Unsupervised Topographic Representations of Motor Primitives in Humanoid Manipulation and Gesture

Radovan Gregor, Igor Farkaš
Department of Applied Informatics, Comenius University
Mlynská dolina 6284, 84248 Bratislava

Abstract

Understanding the computational basis of action recognition is a key challenge in social cognition as well as human–robot interaction. Inspired by the mirror neuron system, we propose a two-level architecture for motor primitive discovery and processing, applied to the humanoid robot. In the first level, two self-organising maps learn topologically structured manifolds over arm kinematics and hand kinematics covering seven manipulation and gesture actions. The second level uses the Echo State Network that encodes the temporal grammar of primitive sequences to predict actions and infer intentions from partial observations. In this paper, we focus on the first level and report preliminary results.

1 Introduction and Related Work

The ability to recognise observed actions and infer the intentions behind them is fundamental to social interaction. An observer who can interpret partner's motor behaviour can proactively adapt, anticipate, and collaborate more effectively. The Mirror Neuron System (MNS) offers a compelling biological model for this capacity: mirror neurons discharge both when an individual executes a motor act and when it observes another performing the same act (Gallese et al., 1996), grounding action understanding in an internal motor vocabulary rather than abstract perceptual classification (Rizzolatti and Craighero, 2004). Neurophysiological studies have shown that parietal mirror neurons encode not only the observed action but also its goal (Fogassi et al., 2005). This goal-specificity is a key feature of the action-understanding hypothesis and motivates the design of computational systems that can similarly distinguish actions at the level of motor primitives and at the level of goal-directed sequences (Oztop et al., 2006).

The classical computational approach encodes observed motion directly from kinematic or visual signals using recurrent architectures trained end-to-end on action labels (Shahroudy et al., 2016). More recent methods replace recurrence with transformer-based architectures, enabling the modeling of long-range temporal dependencies through self-attention (Bertasius et al., 2021). However, neither approach embeds the biological insight that action understanding operates across two levels simultaneously: a repertoire of elementary motion primitives and the goal-directed sequences they compose (Wolpert et al., 2003). Computational models of the MNS have addressed this by modelling the mirror system as a self-organising map (SOM) (Kohonen, 1982) that receives both motion and context inputs, showing that goal-specific neuron pools emerge naturally from the geometric relationships between these inputs without any explicit goal representation (Thill et al., 2011). The same principle underpins chain models of intention understanding (Chersi et al., 2011; Erlhagen et al., 2006), in which sequences of motor primitives activate a learned chain that predicts the overall action goal. In our previous MNS-related work, we also exploited topographic mapping as an organizing principle (Pospíchal et al., 2019; Rebrová et al., 2013).

Here we focus on topographic representation of proprioceptive signals whose sequence describes the moving arm. Proprioception is one of the least understood senses in cognitive neuroscience, yet fundamental for the control of movement. There is evidence that topography of cortical representations is a widespread organising principle of the brain (deCharms et al., 2000). It has also been proposed for proporioception at the level of latent representations of the variational autoencoder (Grogan et al., 2024).

We propose a two-level MNS-inspired architecture grounded in these findings, using topographic organisation at the lower level. There, two SOMs discover a vocabulary of reusable motion primitives from arm and hand kinematics independently – a dual-stream design consistent with the observation that motion and context inputs to parietal mirror neurons originate from distinct neural pathways (Thill et al., 2011). The topology-preserving projection naturally groups temporally similar movement episodes into prototype nodes, which serve as a discrete primitive vocabulary (Hemeren and Thill, 2011). At the higher level, an Echo State Network (Lukoševičius and Jaeger, 2009) encodes the temporal ordering of these primitives into state-level representations from which actions and intentions can be inferred, extending our previous work on arm trajectory prediction (Gregor et al., 2025).

Our central research question is: How much of a robot's motor sequence structure is determined by motion primitives alone, and how much additional contex-

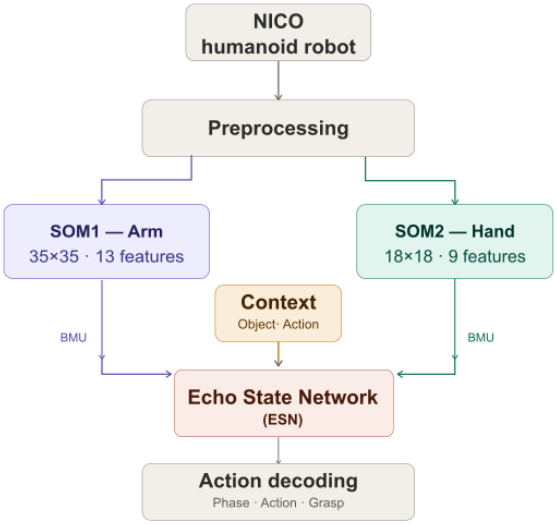

Fig. 1: The overall concept of the presented work.

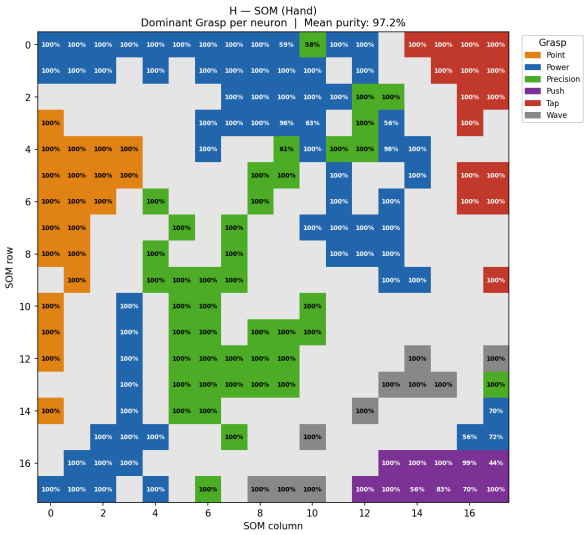

Fig. 2: Topographic organisation (in H-SOM) of dominant grasp types across neurons.

tual information (object properties, action intent) is required for accurate motion prediction?

2 Data and Methods

2.1 Simulation and Data Collection

To build the dataset for motion primitive discovery, we employ a physics-based Unity simulation environment featuring the NICO robot (Kerzel et al., 2017), a humanoid platform for multi-modal interaction. In our setup, the robot performs 7 manipulation and gesture actions on 14 physically diverse objects. Object manipulation actions (Pick, Eat, Place) are executed with either a Power or Precision grasp. Motion trajectories are generated through inverse kinematics, while the hand is controlled via four synergy parameters inspired by human grasping studies.

2.2 Feature Analysis and Selection

Prior to SOM training, a feature selection pipeline was applied to candidate signals to identify and remove redundancy and features that are workspace-dependent and do not generalise across object locations. Based on correlations and hierarchical clustering we selected 22 non-redundant features, which are divided into two functional sets: (1) 13 arm features for arm SOM (A-SOM): 3 TCP (Tool Center Point) palm velocities, 6 arm joint velocities, palm orientation quaternion; and (2) 9 hand features for hand SOM (H-SOM): aperture and aperture velocity, 3 thumb and middle finger positions, 4 finger joint velocities.

2.3 Topographic organisation

We measured the behaviour of the trained A-SOM by computing the so-called *phase purity*, as the average proportion of the dominant phase label across all active neurons, to see how consistently each neuron represents a single movement phase, based on the dominant phase label of its assigned samples. This information helped to determine the proper size of the grid. Increasing the SOM grid from 25×25 to 35×35 improved this consistency (from 68% to 74%) by better separating similar phases while maintaining high neuron utilisation, making it the optimal trade-off between resolution and map coverage. H-SOM uses a 18×18 grid. Testing showed that the hand configuration space is low-dimensional, the synergy controller produces approximately six distinct hand poses, limiting the effective vocabulary size.

3 Results

A-SOM learned to use up to 90% of its neurons (as winners for any input). Phase purity (74%) is the strongest label dimension, with large coherent zones emerging without supervision. The map encodes a directional reaching topology: approach phases occupy the left half and post-action phases the right, a gradient learned entirely from kinematics with no temporal supervision. Action purity (67%) and grasp purity (66%) are lower, as expected, since arm kinematics alone cannot distinguish actions sharing identical reach trajectories, nor encode grip type, both of which are delegated to H-SOM.

H-SOM achieves 51.5% neuron utilisation, lower than A-SOM by design – the hand synergy controller produces discrete, non-overlapping configuration

spaces, so dead neurons act as hard boundaries between isolated action islands rather than representing wasted capacity (see Fig. 2). Grasp purity (97.2%) is the best result overall. Point, Tap, Wave, and Push each form perfectly pure isolated clusters, whereas Power and Precision diverge into distinct, non-overlapping regions, with impurities restricted to a narrow transition band in the mid-close range. Each hand pose occupies its own island with no kinematic path between them, reflecting the hard discreteness of the synergy controller. Action purity (72.9%) and phase purity (76.1%) confirm that finger-level signals carry complementary discriminative structure to the arm-level primitives of A-SOM.

4 The upper level

The lower-level SOM outputs serve as input representation for the upper-level Echo State Network. By encoding sequences of SOM activations, the ESN aims to predict actions and infer intentions from partial observations, extending our prior work on trajectory prediction shown in Figure 3 (Gregor et al., 2025).

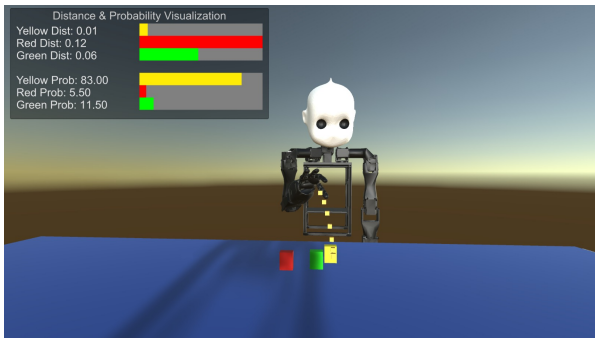

Fig. 3: ESN prediction of future arm trajectories.

The network will be evaluated under three conditions of increasing contextual information: (1) *blind* — BMU sequences only; (2) + *object size and relative distance*; (3) + *action one-hot*. This design directly probes the central scientific question: how much of motor sequence structure is encoded in the primitive vocabulary alone? We expect the blind condition to reveal action-discriminative structure already present in the primitive sequences, with object context and action context providing incremental gains corresponding to the residual ambiguity quantified in the SOM purity metrics.

5 Conclusion

The dual-SOM results demonstrate that arm and hand kinematics encode complementary aspects of motor behaviour not recoverable from either stream alone. A-SOM captures the spatial structure of arm movement (*where* and *how* of reaching), while H-SOM captures the configurational structure of hand shaping (*what* and *how firmly* of grasping). This functional specialisation mirrors the dorsal/ventral stream division in primate motor cortex (Rizzolatti and Craighero, 2004). The complementarity of the two SOMs provides a biologically grounded dual representation supporting both primitive-level motor resonance and sequence-level intention inference, consistent with the dual role attributed to the MNS.

A key limitation is that the SOM is currently trained on balanced, stratified windows from a single observer executing its own actions. Extension to cross-embodiment observation, where NICO observes a human or a different-size robot, will require the mirroring mechanism to resolve the viewpoint and embodiment gap.

Acknowledgment

This research was supported by Slovak Research and Development Agency, project APVV-21-0105.

References

Bertasius, G., Wang, H., and Torresani, L. (2021). Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, pages 813–824.

Chersi, F., Ferrari, P. F., and Fogassi, L. (2011). Neuronal chains for actions in the parietal lobe: A computational model. *PLOS ONE*, 6(11).

deCharms, R. C., , and Zador, A. (2000). Neural representation and the cortical code. *Annual Review of Neuroscience*, 23:613–647.

Erlhagen, W., Mukovskiy, A., and Bicho, E. (2006). A dynamic model for action understanding and goal-directed imitation. *Brain Research*, 1083:174–188.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science*, 308:662–667.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119:593–609.

Gregor, R., Farkaš, I., Malinovska, K., Sobota, B., and Kerzel, M. (2025). Mirroring-based prediction of robot arm movements using an echo state network. In *International Conference on Automation, Control and Robots*, pages 49–54.

Grogan, M., Blum, K. P., Wu, Y., Harston, J. A., Miller, L. E., and Faisal, A. A. (2024). Predicting proprioceptive cortical anatomy and neural coding with topographic autoencoders. *PLOS Computational Biology*.

Hemeren, P. E. and Thill, S. (2011). Deriving motion primitives through action segmentation. *Frontiers in Psychology*, 1:243.

Kerzel, M. et al. (2017). NICO – Neuro-Inspired COmpanion: A developmental humanoid robot platform for multimodal interaction. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 113–120.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149.

Oztop, E., Kawato, M., and Arbib, M. A. (2006). Mirror neurons and imitation: A computationally guided review. *Neural Networks*, 19:254–271.

Pospíchal, J., Farkaš, I., Pecháč, M., and Malinovská, K. (2019). Modeling self-organized emergence of perspective in/variant mirror neurons in a robotic system. In *IEEE International Conference on Development and Learning*.

Rebrová, K., Pecháč, M., and Farkaš, I. (2013). Towards a robotic model of the mirror neuron system. In *IEEE International Conference on Development and Learning and on Epigenetic Robotics*.

Rizzolatti, G. and Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition)*, pages 1010–1019.

Thill, S., Svensson, H., and Ziemke, T. (2011). Modeling the development of goal-specificity in mirror neurons. *Cognitive Computation*, 3(4):525–538.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B*, 358(1431):593–602.