Towards a Mirror Neuron System via Dual Channel Conditional Neural Movement Primitives

M.Yunus Seker[1], Erhan Oztop[2,3], Mete Tuluhan Akbulut[1], Minoru Asada[3], and Emre Ugur[1]

Abstract— **Mirror neuron system often refers to the brain mechanism of establishing and use of the equivalence of action observation and action execution. Mirror neurons were originally found in monkeys; but, recent neuroimaging data indicate that the adult human brain is endowed with a mirror neuron system, containing mirror neurons and related circuits for matching the observation and execution of actions. Exact mechanism of the mirror system is far from known although computational models have been proposed to explain certain functions of the system in the past. In this paper, we propose a mirror neuron system based on a novel computational system called Conditional Neural Movement Primitives (CNMPs), and report our preliminary findings. In the proposed system, the visual data and motor signals generated during self-action are fused together via CNMP learning, which allows sharing and mirroring of the relevant information from the two domains. After learning, the system can predict the full action trajectory and visual scene together, given the partial observation of an action, and generalize the knowledge it learned to different scene configurations.**

I. INTRODUCTION

Mirror neurons were originally found in macaque monkeys, in the ventral premotor cortex, area F5 [1] and later also in the inferior parietal lobule [2]. There are F5 visuomotor neurons that selectively discharge to the visual presentation of a given object, which also discharge selectively during grasping of that object [3]. Recent neuroimaging data indicate that the adult human brain is endowed with a "mirror neuron system", containing mirror neurons and other neurons, for matching the observation and execution of actions. Mirror neurons may serve action recognition in monkeys as well as humans, whereas their putative role in imitation and language may be realized in human but not in monkey.

Computational methods were used to produce and explain the known effects of mirror neurons in recent years (e.g. [4]–[6]). In [7], a computational model was introduced to explain the psychological findings which indicate that there is a correlation between infant's ability to predict others' action goals and development of their own motor ability to produce similar actions. Here, they transferred the sensorimotor information which consists of visual signals, tactile signals, and joint angles to a shared latent space by using an auto-encoder architecture in order to replicate a mirror neuron system that allows making action goal predictions.

In this paper, we propose a system that learns a representation that simultaneously encodes the actions of an agent and its observations. This system self-interacts with the objects around through the agent's grasp and push actions, and learns from self-observed image sequences and the motor commands, i.e. joint angles, that generate the corresponding image sequences. Our system demonstrates properties similar area F5 neurons, including canonical neurons. Given the object to be manipulated or the image of the manipulator that interacts with the object the related action and the corresponding motor program are generated.

Our system is based on a recent movement framework, namely Conditional Neural Movement Primitives that are typically used for Learning for Demonstration (LfD). LfD [8] has been applied to various robotic learning problems including object grasping and manipulation [9]–[13]. Among others, learning methods that are based on dynamic systems [14] and statistical modeling [15] have been popular in recent years. Dynamic Movement Primitives (DMPs) [14] encode the demonstrated trajectory as a set of differential equations, and offers advantages such as one-shot learning of non-linear movements, real-time stability and robustness under perturbations with guarantees in reaching the goal state, generalization of the movement for different goals, and linear combination of parameters.

Specifically, in this paper, we propose a dual channel CNMP that conditions the distribution of manipulation actions with desired images and produces a sequence of predicted images and motor values together. We propose that the images and the motor commands can share related information between each other while training, thus, enhancing the learning procedure while simultaneously coupling the information between these two domains, and constructing a mirror-neuron-like system where the visual information by itself can lead to the activation of the corresponding motor commands.

The experiments that were performed with a simulated robotic hand and a top-down camera showed that our system is able to

- predict the full manipulation trajectory and images of the scene at the same time, given an image of the start of the movement as an observation,
- generalize the knowledge it learned in the training to different scene configurations

after trained with a small interaction set of grasp and push actions.

[1]Computer Engineering, Bogazici University, Istanbul, Turkey; [2]SISReC, OTRI, Osaka University, Japan; [3]Computer Engineering, Ozyegin University, Istanbul, Turkey

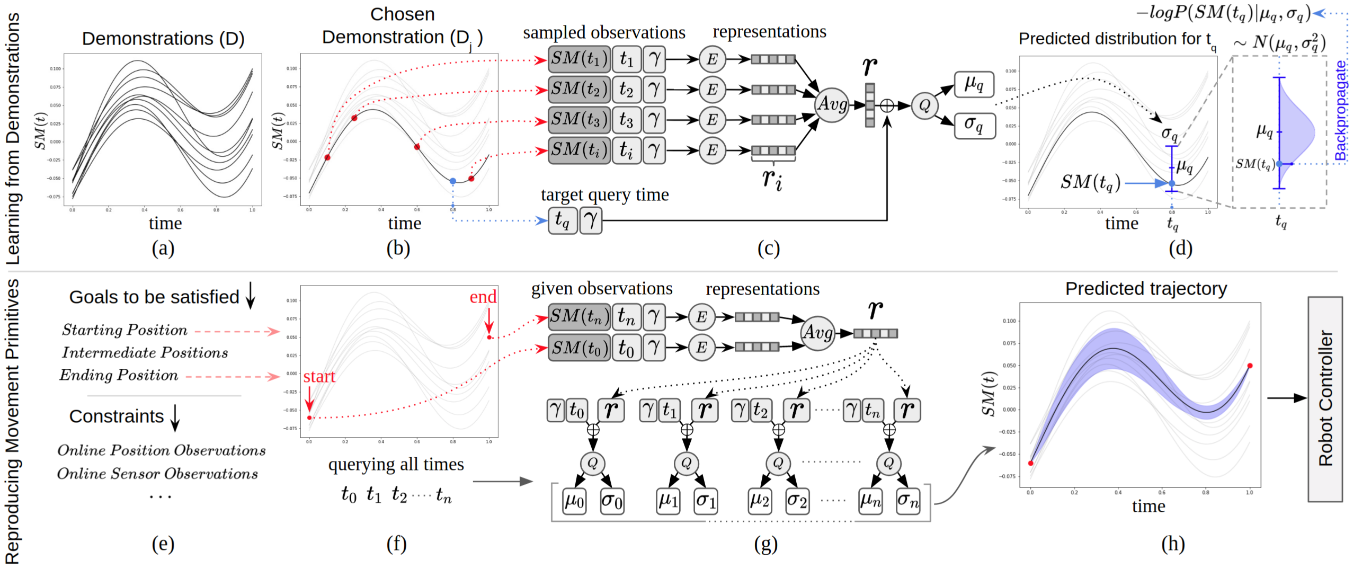

Fig. 1. Training and trajectory generation steps of Conditional Neural Movement Primitives. See Section II.A for details.

II. Method

In this section, first, we provide a summary of Conditional Neural Movement Primitives that our system is based on. For more detailed information and capabilities of CNMPs readers are referred to our paper [16].

A. Background: Learning from Demonstration with CNMPs

CNMP is a learning from demonstration framework that learns multi-modal temporal relationships of the trajectory distributions shown by the experts. It can generate trajectories conditioned on the desired positions at any time step or external parameters in the task or joint space. Fig. 1. provides the general framework of CNMPs for an example 1D scenario. Fig. 1.1a shows the demonstration set, D. At each training step a demonstration is sampled randomly from the demonstration set. A changing number of observation tuples and a target query tuple which consist of sensorimotor data and corresponding time value are sampled from the selected demonstration (Fig. 1.1b). Fig. 1.1c illustrates the neural network architecture for the sampled observations. If external task parameters (γ) are used, they are concatenated with observation tuples and target time value. Each observation sample is passed through the parameter sharing Encoder Network (E) and the corresponding latent space representations (r_i) are obtained. After applying an averaging operation on the latent space representations, a general representation (r) is formed. The produced general representation and the target query time (t_q) are given to the Query Network (Q) as input and the Query Network produces a mean and variance of a Gaussian distribution which represent the distribution of sensorimotor values at the target time step t_q(Fig. 1.1d). Neural network parameters (θ and ϕ) of both Encoder and Query network are optimized end-to-end with the loss function below using the stochastic gradient descent algorithm:

$$\mathscr{L}(\theta,\phi) = -\log P(SM(t_q) \mid \mu_q, \text{softplus}(\sigma_q)) \qquad (1)$$

where μ_q and σ_q are predicted distribution parameters for target time-step t_q, and $SM(t_q)$ is the ground truth sensorimotor value at target time-step for the sampled demonstration in that training iteration.

After the training is over, CNMP can produce trajectories conditioned on any starting, intermediate, and ending position or other constraints such as online position or sensor observations (Fig. 1.2). Condition positions are given as observations and external constraints like online sensor readings are given as task parameters (γ) to the model. In order to generate a full trajectory, Query Network predicts a mean and variance for all time steps, which are given as target query time to the model. In the end, the desired trajectory is obtained for the given conditions and it can be given as an input to any robot controller to execute the motion.

B. Proposed Method: Towards A Mirror Neuron System via Dual Channel CNMPs

In this work, we propose a dual channel CNMP architecture in order to be able to learn and mirror the visual and proprioception information together while sharing information between these two different domains. A dual channel CNMP means that there are two separate CNMPs trained with a shared representation latent space. This allows both CNMPs to share related and important information between the networks while training, thus, mirroring the learned information of the same timesteps. Figure 2 shows an example dual channel CNMP where the above row is the image learning

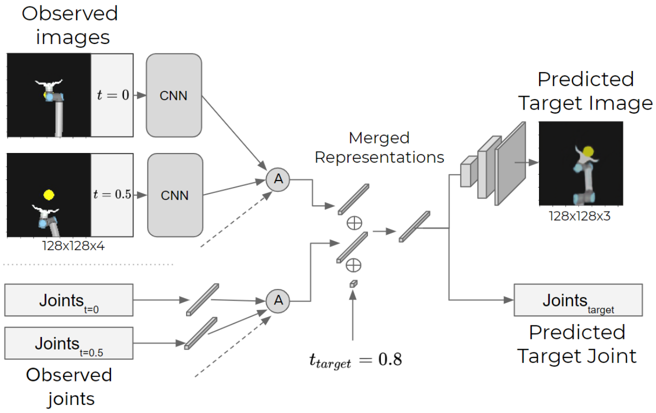

Fig. 2. Proposed mirror neuron framework. Image and joint observations are turned into their latent representations and merged into a shared latent space in order to predict the image and joint positions given at another target timestep.

CNMP channel, and the below row is the joint value learning CNMP channel. Note that both general representations which are obtained after aggregating the observations are mixed into a single representation space before separated again into two in order to predict the relevant image and joint values respectively.

More specifically, we define our demonstration set as D and a demonstration as $D_j = \{(t_i, I_i, J_i)\}_{i=0}^{T}$ where t_i is the time, I_i is the image of the scene collected by the vision sensor, and J_i is the joint data values of the robot at the ith movement step. Figure 2 shows the general structure of the proposed system. At every training step, an observation set is sampled from a randomly selected demonstration D_j. We define the observation set as $O = \{(t_{s_i}, I_{s_i}, J_{s_i})\}_{i=1}^{n} \in D_j$ where n is the random observation number of that iteration $n \in [1, n_{max}]$ and s_i is the ith sampled timestep from the selected demonstration. As seen in the figure, time information of the observations is concatenated to the image as the fourth layer. In order to merge the image and the joint information of the two channels into one single representation, first, the dimensions of the image data are reduced. Standard V-shaped CNN architecture is used to reduce the dimensions of the image into a vector. After reducing the dimensions, image and joint representations are aggregated into their general image and joint representations by using a mean operator (A in the Figure 2). After obtaining the general representations of the image and joint information separately, both representations are concatenated all together with the target time information and mixed into a shared latent space after passed through a dense layer. At this point, a high-level information that involves mixed information of the observed images, observed joints, and the target time that is wanted to be predicted is obtained through a single representation. Finally, mixed representation is copied and separated to both channels in order to predict image and joint position at the target timestep. For both channels, the same loss function defined in the CNMPs is used. Loss values are calculated separately but back-propagated together in order to train the neural network end-to-end and all together.

After the training, the system can be queried by any visual observation that is desired to be satisfied. According to the representations that are encoded through the given observations, dual CNMP can produce full trajectory image and joint trajectory prediction together at the same time because of the information mirroring during the training between the image and joint prediction channel.

III. Experimental Results

To show the capabilities of our system, we designed an environment where the robot's actions can be foreseen by the visual observations taken during the start of the movement execution. A simulated experiment environment using V-REP is built. The setup consists of a UR10 robot equipped with a 3 finger gripper, an object on a table, and a vision sensor that collects information about the scene during the movement. Two motions, pushing and grasping, are defined as the movement primitives and data is collected as follows for each interaction: At the start of the interaction, the robot moves its wide-open hand to an initial point. An object is placed in the middle of the table. If the selected action is push, a random pushing angle is defined and the robot pushes the object to a constant distance from that angle with open gripper. If the selected action is grasp, a random grasping angle is defined and the hands starts to close on the way while the hand is approaching to the object. After closing down the gripper completely and grasping the object, the robot moves the object up to a constant height. At each time step, the joint data of the robot and the image recorded using the vision sensor are collected. In the end, 40 push and 40 grasp interactions are collected using the simulator.

A. Predicting the Approach Angle and the Type of Action According to Visual Observations

In this experiment, we verify whether our system can produce visual sequence of the expected observation and the required joint values given a single image obtained from the beginning of the demonstration. Figure 3 shows example results in predicting the type of motion and the approach angle by using a single image taken from the beginning of the motion. The observed images are shown in Figure 3-left. It can be seen that exploiting to the position of the robot hand in the observed image, our system could successfully predict the approach angle to the object and produce correct images for the rest of the time-steps accordingly. Although the approach angles of the row 1-2 and row 3-4 in the figures are the same in between, our system could successfully predict the type of the action and produce grasping or pushing actions according to the state of the gripper in the observation images which was wide open or closing. The ground truth and the predicted images at the four target times-teps are shown in Figure 3-middle-right.

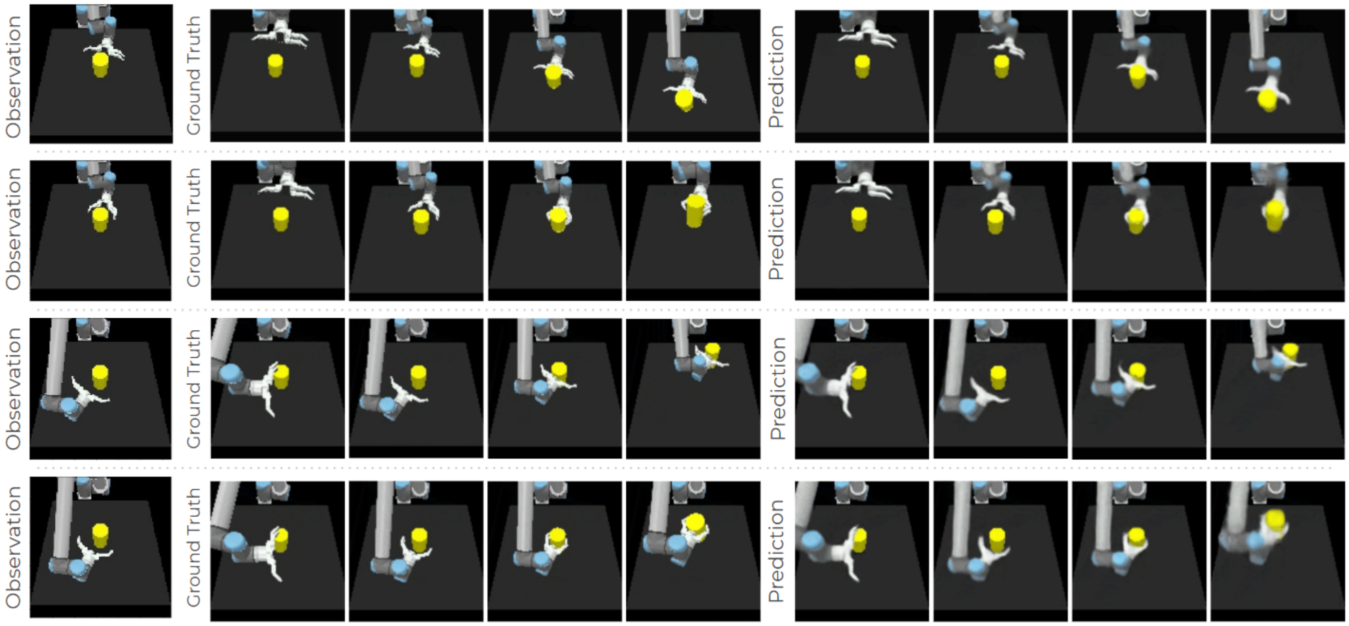

Fig. 3. Image prediction results of the proposed framework. Left: The image that is used as the observation to predict the images of other four timesteps. Middle: The ground truth images of the target timesteps. Right: The images that are predicted for the target timesteps through dual channel CNMPs

B. Generalization of the System to the Novel Visual Information

We tested our system with different novel scenes that have different properties that are not in the training set. Figure 4 shows the generalization performances of the two different configurations. The first row shows a scenario in which the color of the object was different from the object in the training data, and the second row shows a configuration where the size of the object was changed. Despite not seeing a big or blue object in the training, our system could successfully predict the correct approaching angle and the action using the observed image in both configurations. It can be seen that the color and the size of the objects are predicted as in the configuration in the training images. This is expected since there was only a single configuration of the object in the training scenes which was yellow and small. Even though the object in the observed image was not the same with the training, our system could generalize the knowledge that is learned in the training procedure to produce a correct output to satisfy the given observation and accomplish the task.

IV. Conclusion

In this paper, a mirror neuron framework based on Conditional Neural Movement Primitives is proposed. Visual data and joint information are trained together to construct a better and more quality representation space. Our future work will study demonstrating the use of predicted motor signals in reproducing an observed action or scene, and assessing the effect of the perspective on the prediction capacity.

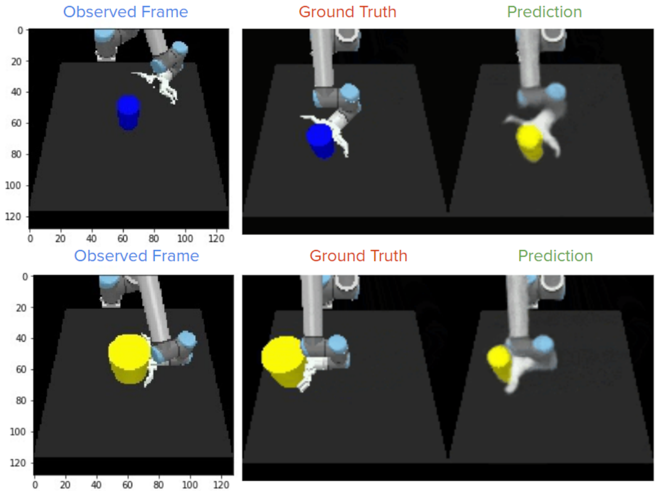

Fig. 4. Generalization performance of the proposed system in two different configurations. First row: the color of the object is blue. Second row: the size of the object is bigger than the original one.

Acknowledgment

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 731761, IMAGINE and JST CREST 'Cognitive Mirroring' (Grant Number: JPMJCR16E2). Additional support is provided by the International Joint Research Promotion Program, Osaka University under the project 'Developmentally and biologically realistic modeling of perspective invariant action understanding'.

References

[1] G. Di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Understanding motor events: A neurophysiological study," 1992.

[2] L. Fogassi, P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti, "Parietal lobe: from action organization to intention understanding," *Science*, 2005.

[3] A. Murata, L. Fadiga, L. Fogassi, V. Gallese, V. Raos, and G. Rizzolatti, "Object representation in the ventral premotor cortex (area f5) of the monkey," *Journal of neurophysiology*, 1997.

[4] E. Oztop and M. A. Arbib, "Schema design and implementation of the grasp-related mirror neuron system," *Biological Cybernetics*, vol. 87, pp. 116–140, 2002.

[5] J. Bonaiuto, E. Rosta, and M. Arbib, "Extending the mirror neuron system model, i - audible actions and invisible grasps," *Biological Cybernetics*, vol. 96, no. 1, pp. 9–38, 2007. [Online]. Available: ¡Go to ISI¿://000244065600002

[6] J. Bonaiuto and M. A. Arbib, "Extending the mirror neuron system model, ii: what did i just do? a new role for mirror neurons," *Biological Cybernetics*, vol. 102, no. 4, pp. 341–59, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20217428

[7] J. Copete, Y. Nagai, and M. Asada, "Motor development facilitates the prediction of others' actions through sensorimotor predictive learning," 2016.

[8] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Rob. and Auto. Sys.*, 2009.

[9] S. Calinon, P. Evrard, E. Gribovskaya, A. Billard, and A. Kheddar, "Learning collaborative manipulation tasks by demonstration using a haptic interface," in *Advanced Robotics, 2009*, 2009.

[10] T. Asfour, P. Azad, F. Gyarfas, and R. Dillmann, "Imitation learning of dual-arm manipulation tasks in humanoid robots," *International Journal of Humanoid Robotics*, 2008.

[11] P. Pastor, L. Righetti, M. Kalakrishnan, and S. Schaal, "Online movement adaptation on previous sensor experiences," in *IROS*, 2011.

[12] H. Ben Amor, O. Kroemer, U. Hillenbrand, G. Neumann, and J. Peters, "Generalization of human grasping for multi-fingered robot hands," in *IROS*, 2012.

[13] M. Mühlig, M. Gienger, and J. J. Steil, "Interactive imitation learning of object movement skills," *Autonomous Robots*, 2012.

[14] S. Schaal, "Dynamic movement primitives-a framework for motor control in humans and humanoid robotics," in *Adaptive Motion of Animals and Machines*. Springer, 2006, pp. 261–280.

[15] S. Calinon, "A tutorial on task-parameterized movement learning and retrieval," *Intelligent Service Robotics*, 2016.

[16] M. Y. Seker, M. Imre, J. Piater, and E. Ugur, "Conditional neural movement primitives," in *Robotics: Science and Systems (RSS)*, 2019.