A Computational Model For Object-Directed Action Prediction Development

Serkan Bugur[1], Yukie Nagai[3], Erhan Oztop[2], and Emre Ugur[1]

[1]Bogazici University, Istanbul, Turkey. [2]Ozyegin University, Istanbul, Turkey.
[3]National Institute of Information and Communications Technology, Osaka, Japan.
Email: serkan.bugur@boun.edu.tr

***Abstract*—Recent studies in neuroscience revealed that a developmental correspondence emerges between predicting goals of others' actions and motor ability to produce the same actions in early infancy. This paper proposes a computational model to show the developmental learning of the action prediction capability. As the infant grows older, his/her experiences towards some actions grows as well. Therefore, an improvement in the action prediction capability of infants is expected. Our preliminary work results proposed computational model correlates the action experience with the action prediction capability of an infant in a developmental manner.**

***Index Terms*—Action Prediction, Autoencoder, CNN (Convolutional Neural Network), LSTM (Long short-term memory), t-SNE Factorization**

I. Introduction

Understanding and predicting the goals of others' actions plays a crucial role in human interaction. Starting from the first year of life, infants learn to predict the goals of others' actions swiftly and with high accuracy even though the actions performed quickly with no clear explanation. How do infants perceive the environment and associate goals to observed actions of others before completion? Which abilities infants should have learned in order for them to infer the goals of others' actions? Before starting, one caveat is that by "other's actions" we specifically refer to object-directed[1] reaching actions performed by an individual other than the observer.

Neuroscientific studies performed on humans and animals showed that action context consists of five broad categories which enables future action prediction: the experimenter performing the action, the observer, the environment, the target object and the object approach [1].

For the remaining of the paper, the observer refers to a model of an infant who watches the experimenter performing an action. Rajmohan et al. [4] stated that others' actions are understood through a direct matching process which is believed to be implemented by mirror neuron system (MNS). Mirror neurons that are core of MNS fire when both an individual performs a particular action and when that individual observes another individual performing the same or similar action. Sommerville et al. [2] supported the idea functioning of MNS by showing that only the infants with the capability of producing the observed action themselves, are capable of inferring the goal of the observed actions. Kanakogi et al. [3] further supported this idea by conducting an experiment and by examining the infant eye movements. They have divided the actions performed by the experimenters into three types. In the first type, infants are shown that an experimenter reaches for an object and grasps the object (GH - goal directed action). In the second one, again an experimenter reaches for an object. However this time the experimenter uses back of his/her hand and does not grasp the object (BH - non-goal directed action). In the third action type, experimenter uses a mechanical claw to reach for and grasp the object (MC - inanimate goal directed action). Neither group of the infants have prior experience with the BH and MC conditions. Therefore, those types of actions were not predicted by the infants. They also stated that four-month-old infants lack the ability to perform grasping actions, hence they did not make any predictive gazes towards the target object of the experimenter's reaching action. However, starting from the age of six-months, infants could perform one-handed grasping and were able to make predictive gazes towards the target object of the GH condition before completion of the reaching action. They concluded their experiment by saying that a developmental correspondence exists between action prediction and the ability to perform the same action. Figure 5a shows the results of their experiment. Copete et al. [5] implemented a computational model and showed that development of the ability to predict action goals is correlated with the development of action production for a robot. They also proved that introducing motor signals improves the prediction of others' goal directed actions.

How a target object is approached plays a crucial role for the infants to attend to the action. Hogan et al. [6] proposed minimizing jerk, in which humans move their hands toward the target points by using the rate of change of acceleration. They stated that between the start and end points, a straight line hand trajectory is formed with a symmetric bell-shaped velocity throughout the motion. Daum et al. [7], made an experiment to show that infants older than six-months are capable of completing the trajectory of the experimenter's hand in the case of a goal-directed action.

Besides the direction of motion, during the first year of life, infants start to attend to the posture of the hand of the experimenter. [3], [8] showed that infants made predictive

[1]An object is a visually observable stuff that an experimenter is able to grasp and push such as a ball or a can.

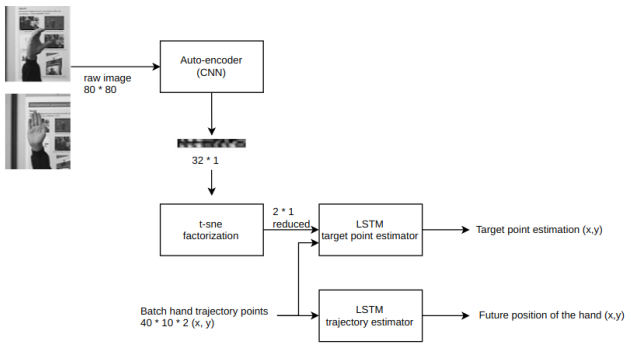


Fig. 1: Overall model architecture.

(a) grasping hand (b) pushing hand

(c) box hand (d) pointing hand

Fig. 2: These figures correspond to examplars from the dataset [11]. We used grasping (upper left) and pushing (upper right) hands in our experiment.

gazes toward the target object in the case of grasping reach, whereas when the back of hand is used they did not respond. [9] also showed that infants made predictive gazes to targets of reaches while experimenter is performing a grasping action whereas in the case of closed fist, infants followed the hand of the experimenter with a delay.

In this work, we propose a computational model in order to explain the development of prediction of others' actions and replicate the experiment results by [3]. In Section II we present the details of the model, in Section III we provide our initial results, and finally in Section IV we discuss how to further improve the model.

II. Model

In this work we aim to show the development of prediction of others' actions, specifically in grasping and pushing. For both of the actions, reaching trajectories towards the target objects are similar [10]. However, infants learn to predict grasping actions of others earlier than pushing actions [3], [8]. Since the trajectories are similar in both actions, the hand posture seems to be the key to differentiate grasping from pushing. What we aim to demonstrate is that even the generated trajectories for both of the actions are similar, progressively as infant learns from his/her experiences, he/she will predict the goals of others' grasping actions rather than the pushing actions.

We start by generating several trajectories following a basic parabola formula. We use a dataset generated by Sebastian Marcel [11]. Dataset includes six different hand posture images from ten people including the grasping and pushing hand postures. An example of a grasping and pushing hand can be seen from Figure 2. We train an autoencoder with the images from grasping and pushing postures, and apply t-SNE factorization [12] for dimensionality reduction, hence reducing feature size to only two.Then we, train two LSTM (Long short-term memory) networks with the trajectories and hand posture features in order to predict both the next trajectory point and the target point of the trajectory. We assume that an object is placed at the end of each trajectory sequence. Therefore if the model predicts the end point of the trajectory as the target point with a small error margin, we say that it predicts the action correctly. Figure 1 shows the overall model architecture. In the following subsections we will present more details of the each module of the model.

Trajectory Generation

We generated 2000 2D trajectories in which 2D corresponds to x and y coordinates respectively, and each trajectory has 50 2D pairs in it. We randomly select a start and target point and fill the points in between those by applying the following formula: $ax^2 + bx + c$. We generate all the trajectories by randomly selecting values for the parameters *a, b ,c* and start and target points.

Autoencoder and t-SNE Factorization

[3], [8] showed in their experiments that infants made predictive gazes only when the experimenter was performing a grasping action. Thus infants made predictions only when the hand of the experimenter is in grasping posture. Infants make the hand posture differentiation before the experimenter starts his/her motion of the action and decide whether he/she attends to the action or not. They have the ability to differentiate the hand postures and correlate them with the actions. In order to replicate this ability, we train an autoencoder and apply t-SNE factorization to each of the hand posture images. After the factorization what we would like to observe is a clear separation between the two sets so that the resulting features could be used to successfully differentiate between grasping and pushing hand. The image dataset [11] we use has 500 training images and 100 test images for each posture type and each image has dimension of 120 x 120 x 3. We first apply a preprocessing on the images and convert each image to 28 x 28. We use convolution layers to reduce the image to 32 x 1. Finally we apply t-SNE factorization to reduced image and get two features that represents whole of the image. Figure 4 shows the results for the test sets of both grasping and pushing hand posture images. We see that two clusters are separated

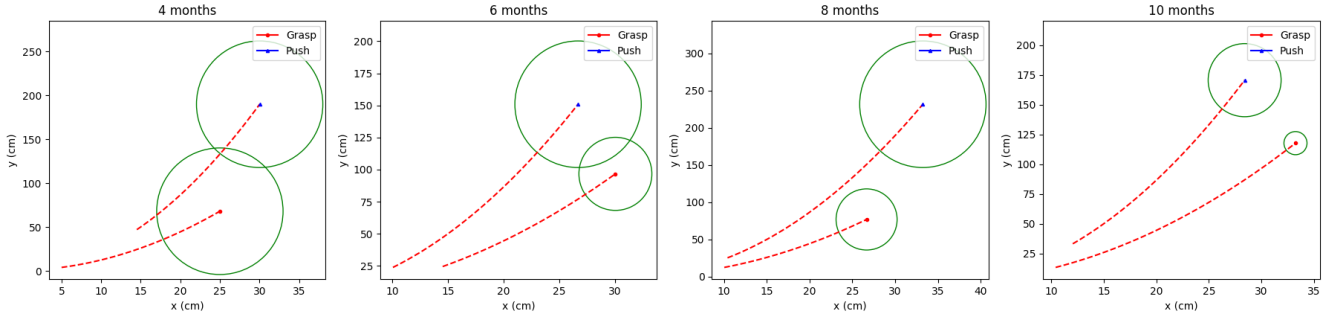


Fig. 3: LSTM training procedure with changing variance, *4 different LSTM networks are trained. In [3], the infants are divided into four groups from 4-months old to 10-months old. Leftmost figure matches the learning procedure of 4-months old infants whereas rightmost figure matches 10 months-old infants. Going from the leftmost figure to rightmost figure, the output variance is reduced for grasping action whereas for the pushing action variance is reduced only when going from 3rd to 4th. The output value of the red trajectory could be any point inside the green circle. The red and blue dots inside the green circle represent the end point of the trajectory and are the mean values for the output.*

almost perfectly. Hence the two resulting features seem a good indicator to differentiate between the grasping and pushing hand postures.

Target Point Estimation with LSTM

Given a partial trajectory,we aim to both predict the future points of the trajectory and the final point of the trajectory which corresponds to the target. Therefore we have trained two different LSTM networks, one for the future point prediction and one for the target point prediction. We divided each trajectory which has 50 2D points into sequences of 10 by sliding one by one. Each trajectory is then replaced with 40 new trajectories with length 10. And we refer to the divided trajectory sequence as the parent of the newly generated trajectory sequences. This operation is performed in order to better predict the future points. The features obtained from t-SNE factorization were used with the trajectory sequences of 10 to predict the target point. And finally we have divided the whole data set into two which correspond to training and test sets. 80% of the data are used as training set whereas the remaining 20% are used as the test set.

In order to show the development of the action prediction ability, we trained four different LSTM networks. We used the amount of variance to model the developmental progression.

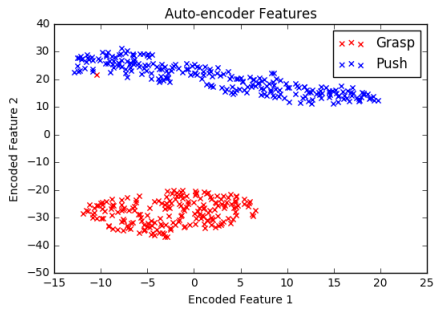


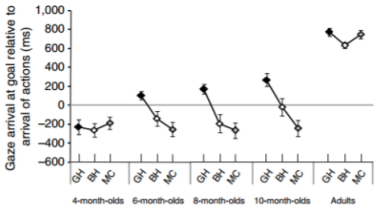
Fig. 4: Autoencoder and t-SNE Factorization Clusters.

Going from the first to last network, we reduced the variances of the output value. When the variance is equal to 0, during the training the output value is equal to the last point of the parent trajectory sequence. Hence we expect that the target point prediction becomes more precise when the variance is lower. In order to replicate the experiment of Kanakogi et al. [3], we start with high variance for each of the actions which are grasping and pushing. In the second and third networks we reduced the variance in grasping action whereas keep the variance in the pushing action same. In the final network we make the variance of grasping equal to 0 and reduced the variance in pushing as well, which corresponds to 10 months-old infants. Figure 3 refers to this idea.

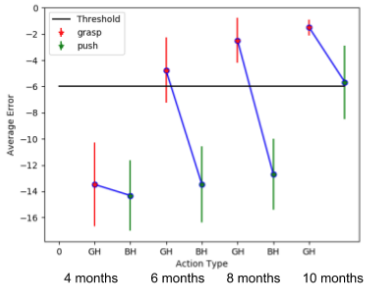
Results

In [3], the prediction accuracy with gaze arrival time is correlated. If the gaze of the infant arrives to the target object earlier than the completion of the hand trajectory, they conclude that infant is able to predict the action. In our case, we correlate the prediction accuracy with average error for each action in the test set. As can be seen from Figure 5b the average errors decrease from first network to last network for grasping, suggesting that the network learns to predict the trajectory of the hand better. One caveat here is that, in order to replicate the experiment results of [3] we have changed the signs of the average errors for each case. We set the error margin at -6 and say that if the average error is less than this value, it means that the prediction looks correct. For the pushing action, since the variance of the output values are higher compared to grasping action the predictions look arbitrary and it is not possible to observe the same learning progress as in the grasping action.

An example of the prediction results are shown in the Figure 6. All the predictions are made by looking at the first 10 points of the whole trajectory points. The little circles correspond to grasping predictions whereas the little triangles correspond to pushing predictions of the network. What is observed here is that, the grasping action predictions overlap



(a) Experiment results of [3], time of gaze arrival at the goal relative to arrival time of each agent's action for each age group. From 4 months-old to 10 months-old infants predict grasping action with better accuracy whereas for pushing action prediction capability remains same.



(b) Our results, instead of time of gaze arrival we used the distance between the actual end point of the trajectory and the predicted distance of our model.We set an error margin at -6 and state that if the error for a prediction is lower than the error margin prediction is successful. For grasping action, a developmental learning from the 4 months-old to 10 months-old is clearly seen whereas for pushing action except the 10 months-old prediction is not successful. Error bars indicate standard deviation for each group.

Fig. 5: Results.

with the final point of the trajectory which is expected since the model learns to predict grasping actions. However in the case of pushing actions, prediction errors are higher than the error margin defined from the final point of the trajectories as expected. Because the variance of the output values are high therefore model could not learn pushing actions as precise as the grasping actions.

Discussions

In this work, we replicate the experiment results of [3] by using the images of the hand postures and the artificial trajectory data. We train four different LSTM networks with changing variance in the output values to show the learning progress for two actions which are grasping and pushing. We use changing variance in order to model the development progression of an infant. Younger infants have less experience with both of the actions. Therefore, we expect their predictions of goal directed actions of others will be less accurate compared to predictions of older infants. Going from first network to last network we reduce the variance in the training for

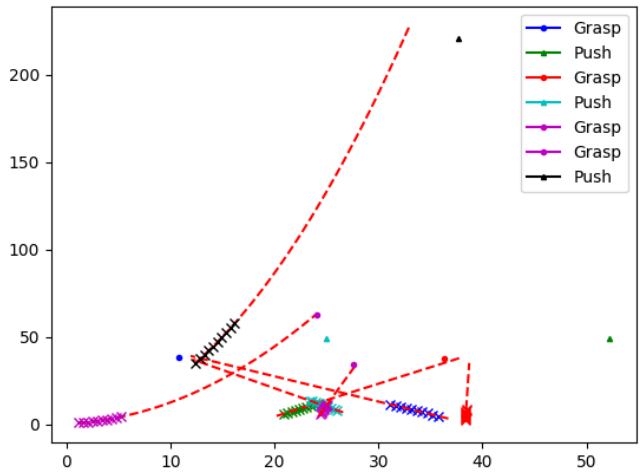


Fig. 6: Prediction results obtained of the final network, *Predictions are done from the first 10 trajectory points. The red lines corresponds to whole trajectories and the first 10 points are marked. The little circles correspond to grasping predictions whereas the little triangles correspond to pushing predictions of the network.*

certain actions and anticipate that the predictions will be more accurate. We introduce an error margin and decide whether a prediction is successful or not by comparing the prediction error with this margin. If the prediction error is smaller than the error margin we count the prediction as successful.

In the future, instead of using variance as the metric for learning, we would like our model to emerge behaviors such as grasping and pushing. What we plan to do is by using a human hand, in a simulation environment, learning the parameters of actions via Reinforcement Learning in the same vein as [13]. In such a learning scenario, we expect the grasping action to emerge before than the pushing action. Because, grasping creates more tactile stimulation and allow object manipulation with high prediction accuracy that is critical for planning. In particular, once the object is in the hand its pose can be predicted reliably, lending a higher utility for grasping over pushing.

During the reinforcement learning phase, after some episodes, we log all the data generated for all of the actions and replicate our work with the logged data. By doing this operation several times, after certain episodes, we plan to show the effect of emergence of behaviors to prediction capability.

Acknowledgment

This work was partially supported by JST CREST "Cognitive Mirroring" (Grant Number: JPMJCR16E2), Japan and by Bogazici Research Fund (BAP) project IMAGINE-COG, no 12721.

REFERENCES

[1] Robson, S.J., & Kuhlmeier, V.A. (2016). Infants Understanding of Object-Directed Action: An Interdisciplinary Synthesis. *Front. Psychol.(2016)*.

[2] Sommerville, Jessica & Woodward, Amanda & Needham, Amy. (2005). Action experience alters 3-month-old Infants' perception of others' actions. Cognition. 96. B1-11. 10.1016/j.cognition.2004.07.004.

[3] Kanakogi, Y., & Itakura, S. (2011). Developmental correspondence between action prediction and motor ability in early infancy. *Nature communications, 2, 341*.

[4] Rajmohan, V., & Mohandas, E. (2007). Mirror neuron system. -*Indian journal of psychiatry*.

[5] Jorge L. Copete, Yukie Nagai, and Minoru Asada, "Motor development facilitates the prediction of others' actions through sensorimotor predictive learning" in Proceedings of the 6th IEEE International Conference on Development and Learning and on Epigenetic Robotics, September 19-22, 2016.

[6] Flash, T. & Hogan, N. (1985) The coordination of arm movements: an experimentally confirmed mathematical model. J. Neursci., 5, 16881703.

[7] Daum, M. M., Prinz, W., & Aschersleben, G. (2008). Encoding the goal of an object-directed but uncompleted reaching action in 6- and 9-month-old infants. Dev. Sci. 11, 607619. doi: 10.1111/j.1467-7687.2008.00705.x

[8] Woodward, A. L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. Infant Behav. Dev. 22, 145160. doi: 10.1016/S0163- 6383(99)00007-7

[9] Gredebäck, G., Stasiewicz, D., Falck-Ytter, T., Rosander, K., and von Hofsten, C. (2009). Action type and goal type modulate goal-directed gaze shifts in 14- month-old infants. Dev. Psychol. 45:1190. doi: 10.1037/a0015667

[10] Hamlin, J. K., Hallinan, E. V., and Woodward, A. L. (2008). Do as I do: 7- month-old infants selectively reproduce others goals. Dev. Sci. 11, 487494. doi: 10.1111/j.1467-7687.2008.00694.x

[11] S. Marcel. Hand posture recognition in a body-face centered space. In Proceedings of the Conference on Human Factors in Computer Systems (CHI), 1999.

[12] L.J.P. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 15(Oct):3221-3245, 2014.

[13] Oztop E, Bradley NS, Arbib MA (2004) Infant grasp learning: a computational model, Exp Brain Res. 158:480-503