Learning from Demonstration Based on a Mechanism to Utilize an Object's Invisibility

Kotaro Nagahama[1] and Kimitoshi Yamazaki[1]

***Abstract*— This paper proposes a novel visual Learning from Demonstration (LfD) system that teaches robots to learn tasks in which an object is put into another object, stacked, or transported by a tool. In such tasks, observation targets become invisible owing to occlusion or frame-out. The proposed "Visual Hierarchy-based Function Estimator (Hi-Fes)" is inspired by the knowledge derived from the field of psychology that uses the visual hierarchy relationships to estimate the changes in observation targets. Hi-Fes employs a mechanism to interpolate the features when the targets are unrecognizable and state variables are incalculable directly. This method facilitated visual learning of complex state changes between multiple targets, in which the target becomes invisible or has a long period of invisibility, difficult for conventional learning methods. The proposed method was implemented in a life-sized humanoid robot and was evaluated in learning based on demonstration experiments. The results demonstrated the effectiveness of our approach.**

I. Introduction

Despite that the number of workers and caregivers have decreased in the recent aging societies, service robots are increasingly expected to take over the work of a person. For example, daily assistive robots help people at home to perform many daily tasks and solve many problems [1]. Such robots should have the ability to learn tasks specifically adapted in each individual home.

Previous studies have developed Learning from Demonstration (LfD, [2]) for a daily assistive robot to learn daily tasks from a person [3]. LfD is a framework that humans use when they naturally learn something from others. Therefore, an LfD system can be expected to be a key technology for such robots because users who do not have technical knowledge regarding robots will be able to easily teach target tasks to a robot with this technology.

However, tasks with a tool and an operational target or multiple objects that can be learned with conventional observational learning system are limited in their ability to move objects on a two-dimensional plane [4], to hit an object, or to sweep it with a brush [5]. The objective of this study was to construct a robotic system that can learn tasks such as inserting, stacking, and transporting of objects visually. These tasks often appear when using daily tools such as boxes (Fig. 1), trays, and bags. To teach robots visually, it is important to understand the hierarchical or the inclusion relationship between the multiple objects. However, when a hierarchical or an inclusion relationship occurs, the observation target becomes invisible owing to occlusion in many cases, which is problematic. Although humans use visual functions and common sense knowledge to estimate the state of invisible objects, this has been one of the biggest challenges for LfD framework.

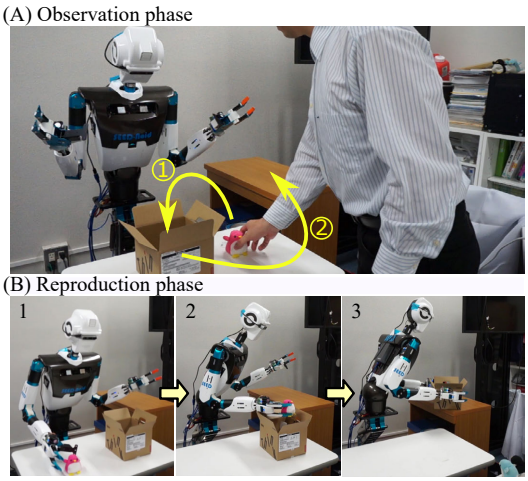

Fig. 1. (A) Observation experiment $T4$ and (B) reproduction experiment $R4$ with Hi-Fes and Aero

This study is inspired by knowledge in the psychological field, using novel findings to develop the LfD framework. This study provides the following primary contributions:
1) An LfD system is proposed to learn tasks of which the observation targets become invisible when being put in, stacked, and transported by a tool.
2) Inspired by psychological knowledge, "Hi-Fes" mechanism is proposed, which has the ability to calculate and interpolate the visual hierarchies and motion features, estimating the state changes of objects when the objects are unrecognized.
3) The proposed system was implemented using a life-sized humanoid robot and its effectiveness was evaluated.

II. Related Work and Our Approach

A. Visual Function of Humans

Studies in the field of cognitive and developmental psychology have shown that humans use visual functions to estimate the state of invisible objects. For example, research in the field of developmental psychology has revealed that humans acquire the ability to understand "object permanence" in early childhood [6]. Humans can also understand that an object continues to exist even when it is covered by a cloth. "The tunnel effect" [7] is the ability to understand that an object that disappears as it moves followed by a sudden appearance of a similar object is really the same

[1]Kotaro Nagahama and Kimitoshi Yamazaki are with AIS Lab., Faculty of Engineering, Shinshu University, 410, Mech. Sys. Eng. Buidling, 4-17-1, Wakasato, Nagano, Nagano, Japan {nagahama, kyamazaki}@shinshu-u.ac.jp

object moving behind another object. In addition, humans can also estimate the state under occlusion from surrounding shapes, such as "subjective contour" [8].

The common sense knowledge presupposed by this visual function is not physically correct under all conditions. Therefore, the illusion may occur cognitively in unusual situations. Since such common sense knowledge is generally correct, humans can accurately estimate the states of objects stacked and transported. If such a visual function can be implemented in a robot, human-like LfD is possible for robots.

However, human visual functions have many unknown parts. If a robotic visual system is constructed based only on already known findings in the psychology field, there may be challenges in estimating the situation sufficiently for robotic LfD. Therefore, in this research, knowledge that is considered common sense for humans and is useful for the estimation of multiple objects' states was experimentally clarified and utilized.

B. Observation of Multiple Objects

Kuniyoshi et al. [9] were the first researchers to start the LfD framework. For visual learning of tasks in which the state between multiple objects changes are important, methods that use camera images to track targets and to imitate the object's position on a two-dimensional plane [4] or the stacking state of visible objects [10] have been employed. Methods have been also proposed to visually learn to push, to tap, to sweep with a brush [5], cutting, and making a sandwich [11][12]. However, these experiments are limited to the learning where observation targets are all visible and recognizable.

In research that uses the information other than images, a system is proposed that uses data gloves to teach pick-and-place tasks [13]. Although the position of an invisible object can be estimated using the hand position, the physical restraint of the teacher is problematic.

Following the studies that add special structures to the observational targets, research has proposed the use of a pressure sensor [14] and the tracking of objects using an electromagnetic motion tracking system to learn pouring and mixing tasks [15]. In addition, a method was proposed for tracking objects using a marker-based motion capture system to learn to stack and to pour [16]. Although these approaches include a possibility that the positions of invisible objects can be estimated correctly, it would be challenging to install such special structures in all objects for daily activities.

C. Computer Vision for Occluded Objects

For robust visual tracking, robust image features [17] and filtering methods [18] are widely used in the field of computer vision. However, tracking becomes unreliable when a majority of the tracked object becomes invisible or the length of invisible time is relatively long. In addition, it is difficult to know exactly when and how the object is occluded. Although it is conceivable to increase the viewpoints by multiple camera system, an object in a bag or in a box becomes invisible from all viewpoints, making the use of multiple cameras useless.

Therefore, special mechanisms have been proposed to track an object partially or temporarily occluded while estimating the occluded area and time. A study [19] proposed a method for dividing a template to detect and track the shielded part of an object. The method proposed by [3] tracks an object by predicting the area that will be occluded at the next time to modify the template. The method proposed by [20] was designed to use the local features on or around the target object when the target object was occluded. The method proposed by [21] was designed to switch a tracking target when the target object was occluded. However, these methods are unable to cope with situations where a target object is occluded for a long time or is completely framed out.

On the other hand, essential inputs for the LfD framework is the higher level information such as the object put into, the object stacked, or how the object is moved, rather than the strict position and the shielded region information.

D. Our Approach

As this study was inspired by the method described by [3], the "visual hierarchy" was one of the features to estimate changes in the state of objects. This makes it possible not only to estimate the state changes that the observation target becomes invisible when it is stacked or inserted, but also to cope with frame-out or failure in visual detection of the target.

The visual hierarchy is a relationship in which an object observed by a camera from a certain viewpoint occludes another object or is occluded by another object. The visual hierarchy comprises "Independent" that there is no overlapping relationship between two objects, "One-way (Obj-k)," where object k is observed in front of another object, and "Mutual," whereby the object A hides object B partially and object B partially occludes object A. Using computer vision, a previous study [3] proposed a method to estimate the visual hierarchy and to estimate an object's function. However, the study failed to arrange the relationship between assumptions and common sense knowledge. Moreover, they were unable to estimate a situation where the target object was completely hidden because their methods lacked pieces of common sense knowledge.

Note that changes in the visual hierarchy are not strictly corresponding to the change of the three-dimensional up-/down/inclusion relationships because they depend on the particular viewpoint. However, humans can obtain only viewpoint-dependent information, which gives a common sense knowledge such as the object's permanence and tunnel effect as discussed previously. This helps the application of appropriate perception in daily activities. From this fact, we hypothesized that this framework could be used in an LfD system for a robot. Thus, visual features, which include visual hierarchy, offer enough information to estimate the state changes of objects with set of common sense knowledge.

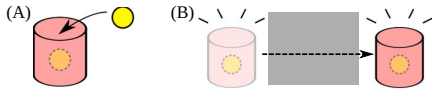

Fig. 2. Example of common sense knowledge on visual hierarchies

However, a visual hierarchy cannot be directly observed by a camera. Humans perceive visual hierarchy using a set of common sense knowledge, such as subjective contour, which is correct in certain assumptions. In this study, a set of common sense knowledge is introduced to interpolate visual features and visual hierarchy.

Fig. 2 shows examples of the above common sense knowledge. When an object becomes invisible after it approaches another object, it is natural to assume that the object is occluded by another object (Fig. 2 (A)). This is a type of knowledge used to estimate the visual hierarchy in this study. When an object becomes invisible and is subsequently detected in a different position, it is natural to assume that the object has moved (Fig. 2 (B)). This is similar to the tunnel effect in human perception and is used for the system considered in this study.

The proposed "visual hierarchy-based function estimator (Hi-Fes)" is a framework that calculates state variables, which includes the visual hierarchy. Moreover, it estimates state changes of multiple objects. The LfD system using the Hi-Fes is described in Sec. III and the details of the Hi-Fes are described in Sec. IV.

III. LfD Using Invisible Information

A. LfD system

Fig. 3 depicts the flowchart of the proposed LfD system. The upper half (green) of Fig. 3 shows the flow of the observation phase, and the lower half (purple) shows the flow of the task reproduction phase.

At the observation phase, the object tracker tracks multiple objects using the images of the robot's camera. The proposed function estimator (Hi-Fes) inputs the tracking results and estimates the state changes of the objects. Hi-Fes outputs the task goals and the objects' usages as the result of the observation.

In the reproduction phase, the inputs of the system are the stored task goals, stored objects' usages, and the initial state of the phase. The task planner generates the executable motion sequences from the inputs; then, sequences are sent to the motion planner, where the inverse kinematics are calculated. Finally, the motor commands are determined and the robot moves.

Even if the observation targets cannot be recognized by the object tracker, Hi-Fes interpolates the state variables during the invisible period and estimates the state changes of the objects. Therefore, the object tracker is not required to output the tracking results at all times. Instead, the Hi-Fes utilizes the object tracker that cannot detect the observational targets to estimate complicated relationships among multiple objects.

Thus, the object tracker does not require a mechanism to estimate targets under occlusion. However, the object

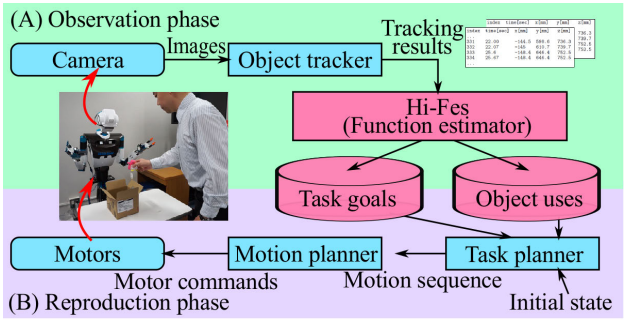

Fig. 3. Proposed robotic LfD system using proposed Hi-Fes

tracker should detect the observation targets after a large occlusion or after a long period of occlusion. Similar to the experiments performed, the Hi-Fes even works with a simple color-based recognition method. It is important to note that the reliability of the system increases with effective robust tracking methods such as LINE-MOD [22].

In this system, the task goals and the object usage are described using Planning Domain Definition Language (PDDL) [23], which is a planning language of STRIPS [24] type. In PDDL, "predicates" are used to describe the initial state, the target state, and the properties of objects. The task planner uses "actions" consist of a set of "preconditions" and "effects" to search the appropriate order of the actions. Finally, the task planner sends action sequences to achieve the task goals to the motion planner. Execution by the motion planner corresponds to each action and the motor commands are calculated, which are used to move the robot.

In this study, three types of predicates are defined to express an object's state, where "at" indicates that an object is located at a particular place; "on" indicates that an object is placed on another object, and "in" indicates that an object is placed inside another object. These predicates enable the system to reproduce the task goals with complicated states between multiple objects. As actions, not only "hold" and "move," but also "place-on-obj" to place an object onto another object and "put-into-obj" to put an object into another object are implemented. Furthermore, four types of predicates to express an object's properties are installed: "as-mat?" indicates that an object can be loaded with another object; "as-storage?" indicates that the object can be inserted in another object; "as-tray?" and "as-container?" indicate that multiple objects with "on" and "in" relationships can be carried together, respectively.

B. Concept of Hi-Fes

Fig. 4 depicts the concept of the proposed Hi-Fes that estimates the state of multiple objects from the temporal transition of the state variables, which are recognizability, visual hierarchy and motion features (Fig. 4, (A)). However, there are state variables that cannot be determined when the observation target cannot be recognized by the object tracker. Therefore, Hi-Fes has a mechanism to mutually interpolate state variables using the knowledge that reflects common sense among humans (Fig. 4, (B)). Arrows in Fig. 4 (B) show Hi-Fes's interpolation of state variables using five types of knowledge $K1$ to $K5$. Whereas $K3$

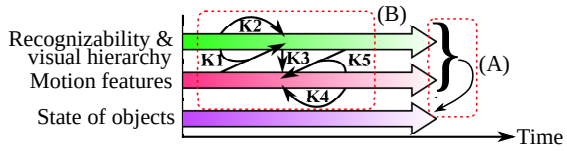

Fig. 4. Concept of Hi-Fes. (A) estimation of objects' states from features; (B) $K1$ to $K5$ are pieces of common sense knowledge for interpolation of features

interpolates motion features from the recognizability and the visual hierarchy at the same time, $K1$ and $K2$ interpolate the state variables at a certain time from the state variables at a past time. Furthermore, $K4$ and $K5$ interpolate the past state variables from the state variables at a future time. These are characteristic structures of Hi-Fes.

The state variables in Hi-Fes and the assumptions that utilize common sense knowledge are described in Sec. IV-A and Sec. IV-C, respectively. Sec. IV-B describes the mechanism to estimate the state changes of multiple objects from the temporal transition of state variables. Sec. IV-D describes the method to mutually interpolate state variables.

IV. Hi-Fes

A. State Variables for Estimation of Objects' State Changes

Five types of state variables are used in this research. Each state variable represents a characteristic of an observation target or a relationship between two objects. Each state variable is defined as follows, where k expresses the index of the observation target.

1) Recognizability of an observation target ($s_V^k(T)$): A state variable defined for each target. "Recognized (Rec.)" is assigned when target k is detected by the object tracker, otherwise "Unrecognized (Unrec.)" is assigned.

2) Movements of an observational target ($s_M^k(T)$): A state variable is defined for each target. "Yes" is assigned when the target is moving, and "No" is assigned when the target is unmoving. If the target is unrecognized and the movements cannot be evaluated, "Unknown" is assigned.

3) Distance between targets ($S_D(T)$): This is one of the state variables representing the motion characteristics of two targets. If the distance between targets is closer than a predefined threshold D_{Near}, "Near" is assigned; otherwise, "Far" is assigned. If one of the targets is unrecognized, "Unknown" is assigned.

4) Co-movements of the targets ($S_C(T)$): "Yes" is assigned when the two targets are moving and the velocity of the targets is similar, and "No" is assigned when the velocities are dissimilar. If one or more of the targets is unrecognized, "Unknown" is assigned.

5) Visual hierarchy ($S_H(T)$): "Independent" is assigned when there is no overlapping relationship between two targets. "One-way (OBJ-k)" is assigned when target k hides another target $\bar{k}$; target $\bar{k}$ is unrecognizable.

$s_M^k(T)$ and $S_H(T)$ are utilized to estimate movement and the hierarchical/inclusion relationships of targets, respectively. $S_C(T)$ is useful for estimating an object carried by another object.

B. Estimation of State Changes from State Variables

In this subsection, a Hi-Fes mechanism is used to estimate the state changes of multiple objects (Fig. 4 (A)).

First, the time set $\mathbb{T}_{start}^k$ when the object k begins to move can be searched in the following manner. The time set $\mathbb{T}_{stop}^k$ when the object stops can be searched in a similar manner.

$$\mathbb{T}_{start}^k = \left\{T_n \mid s_M^k(T_{n-1}) = \text{"No"}, s_M^k(T_n) = \text{"Yes"}\right\} \quad (1)$$

Here, T_n is expressed by $T_{n+1} = T_n + \Delta T$.

Second, the Hi-Fes detects a state change when an object is put onto another object or put into another object. The following common sense knowledge is utilized for the estimation.

1) When Object A approaches another Object B and "One-way (A)" visual hierarchy is detected, A is on B.

2) When Object A approaches another Object B and "One-way (B)" visual hierarchy is detected, A is in B.

As described in Sec. II-D, these pieces of knowledge are not always physically correct. However, it is also impossible for humans to visually check the true constraint, so use of these different types of knowledge are considered to be commonsense for humans.

According to the above knowledge and using Eq. (2) for the calculation, the time set $\mathbb{T}_{LOAD}^k$ when a target is put onto another target:

$$\begin{aligned}\mathbb{T}_{LOAD}^k \leftarrow \{T_n \mid T_n \in \mathbb{T}_{stop}^k, S_D(T_n) = \text{"Near"},\\ S_H(T_n) = \text{"One-way (Obj-k)"}, S_C(T_{n-1}) \neq \text{"Yes"}\}\end{aligned} \quad (2)$$

The time set $\mathbb{T}_{BOX}^k$ when a target is put into another target, this is calculated using Eq. (3):

$$\begin{aligned}\mathbb{T}_{BOX}^k \leftarrow \{T_n \mid T_n \in \mathbb{T}_{stop}^k, S_D(T_n) = \text{"Near"},\\ S_H(T_n) = \text{"One-way (Obj-}\bar{k}\text{)"}, S_C(T_{n-1}) \neq \text{"Yes"}\},\end{aligned} \quad (3)$$

where $\bar{k}$ indicates the index of the target that is not k.

Using the above equations, the Hi-Fes estimates the state changes where an occlusion relationship occurs. Using the visual hierarchy for estimating state changes is one of the primary features of the Hi-Fes. However, the state variables must be determined for these estimations.

C. Assumptions to Interpolate State Variables

Hi-Fes has a mechanism to interpolate state variables based on the common sense knowledge when the variables cannot be directly observed or calculated. In this study, the following four assumptions are set to utilize the knowledge.

Assumption $A1$: In the initial state, the robot can detect all the observation targets whose state changes are important for the task.

Assumption $A2$: The observation targets can be detected at a nearby time when an important appearance or disappearance happens.

Assumption $A3$: The distance between the observation targets is unchanged when one target is hiding another target.

Assumption $A4$: State changes, which are important for the task goals, do not occur while the observation targets are invisible.

These assumptions are inapplicable when a person executes a task while he/she is hiding the observation targets behind a robot. However, in situations of observation, it is natural for the teacher to execute a task to help learners to observe the targets easily. Therefore, these assumptions are considered to be reasonable.

D. Mechanism to Interpolate State Variables

This subsection describes the Hi-Fes's mechanism used to interpolate state variables indirectly. The following five types of common sense knowledge are used for the interpolation:
Knowledge $K1$: When only one target is detected after two targets get close to one another, it is assumed that One-way visual hierarchy occurs and continues until the occluded target appears again. Assumptions $A2$ and $A4$ support this knowledge.
Knowledge $K2$: If both targets become unrecognizable after One-way visual hierarchy is detected, it is assumed that the detection of the target in front failed. Assumptions $A2$ and $A4$ support this knowledge.
Knowledge $K3$: If One-way visual hierarchy continues after both targets become close, it is assumed that both targets continue to be close. Assumptions $A3$ and $A4$ support this knowledge.
Knowledge $K4$: If a target not occluded becomes unrecognizable and then detected again, failure of detection of the target is assumed during that time period. Assumption $A4$ supports this knowledge.
Knowledge $K5$: If target A is hiding target B and that One-way visual hierarchy continues after B is moved enough, it is assumed that B has been moved with A. Assumptions $A2$ and $A3$ support this knowledge.

These types of knowledge interpolated the undefined state variable using other state variables at the time or before and after the time as shown in Fig. 4 (B). Especially, Knowledge $K4$ and $K5$ are inspired by human tunnel effect and object permanence, respectively.

1) Knowledge $K1$-$K3$ and the Evaluation of Co-movements:

Algorithm based on Knowledge $K1$: An algorithm based on Knowledge $K1$ was used to estimate the current visual hierarchy using past motion features and past visual hierarchy. Interpolation of state variables using Knowledge $K1$ was implemented with the following algorithm. Note that $T_{init}(\mathbb{T})$ and $T_{fin}(\mathbb{T})$ are the first and last time of each time sequence $\mathbb{T}$:

1) Time sequences rewritten $\mathbb{T}_n$ are searched, where

$$\begin{aligned}&\mathbb{T}_n \subset \mathbb{T}_{All}, S_D(T_{init}(\mathbb{T}_n) - \Delta T) = \text{“Near”},\\ &s_V^0(T_{init}(\mathbb{T}_n)) = ... = s_V^0(T_{fin}(\mathbb{T}_i)) = \text{“Rec.”},\\ &s_V^1(T_{init}(\mathbb{T}_n)) = ... = s_V^1(T_{fin}(\mathbb{T}_n)) = \text{“Unrec.”}.\end{aligned}$$

2) If there exists $T_n \in \mathbb{T}_i$ such that $S_H(T_n) =$ “Unknown”, S_H is rewritten as $S_H(T_n) \leftarrow$ “One-way (Obj-0)”.
3) Target 0 and 1 are replaced and 1) and 2) are performed again.

Algorithm based on Knowledge $K2$ and $K3$: An algorithm based on Knowledge $K2$ was used to estimate the current visual hierarchy from the past visual hierarchy. An algorithm based on Knowledge $K3$ was used to estimate the distance of targets using visual hierarchy. Interpolation of state variables using Knowledge $K2$ and $K3$ are implemented like the algorithms employed for $K1$.

Note that the interpolation using Knowledge $K1$, $K2$, and $K3$ must continue repeatedly until convergence is achieved.

Evaluation of Co-movements: Subsequently, co-movements of the two targets were evaluated. In case the velocities of both targets are calculated, co-movements are evaluated using the two velocities as follows:

$$\begin{aligned}&S_C(T_n) \leftarrow\\ &\begin{cases}\text{“Unknown”, if } \neg h_M(T_n,0) \text{ and } \neg h_M(T_n,1)\\ \text{“No”, if } (h_M(T_n,0) \text{ and } h_M(T_n,1) \text{ and } \neg h_c(T_n))\\ \quad \text{or } (h_M(T_n,0) \text{ and } \neg h_M(T_n,1))\\ \quad \text{or } (\neg h_M(T_n,0) \text{ and } h_M(T_n,1))\\ \text{“Yes”, otherwise}\end{cases}\\ &h_M(t,k) \triangleq \overline{||\boldsymbol{v}^k(t)||} \geq v_{thr}, h_c(t) \triangleq ||\boldsymbol{v}^0(t) - \boldsymbol{v}^1(t)|| \geq E_a,\end{aligned}$$

where $\boldsymbol{v}^k$ is the velocity of the Object k; v_{thr} and E_a are positive constants to evaluate the presence of movements and co-movements, respectively. $S_C(T)$ cannot be calculated when the velocity of the target is not calculated. If so, “Unknown” is assigned to $S_C(T)$, and the algorithm using Knowledge $K5$ interpolates the variable.

2) Knowledge $K4$ and $K5$:

Algorithm based on Knowledge $K4$: An algorithm based on Knowledge $K4$ was used to estimate the current motion features using future motion features. This algorithm is described as follows:

1) Time section $\mathbb{T}_i$, which rewritten is searched

$$\begin{aligned}&\text{as: } \mathbb{T}_i \subset \mathbb{T}_{All},\\ &s_V^0(T_{init}(\mathbb{T}_i)) = ... = s_V^0(T_{fin}(\mathbb{T}_n)) = \text{“Unrec.”},\\ &S_H(T_{init}(\mathbb{T}_n)), ..., S_H(T_{fin}(\mathbb{T}_n))\\ &\qquad \neq \text{“One-way (Obj-1)”}.\end{aligned}$$

2) At each time $T_n \in \mathbb{T}_i$, $s_M^0(T_n) \leftarrow$ “Yes” if $||\boldsymbol{x}^0(T_{init}(\mathbb{T}_i)) - \boldsymbol{x}^0(T_{fin}(\mathbb{T}_i))|| > A$. $s_M^0(T_n) \leftarrow$ “No,” otherwise. $\boldsymbol{x}^0$ and A are the position of Object 0 and a positive constant, respectively.
3) Target 0 and 1 are replaced, and 1) and 2) are performed again.

Algorithm based on Knowledge $K5$: An algorithm based on Knowledge $K5$ was used to estimate the current motion features using future visual hierarchy and future motion features. Interpolation of state variables using Knowledge $K5$ can be implemented like the algorithms for $K4$.

The Hi-Fes estimated the task goals and the objects' usages by interpolating the state variables with the above-mentioned algorithms; then, estimates the state changes as described in Sec. IV-B.

V. Experiments and Discussion

A. Experimental Setup

To evaluate the proposed method, a life-sized humanoid robot "Aero" ([25], Fig. 1), which we are currently developing, is used for the experiments. Aero has an RGB-D camera mounted on its head, and a total of 26 DoFs (Degrees of Freedom) in the upper body, including two 8-DoF arms, a lower body with two parallel links, and omni-directinal wheels.

In the experiments, a person performed the following four tasks independently.

Task $T1$: Two dishes on a table were carried and placed onto a shelf, respectively.

Task $T2$: Two dishes on a table were stacked and carried, then placed on a shelf.

Task $T3$: A stuffed animal (penguin) and a box on a table were carried and placed on a shelf, respectively.

Task $T4$: A penguin was put into a box on a table and carried, then placed on a shelf (Fig. 1 (A)).

The two dishes were the observation targets in Task $T1$ and $T2$; the penguin and the box were the targets of the Task $T3$ and $T4$. The object tracker detected each object by thresholding the color and the size of the points from the RGB-D camera. The threshold D_{Near} to evaluate the distance was 120 mm, which was defined by the size of the largest target object. The top rows of Fig. 5 ($T1$), ($T2$), and ($T4$) show the images taken by the RGB-D camera during Task $T1$, $T2$, and $T4$, respectively. When the dishes were stacked, one became invisible to the camera and the penguin became invisible when it was put into the box. Note that there were time periods when no observation targets were visible for several seconds due to frame-out, or there was a failure in detecting targets during each task.

After the observation phase, task reproduction experiments were performed using the information associated with each task. The purpose of each reproduction was to reproduce the learned goal using the learned objects' usages. The difference in the structure of the human teacher and the robot learner was taken into account by the task planner in this reproduction. Solver FF(h_a) (FF planner but with a different heuristic h_a [26]) was used as the task planner. The task planner searched for a plan to minimize the number of actions.

Note that the information acquired during the observation experiments is symbolic, which expresses the hierarchical or the inclusion relationship. In the reproduction phase, symbolic action sequence with minimum steps to achieve the goal is searched, which is supposed to be an efficient reproduction. Therefore, geometric information, which includes the range of the table and detailed motion parameters, is known in advance.

B. Experimental Results

The middle rows of Fig. 5 ($T1$), ($T2$), and ($T4$) show the results of estimating the transition of the state variables when using Hi-Fes. The estimated task goal and the objects' usages are shown at the lower part of each figure. "Recog." and "Move." in each figure are referred as Recognizability and Movements, respectively.

In Task $T1$ (Fig. 5 ($T1$)), sometimes two dishes disappeared from the camera view owing to frame-out. However, "Move.=Yes" was correctly estimated by the effect of interpolation using Knowledge $K4$. The same effect was seen in the time of frame-out in Task $T2$, $T3$, and $T4$. Consequently, the $T1$'s task goal is to place two dishes on the shelf, correctly estimated by the Hi-Fes. The Hi-Fes also estimated the $T3$'s task goal of the penguin, and the box was put onto the shelf.

In Task $T2$ (Fig. 5 ($T2$)), the green dish became invisible when the teacher stacked the orange dish onto the green one. However, the visual hierarchy, where the orange dish was hiding the green one was correctly estimated because Knowledge $K1$ was used. The state of the orange dish carried by the green dish was also estimated correctly because Knowledge $K3$ and $K5$ were properly employed. Consequently, the task purpose of stacking the orange dish onto the green one and putting them on the shelf was correctly estimated by Hi-Fes. Furthermore, the permitted usage of the green dish was correctly estimated, namely that the robot could stack an object onto the green dish (as-mat?) and could carry them together (as-tray?). These are useful pieces of information for a robot to reproduce tasks consistently, safely, and efficiently.

In Task $T4$ (Fig. 5 ($T4$)), the penguin initially became invisible. However, the fact that the box and the penguin were near, that the box was hiding the penguin, and that they were moved and put onto the shelf together were correctly estimated based on the Hi-Fes's mechanism that interpolated the state variables using five pieces of knowledge. Finally, the task purpose to put the penguin into the box and to place the box on the shelf were estimated correctly. Furthermore, the box's usage was correctly estimated, namely that the robot can put an object into the box (as-storage?) and can carry them together (as-container?).

Through these observation experiments, the Hi-Fes demonstrated that it could correctly estimate the state changes even if an observation target became invisible when the targets were stacked, one target was put into another target, the targets were transported together, or by frame-out. The Hi-Fes could also estimate the purpose of the task and the objects' usages correctly.

Then, experiments were performed to reproduce the observed tasks. The experimental conditions were as follows:

Reproduction $R1$: Tidying up the two dishes on the table using Task $T1$'s observation result.

Reproduction $R2$: Tidying up the two dishes on the table using Task $T2$'s observation result.

Reproduction $R4$: Tidying up the penguin and the box on the table using Task $T4$'s observation result.

Fig. 6 shows the outputs of the task planner in Reproduction $R1$, $R2$, and $R4$. In Reproduction $R1$, a plan to carry the dishes without stacking them is selected since it is based on the observation of transporting the dishes independently. Note that both hands of Aero's are designed to hold a dish for

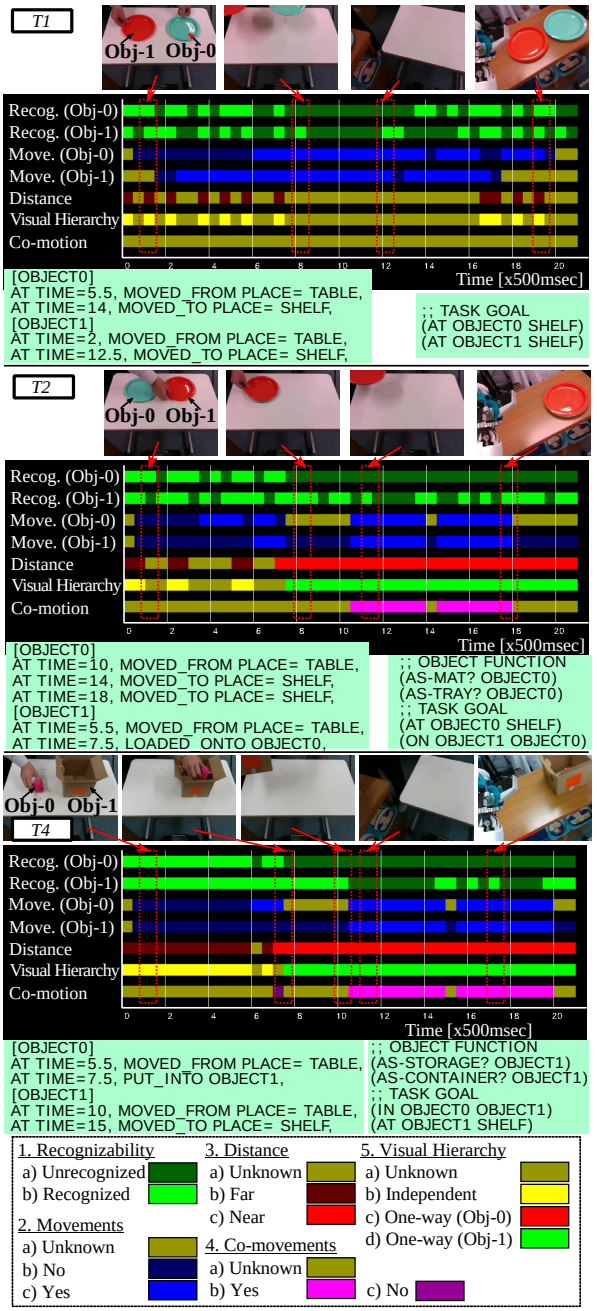

Fig. 5. Result of observation of Task $T1$, $T2$, and $T4$, respectively. The top row of each figure shows the images taken by the RGB-D camera. The middle row of each shows the state variables estimated by Hi-Fes. Color legends for the figures are shown in the bottom. The bottom low of each shows the estimated task goals and the objects' usages.

safety. Therefore, the Task planner decided to carry dishes one-by-one to the shelf although a human teacher held one dish in each hand and carried them at the same time. On the other hand, Reproduction $R2$ and $R4$ were based on the observations in which a human teacher stacked dishes or put an object into a box. For Aero, it is more efficient to carry multiple objects at the same time. For that reason, the plan to carry dishes after stacking (place-on-obj) and the plan to carry the box after inserting the penguin (put-into-obj) were generated. Fig. 7 and Fig. 1 (B) show the result of Aero's execution of $R1$ and $R4$, respectively. The experimental results proved that the proposed LfD system could generate a plan to achieve the task goal of minimum action steps according to the characteristics of the robotic body structure while employing the usage of the object taught by the person.

C. Discussion

The proposed Hi-Fes has the capacity to use knowledge considered natural for human beings and has the ability to use visual hierarchy. These two important elements enable the system to estimate the state changes such as putting in, stacking and carrying. By increasing the number of types of knowledge, the Hi-Fes are capable of learning more complicated state changes that human beings learn naturally. In this case, it is important that different types of knowledge uncontradict each other. One solution is to actively use the knowledge that becomes obvious in the field of psychology.

Regarding the outputs of the Hi-Fes, the task purpose and the objects' usages were independently recorded and used for task planning in this study. This assists in reproducing task goals according to the constraints of the body structure of the robot. In addition, it makes it possible to combine the task objectives learned from multiple observations and to plan action sequences to complete complex tasks.

However, the proposed method cannot currently handle some points. The proposed Hi-Fes worked without any problems when there was a period that the object tracker failed to recognize the observation targets because the Hi-Fes has a mechanism to deal with a false negative result of the tracker. On the other hand, the Hi-Fes cannot currently cope with a false positive, that is, any wrong position's outputs when the target is actually invisible. Therefore, the object tracker must limit the false-positive results. To facilitate more general object tracking methods, mechanisms to extract only reliable data from the tracking results play an important role.

In addition, the current Hi-Fes utilizes a visual hierarchy where an object is completely hidden to estimate putting-in and stacking states. Therefore, this cannot deal with the situations where an object is not completely hidden when it is put into anything or stacked. To address this problem, the proposed method [3] to estimate hidden states from spatial neighborhood features are useful. More reliable and robust state estimations are considered by integrating this method.

Estimating some target objects are unrecognizable in the initial state; however, this is out of Assumption $A1$ and will be considered in a future task. For this estimation, another piece of knowledge is required. If an unrecognizable object A becomes recognizable and it is close to another B at that time, A must have been occluded by B.

In this study, the Hi-Fes estimates the above-mentioned four types of object uses. The uses "as-mat?" and "as-tray?" were employed independently as tools to prevent the Aero from carrying items such as a table cloth. However, further discussion is critical for the best description method of an object's use. The questions arise: should the use be stored for each object or object class or should the pair of the objects be stored? For example, in case household rules indicate that

```
[R1](HOLD AERO DISH-GREEN ARM TABLE)
   (MOVE AERO TABLE SHELF)
   (PLACE AERO DISH-GREEN ARM SHELF)
   (MOVE AERO SHELF TABLE)
   (HOLD AERO DISH-ORANGE ARM TABLE)
   (MOVE AERO TABLE SHELF)
   (PLACE AERO DISH-ORANGE ARM SHELF)

[R2](HOLD AERO DISH-ORANGE ARM TABLE)
   (PLACE-ON-OBJ AERO DISH-ORANGE
                 ARM DISH-GREEN TABLE)
   (HOLD AERO DISH-GREEN ARM TABLE)
   (MOVE AERO TABLE SHELF)
   (PLACE AERO DISH-GREEN ARM SHELF)

[R4](HOLD AERO RED-PENGUIN ARM TABLE)
   (PUT-INTO-OBJ AERO RED-PENGUIN
                 ARM ORANGE-BOX TABLE)
   (HOLD AERO ORANGE-BOX ARM TABLE)
   (MOVE AERO TABLE SHELF)
   (PLACE AERO ORANGE-BOX ARM SHELF)
```

Fig. 6. Experimental results of task planning for the reproduction of the learned tasks

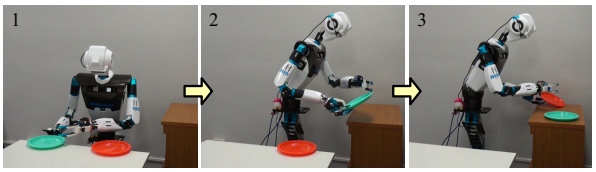

Fig. 7. Experimental results of reproducing Task $T1$ (Reproduction $R1$)

cleanly washed dishes can be stacked for tidying, dirty dishes with oil should not be stacked. Thus, the condition of the object should be learned and stored. To realize such adaptive learning, a framework for determining attributes should be learned by human interactive instructions.

VI. Conclusion

An LfD method has been proposed, whereby a robot learned a task using a tool or multiple objects even when the observation targets are completely invisible. One of the primary features of the proposed method was to use recognizability and the visual hierarchy of targets as a clue to estimate the targets' states. Another feature was to interpolate the state variables and estimate the states of invisible targets using knowledge, inspired by psychological findings and common human sense. This method facilitated LfD operation when an object is put into another object, when an object completely overlaps another object, and when multiple objects become invisible for long period owing to frame-out. The effectiveness of the proposed method was confirmed by experiments of tidying tasks performed by a life-sized humanoid robot.

References

[1] K. Yamazaki, R. Ueda, S. Nozawa, et al., Home-Assistant Robot for an Aging Society, Proc. of the IEEE, Vol. 100, Issue 8, pp. 2429–2441, 2012.

[2] B. D. Argall, S. Chernova, M. Veloso, and Brett Browning, A survey of robot learning from demonstration, Robotics and Autonomous Systems, Vol. 57, Issue 5, pp. 469–483, 2009.

[3] K. Nagahama, K. Yamazaki, K. Okada, and M. Inaba, Hierarchical Estimation of Multiple Objects from Proximity Relationships Arising from Tool Manipulation, Proc. of the 12th IEEE-RAS Int. Conf. on Humanoid Robots, pp. 666–673, 2012.

[4] S. R. Ahmadzadeh, A. Paikan, F. Mastrogiovanni, L. Natale, P. Kormushev, and D. G. Caldwell, Learning Symbolic Representations of Actions from Human Demonstrations, Proc. of the 2015 IEEE Int. Conf. on Robotics and Automation, pp. 3801–3808, 2015.

[5] R. Fukano, Y. Kuniyoshi, and A. Nagakubo, A Cognitive Architecture for Flexible Imitative Interaction Using Tools and Objects, Proc. of the 2006 IEEE-RAS Int. Conf. on Humanoid Robots, pp. 376–381, 2006.

[6] J. Piaget, The construction of reality in the child, New York: Basic Books, 1954.

[7] J. I. Flombaum, S. M. Kundey, L. R. Santos, et al., Dynamic Object Individuation in Rhesus Macaques –A Study of the Tunnel Effect–, Psychological Science, Vol 15, Issue 12, pp. 795–800, 2004.

[8] G. Kanizsa, Margini quasi-percettivi in campi con stimolazione omogenea, Rivista di Psicologia, Vol. 49(1), pp. 7–30, 1955.

[9] 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.

[10] B. Hayes and B. Scassellati, Discovering Task Constraints Through Observation and Active Learning, Proc. of the 2014 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 4442–4449, 2014.

[11] K. Ramirez-Amaro, M. Beetz, and G. Cheng, Transferring skills to humanoid robots by extracting semantic representations from observations of human activities, Artificial Intelligence, Vol. 247, pp. 95–118, 2017.

[12] E. E. Aksoy, M. J. Aein, M. Tamosiunaite, and F. Wörgötter, Semantic parsing of human manipulation activities using on-line learned models for robot imitation, Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 2875–2882, 2015.

[13] A. Skoglund, B. Iliev, B. Kadmiry, and R. Palm, Programming by Demonstration of Pick-and-Place Tasks for Industrial Manipulators using Task Primitives, Proc. of the 2007 Int. Symposium on Computational Intelligence in Robotics and Automation, pp. 368–373, 2007.

[14] K. Matsuo, K. Murakami, T. Hasegawa, K. Tahara, and R. Kurazume, Segmentation method of human manipulation task based on measurement of force imposed by a human hand on a grasped object, Proc. of the 2009 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 1767–1772, 2009.

[15] K. Ogawara, Y. Tanabe, R. Kurazume, and T. Hasegawa, Detecting Repeated Patterns using Partly Locality Sensitive Hashing, Proc. of the 2010 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 1353–1358, 2010.

[16] M. Wächter and T. Asfour, Hierarchical segmentation of manipulation actions based on object relations and motion characteristics, Proc. of the IEEE Int. Conf. on Advanced Robotics, pp. 549–556, 2015.

[17] D. G. Lowe, Object Recognition from Local Scale-Invariant Features, Proc. of the 1999 IEEE Int. Conf. on Computer Vision, pp. 1150–1157, 1999.

[18] M. Isard and A. Blake, CONDENSATION – conditional density propagation for visual tracking, Int. Journal on Computer Vision, Vol. 28, No. 1, pp. 5–28, 1998.

[19] K. Ito and S. Sakane, Robust View-based Visual Tracking with Detection of Occlusions, Proc. of the 2001 IEEE Int. Conf. on Robotics and Automation, pp. 1207–1213, 2001.

[20] H. Grabner, J. Matas, L. Van Gool, and P. Cattin, Tracking the Invisible: Learning Where the Object Might be, Proc. of the 2010 IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1285–1292, 2010.

[21] J. Imai and M. Kaneko, Visual Tracking in Occlusion Environments by Autonomous Switching of Targets, IEICE Transactions on Information and Systems, Vol. E91-D, No. 1, pp. 86-95, 2008.

[22] S. Hinterstoisser, S. Holzer, C. Cagniart, et al., Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes, Proc. of the 2011 IEEE Int. Conf. on Computer Vision, pp. 858–865, 2011.

[23] M. Ghallab, A. Howe, C. Knoblock, et al., PDDL – The Planning Domain Definition Language, Yale Center for Computational Vision and Control, CVC TR-98-003/DCS TR-1165, 1998.

[24] R. E. Fikes and N. J. Nilsson, STRIPS: A NEW APPROACH TO THE APPLICATION OF THEOREM PROVING TO PROBLEM SOLVING, Proc. of the 1971 Int. Joint Conf. on Artificial Intelligence, pp. 608–620, 1971.

[25] K. Sasabuchi, H. Yaguchi, K. Nagahama, et al., The Seednoid Robot Platform: Designing a Multipurpose Compact Robot From Continuous Evaluation and Lessons From Competitions, IEEE Robotics and Automation Letters, Vol. 3, Issue 4, pp. 3983–3990, 2018.

[26] E. Keyder and H. Geffner, The FF(ha) Planner for Planning with Action Costs, Document for Sixth Int. Planning Competition, 2008.