An Ecologically Valid Reference Frame for Perspective Invariant Action Recognition

Berkay Bayram
Dept. of Computer Science
Ozyegin University
Istanbul, Turkey
berkay.bayram@ozu.edu.tr

Emre Ugur
Dept. of Computer Science
Bogazici University
Istanbul, Turkey
emre.ugur@boun.edu.tr

Minoru Asada
OTRI/SISREC
Osaka University
Osaka, Japan
asada@otri.osaka-u.ac.jp

Erhan Oztop[1,2]
[2]*Ozyegin University*
[1]*OTRI/SISREC, Osaka University*
Osaka, Japan
erhan.oztop@otri.osaka-u.ac.jp

***Abstract*—In robotics, objects and body parts can be represented in various coordinate frames to ease computation. In biological systems, body or body part centered coordinate frames have been proposed as possible reference frames that the brain uses for interacting with the environment. Coordinate transformations are standard tools in robotics and can facilitate perspective invariant action recognition and action prediction based on observed actions of other agents. Although it is known that human adults can do explicit coordinate transformations, it is not clear whether this capability is used for recognizing and understanding the actions of others. Mirror neurons, found in the ventral premotor cortex of macaque monkeys, seem to undertake action understanding in a perspective invariant way, which may rely on lower level perceptual mechanisms. To this end, in this paper we propose a novel reference frame that is ecologically plausible and can sustain basic action understanding and mirror function. We demonstrate the potential of this representation by simulation of an upper body humanoid robot with an action repertoire consisting of push, poke, move-away and bring-to-mouth actions. The simulation experiments indicate that the representation is suitable for action recognition and effect prediction in a perspective invariant way, and thus can be deployed as an artificial mirror system for robotic applications.**

***Index Terms*—Mirror Neurons, Perspective Invariant Action Recognition, Robotics, Reference Frame**

I. Introduction

How infants learn to develop the skill to detect equivalence (parity) between their own action and the observed ones is still an open scientific question [1], which is interesting not only for biological sciences but also for robotics. One idea proposed for the development of this skill is that infants first learn how to perform 'coordinate transformation', which is the skill of rotation and translating of a 3D object so as to predict how it would look from the infant's own perspective. In robotics this is a common operation; one can obtain the self-perspective 3D pose of an object given in any arbitrary coordinate frame by using a straightforward transformation. Although adult humans seem to have this ability as an explicit skill (usually called metal rotation) [2], it is unknown how this develops and whether it is directly related to action understanding. To be concrete, it is unknown whether it is the precursor of perspective invariant action understanding although often it is accepted as such.

Animals necessarily are aware of the effects of gravity and thus it is reasonable to postulate that one of the axes they would use to assess action and effect is formed by the direction of gravity. In fact, it is shown that monkey brain uses gravity direction dependent representations of object orientations [3]. For primates who are equipped with dexterous hand use ability, it is critical to monitor moving hands for error correction in the case of self execution and for predicting others' action goal for appropriate social behavior. This intuition is supported by neurophysiological findings showing that a special brain area in the superior temporal cortex is evolved for hand movement detection regardless of the actor [4]. Therefore it makes sense that the prediction of the effect of an action be defined with respect to the movement of the hand.

Combining these two pieces of information, in this paper we propose a novel reference frame that is ecologically plausible, and present our results on its suitability to sustain action understanding and mirror neuron function. We call this the Action Reference Frame or Action Frame (AF) in short. We further propose that predictive learning can use AF to represent the data generated by self-action and self-observation so that generalization to others is possible, without rejecting the possibility of parallel predictive systems that employ other reference frames. To complete the reference frame definition, it is also necessary to state the origin of the AF. Two main possibilities exist: either AF is placed on the moving hand or on the object that is the target of the action. Although, in the multiple object scenarios, the latter might bring ambiguity during the initial portions of an action, in the current report we took this approach due the more straightforward analysis it allows.

In addition to the biological relevance of this idea, we are interested in implementing an action recognition capability for a self-learning robotic system. To this end, all of the work is implemented on an upper body humanoid robot, TOROBO[1] simulation and the processing steps are kept at a feasible level for physical robot deployment. Also to show that the results obtained are not due to the high fidelity information available from the simulator, a depth camera based color coded object and hand detection system is integrated into the prediction network. Overall, the results indicate that the proposed AF based prediction system can undertake action recognition and

[1]Tokyo Robotics, https://robotics.tokyo/products/torobo/

Manuscript 63 submitted to 2021 11th IEEE International Conference on Development and Learning (11th ICDL-EpiRob). Received March 17, 2021.

effect prediction in a perspective invariant manner. Thus, it may serve as a mirror neuron system architecture that is amenable to robotic implementation.

The paper is organized as follows. The next section briefly describes the related work in the literature. Then, the construction of AF, the simulation environment, robot actions and learning details are given. Finally, the results of simulation experiments and conclusions with a brief discussion on limitations and future directions are given.

II. Related Work

Action recognition is a widely studied area which attracts interest from a wide range of fields including computer vision and robotics [5], [6], as it facilitates human behaviour analysis [7] and human-robot interaction [8].

In computer vision, action recognition, in particular, pose estimation is a well studied topic with approaches including monocular [9] or stereo-vision, and depth camera based point cloud approaches [10]. These approaches can either be based on manually designed feature matching with pose search or on gradient based learning. For example, Keskin et al. [11] estimated hand positions with high accuracy using pose estimators that exploit multi-layered randomized decision forests.

Keeping this in mind, multiple studies of the first person (egocentric) perspective human action recognition studies have been conducted. In [12] authors used Convolutional Neural Network (CNN) [13] architecture to predict executed actions which are learned from hand motion cues. Ma et al. [14] also utilized a CNN architecture to learn scene and hand motion information. Garcia-Hernando et al. [15] used RGB-D images to train multiple state of the art recurrent neural networks with a large corpus of video and mo-cap data to recognize different hand actions such as tearing, flipping and pouring.

In some robotic applications it may be possible to apply state estimation techniques to map observed behavior into a sequence of state descriptions compatible with the observer. This allows behavior understanding through estimating the value function of the demonstrator with the aid of adopting a reinforcement learning framework as exemplified by Takahashi et al. [16]. In general, in such scenarios, a range of methods from dynamic time warping (e.g. [17]) to inverse reinforcement learning can be applied (see [18]). In our work, we consider raw perceptual input and do not assume the existence of any state estimation capability.

Research in neuroscience has shown that primate brains are endowed with multi-modal neurons (mirror neurons) that become active during the execution of hand actions as well during the observation of a similar actions when executed by a conspecific or experimenter [19]–[21]. It is generally accepted that mirror neurons encode goal directed actions and play a significant role in understanding of observed action of others' [22]. However the underlying mechanisms and representations are still far from clear [23], [24]. Several mirror neuron models exist in the literature which may be considered as biologically realistic action recognition models (e.g. [25]–[28]). For example Oztop and Arbib [25] achieves perspective invariance by defining a 'hand state' that describes the relation of hand with respect to an object, which may be considered as a special case (as it focuses only grasping and not other actions) of the proposal pursued in this paper.

Besides the existence of mirror neuron mechanism for perspective invariant action understanding, additional line of research lends strong support to the proposed Action Frame concept. Experimental evidence shows that the brain uses multiple spatial representation and reference frames for action production [29] and location recognition [30]. These reference frames are used to encode spatial information with respect to a collection of reference frames, including the egocentric ones such as head, eyes, hand and body, as well as the allocentric ones such as the position of an object in attention. Finally, a more specific support to the proposed Action Frame is provided by the recent findings showing that the gravity direction is well incorporated in the parietal representation of object orientations [3] suggesting that reference frames incorporating the gravity direction exists in the brain.

III. Method

With the goal of realizing an action recognition system based on the proposed action frame and assessing its feasibility for perspective invariant recognition, we designed a robotic simulation setup. The following subsections describe this setup and the experiments conducted throughout this study.

A. Simulation Environment and Task Setup

We used the Gazebo simulator [31] for the dynamic simulation of the TOROBO robot and Robot Operating System (ROS) [32] robot control architecture for communicating with the robot. TOROBO is an upper body humanoid robot which has bi-manual manipulation capability with 22 degrees of freedom. In this study, we used only single hand manipulation and considered a single object, i.e. a cylinder with radius of 3cm and height of 13cm that the robot can interact with. For modeling the case of action observation from different perspectives, we assumed that there are virtual observers around the table that might be observing the actor. The perception of those observers were either emulated by direct access to the simulator data (using an appropriate homogeneous transformation matrix to calculate observer view points), or obtained by simulated depth cameras around the table. The latter aimed at assessing the applicability of AF based action recognition in physical robotic setups.

B. Action Repertoire of the Robot

To test the development of AF-based action recognition capability based on self-observation, we focused on four predefined parameterized actions. Out of the four, the two were simple actions of *push* and *poke*, and the other two were relatively more complex actions of *move-away* and *bring-to-mouth*. The trajectory and outcome of each action type were determined by two parameters: the angle of the robot gripper with respect to the object (action angle) and the location of the object prior to interaction. The execution trajectories of

Manuscript 63 submitted to 2021 11th IEEE International Conference on Development and Learning (11th ICDL-EpiRob). Received March 17, 2021.

these actions were assumed to start from a fixed initial robot configuration. Given the action type and its parameters, the

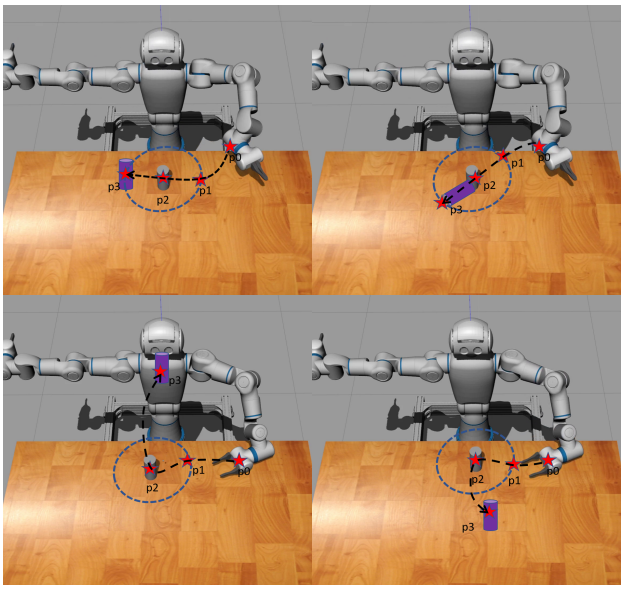

Fig. 1. Table top simulation environment used is shown. Top-left: Push, Top-right: Poke, Bottom-left: Bring-to-Mouth, Bottom-right: Move Away.

joint trajectory to be followed by the robot is constructed and then communicated to the robot through ROS. First, four Cartesian space via-points are determined in an action specific way, which are then converted into joint angles by using the inverse kinematics of the left arm of the robot. Then the obtained set of joint angles are fitted with cubic splines to obtain smooth continuous curves for each joint of the robot. Finally, these curves are sampled at 20 equal intervals and fed to ROS to drive the robot arm and gripper. For all the actions considered, the robot is assumed to start its action at the same initial joint configuration determining the first via-point (P0); the second via point (P1) is determined by the angle parameter, which specifies a point on an imaginary circle (r=6cm) centered on the object. The next via-point (P2) is taken to be the center of the object, and the final via-point (P3) is determined according to the action type as described below.

1) ***Push and Poke****:* For these simpler actions, the object is assumed to be at a fixed location in front of the robot (see Fig 1). The via-points for push and poke are formed by the initial gripper position (P0), and two symmetrical points (P1, P3) on the imagery circle centered at the mid-point of the object, which serves also as the middle via-point (P2). The only difference between the push and poke is that the vertical (z) coordinates of the P1, P2 and P3 are set differently. The z-coordinate is taken as 6.5cm for the push and 11.5cm for the poke actions. Since the object has a height of 13cm, the push action stably translates the object while the higher contact with the poke action often knocks over the object. While executing these actions the gripper is kept fully closed and the side of the gripper is used to form the contact with the object. This facilitates a more robust contact and generates repeatable effects compared to using the tip of the gripper for establishing contact.

2) ***Move-away and Bring-to-mouth****:* For these complex actions, the goals of the actions are taken as specific locations that the object must be brought. In the case of bring-to-mouth action, the target of action is a fixed point in the proximity of the facial area of the robot, which determines the last via-point (P3). Likewise, in the case of move-away action the P3 is a fixed point which is away from the robot, near the boundary of the workspace of the robot (see Fig 2). As in the push and poke actions, the first via point (P1) is determined by the angle parameter of the move-away and bring-to-mouth actions. Similarly, the second via-point P2 is taken as the mid-point of the object. To enable grasping and transportation of the object to the desired target location, the gripper is commanded to enclose when the robot hand reaches P2.

C. Data Generation and Collection

As discussed in the previous section each action takes two parameters: action angle and position of the target object. For the Push and Poke actions, action angle is sampled uniformly at random within a range of [-75, -105], and object position is taken as a fixed position in front of the robot ([0.47, 0, 1.08]). The angle parameters for the Move-away and Bring-to-mouth actions are sampled within the range of [-35, -110], and the object position parameter is sampled uniformly from the interior of a circle (r = 7cm) that is parallel to the table and centered at [0.47, 0, 1.08] in world coordinates.

The simulated robot executed each action in 1000 different settings. While executing the actions, the robot is commanded through ROS and the desired joint angles are send to the robot at 20Hz for one second (i.e. actions are assumed to take one second to complete). Data collection is carried out using the same interval, and the final object position is recorded after the action ends. The arm and torso joint angles are also recorded to be able to recreate the results of the experiments later.

In order to overcome computational limitations in synchronized data collection, observer experience (i.e. positions with respect to the Egocentric Frame of each observer) are calculated after the action is completed by using the data of the actor via appropriate homogeneous coordinate transformations.

D. Object and Hand Localization via Emulated Depth Camera

Since we plan to deploy the developed perspective invariant action recognition system in the real world, and explore its possible use cases in robotics as a functional mirror neuron architecture, we designed a simple color-based perception system so that the robot can 'see' and track the target object and the hand in action. This way, it would be easier to transfer the learning and prediction to the real world. During action execution in the simulator, we also performed data collection of hand and object positions by using the emulated Kinect depth cameras. In order to gather this data, the depth camera outputs are processed with the help of OpenCV library [33]. For computational convenience, the processing for object detection is based on 3D color segmentation. Consequently,

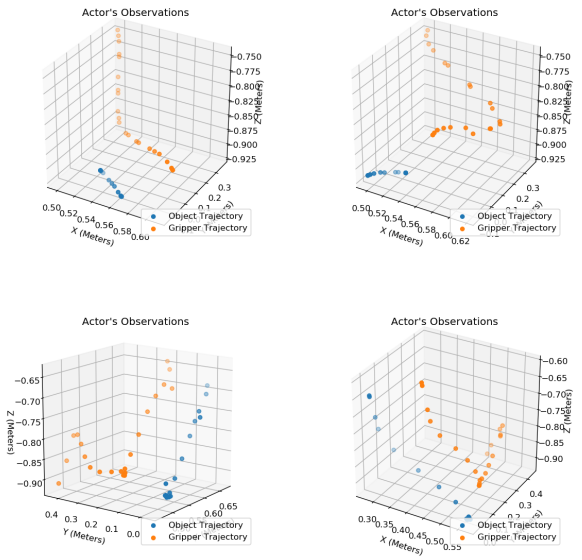

Fig. 2. Sample object and gripper trajectories for each action. Top-left: Push, Top-right: Poke, Bottom-left: Bring-to-Mouth, Bottom-right: Move Away.

the object and the gripper are given distinct colors, red and green, respectively. Given an image frame for color filtering is applied using the 'inRange' method of OpenCV which gives a set of points for each color. The centers of the extracted point clouds are then found by using the 'findContours' and 'moments' methods. In order to accelerate the computations, both image and point cloud data are down-sampled by 8. One limitation of using object detection in this fashion is that view-dependent occlusions may create offsets in the point-cloud centers corresponding to the gripper and the target object.

E. Action Coordinate Frame (AF) Construction

The AF is constructed based on the Gravity vector (g) and the velocity vector (v) pertaining to a hand in action. The velocity vector is in general a function of time over the action period. In this study, we take the velocity vector of the hand at the moment when it enters the vicinity, i.e. 20cm proximity of the target object. The velocity vector is estimated by numerical differentiation, and its projection to the horizontal plane (v_{proj}) is used for setting up the Action Reference Frame (AF). Fig 3 shows the AF overlaid on the initial position of the object. In detail, AF is calculated as follows: $x_{axis} = v_{proj}/||v_{proj}||$, $z_{axis} = -g = [0, 0, 1]$, $y_{axis} = x_{axis} \times z_{axis}$.

F. Predictive Learning Network

To model a predictive system that can be trained by self-observation, we formalized the problem as learning to predict the action-code, action-parameters and the effects given an object and hand in action. In particular, the input for the predictive system is taken as 3D hand positions of 5 consecutive frames represented in either EF or AF. The output, on the other hand, corresponds to the effect that would be generated, the

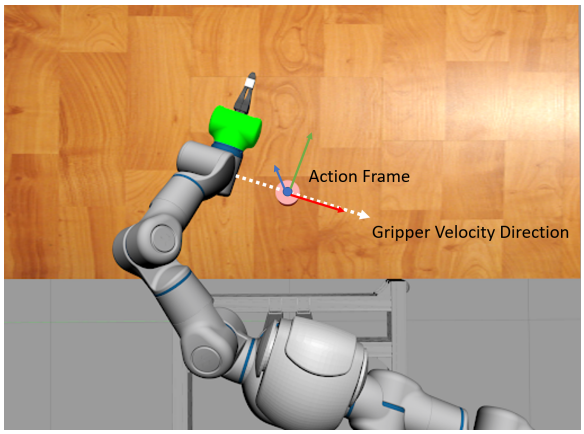

Fig. 3. Illustration of the Action Frame (AF). Gravity and hand velocity vectors are used to construct the AF (red:x-axis; green:y-axis; blue:z-axis).

action type and action parameters corresponding to the action being observed. Thus, the size of input to the neural network is 15 (5 positions) and the size of the output size is 11 (the effect encoded as a 3D offset vector (3), the one-hot action code (4), and the action parameters encoded with the action angle (1) and the initial position of the object (3)). The system starts storing hand positions after the hand approaches to the vicinity of the target object (enters within 20cm range of the object), and after 5 observations are done, the system produces its prediction.

With this input-output specification the prediction system is implemented with a three layer fully connected Artificial Neural Network (ANN) with 16 neurons in each layer. We used *rectified linear units* (Relu) [34] for the network activations. The size of the network is empirically tuned to be small yet capable of learning the prediction problem targeted.

The training and testing data is scaled between 0 and 1 using a min-max scaler. For every experiment the network is trained 4000 epochs with learning rate 1e-3, batch size of 64 and the same random seed is used in training to exclude randomness. Finally, ADAM-optimizer [35] is used. No regularization technique is used since the networks are already pretty shallow.

G. Experiments conducted

By using self-observation data, we trained two separate networks: one that represents the data in the Action Frame (AF), and one that uses the Egocentric Frame (EF) representation. After we ensured that the predictions with self-collected data and both representations are successful (Results-A), we switched our attention to contrast the capabilities of AF- and EF-based predictions when they are used to make predictions of others' actions (Results-B). Note that, in general, an egocentric reference frame has an origin aligned with the observing agent. However, training a system with self-observation and then attempting to do prediction based on the observation of others would create large offsets in hand positions due to the simple fact that one's own hand is often

much closer than others' hands. Therefore, to improve the prediction capability of EF-based prediction, we translated the origin of EF to the object center, as we did for the case of AF.

Every action is sampled 1000 times as described in the Data Generation and Collection section. We used a %20 train test split on the data. The training set includes 800 randomly selected samples from 1000 samples generated during simulation for every action, therefore the size of the training set is 3200. Test set has 2400 samples gathered from each observer and the actor (12 observers in total). The network prediction error is calculated using *mean squared error* (MSE) loss function.

Finally, to test the performance of both EF- and AF-based predictions of others' actions, we repeated the latter experiment by using the emulated depth camera image.

IV. Results

In this section we present the predictive learning results based on self-observation learning by using representations in Egocentric and Action Frames. Then we present the results showing the generalization ability induced by these reference frames for observation actions of others. Finally, towards a real world implementation we present the learning and generalization results based on depth camera based object and hand perception.

A. AF and EF based learning of self-generated data

Figure 4 shows the RMSE error for predicted action angle and Euclidean distances of effect throughout the training process for both AF and EF networks. It is evident that both networks show a convergent learning regime with the loss approximately stabilizing towards the final epochs. So we can deduce that the networks designed are suitable for learning the data derived from our setup.

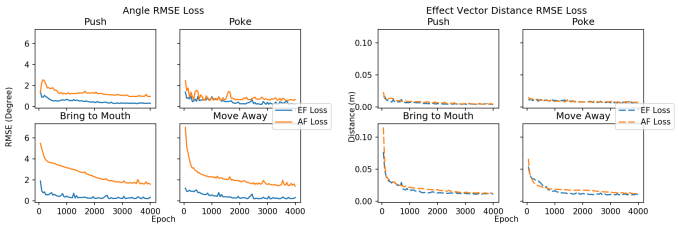

Fig. 4. Test RMSE loss for action angle and effect distance for both networks. AF is Action Frame and EF is Egocentric Frame losses.

B. Observing Others via AF- vs EF-based Prediction System

After training is completed on self observations, both AF and EF networks are tested on previously unseen data perceived by the observers. Each observer perceives the world through its eyes (or cameras), thus the eye-centered/egocentric representation is dependent on the pose of the observer. Note that in the stage we are considering, each observer can only learn from its own actions. Therefore, a change in the pose of the observer considerably affects the prediction capabilities of the observer's prediction if it is based on an Egocentric representation. For understanding the actions of others, additional mechanisms or different representations seem necessary. We took the latter alternative and proposed the Action Frame. When the actions are seen through the Action Frame, what the observer and the actor 'see' is very similar, and indeed in a noise-free simulation environment it is identical. In the real world, there would be perceptual noise, occlusions and distortions that would create imperfections. The results in this section, shows these arguments quantitatively.

The leftmost plot in Figure 5 shows the action prediction accuracy of the network trained with data using the EF representation. As can be seen in the graph, the more the observer wanders away from the viewpoint of the actor (i.e.position around the table), the worse the prediction accuracy gets. This is outcome expected since the observer has no experience related to the other viewpoints. Still, the generalization capability of the neural network generates somewhat correct predictions related to the observed action for neighbor observers (i.e. viewpoints); but, for most of the actions, the accuracy goes to zero when the observer is for example, directly opposite from the actor. In the same Figure, rightmost graphs show the action angle prediction error for each action as root-mean-square error (RMSE) with standard-deviation. Similar to the action recognition performance, the further the observer wanders away from the actor's perspective, the more the loss increases. Note that, for error calculation, only the action samples that were correctly predicted were used. The low angle errors and standard-deviation at viewpoints where action recognition error is high indicates that for those observers only a few actions can be recognized but when they are recognized their action parameters could be reliably retrieved.

When the network is trained with AF representations, we get perfect prediction accuracy since the network is trained with noise-free data and thus the observer experience is the same as the actor in AF representation. Similarly, the action angle error is close to zero. The similar contrast between AF and EF effect prediction can be seen in Figure 6.

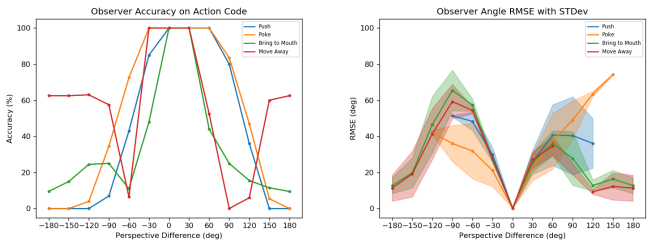

Fig. 5. Action understanding performance with EF based representations by using noise-free data. Left: recognition accuracy as a function of viewpoint difference. Right: Action parameter (angle) prediction accuracy for those actions recognized correctly. The shades indicate standard deviation.

C. Observing Others via AF- vs EF-based Prediction System using Depth Camera Input

We conducted the same experiments of the previous section by using object and hand position data obtained via the Kinect based perception system. Figures 7 and 8 show action recognition performance for EF and AF based learning respectively. As expected, the networks trained with the emulated

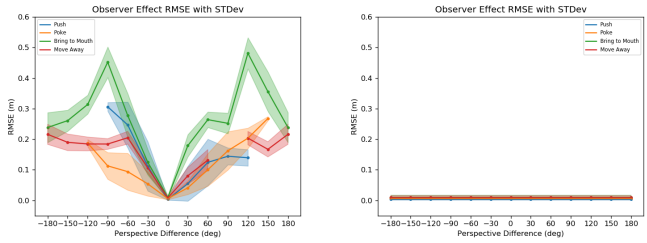

Fig. 6. Effect prediction performances with EF and AF based representations. Left: accuracy based on EF, Right: accuracy based on AF as a function of observer viewpoint difference. The shades indicate standard deviation.

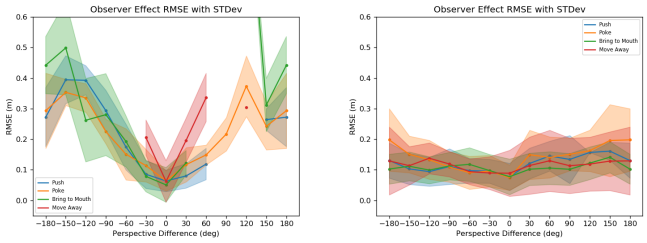

Fig. 9. Effect prediction performances with EF and AF based representations for emulated Kinect data. Left: accuracy with EF, Right: accuracy with AF as a function of viewpoint difference. The shades indicate standard deviation.

Kinect data show a poorer performance compared to the results obtained by using noise-free data from the simulator. Even though depth-sensing is itself a simulation, there are certain perceptual biases and occlusions depending on the viewpoint. The results from our experiments suggests that with non-perfect perception still a high level of perspective invariant action recognition capability can be obtained if we use AF-based representations. This observation is also valid for effect prediction as shown in Fig 9). It is worth noting that the asymmetric performance drop seen in Figure 7, when compared with the symmetric performance drop of Figure 5, indicates that the imperfections in the implemented hand and object perception system manifests itself differently for each action.

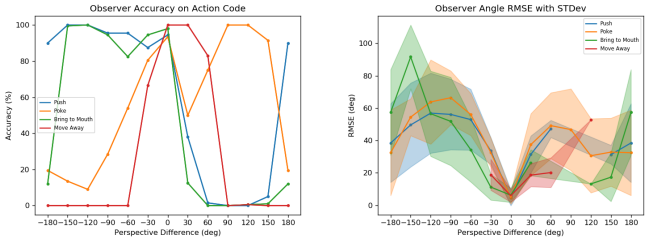

Fig. 7. Action understanding performance with AF based representations for emulated Kinect data. Left: recognition accuracy as a function of observer viewpoint difference. Right: action parameter (angle) prediction accuracy for those actions recognized correctly. The shades indicate standard deviation.

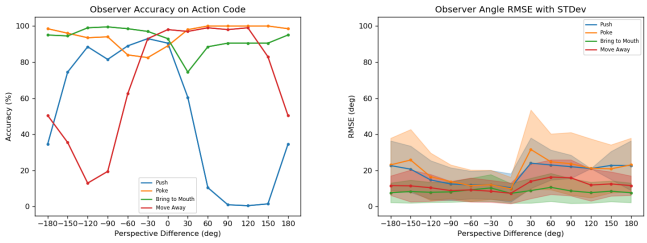

Fig. 8. Action understanding performance with EF based representations for emulated Kinect data. Left: recognition accuracy as a function of observer viewpoint difference. Right: action parameter (angle) prediction accuracy for those actions recognized correctly. The shades indicate standard deviation.

V. Conclusion and Discussion

Mirror neurons found in the ventral premotor cortex of primate brain encode goal directed actions in a multi-modal way, discharging both during the observation and execution of similar actions [21]. Most of these neurons show perspective invariant responses, thereby forming a basis for perspective invariant action understanding [36]. Although it is accepted that the brain employs several reference frames for interacting with the environment [3], [37], it is not clear whether the mirror neuron system is linked to a representation that use a particular reference frame. In this study, we proposed 'Action Frame' (AF) which is an ecologically and biologically plausible reference frame that represents action predictions and effects with respect to the gravity and the approach direction of a manipulator, i.e. hand. AF facilitates development and learning of perspective invariant action recognition based on self-observation. Since primates are endowed with special neural circuits for visually processing hands [38] and representing gravity [3], AF may form the basis for mirror neuron system development.

To show the efficacy of AF in action understanding, we conducted experiments in a simulation environment with a humanoid robot equipped with the actions of *push, poke, bring-to-mouth* and *move-away*. To be concrete, we trained a prediction network based on the robot self-observation of executed actions by using either AF or EF based representations. Then, the network was asked to make predictions for observations from different perspectives. As expected, the prediction system based on EF gave declining performance as the viewing angle deviates from the self-action view. In contrast, with AF, observed actions, their parameters and the effect that would be generated could be accurately predicted. To verify that the results obtained were not simply due to the noise-free simulator data, similar experiments were conducted by using emulated depth-cameras to serve as the 'eyes' of the actor and the observers. Thus, the training and testing results were subject to the imperfections of the proof-of-concept visual processing employed. Although the results were affected by the imperfections, capability of the perspective invariant action recognition system stayed at an acceptable level. This suggests that (1) The mirror neuron system may be the result of the development of a predictive system based on such a reference frame, and (2) development of a predictive capability based on self-observation can be readily realized as part of a real robotic system.

One limitation for the current system is that the tested en-

vironment uses few actions and a single object. To strengthen the results, the experiments should be expanded to include different objects with different poses, and more complex actions must be introduced. Although the current actions can generate infinitely many movements since the actions are parametric, the learning task for the prediction network is not demanding as the actions used can be differentiated with a low number of features. A more complex realistic environment would require more robust visual processing and more powerful prediction networks. With these in place, for physical robots, it would be possible to naturally interleave action learning with action understanding, i.e. mirror neuron development for emergent human-robot interaction.

ACKNOWLEDGMENT

This work is supported by the International Joint Research Promotion Program, Osaka University under the project "Developmentally and biologically realistic modeling of perspective invariant action understanding" and Ozyegin University. We thank Tokyo Robotics Inc. for technical assistance.

REFERENCES

[1] R. Peeters, L. Simone, K. Nelissen, M. Fabbri-Destro, W. Vanduffel, G. Rizzolatti, and G. A. Orban, "The representation of tool use in humans and monkeys: Common and uniquely human features," vol. 29, no. 37, pp. 11 523–11 539, 2009.

[2] D. I. Perrett, M. H. Harries, R. Bevan, S. Thomas, P. J. Benson, A. J. Mistlin, A. J. Chitty, J. K. Hietanen, and J. E. Ortega, "Frameworks of analysis for the neural representation of animate objects and actions," vol. 146, no. 1, pp. 87–113, 1989.

[3] A. Rosenberg and D. E. Angelaki, "Gravity influences the visual representation of object tilt in parietal cortex," vol. 34, no. 43, pp. 14 170–14 180, 2014.

[4] M. Oram and D. Perrett, "Responses of anterior superior temporal polysensory (stpa) neurons to "biological motion" stimuli," *Journal of cognitive neuroscience*, vol. 6, no. 2, pp. 99–116, 1994.

[5] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[6] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[7] N. Jaouedi, N. Boujnah, O. Htiwich, and M. S. Bouhlel, "Human action recognition to human behavior analysis," in *2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*. IEEE, 2016, pp. 263–266.

[8] Y. Ji, Y. Yang, F. Shen, H. T. Shen, and X. Li, "A survey of human action analysis in hri applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2114–2128, 2019.

[9] M. de La Gorce, D. J. Fleet, and N. Paragios, "Model-based 3d hand pose estimation from monocular video," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 9, pp. 1793–1805, 2011.

[10] C. Choi, S. Ho Yoon, C.-N. Chen, and K. Ramani, "Robust hand pose estimation during the interaction with an unknown object," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3123–3132.

[11] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Hand pose estimation and hand shape classification using multi-layered randomized decision forests," in *European Conference on Computer Vision*. Springer, 2012, pp. 852–863.

[12] S. Singh, C. Arora, and C. V. Jawahar, "First person action recognition using deep learned descriptors," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2620–2628.

[13] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, *Object Recognition with Gradient-Based Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 319–345.

[14] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," 2016.

[15] G. Garcia-Hernando, S. Yuan, S. Baek, and T. Kim, "First-person hand action benchmark with RGB-D videos and 3d hand pose annotations," *CoRR*, vol. abs/1704.02463, 2017.

[16] "Emulation and behavior understanding through shared values," *Robotics and Autonomous Systems*, vol. 58, no. 7, pp. 855–865, 2010, advances in Autonomous Robots for Service and Entertainment.

[17] S. Sempena, Nur Ulfa Maulidevi, and Peb Ruswono Aryan, "Human action recognition using dynamic time warping," in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, 2011, pp. 1–5.

[18] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," 2018.

[19] A. Compston, "Action recognition in the premotor cortex. By Vittorio Gallese, Luciano Fadiga, Leonardo Fogassi and Giacomo Rizzolatti. Brain 1996: 119; 593–609." *Brain*, vol. 132, no. 7, pp. 1685–1689, 06 2009.

[20] "Premotor cortex and the recognition of motor actions," *Cognitive Brain Research*, vol. 3, no. 2, pp. 131–141, 1996, mental representations of motor acts.

[21] G. Di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Understanding motor events: a neurophysiological study," *Experimental brain research*, vol. 91, no. 1, pp. 176–180, 1992.

[22] L. Bonini, M. Maranesi, A. Livi, L. Fogassi, and G. Rizzolatti, "Space-dependent representation of objects and other's action in monkey ventral premotor grasping neurons," vol. 34, no. 11, pp. 4108–4119, 2014.

[23] E. Oztop, M. Kawato, and M. Arbib, "Mirror neurons and imitation: a computationally guided review," *Neural Networks*, vol. 19, no. 3, pp. 254–71.

[24] E. Oztop, M. Kawato, and M. A. Arbib, "Mirror neurons: Functions, mechanisms and models," *Neuroscience Letters*, vol. 540, pp. 43–55.

[25] E. Oztop and M. A. Arbib, "Schema design and implementation of the grasp-related mirror neuron system," *Biological cybernetics*, vol. 87, no. 2, pp. 116–140, 2002.

[26] Y. Demiris* and M. Johnson†, "Distributed, predictive perception of actions: a biologically inspired robotics architecture for imitation and learning," *Connection Science*, vol. 15, no. 4, pp. 231–243.

[27] G. Metta, G. Sandini, L. Natale, L. Craighero, and L. Fadiga, "Understanding mirror neurons: A bio-robotic approach," *Interaction Studies*, vol. 7, no. 2, pp. 197–232.

[28] Y. Nagai, Y. Kawai, and M. Asada, "Emergence of mirror neuron system: Immature vision leads to self-other correspondence," in *2011 IEEE International Conference on Development and Learning (ICDL)*, vol. 2. IEEE, 2011, pp. 1–6.

[29] C. L. Colby, J.-R. Duhamel, and M. E. Goldberg, "Visual, presaccadic, and cognitive activation of single neurons in monkey lateral intraparietal area," *Journal of neurophysiology*, vol. 76, no. 5, pp. 2841–2852, 1996.

[30] "Heterogeneity of extrastriate visual areas and multiple parietal areas in the macaque monkey," *Neuropsychologia*, vol. 29, no. 6, pp. 517–537, 1991.

[31] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, vol. 3, 2004, pp. 2149–2154 vol.3.

[32] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system."

[33] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[34] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines." in *ICML*, J. Fürnkranz and T. Joachims, Eds. Omnipress, pp. 807–814.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[36] G. Rizzolatti, L. Fogassi, and V. Gallese, "Neurophysiological mechanisms underlying the understanding and imitation of action," *Nature Reviews Neuroscience*, vol. 2, no. 9, pp. 661–70.

[37] C. L. Colby, "Action-oriented spatial reference frames in cortex," *Neuron*, vol. 20, no. 1, pp. 15–24, 1998.

[38] "Neural representation for the perception of the intentionality of actions," *Brain and Cognition*, vol. 44, no. 2, pp. 280–302, 2000.