研 究 解 説

Recognition of Human Behaviour using Stereo Vision and Data Gloves

Koichi OGAWARA*, Soshi IBA**, Tomikazu TANUKI***, Yoshihiro SATO***
Akira SAEGUSA****, Hiroshi KIMURA**** and Katsushi IKEUCHI*

This paper presents a novel method of constructing a human behaviour model by attention point (AP) analysis. The AP analysis consists of two steps. At the first step, it broadly observes human behaviour, constructs rough human behaviour model and finds APs which require detailed analysis. Then at the second step, by applying time-consuming analysis on APs in the same human behaviour, it can enhance the human behaviour model. This human behaviour model is highly abstracted and is able to change the degree of abstraction adapting to the environment so as to be applicable in a different environment. We describe this method and its implementation using data gloves and a stereo vision system. We also show an experimental result in which a real robot observed and performed the same human behaviour successfully in a different environment using this model.

1. Introduction

If a robot can learn a human behaviour through an observation and automatically increase the repertoire of its behaviour model, it would be possible to dramatically extend the area of robot applications in a human co-existent environment.

To obtain a human behaviour, a robot must construct some human behaviour model, which then can be applied to a robot to perform the same task, or to perform a cooperative task between a human and a robot.

So far, vision-based robot learning [1, 2] and vision-based cooperation between a human and a robot [3] have been proposed.

In these approaches, once a robot analyzes human behaviour and constructs a human behaviour model, the robot never turns its attention back for a closer analysis. However, it is impractical to apply detailed analysis over the entire human behaviour sequence to obtain human behaviour, though a rough analysis over the entire human behaviour turns out to be inadequate on parts which require detailed analysis. Therefore, we divided the analysis into two steps and proposed a novel method of constructing a human behaviour model by attention point (AP) analysis (Fig. 1).

An attention point (AP) requires close observation to learn a particular behaviour. These include start and stop points of segmented human actions or parts which need closer attention to successfully perform the task.

In the step 1, the robot analyzes human behaviour broadly, and constructs a rough human behaviour model. Then the robot finds APs along the time series. In the step 2, the robot applies a detailed analysis on the same human behaviour that may include data from different input devices. The detailed analyses are applied at around each AP detected during the step 1. This way, the robot can analyze the necessary parts of the human behaviour with much time and care, which helps enhancing the model efficiently.

In the latter chapters, we describe the method mentioned above in detail and the implementation of human behaviour

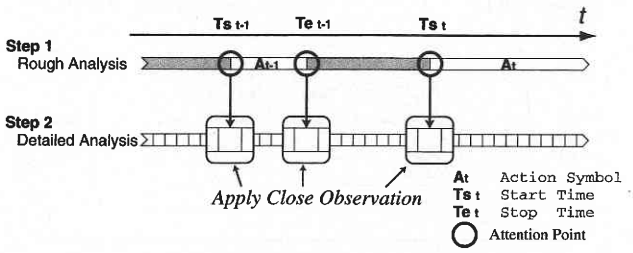

Fig. 1 Two Steps Analysis using Attention Point

* Institute of Industrial Science, University of Tokyo
** The Robotics Institute, Carnegie Mellon University
*** Research Division, Komatsu Ltd.
**** Univ. of Electro-Communications

modeling using data gloves and 9-eye stereo vision system. Then we show an experimental result which demonstrates the applicability of the model in a different environment using a real robot.

2. Recognition of Human Behaviour

In this study, we limited the possible tasks to those on a table. We utilized data gloves (CyberGlove with Polhemus position sensor) and a stereo vision as recognition devices. The robot constructs a rough human behaviour model based on the input data from data gloves and extracts APs from the model. At the second step, the robot observes the human behaviour again. But this time it pays attention only around these APs. The robot fetches an image segment from the vision at around each AP and analyzes it. An analysis on images is much time-consuming compared to an analysis on data gloves, so the AP analysis can enhances the human behaviour model efficiently.

2.1 Rough Human Behaviour Model

The robot constructs a rough human behaviour model by analyzing a sequence of data stream which represents human hand-work by data gloves.

We defined human hand-work as a series of hand actions and classified the attributes of a hand action as shown in Table 1.

First the robot segments the entire human hand-work in series of "Action Symbols" (A_t) by gesture spotting using Hidden Markov Model (HMM) technique as described in chapter 3.

Then the robot indicates each "Action Symbol" by the corresponding hand action and assigns the other 3 attributes "Time Stamp (Ts_t, Te_t)", "Hand" and "Position".

This abstract symbol sequence is the rough human behaviour model (shown in Fig. 2). The robot can simulate nearly the same hand behaviour using this model under the same environment in which the human demonstration was performed, but the robot will easily fail the task in a much complicated environment because information of the target objects is missing.

Table. 1 Attributes of Hand Action

Attributes	Priority	value
Time Stamp	1(low)	Absolute Time (start and stop time)
Action Symbol	3(high)	Power Grasp, Precision Grasp Release, Pour, Hand Over
Hand	2	Right, Left, Both
Position	1	Absolute Position in 3D space
Object Model	3	Type of the Manipulated Object

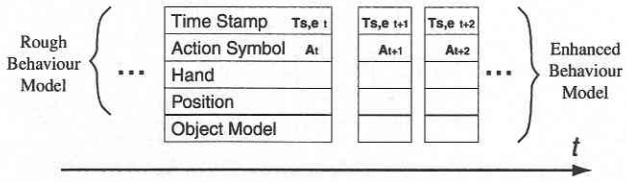

Fig. 2 Human Behaviour Model

2.2 Attention Point (AP)

This study chose the start and stop point (Ts_t, Te_t) of each hand action as an AP. This is because around these points a human-changes his action significantly and this also means that objects may be manipulated dynamically at this point according to the syntax of the behaviour model. By analyzing the human behaviour around these points closely by vision, the robot can obtain information about the manipulated objects used to enhance the rough human behaviour model.

For example, when a human changed his action to "Power Grasp", the robot can tell where and how the object was grasped from the rough behaviour model, but the detailed information of the object is not known. It is difficult to recognize the grasped object by vision after human actually grasped it because of occlusion problems. So we set an AP at this time instance and request a detailed analysis around this point later.

In this case, the robot knows the position where the action was performed and the fact that the target object was grasped from the rough model. Therefore in the step 2, the robot recovers the image before the object is occluded by the grasping hand and recognizes the object by analyzing the specific part of the image where the grasping hand will be placed.

2.3 Detailed Analysis on Attention Points

To realize AP analysis, instead of observing the human behaviour two times, we adopted two kinds of input data streams from a single human behaviour simultaneously, one for a rough analysis and the other for a detailed analysis. The robot records the series of raw depth data of the human behaviour produced by the stereo vision system (for a detailed analysis) while analyzing input data from data gloves (for a rough analysis). Each recorded depth data has time stamp synchronized with "Time Stamp" in the human behaviour model. Therefore, when the robot tries to analyze the human behaviour in detail at an AP, it can fetch the depth data recorded at the moment of the AP (Figure 3). After the robot obtains the proper depth data, the robot recognizes the manipulated objects by 3D Template Matching (3DTM) technique as described in chapter 4. In this study, the robot has 3D CAD models of the objects and can determine which model has the highest likelihood of being the target object in the depth data by 3DTM. With this information, the robot can add the "Object Model" attribute to the behaviour model.

2.4 Priority

Each attribute has the priority term which tells the extent of its importance. First, the robot tries to perform the task exactly as the

model tells. If it fails, the robot ignores the attributes with lower priority for completing the task. For example, if the robot fails to perform the hand action, "Grasp the object A at the place X by the Left Hand", for some reason, it omits the attribute "Position" and "Hand" in order, and tries to grasp the object A with the right hand. Thus, by reducing the constraint before giving up the entire task, the robot can avoid discontinuance of the task.

The AP analysis constructs a human behaviour model which is highly abstracted and is able to change the degree of abstraction adapting to the environment by the priority term. So the model can be applicable in a different environment. We show the example of this applicability in chapter 5.

3. Construction of Rough Human Behaviour Model by Gesture Spotting

To obtain "Action Symbol" for the rough human behaviour model, we aimed at spotting human gestures while a human is performing some hand-works. In this study, we selected 6 gestures (5 described in Table 1 + OK-sign for training) as "Action Symbols" and tried to symbolize human behaviour by gesture spotting with Hidden Markov Models (HMMs) from the input streams from two data gloves. Gestures from each hand are spotted in parallel, while two-handed gestures are spotted by combining results from both hands.

In this chapter, we will go over the general concept of HMM, gesture spotting, and the description of the gesture spotting system.

3.1 HMM

HMMs are used to model a signal with variability in parameter space and time. HMMs model doubly stochastic processes, that are first-order Markov processes whose internal states are not directly observable and, thus, the term "hidden" is used. The observable output signal depends on probability distributions, fixed for each internal state. Since the model can disregard noise through a stochastic framework and it allows us to deal with the highly stochastic underlying structure of the process [4].

3.2 Gesture Spotting

Gesture spotting refers to the recognition and extraction of a meaningful segment corresponding to gestures from input signals that vary in both time and space. By using a gesture spotter, the user is able to interact with the system without keeping start and end of gestures in mind.

HMM-based pattern spotting is done by placing keywords modeled by HMMs and filler models in parallel in a loop. This way, the keyword can be recognized while rejecting non-keywords through filler models [5].

3.3 Recognition using Data Gloves

We used right and left data gloves (CyberGlove), and 6-DOF position sensors (Polhemus) as input devices to perform HMM-based gesture spotting. Part of the system is based on the Hidden Markov Model Toolkit (HTK) [6].

As observable features of the HMMs, we are using 48 dimensional features per hand at time t. The feature vector consists of 18-dimension joint angles, $\{r_1 ... r_{18}\}_t$, 6-dimension hand velocity, ${}^{t-1}P_t = {}^{t-1}\{x, y, z, \alpha, \beta, \gamma\}_t$ which in fact is a velocity referenced from the previous hand coordinate, and differentials of the above 24 features.

We defined a gesture as an attachment of primitive HMMs (Table 2). By sharing primitives, each gesture can use a small number of training data with better efficiency. We defined 9 primitives: *cls*, *prc*, *roll*, *forw*, *opn*, *ok*, are defined as 5-state left-right HMMs (models with single way transition from the start to the end). *sil* is a silent state used at a time of training, *sp* is a short pause which tends to be there at the end of the gesture, and *gb* is a garbage collector that is trained on arbitrary non-gesture movements.

3.4 Recognition Result

Our system can sample the data from the right and left data gloves in 30Hz and spot gestures in parallel without delay. Fig. 4

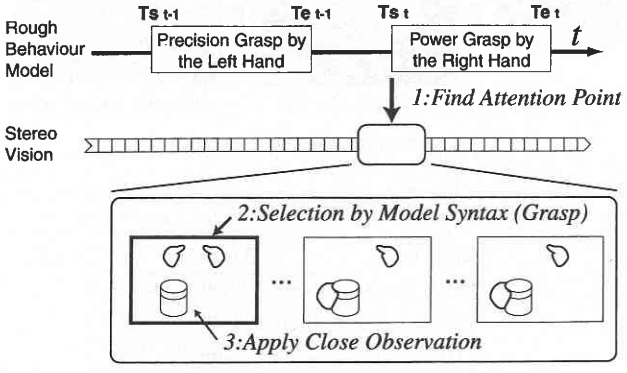

Fig. 3 Example of Applied Model

Table 2 Gesture definitions

Gesture	Primitives	Action
Power Grasp	cls+sp	Power-grasp from open position
Precision Grasp	prc+sp	Precision-grasp from open position
Pour	cls+roll+sp	Power-grasp, and roll the wrist
Hand-over	prc+forw+sp	Precision-grasp, move forward, and back
Release	opn+sp	Open a grasp hand
OK-sign	ok+sp	Make a circle with thumb and index finger
Garbage	gb	A filler model for spotting
Start,End	sil	Silence at the start and end

has information on the HMM grammar network used in gesture spotting.

The model parameters are estimated separately between right and left hands. For right hand gesture modeling, we prepared 5520 frames (20 data sets, 184 sec) of training data which follow the HMM grammar explained in the previous section. Gesture spotting experiments use 15 test data sets (109 sec) with different appearance order of gestures. The recognition results for both hands are shown in Table 3.

4. Detailed Analysis with Stereo Vision System

To obtain the enhanced human behaviour model, the combination of stereo vision and 3D template matching technique is used to recover the type and the position of the object. In this study, we utilized 9-eye multi-baseline stereo vision system [7] to produce depth data in real-time. We extended the baseline and tilted the exterior cameras inward so that the stereo system can produce high resolution depth data of short distance, which is suitable for our robot. The measurable range is changeable and in this study we set up the valid range from 510mm to 1010mm, which is the closest.

In this chapter, we describe 3D Template Matching (3DTM) technique which is adopted to recognize objects in a depth data. Then we describe the method of detailed analysis using stereo vision and 3DTM.

4.1 3D Template Matching (3DTM)

3D Template Matching (3DTM) [8] is a technique to find the precise position and orientation of the target object in a depth data by projecting the corresponding 3D model.

It assumes that the 3D geometric model (template) of a target object and the initial position of the target object is known in advance. It projects the 3D model into the 3D space generated from a depth data. Then it calculates the matching likelihood between the 3D model and the 3D data by summing the weighted distance between each vertex in the template model and the closest 3D point. We adopted M-estimator which is a generalized least squared method as a weight function.

3DTM iteratively moves the model in 6D parameter space (position and orientation) to decrease the distance until it converges.

4.2 Detailed Analysis using 3DTM

The robot performs a detailed analysis as follows. (i) It extracts regions corresponding to objects by removing the background and the table surface from the depth data. (ii) It applies 3DTM in an extracted region using a set of known models (Fig. 5). (iii) It determines the most appropriate model with the best likelihood for that region.

3DTM is sensitive to the initial position of the projected model and produces a better result when the target object is not occluded. From the rough human behaviour model, the robot can select the proper depth data (in which the target object is probably not occluded) in the neighborhood of the AP (from "Action Symbol" syntax) and get the initial position for 3DTM (from "Position" attribute).

Table 4 shows the result of 3DTM applied to the objects used in

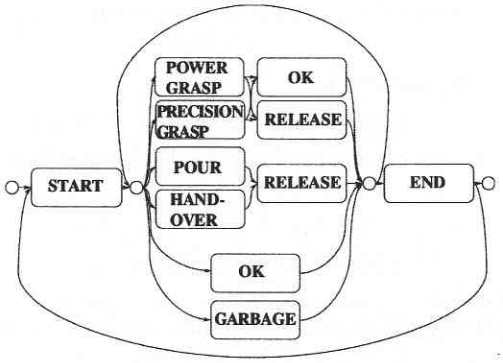

Fig. 4 Gesture Transition Network

Table 3 Gesture Recognition

	Left	Right
% 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)umber of gestures, (D)eletion error, (S)ubstitution error, (I)nsertion error

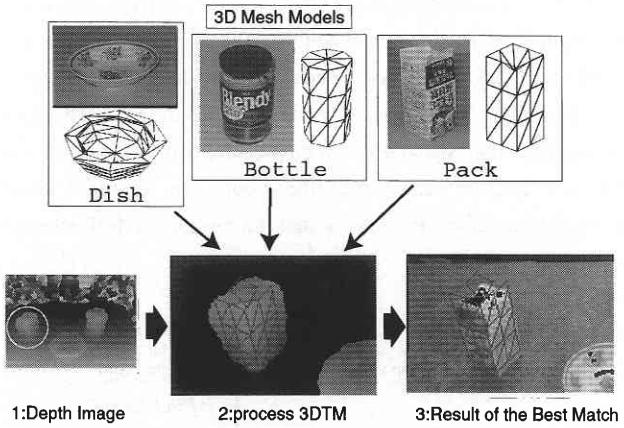

Fig. 5 Recognition of objects with 3 DTM

Table 4 Result of 3 DTM: M-estimator (Lorentzian)

Objects (in Depth Data)	Models		
	Pack	Dish	Bottle
Pack	0.25	1.30	0.55
Dish	2.08	0.65	1.43
Bottle	0.92	1.20	0.37

our experiment. The value indicates normalized weighted distance by M-estimator and the underlined value is the best matching result. The result shows that the objects were correctly registered.

5. Performance and Recognition by Robot

5.1 Platform

We have developed a robot (Fig. 6) as an experimental platform for robot learning and cooperative tasks between a human and a robot. In our research, we focus attention on the learning and the performing of human hand-work by a robot, therefore the platform must have similar capabilities to humans, including vision, dual arms and upper body.

The main features of this robot are as follows.

- It is equipped with 9-eye stereo vision system for 3D recognition (vision).
- It has dual 7DOFs robot arms. The right arm has a hand with 4 fingers and the left arm has a hand with 3 fingers. Each finger has 3DOFs and a Force/Torque sensor on its tip (arms and hands).
- The robot body can freely move on 2D plane in order to move the view point and the arms in any position (upper body).
- CORBA [9] based software architecture enables the robot to be connected easily from exterior devices such as data gloves.

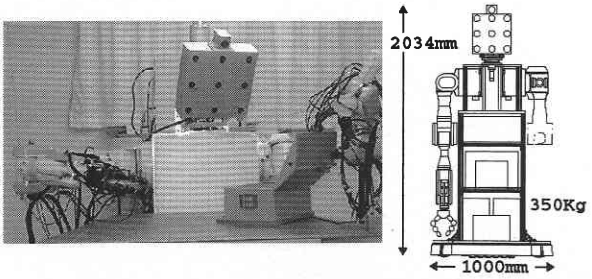

Fig. 6 Platform

5.2 Experiment

To examine the validity of the human behaviour model, we set up an experiment. In this experiment, a human held the container A in one hand and poured the content of B, which was held by the other hand, into the container A. The robot observed the task and constructed a human behaviour model. Then the robot performed the same task in a different environment using the constructed model.

5.2.1 Recognition of Human Behaviour

First, the robot observed the human task through data gloves. As the step 1, the robot constructed a rough human behaviour model by HMM-based gesture spotter in real-time and then found APs (upper partof Fig. 7).

Second, the robot applied detailed analysis with 3DTM on the depth data at each AP as the step 2 (middle part of Fig. 7). This analysis added objects information to the human behaviour model.

5.2.2 Performance by Robot

Then the robot performed the same task using the constructed human behaviour model.

To examine the applicability of this abstract human behaviour model in a different environment, we added a new object "Dish" which was not present at the time of training, and changed the arrangements of the objects on the table.

To recognize objects on the table by the robot, we adopted the same technique described in chapter 4.

The result shows that the robot properly chose the right objects and completed the task (lower part of Fig. 7). The result also shows two kinds of effectiveness of the priority term as described below.

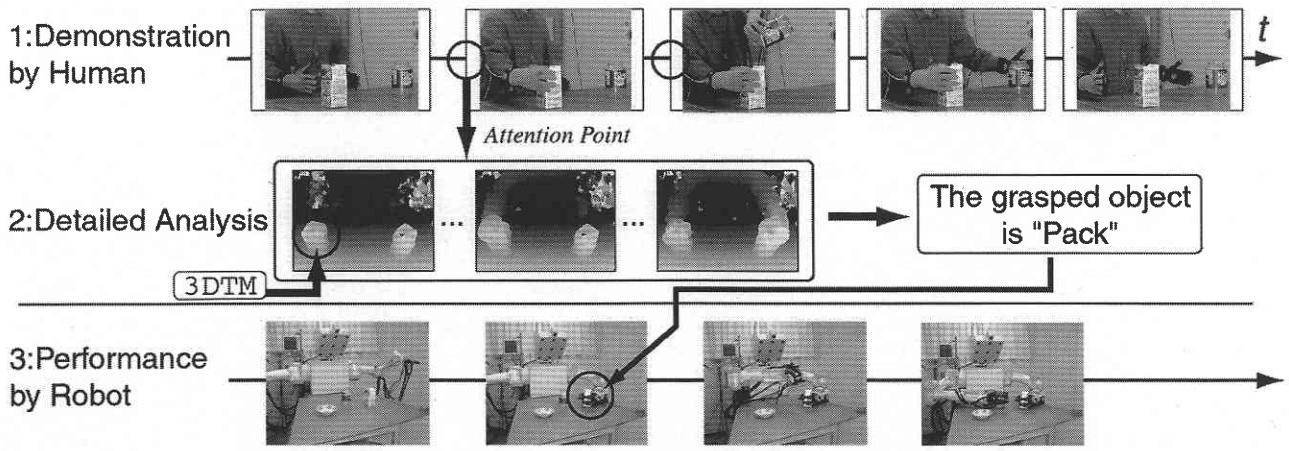

Fig. 7 Experiment

- When the target object was not found in the position described in the model, the robot omitted the "Position" attribute whose priority is low and tried to find the target in the entire view. This feature is especially effective when there are several objects whose geometric shape is similar to each other.
- When the target object was out of the reach of the arm which was described in the model, the robot omitted the "Hand" attribute whose priority is low and tried to reach the object by the other arm.

The priority term changes the degree of the abstraction of each hand action adapting to the environment, so the robot can complete the task while maintaining the model description.

6. Conclusion

In this paper, we presented a novel method of constructing an abstract human behaviour model by attention point analysis using data gloves and 9-eye stereo vision system, which, featured by symbolization and priority term, is highly applicable in a different environment. We showed the validity of this model by an experiment with a real robot. In that experiment, the robot constructed a model of human hand-work from observation and performed the same task successfully in a different environment using this model.

(Manuscript received, March 13, 2000)

References

1) K. Ikeuchi and T. Suehiro, "Toward an Assembly Plan from Observation Part I: Task Recognition With Polyhedral Objects," *IEEE Trans. Robotics and Automation*, 10 (3): 368-384, 1994.
2) Y. Kuniyoshi, *et al.*, "Learning by watching," *IEEE Trans. Robotics and Automation*, 10 (6): 799-822, 1994.
3) H. Kimura and T. Horiuchi and K. Ikeuchi, "Task-Model Based Human Robot Cooperation Using Vision," *IROS* '99, 2: 701-706, 1999.
4) T. Starner and A. Pentland, "Real-time American Sign Language recognition from video," *IEEE International Symposium on Computer Vision*, Coral Gables, FL. 265-270, 1995.
5) K. M. Knill and S. J. Young, "Speaker Dependent Keyword Spotting for Accessing Stored Speech," *Cambridge University Engineering Dept., Tech. Report*, No. CUED/F-INFENT/TR 193, 1994.
6) S. J. Young, "Hidden Markov Model Toolkit V 2.2.," Entropic Research Lab Inc., Washington DC, January 1999.
7) http://www.komatsu.co.jp/research/study56.htm
8) M. D. Wheeler and K. Ikeuchi, "Sensor Modeling, Probabilistic Hypothesis Generation, and Robust Localization for Object Recognition", *IEEE Trans. PAMI*, 17 (3): 252-265, 1995.
9) Common Object Request Broker Archictecture, OMG, July, 1995.