域の有用性を検証するために学習された場所領域と人間が認識している場所領域を比較する．比較方法として，まず，一人のユーザが学習に用いられた画像をそのデータが取得できると推測する場所ごとに分類する．そのとき，ユーザには別の三人のユーザが決めた場所の名前とその範囲が提示される．本稿では場所の数を 19 とした．図 3，4 が三人のユーザが決めた場所の範囲と場所の名前のリストである．

ユーザによって画像が分類された結果を教師ラベルとし，提案手法によって各場所領域に割り当てられた結果のクラスタリング精度を検証する．本稿ではクラスタリング精度の指標として Adjusted rand intex (ARI) を用いる．また，画像情報だけによって画像が分類された結果のクラスタリング精度と比較

Algorithm 1 The Learning Place Concepts by Gibbs sampling

$X = \{x_1, x_2, \cdots, x_N\}$
$F = \{f_1, f_2, \cdots, f_N\}$
procedure LEARNING PLACE CONCEPTS(X,F,L,N)
  $C_t = \{C_{t_1}, C_{t_2}, \cdots, C_{t_N}\}$
  //(1)Initializing $\boldsymbol{\mu}, \boldsymbol{\Sigma}$
  $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ = Initialize()
  //(5)Repeating from (2) to (4)
  for $i = 1$ to $iteration\_number$ **do**
    //(2) Allocating the data to each place concept by the sampling of the index of place concept C_t
    for $t = 1$ to N **do**
      $C_t \sim p(C_t \mid \boldsymbol{x}_t, f_t, \varphi, \mu, \Sigma)$
    end for
    //(3) Sampling parameters of Gaussian distributions μ and Σ
    for $c = 1$ to L **do**
      $\mu_c, \Sigma_c \sim p(X_c \mid \mu_c, \Sigma_c)p(\mu_c, \Sigma_c)$
    end for
    //(4) Sampling parameter of a multinomial distribution φ
    for $c = 1$ to L **do**
      $\varphi_c \sim p(F_c \mid \varphi_c)p(\varphi_c)$
    end for
  end for
  return μ, Σ, φ
end procedure

し，位置と視覚情報によって場所領域を学習することの有用性を検証する．本実験では比較手法はクラスタリングのために k-means 法と特徴ベクトルとして提案手法と同様に CNN の物体認識結果を用いた．

3.1 実験条件

本実験では場所領域を学習させるために 1913 セットの位置と画像のデータを実環境で集めた．その際ロボットは SLAM によって推定された地図を持っていおり MCL によって位置推定しながら環境内を自律移動することでデータを取得した．ロボットとしては Turtlebot 2 を用いた．また，CNN を用いるために，学習済みの CNN のフレームワークである Caffe を使用した [Jia 14].

3.2 実験結果

図 5 に提案手法によって学習された場所領域の結果とその場所領域に割り当てられた画像の例を示す．学習された結果の橙色に色付けされた場所領域は図 3，4 の A (In front of the door) の場所領域を学習した結果だと考えられる．また，青，緑，赤に色付けされた場所領域はそれぞれ図 3，4 の B (In front of the book shelf)，O (In front of the white shelf)，Q (In front of the white board) と対応していると考えられる．また，提案手法のクラスタリングと画像の物体認識結果のみによってクラスタリングの比較実験の結果が図 6 である．結果より，提案手法によってクラスタリングした結果の方が ARI の値が高いがわかる．これらの結果から提案手法は画像の物体認識結果だけを用いてクラスタリングする結果よりも人間によってクラスタリングされた結果と類似していることが分かった．

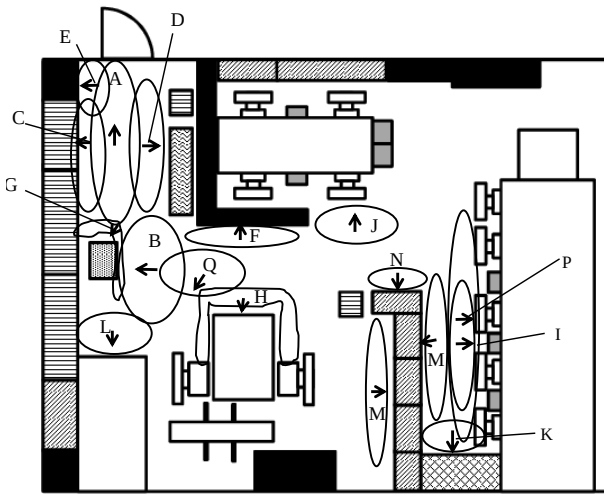

図 3: 三人のユーザが決めた場所の範囲.

A : In front of the door
B:In front of the book shelf
C:In front of the book stand
D:The side of chair yard
E:In front of the umbrella stand
F:In front of the partition
G:In front of the air cleaner
H:Meeting space
I:In front of the cabinet (Students working space)
J:In front of the cabinet(secretaries working space)
K:In front of the Sever machine
L : The place for something
M:In front of the white shelf (the side of meeting space)
N:In front of the white shelf(The side of secretaries working space)
O:In front of the white shelf(The side of student working space)
P:In front of the desk
Q: In front of the white board
R:Others

図 4: 三人のユーザが決めた場所の名前のリスト.

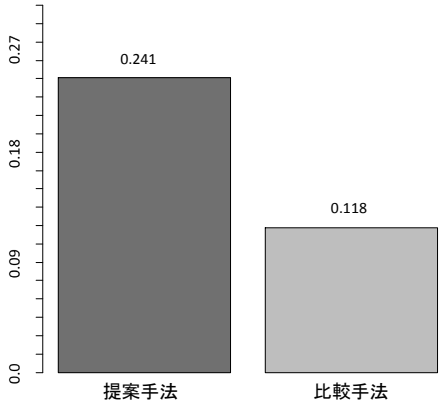

図 6: クラスタリング精度の比較実験結果

4. おわりに

本論文では MCL に Convolutional Neural Network の物体認識結果を確率的に統合した，場所領域推定手法を提案した．

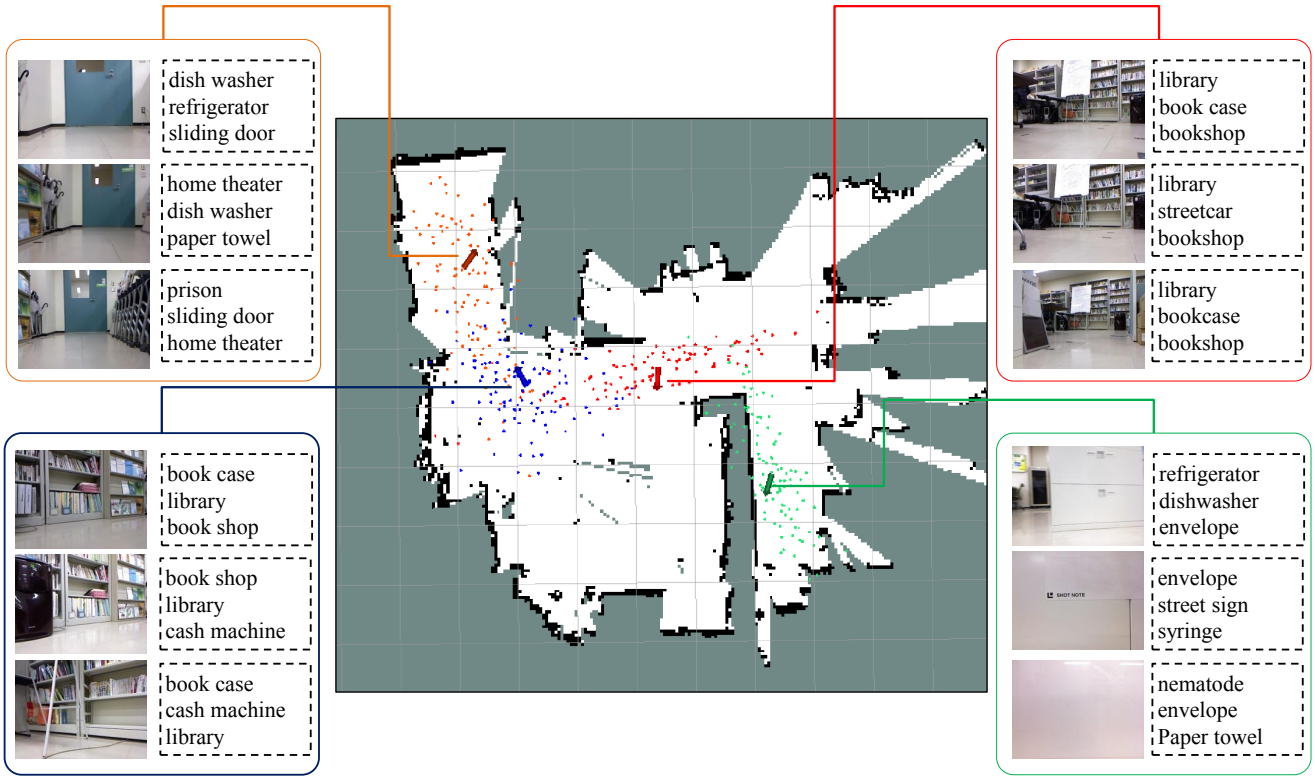

図 5: 提案手法によって学習された場所領域とその場所領域に割り当てられた画像の例，各画像の右のリストはその画像を物体認識したときに確率が高いカテゴリの上位クラスである．各矢印が場所領域のガウス分布の平均であり同じ色で色付けされた点がその場所領域の空間的広がりを表現している．

実験では分類された画像に基づいて，提案手法によって推定された場所領域の結果と人間が持つ場所領域との比較を行った．さらに画像の物体認識結果だけを用いてクラスタリングした結果との比較を行った．実験の結果，提案手法方が画像の物体認識結果だけを用いたクラスタリングよりも人間によってクラスタリングされた結果と類似していることが分かった．このことより，提案手法によって位置情報と画像情報の両方を用いて場所領域を生成することにより，より人間が持っているものに近い場所領域を生成できると考える．本研究では画像情報とロボットの位置情報から場所領域を生成しているが，人間からの「キッチンに行って」や「本棚から本を取ってきて」などのタスクを行うためには「キッチン」や「本棚」といった人間の場所を意味する言語情報が必要であると考える．そのため，人間からの言語情報などをマルチモーダルに統合することによりさらに人間に近い場所領域の生成を行えるようにすることが今後の課題である．

参考文献

[Jia 14] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T.: Caffe: Convolutional architecture for fast feature embedding, in *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678 (2014)

[Krizhevsky 12] Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, Inc.Curran Associates (2012)

[Thrun 05] Thrun, S., Burgard, W., and Fox, D.: *Probabilistic robotics*, Inc.MIT press (2005)