Communications in Information and Systems
Volume 22, Number 4, 431–458, 2022

Topological mapping based on perceiving-acting cycle in sharing cognitive environments for robot partners

Fernando Ardilla, Mohamad Yani, Azhar Aulia Saputra, Weihong Chin, and Naoyuki Kubota

Various kinds of human-friendly robot partners have recently been developed to provide humans with superior services. Manipulation skills, including grasping, arranging, and delivering, are essential for home applications. A robot partner is designed to grasp the meaning of human behavior and their intention in shared spaces to assist older people at home. As a result, the robot partner requires the cognitive ability to comprehend states of the environment based on both people's and robot partners' physical and sensory embodiment. This research presents a human-robot interaction technique for handover behaviors based on cognitive contexts. First, we describe how to share a person's cognitive environment with a robot companion using the relevance idea presented in Cognitive Pragmatics. The perceived cognitive environment of humans contains a type of spatial topological structure, such as relative placement and proximity among objects. Furthermore, the human cognitive environment is continually updated due to the cyclic process of perception and action. As a result, we will look at how to apply topological mapping approaches in cognitive contexts. Next, using the idea of the perceiving-acting cycle presented in Ecological Psychology, we apply topological mapping methods of Growing Cell Structure (GCS) and Growing Neural Gas (GNG). The GCS represents the effectivity in the action system. In contrast, the GNG represents the human and robot task space. The experimental findings and real-world robot application examples indicate that the robot can correctly estimate human intention and conduct handover actions. Finally, we examine the effectiveness of the proposed approach and future research directions in the human-robot interaction based on the perceiving-acting cycle.

1. Introduction

Recently, various types of human-friendly robot partners have been developed to realize sophisticated service to people. Both capabilities of verbal and nonverbal communication are needed for robot partners [1, 2]. However, the capability of environmental perception is also important to realize unconstrained mutual interaction. In general, human communication is restricted by their surrounding environments.

In general, human communication is restricted by their surrounding environments. According to the relevance theory, each person has his or her cognitive environment. Communication starts with making a person pay attention to a specific target object, events, or person. Thus, the cognitive environment of the other person can be enlarged. The shared cognitive environment is called a mutual cognitive environment.

A robot partner should have a cognitive environment to achieve natural interaction. The robot should keep upgrading the cognitive environment based on existing knowledge acquired from experiences with people. A robot partner needs cognitive capabilities of human detection and recognition, verbal communication, nonverbal communication, object, and environmental recognition through interaction with people. In this paper, we focus on how to share cognitive environments. In Figure 1a, an older person approaches the table and looks at the table near the robot. If the person points to a blue cup, the person wants to get the blue cup. This is the ability to try to share the intention with others, but it is difficult to specify the meaning of a gesture as either "can you hand me the blue cup?" or "may I have the blue cup?". However, the meaning of such a pointing gesture can be estimated easier if the robot partner can estimate the reachable human range (Figure 1b). And then, the robot partner should also estimate the possible handover area. In the viewpoint of ecological psychology [3], the surrounding environmental conditions strongly influence the cyclic process of human perception and action based on the physical embodiment. A typical handover behavior is done as a result of a mutual perceiving-acting cycle [4].

This paper proposes a method to estimate the human cognitive environment according to their physical embodiment to realize handover behaviors. We utilize two topological mapping methods i) Growing Neural Gas (GNG) [5–9] to estimate human cognitive environment; ii) Growing Cell Structure (GCS) [10] to estimate human reachable hand range.

This paper is organized as follows. Section 2 explains related work with another research and the background understanding on perception and action of this research to realize the sophisticated service. Section 3 proposes a method of topological mapping based on the perceiving-acting cycle to share mutual cognitive environments. Section 4 shows experimental results to show the effectiveness of the proposed method. Section 5 concludes this paper and discusses the future works to improve the methodology on the perceiving-acting cycle.

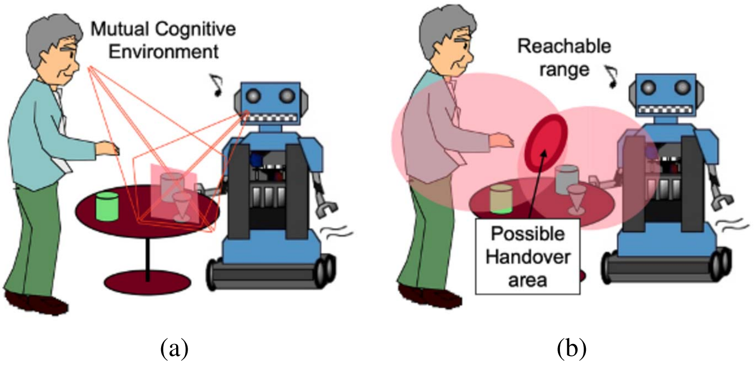


Figure 1: (a) Mutual cognitive environment through the visual and non-verbal communication. (b) An example of handover behaviors based on the estimation of human intention.

2. Related work

How robots can evaluate and comprehend human intents and behaviors has been extensively investigated in various fields, with a variety of foci and methodologies. Numerous disciplines have established frameworks for describing mutual modeling capability [11]. [12] developed robots capable of cooperating with a human user in a confined experimental context by sharing intents with her and accomplished by letting the robot observe a human's goal-directed behavior and then adopt the user's strategy. Thus, the robot demonstrates the capability of determining and recognizing the intents of other agents and sharing intentions with the human user. In [13], the concept of mutual understanding was established by defining it as an agent's ability to predict others and be predicted by others. In a pick-and-place scenario involving a robot and a coworker, the robot arm should approach things and places reasonably [14]. [15] presented online learning of deliberate motion patterns and intention prediction using Hidden Markov Models, enabling fast inference in real-time. The HMM can be gradually trained to cope with novel motion patterns concurrently with projection [16]. Existing systems of human-robot interaction can be categorized according to the type of monitoring employed. One class of devices detects mechanical forces and displacements produced by the robot during physical interaction [17, 18]. Another type of technology is involved in monitoring human communication signals [19]. These systems can also be classified as visual monitoring systems or physiological monitoring systems. Visual monitoring systems capture video data of the human involved in robot-machine interaction and use this data to guide the machine response to the interaction. This monitoring may involve visual tracking of the user's eye-gaze direction [20, 21] and head position, as well as facial expression classification [22], as well as hand and body motion classification [23, 24].

In ecological psychology based on the concept of perceiving-acting cycle [25, 26] related with the perception and action we explain multi-scopic cognitive understanding [8]. In the field of robotics, the concept of perceiving-acting cycle has widely implemented for social robots [26], robot locomotion [27], teleoperated robots [28], intelligent control [29], and object grasping recognition [8]. The measurement or sensing is done at the lowest level by using sensors (Figure 2a). This motion control in the layer directly faces the surrounding environment, and the sensing corresponds to the information measurement. Essentially, reactive motion and sensory-motor coordination are performed at the lowest level or in the most direct way without decision-making. As sensory inputs, the robot measures the necessary environmental data and conducts the corresponding motion control. This level could be assigned to the subsumption architecture because each layer is selective based on direct sensory inputs [30, 31]. The selection mechanism in the original subsumption architecture, on the other hand, covers high-level decision-making. The active perception is the next cognitive level (D). The perceiving-acting cycle in an intentional context must be the lowest unit of analysis in ecological psychology [4]. Here, selective attention plays an important role in extracting external sensory information to continue making a series of motions to take intentional action.

Moreover, the time series of action outputs constructs the spatiotemporal context to persist the specific perception with the dynamics of environments. Therefore, the coupling process of perception and action is called the perceiving-acting cycle (Figure 2b). Furthermore, affordance is an opportunity for action offered by the environment. Here the effectivity is defined as the possibility of realizing action restricted by the current posture. If the posture is changed, its corresponding possible action is changed. Therefore, a suitable posture is required to specify the affordance. The effectiveness of the proposed of affordance-effectivity integration has been proved in robot control [32], object manipulation tasks [33], imitation learning [34]. In our previous research, we have done the affordance-effectivity learning for robot climbing behaviors [35, 36] and safe robot manipulations [37].

The goal-specific information is specified as affordance in an intentional behavior, while the goal-relevant control is specified as effectivity. In this paper, an action is defined as a motion sequence observed by an internal description, while the behavior is defined as a motion sequence followed by an

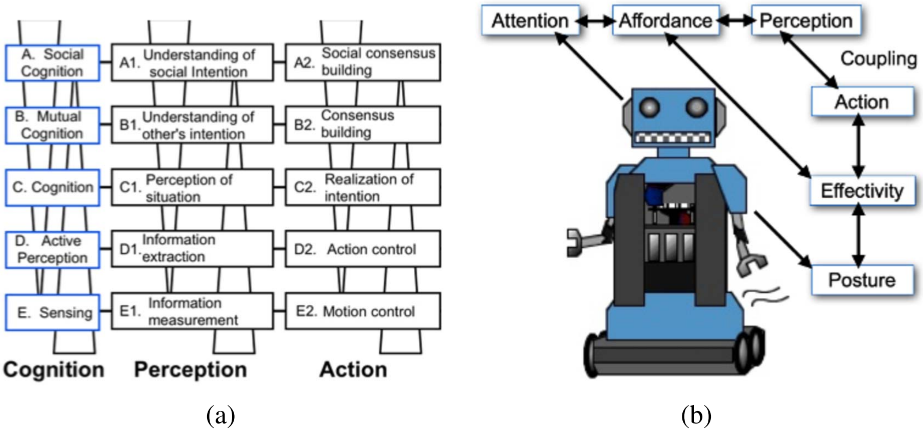


Figure 2: (a) Multi-scopic cognitive understanding on perception and action. (b) Perceiving-acting cycle based on attention and posture.

external description. The next level is based on cognition (C). The situated perception enables the prediction suitable to the spatiotemporal context of the environment. Furthermore, the estimation of human cognitive and physical capacities is useful to predict human goal-directed behaviors. Therefore, the meaning of goal-directed behaviors can be shared between a person and robot partner. In the level of mutual cognition based on relevance theory (B), consensus building or intention sharing is conducted through the communication and interaction between a person and robot partner. As a result, they are able to engage in cooperative behaviors based on the trinomial relationship of humans, robot partners, and objects. Social cognition is linked to ontology and social knowledge at the highest level. Social learning in this context refers to the acquisition of common sense or social value. The goal of robot partners in social interaction is to express social logical thinking or social attitudes to people or establish robot social identities in a human-robot coexisting society.

The most important thing to consider in this figure is the bottom-up construction and top-down constraints. For instance, in B2, a consensus is achieved through bottom-up estimation of others' intentions in C2 while adhering to the top-down constraint of social common sense in A2. The mechanism of cognitive development in structured learning is therefore clarified by top-down constraint and bottom-up building. In the example of Figure 2a, the interpretable intention of a pointing gesture in B2 is restricted by social consensus (common sense) in A2, and the situation is recognized according to the possible behaviors in C2 arising from the conjunction of actions (effectivities) in D2. We explain the procedure of this example in Section 3.

3. Proposed method

Recently, various approaches to Digital Transformation [38], Cyber-Physical Systems [39], and Digital Twin [40] have been proposed and discussed based on the integration of information, intelligence, communication, and robot technologies. The essence of these approaches is to realize super real-time measurement, monitoring, simulation, prediction, search, adaptation, and control integrated mutually from micro-, meso-, and macro-scopic points of view. Especially, feature extraction from big data is important to realize super real-time information processing. The methodology on topological mapping [41], knowledge graph [42] and graph neural networks [43] is very useful to deal with feature-based information processing. Topological mapping methods can extract hidden relationships among features and deal with hidden relationships explicitly. Topological mapping methods are used for 3D modeling available for accurate physics simulation from the microscopic point of view. In contrast, graph-based methods are used for knowledge representation available for huge-scale rule-based inference from the macroscopic point of view. Furthermore, we can build a topological model and knowledge according to a mesoscopic modeling and simulation approach to integrate microscopic models and macroscopic knowledge, called Topological Twin. In this way, topological twin can explicitly deal with relationships hidden in real data to realize digital twin by real world simulation.

3.1. Topological approaches for perception and action

The proposed method comprises two main components: the Growing Cell Structures (GCS) and the Growing Neural Gas (GNG) inspired by the perceiving and acting cycle. GCS and GNG are unsupervised topological mapping algorithms that employ competitive learning to dynamically update the nearby relation (edge) referring to the neighboring node's ignition frequency. The limitation to adding and deleting nodes and edges is the fundamental distinction between GNG and GCS. GCS can be made up of k-dimensional simplices, where k is a positive number that has been selected in advance. GNG, on the other hand, is capable of node deletion and edge structuring but not of maintaining k-dimensional simplices. GNG may partition a data set into several data segments as a clustering approach. In this study, we propose a modified learning method with distance-based node update

and error-based node deletion to enhance rapid online adaptability. We explain the basic algorithm of GCS used in our previous paper [9, 36]. In the algorithm, parameters w_i, A, N_i, and $c_{i,j}$, represent the n-th dimensional reference vector of i-th node, a set of nodes numbers, a set of nodes connected to the i-th node, and edge between i-th and j-th nodes nodes where we assume $c_{i,j} = 0$, respectively.
Step 0: Generate three nodes at random position w_1, w_2, and w_3 in R^n. Initialize the connection set $f(c_{1,2} = 1,\ c_{1,3} = 1,\ c_{2,3} = 1)$.
Step 1: Generate at random an input data p
Step 2: Select the nearest nodes (winner) s by

$$s = \arg\max_{i \in A} \|p - w_i\| \tag{1}$$

Step 3: Update the reference vectors of the winner and its direct topological neighbors by the learning rate μ_1, μ_2 and coefficient μ_3, respectively.

$$w_{si} \leftarrow w_{si} + \mu_1(p - w_{si}) \tag{2}$$

$$w_j \leftarrow w_j + \mu_2 \xi_j (p - w_j) \ \text{ if } c_{s,j} = 1 \tag{3}$$

where $\xi_j = exp(-\mu_3||v - w_j||)$ as a distance-based learning method.
Step 4: Add the squared distance between the input data and the winner to a local error variable.

$$E_s \leftarrow E_s + \|p - w_i\|^2 \tag{4}$$

Step 5: If the number of input data generated so far is an integer multiple of a parameter λ, insert a new node as follows.
i. Select the node q with the maximum accumulated error.

$$q = \arg\max_{i \in A} E_i \tag{5}$$

ii. Select the node f with the maximum accumulated error among the neighbors of q.
iii. Add a new node r to the network and interpolate its reference vector from q and f.

$$w_r = 0.5(w_q + w_f) \tag{6}$$

iv. Insert the edges connecting the new node r with nodes q and $f(c_{r,q} = 1,$ $c_{r,f} = 1)$.

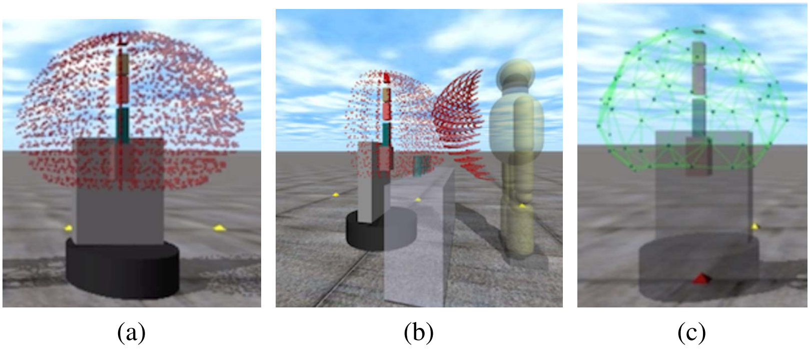


Figure 3: (a) Teaching data robot (b)Teaching data robot and human (c)Learning result.

v. Decrease the error variables of q and f by a fraction α.

$$E_q \leftarrow E_q - \alpha E_q \tag{7}$$

$$E_f \leftarrow E_f - \alpha E_f \tag{8}$$

vi. Interpolate the error variable of r from q and f

$$E_f = 0.5(E_q + E_f) \tag{9}$$

Step 6: Decrease the error variables of all nodes

$$E_i \leftarrow E_i - \beta E_i(\forall_i \in A) \tag{10}$$

Step 7: Continue with step 2 if a stopping criterion (e.g., net size or some performance measure) is not yet fulfilled.

We apply GCS to estimate the effectivity based on movable range. First, we generate teaching data generated by the direct Kinematics (the gripper position (x,y,z) calculated from 5 joint parameters). In this way, GCS can learn Inverse Kinematics. GCS can cover the movable range of the arm of the robot partner. Figure 3 shows a learning result of inverse Kinematics by GCS. The j-th node includes the joint parameters ($v_{j,k}$, k= 1, 2, . . . , o) to move the arm to the target point. The target trajectory from the current position to the target position is generated in the topological map. The nearest node toward the target position is selected sequentially. In order to improve the continuity of a trajectory, the joint parameters are calculated by the distance-based weighted average of the nearest node and its connected nodes step by step.

$$V_k = \frac{\sum_{j \in N_s} \xi_j . v_{j,k} + v_{s,k}}{\sum_{j \in N_s} \xi_j + 1} \tag{11}$$

where s is the selected nearest node.

Next, we explain the learning algorithm of GNG according to the difference with the above GCS. The main difference with GCS is how to update the edge and to delete the nodes and edges. Furthermore, the concept of age is introduced to control the edge connectivity in GNG. Step 2 in GCS is modified in the following.
Step 2'-i. We select the nearest node of GNG (winner) (s1) and the second-nearest unit (s2) of GNG calculated as follows:

$$s_1 = \underset{i \in A}{\operatorname{argmax}} \parallel p - w_i \parallel \tag{12}$$

$$s_2 = \underset{i \in A \backslash \{s_1\}}{\operatorname{argmax}} \parallel p - w_i \parallel \tag{13}$$

Step 2'-ii. We create the connection between s_1-th node and s_2-th ($c_{s1,s2}=1$) node. Then, the age of the connection between s_1-node and s_2-node is set to zero;

$$a_{s_1,s_2} = 0 \tag{14}$$

Step 3 in GCS is modified in the following to deal with age-dependent processing. Step 3'-i. Update the reference vectors of the winner and its direct topological neighbors by (2) and (3).
Step 3'-ii. Increment the age of all edges connecting from s_1.

$$a_{s_i,j} \leftarrow a_{s_i,j} + 1 \quad \text{if } c_{s_{i,j}} = 1 \tag{15}$$

Step 3'-iii. Delete edges with an age larger than a_{max} ($c_{i,j} = 0$). If this results in nodes with no more connecting edges, delete those nodes as well.
Step 6 in GCS is modified in the following to improve the online adaptivity of topological mapping to changing environment.
Step 6'-i: Decrease the error variables of all nodes by (10).

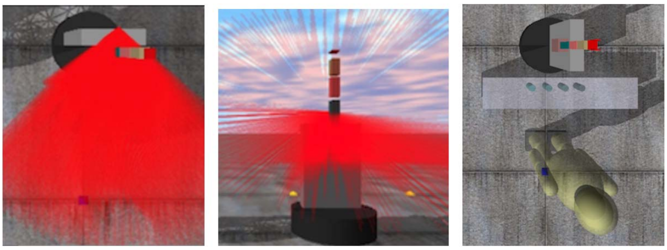


Figure 4: Point cloud measured by virtual 3D-LRF (a) Laser Top View (b) Laser front view (c)An example of layout.

Step 6'-ii: Delete the i-th note and its connecting edges ($c_{i,j} = 0, j \in N_i$), if E_i is less than E_{min} where E_{min} is a deletion criterion.

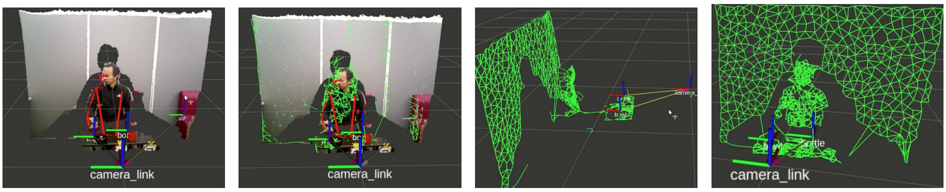
We apply the adaptive GNG to estimate the position of the target object. The input to GNG is the 3D position of the points cloud (x,y,z). The adaptive GNG can conduct online fast clustering and topological mapping. Figure 4a shows the measurement direction of a simulated 3D Laser Range Finder (virtual 3D-LRF) and an example of point cloud measured by virtual 3D-LRF, respectively. The resolution of the point cloud is 200 by 200 on the XY-axis. Figure 4. shows measurement results of the virtual 3D-LRF with down sampling of 5 points on each axis to objects shown in Fig. 4c

3.2. Sharing cognitive environments

In general, human communication is restricted by their surrounding environments, as explained in Chapter 1. According to the relevance theory [44], each person has their cognitive environment, and the communication starts with making a person pay attention to a specific target object, events, or person. Human communication plays a role in achieving consensus. For example, while saying different things, one person can communicate with another and interpret the sense of an unfamiliar word spoken by the other since the symbol corresponds to the perception. The significance principle [45] helps discuss the verbal speech. According to the significance principle, human thinking is exchanged rather than communicated between two beings. The ability of utterances or gestures to draw an individual's attention to a particular target object or entity is one of their most essential functions. The cognitive environment of the other individual may be enhanced as a product of attention's representation. A mutual cognitive environment is a shared cognitive environment in which two or more people share their thoughts. A human-friendly robot should have such a cognitive environment to achieve natural communication. The robot could keep updating the cognitive environment according to current perception through contact with a human.

The research reported in [46] examines how one human's non-verbal communication clues enable others to discern his behavior intentions. The actor's non-verbal communication was captured utilizing a motion-tracking device to capture the actor's body movement and a head-mounted eye tracker to capture the actor's eye gaze behavior. One actor interacted with three individuals and performed one of two actions: placing an object on a table or handing the thing to one of the persons facing him. The activities of putting and giving were chosen because they fell within two types of actions outlined in micro-sociological studies [47]. The placement action is an example of individual activity, but the providing step is an example of interactional action, requiring communication between the interaction partners. Duarte et al. [46] increased focus on the value of nonverbal communication indicators such as arm movement, head movement, and eye movement. Human volunteers were exposed to brief segments of footage of the actor doing one of two possible behaviors. These pieces contain varying levels of information about non-verbal signals, and the goal was to determine the effect of each cue on the capacity to "read" the actor's intents. The obtained data was utilized to simulate armed behavior for the two types of activities and to suggest a gaze controller that, when paired with arm movement, may create human-like motions similar to those found in human-human interaction (HHI) tests. However, the work is incomplete since it only investigates one aspect of the connection. As a result, the next essential step is to investigate not just the nonverbal communication of the human performing the activity, but also the information shared by the environment.

In our work, the sharing information environment is called the shared cognitive environment. A cognitive environment is an area perceived by human attention, and effectivity is represented by topological maps. Furthermore, the area paid more attention to is characterized by a higher density of nodes in topological maps. A robot partner also has a cognitive environment. Each cognitive site is updated through communication and interaction between the human and robot partner. The shared cognitive environment (mutual cognitive environment) is represented by the overlapping human and robot cognitive areas. Figure 5d visualizes the attention of humans and robots by showing the density of nodes, and higher density occurs in the position of humans, tables, and objects around them. Higher density shows

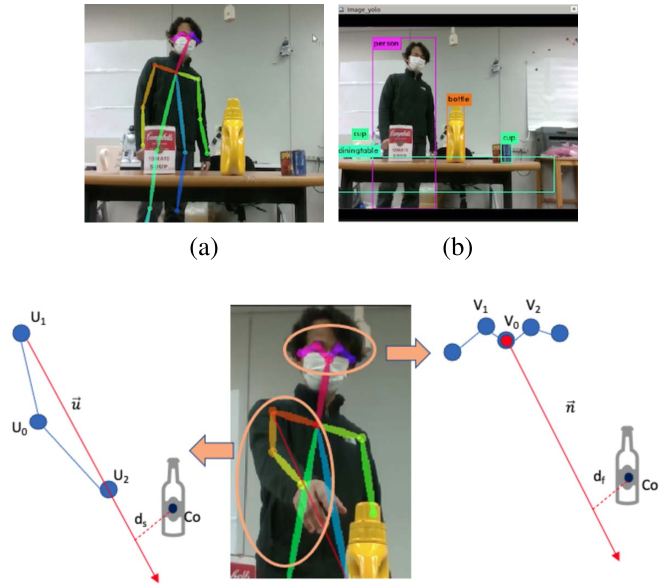


(a) (b) (c) (d)

Figure 5: Updates the attention ranges as a cognitive environment. (a). Human pose detection, red lines depict skeleton poses of humans. (b). Point-cloud and adaptive GNG show in rviz. (c) Adaptive GNG show in rviz. (d). adaptive GNG node density level in human images illustrates the increasing attention of robots in humans.

the intersection of human and robot attention. Human attention leads to the table, and robot attention leads to humans and the table.

Human detection by human pose skeleton and clustering using adaptive GNG. The robot requires intelligent capabilities such as human visual awareness, gesture recognition, and object recognition to share a cognitive environment. As sensory inputs, the robot obtains environmental data from the RGB-D camera to detect human and object recognition. Human posture is identified using the skeleton data generated from the OpenPose [48–50] library. The OpenPose produces 18 annotated 2D main points linked to the human body's posture. A subset of these critical points and their pair-wise anatomical relationships are created for humans, as shown in Figure 6a. Only the point on the face and the right hand is used to identify human intentions in pointing at a target object. The surrounding objects are also recognized simultaneously in addition to human recognition. The method for object recognition is shown in Figure 6b. We use "You only look once" (YOLO) [51]. The robot recognizes the intention from the pointing gesture in Figure 6c according to three conditions in the following; i) the angle at the elbow calculated by eq. 16 is the threshold; θ_{max} ii) both of the distance d_s between human hand direction and the same target object and between the gaze direction and that are shorter than the threshold; d_{min} shown in Figure 6c where the gaze direction is estimated by calculating the normal vector [52] f_0 eq. 17, iii) the time duration is longer than the threshold; t_{max}. When the above three conditions are satisfied, the robot recognizes human intention and conducts a handover behavior of the target object. The robot raised the arm based on the predicted object position after the person demonstrated actions to hit the object for several seconds. At the



(a) (b)

(c)

Figure 6: (a) Human pose estimation (b) Object recognition using YOLO (c). Pointing gesture estimation where the node u is the position on the human arm and v is the position of feature points on face.

end of the learning process, the robot began rotating its arm to grab the desired object. In this way, the robot's predictive ability allows it to have a seamless relationship with a person. In addition, the robot shares a cognitive environment with a person.

$$\theta = \cos^{-1}\left(\frac{(u_1 - u_0)\cdot(u_2 - u_0)}{||(u_1 - u_0)||\cdot||(u_2 - u_0)||}\right) \tag{16}$$

$$N = \frac{(v_0 - v_1)\times(v_0 - v_2)}{\|(v_0 - v_1)\| \times \|(v_0 - v_2)\|} \tag{17}$$

3.3. Handover behaviors

A handover behavior starts by recognizing the target object by utilizing ROS (Robot Operating System). A pointing gesture consists of an arm's motion to point to a target in space and emphasize it to other people through these

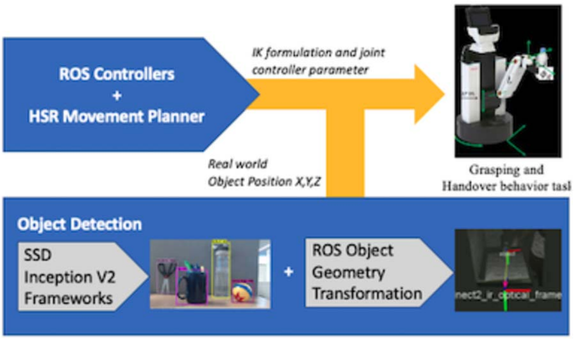

ROS Controllers
+
HSR Movement Planner
IK formulation and joint controller parameter
Real world Object Position X,Y,Z
Grasping and Handover behavior task
Object Detection
SSD Inception V2 Frameworks
+
ROS Object Geometry Transformation
nect2_ir_optical_frame

Figure 7: Grasping and handover behavior experimental flow.

movements without explaining its location verbally. The relation connecting the two joints mentioned above and the named pointing vector is shown in Figure 6c were measured to define a pointing gesture. Figure 7 shows the system architecture used for handover behavior by HSR. First, the center of the detected object is calculated by using an image geometry message on ROS. Next, we will transform it into a 3D Position in the real world. To see the center of the 3D Position in each object, we use an Xtion camera attached to the head of HSR. To realize the grasping task in the handover behavior, HSR detects the 3D central point of the object and solves the inverse kinematics. Moreover, we use the motion planner of HSR to perform post-grasp movement according to the 3D central point of the target human.

4. Simulation and experimental results

4.1. Preliminary simulation results of perception and action

This subsection shows preliminary simulation results of perception and action in the handover behavior by a robot partner. Figure 8a shows a simulation result where the number of nodes in GCS is 108 (left side) and that in GNG is 200 (right side). The second object from the right is selected as the first target, and the node nearest to the target object is set in the perception. Figure 8b shows an estimated path from the current arm position to the target point on the right side of effectivity estimation. However, the resolution for both the perception and the control is not enough. Furthermore, the reached position of the gripper is a little far away from the target point in Figure 8c. The node of GCS can be increased if the distance between

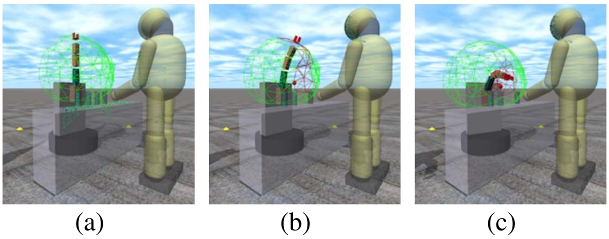


Figure 8: Preliminary simulation results: The trained topological map of GCS is shown virtually at the left side and that of GNG at the right side in each figure.

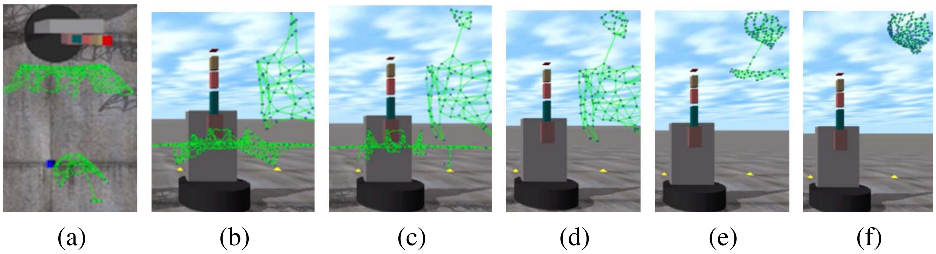


Figure 9: Adaptability of adaptive GNG.

the target and the point of GCS perception is less than the threshold. The threshold value is the minimum distance that the robot's hand can take the object.

Figure 9 shows a simulation result of adaptability of adaptive GNG. The clusters of nodes move to the face from the overall view and picture the transition of the attention range from the human face to the human hand and its near objects. And then, the second object from the right is selected as the target object. When the attention range is transited, nodes decrease quickly, and GNG is adaptive to the transited target. The robot updates the attention ranges as a cognitive environment through verbal and nonverbal communication with a person. The robot partner confirms the facial direction and hand gesture of the person. The attention range is shown in Figure 10, from the left side of the attention, is directed towards the head, then the attention moves towards the human hand and the object. The attention range is shown by the density value of the GNG nodes. Figure 11 offers the numerical value of the number of nodes at each attention and adaptability of the adaptive GNG—the number of nodes changes dependent on the attention of the robot partner. Figure 12 shows the point density level

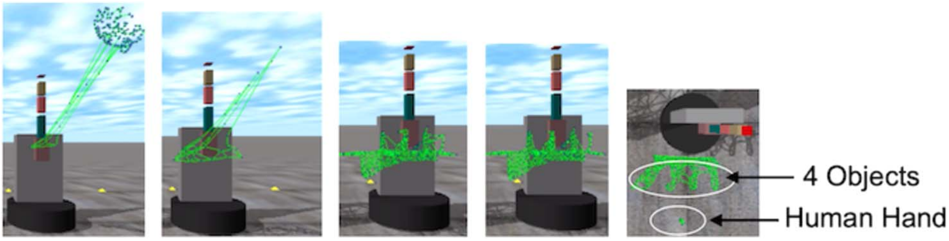
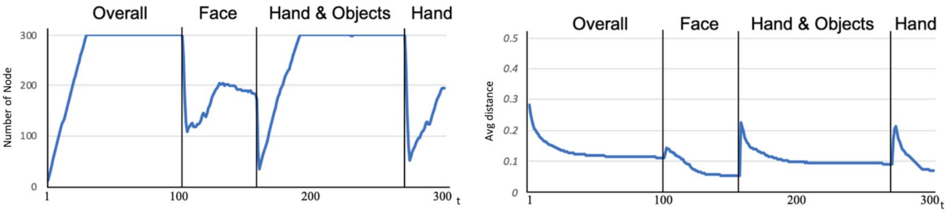


Figure 10: Transition of attention range from human face to the human hand and its near objects.



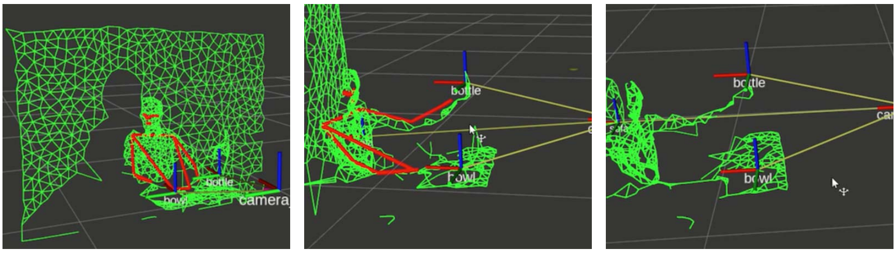
(a) (b)

Figure 11: The learning state of GNG according to the transition of attention range. (a) The number of active nodes. (b). average error (distance) to all sampling data.

as the level of attention. The density of the nodes increases in the human area as the main object. Node density has risen fourfold in the human area than in other regions to reduce the computational cost of recognizing human behavior.

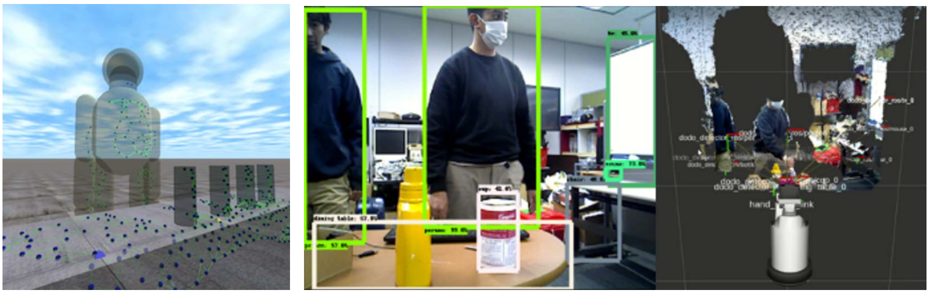
4.2. Experimental results of visual perception

Scenarios are prepared in advance for the experimental environment set up to demonstrate the visual perception of the proposed system, as shown in Figure 13. We show experiments by both simulation and real environment. The simulation uses the Open Dynamic Engine (ODE) to visualize the robot's perception based on the topological map of the adaptive GNG shown in Figure 13a. The blue dot shows the node, and the green line shows the connected edge nodes. The total number of nodes formed from the image is 300, spread across tables, objects, and people. These nodes are used to find out the target object's position for handover behavior to the person. s position for grasping by the robot partner.



(a) (b) (c)

Figure 12: GNG adaptive node density shows the attention range, and high density is in the human position. the thickness is four times more than the other areas.



(a) (b)

Figure 13: (a). Simulation of visual perception (b). Visual perception with SSD Inception in real environment.

In real life environment illustration, the scenario can be seen in Figure 13a. The robot is placed in front of the table with several objects on it. For each attempt, one actor executes the action of trying to reach or point to one of the objects on the table (left/center/right). The distance of each object is between 20–25 cm. The actor was instructed to act as normally as possible when performing those actions. The actor will try to grab the object randomly to prevent the actor from adapting its posture prior to initiation. A total of 60 experiments were carried out with the action configuration: left, center, and right were performed 20 times each. The RGB-D camera tracking system records world camera videos at 30-35 Hz. Furthermore, to recognize the human intention in reaching the surrounding object, it is carried out with the following stages, the human position is identified using Openpose. The robot's attention will increase in the human area, described by the increasing density of GNG nodes around the skeleton. If the time duration of the human arm position shown in Figure 6c exceeds the limit time, then the human gaze direction will be calculated to determine the object's to be achieved. The accuracy of determining the target object is improved by calculating the distance of the GCS and GNG nodes, where the GCS node is the effectivity of humans and GNG is the object's position. Experiments results in the real environment are shown in Figure 13b. Rviz is used to display the results of object recognition experiments. In addition, we analyze the success rate of recognizing human actions to reach objects in the vicinity. The success rate gets 93.3% when Pointing at the object on the right of humans, 85% when pointing at the object in the middle, and the lowest success rate is obtained at the object on the left, which is 78.3%. This happens because the accuracy of the calculation of the angle at the elbow is inaccurate and occlusion occurs when the right-hand points to the left. Errors also often occur when detecting the direction of the human gaze. The points of the facial skeleton often go undetected. As a comparison with other researchers, Duarte et al. [46] read human intent by using motion capture and glasses that can detect the gaze direction. The results are more accurate but require more equipment and costs, such as room-mounted cameras and eyeglasses to see pupils, and also the installation of instruments on humans is considered unnatural and uncomfortable. Luo et al. [53] predict human intention by hand motion. The experiment was conducted similarly with us, using a tabletop manipulation task with three initial positions and four target positions. The experiment was similar to ours, using a table manipulation task with four target positions. Still, the recognition was only on hand movements without considering the object to be addressed. Huang et al. [54] showed that the characteristics of gaze cues, especially duration and frequency, only focused on using gaze cues to predict customer intentions without considering environmental conditions. Shi et al. [55] also experimented with pointing at some objects on the table and recognizing intentions by looking at the Earth Moving Distance (EMD). The results still showed a gaze drift error and could not be applied in real-time.

4.3. Experimental results of handover behaviors

This subsection shows experimental results of handover behaviors. Figure 14. show a simulation result of grasping and handover behaviors; (a) the robot partner detected a person, (b) the robot partner confirmed the facial direction and hand gesture, this is indicated by the density of the topological map

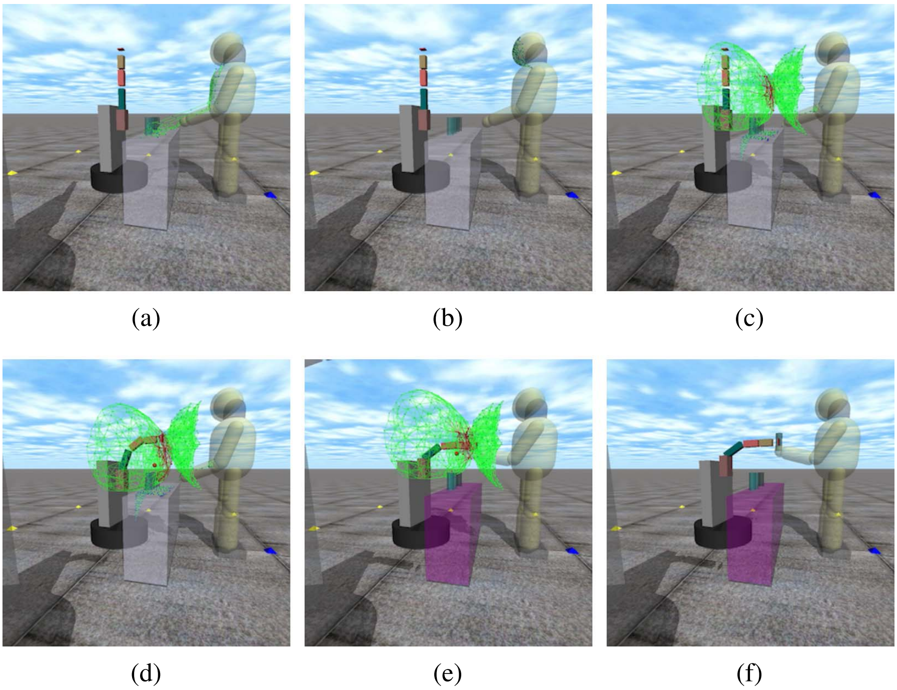
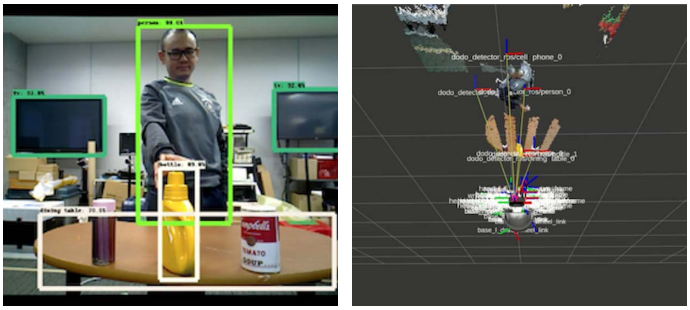


Figure 14: Simulation handover behavior.

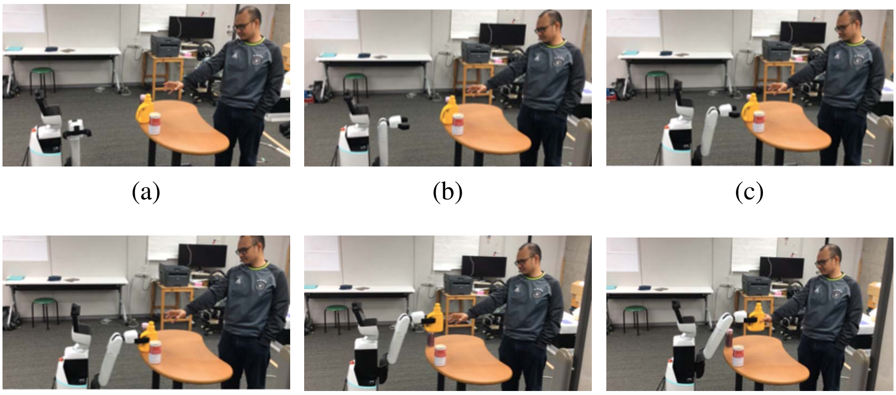
on the human head, (c) the robot partner estimated and shared human cognitive environment (gazing at the human hand and objects). The red nodes and lines show the area where the effectivity of humans and robot intersects mutually or the possible handover area. The robot partner estimated human intention and decided to grasp the second object, (d) the robot partner moved the arm to the second object, and (e,f) the robot partner gave it to the person, the table, a change in the table color indicates a handover behavior condition.

In the real environment, we performed grasping and handover experiments using the Toyota HSR. We used HSR mobile and manipulator movement to achieve the grasping and handover behavior. The object detection framework was run in Python with Tensorflow API and on an Intel i7 with GPU. Figure 15a shows an experimental result of human and object detection to start a handover behavior. To solve the inverse kinematics for the grasping and handover behavior, HSR uses the detected object and human position concerning the real-world situation based on the ROS geometry transformation calculation shown in Figure 15b. The sequential movement


person: 99.0%
tv: 92.8%
tv: 92.8%
bottle: 89.6%
dining table: 20.0%
dodo_detector_ros/cell_phone_0
dodo_detector_ros/person_0
dodo_detector_ros/dining_table_0
base_l_drive_wheel_link

(a) (b)

Figure 15: 2D to 3D object and human transformation and distance visualization in Rviz (ROS Visualization).



(a) (b) (c)

(d) (e) (f)

Figure 16: HSR grasping and handover sequential movement.

on grasping and handover behavior is shown in Figure 16. (a) HSR estimated the position of both the target person and the object nearest to the extracted human hand. (b,c) HSR performed moving forward for the grasping task, (d) HSR grasped the object, (e) HSR performed the post-grasping movement to deliver the object based on the central position of the target person in Figure 16f. In addition, we analyzed the success rate of robots in picking up and handing objects to humans. The success of handover an object is judged by whether humans can reach the object without moving the body; only the hands move. Each experiment was carried out 20 times for each object so that a total of 60 times by taking objects at random. The retrieval success rate is about 86.6%. An error occurred in converting the point cloud object with frame geometry to a gripped position. The grasping position is calculated from the center point of the bounding box object and the average value of the 8 points around the center point. The handover success rate is 90%, errors occur when estimating the human position, and the robot arm hits the table. Arm collision with the table occurs because it ignores the position of the edge of the table with the arm.

5. Summary

This paper proposed a human-robot interaction method based on a mutual cognitive environment for handover behaviors. First, we discussed how to share the cognitive environment between a person and robot partner based on relevance theory discussed in Cognitive Pragmatics. Next, we applied topological mapping methods of growing cell structure (GCS) and growing neural gas (GNG) based on the concept of perceiving-acing cycle discussed in Ecological Psychology. The growing cell structure is used to represent the effectivity in the action system. The essence of GCS is in the unsupervised learning of continuously regular topological structure according to the distribution of input data. Therefore, the GCS can be applied to the smooth control of a robot manipulator and human arm. On the other hand, GNG is used to visualize the human and robot task space. The essence of GNG is in the features of clustering and topological mapping. The clustering realizes the segmentation into several different objects, while the topological mapping realizes the surface feature of objects. Furthermore, we propose adaptive GNG using distance-based node update and error-based node deletion to improve online stability and adaptability in a dynamic environment. In this way, we applied two different topological mapping methods to the perceptual system and action system in this paper. We showed the effectiveness of the proposed method through several simulation results in the learning of effectivity and change of attention in human-robot interaction. Finally, we conducted a real robot application of handover behavior by HSR. In future work, we intend to discuss the online learnability of GCS and GNG based on the perceiving-acting cycle in other human interaction tasks. Furthermore, we integrate the proposed method with gesture recognition methods to realize different types of non-verbal communication.

Acknowledgment

This work was partially supported by JST [Moonshot RnD][Grant Number JPMJMS2034]

References

[1] S. Saleh and K. Berns, “Nonverbal communication with a humanoid robot via head gestures,” in *Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, 2015, pp. 1–8.

[2] A. Aly and A. Tapus, “Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction,” *Autonomous Robots*, vol. 40, no. 2, pp. 193–209, 2016.

[3] R. W. Gibbs Jr, “Mutual knowledge and the psychology of conversational inference,” *Journal of Pragmatics*, vol. 11, no. 5, pp. 561–588, 1987.

[4] M. F. Young, S. A. Barab, and S. Garrett, “Agent as detector: An ecological psychology perspective on learning by perceiving-acting systems,” *Theoretical Foundations of Learning Environments*, pp. 147–173, 2000.

[5] B. Fritzke *et al.*, “A growing neural gas network learns topologies,” *Advances in Neural Information Processing Systems*, vol. 7, pp. 625–632, 1995.

[6] Y. Prudent and A. Ennaji, “An incremental growing neural gas learns topologies,” in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*, vol. 2. IEEE, 2005, pp. 1211–1216.

[7] D. Fišer, J. Faigl, and M. Kulich, “Growing neural gas efficiently,” *Neurocomputing*, vol. 104, pp. 72–82, 2013.

[8] A. A. Saputra, C. W. Hong, and N. Kubota, “Real-time grasp affordance detection of unknown object for robot-human interaction,” in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 3093–3098.

[9] Y. Toda, Z. Ju, H. Yu, N. Takesue, K. Wada, and N. Kubota, “Real-time 3d point cloud segmentation using growing neural gas with utility,” in *2016 9th International Conference on Human System Interactions (HSI)*. IEEE, 2016, pp. 418–422.

[10] B. Fritzke, “Unsupervised clustering with growing cell structures,” in *IJCNN-91-Seattle International Joint Conference on Neural Networks*, vol. 2. IEEE, 1991, pp. 531–536.

[11] S. Lemaignan and P. Dillenbourg, "Mutual modelling in robotics: Inspirations for the next steps," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 303–310.

[12] P. F. Dominey and F. Warneken, "The basis of shared intentions in human and robot cognition," *New Ideas in Psychology*, vol. 29, no. 3, pp. 260–274, 2011.

[13] A. Jacq, W. Johal, P. Dillenbourg, and A. Paiva, "Cognitive architecture for mutual modelling," *arXiv preprint arXiv:1602.06703*, 2016.

[14] F. Stulp, J. Grizou, B. Busch, and M. Lopes, "Facilitating intention prediction for humans by optimizing robot motions," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 1249–1255.

[15] D. Vasquez, T. Fraichard, O. Aycard, and C. Laugier, "Intentional motion on-line learning and prediction," *Machine Vision and Applications*, vol. 19, no. 5, pp. 411–425, 2008.

[16] D. Vasquez, T. Fraichard, and C. Laugier, "Growing hidden markov models: An incremental tool for learning and predicting human and vehicle motion," *The International Journal of Robotics Research*, vol. 28, no. 11-12, pp. 1486–1506, 2009.

[17] Y. Maeda, A. Takahashi, T. Hara, and T. Arai, "Human-robot cooperation with mechanical interaction based on rhythm entrainment-realization of cooperative rope turning," in *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, vol. 4. IEEE, 2001, pp. 3477–3482.

[18] Y. Yamada, Y. Umetani, H. Daitoh, and T. Sakai, "Construction of a human/robot coexistence system based on a model of human will-intention and desire," in *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C)*, vol. 4. IEEE, 1999, pp. 2861–2867.

[19] Z. Z. Bien, J.-B. Kim, D.-J. Kim, J.-S. Han, and J.-H. Do, "Soft computing based emotion/intention reading for service robot," in *AFSS International Conference on Fuzzy Systems*. Springer, 2002, pp. 121–128.

[20] V. J. Traver, A. P. Del Pobil, and M. Pérez-Francisco, "Making service robots human-safe," in *Proceedings. 2000 IEEE/RSJ International*

Conference on Intelligent Robots and Systems (IROS 2000)(Cat. No. 00CH37113), vol. 1. IEEE, 2000, pp. 696–701. MR1797421

[21] Y. Matsumoto, J. Heinzmann, and A. Zelinsky, "The essential components of human-friendly robot systems," in *International Conference on Field and Service Robotics*, 1999, pp. 43–51.

[22] W.-K. Song, D.-J. Kim, J.-S. Kim, and Z. Bien, "Visual servoing for a user's mouth with effective intention reading in a wheelchair-based robotic arm," in *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, vol. 4. IEEE, 2001, pp. 3662–3667.

[23] Y. Xiao, Z. Zhang, A. Beck, J. Yuan, and D. Thalmann, "Human–robot interaction by understanding upper body gestures," *Presence*, vol. 23, no. 2, pp. 133–154, 2014.

[24] P. Pławiak, T. Sośnicki, M. Niedźwiecki, Z. Tabor, and K. Rzecki, "Hand body language gesture recognition based on signals from specialized glove and machine learning algorithms," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1104–1113, 2016.

[25] N. Kubota and H. Masuta, "Action learning of a mobile robot based on perceiving-acting cycle," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, vol. 2. IEEE, 2003, pp. 1222–1227.

[26] H. Masuta and N. Kubota, "The intelligent control based on perceiving-acting cycle by using 3d-range camera," in *2009 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2009, pp. 929–934.

[27] A. A. Saputra, J. Botzheim, A. J. Ijspeert, and N. Kubota, "Combining reflexes and external sensory information in a neuromusculoskeletal model to control a quadruped robot," *IEEE Transactions on Cybernetics*, pp. 1–14, 2021.

[28] B. Mantel, P. Hoppenot, and E. Colle, "Perceiving for acting with teleoperated robots: ecological principles to human–robot interaction design," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 42, no. 6, pp. 1460–1475, 2012.

[29] K. Nishiwaki, T. Sugihara, S. Kagami, F. Kanehiro, M. Inaba, and H. Inoue, "Design and development of research platform for perception-action integration in humanoid robot: H6," in *Proceedings. 2000*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000) (Cat. No.00CH37113), vol. 3, 2000, pp. 1559–1564.

[30] D. Toal, C. Flanagan, C. Jones, and B. Strunz, "Subsumption architecture for the control of robots," *IMC-13, Limerick*, 1996.

[31] M. A. Arbib, "Perceptual structures and distributed motor control," *Comprehensive Physiology*, pp. 1449–1480, 2011.

[32] E. Şahin, M. Cakmak, M. R. Doğar, E. Uğur, and G. Üçoluk, "To afford or not to afford: A new formalization of affordances toward affordance-based robot control," *Adaptive Behavior*, vol. 15, no. 4, pp. 447–472, 2007.

[33] B. Moldovan, P. Moreno, M. Van Otterlo, J. Santos-Victor, and L. De Raedt, "Learning relational affordance models for robots in multi-object manipulation tasks," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 4373–4378.

[34] M. Lopes, F. S. Melo, and L. Montesano, "Affordance-based imitation learning in robots," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, pp. 1015–1021.

[35] A. A. Saputra, Y. Toda, N. Takesue, and N. Kubota, "A Novel Capabilities of Quadruped Robot Moving through Vertical Ladder without Handrail Support," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov 2019, pp. 1448–1453.

[36] A. A. Saputra, W. H. Chin, Y. Toda, N. Takesue, and N. Kubota, "Dynamic density topological structure generation for real-time ladder affordance detection," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Nov 2019, pp. 3439–3444.

[37] M. Yani, A. R. A. Besari, N. Yamada, and N. Kubota, "Ecological-inspired system design for safety manipulation strategy in home-care robot," in *2020 International Symposium on Community-centric Systems (CcS)*. IEEE, 2020, pp. 1–6.

[38] A. Richert, M. Shehadeh, F. Willicks, and S. Jeschke, "Digital transformation of engineering education-empirical insights from virtual worlds and human-robot-collaboration," 2016.

[39] S. M. Rahman, "Cognitive cyber-physical system (c-cps) for human-robot collaborative manufacturing," in *2019 14th Annual Conference System of Systems Engineering (SoSE)*. IEEE, 2019, pp. 125–130.

[40] A. A. Malik and A. Brem, “Digital twins for collaborative robots: A case study in human-robot interaction,” *Robotics and Computer-Integrated Manufacturing*, vol. 68, p. 102092, 2021.

[41] F. Fraundorfer, C. Engels, and D. Nistér, “Topological mapping, localization and navigation using image collections,” in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, pp. 3872–3877.

[42] Y. Ding, W. Xu, Z. Liu, Z. Zhou, and D. T. Pham, “Robotic task oriented knowledge graph for human-robot collaboration in disassembly,” *Procedia CIRP*, vol. 83, pp. 105–110, 2019.

[43] Q. Li, F. Gama, A. Ribeiro, and A. Prorok, “Graph neural networks for decentralized multi-robot path planning,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 11 785–11 792.

[44] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, “Understanding and sharing intentions: The origins of cultural cognition,” *Behavioral and Brain Sciences*, vol. 28, no. 5, pp. 675–691, 2005.

[45] N. Kubota, “Computational intelligence for structured learning of a partner robot based on imitation,” *Information Sciences*, vol. 171, no. 4, pp. 403–429, 2005.

[46] N. F. Duarte, M. Raković, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor, “Action anticipation: Reading the intentions of humans and robots,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4132–4139, 2018.

[47] C. Bassetti, “Social interaction in temporary gatherings: A sociological taxonomy of groups and crowds for computer vision practitioners,” in *Group and Crowd Behavior for Computer Vision*. Elsevier, 2017, pp. 15–28.

[48] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *CVPR*, 2017.

[49] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, “Hand keypoint detection in single images using multiview bootstrapping,” in *CVPR*, 2017.

[50] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in *CVPR*, 2016.

[51] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE*

Conference on Computer Vision and Pattern recognition, 2016, pp. 779–788.

[52] A. A. Saputra, W. H. Chin, Y. Toda, N. Takesue, and N. Kubota, "Dynamic density topological structure generation for real-time ladder affordance detection," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3439–3444.

[53] R. C. Luo and L. Mai, "Human intention inference and on-line human hand motion prediction for human-robot collaboration," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5958–5964.

[54] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration," *Frontiers in Psychology*, vol. 6, p. 1049, 2015.

[55] L. Shi, C. Copot, and S. Vanlanduit, "Gazeemd: Detecting visual intention in gaze-based human-robot interaction," *Robotics*, vol. 10, no. 2, p. 68, 2021.

Fernando Ardilla
Graduate School of System Design
Tokyo Metropolitan University
6-6 Asahigaoka, Hino
Tokyo 191-0065, Japan
E-mail address: fernando-ardilla@ed.tmu.ac.jp
Department of Information and Computer Engineering
Politeknik Elektronika Negeri Surabaya
Surabaya, Indonesia
E-mail address: nando@pens.ac.id

Mohamad Yani
Graduate School of System Design
Tokyo Metropolitan University
6-6 Asahigaoka, Hino
Tokyo 191-0065, Japan
E-mail address: mohamad-yani@ed.tmu.ac.jp
Department of Computer Engineering
Faculty of Electrical Engineering
Institut Teknologi Telkom Surabaya
Surabaya, Indonesia

Azhar Aulia Saputra
Graduate School of System Design
Tokyo Metropolitan University
6-6 Asahigaoka, Hino
Tokyo 191-0065, Japan
E-mail address: azhar.aulia.s@gmail.com

Weihong Chin
Graduate School of System Design
Tokyo Metropolitan University
6-6 Asahigaoka, Hino
Tokyo 191-0065, Japan
E-mail address: weihong@tmu.ac.jp

Naoyuki Kubota
Graduate School of System Design
Tokyo Metropolitan University
6-6 Asahigaoka, Hino
Tokyo 191-0065, Japan
E-mail address: kubota@tmu.ac.jp

Received January 13, 2021