Available online at www.sciencedirect.com
SCIENCE DIRECT®
Robotics and Autonomous Systems 47 (2004) 171–175
Robotics and Autonomous Systems
www.elsevier.com/locate/robot

Towards multimodal neural robot learning

S. Wermter [a,*], C. Weber [a], M. Elshaw [a], C. Panchev [a], H. Erwin [a], F. Pulvermüller [b]

[a] *Center for Hybrid Intelligent Systems, School of Computing and Technology, University of Sunderland, St Peter's Way, Sunderland SR6 0DD, UK*
[b] *Medical Research Council, Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, UK*

Abstract

Learning by multimodal observation of vision and language offers a potentially powerful paradigm for robot learning. Recent experiments have shown that 'mirror' neurons are activated when an action is being performed, perceived, or verbally referred to. Different input modalities are processed by distributed cortical neuron ensembles for leg, arm and head actions. In this overview paper we consider this evidence from mirror neurons by integrating motor, vision and language representations in a learning robot.

Keywords: Neural networks; Learning robots; Multimodal integration; Vision; Language

1. Introduction

There has been some initial research in learning by language instruction or demonstration [1–3], but this has only played a minor role in intelligent robotics so far. In response to this, our approach [6,7] studies robot learning based on multimodal learning and topological memory organisation. In this paper we show how representations of demonstrating motor actions and language instructions can be integrated and outline an architecture for the integration of motor actions, vision and language representations.

2. Associating multiple modalities

First we provide a general outline of the overall architecture. In the network of Fig. 1 mirror neuron properties [5] evolve among some of the neurons in the top layer. They carry an internal representation $\vec{r}$ of all the inputs below. The inputs are from multiple modalities including higher level representations.

The vector $\vec{l}$ contains language input information. $\vec{pv}$ contains the visual perception which includes the identity and perceived location of a target to be grasped. $\vec{m}$ are the motor unit activations including wheels, gripper and pan-tilt camera. $\vec{ms}$ denotes motor sensory unit activations. $\vec{i}$ are other internal states such as the goal-related value function of the critic used in reinforcement learning.

Thick lines with arrow heads denote the weights. The vertical connections are trained with a sparse coding unsupervised learning scheme similar to the

* Corresponding author. Tel.: +44-191 515 3279; fax: +44-191 515 3553.
E-mail address: stefan.wermter@sunderland.ac.uk (S. Wermter).

0921-8890/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2004.03.011

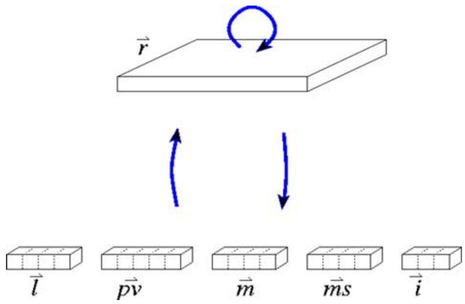

Fig. 1. The overall associative architecture.

Helmholtz machine which we describe for image processing later. The inputs are collected from real robotic actions (after exercising with simulated data) which are performed interactively in the environment. The data is only instantaneous information, i.e. the whole action sequence is not known. Therefore, these neurons do not necessarily fire over a sustained period in time as do mirror neurons. However, since $\vec{r}$ is a distributed code, some of the units may specialise to code for longer sequences. The horizontal recurrent connections (depicted as open circle) are trained as an autoassociator neural network. They are used in a neural activation relaxation procedure which removes noise from the representation $\vec{r}$ and may also encourage prolonged firing. As a possible extension, associator recurrent connections may also feed back to the input. This would be interesting for the cortical feed back to the motor units, because of implications for motor control.

3. Associating motor actions with action verbs

Two concrete examples of this overall architecture have been demonstrated. The MIrror-neuron Robot Agent (MIRA) (see Fig. 2) robot was set up to perform various actions that are associated with the leg, head or hand. Sensor readings were taken while performing a sequence of sub-actions that corresponds to these actions.

First, we associated internal representations of demonstrated actions with a word description. The system accepted two kinds of input: words using a representation of phonemes and demonstrated actions based on sensor readings to represent the semantic features of the action.

Fig. 2. The MIRA recognising and tracking an orange with its camera.

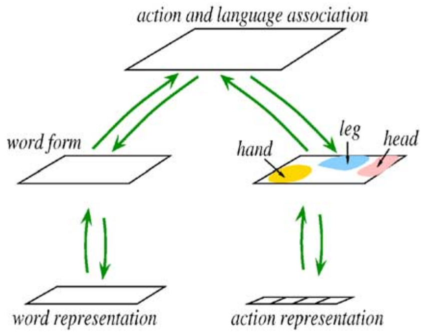

Fig. 3. The self-organising associative architecture.

As can be seen in Fig. 3 the associative architecture uses self-organising networks to associate actions with the appropriate body part and then associates the word form with the action. By associating the action representation with the word form the robot can then produce the action word when receiving the corresponding action input, and vice versa.

Fig. 4 shows an example of self-organising network with 12×12 units. Once this network architecture was trained there was a clear clustering into the three body parts (see Fig. 4). The hand action words were at the bottom of the output layers in the hand body part region, with the head actions slightly below and to the right of the leg region.

4. Associating vision and motor representations

The second concrete example is an associator neural network to localise a recognised object within the visual field. This is an essential basic skill for robotic

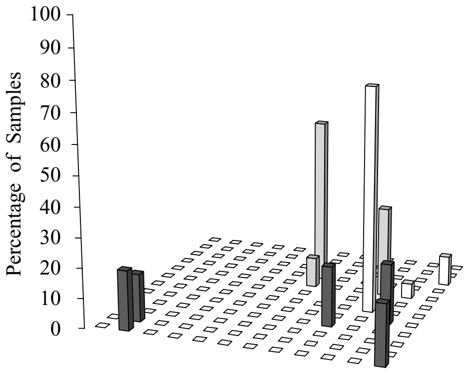

Fig. 4. The percentages for the test samples for the body parts that have the highest activation for each unit on a network (black – hand, white – head, grey – leg).

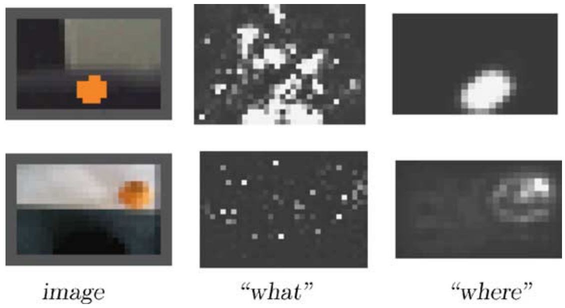

Fig. 6. Example representations on the image, 'what' and 'where' areas. The *image* is originally in colour, where in the upper row, the orange fruit target is artificially generated. The networks of the upper and lower row were trained and activated with different parameters. For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.

learning by demonstration which we solve by a neuron reinforcement approach. The model, depicted in Fig. 5, extends the use of lateral associator connections within a single cortical area to their use between different areas. The first cortical area is the visual area V1 which encodes an internal 'what' representation of images. The weights connecting it to the image are trained by a sparse coding Helmholtz machine. We extend the lateral connections to also span a second cortical area, the 'where' area which is laterally connected to the simulated V1. The lateral weights are trained to associate the V1 representation of the image to the location of an object of interest which is given on the 'where' area.

Fig. 6 shows the network activities after initialisation with sample stimuli of an orange and relaxation to a steady state. The relaxation procedure which spans the 'what' and the 'where' area then completes the pattern by displaying the location of the object of interest as a Gaussian activity hill.

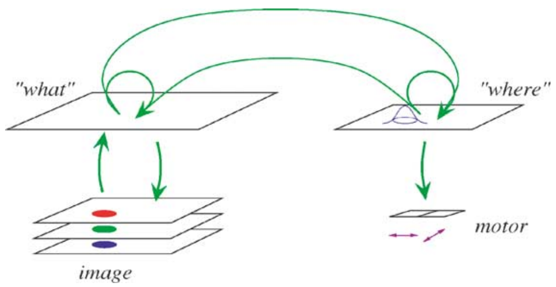

Fig. 5. Model architecture. The hidden representation 'what' of the image including the target object is associated to the location 'where' of the target which is relevant for motor action.

Once that an object of interest appears in the visual field, it is first necessary to localise its position within the visual field. Then, usually the centre of sight is moved towards it, and a grasping movement prototype will be activated.

We connected the 'where' area to motor neuron's output which control the robot camera's pan-tilt motors to centre the orange object. These move the camera so that the orange fruit is located in the centre of the 'where' area (Figs. 5 and 6). Fig. 2 shows the MIRA robot performing the tracking with an orange.

Additionally, using reinforcement learning, we have successfully implemented the task of robot 'docking' at a table so that it can grasp an object which lies at the border of the table with its grippers. The input to the reinforcement-trained network is the perceived target location (from the 'where' area) and the robot rotation angle of the robot relative to the table. Outputs are the four motor units and a critic unit which has a positive value if the target is perceived at the middle of the lower edge of the visual field and the rotation angle is zero. The weights to the value function unit and those to the motor units develop concurrently such that an optimal strategy towards reaching the target will be performed. The data delivered during these actions will be used for the training and verification of mirror neurons.

5. Conclusion

We have developed neural solutions for tasks that need to be solved by a robot that learns by demonstration and instruction. The robot sensor inputs to the modular, self-organising network were partitioned in a way that they match the three body areas ‘leg’, ‘head’ and ‘hand’. This network realises aspects of modularity, because different types of semantic information – head, arm and leg-related information – are projected to different parts of the network and representational space. At the same time, the network processes perceptions, actions and words by distributed neural units that have been linked together in a learning process. The network can in principle realise the findings of Pulvermüller by identifying the semantic features from the actual sensor readings for the individual action verb classes that were specific to the appropriate body part [4].

A recurrent associator network with distributed coding was developed for the visually related part of the task. Such associator networks form the neural basis for multimodal convergence and at the same time can supply a distributed representation across modalities as has been proposed for linguistic structures. Multimodal representations furthermore allow for mirror neuron-like response properties which emerge in our bio-mimetic mirror neuron-based robot.

We think that visual observation and language instructions are complementary forms of guiding robots in a natural manner to perform and link their performance to their own underlying actions. An associative neural organisation of the internal memory may therefore be advantageous for a robot’s learning of visually described actions or verbally instructed actions.

Acknowledgements

This work is partially supported by the MirrorBot project funded by the EU in the FET-IST programme under grant IST-2001-35282.

References

[1] A. Billard, Imitation: a means to enhance learning of a synthetic proto-language in an autonomous robot, in: K. Dautenhahn, C. Nehaniv (Eds.), Imitation in Animals and Artifacts, Academic Press, 2001, pp. 281–311.

[2] Y. Demiris, G. Hayes, Imitation as a dual-route process featuring prediction and learning components: a biologically plausible computational model, in: K. Dautenhahn, C. Nehaniv (Eds.), Imitation in Animals and Artifacts, MIT Press, 2002, pp. 327–361.

[3] G. Maistros, G. Hayes, An imitation mechanism for goal-directed actions, in: Proceedings of the Conference TIMR 2001 on Towards Intelligent Mobile Robots, Manchester, 2001.

[4] F. Pulvermüller, The Neuroscience of Language: On Brain Circuits of Words and Serial Order, Cambridge University Press, 2003.

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

[6] C. Weber, S. Wermter, Object localisation using laterally connected “What” and “Where” associator networks, in: Proceedings of the International Conference on Artificial Neural Networks, Istanbul, Turkey, 2003, pp. 813–820.

[7] S. Wermter, M. Elshaw, Learning robot actions based on self-organising language memory, Neural Networks 16 (5–6) (2003) 661–669.

S. Wermter holds the Chair in Intelligent Systems and is leading the Intelligent Systems Division at the University of Sunderland, UK. His research interests are in intelligent systems, neural networks, cognitive neuroscience, hybrid systems, language processing, and learning robots. He has a diploma from the University of Dortmund, Germany, an MSc from the University of Massachusetts, USA and a PhD and Habilitation from the University of Hamburg, Germany, all in Computer Science. He was a Research Scientist at Berkeley, USA before joining the University of Sunderland. Professor Wermter has written, edited or contributed to 8 books and published about 80 articles on this research area, including books like ‘Hybrid Connectionist Natural Language Processing’ or ‘Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing’, ‘Hybrid Neural Systems’ and ‘Emergent Neural Computational Architectures based on Neuroscience’.

C. Weber graduated in Physics at the University of Bielefeld, Germany, in 1995 and received his PhD from the Technische Universität Berlin in 2000. He was a Postdoc in Brain and Cognitive Sciences at the University of Rochester, USA, in 2001. Since 2002 he is a Research Scientist in Hybrid Intelligent Systems at the University of Sunderland, UK, working on the biomimetic multimodal learning in a mirrorneuron-based robot (MirrorBot) project. His research interests are in computational neuroscience. Using mathematically and biologically motivated approaches such as Helmholtz machines he models learning in the cerebral cortex with particular interest in the visual system and applications in robotics.

M. Elshaw obtained an MSc in Applied Artificial Intelligence from the University of Sunderland, UK in 1996 and MPhil using intelligent approaches to assess the performance of electronic odour sensing systems. Since 1999, he has been a Research Assistant on the Emergent Neural Computational Architectures based on Neuroscience (EmerNet) and Biomimetic Multimodal Learning in a Mirror Neuron-based Robot (MirrorBot) projects at the University of Sunderland, UK. He is also completing a PhD related to natural self organisation for modelling language robot agent behaviour at the University of Sunderland. His research interests are intelligent robots, language processing, grounding language with action association in robots, and cognitive neuroscience.

C. Panchev is a Senior Lecturer at the School of Computing and Technology at the University of Sunderland, UK and in the process of completing his PhD. He received an MSc in Computer Science from the University of Sofia, Faculty of Mathematics and Informatics in 1995. He has done research at the Department of Mathematical Linguistics at the Institute of Mathematics and Department of Artificial Intelligence at the Institute of Information Technologies, Bulgarian Academy of Sciences Sofia, Bulgaria (1996–1997), Faculty of Mathematics and Informatics, and Faculty of Economics and Business Administration, University of Sofia, Bulgaria (1997–1998) and School of Computing and Technology, University of Sunderland, UK (1998–to date). His research interests are neuroinformatics, pulsed neural networks, connectionist modelling, sequential processing and learning in context, natural language processing, intelligent robotics.

H. Erwin is a Senior Lecturer at the School of Computing and Technology, University of Sunderland, UK since 2001. He graduated in Mathematics, at the University of California at Davis, 1968, and gained his PhD in Computational Neuroscience in the Institute for Computational Science and Informatics at George Mason University in 2000. His career in industry included such areas as security engineering for large scale systems, software metrics, performance engineering, software engineering management, and system simulation. He is a Member of the Society for Neuroscience, the ACM, the IEEE Computer Society, and the IEEE. His main research interests are auditory neuroethology, computational neuroscience, and biomimetic robotics.

F. Pulvermüller is Senior Scientist in the Cognitive Neuroscience of Language at the MRC Cognitive and Brain Sciences Unit in Cambridge, UK. He has developed a model of language processing in the human brain which specifies neural circuits underlying the processing of words, their meaning and serial order in sentences. His research focuses on studying language with multiple imaging techniques, including EEG, MEG, fMRI and TMS. He has an MA in Biology, PhD in Linguistics and Psychology, and Habilitation degrees in Psychology and Medicine. He is the author of over 100 publications in the area of cognitive neuroscience including his recent book the 'Neuroscience of Language' published by Cambridge University Press.