A Multimodal Hierarchical Approach to Robot Learning by Imitation

Cornelius Weber, Mark Elshaw, Alex Zochios and Stefan Wermter
Hybrid Intelligent Systems
School of Computing and Technology
University of Sunderland UK
[Cornelius.Weber,Mark.Elshaw,Stefan.Wermter]@sunderland.ac.uk

Abstract

In this paper we propose an approach to robot learning by imitation that uses the multimodal inputs of language, vision and motor. In our approach a student robot learns from a teacher robot how to perform three separate behaviours based on these inputs. We considered two neural architectures for performing this robot learning. First, a one-step hierarchical architecture trained with two different learning approaches either based on Kohonen's self-organising map or based on the Helmholtz machine turns out to be inefficient or not capable of performing differentiated behaviour. In response we produced a hierarchical architecture that combines both learning approaches to overcome these problems. In doing so the proposed robot system models specific aspects of learning using concepts of the mirror neuron system (Rizzolatti and Arbib, 1998) with regards to demonstration learning.

1. Introduction

Intelligent robots that are easy to use require a learning approach such as imitation learning which allows the observer to gain skills, by creating an abstract representation of the teacher's behaviour, understand the aims of the teacher and produce the solution (Infantino et al., 2003). There is growing interest in imitation as it offers a flexible way to programme robots by having the robot observe and imitate either another robot or a human.

Multimodal inputs are used in our robot learning model as it is only through the combination of language, vision and motor actions, that robots will be able to become service robots to benefit humans. By combining multimodal inputs social robots should adapt to changes in their environment and improve their decision-making.

In response to this various multimodal approaches have been used. For instance, (McGuire et al., 2002) developed a robot to perform grasping operations based on language, gestures and vision. (Roy and Pentland, 2002) develop a language acquisition model that is able to learn words based on raw multimodal sensory data. A mirror neuron approach using multimodal inputs devised by (Demiris, 2002) was applied to behaviour prediction. In this approach the behaviour model was given information on the current state and the goal and produces the required motor commands. The forward model then created the expected next state based on the output from the behaviour model. The predicted state was compared with the actual state of the demonstrator that produced an error signal to establish confidence values for particular behaviours. Our approach adds the language element in the mirror neuron system to achieve learning by imitation.

A class of the neurons in the F5 motor area of the monkey cortex not only fire when performing an action but when seeing or hearing the action performed. The role of these *mirror neurons* is to represent actions in an abstract sense so they are understood or can be imitated (Rizzolatti et al., 2002). Mirror neurons in humans (Gallese and Goldman, 1998) have been associated with Broca's area (Rizzolatti and Arbib, 1998) which indicates the role played by mirror neurons for language development.

2. Methods

A robot simulator was produced with a teacher robot performing 'go', 'pick' and 'lift' actions one after another in an environment (Fig. 1).

The student robot observed the teacher robot performing the behaviours and was trained by receiving multimodal inputs. These multimodal inputs were:

- higher-level visual inputs which were the x and y coordinates and the rotation angle φ of the teacher robot relative to the front wall,
- the motor directions of the robot ('forward', 'backward', 'turn left' and 'turn right') and

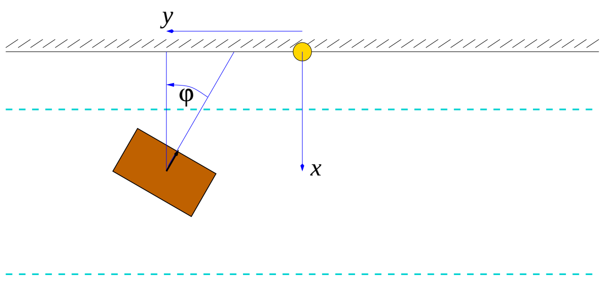

Figure 1: The simulated environment containing the robot at coordinates x, y and rotation angle φ. The upper dashed line indicates where the robot turns away from the wall in the 'go' behaviour. The lower dashed line indicates where the robot turns in the 'lift' behaviour.

- a language description stating the behaviour the teacher is performing ('go', 'pick' or 'lift').

The coordinates x and φ in the 'go' behaviour ensure that the robot avoids the wall, irrespective of y. Coordinates x, y and φ are relevant for the 'pick' action where the robot moves to the coordinate origin to grasp a target. For the 'lift' behaviour, coordinates x and φ determine how far to move backward and in which direction to turn around. These coordinates which are shared by teacher and learner are chosen such that they could be retrieved once the imitation system is implemented on a real robot.

The first behaviour, 'go', involves the robot moving forward in the environment until it reaches a wall and then turns away from it. The second behaviour, 'pick', involves the robot moving toward the target object depicted in Fig. 1 at the top of the arena. This "docking" procedure is produced by a reinforcement approach as described in (Weber et al., 2004). The final behaviour, 'lift', involves moving backward to leave the table and then turning around to face toward the middle of the arena.

The simulated teacher robot performs the three behaviours one after another in a loop. At the start of each behaviour the robot is initialised at random x, y and φ co-ordinates. When receiving the multimodal inputs corresponding to the teacher's actions the student robot was required to learn these behaviours so that it could recognise them in the future or perform them based on a language instruction. Two neural architectures were considered for performing the imitation learning.

2.1 Choice of Architecture

The first architecture depicted in Fig. 2 was run with two different learning algorithms. In (Elshaw et al., 2004) we have used a winner-take-all mechanism on the hidden area. In this self-organising model, however, any hidden unit must "explain" all input modalities at once, i.e. if there are differences in the input in only one modality, then additional hidden units are needed. If one behaviour would be described by several different words, then it would have to be represented several times on the hidden area.

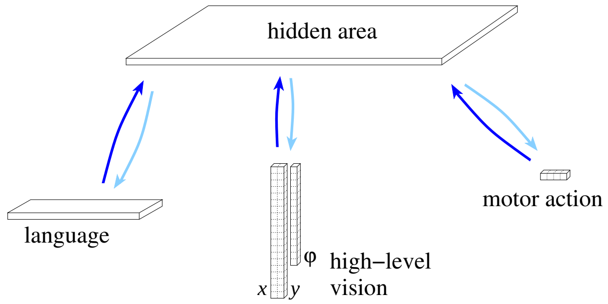

Figure 2: A single-step (3-to-1) architecture.

For a more efficient hidden representation, we considered a distributed code to be chosen for the hidden area of Fig. 2 which was done by using a Helmholtz machine learning algorithm rather than the winner-based self-organising map. Here a small number of units might account for the word while others account for the visual-motor representation. This means that some units specialise to account for just one input modality.

Our goal, however, is that if certain sensory input arrives which would lead to different motor output dependent only on the language input then the language input shall deliver the necessary bias to cause the differential activation pattern. This bias needs to be situation dependent, since behaviours differ in different situations, by activating different motor units. But we found that a language unit projects a certain input pattern onto the hidden units, i.e. it biases the hidden unit's activations dependent of the static connection pattern toward them. This bias is thus not situation dependent, since the language area does not receive sensory input.

In response to these identified problems we will concentrate on the architecture represented in Fig. 3. It associates the motor and high-level vision inputs using the first hidden layer, denoted HM area. The activations of the first hidden layer are then associated with the language region input at the second hidden layer, denoted SOM area. The first hidden layer uses a Helmholtz machine learning algorithm (Dayan, 2000) and the second hidden layer uses Kohonen's self-organising map algorithm (Kohonen, 1997). Training of the SOM area weights was done after the HM area weights learning was completed. Such an architecture allows the features created on the Helmholtz machine hidden layer to relate a specific action for one of the three behaviours given the particular higher visual information to "flexible" associations of pairs/patterns of

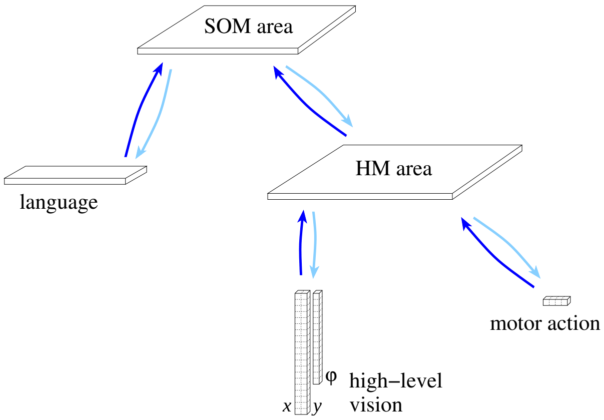

Figure 3: A two-layer hierarchical architecture was used.

activations on the hidden area.

2.2 Training

The multimodal inputs included first the high-level vision which represents the x and y coordinates and rotation angle φ of the robot. The x, y and φ coordinates in the environment were represented by two arrays of 36 units and one array of 24 units, respectively. The centre of a broad Gaussian hill of activation denotes the corresponding coordinate as a neural population code. The language region input was based on a representation of phonemes. This approach used a feature description of 46 English phonemes, as developed by partners in Cambridge based on the phonemes in the CELEX lexical databases (http://www.kun.nl/celex/). A region of 4 rows by 20 columns was used to represent the words with each row representing a phoneme which had 20 phonetic features each. The robot motor directives were presented on the 4 motor units ('forward', 'backward', 'turn right' and 'turn left') with only one active at a time.

The size of the HM hidden layer was 32 by 32 units and the SOM layer was 24 by 24 units. The number of training examples was around 500000. The duration of a single behaviour depended on the initial conditions and may average at around 25 consecutive steps, before the end condition (robot far from wall or target object reached) was met.

During training the student robot received all the inputs, however when testing, either the language area or the motor inputs were not provided. When the student network had to recognise the behaviour that was performed, then the language input was omitted. Recognition was verified by comparing the units which are activated on the language area via W^{td} (depicted light in Fig. 3) with the activation pattern belonging to the verbal description of the corresponding behaviour. When the student robot was required to perform the learnt behaviours based on a language instruction, then the motor input was omitted. It then continuously received its own current x, y and φ coordinates and the language instruction of the behaviour to be performed. Without motor input it had to produce the appropriate motor activations via W^{td} which it had learnt from observing the teacher to produce the required behaviour.

3. Results

First, we have trained a HM area to perform a single behaviour, 'pick', without a SOM area. The robot thereby self-imitates a behaviour it has previously learnt by reinforcement (Weber et al., 2004). Example videos of its movements can be seen on-line at: `www.his.sunderland.ac.uk/supplements/NN04/`

Finally, the full network was trained and sample weights are shown in Fig. 4. One can see that a) the motor units receive input from only small regions in the HM area while b) the SOM units are connected to larger regions. These larger input regions of the SOM units generally comprise one region devoted to one of the motor units and in addition regions devoted to x,y,φ input from high-level vision. The SOM units thus perform feature binding, or association of visually perceived input to a motor command.

The circles drawn into Fig. 4 show in b) that as one progresses along the SOM area along four neurons, their association toward motor units changes: the left unit has a RF overlap with that of the 'turn left' motor unit, but not with that of the 'forward' unit, while the right unit behaves opposite. All of these four selected SOM units are activated during the 'go' behaviour, thus their differential activation reflects different phases within that behaviour.

A part of the 'go' performance of the learner is

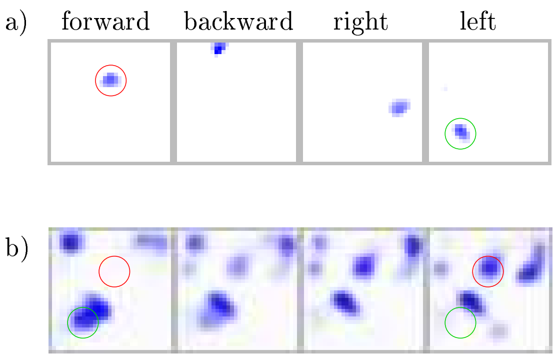

Figure 4: a) The four motor units' receptive fields (RF) in the HM area. Strong weights are depicted dark. Each unit receives input from a narrowly confined region in the HM area. b) Four neighbouring SOM units' RFs in the HM area. Circles indicate that the leftmost units' RFs have an overlap with those of the 'left turn' motor unit while the rightmost unit's RF overlaps with the RF of the 'forward' motor unit.

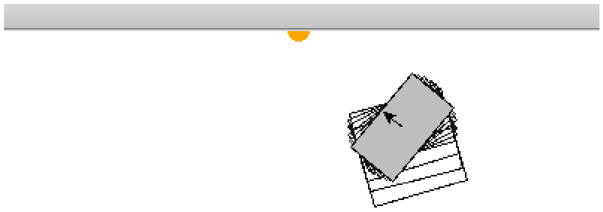

Figure 5: The ten steps of the learnt 'go' behaviour, as the robot starts to turn away from the wall.

shown in Fig. 5. The robot approaches the wall and starts to turn away from it.

4. Discussion

It is suggestive to identify the HM area of the model with area F5 of the primate cortex and the SOM area with F6. F5 represents motor primitives where the stimulation of neurons leads to involuntary limb movements. F6 rather acts as a switch, facilitating or suppressing the effects of F5 unit activations but it is itself unable to evoke reliable and fast motor responses. In our model, the HM area is directly linked to the motor output and identifiable groups of neurons activate specific motor units while the SOM area represents the channel through which a verbal command must pass in order to reach the motor units.

Mirror neurons have so far been reported in F5. By design, our model uses the HM area for both, recognition and production, so an overlap in the activation patterns as observed in mirror neurons is expected. This overlap is mainly due to those neurons which receive high-level vision input. This perceptual input is tightly related to the motor action as it is necessarily present during the performance of an action and contributes to the "motor affordances" (Gallese and Goldman, 1998). The decisive influence on the motor action, however, is localised in our model on smaller regions on the HM area, as defined by the motor units' receptive fields (Fig. 4 a)). Units in these regions would correspond to the canonical motor neurons which do not have mirror neuron properties and which are also found in F5.

In summary, we have developed a hierarchical approach to robot learning by imitation that combines Helmholtz machine and self-organising map learning algorithms in a hierarchical model. The model offers multimodal input processing of vision, language and action, and suggests analogies to the organisation of motor cortical areas F5 and F6 and to the properties of mirror neurons found in these areas.

Acknowledgements This work is part of the MirrorBot project supported by the EU in the FET-IST programme under grant IST- 2001-35282. We thank Prof. Friedemann Pulvermüller and Fermín Moscoso del Prado Martín at the Cognition and Brain Science Unit in Cambridge for their assistance with developing the language phoneme representation.

References

Dayan, P. (2000). Helmholtz machines and wake-sleep learning. In Arbib, M., (Ed.), *Handbook of Brain Theory and Neural Network*. MIT Press, Cambridge, MA.

Demiris, Y. (2002). Biologically inspired robot imitation mechanisms and their application as models of mirror neurons. In *Proceedings of the EPSRC/BBSRC International Workshop on Biological Inspired Robotics*.

Elshaw, M., Weber, C., Zochios, A., and Wermter, S. (2004). An associator network approach to robot learning by imitation through vision, motor control and language. In *Proceedings of International Joint Conference on Neural Networks*.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Science*, 2(12):493–501.

Infantino, I., Chella, A., Dzindo, H., and Macaluso, I. (2003). A posture sequence learning system for an anthropomorphic robotic hand. In *Proceedings of the IROS-2003 Workshop on Robot Programming by Demonstration*.

Kohonen, T. (1997). *Self-Organizing Maps*. Springer Verlag, Heidelberg.

McGuire, P., Fritsch, J., Steil, J., Röthling, F., Fink, G., Wachsmuth, S., Sagerer, G., and Ritter, H. (2002). Multi-modal human-machine communication for instructing robot grasping tasks. In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1082–1088.

Rizzolatti, G. and Arbib, M. (1998). Language within our grasp. *Trends in Neuroscience*, 21(5):188–194.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2002). Motor and cognitive functions of the ventral premotor cortex. *Current Opinion in Neurobiology*, 12:149–154.

Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26:113–146.

Weber, C., Wermter, S., and Zochios, A. (2004). Robot docking with neural vision and reinforcement. *Knowledge-Based Systems*, 17(2-4):165–72.