3D1-04

HMM による協調動作の模倣学習

Imitation of Cooperative Behavior by Hidden Markov Model

野田五十樹[*1*2]
NODA, Itsuki

[*1]産業技術総合研究所サイバーアシスト研究センター
Cyber Assist Research Center, AIST

[*2]科学技術振興事業団 さきがけ研究２１
PRESTO, JST

This paper addresses agents' intentions as building blocks of imitation learning that abstract local situations of the agent, and proposes a hierarchical hidden Markov model (HMM) to represent cooperative behaviors of teamworks. The key of the proposed model is introduction of gate probabilities that restrict transition among agents' intentions according to others' intentions. Using these probabilities, the framework can control transitions flexibly among basic behaviors in a cooperative behavior.

1. はじめに

模倣学習は、人間やエージェントの複雑な振る舞いを獲得し、さらに強化学習などの機械学習手法の初期状態を与える方法として注目されており、これまで主に単一のエージェントの振る舞いを中心に研究されてきた [4, 8, 5]。一方、状態の多様さや相互作用の扱いの難しさなどから、複数のエージェントの協調作業などの模倣はあまり取り上げられてこなかった。本研究ではマルチエージェントシステムの模倣学習におけるこれらの問題を解決するため、エージェントの**意図**を模倣動作の基本単位とし、隠れマルコフモデル (HMM) により意図の基づく動作および意図の変化を表現し、チームワークなどの協調動作の模倣を実現する枠組みを提案する。

2. チームワークと意図

チームワークのもとに人間が協調的に振る舞っている時、相互の意図が重要な役割を果たしている。たとえばサッカーにおいて二人のプレーヤーがパスを行っている場合、パサー (パスを出すプレーヤー) はレシーバー (パスを受けるプレーヤー) がパスを受けられる状態にあることを知っており、レシーバーは、パサーがパスを出す先を求めていることを知っている必要がある。このように、パスというチームワークをきめるためには、それに参加するプレーヤー同士がお互いの意図とそのタイミングを知っている必要がある。以下、この節では、この共有しているお互いの意図に着目し、意図をチームワークを構成する基本単位として形式化する。

2.1 意図とプレー

ここでは、**意図**を比較的短期間に、一人のエージェントがある条件下である目標状態を達成しようとする考えとみなす。たとえばサッカーの場合、「ボールをある方向にドリブルで運ぶ」という意図は、プレーヤーがボールとともにある方向に位置を移すという状態を達成する考えを表す。なお、ここでは**意図**は個々のエージェントに閉じているものとし、他のエージェントの行動などは達成する状態には反映されないものとする。

A: 野田五十樹、産業技術総合研究所サイバーアシスト研究センター，東京都江東区青海 2-41-6、Tel:03-3599-8230、Fax:03-5530-2067、i.noda@aist.go.jp

ある特定の**意図** を達成するために、エージェントはそれに対応する**プレー**を行うものとする。この**プレー** は一連の基本動作の系列からなり、エージェントの全振る舞いを構成する基本単位となる。たとえばサッカーの場合、'ドリブル' は「ボールをある方向にドリブルで運ぶ」という**意図**を達成するための**プレー**であり、それは 'turn'、'dash'、'kick' などの基本動作の列として表される。**意図**と同様、**プレー**も他のプレーヤーの行動は考慮しない、単一のエージェントで完結したものとする。また、ここでは**意図**と**プレー** は 1 対 1 対応しているものとする。

意図は全振る舞いの構成要素であるとともに、環境の状態を抽象化、簡単化する基本単位としても扱われる。すなわち、あるエージェントがある**意図**を持っていることを、エージェントの環境が (チームプレーを同期させるための) ある条件を満たしている状態と見なす。たとえば、あるプレーヤーが 'ドリブル' をする意図を持っている場合、これは、そのプレーヤーがボールを持っており、ドリブルをしているあるスペースに向かおうとしており、さらに、そのプレーヤーは (その方向に近い位置でパスを受けるなど) 何らかのサポートを必要としている、という条件を表していると見なす。以下では、この**意図** による環境条件の抽象化をチームプレーの同期のトリガとして用いる。

2.2 チームプレー

チームプレー は複数のエージェントによって行われる一連の **プレー** の集まりとして表される。前節で述べているように、意図およびプレーはエージェント単体で閉じたものとして扱われる。よって、エージェントはある意図を持ち続けている間は他のエージェントと相互作用することはない。そのかわり、エージェントが意図を変化させる (転意) 時点で他のエージェントとの相互作用が行われ、チームプレーが成立すると考える。たとえばサッカーで二人のプレーヤーがドリブルからパス交換をするというチームプレーを考えた場合、各プレーヤーの意図 (プレー) は以下のように変化すると考える。

パサー	レシーバー
ドリブル	:
↓	サポート
レシーバー確認	↓
↓	フリーラン
パス	↓
:	トラップ

この例においてプレーの同期は、各々のプレーヤーが転意する

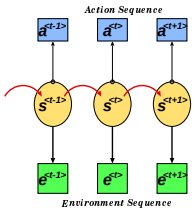

図 1:Single Behavior Model

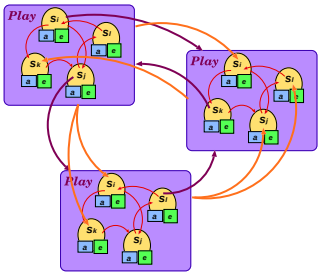

図 2:Cooperative Behavior Model

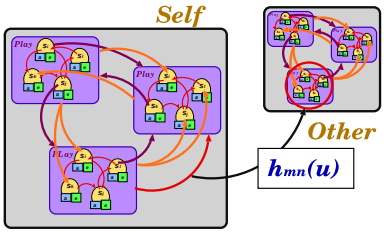

図 3: Joint Behavior Model

時点を相手の意図が制御すると考える。すなわち、パサーが「レシーバー確認」から「パス」に転意させるためには、相手が「フリーラン」を行っていなければならないと考える。このように、各々のエージェントの意図が別のエージェントの転意のタイミングを制御する形でチームワークが成立するものとみなす。

3. チームプレーの階層的隠れマルコフモデル

3.1 単純行動モデル

あるエージェントが行うプレー m は、Moore 型の HMM $\mathbf{\Pi}_m = \langle \boldsymbol{S}_m, \boldsymbol{V}_m, \boldsymbol{P}_m, \boldsymbol{Q}_m, \boldsymbol{R}_m \rangle$, により表現される [7, 6]。ただし、$\boldsymbol{S}_m$ はプレー m の状態集合、$\boldsymbol{V}_m$ は状態の出力 (センサーおよび基本行動の対) 集合、$\boldsymbol{P}_m = \{p_{mij} | i, j \in \boldsymbol{S}_m\}$、$\boldsymbol{Q}_m = \{q_{mi}(v) | i \in \boldsymbol{S}_m, v \in \boldsymbol{V}_m\}$ および $\boldsymbol{R}_m = \{r_{mi} | i \in \boldsymbol{S}_m\}$ は各々、状態遷移、状態出力、初期状態確率の行列を表す。

3.2 プレー間遷移のモデル

2.2 節で述べたように、ここではチームプレーは複数のエージェントの意図の変化、すなわち、前節で定義したプレー間の転意として定義される。これを表現するために、まず、エージェント一人の転意を表現するために、Mealy 型の HMM $\mathbf{\Theta} = \langle \boldsymbol{M}, \boldsymbol{U}, \boldsymbol{E}, \boldsymbol{F}, \boldsymbol{G}, \boldsymbol{H} \rangle$, を導入する。ただし、$\boldsymbol{M} = \{\mathbf{\Pi}_m\}$ はプレーの集合、$\boldsymbol{U}$ は出力集合 (通常、$\boldsymbol{M}$ と同じ)、$\boldsymbol{E} = \{e_m | m \in \boldsymbol{M}\}$ は初期プレーの確率分布、$\boldsymbol{F} = \{f_{min} | m \in \boldsymbol{M}, i \in \boldsymbol{S}_m, n \in \boldsymbol{M}\}$ はプレーからの転出確率、$\boldsymbol{G} = \{g_{mnj} | m \in \boldsymbol{M}, n \in \boldsymbol{M}, j \in \boldsymbol{S}_n\}$ はプレーへの転入確率、$\boldsymbol{H} = \{h_{mn}(u) | m \in \boldsymbol{M}, n \in \boldsymbol{M}, u \in \boldsymbol{U}\}$ はプレー間の転意確率である。これらの確率により、実際にプレー m の状態 i からプレー n の状態 j へ遷移する確率は次のように計算される。

$$p'_{minj} = Pr(s_{nj}^{\langle t \rangle} | s_{mi}^{\langle t-1 \rangle}) = \begin{cases} f_{mim} p_{mij} & ; m = n \\ f_{min} g_{mnj} & ; m \neq n \end{cases}$$

3.3 協調行動のモデル

最後に、上で導入した HMM $\mathbf{\Theta}$ を複数カップリングすることで、複数のエージェントからなる協調行動を表現する。カップリングは転意確率 $\boldsymbol{H}$ を介して行われる (図 3)。たとえば、エージェント X とエージェント Y がお互いに協調している場合、エージェント X の$\mathbf{\Theta}$ の $h_{mn}(u)$ は、エージェント X が m から n へプレーを転化させる時に、エージェント Y がとっているプレーが u である確率を示すものとする。この確率を用いて、プレーの再生時にプレー間の転意がどうかの尤度を求めることができる。

3.4 学習法

上で提案したモデルは通常の HMM であるため、学習方法としては Baum-Welch のアルゴリズムを用いることができる。HMM $\mathbf{\Theta}$ の具体的な学習法は以下のようになる。

$$e_m \leftarrow \sum_i \gamma_{mi}^{\langle 0 \rangle} \qquad g_{mnj} \leftarrow \frac{\sum_t \xi_{mnj}^{\langle t \rangle}}{\sum_t \gamma_{mn}^{\langle t \rangle}}$$

$$f_{min} \leftarrow \frac{\sum_t \xi_{min}^{\langle t \rangle}}{\sum_t \gamma_{mi}^{\langle t \rangle}} \qquad h_{mn}(u) \leftarrow \frac{\sum_{t, u^{\langle t \rangle} = u} \gamma^{\langle t \rangle}}{\sum_t \gamma_{mn}^{\langle t \rangle}},$$

$$\begin{aligned} \xi_{min}^{\langle t \rangle} &= \alpha_{mi}^{\langle t-1 \rangle} f_{min} h_{mn}(u^{\langle t-1 \rangle}) \bar{\beta}_{mn}^{\langle t \rangle} \\ \xi_{mnj}^{\langle t \rangle} &= \bar{\alpha}_{mn}^{\langle t-1 \rangle} g_{mnj} q_{nj}(v^{\langle t \rangle}) \beta_{nj}^{\langle t \rangle} \\ \gamma_{mi}^{\langle t \rangle} &= \alpha_{mi}^{\langle t \rangle} \beta_{mi}^{\langle t \rangle} \\ \gamma_{mn}^{\langle t \rangle} &= \bar{\alpha}_{mn}^{\langle t \rangle} h_{mn}(u^{\langle t \rangle}) \bar{\beta}_{mn}^{\langle t+1 \rangle}. \end{aligned}$$

であり、$\alpha_{mi}^{\langle t \rangle}$ および $\beta_{mi}^{\langle t \rangle}$ は各時点における各状態の前方および後方確率である。

3.5 模倣学習の手順

上記で定義した HMM を用いて、以下のような手順により模倣学習を行う: [L-1] 各々の基本プレーに関して、HMM $\mathbf{\Pi}$ を学習により構成する。[L-2] 学習によって獲得された $\mathbf{\Pi}$ を用いて、HMM $\mathbf{\Theta}$ を構成する。ただし $\mathbf{\Theta}$ の各確率はランダムに設定する。[L-3] 模範演技者達の行動と環境の変化を観察し、Baum-Welch の方法にしたがって前方、後方尤度 ($\alpha_{nj}^{\langle t \rangle}$、$\beta_{mi}^{\langle t \rangle}$) を求める。[L-4] $\mathbf{\Theta}$ の初期、転出、転入確率 ($\boldsymbol{E}$、$\boldsymbol{F}$、$\boldsymbol{G}$) を求める。[L-5] 上記の 3 から 4 のステップを繰り返す (**観察フェーズ**)。[L-6] 最終的な前方、後方尤度から転意確率 ($\boldsymbol{H}$) を求める (**抽出フェーズ**)。[L-7] 求められた確率に基づき、基本プレー間および基本プレー内の遷移を求め、行動を決定する (**生成フェーズ**)。この最後の生成フェーズにおいて、行動の決定は以下のように行われる。時刻 t においてエージェントが基本プレー m の状態 i にあり、他のプレーヤーが基本プレー $u^{\langle t \rangle}$ を行っている場合、次の時刻 $t+1$ における基本プレー n およびその状態 j は以下の尤度 L に従って決定される。

$$L(s_{nj}^{\langle t+1 \rangle}) = \begin{cases} f_{mim} p_{mij} q_{mj}(\hat{v}^{\langle t+1 \rangle}) & ; m = n \\ f_{min} g_{mnj} h_{mn}(u^{\langle t \rangle}) q_{nj}(\hat{v}^{\langle t+1 \rangle}) & ; m \neq n \end{cases} \quad (1)$$

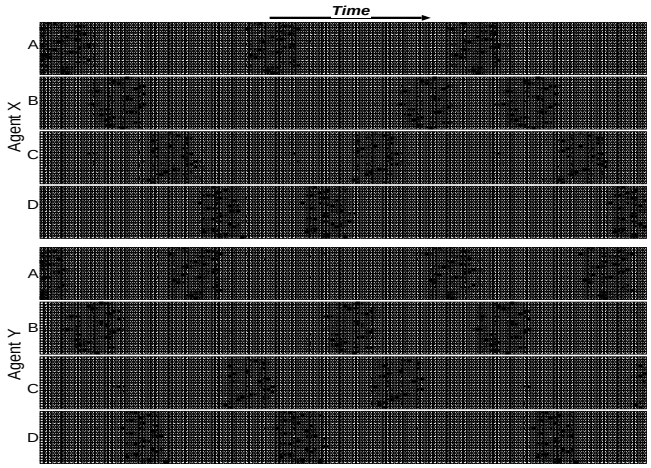

図 5: Exp. 1: Result of Recognition of Mentor's Behaviors

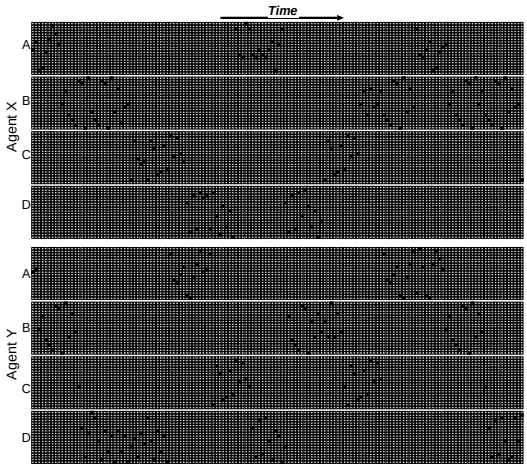

図 6: Exp. 1: State Transitions Generated by Learned HMM

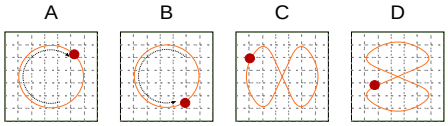

図 4: Basic Plays used in Exp. 1

4. 実験

4.1 Exp.1: 平面上の周期的行動の同期

前節までに提案してきた手法の有効性を検証するため、平面上のエージェントの周期的行動を用いた実験を行った。実験では、2 つのエージェント (Agent-X および Agent-Y) が互いに同期しながら、平面上の周期的な 4 種類の行動を転意させていく。4 種類の行動は図 4 に示すように、円軌道を時計回り (A)、反時計回り (B)、横 8 の字運動 (C)、縦 8 の字運動 (D) である。エージェントの行動は平面上の位置として観測される*1。2 人の模範演技者は以下に示す順番で交互に基本プレーを転意させるものとする。

Agent X	→ A → B → C → D → A → D → C → B → A
Agent Y	A → B → D → A → C → D → B → C → A →

また、模倣学習の手順 1 として、基本プレーは各々、20 状態の HMM $\mathbf{\Pi}$ により学習されているものとする。

図 5 は **観察フェーズ** によって求められた各々のエージェントの各時点における基本プレーおよびその状態の相対尤度を示している。この図の構成は以下の通りである。図は上半分が Agent-X、下半分が Agent-Y に対応し、各々が多数の小さい正方形からなる 4 つの横方向の帯から構成される。これらの帯は各々基本プレー (A∼D) に相当し、その縦方向の一列には 20 個の正方形が並んでいる。これらの正方形が各々基本プレーの状態に対応しており、黒い部分の面積がその尤度の相対的な大きさを示している。図の右方向が時間軸となっており、縦 1 列がある時刻 t の各状態の尤度を表していることになる。

この図から、学習者は両エージェントの基本プレーの転意に関して、正しい順でプレーが転意すると推定していることがわかる。また、転意のタイミングに関しても、模範演技と同じタイミングで行っていると推定できていることがわかる。

*1 実際の実験では、平面は 7 × 7 の 49 区画に分割され、どの区画にいるかが観測される。

このように獲得された転意確率を用いてエージェントの行動を生成すると、図 6 のようになる。この図は図 5 と同様の構成をとっているが、各々のエージェントについて縦一列あたり 1 つの正方形のみ塗り潰されている。これは、各時刻 t において転意確率および遷移確率に従って各エージェントの状態を決定しているためである。この図に示されているように、学習の結果再生されたエージェントの行動は、細部は異なるものの、基本プレーの単位では正しく順番および転意のタイミングが再現されている。

また、模範演技では、相手の意図によりおなじ意図から異なる転意を行うようになっている。たとえば agent X はプレー A から 2 種類の転意、すなわち A → B と A → D の可能性があり、各々の転意は agent Y の意図によって条件付け、すなわち、agent Y がプレー B を行っている場合は A → B、プレー D の場合は A → D という転意になっている。このような場合においても図 6 に示すように、提案手法では正しい転移パターンを獲得している。実際、獲得された agent X の転意確率では、$h_{AB}(B) = 0.97$、$h_{AD}(B) = 0.00$ というように、模範演技に準じた条件付けが獲得されていることがわかる。

4.2 Exp. 2: サッカーにおける協調プレー

次に、このフレームワークを簡単なサッカーの協調プレーに適用した実験を示す。模範演技者は図 7 に示すようなドリブルおよびパスによるチームプレーを行う。このなかで、まず第 1 プレーヤー (agent X) はフィールドの左側より右側に向かってドリブルをはじめ、第 2 プレーヤー (agent Y) はその横をサポートの形で平行して走る。agent X はパスを出す直前にスローダウンし、パスの受け手 (agent Y) を確認したあと、パスを出す。agent Y はこのタイミングでフリーランを開始し、agent X は agent Y が向かっているオープンスペースへパスを出す。その後、二人のプレーヤーは役割を交代し、プレーを続ける。

このチームプレーを模倣するため、学習者はあらかじめ 'ドリブル'、'ルックアップ'、'パス'、'フリーラン'、'ボールチェイス'、'サポート' という基本プレーに対応する 6 つの HMM $\mathbf{\Pi}$ の学習を行う。ただし、$\mathbf{\Pi}$ の各状態出力集合 $\boldsymbol{V}_m$ は相対的なボールの位置と速度、およびプレーヤーの行った基本動作 ('turn', 'dash', 'small-kick', 'long-kick', 'trap', 'look') であり、各$\mathbf{\Pi}$ の状態数は 5 とした。これらの $\mathbf{\Pi}$ をもとに agent X および Y 用の HMM $\mathbf{\Theta}$ を構成し、模範演技者のプレーを 1 つ観察させ、学習を行う。

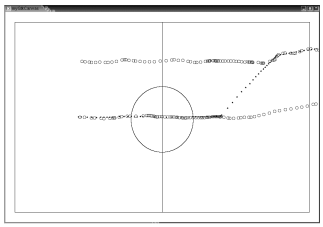

図 7: Exp. 2: Dribble and Pass Play by a Mentor

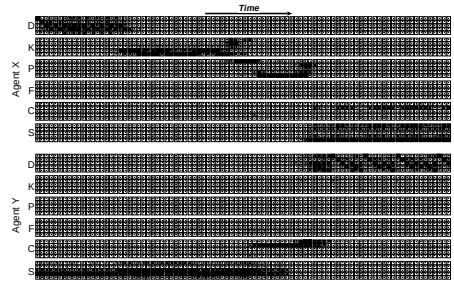

図 8: Exp. 2: Result of Recognition of a Mentor's Behaviors

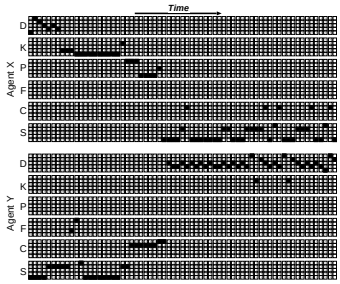

図 9: Exp. 2: State Transitions Generated by Learned HMM

図 7 に観察による状態推定の結果を示す。この図のなかで各 $\mathbf{\Pi}$ に対応する帯につけられている文字 (D, K, P, F, C, S) は各々、'ドリブル (D)'、 'ルックアップ (K)'、 'パス (P)'、 'フリーラン (F)'、 'ボールチェイス (C)'、 'サポート (S)' に対応する。この図から、学習者は agent X、 Y が各々、'ドリブル' →'ルックアップ' →'パス' →'サポート' および 'サポート' →'ボールチェイス →'ドリブル' を行ったと推定している。これにより得られた転意確率に基づきこの協調プレーの再生を行うと、図 9 に示したような状態遷移が生成される。この例では必ずしも模範演技者と同じ状態遷移が行われていないところがいくつか見受けられる。例えば agent X がルックアップの最中に、agent Y はたびたびフリーランを試みている。これは、抽出フェーズにおいて得られる転意の規則が確率的なものであり、かつ、サポートとフリーランやボールチェイスが類似の行動であることから、比較的曖昧な転意規則が獲得されていることを示している。この曖昧さは間違った転意にもつながるが、逆に学習の汎化能力を示していると言え、今後検討が必要である。

5. 関連研究と議論

HMM のカップリングの手法については、マルチエージェントの協調の確率的表現としていくつかの研究が行われてきている [3, 1, 2]。これらの研究では各エージェントが一つの HMM を構成し、それらの状態遷移を相互作用による条件付き確率として表す。このため、調整すべき確率パラメータの数が膨大なものとなり、数少ない例しか与えられない模倣学習などでは大きな問題となる。それに対し我々の手法では、相互作用は転意確率 $\boldsymbol{H}$ に集約されて表現され、また、他のエージェントの状態については基本プレー、すなわち意図として抽象化されるため、比較的少ない数の例からでも適切な学習が行えることが特徴となっている。

意図による抽象化はエージェント間のコミュニケーションの利用にも示唆を与える。ここで提案したモデルではエージェント間の直接的コミュニケーションはないものとしてきた。しかし、獲得したモデルを用いてプレーを再生する場合、HMM $\mathbf{\Theta}$ で他のエージェントの内部状態をシミュレーションすることで、自分あるいは相手の意図に対する相対尤度にめりはりがなくなる状態を検出することができる。これはエージェント間で相互の意図が十分に共有されていないことを示しており、このような場合に意図を交換するコミュニケーションを行うことは有効であると考えられる。

また、提案した方法における問題として、意図の設計がある。本稿で述べてきたように提案手法では意図が、行動の抽象化および他のエージェントを中心とした環境の部分的状況を代表するものとして、重要な役割を担っている。このため何を意図として採用するかが提案手法の能力をきめることになり、その設計をどのように行うかを検討する必要がある。

参考文献

[1] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden markov models. *Machine Learning*, 29:245–275, 1997.

[2] Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[3] Michael I. Jordan, Zoubin Ghahramani, and Lawrence K. Saul. Hidden markov decision trees. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 501. The MIT Press, 1997.

[4] Y. Kuniyoshi and H. Inoue. Qualitative recognition of ongoing human action sequences. In *Proc. IJCAI93*, pages 1600–1609, 1993.

[5] Hiroyuki Miyamoto and Mitsuo Kawato. A tennis serve and upswing learning robot based on bi-directional theory. *Neural Networks*, 11:1331–1344, 1998.

[6] Itsuki NODA. Hidden markov modeling of multi-agent systems and its learning method. In *Proc. of RoboCup 2002 International Symposium*, 2002.

[7] Itsuki NODA. Segmentation of environments using hidden markov modeling of other agents. In *Proc. of AAMAS-2002*, pages 1395–1396, July 2002.

[8] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3(6):233–242, Jun. 1999.