コンピュータビジョンと
イメージメディア 121－16
（2000．3．23）

9眼ステレオとデータグローブを用いた人間行動の認識

小川原光一　射場総司　木村浩　池内克史

東京大学生産技術研究所 第3部
〒 106-8558 東京都港区六本木 7-22-1 Tel：03-3401-1433
E-mail：{ogawara,ki}@iis.u-tokyo.ac.jp, iba+@cmu.edu, hiroshi@kimura.is.uec.ac.jp

あらましロボットが人間の行動を獲得しそれを再現するためには、適用範囲が広く汎用性のある形で人間行動をモデル化できる必要がある。本稿では人間の手作業を対象とし、データグローブと HMM を用いて、人間の手の動作の「手・把握の状態・動き・3次元空間上の位置関係」の4属性に着目して抽象化を行い、抽象化された動作を時系列に沿って離散的に並べることで人間行動をモデル化する手法について述べる。また、9眼ステレオビジョンと3次元テンプレートマッチング法を用いて、生成された行動モデルからロボットによる人間の手作業の再現を行う手法について述べ、実機により検証した結果について報告する。

Recognition of Human Behaviour with 9eye Stereo Vision and Data Glove

Koichi OGAWARA　Soshi IBA　Hiroshi KIMURA　Katsushi IKEUCHI

Institute of Industrial Science
University of Tokyo
7-22-1 Roppongi, Minato-ku, Tokyo JAPAN 106-8558 Tel：03-3401-1433
E-mail：{ogawara,ki}@iis.u-tokyo.ac.jp, iba+@cmu.edu, hiroshi@kimura.is.uec.ac.jp

Abstract When a robot recognizes human behaviour and tries to perform the same behaviour afterward, a human behaviour model which is highly applicable in different environment is required. In this paper, we limits the task field to hand operation and describe a novel method of constructing a behaviour model of hand operation using data-glove and HMM, which abstracts human hand action from four attributes "which hands", "state of grasping", "hand motion" and "relative position of target objects and hands in 3D space" along the time series. We also describe a method of performing the same human operation by robot with 9eye stereo vision system and 3D template matching, and finally we show the experimental results using a real robot.

1 はじめに

ロボットが人間の行動を観察することで、自動的に人間の動作を習得し、ロボット自身の行動のレパートリーを増やしていくことができれば、人間の生活する社会において、ロボットの適用範囲を飛躍的に拡大させることが可能になる。

このようなロボットを実現する上で重要なことの一つは、人間の動作を認識する際に、いかに汎用性のある形で行動をモデル化できるかにある。人間が動作を行った時と全く同じ環境でないとロボットが動作を再現できないのでは意味がなく、様々な状況で再利用できる形で行動を獲得しておくことが重要となる。

このように獲得された行動は一種の技能と呼ぶこともでき、人間の行う作業のロボットによる補助や、ひいては専門家の持つ技のロボットによる継承につなげることができると考えている。

本稿では、本研究の位置付けについて簡単に述べた後、データグローブと9眼のマルチベースラインステレオを使用したロボットによる人間行動の認識・モデル化及び再現手法について述べ、新たに開発した人間型ロボットを用いて手法の検証を行った結果について報告する。

2 人間行動認識システム

2.1 研究の位置付け

従来、ロボットによる人間動作の認識・獲得というテーマは、視覚処理なら視覚処理、行動計画なら行動計画、動作なら動作と分野毎に個別に研究が行われていることが多く、その場合個々の分野では入力にある仮定を置いて研究を行うため、それらの成果を全部繋げた時に、本当に実世界において機能するものが出来るのであろうかという疑問があった。

その反省から、認識から実現に至るまで常に実世界との相互作用を考慮した研究が重要視されてきている。

これまでに、人間の行う作業を認識し、獲得した動作をロボットに再現させる手法として、Assembly Plan from Observation system(APO) [1] が提案されている。APO の枠組みでは、ロボットが視覚によって人間の行う作業を観察し、扱われる物と物との接触状態の遷移に着目して時系列に沿った接触状態の変化を解析することで動作の認識を行う。そして、ロボットの動作実行時において、解析結果と実世界からの情報の同期を取ることで、ロボットによる動作の再現を実現している。

また、ロボットの視覚を用いて人間の作業手順を認識し、作業中の各動作単位の実行時における依存関係を解析して作業モデルを自動生成する研究がある [2]。この研究では、生成された作業モデルによって、ロボットと人間の柔軟な協調作業が実現されている。

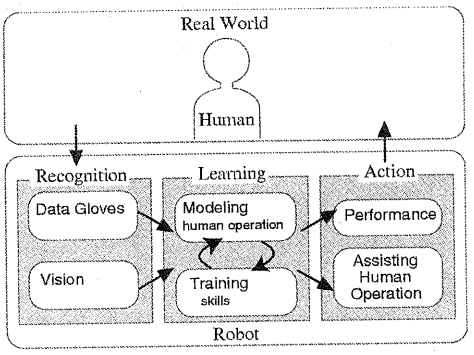


図 1: 人間行動認識システム

上述の研究の流れを踏まえ、本研究では図1に示すように、実世界との相互作用を含んだループの中でロボットによる人間行動の認識・学習・動作の実現を図っていく。本稿は、その枠組みのうち、ロボットによる人間行動の認識と再現について研究を行った結果についての報告である。

人間行動の認識と再現においては、いかに人間の動作を汎用性のある形でモデル化できるかが課題であり、次節以降では、人間の行動のうち手作業にタスクを限定し、そのモデル化手法及び実現方法について述べる。

2.2 人間の手作業のモデル化

従来の主なアプローチは、人間の行った動作を、基本的にその時と全く同じ環境でロボットに行わせようというものであった。

本研究では、以下に示すように人間動作の観察時に得られる生データを抽象化することで、人間が行った時とロボットが再現する時との環境の差異を吸収し、獲得した人間動作の適用範囲を拡大することを試みた。

- **手の動作の抽象化**

人間の手の使い方を分類分けし、人間の行う一連の手作業を、分類分けされた動作（シンボル）の組み合わせとして表現する。

本研究では、手の取りうる動作を、「右手・左手・両手同時」「握力把握・精密把握[3]」、及びその状態における「手の位置・姿勢の変化」の3属性から作られるシンボルとして分類分けし、手の動作の抽象化を行う。

- **位置関係の抽象化**

  操作対象物や手の正確な位置情報のみを動作モデルに組み込むのではなく、3次元空間における対象物と両手の位置関係（前後・左右・上下）をもモデルに組み込むことで、ロボット動作時において物体の配置が異なる状況への対応が可能となる。

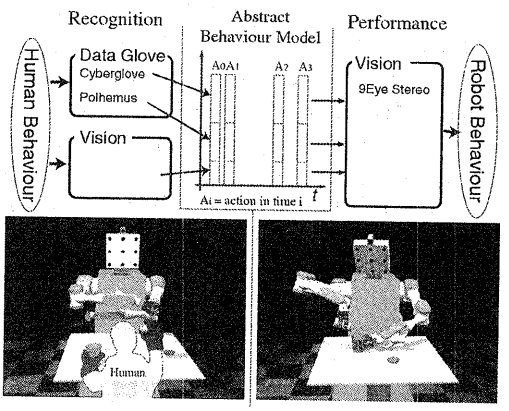


図 2: 人間行動の認識と再現

2.3 人間行動の認識と再現

手の動作の抽象化として、3章で述べるようにデータグローブを用いて、Hidden Markov Model(HMM) によって動作の認識及び動作シンボルの生成を行う。

また位置関係の抽象化のために、データグローブの手首に搭載されている Polhemus 位置センサの情報を用い、各動作の開始・終了時点における手の絶対位置情報と共に、もう一方の手や時間的に一回前の手の位置との相対位置関係を記録する。

人間行動認識に視覚を用いることは、現段階では速度や精度の面で未実装である。

これらの情報を、両手いずれかの動作シンボルが切り替わる時点でセグメンテーションを行い、図2に示すように時間方向に対して離散的なシンボル列 (A_i) の形に変換し、人間行動モデルとする。

この人間行動モデルの各動作 (A_i) を、時系列に沿って順番にロボットが行うことによって、人間行動の再現を行う。

ロボットによる再現時には、4章で述べるように9眼のマルチベースラインステレオビジョンシステムを用い、まず環境内の物体の認識を行う。抽象化された行動モデルと認識された物体の対応を取ることで、ロボットは人間が作業を行った環境と物体の配置や数が異なる環境においても、正確に人間の行動を再現することが可能となる。

3 HMM を用いた人間行動のモデル化

人間とロボットの協調作業を行う上で、ジェスチャ認識のリアルタイム性とスポッティングが重要となる。本研究では、人間の手を使った動作を握力把握 (Power Grasp)・精密把握 (Precision Grasp) 及びその時の手の動きで分類し、HMM でモデル化してスポッティング認識を行った。右手と左手でのジェスチャは、それぞれ別の入力プロセスとして扱い、両手を用いるジェスチャは左右スポッティング認識の結果を合わせることで行動モデルの抽出を行っている。モデル化の対象となるジェスチャは「Power Grasp (握る)」「Precision Grasp (摘む)」「Release (離す)」「Pour (注ぐ)」「Hand-over (差し出す)」「OK sign」の六種である。本研究ではこれら六種のジェスチャを基本要素に分解し、HMM によるモデル化を行なった。

この章では HMM の概要、スポッティング認識、そして Hidden Markov Model Toolkit(HTK) と両手のデータグローブ (CyberGlove) を用いた実装レベルでのシステムの説明を行う。

3.1 HMM

HMM は不確定な時系列のデータをモデル化するために有効な統計的手法であり、音声認識における重要な基礎技術の一つとなっている [4]。HMM の特徴は、単純マルコフモデルの状態間の遷移と状態における出力がどちらも確率的であり、状態が出

力系列から一意に決まらないところにある。これが「隠れ」マルコフモデルと呼ばれる所以である。出力系列に含まれるノイズを確率的に処理出来る上、モデルのシンボルと時系列データを結びつける各状態の理論的展開が容易な事から、音声のみならずジェスチャ認識にも使われる事が多い [5, 6]。

HMM では、モデルのパラメータ推定法 (Baum-Welch)、最適状態偏移系列の算出法 (Viterbi)、モデルが出力系列を出力する確率を求める方法 (forward-backward) などの基本アルゴリズムが提案されており、これらは HMM に基ずく基本問題の三つの解として知られている。

3.2 ジェスチャのスポッティング認識

ジェスチャのスポッティング認識とは連続した動作の中から時間的位置が未知のジェスチャの判別と、時系列上のセグメンテーションを同時に行うものである。これにより、ジェスチャを動作の開始と終了を意識することなく認識する事が可能となる。

スポッティング認識には連続 DP を用いたもの、HMM を用いたもの、ニューラルネットを用いたもの等、様々な手法が提案されている。連続 DP は始点固定で終点フリーの DP マッチングを連続して行う方法で、計算量が HMM に比べて極めて大きくなる為、さまざまな工夫が提唱されている [6]。標準的な HMM によるスポッティング認識は、キーワードとなる HMM と Filler (Garbage) HMM を並列に繋げ、対象外の動きを Filler HMM 内の状態に落とし込む方法がとられる [7]。音声の場合、Filler HMM を全音素の HMM から用意する事が可能だが、ジェスチャの場合、明確な要素となる動作の定義は無い。その為、アドホックに作成するか、閾値を出す為の閾値モデルを作るなどの工夫が必要となる [8]。本研究ではスポッティング対象の語彙数が少ない為、Filler HMM を独自に作成した。

3.3 データグローブと位置センサを入力とする HMM によるスポッテング認識

本研究では左右のデータグローブ (CyberGlove) と三次元位置センサ (Polhemus) を入力として、HMM を用いてスポッティング認識を行った。なお、システムの一部は、Hidden Markov Model Toolkit(HTK)[9] に基づいている。

出力データとして、片手につき 48 次元の特徴点を使用している。CyberGlove から得られる 18 次元の手指の曲げ $\{r_1 ... r_{18}\}_t$、Polhemus から得られる 6 次元の手の姿勢 $P_t = \{x, y, z, \alpha, \beta, \gamma\}_t$、それらに速度情報を加えて、計 48 次元となっている。Polhmus からの位置情報 P_t は、作業者の向きや位置に依存しない認識を可能とする為、常に 1 フレーム前の状態、P_{t-1}、を座標の起点とするフレーム間の速度、${}^{t-1}P_t$、に変換してから特徴点として使用している (図 3)。

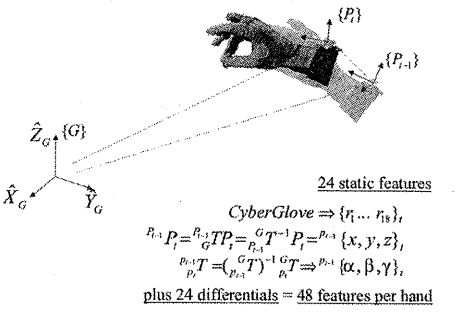


図 3: 特徴量の定義

行動	構成要素	備考
Power Grasp	cls+sp	開いた状態から握力把握
Precision Grasp	tsu+sp	開いた状態から精密把握
Pour	cls+roll+ sp	握力把握から手首をロールして中身を注ぐ
Hand-over	tsu+mae+sp	精密把握した物体を前に差し出し、もどす
Release	opn+sp	把握状態の手を開く
OK-sign	ok+sp	親指と人差し指で丸をつくってから開く
Garbage	gb	スポッティング認識の為のガベージモデル
Start,End	sil	始め、終わりの静止状態

表 1: ジェスチャの定義

ジェスチャの定義では各々に独立した HMM を割り当てず、構成要素の HMM を定義し、それらを繋げたものを使用した (表 1)。各ジェスチャが要素を共用する事により、学習データの不足を補い、学習効率を向上させる事が可能となっている。各要素、(*cls, tsu, roll, mae, opn, ok,gb, sil, sp*) はそれぞれ 5 か 3 state HMM でモデル化されている (図 4)。*cls, tsu, roll, mae, opn, ok* は 5-state left-right (初めから終りまで遷移が一方通行) HMM でモデル化されている。*sil* は トレーニングの際の開始&終了状態、*sp* はジェスチャ終了時の短い静止

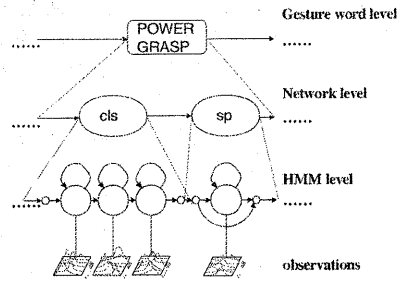


図 4: ジェスチャ定義から HMM への展開例

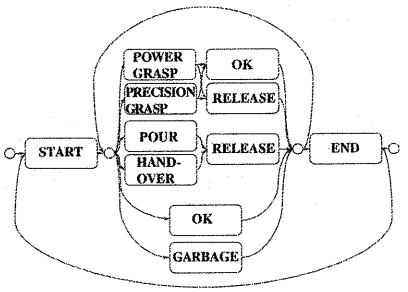


図 5: ジェスチャの状態遷移

状態、*gb* はジェスチャ間の動きを他の要素よりも高い確率で拾うガベージコレクタとして使用する。

HMM の学習では、時間情報付きラベルデータを用いるのが最も望ましい。しかし、本研究が対象とするジェスチャや人間のスキルなどは、データベースが存在しない。そのため、初期モデルの平均ベクトルと分散を学習データ全体から求め、フラットスタート処理を行なった。初期化された HMM は各々連結学習により最適化する。連結学習では、繋げられた複数の HMM が学習データ全体で同時に最適化される。時間情報付きラベルデータは必要無く、学習データに対応したジェスチャの出現順序を記したラベルデータがあれば良い。

本システムは認識に Viterbi アルゴリズムを連続信号認識向けに最適化した Token Passing model を採用している。Viterbi アルゴリズムは、HMM において与えられた時系列データに沿った最適な状態系列と系列上での確率を求めるアルゴリズムである。Token Passing では、データの更新と同時に、遷移情報と確率を含んだ各 Token を更新をしながら最尤 Token を遷移先にコピーしていくことで、Viterbi アルゴリズムと同じ問題を解く。この同時性を使うことで、スポッティング認識を行った。更に、Token にジェスチャの最終状態や遷移時間を含むことで、時系列セグメンテーションも可能である。本研究では、30Hz で入力される両手の特徴ベクトルに対し、遅延無しでジェスチャのスポッティング認識を左右同時に行なっている。なお、Token Passing に使うネットワークの連結情報は指定された文法と辞書から構築される。本研究で採用した文法は図 5 の通りである。

	左	右
% Accuracy	98.89%	95.56%
N,D,S,I	90,0,0,1	90,0,4,0

$$\%\ \text{Accuracy} = \frac{N-D-S-I}{N} \times 100\%$$

(N)number of gestures, (D)eletion error, (S)ubstitution error, (I)nsertion error

表 2: ジェスチャの認識

3.4 認識結果

モデルの学習は片手ずつ行われ、右腕は文法にそったトランスクリプト付き学習データを四パターン（並び方）各五回分とガベージモデル用データを十回分、計 184 秒 (5520 フレーム、内ガベージ用は 1670 フレーム) 分使用。ジェスチャの単語認識は学習データとは別パターン（並び方）のテストセットで実験を行った。テストセットは三パターンのジェスチャ配列を五回ずつ、計 109 秒分使用した。結果を表 2 に示す。

4 9 眼ステレオによる物体認識

本章では、物体認識に用いた 9 眼のマルチベースラインステレオビジョンについて説明し、距離画像に対する 3 次元幾何モデルのマッチング度合いによって物体を認識し 3 次元位置を推定する手法について述べ、ロボットによる物体認識への適用について述べる。

4.1 ステレオビジョンシステム

本研究で使用した 9 眼ステレオビジョンシステム（図 6 右）は、マルチベースラインステレオ [10] のアプローチに基づきコマツによって開発されたもの

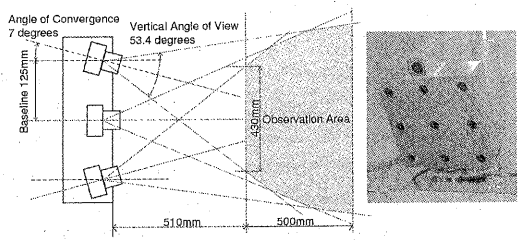
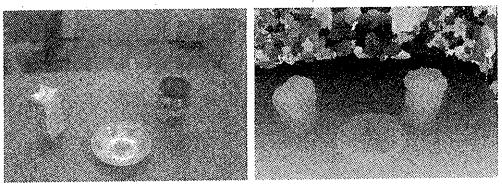


図 6: ステレオビジョンシステム

であり、(i) 8つの画像対を使うためオクルージョンに強い (ii) ハードウェアでリアルタイム (15fps to 30fps) に距離画像が生成可能 (iii) ロボットに一式搭載できるほどコンパクトであるという特徴を持つ [11]。

本研究では、ロボットの視覚として近距離の物体の認識を精度よく行うために、図6左に示すように、ベースラインを広く取りつつ、かつ外周のカメラを内側に傾けるようにカメラ配置の設計を行った。

測定レンジは可変であり、今回は最も近い距離が測定可能なようにレンジを設定した（510mm - 1010mm）。このレンジでの距離画像例を図7に示す。



Intensity　　Depth

図 7: 輝度と距離画像 (白：近い点, 黒：遠い点)

4.2 3次元テンプレートマッチング

獲得した距離画像中で、3次元的に物体の認識を行うために、3D Template Matching(3DTM)[12] の手法を用いた。3DTM では、認識したい物体の3次元形状を既知とし、その3次元幾何モデル（テンプレート）をあらかじめ生成しておく。次に、得られた距離データから構成される3次元空間中に適当に初期位置を設定してモデルを投影し、モデルの各点とその最近傍の3次元点の距離が最小となるようにモデルを6次元空間（位置と姿勢）中で移動させることで、対象物の3次元位置・姿勢を推定する。

モデルと距離データ上の対象物との距離の推定には、最小2乗法の一般形であるM推定量 (Lorentz) を用いる。M推定量では、モデル上の点とそれに対応する距離データ上の点の距離が著しく遠い場合に、その重みを減少することで影響を低減し、モデルの局所的なマッチング精度を高めることが可能である。

4.3 物体の認識

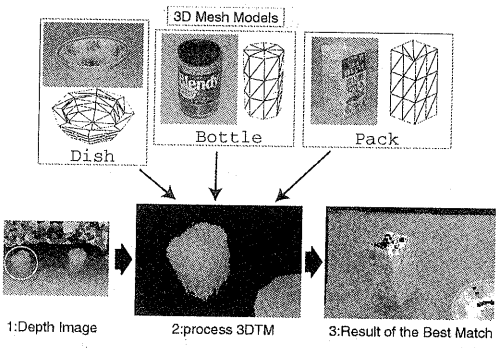


図 8: 3DTM による物体の認識

9眼のステレオビジョンと 3DTM の手法を用いて、次の手順により物体の認識を行う。

まずあらかじめ、図8に示すように、CAD によって対象物体の3次元幾何モデルを作っておく。

次に、ステレオシステムから得られた距離データから物体の切り出しを行う。今回はテーブル上での作業という仮定を置くため、最小2乗法により平面を抜き出し、物体を抽出する。

次に、抽出された全ての物体候補に対して、保持しているモデル全てと 3DTM を用いたマッチング計算を行い、モデルの中で最も一致度の高いものを対象物のモデルであるとして、認識を行う。

今回の実験で用いる図8の3つの物体についてマッチング計算を行い、収束した状態でのマッチング度合いを計算した結果を表3に示す。上の段がM推定量による正規化された距離を示し、下の段が単純な平均距離 (cm) を示している。全てのモデルについて正確な認識がなされていることが分かる。

Model	Image (Depth Data)		
	Pack	Dish	Bottle
Pack	0.25	2.08	0.92
	1.36	10.45	3.62
Dish	1.30	0.65	1.20
	5.78	2.65	5.13
Bottle	0.55	1.43	0.37
	2.27	5.95	1.73

Upper Row: Lorentz (M-estimators)
Lower Row: Averatge Distance (cm)

表 3: 認識の結果

5 ロボットによる人間行動の認識と再現

5.1 プラットフォーム

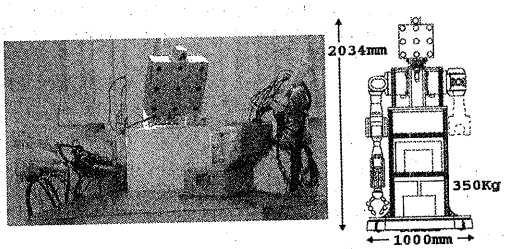


図 9: プラットフォーム

人間の行動を真の意味で獲得したと言うためには、その獲得した行動が実際に実世界において有効に機能するか否かを検証することが欠かせない。人間行動の内、本研究が対象とする手作業においては、眼で対象を認識し、両手を用いて対象物を操作することが作業の大半を占める。そのため、実現プラットフォームには人間の眼と両手に相当する機構を持たせる必要があり、本研究では新たに人間の上半身の機能を持たせたロボット（図 9）を開発した。

本ロボットの特徴は次の通りである。

視覚 3 次元認識のために 9 眼ステレオビジョンシステムを搭載。

腕 腕として、7 自由度のロボットアームを左右に搭載。また手として、右手 4 本指・左手 3 本指（指 1 本につき 3 自由度・指先に 6 軸力力覚センサ）のロボットハンドを搭載。

外部 ソフトウェア的に外部デバイスを容易に拡張・追加できる構成になっており、現在両手分のデータグローブが接続されている。

これらは全てオンラインでつながっており、認識から動作までロボットが連続的に処理を行うことが可能である。

5.2 実験例

実験例として、人間が容器 A を持ちその中身を容器 B の中に注ぐ動作を行い、ロボットによる認識と再現を行った例を示す。

5.2.1 人間行動の認識

人間は両手のデータグローブを装着し、左手で容器 A を持ちその中身を右手で持った容器 B に注ぐ動作を行った。この動作（図 10 の左）は、第 3 章で述べた手法によって、時間軸に沿って離散化したシンボル列の形でモデル化（図 10 の中央）された。

この例では、人間の最後の動作（左手を離す）は認識されていない。この理由は、径の大きな物の把握・離しは、手の形状がさほど変化しないため認識を誤る確率が高いからであると考えられ、これは今後の課題である。

5.2.2 ロボットによる再現

図 10 の中央に示す行動モデルに従って連続して動作を行うことで、ロボットは人間行動の再現を行った。行動モデルは抽象化されているため、人間が行った時と物体の配置や数が異なる場合にも再現することができる。

この例では、図 11 に示すように、人間動作時には無かった皿が増えており、また、容器 A と容器 B の配置も人間動作時とは異なっていた。

4 章で述べた手法により、ロボットは目の前の物体を認識し、その種類及び 3 次元位置を獲得することができた。

獲得した行動モデルから、人間が握力把握を使って物体を掴んでいることが分かったため、ロボットは対象物が皿ではなく容器であると推定することができた。また同じく行動モデルから、動作時にお

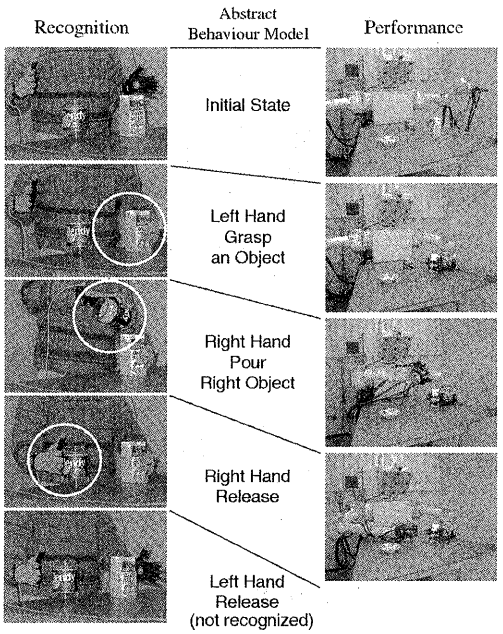


図 10: 実験例

ける両手と対象物の位置関係が分かり、この場合はまず左にあるものを掴み、次に右にあるものを把握し、注いでいることが行動モデルから推定された。

その結果、ロボットは図 10 の右に示すように、まず左側にある容器に対して握力把握を行い、次に右側にある容器に対して注ぐ動作を行うことによって、人間の動作の再現を行うことができた。

6 まとめ

本稿では、人間の手作業をシンボル化された抽象度の高い行動モデルに変換することで、適用範囲の広い人間行動モデルを獲得する手法について述べ、実ロボットを用いた人間行動再現実験による手法の検証例を紹介した。

今回は人間行動の認識に視覚を用いていないため、人間の操作対象物体の詳細な情報は不明であった。今後は、行動観察時に視覚を用い、人間の手と共に抽象化された操作対象物体の情報をも行動モデルに組み込む必要がある。

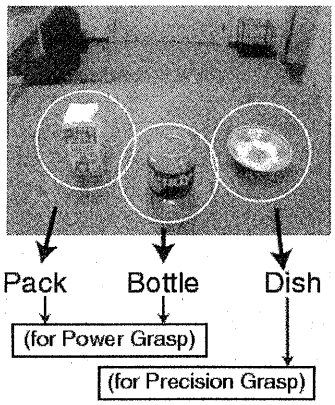


図 11: ロボットの視界

参考文献

[1] K. Ikeuchi and T. Suehiro. Toward an assembly plan from observation part i: Task recognition with polyhedral objects. In *IEEE Trans. Robotics and Automation*, pp. 368–384, 1994.

[2] H. Kimura, T. Horiuchi, and K. Ikeuchi. Task-model based human robot cooperation using vision. In *IROS '99*, pp. 701–706, 1999.

[3] 鎌倉矩子. 手のかたち手のうごき. 医歯薬出版株式会社, 1989.

[4] 中川聖一. 確率モデルによる音声認識. 電気通信学会編.

[5] T. Starner and A. Pentland. Real-time american sign language recognition from video. In *IEEE International Symposium on Computer Vision*, pp. 265–270, 1995.

[6] 西村拓一, 向井理朗, 野崎 俊輔岡隆 一. 低解像度特徴を用いた複数人物によるジェスチャの単一動画像からのスポッティング認識. 電子情報通信学会論文誌 D-II, pp. 1563–1570, 1997.

[7] K. M. Knill and S. J. Young. Speaker dependent keyword spotting for accessing stored speech. In *Cambridge University Engineering Dept., Tech. Report*, p. 193, 1994.

[8] H. K. Lee and J. H. Kim. Gesture spotting from continuous hand motion. In *Pattern Recognition Letters*, pp. 513–520, 1998.

[9] S. J. Young. *HTK: Hidden Markov Model Toolkit V2.2*. Entropic Research Lab Inc., Washington DC, 1999.

[10] 奥富正敏, 金出武雄. 複数の基線長を利用したステレオマッチング. 電子情報処理学会誌 (D-II), pp. 1317–1327, 1992.

[11] 三輪浩史, 新保哲也, 山口博義. リアルタイム多眼ステレオシステム. In *O plus E*, pp. 1259–1264, 1998.

[12] Mark d. Wheeler. Sensor modeling, probabilistic hypothesis generation, and robust localization for object recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 252–265, 1995.