没入型インタフェースを伴うクラウド仮想環境における HRIに基づく場所の学習

浅田 和弥[1,a)]　萩原 良信[2,b)]　谷口 忠大[2,c)]　稲邑 哲也[3,d)]

概要：介護や案内といった，人の生活支援を行うロボットの需要が近年増加している．このようなロボットは，活動環境を構成する場所の領域と，その領域を表現する語彙の知識を獲得する必要があると考えられる．本研究では，この場所に関する知識を場所の概念とし，仮想環境において収集したマルチモーダル情報から場所の概念を形成するクラウド概念学習システムを構築する．本稿では，人間の代替となるアバターの追加や仮想環境への没入等の，システムに対する改良を報告する．

Learning Based on HRI in Cloud Virtual Environment with Immersive Interface

Asada Kazuya[1,a)]　Hagiwara Yoshinobu[2,b)]　Taniguchi Tadahiro[2,c)]　Inamura Tetsunari[3,d)]

Abstract: Demand for robots that support people's lives, such as nursing care and guidance, has been increasing in recent years. It is thought that the robot needs to acquire the area of the place constituting the human activity environment and the knowledge of the vocabulary expressing it. In our research, we will construct a cloud system for concept learning which forms the location concept from multimodal information collected in the virtual environment, with the knowledge on this place as the location concept. In this paper, we improved the system such as addition of avatar as a substitute for human and immersion into virtual environment.

1. はじめに

近年，介護や案内といった，人間の生活を支援するサービスロボットの需要が増加している．このようなロボットが家庭環境において人間と共存する場合，「玄関」や「台所」など，そのロボットが導入された環境内における場所の名前や，その名前が意味する空間的な領域といった，活動環境を構成する場所の領域と，その領域を表現する語彙の知識を獲得する必要があると考えられる．

このような場所の概念をロボットに獲得させる手法として，石伏らは MCL(Monte Carlo Localization) および CNN(Convolutional Neural Network) による物体認識結果を統計的に統合することで，似た物体認識結果が得られる場所の領域を獲得させるモデルを提案している [1]．しかし，このモデルにおいては，ある 1 つの環境において 1 人の教師がロボットに対して学習をさせるのに数日を要している．このため，多様な環境および多数の教師がロボットに対して学習をさせることは，大きな負担となる．

そこで本研究では，仮想環境において収集したマルチモーダル情報から場所の概念を形成する，クラウド概念学習システムの構築を行う．これにより，ある 1 つの環境において，1 人の教師がロボットに対して学習をさせるのに要する時間の短縮が見込まれ，複数の教師および環境にお

[1] 立命館大学大学院情報理工学研究科
Graduate School of Information and Science Engineering, Ritsumeikan University
[2] 立命館大学情報理工学部
College of Information and Science Engineering, Ritsumeikan University
[3] 国立情報学研究所/総合研究大学院大学
National Institute of Informatics/The Graduate Univ. for Adv. Studies
[a)] k.asada@em.ci.ritsumei.ac.jp
[b)] yhagiwara@em.ci.ritsumei.ac.jp
[c)] taniguchi@em.ci.ritsumei.ac.jp
[d)] inamura@nii.ac.jp

図 1 システムの概要図
Fig. 1 System overview

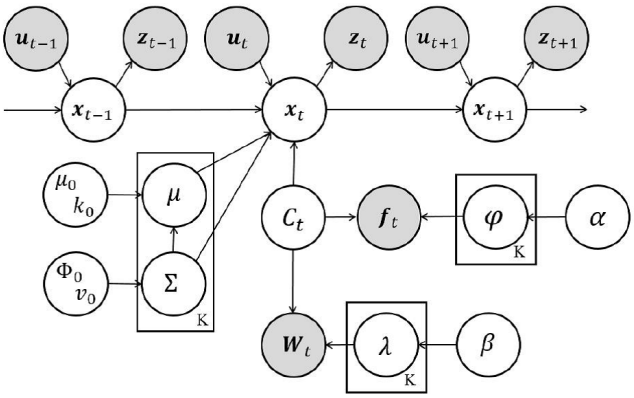

図 2 福井らの提案した手法のグラフィカルモデル
Fig. 2 The graphical model of method presented by Fukui.

ける概念の学習が容易となる．また，使用した環境および取得したデータや学習の結果を，クラウド上で共有することが可能となる．さらに，これらの共有された学習結果を統合することで，複数の教師による知識を統合した社会的マルチモーダル概念の形成が期待できる．この統合により形成された概念は，ロボットが未知の環境に導入された場合に事前に与えることができる，汎化された知識としての利用ができると考えられる．

これまで著者らは，仮想環境内における場所の概念の形成に関して研究を進めてきた [2]．これは本システムの構築の一環として，仮想環境内において画像やロボットの自己位置等の情報を収集し，それらの情報を用いて場所の概念の形成に関する検証を行うものであった．しかし，その検証実験の条件と，想定しているシステムとの間には，環境の観察に用いるインタフェース等，現状では大きな差があるという問題点が挙げられた．本稿では，このような問題点に対する改良について報告する．

2. 提案するシステム

2.1 提案するシステムの概要

提案するシステムの概要図を図 1 に示す．

仮想環境を用いる利点としては，実世界上で環境を構築するうえでかかる時間や金銭におけるコストが削減できること，また，教示をする者自身もその実世界の環境へ直接加わらなければならないという，物理的な距離の制限を無視できることなどが挙げられる．システムへログインしたユーザは，仮想環境内のロボットやアバターを joystick のようなコントローラ等により操作し，仮想環境内における画像や自己位置，語彙の情報を収集する．これら収集した情報をオフラインにて学習し，場所の概念の形成を行う．また，形成された概念や使用した環境は，クラウド上で共有することが可能である．仮想環境は対話・力学・知覚の統合的なシミュレーションが可能なシミュレータである SIGVerse [3] を用いる．SIGVerse は，ユーザの全身運動を仮想環境中のアバターに反映させるために提供されている Microsoft Kinect 用インタフェースプラグインの利用のほか，HMD(Head Mounted Display) の使用によるアバターの頭部の動きの制御およびアバターの視点からの映像の投影により，アバターとして仮想環境へ没入しているかのようなインタラクションが可能である．

表 1 福井らの提案した手法の各要素
Table 1 Parameters of method presented by Fukui.

x_t	ロボットの自己位置
u_t	制御値
z_t	計測値
C_t	場所概念の index
$\boldsymbol{\mu}, \boldsymbol{\Sigma}$	各ガウス分布の平均および共分散
φ	画像特徴の多項分布のパラメータ
λ	語彙の多項分布のパラメータ
α	画像特徴のディリクレ分布の ハイパーパラメータ
β	語彙のディリクレ分布の ハイパーパラメータ
$\boldsymbol{W}_t$	語彙のベクトル
$\boldsymbol{f}_t$	画像の特徴ベクトル
μ_0, k_0	ガウス分布のハイパーパラメータ
Φ_0, v_0	逆ウィシャート分布のハイパーパラメータ

2.2 場所概念の獲得モデル

本研究における場所の概念をロボットに獲得させるモデルは，福井らの提案したモデルを参考とする [4]．図 2 に福井らの提案した手法のグラフィカルモデルを示す．また，**表 1** にグラフィカルモデルの各要素の概要を示す．

福井らのモデルでは，CNN による物体認識結果と自己位置および場所を表す語彙の教示により，場所の概念を決定する．ただし，福井らはロボットの自己位置を MCL [5] とよばれる手法により推定しているが，本研究ではロボットの自己位置は確定的に得られるものとしている．故に，図 2 における u_t, z_t は観測されないものとなる．また，CNN とは，ニューラルネットワークの隠れ層を複数重ねあわせた

構造を持つ，深層学習の手法の 1 つである [6]．この手法は畳み込み層とプーリング層という 2 種類の層を繰り返し重ねることで構成されており，一般物体認識などにおいて利用されている．

図 2 において，それぞれの場所概念のインデックスを C_t とし, 次のように定義する．

$$C_t \in \boldsymbol{C} = \{1, 2, \cdots, K\} \tag{1}$$

K は場所概念の総数であり，C は場所概念のインデックスの集合である．それぞれの場所概念は多変量ガウス分布で表され，次の式により定義される．

$$p(x_t|\mu_{C_t}, \Sigma_{C_t}) \propto \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_{C_t}|^{\frac{1}{2}}} \times \exp(-\frac{1}{2}(x_t - \mu_{C_t})^T \Sigma_{C_t}^{-1}(x_t - \mu_{C_t})) \tag{2}$$

x_t はロボットの自己位置と向きを示しており，$(x, y, \sin\theta, \cos\theta)$ で表される．x, y は二次元座標で表されたロボットの位置座標であり，θ はロボットの向きを示す．

$\boldsymbol{f}_t$ は CNN によって得られた物体認識結果を示しており，I を CNN が学習した物体の総数として，

$$\boldsymbol{f}_t = \{f_t^1, f_t^2, \cdots, f_t^I\} \tag{3}$$

と定義される．

$\boldsymbol{W}_t$ は語彙情報を表しており，画像特徴ベクトル $\boldsymbol{f}_t$ が学習した物体の総数分の次元数を有していることと同じく，$\boldsymbol{W}_t$ においても学習した語彙の総数分の次元数を持たせる．ゆえに，$\boldsymbol{W}_t$ は J を学習した語彙の総数として，

$$\boldsymbol{W}_t = \{W_t^1, W_t^2, \cdots, W_t^J\} \tag{4}$$

と定義される．

μ, σ は，それぞれ多変量ガウス分布と逆ウィシャート分布よりサンプリングを行う．すなわち，$N(\cdot)$ をガウス分布，$IW(\cdot)$ を逆ウィシャート分布とすると，

$$\Sigma \sim IW(\Sigma|\Phi_0, v_0) \tag{5}$$

$$\mu \sim N(\mu|\mu_0, (k_0\Sigma^{-1})^{-1}) \tag{6}$$

である．

画像特徴と語彙の多項分布は，それぞれ場所概念ごとに存在し，次のように表される．

$$p(\boldsymbol{f}_t|\varphi) = \mathrm{Multi}(f_t^1, f_t^2, \cdots, f_t^I|\varphi, N) \tag{7}$$

$$= \frac{N!}{f_t^1! f_t^2! \cdots f_t^I!} \prod_{i=1}^{I} \varphi_i^{f_t^i} \tag{8}$$

$$p(\boldsymbol{W}_t|\lambda) = \mathrm{Multi}(W_t^1, W_t^2, \cdots, W_t^J|\lambda, M) \tag{9}$$

$$= \frac{M!}{W_t^1! W_t^2! \cdots W_t^J!} \prod_{j=1}^{J} \lambda_j^{W_t^j} \tag{10}$$

ここで，N, M は

$$N = \sum_{i=1}^{I} f_t^i \tag{11}$$

$$M = \sum_{j=1}^{J} W_t^j \tag{12}$$

である．

また，$p(\boldsymbol{f}_t|\varphi), p(\boldsymbol{W}_t|\lambda)$ のパラメータ φ, λ はディリクレ分布事後分布からのサンプリングにより得る．それぞれのディリクレ分布の式を次に示す．

$$\boldsymbol{\varphi} \sim \mathrm{Dir}(\boldsymbol{\varphi}|\boldsymbol{\alpha} + \boldsymbol{p}) \tag{13}$$

$$= \frac{\Gamma(\alpha_0 + p_0)}{\Gamma(\alpha_1 + p_1) \cdots \Gamma(\alpha_I + p_I)} \prod_{i=1}^{I} \varphi_i^{\alpha_i + p_i - 1} \tag{14}$$

$$\boldsymbol{\lambda} \sim \mathrm{Dir}(\boldsymbol{\lambda}|\boldsymbol{\beta} + \boldsymbol{q}) \tag{15}$$

$$= \frac{\Gamma(\beta_0 + q_0)}{\Gamma(\beta_1 + q_1) \cdots \Gamma(\beta_J + q_J)} \prod_{j=1}^{J} \lambda_j^{\beta_j + q_j - 1} \tag{16}$$

ここで，$\boldsymbol{\alpha}, \boldsymbol{\beta}$ はそれぞれ画像特徴と語彙のディリクレ分布のハイパーパラメータであり，

$$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_I)^T \tag{17}$$

$$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_J)^T \tag{18}$$

と表される．さらに，$\boldsymbol{p}, \boldsymbol{q}$ は，それぞれ場所概念ごとで画像特徴ベクトルと語彙ベクトルを Bag-of-Features，Bag-of-Words 化したものであり，

$$\boldsymbol{p} = (p_1, \cdots, p_I)^T \tag{19}$$

$$\boldsymbol{q} = (q_1, \cdots, q_J)^T \tag{20}$$

と表す．また，$\alpha_0, \beta_0, p_0, q_0$ は，それぞれ $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{q}$ の要素の総和である．

2.3 場所概念の学習

本研究では，仮想環境内におけるロボットの自己位置とその自己位置で得た画像特徴および語彙のデータを収集し，その上で場所の概念ごとのパラメータをオフラインで学習を行う．本研究においては，これらのパラメータの学習を Gibbs Sampling により行う．

3. システムの改良

3.1 これまでの実験環境

著者らはこれまでの研究において，SIGVerse 内においてロボットを移動させ，画像や自己位置，語彙の情報を収集し，それらに基づく場所の概念の形成することを試みた [2]．この試みにおいて，SIGVerse 内のロボットを移動させるインタフェースとして，ソニー・インタラクティブエンタテインメント社の Dualshock3 コントローラを用い，環境の観察には通常のディスプレイを用いた．ロボットは SIGVerse 上に作成された室内を移動しながら，画像とその

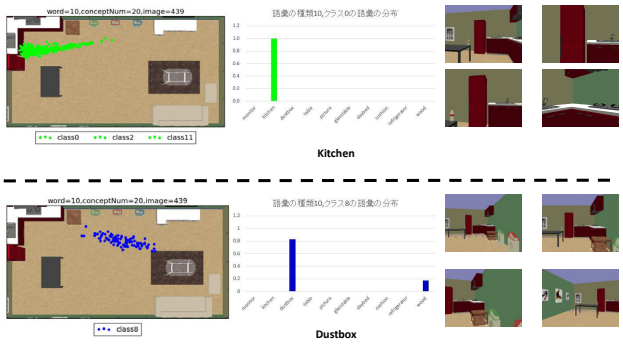

図 3 場所の概念の形成例
Fig. 3 Examples of forming location concepts

撮影位置を取得する．また，語彙の情報として，場所の名前を表すテキストのメッセージを付与した．取得した画像から画像特徴ベクトルを得るため，学習済みの CNN ツールである Caffe [7] を用いた．この実験において形成された場所の概念について，最も生起確率が高い語彙を表した場所の概念の位置分布，およびその生起確率と得られた画像の例を図 3 に示す．

しかし，上記の通り，この実験環境では仮想環境の観察に通常のディスプレイを用いる，ロボット自体を直接操作する等の点で，想定しているシステムとは大きな差があることが問題点として挙げられる．

3.2 改良を行った点

先述したような点を改良するために，変更を行った部分を述べる．

3.2.1 ロボットのアバターへの追従

これまでの環境では，仮想環境内には人間の分身であるアバターは存在せず，教示者はロボット自体をコントローラで操作し，仮想環境内の情報を取得していた．これは，教示者とロボットが仮想環境内において同一の存在となっている不自然な状態であり，教示者とロボットは仮想環境内に別々に存在することが本来は望ましいと思われる．

そこで，仮想環境内にアバターを追加し，教示者はそのアバターをコントローラにより操作する仕様に変更した．一方，ロボットはそのアバターの立っている位置を認識し，アバターが移動するとその背後を追従移動する仕様に変更した．ロボットは移動と同時に自己位置を取得し，さらにその位置で頭部に設置されたカメラから画像を取得する．アバターに追従するにあたり，正面のカメラのみから画像を取得する場合，常にアバターが画像に写り込んでしまうため，頭部の前後左右，計 4 箇所にカメラを設置した．図 4 に，ロボットのカメラから撮影した画像の一例を示す．

3.2.2 仮想環境への没入

仮想環境の観察に通常のディスプレイを用いる場合，第三者の視点でアバターやロボットを見ることとなる．しかし，実世界上においては，ロボットと対面して対話を行うことが予想される．ロボットと一人称の視点で向き合うことに関して重要な事項の 1 つとして，共同注視の実現が挙げられる．現状は，ロボットは移動した際の自己位置と，その位置で得られた画像の情報を取得している．この状態では，ロボットは，教示者が特に教示したいと考えている対象が何であるかを把握することができず，また，ロボット自身が移動可能な場所の情報しか得ることができない．しかし，ロボットと教示者による共同注視が実現できれば，例えば「台所」という場所に関して，さらにそこからロボットが直接移動することのできない「シンク」や「オーブン」といった対象を教示者が指し示し，その位置などの教示を行うことが可能となる．このことから，アバターの視点で仮想環境に没入することが望ましいと考えられる．

そこで，ヘッドマウントディスプレイを用いて，アバターの視点での仮想環境への没入が可能となるように仕様を変更した．アバターの視点の画像の一例を図 5 に示す．

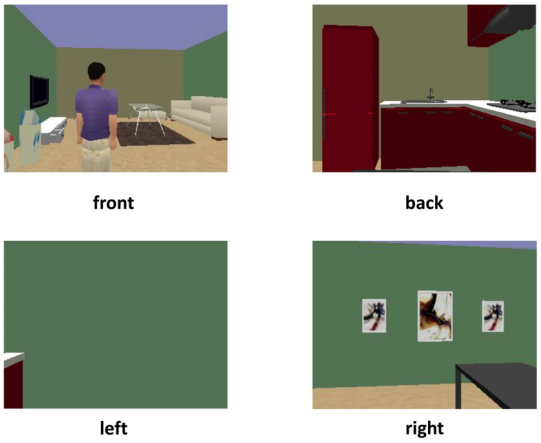

図 4 ロボットのカメラから撮影した画像例
Fig. 4 Examples of images taken from robot cameras

図 5 アバターの視点の画像例
Fig. 5 An example of the avatar's viewpoint

4. 今後に関して

仮想環境において収集したマルチモーダル情報から場所の概念を形成する，クラウド概念学習システムの構築に関して，本稿では，特に仮想環境内への没入に対する重要性と，システムに対する没入の実行に関して述べた．また，仮想環境においてロボットがアバターに追従すると同時に，前後左右の画像が取得できていることを確認した．今後は，これらの得られた自己位置等の情報から場所の概念を形成することを行うとともに，音声対話の追加等の，没入感をより深めるためのインタフェースの充実を行う．また，社会的マルチモーダル概念の形成を行うため，クラウド上に共有された概念を統合するための理論や手法の提案を行う．

謝辞 本研究は JSPS 科研費 JP16K16133 の助成を受けたものである．

参考文献

[1] Satoshi Ishibushi, Akira Taniguchi, Toshiaki Takano, Yoshinobu Hagiwara, and Tadahiro Taniguchi. : Statistical localization exploiting convolutional neural network for an autonomous vehicle., *Industrial Electronics Society, IECON 2015-41st Annual Conference of the IEEE.* IEEE, pp. 1369-1375, 2015.

[2] 浅田和弥， 萩原良信，谷口忠大，稲邑哲也：クラウド仮想環境におけるマルチモーダル情報に基づいた場所概念の形成，計測自動制御学会 システム・情報部門 学術講演会 2016，2016.

[3] Tetsunari Inamura, et al. : Simulator platform that enables social interaction simulation—SIGVerse: SocioIntelliGenesis simulator., *System Integration (SII), 2010 IEEE/SICE International Symposium on.* IEEE, pp. 212-217, 2010.

[4] 福井隆士，石伏智，萩原良信，谷口忠大，高野敏明：移動ロボットによる画像特徴と語彙を統合した教師なし場所概念獲得，第 60 回システム制御情報学会研究発表講演会，2016.

[5] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. : *Probabilistic Robotics.*, The MIT Press, 2005.

[6] Krizhevsky Alex, Ilya Sutskever, and Geoffrey E. Hinton. : Imagenet classification with deep convolutional neural networks., *Advances in neural information processing systems.*, pp. 1097-1105, 2012.

[7] Jia Yangqing, et al. ： Caffe: Convolutional architecture for fast feature embedding., *Proceedings of the 22nd ACM international conference on Multimedia.* ACM, pp. 675-678, 2014.