Learning Visual-Motor Cell Assemblies for the iCub Robot using a Neuroanatomically Grounded Neural Network

S. V. Adams, T. Wennekers, A. Cangelosi
Centre for Robotics and Neural Systems
Plymouth University, Plymouth, UK
samantha.adams@plymouth.ac.uk

M. Garagnani, F. Pulvermüller
Brain Language laboratory
Freie Universität Berlin
Berlin, Germany

Abstract— In this work we describe how an existing neural model for learning Cell Assemblies (CAs) across multiple neuroanatomical brain areas has been integrated with a humanoid robot simulation to explore the learning of associations of visual and motor modalities. The results show that robust CAs are learned to enable pattern completion to select a correct motor response when only visual input is presented. We also show, with some parameter tuning and the pre-processing of more realistic patterns taken from images of real objects and robot poses the network can act as a controller for the robot in visuo-motor association tasks. This provides the basis for further neurobiological experiments on grounded language learning.

Keywords—Neurorobotics; Cell Assemblies; Visual-Motor Learning

I. INTRODUCTION

The wider context for the current work is the design of a biologically-inspired architecture for brain embodied language skills in robots. This is the goal of the interdisciplinary BABEL project, which aims at the synthesis of computational neuroscience modelling (informed by and informing actual fMRI experiments), of the SpiNNaker neuromorphic hardware [1] and the humanoid robot iCub [2] for language learning. In particular an embodied approach to language acquisition is chosen where visual input and motor interaction with the world constitute the grounding for language. Research in behavioural and cognitive neuroscience demonstrates that language, action and perception are closely linked in the brain [3-5]. There is now substantial neuroscientific evidence that the use of language activates brain areas closely joining together motor, perceptual and speech-language mechanisms (see review in [6]). This is consistent with the embodied view of cognition in psycholinguistics and cognitive science where cognitive functions, such as language, are closely integrated with sensorimotor knowledge [7], and with the situated learning approach to studying language in context. Consequently, there have been advances in both computational neuroscience in the design of neuroanatomically grounded models of word acquisition [8-10] and also in cognitive robotics by the use of such brain-inspired models in the integration of action and language learning in robots [11,12]. In [11], Morse et al. trained the humanoid robot iCub to learn the names of objects, replicating the same phenomena observed in child language experiments. The robot’s neural control architecture learned to integrate visual, motor and linguistic representations through Hebbian-linked Kohonen maps. In [12], Caligiore et al. developed the TRoPICALS model to study how vision, action and language are integrated in the representation and activation of affordances. The TRoPICALS model architecture is based upon real brain areas with a path relating to object identity, name and current goal (PFC: prefrontal cortex, STC: superior temporal cortex, VOT: ventral occipito-temporal area) and a path relating to object shape, position and desired position of the arm and hand to form the correct grasp (PMC: premotor cortex, PC: parietal cortex, SSC: somatosensory cortex; VC: visual cortex). However, these robotic models are limited to the use of relatively small numbers of neural elements to model cortical processing. As its core model for language learning, BABEL uses the existing work of [9] which is a multi-area neural network model whose architecture is neuroanatomically grounded in the left perisylvian language cortex and includes biologically inspired learning. In its original formulation, the model had six areas and was used to investigate the association of auditory and articulatory stimuli. This model is currently being extended to 12 areas to incorporate an extra ‘semantic’ set of 6 areas for visual and motor (action) input. The ultimate aim of BABEL is to produce a spiking neuron implementation of the 12 area architecture so that it can be deployed on SpiNNaker hardware and integrated with the iCub robot, delivering a better real time performance than a software neural network running on a host PC would be able to achieve. This integrated platform will be used in experiments for object naming and word associations with action composition. As a first step in the roadmap for robotics experiments we investigated if it was possible to integrate the existing standalone neural model with the iCub.
Fig. 1. Relevant brain areas and model architecture. (A) Sets of cortical areas modelled. Three areas relevant for the planning and execution of manual action (PF – prefrontal; PM – premotor; M1 – primary motor) in the frontal lobe, and three for processing information concerning visual object identity (AT – anterior-temporal; TO – temporo-occipital; V1 – primary visual) in the occipital and temporal lobes, are identified. Numbers indicate Brodmann Areas. (B) Overall between-area connectivity realised by the model. Colours indicate the mapping between model and corresponding cortical areas. A link (arrow) between two model areas indicates the presence of random and sparse – but not all-to-all – synaptic connections between cells of these areas. Links between model correlates of cortically adjacent areas are depicted in black; longer-distance cortico-cortical connections (i.e., between areas that are not cortically adjacent) in purple.

simulator to learn associations between the visual representation of an object and an appropriate grasping action. The structure of this paper is as follows: firstly we briefly describe the existing neural model and modifications that were required to allow learning of visual and motor associations with the iCub simulator. Next we introduce the iCub simulator and its communications protocol, YARP (Yet Another Robotic Platform) as well as describe the integration between the robot and the neural model. Lastly we describe experiments performed with this setup and present the results to show how well the system enabled iCub to learn the association of a visual object and the required grasping motion. We conclude with an evaluation of this prototype and future work required to enhance it in the wider context of the BABEL project aims.

II. LEARNING VISUAL-MOTOR ASSOCIATIONS

A. The Neural Model

The model architecture is shown in Fig. 1.B. In this section we present only a brief overview of salient features of the model—the mathematical formulation of the model, structural and connectivity details, and learning mechanisms, have been previously described elsewhere [9,13-16]. We implemented a network consisting of six areas of artificial neurons (graded response cells) with reciprocal connections between and within areas. The model was constructed so as to mimic a representation of an object and an appropriate grasp well the system enabled iCub to learn the association of a visual object and the required grasping motion. We conclude with an evaluation of this prototype and future work required to enhance it in the wider context of the BABEL project aims.

5. Synaptic weight changes mediated by Hebb-type learning (simulating long-term potentiation and depression, LTP, LTD);

6. Neurophysiological dynamics of single cells including adaptation, spatial and temporal summation of inputs, nonlinear transformation of membrane potential into neuronal output (firing rate);

7. Constant presence of uniform uncorrelated white noise in all parts of the network.

In previous work [9,13-16] we have shown that networks that include this range of realistic neurobiological features tend to exhibit the formation of (input-specific) memory circuits, corresponding to what Hebb once postulated and labelled “Cell Assemblies” [17] or CAs. Below we identify the brain areas that the model simulated, and review the neuroanatomical evidence that we used as a basis to establish the presence, in the network, of direct links between pairs of such areas. The set of sensorimotor brain areas modelled include the left ventral visual “what”-system (blue-shaded areas in Fig. 1A) and the hand/arm motor system (red-shaded areas), involved in processing, respectively, visual object identity [18,19] and manual actions [20-23]. We take it that associative learning mechanisms at work in this system support the development of CAs binding together specific visual stimuli to hand/arm movements, as required, e.g., to pair up visual identity of objects with associated reaching and grasping actions [24-27], or, more generally, to acquire conditional visuomotor associations arbitrarily mapping visual features of objects to specific actions [28-31]. The neural model implemented (Fig. 1.B) reflects structure and connectivity of the above cortical areas. In particular, in the frontal lobe the hand/arm motor area M1 (in Brodmann Area, BA 4), adjacent PM (BA 6) and more rostral PF (BA 8/9/46) areas are reciprocally connected [20-22, 32-34], with cortico-cortical links documented between PF and M1 [35,36]. In addition, the primary visual, V1 (BA 17), temporo-occipital, TO (occipital lobe, inferior parts of BA 18/19, and posterior parts of the inferior and middle temporal gyri, BA 37) and anterior-temporal, AT (temporal pole, BA 38, and middle...
parts of the middle and inferior temporal gyri, BA 20/21) areas (Fig. 1, blue-shaded areas) are reciprocally connected [37,38]. Direct (jumping) links from V1 to anterior temporal (AT) regions via the inferior longitudinal fascicle have also been documented [39, 40]. Neuroanatomical [41, 42], inactivation [43-45] and lesion [29,46,47] studies in the monkey indicate the presence of direct connections also between anterior-temporal and prefrontal cortices (AT and PF in Fig. 1.B). Evidence suggests the presence of direct links (via the external capsule) also between temporoorbital (TO) and prefrontal areas (PF) [48-50] and between anterior/middle temporal (AT) and premotor (PM) areas [32]. This between-area connectivity structure is accurately replicated in the model (see Fig. 1.B). In addition, local, within-area connections realised in the model reflect connectivity features that are shared by sensory and motor systems of the mammalian brain [51,52].

B. Input Patterns

We used three different types of input pattern depending upon the experiment. These are described in the subsections below.

1) Neural Model Representation (NMR)

In the standalone model of [9] input patterns are generated within the model code itself as 25x25 pixel pseudo-random patterns with minimal overlap to guarantee best training of CAs (henceforth referred to as Neural Model Representations or NMRs). We used this type of pattern in Experiment 1 for verification of the Cell Assembly learning (see section IV A). See Fig. 2 for an example of an NMR encoding for one pattern pair.

![Example Neural Model pattern pair](image)

**Fig. 2.** Example Neural Model pattern pair a) visual, b) motor

2) iCub Simulator Representation (ISR)

We also required encodings to represent images of objects and grasp poses within the simulator. We used a repertoire of 6 objects with shapes loosely matching those of everyday real objects we plan to use in the final experiments with real iCub (Ball, Can, Cup, Pen, Peg, Phone) and that require different grasping actions (‘power’ or ‘precision’). Fig. 5 shows an example of iCub looking at one of the objects generated in the simulated world. The visual encoding was generated from iCub’s view of an object by performing the following image-processing steps using OpenCV: convert from RGB to greyscale, Canny edge detection, thresholding to produce a binary image and resizing to 25x25. Grasping poses were generated by taking the joint angles from the iCub head, right arm and torso required to make a particular grasping action, totalling 18 Degrees Of Freedom (DOF). These were then encoded as a 25x25 pixel pattern by representing the DOF by rows 0-17 (leaving 7 rows unused but with the potential to add more DOF later) and the value of the joint angle as the x pixel using the formula in (1).

\[
x = \text{floor} \left( \frac{(\text{angle} - \text{min angle})}{(\text{max angle} - \text{min angle})} \times \text{max x} \right)
\]

where: \(x\) is the calculated x pixel value for the DOF, \(\text{angle}\) is the joint angle for the DOF, \(\text{min angle}\) and \(\text{max angle}\) are the allowed joint ranges for the DOF, \(\text{max x}\) is the maximum x pixel used (as the joint ranges vary for each DOF we do not always use all 25 x pixels in order to make the calculation more convenient). The ‘floor’ operation rounds down to the nearest whole pixel value. These patterns are henceforth referred to as iCub Simulator Representations or ISRs. For the simulator experiments the ISR patterns were used as an intermediate step in identifying the training information to be sent to the neural model: the actual neural model training was done with patterns in NMR form as described in Section II B 1). The reason for this was that the neural model had previously only been trained with the pseudo-random patterns and more realistic data presented problems in that a small overlap between patterns and a similar number of ‘on’ pixels per pattern could not be guaranteed. In the iCub controller code a simple translation between representations is done so that recognition of the object seen and the correct grasp pose to execute are done using the ISR but the data sent to the neural model are in NMR form. See Fig. 3 for an example of the ISR encoding for the ‘Ball’ object and its corresponding grasp pattern.

![Simulator representation for the Ball object](image)

**Fig. 3.** Simulator representation for the Ball object a) visual, b) motor

3) Real data Representation

Given the aforementioned issues with using real data, a separate strand of our work has been to investigate a parameter tuning of the model and/or data pre-processing that allows training with real world data so that we can use the real iCub and real visual input directly with the model rather than using intermediate steps as described in the previous section.
We report our preliminary results for this in Experiment 3 (Section IV C). For this we created a visual encoding from six real images selected from the Amsterdam Library of Object Images (ALOI)\(^3\) (Ball, Can, Cup, Peg, Pen, Phone) [53]. OpenCV was used to perform basic image processing (convert from RGB to greyscale, Canny edge detection, thresholding to produce a binary image and resizing to 25x25). To achieve successful training with the neural model also required that some noise be added to the resulting pixel patterns (some pixels added and some removed) to break up the outlines. See Fig. 4 for an example of the processing stages from image to final pattern for the Cup object. We used the same encoding for the motor part of the pattern as described in Section II B 2) without any further processing.

\[\text{Fig. 4. Image processing of a real image of a cup to training pattern}\]

III. THE ICBU SIMULATOR AND YARP

The iCub simulator is an accurate physical simulation of the real iCub allowing prototyping of applications that can easily be transferred directly to the real robot with minimal changes. See Fig 5. As well as basic motor control and visual processing, for more complex tasks iCub is integrated with external libraries such as the OpenCV image processing framework. For communications, iCub uses YARP, a generic and flexible protocol which we have used to connect with the neural model (a standalone C program).

\[\text{Fig. 5. The simulated iCub and world}\]

Fig. 6 gives an overview of the integrated system. The communication is handled through a main iCub/YARP C++ program which retrieves visual and motor information from the simulator via YARP ports (oval shapes in Fig. 6.). This program also manages the iCub ‘world’ by adding and removing objects and issuing movement commands. The main program also creates and manages the YARP port that the external model connects to in order to receive training data. At the model end this port is connected to as if it were a regular socket but with some special handshaking commands required to establish the protocol for exchanging information with a YARP port.

IV. EXPERIMENTS AND RESULTS

A. Experiment 1 - Verification of Cell Assembly learning

We firstly wanted to verify that the neural model could learn associations between a candidate set of 6 visual and motor inputs (in NMR form) well enough for use in a practical robotics scenario. Firstly, that the network could encode 6 robust Cell Assemblies concurrently and secondly that the pattern completion ability was good enough so that when only the visual part of the pattern was presented, the network would produce an unambiguous motor response. To achieve robust learning of 6 CAs required some parameter retuning from the original study of [9] which was optimised for 12 patterns. We investigated various parameters (e.g. Noise, Global inhibition, Learning Rate) and found that the Noise parameter had the greatest effect on learning. We increased noise from 145 to 200 in our networks which gave a robust and consistent performance. We also found that fewer training presentations were required: minimum 150-200 presentations of each pattern to see learning of CAs as opposed to a minimum of 500 in [9]. Training in the neural model proceeded as described in [9]. In summary: A pattern pair was selected randomly from the set of 6 (with a mechanism to ensure that roughly equal numbers of each have been presented by the end of the run). The visual part was applied to area 1 (V1 in Fig. 1.B) while the motor part was presented simultaneously to layer 6 (M1 in Fig 1.B). This input was presented for a fixed number of model timesteps followed by a fixed number of ‘noise’ presentations. Then the process was repeated for another pattern until the desired number of presentations of each pattern had been achieved. For Experiment 1 we performed 7 independent runs to be sure that the performance was robust to different network initialisations, and trained each network with 150 presentations of each pattern pair (900 presentations in total per run). For a first analysis we presented only the visual half of the input and observed the network response. See Fig. 7 for an example of the network activation when the visual component of a pattern was presented and the target motor component for comparison. In this figure the left hand side shows the visual input (first box) and activation in area 1 (V1) through to area 6 (M1). The right hand side of the figure shows the visual pattern and the target motor pattern. By visual inspection, for all of the 6 patterns a cell assembly was present (a clear subset of neurons in each area were activated all the way through to area 6). For a quantitative analysis we calculated the percentage pixel match between the activated motor response and all possible target patterns – here we were trying to assess if the pattern completion was good enough to distinguish between the possible responses.

\(^3\) http://aloi.science.uva.nl/
Table I shows a matrix of all possible matches with values averaged over all runs (to nearest whole %). Although in many cases there is some activation of pixels from multiple target patterns, in all cases the pattern match with the intended target is the highest (diagonal in Table I). Therefore we concluded that the network could learn Cell Assemblies for 6 patterns well enough for use with the iCub simulator.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>71</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>84</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>54</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>76</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>71</td>
</tr>
</tbody>
</table>

B. Experiment 2 - Learning with the iCub simulator

This experiment consisted of two parts: firstly training online with the iCub simulator generating the inputs to the network in real time and secondly testing where only the visual input was applied to see if the network could generate the correct motor response to pass back to the simulator.

The training procedure:
1. Set iCub in ‘ready’ pose (See Fig. 5)
2. Select a random number between 1-6
3. Create the corresponding object in the world
4. iCub looks at object
5. Image of object is processed to work out what iCub saw (ISR)
6. iCub grasps object using appropriate pose (ISR)
7. Appropriate NMR data sent to neural model
8. Network is trained
9. Set iCub back to ready pose
10. Repeat 2-9 as required

The testing procedure:
1. Set iCub in ‘ready’ pose (as for training sequence)
2. Select a random number between 1-6
3. Create the corresponding object in the world
4. iCub looks at object
5. Image of object is processed to work out what iCub saw (ISR)
6. NMR data for visual input is sent to neural model (motor is blank)
7. Neural model produces a motor response
8. Completed pattern pair sent back to iCub
9. iCub translates NMR to ISR representation
10. iCub grasps object
11. Set iCub back to ready pose
12. Repeat 2-11 as required

Fig 8. Shows example screenshots of the training and testing scenarios. We trained a network for 900 presentations (150 for each of the 6 objects) and then in testing mode presented 50 randomly selected objects. We collected data on the actual object presented, whether it was recognised correctly as one of the candidate objects and whether the motor response back from the neural model was correct. We found that the object was recognised correctly in 100% of the cases. The motor response generated by the model was correct for 94 % of the cases. Because the NMR representation is translated to ISR at the iCub end a valid grasping pose is always executed even if it is the incorrect pose for the object.
C. Experiment 3 – Training with Real Data

As mentioned previously, the real dataset inevitably has more overlap between patterns and (for the visual inputs) different numbers of ‘on’ pixels per pattern. We found that the best regime to learn the CAs stably and avoid merging of the responses was with further increased noise (220) and increased global inhibition (100 instead of the original value of 85). We also found that more training presentations were required (approximately 400 per pattern; 2400 in total). Similarly to Experiment 1 we performed 7 independent runs and trained each network with 400 presentations of each pattern pair. Again patterns were selected randomly, but balanced to ensure roughly equal presentations of each pattern. On presentation of only the visual input to a trained network, we calculated the percentage pixel match between the activated motor response and all possible target patterns. Table II shows a matrix of all possible matches with values averaged over all runs (to nearest whole %). In most cases the diagonal values are the highest with the exception of the match of the Pen object which is also activating the Peg object quite strongly.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Ball</th>
<th>Can</th>
<th>Cup</th>
<th>Peg</th>
<th>Pen</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ball</td>
<td>38</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Can</td>
<td>6</td>
<td>33</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Cup</td>
<td>2</td>
<td>4</td>
<td>38</td>
<td>18</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Peg</td>
<td>1</td>
<td>1</td>
<td>24</td>
<td>63</td>
<td>38</td>
<td>7</td>
</tr>
<tr>
<td>Pen</td>
<td>1</td>
<td>1</td>
<td>28</td>
<td>51</td>
<td>51</td>
<td>8</td>
</tr>
<tr>
<td>Phone</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>34</td>
</tr>
</tbody>
</table>

Although they are generally good enough to distinguish between the different CAs, the pixel match percentages are lower than those for Experiment 1 reflecting the difficulties with using more realistic data.

V. DISCUSSION AND CONCLUSIONS

We demonstrated how it was possible to ‘plug and play’ an existing neural model with a robotic simulator. However, there was a trade-off between the convenience of re-using an existing model and the compromises needed to get it working with ‘real’ data. Although minimal changes were needed to the neural model code (mainly in the communications and I/O) there were limitations in the kind of patterns that could be successfully learned as the model was not originally designed for use in this type of scenario. We have made progress in showing that, with suitable processing and parameter tuning it would be possible for the model to learn more realistic data but more work needs to be done in fully integrating this with real iCub: in particular, a method for online presentation and processing of the images of real objects. We also found (at least with the current setup) that motor patterns are never 100% completed (best individual case was 94 %). In the current simulator setup the motor response from the network is translated back via its ISR form thereby ensuring that iCub always receives a valid motor command, but for future implementations it would be desirable to use the output directly from the network without recourse to intermediate processing. The issue of appropriate pre-processing and encoding of visual, motor and auditory data is relevant to the next steps for BABEL as it is likely that any spiking version of the model will also be sensitive to overlaps and varying information content in the input patterns. The principle of including noise and employing a global inhibition mechanism may also work well for this type of network.

An important next step for our prototype is to extend to the upcoming 12-layer architecture incorporating auditory and motor articulatory input so that experiments with object naming can be done.

As well as the presented results, our takeaway message is that there are advantages to be had in this type of plug and play integration of neural models and robots (simulated or otherwise) rather than creating whole systems from scratch. In technical terms integration need not be difficult but the biggest issue is the compromises required depending upon assumptions made in the respective components. Even if such systems are not likely to provide robust permanent solutions they are ideal as intermediate steps to explore candidate neural models.

REFERENCES


