1I4-OS-09a-1

ユーザ行動の予測と命令解釈の統合に基づくロボットの行動決定手法

Action Decision Method for Service Robots based on Integration of Predicted User Behavior and Command Interpretation

アッタミミ ムハンマド *1 Muhammad Attamimi　中村 友昭 *1 Tomoaki Nakamura　長井 隆行 *1 Takayuki Nagai　岩橋 直人 *2 Naoto Iwahashi　奥乃 博 *2 Hiroshi Okuno

*1電気通信大学　情報理工学研究科
The University of Electro-Communications, Faculty of Informatics and Engineering

*2京都大学 情報学研究科
Kyoto University, Graduate School of Informatics

Generally, service by robots will be determined by the speech command of the user. In order to interpret speech commands correctly, the context might help in many cases. In this study, a situation of a robot which is in close contact with the life of person is considered. Human daily behavior is observed by robots and an unsupervised learning of those behavior patterns is made to realize action decision. The decision made by robots is used not only for providing a forestalled service but also used as one of the context that is incorporated with other contexts to boost the accuracy of speech commands to accomplish the task. To validate the proposed method, experiments have been conducted.

1. はじめに

近年，ロボット技術の進歩と共に，生活支援ロボットの研究開発が盛んになりつつある．特に家庭環境において，環境の把握だけではなく，ユーザに適したサービスを提供することのできるロボットが望まれる．一般に，ロボットはユーザの命令に応じて行動する．ユーザの命令は一つの行動に対して様々な形で存在し得るため，ロボットは命令を解釈して適切な行動を決定しなければならない．また音声による命令では，音声認識誤りが生じる可能性がある．こうした状況の中で柔軟に対応できることが，人々の生活を支援するロボットに要求される．

ロボットがユーザの命令を正しく解釈するための手がかりとして，命令を受けた際の文脈が考えられる．例えば，ユーザが普段ソファーでテレビを見ているときに，お菓子を食べながらお茶を飲んでいるということを知っていれば，ユーザが「お菓子を持ってきて」とロボットに命令した際の音声認識に誤りが生じたとしても，そのときにソファーでテレビを見ていてお茶を飲んでいるという文脈を用いれば，ロボットが適切に判断をして正しい行動をとることができるかもしれない．また，ユーザが日々行っている行動をロボットが学習できれば，ユーザからの命令がなくても，ユーザの次の行動をロボットが予測し，適切なサービスの提供が実現できると考えられる．

本研究ではそのようなシナリオを想定し，ロボットがユーザの生活に密着し行動を観測することで，行動パターンを学習することを考える．この学習した行動パターンを，次の行動の予測に利用する．ロボットは、この予測した次の行動を文脈として音声命令や場所などと統合することで最終的なサービス行動を決定する．図 1 に全体的な考え方を示す．本稿で考える重要な問題は，1) ユーザの観測，2) モデル化・予測，3) サービス・行動決定である．1) に関しては，人を常に観測し続けることについての技術的な困難さがある．スマートホームのような環境を考えることもできるが，そうでない環境においてもロボットがなるべく人をセンシングできれば有用であろう．本稿ではこの問題に対して，センシングを考慮した人追跡手法を用いることを考えている．2) は，どのように行動をモデル化・予測するのかということである．これについても様々考えられるが，ここでは人の動きとその際に使う物体の関係性を教師なしで学習し予測することを試みる．従って重要なのは，動作の学習と動作と物体の結びつきの学習である．動作時系列の学習には階層ディリクレ過程隠れマルコフモデル (HDP-HMM) [Beal 01] を，物体と行動の関係性の学習には多層マルチモーダル Latent Dirichlet Allocation (mMLDA) [Fadlil 13] を用いる．また，mMLDA によって記号化された行動を N-gram (行動言語モデル) で表現し予測を行う．さらに，3) については様々考えられるが，ここでは「何かを持ってくる」というサービスを実現する．つまり，ロボットは人の行動パターンから次を予測し，必要となるであろうものを持ってくるというサービスを行うが，人からの「〜持ってきて」といった命令があった場合は，その解釈と予測を統合することでより精度よくサービスが実行できるのではないかと考える．

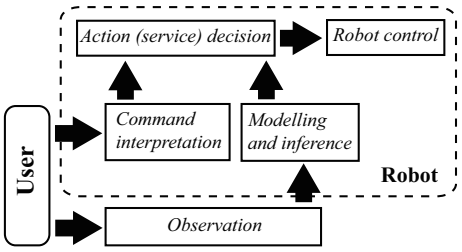

図 1: 提案手法の概要

関連研究として，人の行動学習・認識や，物体と動きの関係性の学習などが挙げられる [Kelly 08]–[Yao 12]. 文献 [Kelly 08, Gehrig 11] において，隠れマルコフモデルや Support Vector Machine (SVM) を用いて人の行動認識及び意図推定を行っている．しかし，人の行動に密室に関係する物体の関係性は考慮されていない．また，教師あり学習に基く動きと物体の関係性の学習が [Koppula 13, Yao 12] において研究されている

連絡先: アッタミミ ムハンマド，電気通信大学 情報理工学研究科 知能機械工学専攻，〒 182-8585 東京都調布市調布ヶ丘 1-5-1，m_att@apple.ee.uec.ac.jp

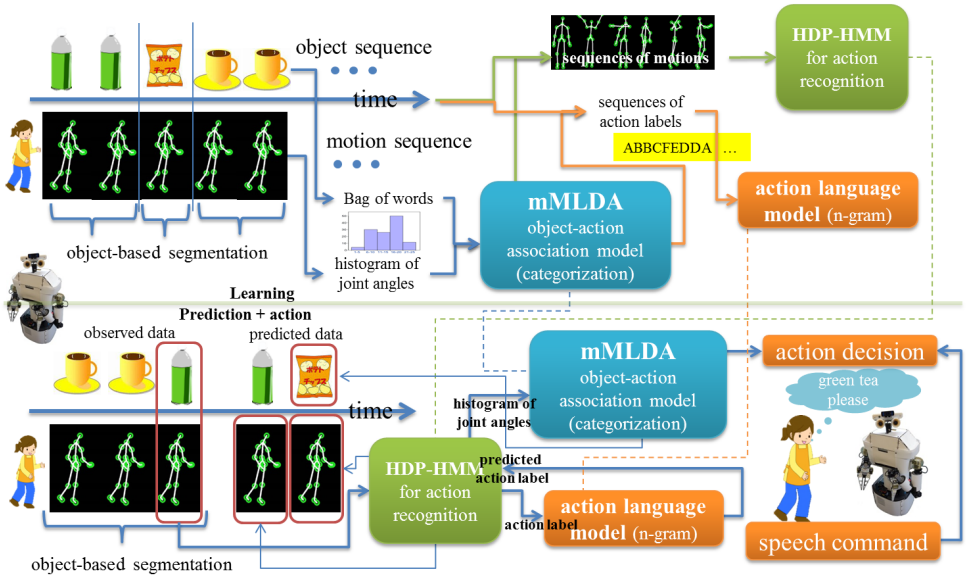

図 2: 提案手法の全体像

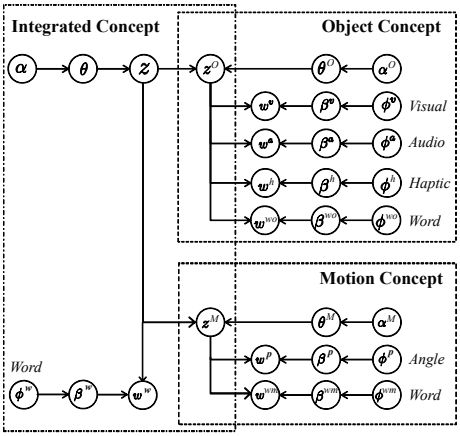

図 3: 多層マルチモーダル LDA のグラフィカルモデル

が，画像内の物体認識精度の向上が主眼となっている．これらに対して本研究では，人の行動を教師なしで学習・認識し，音声命令や場所などの文脈と統合し，適したサービスを決定する問題を扱う．

2. 提案手法

2.1 提案手法の概要

図 2 に提案手法の全体像を示す．本稿では，ロボットは家庭でユーザと共に暮らしていることを想定し，人の音声命令を聞きながらその人の動作と動作を行っている際に関係している物体及び位置を観測する．そして，ユーザの行動パターンを教師なしで学習するのであるが，その際にまず使っている物体を認識しトラッキングすることで，動作の分節化を行う．これは，同じ物体を使っている間が一つの行動としての塊であると仮定し，時系列パターンを区切ることを意味する．そのように区切った動作時系列（関節角の情報）と物体の関係性は，mMLDA [Fadlil 13] によってモデル化（カテゴライズも含む）することができる．つまり現在の行動を認識し，その後に起こる行動を予測できれば，mMLDA によって確率的にユーザが使うであろう物体を予測し，それを持ってくるというサービスを実現可能である．これを “動作－物体関係モデル” と呼ぶ．現在の動作の認識は，分節化された動作の時系列を HDP-HMM を用いてモデル化することで実現する．これを “動作認識モデル” と呼ぶ．また行動の予測については，学習データから行動の N-gram である，“行動言語モデル” を計算することで実現する．こうした動作認識モデルや行動言語モデルは，動作を分節化し，mMLDA を用いたカテゴライズに基づく記号化によって実現されることに注意されたい．

2.2 問題設定

本稿で実現したいサービスは，予測した行動に必要となる物体をユーザに先回りして届けることである．その際，ユーザがどこで動作を行うかという場所文脈や命令された音声文脈などを考慮すればより正確なサービスが行えるであろう．この問題設定は，全ての情報が与えられた際，持ってくるべき物体を推定する問題に置き換えることができる．すなわち，現在の時刻 $t-1$ にロボットが観測したユーザの動き $\boldsymbol{m}^{(t-1)}$，物体 $o^{(t-1)}$，位置 x，音声 S に対して以下の問題を解くことになる．

$$\hat{o}^{(t)} = \underset{o^{(t)}}{\operatorname{argmax}}\, P(o^{(t)} | \boldsymbol{m}^{(t-1)}, o^{(t-1)}, x, S) \tag{1}$$

上式を直接的に計算するのは困難であるため，次のように近似する．

$$\hat{o}^{(t)} = \underset{o^{(t)}}{\operatorname{argmax}}\, P(o^{(t)} | \boldsymbol{m}^{(t-1)}, o^{(t-1)})^{\alpha} P(o^{(t)} | x)^{\beta} P(o^{(t)} | S)^{\gamma} \tag{2}$$

ただし，$P(o^{(t)} | \boldsymbol{m}^{(t-1)}, o^{(t-1)})$, $P(o^{(t)} | x)$, $P(o^{(t)} | S)$ はそれぞれ，行動文脈，場所文脈及び音声命令を表しており，α，β，γ は各文脈に対する重みである．

各文脈の重みの決め方は様々な手法が存在する．例えば，重みのアクティブな学習 [Sugiura 11] を考えることもできるが，ここでは SVM による学習を用いる．具体的には，次節で説明する各文脈 $C \in \{C_1 = \text{行動}, C_2 = \text{場所}, C_3 = \text{音声}\}$ より予測される物体の確率分布 $\boldsymbol{P}^C = (p_1^C, p_2^C, \cdots, p_O^C)$ を一つのヒストグラム $\boldsymbol{h}^C = (\boldsymbol{P}^{C_1}, \boldsymbol{P}^{C_2}, \boldsymbol{P}^{C_3})$ として，SVM の入力デー

タとする．ただし，O は物体カテゴリ数を表す．学習フェーズにおいて，入力データ $\boldsymbol{h}^C$ と正解となる物体カテゴリ o^C の組を用意し，SVM を用いて学習する．認識フェーズでは，与えられた入力データ $\boldsymbol{h}^C_{\rm in}$ に対して，SVM で学習したモデルを用いて認識する．

3. 行動文脈

行動文脈 $P(o^{(t)}|\boldsymbol{m}^{(t-1)}, o^{(t-1)})$ は，現在の行動から次の予測された行動に対する物体の関係性を表す．ここで，現在の時刻 $t-1$ に観測したユーザの動き $\boldsymbol{m}^{(t-1)}$，物体 $o^{(t-1)}$ が与えられた場合，行動文脈を以下のように定義する．

$$P(o^{(t)}|\boldsymbol{m}^{(t-1)}, o^{(t-1)}) = \sum_{a_i^{(t)}, a_j^{(t-1)}} P(o^{(t)}, a_i^{(t)}, a_j^{(t-1)}|\boldsymbol{m}^{(t-1)}, o^{(t-1)}) \quad (3)$$

ただし，$a_i^{(t)}, a_j^{(t-1)} (a_* \in \boldsymbol{A}^{\rm K} = \{a_1, a_2, \cdots, a_K\})$ はそれぞれサイズ K の行動集合 $\boldsymbol{A}^{\rm K}$ に対して，時刻 $t-1$ と時刻 t における行動を表している．また，チェインルールと独立性を用いて，$P(o^{(t)}, a_i^{(t)}, a_j^{(t-1)}|\boldsymbol{m}^{(t-1)}, o^{(t-1)})$ を以下のように書き表すことができる．

$$\begin{aligned} P(o^{(t)}, a_i^{(t)}, a_j^{(t-1)}|\boldsymbol{m}^{(t-1)}, o^{(t-1)}) \propto \\ P(o^{(t)}|a_i^{(t)}) \times P(a_i^{(t)}|a_j^{(t-1)}) \times \\ P(\boldsymbol{m}^{(t-1)}, o^{(t-1)}|a_j^{(t-1)}) \times P(a_j^{(t-1)}) \end{aligned} \quad (4)$$

ただし，$P(o^{(t)}|a_i^{(t)})$ は動作－物体の関係性を表しており，動作－物体関係モデルより算出する．$P(\boldsymbol{m}^{(t-1)}, o^{(t-1)}|a_j^{(t-1)})$ は動作認識の尤度を表し，動作認識モデルを用いて計算する．また，$P(a_i^{(t)}|a_j^{(t-1)})$，$P(a_j^{(t-1)})$ はそれぞれ，動作言語モデルの 2-gram と 1-gram である．

3.1 動作認識モデル

mMLDA によって分類された時系列データを，それぞれ Multimodal HDP-HMM（MHDP-HMM）[中村 13] で学習し行動モデル集合 $\boldsymbol{A}^{\rm K}_M \in \{a^1_M, a^2_M, \cdots, a^K_M\}$ を作成する．ただし，行動モデル集合の各要素 a^*_M は行動集合 $\boldsymbol{A}^{\rm K}$ の a_* と対応している．MHDP-HMM は隠れマルコフモデル (HMM) にディレクレ過程を導入し無限の状態を持つモデルへと拡張した HDP-HMM の各状態から，複数の観測を仮定したモデルである．HDP-HMM の利点としては，状態数を事前に与える必要がない点にある．本稿で用いる観測データは，行動する際のユーザの動き（関節角）$\boldsymbol{m}$ と使っている物体 o であり，それぞれガウス分布と多項分布から生成されるものと仮定する．特徴量としては，関節角とその動的特徴を利用する．

学習した行動モデル集合 $\boldsymbol{A}^{\rm K}_M$ を用いて，入力データ $\boldsymbol{m}_{obs}$，o_{obs} に対する行動 a_i の尤度 $P(\boldsymbol{m}_{obs}, o_{obs}|a_i)$ を計算する．ここで，入力データの時系列は学習データの時系列の途中までであることに注意が必要である．つまり，ユーザの行動が途中であっても次の行動の予測を行う必要がある．

3.2 行動言語モデル

行動のパターンをモデル化するために，N-gram 言語モデルを使用する．これは，mMLDA によって記号化された学習データから，各行動の生起回数を数えることで計算することができる．N をいくつに設定するかは性能や計算量，実際の人の行動がどの程度直前の行動に依存するかによって決まる．本稿では，2-gram 及び 1-gram を行動言語モデルとして用いる．

3.3 動作－物体関係モデル

多層 MLDA (mMLDA) は，下位層に物体と動きの分類モデルであるマルチモーダル LDA (MLDA) を，上位層にそれらを統合するモデル MLDA を配置することによって，物体と動きそれぞれのカテゴライズを行うと同時に，それらの関連性を教師なしで学習する統計モデルである．図 3 に，mMLDA のグラフィカルモデルを示す．このモデルの学習・認識の詳細については文献 [Fadlil 13] を参照されたい．従って，下位層では，何かを飲むような動きといった動作のカテゴリや，ペットボトルという物体のカテゴリが形成され，それらを統合する上位層では，ペットボトルを飲むといった行動が表現される．これは統計モデルであるため，動作の情報を入力することで関連する物体を確率的に予測することができる．本稿では，行動 a_i に対する動作－物体の関係性 $P(o|a_i)$ を，学習した mMLDA を用いて計算する．

4. 場所文脈

本稿における場所文脈はユーザが行動を行う際の位置に対する物体の関係性を表す．ユーザがある場所 $\ell \in \{1, 2, \cdots L\}$ からなる部屋で，ある位置 x において物体 o を使って行動することを想定する場合，場所文脈は以下の式より算出する．

$$P(o|x) = \sum_{\ell} P(x|\ell) P(o|\ell) P(\ell) \quad (5)$$

ただし，$P(\ell)$ は一様分布とし，L は部屋内の場所の数であり，$P(x|\ell)$ は場所 ℓ の尤度であり，2 次元ガウス分布で表現する．また，$P(o|\ell)$ は場所 ℓ において，物体 o が使われる確率である．

5. 音声命令

一方，ユーザは次に必要となるものをロボットに音声で命令することが考えられる．その際ロボットは，これを認識し実行する必要がある．しかし一般的な問題として，音声認識が雑音の多い環境では難しいという点が挙げられる．また，本稿では考慮しないが，命令に曖昧性があり解釈を要する場合も少なくない．こうした状況では，上述の文脈情報，つまり現在のユーザの行動と次の行動の予測が役に立つ．

ここでは，物体を届けるタスクのみを想定しており，ユーザは物体名のみを発話すると仮定する．音声命令の具体的な処理は次の通りである．ユーザの音声 S を認識し，上位 D 個の結果 $\boldsymbol{W}^D$ を用いて以下の式を計算する．

$$P(o = z^o|S) = \sum_{w^{wo} \in \boldsymbol{W}^D} P(z^o|w^{wo}) P(w^{wo}|S) \quad (6)$$

$$P(z^o|w^{wo}) = \frac{1}{P(w^{wo})} \sum_z P(w^{wo}|z^o) P(z^o|z) P(z) \quad (7)$$

$$P(w^{wo}) = \sum_{z^o, z} P(w^{wo}|z^o) P(z^o|z) P(z) \quad (8)$$

ただし，$P(w^{wo}|S)$ は音声 S より抽出された物体名 w^{wo} に対する音声認識尤度である．また，式（8）における詳細な計算式は文献 [Fadlil 13] を参照されたい．

6. 実験

ここでは前節で述べた状況を考慮して，提案手法の基礎的な検証を行う．想定したシナリオとして，ロボットが人を観測

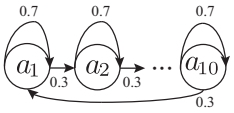

図 4: シミュレーション実験に用いた行動の遷移図

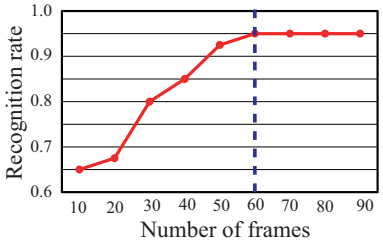

図 5: 観測されたフレーム数に対する動作認識率

し，人の動き，位置及び動作最中に関係する物体情報をキネクトより取得して，ある長さ F フレームのデータを手がかりとして，次の行動に関連する物体を予測する．ただし，ロボットは十分に人を観測し，学習が済んだ段階であるとする．このシナリオを実現するために，擬似データを生成した．まず，文献 [Fadlil 13] に用いたデータセットを用いて動作－物体関係モデルと動作認識モデルを作成した．ここで，行動集合のサイズ K は文献 [Fadlil 13] に従って 10 個とした．また，ユーザの部屋内の場所数を $L = 3$ と設定し，各場所 ℓ に対してガウス分布のパラメータを与えた．次に人の行動の時系列データの生成は，図 4 に示す行動遷移図を用いて行った．図より，各行動 a_* はそれぞれ自己遷移確率 0.7 を与えた．ただし，初期遷移は a_1 からとする．このようなパラメータを用いて，2000 個の行動を生成した．図 4 より生成された各時系列データに対して，$P(o|a)$ を用いて持ってくる物体を生成する．最後に，時系列データを学習用と認識用に分けた．ただし，音声命令において，文献 [Fadlil 13] で与えられたカテゴリ名を物体名とし音声命令を録音した．録音した音声に SNR 100 [dB], 6 [dB], 3 [dB], 0 [dB] の白色雑音をそれぞれ付加した．学習用のデータを用いて，動作言語モデルと場所に対する物体頻度をそれぞれ計算した．また，行動文脈，場所文脈，音声文脈を計算し，それらを一つのヒストグラムとして SVM で学習した．学習したパラメータを用いて，認識用のデータを用いて行動文脈，場所文脈，音声文脈を計算し，学習と同様な方法を用いて認識した．

まず動作認識におけるフレームの長さ F の影響を検討するために，動作認識モデルを用いて認識用のデータで動作認識を行った．図 5 より，F を 60 に設定すれば，95%の認識率が得られた．ロボットの行動決定実験はこの値を用いて行った．その結果を図 6 に示す．図より，全ての SNR に対して平均した結果について，単一の文脈を用いる場合は 70%以下の認識率となることが分かる．一方，SVM を用いて文脈を統合した場合，94.2%まで認識率が向上した．従って，行動文脈や場所文脈など様々な文脈を統合することでよりロバストな行動決定が行えると言える．

7. まとめ

本稿では，行動文脈，場所文脈及び音声命令を統合した，ロボットの行動決定手法を提案した．提案手法ではロボットがユーザの生活に密着し行動を観測することで，行動パターンを

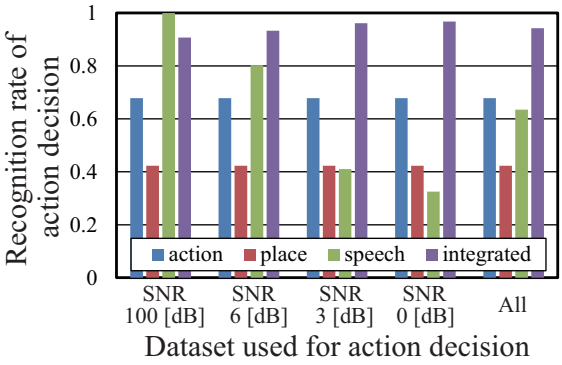

図 6: ロボットの行動決定結果

学習する．学習したユーザの行動パターンを利用して，次の行動を予測し，予測した次の行動を文脈として音声命令，場所情報と統合することで最終的なサービス行動を決定する．本稿では，全ての文脈を確率的枠組みで定義し，SVM を用いて統合する手法を提案した．提案アルゴリズムの基礎的な検討として，ノイズのある環境でロボットの行動決定の実験を行い，その実現可能性を示した．今後の課題として，より現実的なシナリオやフィールドでの学習・行動決定実験が挙げられる．

謝辞

本研究は，科研費（新学術領域研究「予測と意思決定」公募研究 24120526, JSPS 科研費 : 24・9813）の助成を受けたものである．

参考文献

[Beal 01] M.J. Beal, Z. Ghahramani, and C.E. Rasmussen, “The infinite hidden markov model”, Advances in neural information processing systems, pp. 577–584, 2001.

[Fadlil 13] M. Fadlil, M. Attamimi, 長井 隆行, 中村 友昭, 船越 孝太郎, “多層マルチモーダル LDA と相互情報量による語意の獲得”, 日本ロボット学会学術講演会, 2C2-06, Sep. 2013.

[Kelly 08] R. Kelley, M. Nicolescu, A. Tavakkoli, C. King, and G. Bebis, “Understanding human intentions via Hidden Markov Models in autonomous mobile robots”, ACM/IEEE Int. Conf. in HRI, pp. 367–374, March 2008.

[Gehrig 11] D. Gehrig, P. Krauthausen, L. Rybok, H. Kuehne, U. D. Hanebeck, T. Schultz, and R. Stiefelhagen, “Combined intention, activity, and motion recognition for a humanoid household robot”, IEEE Int. Conf. on IROS, pp. 4819–4825, Sep. 2011.

[Koppula 13] H. Koppula, R. Gupta, and Ashutosh Saxena, “Learning Human Activities and Object Affordances from RGB-D Videos”, Int. Journal of Robotics Research, Vol. 32, No. 8, pp. 951–970, 2013.

[Yao 12] B. Yao, and L. Fei-Fei, “Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses”, IEEE Trans. on PAMI, vol.34, pp.1691–1703, Sep. 2012.

[中村 13] 中村 友昭, 船越 孝太郎, 長井 隆行, “HDP-HMM を用いたロボットによる物体軌道の学習と予測”, 日本ロボット学会学術講演会, 2C1-05, Sep. 2013.

[Sugiura 11] K. Sugiura, N. Iwahashi, H. Kawai, and S. Nakamura, “Situated spoken dialogue with robots using active learning”, Advanced Robotics, Vol. 25, No. 17, pp. 2207–2232, 2011