Motion Recognition and Generation by Combining Reference-Point-Dependent Probabilistic Models

Komei Sugiura and Naoto Iwahashi

Abstract— **This paper presents a method to recognize and generate sequential motions for object manipulation such as placing one object on another or rotating it. Motions are learned using reference-point-dependent probabilistic models, which are then transformed to the same coordinate system and combined for motion recognition/generation. We conducted physical experiments in which a user demonstrated the manipulation of puppets and toys, and obtained a recognition accuracy of 63% for the sequential motions. Furthermore, the results of motion generation experiments performed with a robot arm are presented.**

I. Introduction

The recognition of motions has gained considerable interest from the computer vision, robotics, and ubiquitous computing communities [5], [9]. In the robotics community, one of the most important applications of motion recognition is imitation learning [1], [6]. Imitation learning research explores methods to teach a robot new motions by user-friendly means of interaction. In the previous studies, machine learning algorithms such as Forward-Inverse Relaxation Model [8], Gaussian mixture models (GMMs) [2], hidden Markov models (HMMs) [4], and recurrent neural networks [10], [13] have been used for motion learning.

For robots aimed at household environments, motions such as "to put the dishes in the cupboard" are fundamental, but difficult to realize. This is because the desired motion depends on the size and shape of the dishes, as well as those of the cupboard, and also on whether the cupboard has a door. In [5], the difficulties involved in learning such motions are discussed. Ogawara *et al.* proposed a method in which the relative trajectories between two objects are modeled by HMMs [11]. Furthermore, we have proposed a motion learning and generation method that is based on reference-point-dependent HMMs, which enabled the learning of motions such as rotating an object, drawing a spiral, and placing a puppet on a box [3], [14].

In this paper, we propose a novel method that recognizes and generates sequential motions for object manipulation such as placing an object on another (place-on) and moving it away (move-away). In this method, motions are learned using reference-point-dependent probabilistic models, which are then transformed and combined. These composite probabilistic models are used for the recognition of the sequential motions performed by a user. Moreover, motions can be generated from the composite probabilistic models in accordance with user instructions, which can then be performed by a robot arm. Fig. 1 shows the hardware platform used in this study. The system has multimodal interfaces such as a stereo vision camera and a microphone.

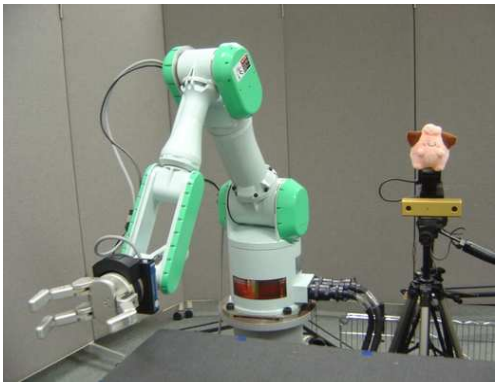


Fig. 1. Hardware platform used in the experiments.

The rest of this paper is organized as follows: Section II first states the problem addressed herein, briefly reviews related work, and introduces our method. Section III describes the motion recognition and generation method in detail. The experimental results for the recognition and generation of sequential motions are presented in Section IV. Section V discusses some problems of our method, and Section VI concludes the paper.

II. Learning Reference-Point-Dependent Motions

A. Reference-Point-Dependent Motions

Motions such as "place-on" and "raise" are dependent on reference points. Let us take the example shown in the left-hand figure of Fig. 2. The figure depicts a camera image in which the green puppet is moved along the dotted line. When the reference point is the blue box, we can provide a label "place-on" to the trajectory. However, the label must be "let the green puppet jump over the green box (jump-over)" when the reference point is the green box.

In cognitive linguistics, a trajector is defined as a participant (object) that is focused on. A landmark has a secondary focus and a trajector is characterized with respect to a landmark. Words representing spatial relationships such as "away" and "left of" are described in terms of a relationship between a trajector and a landmark [7].

Komei Sugiura and Naoto Iwahashi are with Spoken Language Communication Research Laboratories, ATR, and with the National Institute of Information and Communications Technology, 2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan. {komei.sugiura, naoto.iwahashi}@atr.jp

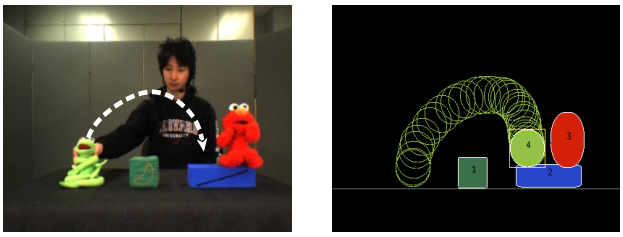


Fig. 2. Left: Example shot of an image stream. The user is manipulating the green puppet. The dotted line represents the trajectory. Right: Preprocessed visual features obtained from the image stream.

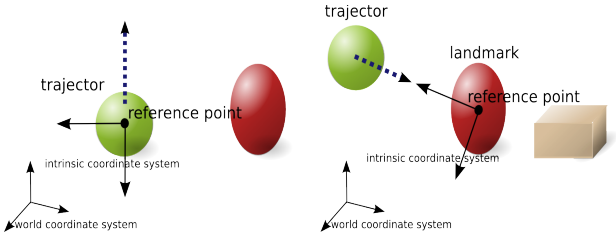


Fig. 3. Relationship between trajector/landmark, a reference point, and an intrinsic coordinate system. The spheres, ellipsoids, and box represent objects, and the arrows represent the axes of the intrinsic coordinate systems. Left: "raise." The small sphere is the trajector, and the reference point is its center. The x axis of the intrinsic coordinate system is horizontal. Right: "move-closer." The direction of the x axis is toward the trajector from the landmark.

Now, we consider the problem of learning reference-point-dependent motions in the framework of imitation learning [1], [6] by a robot. Here, clustering manipulation trajectories and mapping them to a verb are not sufficient for the learning if the trajectories are considered only within the camera coordinate system. For simplicity, we assume that the mapping between the camera coordinate system and the world coordinate system is given, and that the user's utterances are accurately recognized.

Regier investigated a model describing the spatial relationship between two objects [12]. He proposed to model verbs as the time evolution of the spatial relationship between a trajector and a landmark. In [11], the relative trajectories between two objects are modeled by using probabilistic models. The probabilistic models are used for the generation of manipulation trajectories.

In contrast, we have proposed a machine learning method for learning object-manipulation verbs by reference-point-dependent probabilistic models [3], [14]. The method estimates (1) the reference point, (2) *intrinsic coordinate system* type, which is the type of coordinate system intrinsic to a verb, and (3) probabilistic model parameters of the motion that is considered in the intrinsic coordinate system. Let us consider two examples, "raise" and "move-closer" (Fig. 3). We can reasonably assume that the reference point of "raise" is the trajector's center of gravity. The intrinsic coordinate system can be a Cartesian coordinate system, as shown in the left-hand figure. In the case of "move-closer," another type of intrinsic coordinate system is necessary. In this case, the x axis of the coordinate system passes through the centers of gravity of the trajector and the landmark. Before we describe the proposed method, in the next subsection, we briefly introduce the basic concepts and notations of the learning method for reference-point-dependent motions.

B. Motion Learning by Reference-Point-Dependent Probabilistic Models

Consider that L kinds of learning data are given for a verb. Let $\mathcal{V}_l$ denote the lth learning data. $\mathcal{V}_l$ consists of the motion information of the trajector, $\mathcal{Y}_l$, and the candidate set of reference points, $\boldsymbol{R}_l$, as follows:

$$\mathcal{V}_l = (\mathcal{Y}_l, \boldsymbol{R}_l), \tag{1}$$

$$\mathcal{Y}_l = \{\boldsymbol{y}_l(t) | t = 0, 1, ..., T_l\}, \tag{2}$$

$$\boldsymbol{y}_l(t) = \left[\boldsymbol{x}_l(t)^\top, \dot{\boldsymbol{x}}_l(t)^\top, \ddot{\boldsymbol{x}}_l(t)^\top\right]^\top, \tag{3}$$

$$\boldsymbol{R}_l = \{\boldsymbol{O}_l, \boldsymbol{x}_l(0), \boldsymbol{x}_{\text{center}}\} \triangleq \left\{\boldsymbol{x}^{r_l} \middle| r_l = 1, 2, ..., |\boldsymbol{R}_l|\right\}, \tag{4}$$

where $\boldsymbol{x}_l(t)$, $\dot{\boldsymbol{x}}_l(t)$, and $\ddot{\boldsymbol{x}}_l(t)$ denote the position, velocity, and acceleration of the trajector, respectively; T_l denotes the duration of the trajectory; and $\boldsymbol{O}_l$ denotes the set of the static objects' centers of gravity. The operator $|\cdot|$ represents the size of a set. The reason why $\boldsymbol{O}$ is included in $\boldsymbol{R}$ is that the static objects are candidate landmarks. We also include the first position of the trajector, $\boldsymbol{x}_l(0)$, in $\boldsymbol{R}$ so that we can describe a motion concept that is dependent only on the object's trajectory. Additionally, the center of the camera image, $\boldsymbol{x}_{\text{center}}$, is added to $\boldsymbol{R}$ to describe motion concepts that are independent of the positions of the objects.

We assume that there are K types of intrinsic coordinate systems, and these are provided by the designer. We denote the type of the intrinsic coordinate system by k. k corresponds to a verb, and the reference point corresponds to each $\mathcal{V}_l$. We obtain the estimated intrinsic coordinate system for the lth data from the estimation of k and the reference point $\boldsymbol{x}^{r_l}$.

Let ${}^{C_k(\boldsymbol{x}^{r_l})}\mathcal{Y}_l$ denote the trajectory in the intrinsic coordinate system $C_k(\boldsymbol{x}^{r_l})$. Henceforth, parameters in a particular coordinate system are written in a similar manner. Now, the index series of reference points, $\boldsymbol{r} = \{r_l | l = 1, 2, ..., L\}$, the type of the intrinsic coordinate system, k, the parameters of a probabilistic model regarding trajectories, λ, are searched for using the following maximum likelihood criterion:

$$(\hat{\boldsymbol{r}}, \hat{k}, \hat{\lambda}) = \underset{\boldsymbol{r},k,\lambda}{\text{argmax}} \sum_{l=1}^{L} \log P(\mathcal{Y}_l | r_l, k, \lambda), \tag{5}$$

$$= \underset{\boldsymbol{r},k,\lambda}{\text{argmax}} \sum_{l=1}^{L} \log P({}^{C_k(\boldsymbol{x}^{r_l})}\mathcal{Y}_l; \lambda), \tag{6}$$

where $\hat{\cdot}$ represents estimation. In [14] and [3], the solution to Equation (6) is explained in detail.

III. Combination of Reference-Point-Dependent HMMs

A. Transformation of HMMs

Now we consider the problem of the recognition and generation of sequential motions based on composite reference-point-dependent HMMs. In speech recognition, HMMs are usually combined by simply aligning them, since they share the same coordinate system. In contrast, in HMM-based speech synthesis, the coordinate systems used in the training phase, C, and that used in the trajectory generation phase, C' are sometimes different. In such cases, the trajectories are generated in C and are then transformed from C to C'.

However, in neither ways can we combine two reference-point-dependent HMMs. This is because the jth HMM parameters are dependent on the $(j-1)$th HMM parameters (Fig. 4). Fig. 5 illustrates an example of the process of combining two reference-point-dependent HMMs. To combine HMMs corresponding to "raise" and "move-closer," the output probability distributions of each HMM must be transformed since they represents distributions on different coordinate systems.

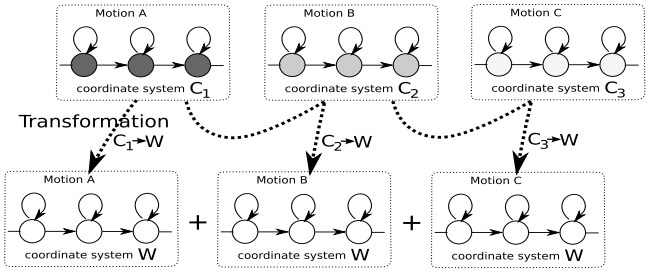


Fig. 4. Schematic of the combination of two reference-point-dependent HMMs. W represents the world coordinate system.

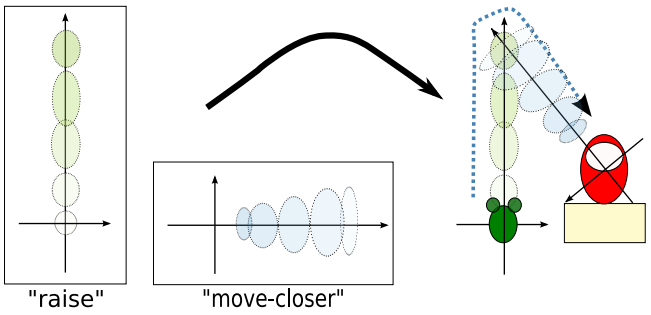


Fig. 5. Example of transformation for the combination of two HMMs, "raise" and "move-closer." Each dotted circle represents the variances of the output probability distribution at each state of a left-to-right HMM. The direction of state transition is indicated by the color darkness. The intrinsic coordinate system for "move-closer" is transformed so that the x axis passes through both the landmark (the reference point of "move-closer") and the last position of the HMM in "raise." The dotted line represents the composite trajectory.

An advantage of transforming intrinsic coordinate systems is the smoothness of the composite trajectories. In our method, velocity and acceleration data are used for learning as well as position data. For safety reasons, changes in the velocity and acceleration data should be continuous. It is therefore important to obtain smooth trajectories of $\dot{\boldsymbol{x}}$ and $\ddot{\boldsymbol{x}}$ when combining two HMMs. Let us consider a case in which verbs dependent only on velocity information, e.g. "throw," are to be combined. If two HMMs were simply aligned to generate the composite trajectory, the velocity changes might be discontinuous in this case. In contrast, our method, which is described in detail below, generates a smooth trajectory.

Now we consider the problem of obtaining a composite HMM from the transformation and combination of reference-point-dependent HMMs. Let $\lambda^{(j)}$ and $C^{(j)}$ denote the parameters and the intrinsic coordinate system, respectively, of the jth HMM, which is a left-to-right HMM. The output probability density function of each state is modeled by a single Gaussian. The mean position vector at state s, ${}^{C^{(j)}}\boldsymbol{\mu}_x(s)$, is transformed by the following homogeneous transformation matrix:

$$\begin{bmatrix} {}^{W}\boldsymbol{\mu}_x(s) \\ 1 \end{bmatrix} = \left[\begin{array}{c|c} {}^{W}_{C^{(j)}}R & {}^{W}\boldsymbol{\mu}_x^{(j-1)}(S_{j-1}) \\ \hline \mathbf{0} & 1 \end{array}\right] \begin{bmatrix} {}^{C^{(j)}}\boldsymbol{\mu}_x(s) - {}^{C^{(j)}}\boldsymbol{\mu}_x(1) \\ 1 \end{bmatrix}, \quad (7)$$
$$(j = 1, 2, ..., D, \quad s = 1, 2, ..., S_j)$$

where ${}^{W}_{C^{(j)}}R$ denotes the rotation matrix from $C^{(j)}$ to the world coordinate system W. Furthermore, $s = 0$ and $s = S_j + 1$ are defined as the initial and final states of the jth HMM, respectively. The mean vector of velocity, $\boldsymbol{\mu}_{\dot{x}}^{(j)}(s)$, and the mean vector of acceleration, $\boldsymbol{\mu}_{\ddot{x}}^{(j)}(s)$, are rotated by using the rotation matrix ${}^{W}_{C^{(j)}}R$.

However, the diagonal items of covariance matrices for position are approximated as follows:

$$\mathrm{diag}\ {}^{W}\boldsymbol{\Sigma}_x(s) = \mathrm{diag}\ {}^{C^{(j)}}\boldsymbol{\Sigma}_x(s), \quad (8)$$

where ${}^{W}\boldsymbol{\Sigma}_x(s)$ and ${}^{C^{(j)}}\boldsymbol{\Sigma}_x(s)$ denote the covariance matrices at state s in coordinate systems W and $C^{(j)}$, respectively. The non-diagonal items of the matrices are equal to zero. The matrices for velocity and acceleration are transformed by the same simple approximation. We do not perform a rotation of the covariance matrix because the HMM-based trajectory generation method [15] we use does not deal with full covariance matrices.

B. Recognition of Motion Sequences by Composite HMMs

Recognition of sequential motions by reference-point-dependent HMMs can be formalized as the problem of obtaining the most likely probabilistic model for trajectory $\mathcal{Y}$ under the condition that verbs, the intrinsic coordinate systems corresponding to verbs, and the HMM parameters corresponding to the verbs are given. Here, let $V = \{v_i | i = 1, 2, ..., |V|\}$ denote a set of verbs, λ_i denote HMM parameters corresponding to verb v_i, and k_i denote the index of intrinsic coordinate systems corresponding to verb v_i.

Suppose that the trajectory of an object, $\mathcal{Y}$, and the candidate set of reference points, $\boldsymbol{R}$, are obtained from an image stream (c.f. Section II-B). Let $(\boldsymbol{i}, \boldsymbol{r}) =$

$(i^{(1)}, i^{(2)}, ..., i^{(D)}, r^{(1)}, r^{(2)}, ..., r^{(D)})$ denote a D-tuple of verb-landmark pairs. We obtain a composite HMM $\Lambda_D(\boldsymbol{i}, \boldsymbol{r})$ from the method explained in Section III-A since the coordinates of the reference points are obtained from $\boldsymbol{R}$ and $r(j)$. The maximum likelihood index sequence of the verb-landmark pairs, $(\hat{\boldsymbol{i}}, \hat{\boldsymbol{r}})$, is searched for through the following equation:

$$(\hat{\boldsymbol{i}}, \hat{\boldsymbol{r}}) = \underset{\boldsymbol{i},\boldsymbol{r},D}{\operatorname{argmax}} P(\mathcal{Y}|\boldsymbol{i}, \boldsymbol{r}, D, \boldsymbol{R}) \tag{9}$$

$$= \underset{\boldsymbol{i},\boldsymbol{r},D}{\operatorname{argmax}} P(\mathcal{Y}|\Lambda_D(\boldsymbol{i}, \boldsymbol{r})) \tag{10}$$

C. Generation of Motion Sequences by Composite HMMs

Now we consider the problem of generating trajectories of sequential motions from composite HMMs. Suppose that a static image and the index of the trajector are given. As in Section II-B, we extract the candidate set of reference points, $\boldsymbol{R}$. Our proposed method deals with two types of motion generation: 1) explicit instruction and 2) target instruction.

1) Explicit Instruction: The trajectory corresponding to the index sequence of the verb-landmark pairs, $(\boldsymbol{i}, \boldsymbol{r})$, is obtained as follows:

$$\hat{\mathcal{Y}} = \underset{\mathcal{Y}}{\operatorname{argmax}} P(\mathcal{Y}|r_{\text{traj}}, Q_D(\boldsymbol{i}), \boldsymbol{r}, \boldsymbol{R}) \tag{11}$$

$$= \underset{\mathcal{Y}}{\operatorname{argmax}} P(\mathcal{Y}|\boldsymbol{x}^{r_{\text{traj}}}, Q_D(\boldsymbol{i}), \Lambda_D(\boldsymbol{i}, \boldsymbol{r})), \tag{12}$$

where $Q_D(\boldsymbol{i})$ denotes the state sequence of the HMM corresponding to verb v_i, and $\boldsymbol{x}^{r_{\text{traj}}}$ denotes the initial position of the trajector. The method explained in [15] provides the maximum likelihood trajectory from the unknown state sequence $Q_D(\boldsymbol{i})$.

2) Target Instruction: Next we consider the problem of obtaining the index sequence of verb-landmark pairs, $(\hat{\boldsymbol{i}}, \hat{\boldsymbol{r}})$, which affords the maximum likelihood trajectory $\hat{\mathcal{Y}}$ from initial position $\boldsymbol{x}^{r_{\text{traj}}}$ to the goal position $\boldsymbol{x}_{\text{goal}}$. We obtain $(\hat{\mathcal{Y}}, \hat{\boldsymbol{i}}, \hat{\boldsymbol{r}})$ by conditioning the right side of Equation (12) with $\boldsymbol{x}_{\text{goal}}$ and then adding $(\boldsymbol{i}, \boldsymbol{r})$ to the search arguments:

$$(\hat{\mathcal{Y}}, \hat{\boldsymbol{i}}, \hat{\boldsymbol{r}}) = \underset{\mathcal{Y},\boldsymbol{i},\boldsymbol{r},D}{\operatorname{argmax}} P(\mathcal{Y}|\boldsymbol{x}^{r_{\text{traj}}}, \boldsymbol{x}_{\text{goal}}, Q_D(\boldsymbol{i}), \Lambda_D(\boldsymbol{i}, \boldsymbol{r})),$$

where the number of combined HMMs, D, is a constant that acts as a search depth parameter. We obtain the solution by applying Tokuda's method [15] as well.

IV. Experiments

A. Experimental Setup

The experiments were conducted with a Mitsubishi Heavy Industries PA-10 manipulator with seven degrees of freedom (DOFs). The manipulator was equipped with a BarrettHand, a four-DOF multifingered grasper. The user's movements were recorded by a Bumblebee 2 stereo vision camera at a rate of 30 [frame/s]. The size of each camera image was 320 × 240 pixels. The left-hand figure of Fig. 2 shows an example shot of an image stream, and the right-hand figure shows the internal representation of the image stream. All the motion data used for learning and recognition were obtained from physical devices. In addition, motion generation results were examined in an environment using the manipulator and physical objects such as puppets and toys.

Motions were taught by the user in a learning phase beforehand, and they were fixed throughout the motion recognition/generation experiments. During the learning phase, the user taught verbs to the robot by uttering them and demonstrating their trajectories. The following verbs were used for learning.

raise, move-closer, move-away, rotate, place-on, put-down, jump-over

For each verb, the number of training data, $L = 9$.

Fig. 6 illustrates some example trajectories in the training set. In the figure, verbs and the estimated type names of the intrinsic coordinate systems are shown below the corresponding figures. We defined the following types of intrinsic coordinate systems:

- C_1 A coordinate system with its origin at the landmark position. C_1 is a transformed camera coordinate system. The x axis is inverted in case the x coordinate of the original position of the trajector is negative after transformation.
- C_2 An orthogonal coordinate system with its origin at the landmark position. The direction of the x axis is from the landmark toward the trajector.
- C_3 A translated camera coordinate system with its origin at the original position of the trajector.
- C_4 A translated camera coordinate system with its origin at the center of the image.

We set the maximum search depth parameter $D_{\max}$ as $D_{\max} = 3$ throughout the motion recognition/generation experiments. In addition, we do not consider collisions between objects in the motion generation experiments.

B. Result (1): Motion Recognition

The user was presented with six pairs of randomly chosen verbs, and performed the motions sequentially. The manipulation trajectories and the positions of the static objects were recorded to obtain a test set. For each pair, five different object settings were given. Therefore, the size of the test set was 30.

Fig. 7 illustrates example trajectories in the test set. In the figures, the top three recognition results for each scene are shown. The bracketed pairs and numbers represent the estimated sequences of verb-landmark pairs and the log likelihood, respectively.

We can see that a correct recognition result was obtained for the left-hand figure of Fig. 7. On the other hand, the correct recognition result for the right-hand figure does not have the maximum likelihood. This is considered to be due to the approximation used earlier (Equation (8)). Here, the point is that only C_2 requires a rotation of the covariance matrix to be combined. And in this case, the correct verb-landmark sequence contains "move-away," which is a C_2 verb.

Table I shows the number of correctly recognized samples. The column labeled "n-best" stands for the number of correct

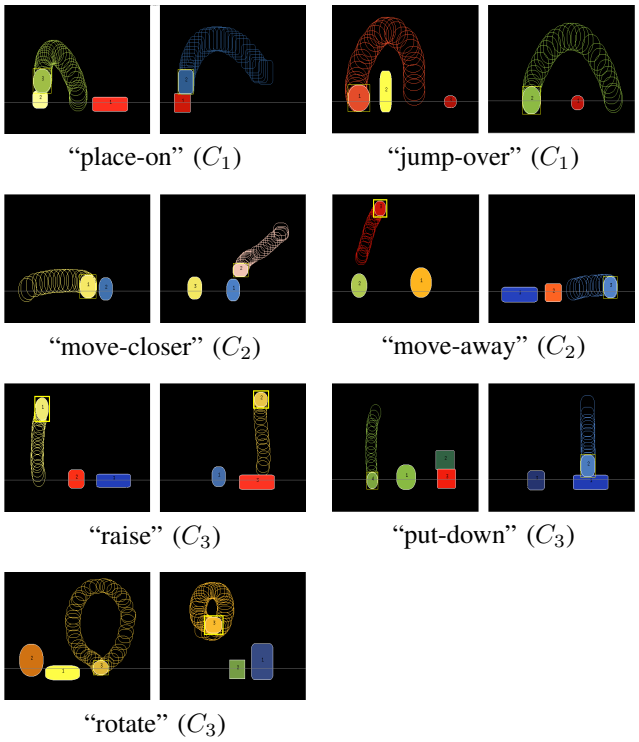


Fig. 6. Examples of training data.

TABLE I

NUMBER OF CORRECTLY RECOGNIZED SAMPLES.

Test set	1-best	2-best	3-best
(1) rotate + rotate	5	5	5
(2) move-away+move-closer	3	3	3
(3) place-on + place-on	3	5	5
(4) rotate + jump-over	4	4	4
(5) rotate + move-away	2	4	4
(6) move-closer + place-on	2	4	5
Total	19/30 (63%)	25/30 (83%)	26/30 (87%)

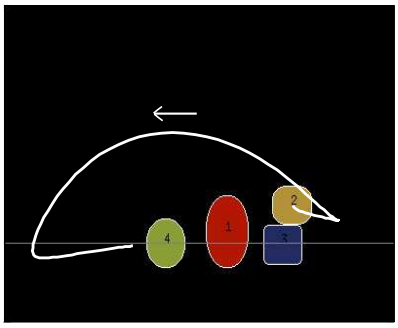
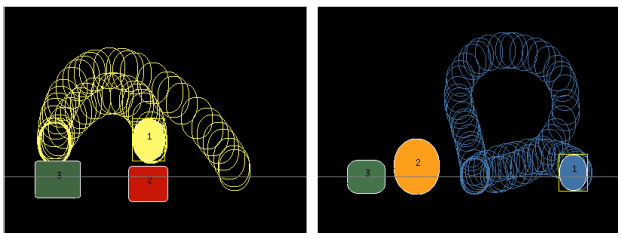


Fig. 9. Generated trajectory in the explicit instruction mode.

answers contained in top n recognition results. The accuracy of 1-best, 2-best, and 3-best recognition results are 63%, 83%, and 87%. In the table, we obtain an accuracy of 80% (12 / 15) for sequences (1), (3), and (4). This is reasonable since we have obtained an accuracy of 90% for the recognition of single motions in preliminary experiments. However, we obtain an accuracy of 47% (7 / 15) for sequences (2), (5), and (6) which contains at least one C_2 verb. This result also supports the fact that the approximation (Equation (8)) deteriorated the recognition accuracy.



1. [place-on, 3] [place-on, 2] : -22.04
2. [jump-over, 2] [place-on, 2] : -23.79
3. [place-on, 3] [move-away, 3] : -28.79

1. [move-away, 2] : -22.18
2. [rotate] [move-away, 2] : -22.65
3. [rotate] : -25.06

Fig. 7. Examples of test set. The recognition results for each scene are shown below the corresponding figure. Left: "place object 1 on object 3, then place object 1 on object 2." Right: "rotate object 1, then move object 1 away from object 2."

C. Result (2): Motion Generation

Fig. 9 shows an example trajectory generated by the proposed method. The solid line represents the trajectory generated in the explicit instruction mode. The input for the explicit instruction mode was as follows:

- trajector ID = 2
- verb-landmark pairs = [move-away, 1] [jump-over, 4] [move-closer, 4]

From Fig. 9, we can see that the proposed method has generated an appropriate trajectory. To support this, the manipulator is shown performing the generated trajectory is shown in Fig. 8.

For the target instruction mode, the top three trajectories are shown in the Fig. 10. The trajector ID was set to 1, and the goal position used is indicated in the figure. In the figure, the solid, broken, and dotted lines represent the best, second-best, and third-best trajectories, respectively. Furthermore, the top three verb-landmark pairs are shown in the figure.

V. DISCUSSION

Now we discuss two causes for the deterioration in the recognition accuracy.

The first one is that sequential motions performed by users tend to be smoothly combined. However, the likelihood for such motions is not always high. This is because the trajectory in the training set always starts from pause ($\dot{\boldsymbol{x}}_l(0) = \mathbf{0}$), and therefore, the composite HMMs contain states representing pauses between motions.

We think that this problem can be solved by using pause HMMs. In this case, a sequence of HMMs comprising a motion HMM sandwiched between pause HMMs are trained[1].

[1]In speech recognition, a sequence of HMMs corresponding to silence (silB), phoneme, and silence (silE) are sometimes used for the training of single phonemes.

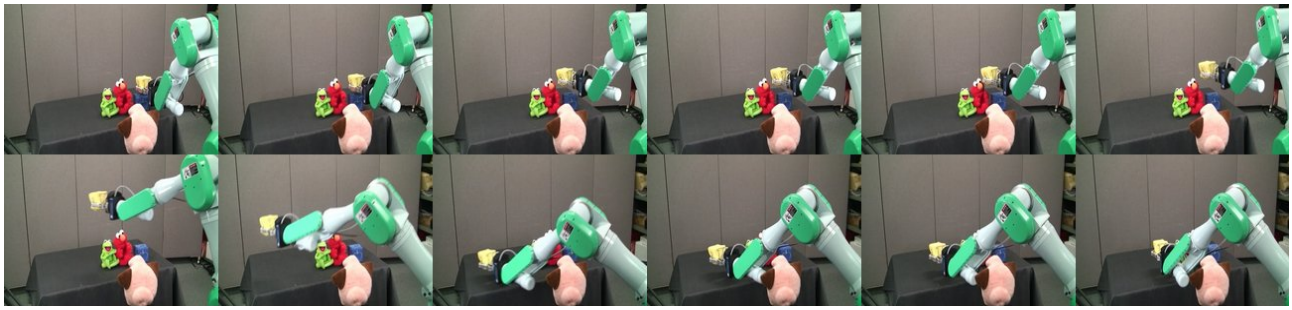
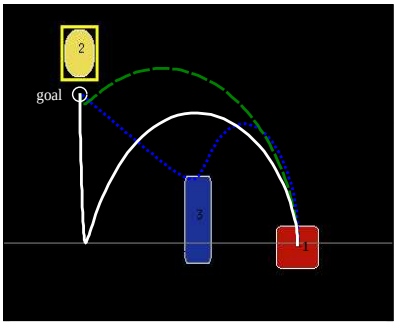


Fig. 8. Sequential photographs of the manipulator executing the trajectory shown in Fig. 9.



1. (solid line) [jump-over, 3] [move-closer, 2] : -16.45
2. (broken line) [jump-over, 3] : -18.66
3. (dotted line) [place-on, 3] [move-closer, 2] : -24.00

Fig. 10. Generated trajectories in the target instruction mode.

Furthermore, we can obtain a composite HMM by aligning the HMMs of pause, motion A, motion B, and pause and thereby combine two motions.

Another problem is that Equation (8) does not consider the full covariance matrices, so the rotation of coordinate systems is ignored. As stated above, we think this approximation deteriorated the recognition accuracy for the C_2 verbs. In the future work, we will perform the rotation of covariance matrices.

VI. Conclusion

Within environments shared by humans and machines, it is important that machines be able to report their internal states to humans in a comprehensive manner. For example, it is critical to the safety of people working around machines that a robot functioning in the same area be able to communicate what it will do next. In this paper, we have described a method to (1) learn motions grounded in real world actions, (2) combine them to recognize human motions, and (3) combine them to generate motions in accordance with user instructions.

VII. Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 20500186, 2008, Tateishi Science and Technology Foundation, and the National Institute of Informatics.

References

[1] C. Breazeal and B. Scassellati, "Robots that imitate humans," *Trends in Cognitive Science*, vol. 6, pp. 481–487, 2002.

[2] S. Calinon and A. Billard, "Incremental learning of gestures by imitation in a humanoid robot," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2007, pp. 255–262.

[3] T. Haoka and N. Iwahashi, "Learning of the reference-point-dependent concepts on movement for language acquisition," in *PRMU2000-105*, 2000, pp. 39–46.

[4] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura, "Embodied symbol emergence based on mimesis theory," *International Journal of Robotics Research*, vol. 23, no. 4, pp. 363–377, 2004.

[5] V. Krüger, D. Kragic, A. Ude, and C. Geib, "The meaning of action: a review on action recognition and mapping," *Advanced Robotics*, vol. 21, no. 13, pp. 1473–1501, 2007.

[6] Y. Kuniyoshi, M. Inaba, and H. Inoue, "Learning by watching: extracting reusable task knowledge from visual observation of human performance," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 6, pp. 799–822, 1994.

[7] R. W. Langacker, *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford Univ Pr, 6 1987.

[8] H. Miyamoto, S. Schaal, F. Gandolfo, H. Gomi, Y. Koike, R. Osu, E. Nakano, Y. Wada, and M. Kawato, "A kendama learning robot based on bi-directional theory," *Neural Networks*, vol. 9, no. 8, pp. 1281–1302, 1996.

[9] T. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[10] T. Ogata, M. Murase, J. Tani, K. Komatani, and H. G. Okuno, "Two-way translation of compound sentences and arm motions by recurrent neural networks," in *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and System*, 2007, pp. 1858–1863.

[11] K. Ogawara, J. Takamatsu, H. Kimura, and K. Ikeuchi, "Generation of a task model by integrating multiple observations of human demonstrations," in *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, 2002, pp. 1545–1550.

[12] T. Regier, *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Bradford Books, 9 1996.

[13] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adaptive Behavior*, vol. 13, no. 1, pp. 33–52, 2005.

[14] K. Sugiura and N. Iwahashi, "Learning object-manipulation verbs for human-robot communication," in *Proceedings of the 2007 workshop on Multimodal interfaces in semantic interaction*, 2007, pp. 32–38.

[15] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 1995, pp. 660–663.