Proceedings of the 2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), Engineer's House Conference Centre, Oslo, Norway, August 19-22, 2019

Modeling Self-organized Emergence of Perspective In/variant Mirror Neurons in a Robotic System

Jakub Pospíchal, Igor Farkaš & Matej Pecháč
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava
Bratislava, Slovak Republic
farkas@fmph.uniba.sk, matej.pechac@gmail.com

Kristína Malinovská
Czech Institute of Informatics, Robotics, and Cybernetics
Czech Technical University
Prague, Czech Republic
kristina.malinovska@cvut.cz

***Abstract*—A major role attributed to mirror neurons, according to the direct matching hypothesis, is to mediate the link between an observed action and agent's own motor repertoire, to provide understanding "from inside". The mirror neurons gave rise to various models but one of the issues not tackled by them is the perspective in/variance. Neurons in STS visual areas can be either perspective selective or invariant and the same variability was later also discovered in premotor F5 area in macaques, showing the existence of different types of mirror neurons regarding their perspective selectivity. We model this as an emergent phenomenom using the data from the simulated iCub robot, that learns to reach for objects with three types of grasp. The neural network model learns in two phases. First, the motor (F5) and visual (STS) modules are trained in parallel to self-organize modal maps using the corresponding data sequences from the self-perspective. Then, F5 area is retrained using the output from the pretrained STS module, to acquire the mirroring property. Using the optimized model hyperparameters found by grid search, we show that our model fits very well empirical observations, by showing how neurons with various degrees of perspective selectivity emerge in the F5 map.**

***Index Terms*—perspective invariance, mirror neurons, cognitive robotics, iCub, neural network**

I. Introduction

Action understanding is undoubtedly a vital component in human–robot interaction. The direct matching hypothesis [1] places the so-called mirror neurons into the role of the medium that matches the observed action with the closest counterpart of one's own motor repertoire. Additionally, the feedback connection between motor and visual areas of the brain with projections from mirror neurons to visual areas could facilitate the very complex task of invariant action recognition in the visual parts of the mirror neuron system circuitry [2]. In our previous research we started to develop a modular hierarchical architecture for a humanoid robot to model rudimentary action understanding based on the outlined theories [3]. In this paper we continue our research with a focus on novel evidence on viewpoint in/variance in the mirror neuron firing [4].

A. Mirror neurons and perspective in/variance

Mirror neurons are cells in ventral premotor area F5 in the macaque brain that encode goal-directed hand and mouth movements, which also fire when the monkey is still and solely observes these actions [5]. Since their discovery in early 1990s, mirror neurons have gained a lot of attention. There are strong theories proposing that mirror neurons play an important role in action understanding, imitation learning and empathy [6], as well as strong opposition against such theories [7].

Mirror neurons have been postulated to serve for direct matching of the observed and executed movement representations to facilitate or at least mediate understanding of actions of others. Brain areas of the macaque brain that constitute the original MNS are frontal and parietal areas F5, PF, PG (PFG) that have been shown to have mirror properties, and the superior temporal sulcus (STS), which is sensitive to a large variety of biological movements, but does not react to stimuli from other modalities (i.e. other than vision), therefore it is not a true part of the MNS.

The role of STS area in the MNS is to project visual information to the system. Neurons of STS respond to various visual stimuli in a view-dependent (in posterior part, STSp), but also view-independent manner (in anterior part, STSa) [8]. By view-dependent neurons we mean sensitive to viewpoint or *perspective*, from which the scene is observed (e.g. front view, side view, etc.), we will also call it perspective-variant. Analogically, the view-independent neurons that react to the same object or movement regardless of the viewpoint will be called perspective-invariant. Interestingly, STSp projects to sector F5c in area F5 through PFG forming a connection that can be seen as a perspective variant path, but also STSa projects to F5a through AIP, which can be seen as an invariant path. The invariant path encodes the actor and the object acted upon, rather than the viewpoint from which it is observed [9].

Perspective-variant and invariant firing properties were also found in area F5 [4]. In a novel experimental design, monkeys were watching short films of grasping actions shown in three different perspectives, the self-observing view ($0°$), the side view ($90°$), and the opposite view ($180°$). Caggiano and colleagues report that 52% of all (389) recorded neurons had mirroring properties, 74% of these visually responsive motor neurons exhibited perspective-dependent and 26% perspective invariant properties. In a follow-up research they explored the rhythmic properties of MN responses with a conclusion that the perspective from which the action is observed is reflected in the firing and that self-observing view produces significantly different response which is more similar to motor execution. Therefore, it is possible that action observation

978-1-5386-8128-2/19/$31.00 ©2019 IEEE

can trigger different processes in F5 based on the viewpoint. In our current work, we develop a mirror neuron model that will, along with commonly modeled perspective-invariant MN, also encompass perspective variant mirror neurons. This process has not been sufficiently addressed in the modeling literature, even though this is an important step in the long-term developmental process of learning abstraction that needs to be computationally explained (which could also help in designing artificial agents).

B. Computational models

Computational models of the MNS can be basically divided into two groups as pointed out in [10], a large overview is offered centered around clustering and confronting the existing models with empirical data. FARS, MNS1 and MNS2, MSI and other models summarized in [11], [12] capture the areas of the monkey brain with precise connections and focus on grasping. Models such as HAMMER [13] or RNNPB [14] aim at endowing the cognitive robotics with action understanding through implementing the mirroring function without explicitly using brain area analogs.

There were several recent approaches trying to model view invariance. For instance, in [15] it is shown how view independence in imitation learning is achieved through a bio-inspired hybrid model (combining neural networks, conditionals and radial basis function networks). In recent cognitive robotics, a Time-Contrastive Network has been proposed for a robot to learn to imitate the human demostrator in a self-supervised manner from various perspectives also utilizing motor knowledge [16]. Invariant perception of faces (identities) based on the recordings from the face-processing network of the macaque brain has been presented in [17] as an alternative approach to end-to-end SGD-trained deep networks in the form of hierarchical architectures trained through biologically plausible Hebbian learning.

In our previous work [3], we aimed at accounting for the existence of view-dependency of neurons in both STS and F5 as a possible outcome of their bidirectional connectivity. We presented a multi-layer connectionist model of action understanding circuitry and mirror neurons, emphasizing the bidirectional activation flow between visual and motor areas pointed out in [2]. We implemented our model in a simulated iCub robot that learned a grasping task. Within two experiments we demonstrated the properties of the model and also discussed further steps to be done to extend the functionality of our model towards achieving view invariant properties of neurons in STSa area. In this work, inspired by the existence of biological data [4], we focus on modeling the emergence of mirror neurons in F5 area, which has not been attempted before, using a very similar neural network model. This is modeled as an emergent self-organized process and the model properties are shown to match well biological data, including the observation that invariance is a graded, rather than a binary, phenomenon. Computational account of this process is the main contribution of this paper.

II. Robotic MNS Model

Our robotic MNS model (Fig. 1) consists of two major connected modules (F5 and STSp), that are fed with data from their respective modalities. The model consists of two levels. At the lower level there are executive modules, namely the motor executive module, feeding sequential motor information like the joint angles, and the visual module, which provides sequential visual information to the system (from a concrete perspective). We assume that sensory-motor links are established between higher level representations, rather than directly between low-level representations of the movement as a temporal sequence of the robot arm's state. At the higher level, there are two modules that process the low-level motor and visual information, and form high-level representation of movement in F5 and STSp, respectively.

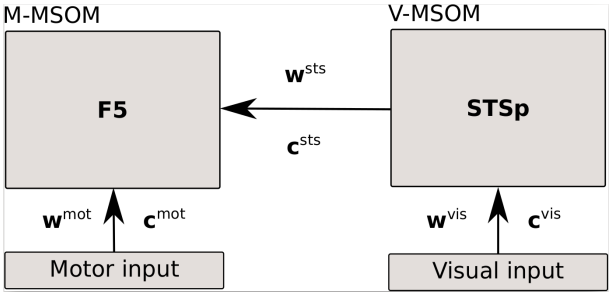


Fig. 1: Model schematics. Input mot and vis connections are updated in Phase A, and sts connections in Phase B.

For experiments we chose the freely available simulator [18] of the iCub robot [19] that is considered to be one of the most accurate humanoid robots. For generating the grasping sequences, we created a preprogrammed routine controlling iCub's movements. Both higher areas, F5 and STSp, are implemented as MSOMs [20], the self-organizing maps[1] that can process sequential data.

In the process of acquiring the whole MNS functionality the robot first learns to produce the three grasps. The information from the motor module is processed with the higher level F5c module (Sec. II-A) and gets organized on the resulting map as clusters of instances of the same movements. At the same time, we assume that the robot observes another robot producing the same actions and creates visual representations of those actions from different perspectives (self, $90°$, $180°$, and $270°$) in STSp and associates them with the motor representations. Then, if the robot observes an action from various perspectives, the motor representation of the action is triggered as well.

A. Merge Self-Organizing Map

To make the paper self-contained, first we describe the MSOM model [20]. MSOM is based on the well-known Kohonen's map but it has recurrent architecture, so it can be used for self-organization of sequential data. Each neuron $i \in \{1, 2, ..., N\}$ in the map has two weight vectors: (1)

[1]Like in our previous work [3], for modeling STSp and F5, we use topographic maps as a ubiquitous organizing principle in the brain [21].

$\mathbf{w}_i \in \mathcal{R}^n$, associated with an n-dimensional input vector $\mathbf{s}(t)$, and (2) $\mathbf{c}_i \in \mathcal{R}^n$, associated with the so-called context descriptor $\mathbf{q}(t)$ specified below. The output of unit i at time t is computed as $y_i(t) = \exp(-d_i(t))$, where

$$d_i(t) = (1-\alpha) \cdot \|\mathbf{s}(t) - \mathbf{w}_i(t)\|^2 + \alpha \cdot \|\mathbf{q}(t) - \mathbf{c}_i(t)\|^2 \quad (1)$$

Hyperparameter $0 < \alpha < 1$ trades off the effect of the context and the current input on the neuron's profile and $\|\cdot\|$ denotes the Euclidean norm. The context descriptor is calculated from the information related to the previous best matching unit ('winner') $b(t-1) = \arg\min_i\{d_i(t-1)\}$, as

$$\mathbf{q}(t) = (1-\beta) \cdot \mathbf{w}^{\mathrm{inp}}_{b(t-1)}(t) + \beta \cdot \mathbf{w}^{\mathrm{ctx}}_{b(t-1)}(t) \quad (2)$$

where hyperparameter $0 < \beta < 1$ trades off the impact of the context and the current input on the context descriptor. The training sequences are presented in natural order, one input vector a time, and in each step both weight vectors are updated using the same form of Hebbian rule:

$$\Delta\mathbf{w}_i(t) = \gamma \cdot h_{ib} \cdot (\mathbf{s}(t) - \mathbf{w}_i(t)), \quad (3)$$

$$\Delta\mathbf{c}_i(t) = \gamma \cdot h_{ib} \cdot (\mathbf{q}(t) - \mathbf{c}_i(t)), \quad (4)$$

where b is the winner index at time t and $0<\gamma<1$ is the learning rate. Neighborhood function h_{ib} is a Gaussian (of width σ) on the distance $d(i,k)$ of units i and b in the map: $h_{ib} = \exp(-d(i,b)^2/\sigma^2)$. The neighborhood width, σ, linearly decreases in time to allow for forming topographic representation of input sequences. As a result, the units (i.e. their responsiveness) get organized according to sequence characteristics, biased towards their suffixes (most recent inputs).

B. Mirror neurons training

The training of mirror neurons was performed in two phases (referred to later in experiments A and B). In Phase A we trained both F5 and STSs on their own input sequences representing the paired motor and visual data (self-learning in a robot). The F5 and STSp module weights (i.e., $\mathbf{w}^{\mathrm{mot}}_i, \mathbf{c}^{\mathrm{mot}}_i$, and $\mathbf{w}^{\mathrm{vis}}_i, \mathbf{c}^{\mathrm{vis}}_i$ pairs, respectively) were updated according to Eq. 3 and 4. After training (for T episodes), Phase B starts when the M-MSOM is again trained, using its own motor inputs (keeping the weights $\mathbf{w}^{\mathrm{mot}}$ and $\mathbf{c}^{\mathrm{mot}}$ fixed), as well as precomputed activation vectors in STSp.[2] Activation of F5 units is based on Eq. 5 (with time indices left out for simplicity) which is again taken as an affine combination of the two sources. After presenting the visual input to V-MSOM, calculating its activations and obtaining the distances d^{v}_i we presented motor inputs to F5, and could calculate the merging distance d^{mir}_i of neuron i in F5 according to Eq. 6 as

$$d^{\mathrm{v}}_i = (1-\alpha)\|\mathbf{y} - \mathbf{w}^{\mathrm{sts}}_i\|^2 + \alpha\|\mathbf{r} - \mathbf{c}^{\mathrm{sts}}_i\|^2 \quad (5)$$

$$d^{\mathrm{mir}}_i = \kappa \cdot d^{\mathrm{m}}_i + (1-\kappa) \cdot d^{\mathrm{v}}_i \quad (6)$$

where $\mathbf{y}$ is the activation vector from STSp, $\mathbf{r}$ is the corresponding context vector (updated according to Eq. 2), and the hyperparameter κ trades off the relative contributions of the two sources.

During pilot simulations we figured out that in order to enforce better selectivity of neurons in F5, two additional mechanisms would be useful in phase B. The first one is the winner-take-all competition among STSp neurons, yielding $b(t) = \arg\min_i\{d^{\mathrm{vis}}_i(t)\}$, i.e. the winner for each input pattern, its activation $y_b = 1$ and activations of other neurons set to zero, resulting in a one-hot vector $\mathbf{y}(t)$. The second mechanism, applied to F5 neurons, aims at eliminating the "distance bias", by calculating the unit's activations as

$$y_i = \frac{1}{1 + \exp(k(d_i - \langle d^{\mathrm{mir}}_j \rangle))} \quad (7)$$

where $\langle . \rangle$ denotes the mean value (of all units for the current input), and k is set empirically (we used $k = 10$). This distance shift increases the mutual differences among competing neurons which is also important for better selectivity.

In summary, a regular MSOM algorithm is applied in two steps. In phase A, M-MSOM and V-MSOM are trained on respective input sequences, using Eq. 3 and 4, yielding $\mathbf{w}^{\mathrm{mot}}, \mathbf{c}^{\mathrm{mot}}, \mathbf{w}^{\mathrm{vis}}, \mathbf{c}^{\mathrm{vis}}$. In phase B, M-MSOM is trained again (using weights from phase A), updating only its visual weights $\mathbf{w}^{\mathrm{sts}}$, $\mathbf{c}^{\mathrm{sts}}$, based on merging distances d^{mir}_i (Eq. 5 and 6), computed from paired inputs (via d^{mot} and d^{vis} activations).

III. Results

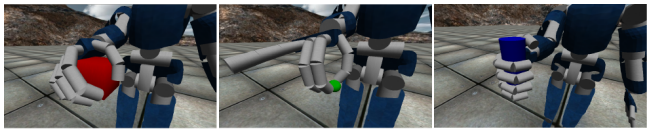
We present results from two experiments. Experiment A encompasses processing of visual (preprocessed Cartesian coordinates) and motor (joint angles) data taken from the trained iCub during grasping an object. The data are self-organized to high-level topographic representations using the MSOM model. In Experiment B we let mirror neurons emerge and develop by linking pretrained motor activations in F5 with STSp activations. After training, the visual input will cause (via STSp) the activation of dedicated mirror neurons in F5.

A. Self-organization of sensory and motor inputs

For training MSOMs, we first needed to generate the input data. Both sensory and motor sequences were artificially generated in the iCub simulator using forward and inverse kinematic module. In motorgui (extension to iCub simulator) we set the final position of arm in one of three different types of grasp: power, precision and side grasp. These grasps (shown in Fig. 2) were generated, with 10 instances per category, in such a way that individual trajectories slightly differed from one another (which was achieved by adding small perturbations to arm joints during the motion execution), (resulting also in slight differences among final positions.

1) Collecting the motor and visual data from iCub: One instance of the grasp lasted for 4 seconds, and each 1/4 of a second the values of the individual joints were stored. Thus, one motor sequence consisted of 16 steps and the representations are based on proprioceptive information provided by all joint values from 16 DoF in robot's right arm which were stored during the motor execution, i.e. input vectors

[2]This presumes that F5 units also have inputs from the visual pathway (STSp map itself) weighted by $\mathbf{w}^{\mathrm{sts}}_i$ and $\mathbf{c}^{\mathrm{sts}}_i$, not utilized in phase A.



(a) Power (b) Precision (c) Side

Fig. 2: Examples of three grasp types.

$\mathbf{s}_{\mathrm{m}} \in \mathcal{R}^{16}$ for motor map. These values are given in degrees, and so prior to storing them we rescaled them to the interval $[-1, 1]$, independently for each DoF. Visual information was obtained from a camera that is available in the iCub simulator (Fig. 3) which has been set to four positions representing four orthogonal perspectives (0°, 90°, 180°, and 270°).

(a) 0° (b) 90° (c) 180° (d) 270°

Fig. 3: Four different perspectives taken by iCub's camera.

As for visual input, for obtaining the coordinates of iCub fingers, it was necessary to use the iCubFinger extension, which provided information for thumb, index and middle fingers assuming they are sufficient for recognition (ring and pinky fingers are underactuated). We used the chains of joints from iKin extension to calculate the final values for each frame. So we could get 3D joint positions with respect to the robot's coordinate system, which were then transformed to the world coordinates. In the next step, using the algorithm, we extracted coordinates from the projection of frame onto the camera lens (Fig. 4). Overall we had 14 points in 2D yielding 28 values as input vectors $\mathbf{s}_{\mathrm{v}} \in \mathcal{R}^{28}$ for visual MSOM. As in the case of motor data, the values were rescaled to $[-1, 1]$, independently for each coordinate. This preprocessing was repeated for each frame of each perspective. In total, we used 30 motor and 120 visual sequences.

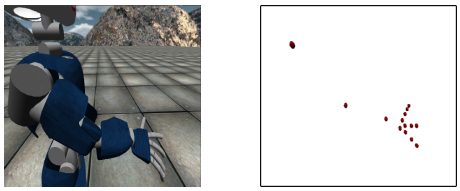


Fig. 4: Snapshot of right iCub's arm (on the left) and its visual kinetic representation in 2D space (on the right).

2) Finding optimal maps: In order to obtain map representations, we systematically searched for optimal MSOM hyperparameters (for maps 12×12 in F5 and 16×16 in STSp). We aimed at getting the map that would optimally distribute its resources (units) for best discrimination of input data. Following the methods from our previous work [22], we calculated three quantitative measures: (1) winner distribution (WD) returns the proportion of different winners (out of all units) at the end of training; (2) entropy (ENT) evaluates how often various units become winners, so the highest entropy means most balanced unit participation in the competition process; (3) quantization error (QE) calculates the average error at the unit as a result of quantization process. To get the best MSOMs, we used grid search for hyperparameters α (eq. 1) and β (eq. 2) in the interval (0,1) and selected the configuration with highest WD and ENT and possibly minimal QE. As a result, we chose $\alpha_{\mathrm{v}} = 0.4$ $\beta_{\mathrm{v}} = 0.8$ for STSp, and $\alpha_{\mathrm{m}} = 0.8$, $\beta_{\mathrm{m}} = 0.9$ for F5. Heatmaps for F5, shown in Fig 5 reveal that the map performance is very sensitive to trading off the context and input (quantified by α_{m}), whereas the other parameter (β_{m}) weighting the winner's components is quite robust. The profiles for STS map were similar.

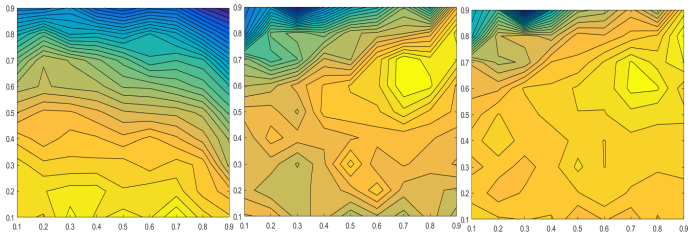
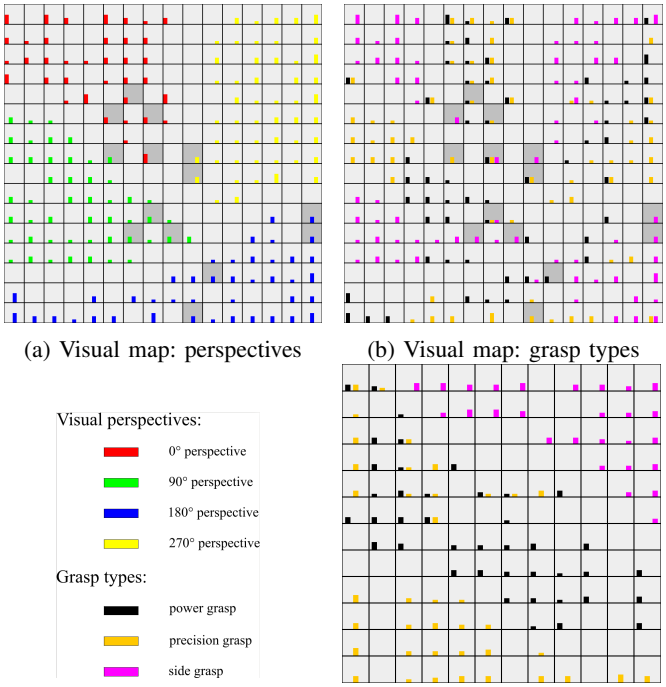


Fig. 5: Contour plot of results of grid search for F5 hyperparameters α_{m} (vertical axis) and β_{m} (horizontal axis), for three evaluated measures (left to right): quantization error, winner distribution and entropy. The lighter the color, the higher the value. The best model has high WD and ENT, and low QE.

Using these hyperparameters, we trained both MSOMs and evaluated winner hits (i.e. the number of times the particular unit became the winner, for three categories of grasp type for both visual (Fig. 6b) and motor dataset (Fig. 6c), and additionally for four perspectives for the visual dataset (Fig. 6a). For better transparency we only plotted units that won at least three times. Topographic organization of unit's sensitivity is evident in all cases. For visual map, the organization on a coarse level is arranged according to the perspectives, and on a more fine-grained level according to the grasp types. Topographic order reflects the natural separability of classes (types of grasps) both in terms of their motor and visual features. Visual maps reveal that perspective is a more strongly distinguishing feature than the type of grasp.

B. Emergence of mirror neurons

Once the F5 and STSp maps have been trained using paired visual and motor sequences, the second step (phase B) of training F5 can be initiated. Here we use a sort of "scenario-based shortcut". It is known that parents often imitate children's immediate behavior providing them with something like a mirror, which may explain how mirror neurons could emerge as a product of associative learning [23]. Hence, in our approach, the robot first produces the self movement, while observing its own arm. Right after it, while the generated motor pattern is assumed to be still residually active, the robot



(a) Visual map: perspectives (b) Visual map: grasp types

(c) Motor map: grasp types

Fig. 6: Optimal trained visual and motor maps. The shaded cells in visual maps denote winners for earlier steps within sequences (reaching phase).

observes the same movement from another perspective (as if it was playing an educational game with its parent).

Hence, during movement observation the connections from STSp to F5 are trained, using the residual motor input to F5. Using grid search, we chose $\alpha_{\text{mir}} = 0.3$ and $\beta_{\text{mir}} = 0.6$ as optimal hyperparameters for phase B. To three measures mentioned above, we added the fourth, *grasp selectivity* defined as the percentage of grasp type selective neurons in the map. Variations in hyperparameter κ (eq. 6) led to different topographic organizations in F5. For κ being close to zero, the visual information became more influential in map organization, and for κ close to one, the contribution of the motor information encoding the grasp type became dominating (over perspective) for final organization. Setting $\kappa = 0.9$ led to best results.

During the evaluation phase, we presented the inputs only to STSp module and its sole activation served as the input to F5 module. The motor input was completely omitted. We then examined the topographic organization by the same method as we used previously. We trained several models and compared the obtained number of invariant and variant neurons with measured data [4]. The maps that best fitted experimental data are presented in Fig. 7, revealing the emergence of neurons responding to various numbers of perspectives (for a given grasp type), ranging from one (perspective-variant), two (bivariant), up to all perspectives (invariant).[3] Most of the neurons are grasp type selective, except a few (based on comparison of both maps). In the top left corner are neurons (dark grey cells) responding to reaching phase of the movements (mostly the first five frames of each sequence), mostly for all grasp types (completely invariant neurons). However, there are also perspective-invariant grasp-selective neurons, in each of the grasp-specific areas. Interestingly, bivariant neurons were always selective for (any of the) two "neighboring" perspectives (i.e. not the opposing perspectives that are visually most different).

The quantitative comparison of the best model with biological measurements is shown in Fig. 8. In the model, multiple views preference includes neurons responsive to two or three perspectives (in [4] it was two perspectives, since 270° was not measured). The graph shows that overall, around 88 neurons (i.e. 60% of 144) had mirroring properties, of these 28 (20%) are invariant, and around 60 neurons (42%) are variant (including also partially invariant neurons). In summary, the graph reveals a good match in relative proportions of different types of mirror neurons.

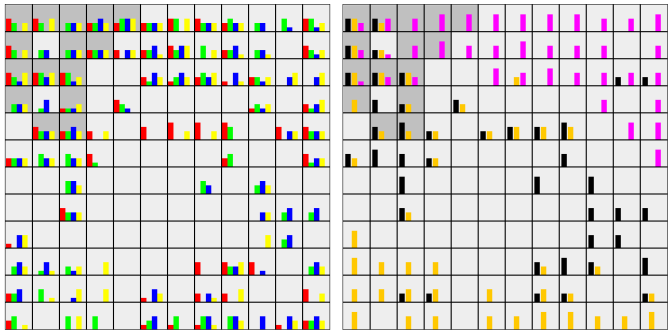


Fig. 7: Responses of mirror neurons to visual inputs. *Left:* Winners for various grasps from different perspectives. *Right:* Winners for different types of grasps.

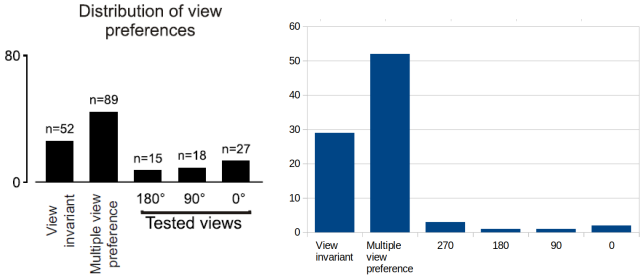


Fig. 8: *Left:* Experimental data (from [4]). *Right:* Neuron preferences in F5 area of the model. Bars denote the numbers of neurons.

[3]Despite the existence of prepared pairs of patterns for learning, we think we can interpret the invariance as an emergent phenomenon whose variability cannot be predicted from the data.

IV. Discussion

We presented pilot results directed toward enriching our computational model with mirror neurons in F5 area by letting the visual representations (STSp) affect its organization. In our approach, we were inspired by empirical findings showing that in STS as well as F5, most of the neurons are perspective-dependent, and that this may be the lowest level at which sensory-motor relationships could be established.

In the first phase, we let the simulated iCub robot learn to reach for and grasp an object with its right hand. Subsequently, we let the F5 area be properly reorganized using visual representations from STSp, to affect the F5 organization by proper trading off motor and visual information. This self-organized process resulted in a variety of motor neurons, some of which also acquired mirror properties to various degrees. Assumption behind this approach is that mirror neurons are not predetermined but emerge in early phase of sensorimotor development. Another assumption behind our approach is the graded degree of invariance of both visual and motor neurons. This is consistent with empirical evidence when it comes to STS [24] and F5 [4]. Indeed, in F5 we observe a variety of neurons regarding their selectivity to perspectives and/or to grasp type.

The desirable neuron selectivity was achieved using two additional mechanisms: (1) winner-based localist output representation of STSp map for reorganization of F5 with mirroring properties, and (2) shifting distances in F5 neurons during competition to enlarge mutual differences. These mechanisms turned out to be necessary, since no lateral inhibition in MSOM map was implemented that could take care of better selectivity, albeit with an increased computational cost (as confirmed by preliminary experiments).

Regarding the limitations of our model, we could have used a more variable data set in terms of final arm and hand positions. Neither did we apply the classical split to training and testing data, that would reveal the generalization property of the mapping. These results are merely a proof of concept, that such a mapping with desirable properties can be learned, given optimized (hyper)parameters. Therefore, we did not provide a more detailed quantitative assessment of the models. A questionable feature of the model is also the assumption about the residual activity of the motor system allowing the pairing of motor and visual sequences which are crucial for learning the mirroring property.

Acknowledgment

This work was supported by grant 1/0798/18 from Slovak Grant Agency for Science (VEGA). Kristína Malinovská was supported by grant VI20172019082, Smart Camera, from Ministry of the Interior of the Czech Republic. The authors thank the three reviewers for precious comments.

References

[1] G. Rizzolatti, L. Fogassi, and V. Gallese, "Neurophysiological mechanisms underlying the understanding and imitation of action," *Nature Reviews Neuroscience*, vol. 2, pp. 661–670, 2001.

[2] G. Tessitore, R. Prevete, E. Catanzariti, and G. Tamburrini, "From motor to sensory processing in mirror neuron computational modelling," *Biological Cybernetics*, vol. 103, no. 6, pp. 471–485, 2010.

[3] K. Rebrová, M. Pecháč, and I. Farkaš, "Towards a robotic model of the mirror neuron system," in *3rd IEEE International Conference on Development and Learning and on Epigenetic Robotics*, 2013.

[4] V. Caggiano, L. Fogassi, G. Rizzolatti, J. K. Pomper, P. Thier, M. A. Giese, and A. Casile, "View-based encoding of actions in mirror neurons of area F5 in macaque premotor cortex," *Current Biology*, vol. 21, no. 2, pp. 144–148, 2011.

[5] G. 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.

[6] G. Rizzolatti and C. Sinigaglia, "The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations," *Nature Reviews Neuroscience*, vol. 11, no. 4, pp. 264–74, 2010.

[7] H. G. and H. M., "(mis)understanding mirror neurons," *Current Biology*, vol. 20, no. 14, pp. R593–4, 2010.

[8] T. Jellema and D. Perrett, "Neural representations of perceived bodily actions using a categorical frame of reference," *Neuropsychologia*, vol. 44, pp. 1535–1546, 2006.

[9] K. Nelissen, E. Borra, M. Gerbella, S. Rozzi, G. Luppino, W. Vanduffel, G. Rizzolatti, and G. A. Orban, "Action observation circuits in the macaque monkey cortex," *The Journal of Neuroscience*, vol. 31, no. 10, pp. 3743–3756, 2011.

[10] Y. Demiris, L. Aziz-Zadeh, and J. Bonaiuto, "Information processing in the mirror neuron system in primates and machines," *Neuroinformatics*, vol. 12, no. 1, pp. 63–91, 2014.

[11] E. Oztop, M. Kawato, and M. Arbib, "Mirror neurons and imitation: A computationally guided review," *Neural Networks*, vol. 19, no. 3, pp. 254–271, 2006.

[12] ——, "Mirror neurons: Functions, mechanisms and models," *Neuroscience Letters*, vol. 540, pp. 43–55, 2013.

[13] 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, 2003.

[14] J. Tani, M. Ito, and Y. Sugita, "Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using RNNPB," *Neural Networks*, vol. 17, no. 8-9, pp. 1273–1289, 2004.

[15] H. Oh, A. R. Braun, J. A. Reggia, and R. J. Gentili, "Fronto-parietal mirror neuron system modeling: Visuospatial transformations support imitation learning independently of imitator perspective," *Human Movement Science*, 2018.

[16] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1134–1141.

[17] J. Z. Leibo, Q. Liao, F. Anselmi, W. A. Freiwald, and T. Poggio, "View-tolerant face recognition and hebbian learning imply mirror-symmetric neural tuning to head orientation," *Current Biology*, vol. 27, no. 1, pp. 62–67, 2017.

[18] V. Tikhanoff et al., "An open-source simulator for cognitive robotics research: The prototype of the iCub humanoid robot simulator," in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*. ACM, 2008, pp. 57–61.

[19] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: an open platform for research in embodied cognition," in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*. ACM, 2008, pp. 50–56.

[20] M. Strickert and B. Hammer, "Merge SOM for temporal data," *Neurocomputing*, vol. 64, pp. 39–71, 2005.

[21] J.-P. Thivierge and G. Marcus, "The topographic brain: from neural connectivity to cognition," *Trends in Neurosciences*, vol. 30, no. 6, pp. 251–259, 2007.

[22] P. Vančo and I. Farkaš, "Experimental comparison of recursive self-organizing maps for processing tree-structured data," *Neurocomputing*, vol. 73, no. 7-9, pp. 1362–1375, 2010.

[23] C. Heyes, "Where do mirror neurons come from?" *Neuroscience and Biobehavioral Reviews*, vol. 34, no. 4, pp. 575–83, 2010.

[24] D. Perrett et al., "Viewer-centred and object-centred coding of heads in the macaque temporal cortex," *Experimental Brain Research*, vol. 86, no. 1, pp. 159–173, 1991.