Grounding Neural Robot Language in Action

Stefan Wermter[1], Cornelius Weber[1], Mark Elshaw[1], Vittorio Gallese[2], Friedemann Pulvermüller[3]

[1] Hybrid Intelligent Systems
University of Sunderland,
School of Computing and Technology
St Peter's Way, Sunderland, SR6 0DD, UK
Email: [Stefan.Wermter][Cornelius.Weber][Mark.Elshaw]@sunderland.ac.uk
www.his.sunderland.ac.uk
[2] Institute of Neurophysiology, University of Parma, Via Volturno, 39/E I-43100 Parma, Italy
Email: vittorio.gallese@unipr.it
[3] Medical Research Council, Cognition and Brain Sciences Unit,
15 Chaucer Road, Cambridge, UK
Email: friedemann.pulvermuller@mrc-cbu.cam.ac.uk

Abstract. In this paper we describe two models for neural grounding of robotic language processing in actions. These models are inspired by concepts of the mirror neuron system in order to produce learning by imitation by combining high-level vision, language and motor command inputs. The models learn to perform and recognise three behaviours, 'go', 'pick' and 'lift'. The first single-layer model uses an adapted Helmholtz machine wake-sleep algorithm to act like a Kohonen self-organising network that receives all inputs into a single layer. In contrast, the second, hierarchical model has two layers. In the lower level hidden layer the Helmholtz machine wake-sleep algorithm is used to learn the relationship between action and vision, while the upper layer uses the Kohonen self-organising approach to combine the output of the lower hidden layer and the language input.
On the hidden layer of the single-layer model, the action words are represented on non-overlapping regions and any neuron in each region accounts for a corresponding sensory-motor binding. In the hierarchical model rather separate sensory- and motor representations on the lower level are bound to corresponding sensory-motor pairings via the top level that organises according to the language input.

1 Introduction

In order to ground language with vision and actions in a robot we consider two models, a single-layer and a hierarchical approach based on an imitation learning. Harnad 1990 and Harnad 2003 [10, 11] devised the concept of the symbol grounding problem in that abstract symbols must be grounded or associated to objects and events in the real world to know what they actually mean. Hence, in order to actually attribute meaning to language there must be interaction with the world to provide relevance to the symbolic representation. In terms of robotics there is a need to ground actions and visual information with symbolic information provided by language to meaningfully portray what is meant [11]. For instance, the action verb 'lift' could be grounded in the real-world robot behaviour of closing the gripper on an object, moving backward and turning around. The importance of grounding abstract representations can be seen from Glenberg and Kaschak 2002 [9] who found that the understanding of language is grounded in the action, how the action can be achieved and the likelihood of the action occurring.

Although the grounding problem is fundamental to achieve the development of social robots, Roy [29] states that there has not been the grounding of language in actions but abstract representations whose meaning must be interpreted by humans. As a result limited progress has been made in the development of truly social robots that can process multimodal inputs in a manner that grounds language in vision and actions. For instance, robots like the tour-guide robots Rhino [5] and Minerva [34] do not consider grounding of language with vision and actions.

We pursue an imitation learning approach as it allows the observer robot to ground language by creating a representation of the teacher's behaviour, and an understanding of the teacher's aims [14]. As a result of the role played by imitation learning in animal and human development there has been a great deal of interest from diverse fields such as neuroscience, robotics, computation and psychology. Imitation learning offers the ability to ground language with robot actions by taking an external action and relating it with the student robot's internal representation of the action [32]. It is a promising approach for grounding robots in language as it should allow them to learn to cope with complex environments and reduces the search space and the number of training examples compared with reinforcement learning [7].

In our language grounding approach we used the concepts of the mirror neuron system by using multimodal inputs applied to predictive behaviour perception and imitation. Mirror neurons are a class of neurons in the F5 motor area of the monkey cortex which not only fire when the monkey performs an action but also when it sees or hears the action being performed by someone else [23]. Mirror neurons in humans [8] have been associated with Broca's area which indicates their role in language development [23]. Their sensory property justifies our use of models designed for sensory systems that use self-organising learning approaches such as the Helmholtz machine wake-sleep algorithm and the Kohonen algorithm.

2 Robot approaches to grounding language in actions

Other approaches based on imitation learning have been developed to ground robot language in neural action. For instance, Billard 1999 [1] used the Dynamic Recurrent Associative Memory Architecture approach when grounding a proto-language in actions through imitation. This approach uses a hierarchy of neural networks and provides an abstract and high-level depiction of the neurological structure that are the basis of the visuo-motor pathways. By using this recurrent approach the student is able to learn actions and labels associated with them. Experiments were performed using a doll-like robot. The robot can imitate the arms and head movements of the human teacher after being trained to perform a series of actions performed by the teacher and to label this series with a name. The name is entered by using a keyboard attached to the robot. This was also expanded to use proto-sentences such as 'I touch left arm' to describe the actions. The experiments showed that the hierarchical imitation architecture was able to ground a 'proto-language' in actions performed by the human teacher and recreated by the robot.

Vogt 2000 [36] considered the grounding of language in action using imitation in robots through playing games. In the experiment two robots play various language games while one follows the other. The robots are required to develop various categories and a lexicon so they are able to ground language in actions such as 'turn left' or 'go forward'. The robots share the roles of teacher and student, and language understanding is not preprogrammed. The experiments consist of two stages. In the development stage the task is to acquire categories and a lexicon related to the categories. In the test stage the aim is to determine how well the robot performs the task when only receiving the lexicon. In this phase the teacher and student swap roles after each language game. In this imitation learning language approach only the motor signals are categorised as the teacher and student robots have different sensory-motor signals to control their actions. The categorisation achieved is found to be much more successful than the naming.

In addition, a non-imitation approach to grounding language with robot actions developed by Bailey et al. 1998 [3] investigates the neurally plausible grounding of action verbs in motor actions, such that an agent could execute the action it has learnt. They develop a system called VerbLearn that could learn motor-action prototypes for verbs such as 'slide' or 'push' that allows both recognition and execution of a learnt verb. Verb Learn learns from examples of verb word/action pairs and employs Bayesian Model Merging to accommodate different verb senses where representations of prototypical motor-actions for a verb are created or merged according to a minimum description length criterion. However, Bailey's approach makes use of discrete values that rely on opinion rather than on real world values.

3 Neurocognitive evidence as basis for robot language neural grounding

The robot language grounding model developed in this chapter makes use of neurocognitive evidence on word representation. The neurocognitive evidence of Pulvermüller states that cortical assemblies have been identified in the cortex that activate in response to the performance of motor tasks at a semantic level

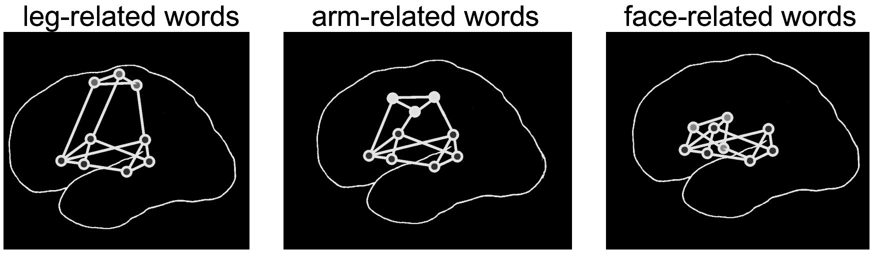

Fig. 1. Based on the brain imaging studies a schematic of the distributed semantic representation in the brain of action verb processing based on the body-parts performing them.

[21, 23, 24]. Accordingly, a cognitive representation is distributed among cortical neuronal populations. Using MRI and CT scans it was found that these semantic word categories elicit different activity patterns in the fronto-central areas of the cortex, in the areas where body actions are known to be processed [24, 12].

Pulvermüller and his colleagues have performed various brain imaging experiments [22, 12] on the processing of action verbs to test their hypothesis on a distributed semantic word representation. From these experiments it has been possible to identify a distributed representation where the activation was different between action verbs based on the body parts they relate to. It was found that there were clustered activation patterns for the three types of action verbs (arm, leg and face) in the left hemispheric inferior-temporal and inferior-frontal gyrus foci. There were also however differences between these three types of action verbs in terms of the average response times for lexical decisions. For instance, the response time is faster for head-associated words than for arm-associated words, and the arm-associated words are faster processed than leg words. Consistent with the somatotopy of the motor and premotor cortex [20], leg-words elicited greater activation in the central brain region around the vertex, with face-words activating inferior-frontal areas, thereby suggesting that the relevant body-part representations are differentially activated when words that denote actions are being comprehended.

These findings suggest the word semantics is represented in different parts of the brain in a systematic way. Particularly, the representation of the word is related to the actual motor and premotor regions of the brain that perform the action. This is evidence for distributed cortical assemblies that bind acoustic, visual and motor information and stresses the role of fronto-central premotor cortex as a prominent binding site for creating neural representations at an abstract semantic level. Fig. 1 shows a schematic view of this distributed representation of regions in the brain activated by leg, arm and face based on the brain imaging experiments.

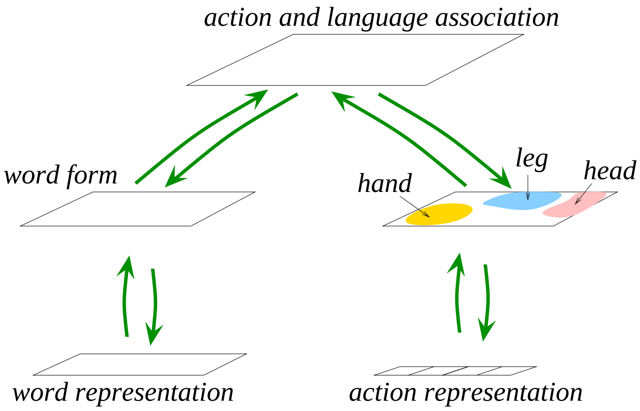

Fig. 2. A neural model of the somatotopy of action words model.

Previously, we have developed a computational model of the somatotopy of action words model that recreates the findings on action word processing [31, 30]. This neural model shown in Fig. 2 grounds language in the actual sensor readings from an autonomous robot. In particular, the actual sensor readings represent semantic features of the action verbs. The approach provides a computational implementation of distributed cell assemblies representing and processing action words along with the actions they can refer to [22]. In the novel architecture presented in this paper, the link between perception and production and between action and language is set up by one single map.

4 Mirror neuron grounding of robot language with actions

The mirror neuron approach offers a biological explanation for the grounding of language with vision and actions. Rizzolatti and Arbib 1998 [23], Gallese and Goldman 1998 [8] and Umilta et al. 2001 [35] found that neurons located in the rostral region of a primate's F5 motor area were activated by the movement of the hand, mouth or both. It was found that these neurons fire as a result of the goal-oriented action but not the movements that make up this action. The recognition of motor actions depends on the presence of a goal and so the motor system does not solely control movement [8, 25]. Hence, the mirror neuron system produces a neural representation that is identical for the performance and recognition of the action [2]. Fig. 3 shows neuronal responses during recognition and performance of object-related actions. The neurons are active during performance of the action (shown for neurons 3 and 4) and during recognition where recognition can be either visual or auditory. These mirror neurons do not fire in response to the presence of the object or mimicking of the action. Mirror neuron responses require the action to interact with the actual object. They differentiate not only between the aim of the action but also how the action is carried out [35]. What turns a set of movements into an action is the goal, with the belief that performing the movements will achieve a specific goal [2]. Such a system requires the recognition of the grasping hand and examination of its movement and an examination of the association of the hand parameters to the position and affordance to reach the object [2].

The role of mirror neurons was to depict actions so they are understood or can be imitated [23]. Furthermore, the mirror neuron system is held to have a major role in the immediate imitation if an action exists in the observer's repertoire [4]. According to Schaal et al. 2003 [33] and Demiris and Hayes, 2003 [7] imitation learning is common to everyday life and is able to speed up the learning process. Imitation can take the form of mimicking the behaviour of the demonstrator or learning how the demonstrator behaves, responds or deals with unexpected events. Complex imitation not only has the capacity to recognise the actions of another person as familiar movements and to produce them, but also to identify that the action contains novel movements that can be approximated by using movements already known. Imitation learning requires learning and the ability to take the seen action and produce the appropriate motor action to recreate the observed behaviour [4].

An explanation proposed by Rizzolatti and Luppino 2001 [26] for the ability to imitate through the mirror neuron system is an internal vocabulary of actions that are recognised by the mirror neurons. Normally the action is recognised even when the final section is hidden [25]. Understanding comes through the recognition of the action and the intention of the indiviudal. This allows the observer to predict the future actions of the action performers and so determine if they are helpful, unhelpful, threatening and to act accordingly [8]. Such understanding of others' actions also allows primates to cooperate, perform teamwork and deal with threats. The mirror neuron system was a critical discovery as it shows the role played by the motor cortex in action depiction [27]. Hence, the observing primate is put in the same internal state as the one performing the action.

The mirror neuron system also exists in humans [8]. Increased excitation was found in the regions of the motor cortex that was responsible for performing a movement even when the subject was simply observing it. Motor neurons in humans are thus excited when both performing and observing an action [8]. The F5 area in monkeys corresponds to various cortical areas in humans including the left superior temporal sulcus of the left inferior parietal lobule and of the anterior region of Broca's area. The association of mirror neurons with Broca's area in human and F5 in primates points to their role in grounding of language in vision and actions [17]. The ability to recognise an action is required for the development of communication between members of a group and finally speech. It is possible that the mirror neuron system was firstly part of an intentional communication system based on hand and face gestures [17] and then in a language based system [23]. Once language became associated with actions it was no longer appropriate for it to be located in the emotional vocalisation centre. It would emerge in the human Broca's area from an F5-like region that had mirror

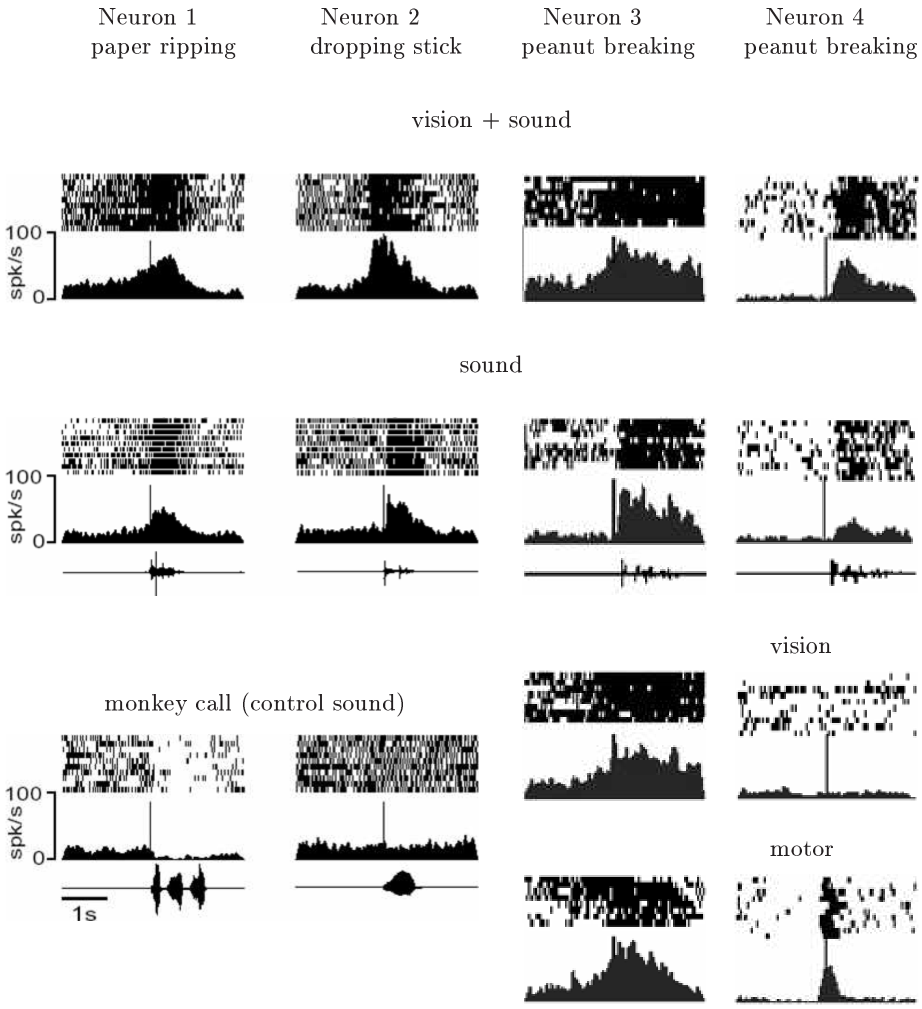

Fig. 3. Responses of macaque F5 mirror neurons to actions. From left to right, the four selected neurons and the chosen stimulus which is in each case a strong driving stimulus. From top to bottom, their responses to vision plus sound, to sound only, and for neurons 3 and 4 to vision of the action only and to the monkey's own performance of the action. For neurons 1 and 2, their reaction to a control sound is shown instead. In each little figure, the above rastergram shows spikes during 10 trials, below which a histogram is depicted (a vertical line indicates the time of the onset of the sound or at which the monkey touches the object). In the case of sound stimuli, an oscillogram of the sound is depicted below the histogram. It can be seen that neurons respond to their driving action via any modality through which they perceive the action, but not to the control stimuli. As an exception, neuron 4 does not respond if the monkey only visually perceives the action. (from Kohler et al. 2002 [17] and Keysers et al. 2003 [15])

neuron features and a gesture system. The importance of gestures reduced until they were seen as an accessory to language [23].

Arbib 2004 [2] examined the emergence of language from the mirror neuron system by considering the neural and functional basis of language and the development of the recognition ability of primates to the full language in humans. In doing so Arbib produced a notion of language development over 7 stages: (i) grasping; (ii) a mirror system for grasping; (iii) a simple imitation system for object grasping; (iv) a complex imitation system that allows the recognition of a grasping action and then repeat; (v) a gesture based language system; (vi) proto-speech and (vii) language that moves from action-object frames to a semantic syntax based approach. Hence, evolution has enabled the language system to develop from the basic mirror neuron system that recognises actions to a complex system that allowed cultural development. This concept of the mirror neurons forms the basis of our models for the grounding of robot language in neural actions. In the remainder of this paper we will consider two models that neurally learn to perform the grounding of language with actions.

5 Methods and Architectures

A robot simulator was produced with a teacher robot performing 'go', 'pick' and 'lift' actions. The actions were performed one after another in a loop in an environment (Fig. 4). The student robot observed the teacher robot performing the behaviours and was trained by receiving multimodal inputs. These multimodal inputs were *(i)* high-level visual inputs which were the x and y coordinates and the rotation angle φ of the teacher robot relative to the front wall, *(ii)* the motor directions of the robot ('forward', 'backward', 'turn left' and 'turn right') and *(iii)* a symbolic language description stating the behaviour the teacher is performing ('go', 'pick' or 'lift').

The first behaviour, 'go', involves the robot moving forward in the environment until it reaches a wall and then turns away from it. The coordinates x and φ ensure that the robot avoids the wall, irrespective of y. The second behaviour, 'pick', involves the robot moving toward the target object depicted in Fig. 4 at the top of the arena. This "docking" procedure is produced by a reinforcement approach as described in [38] and uses all, x, y and φ coordinates. The final behaviour, 'lift', involves moving backward to leave the table and then turning around to face toward the middle of the arena. 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.

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 from a language instruction. Two neural architectures were considered.

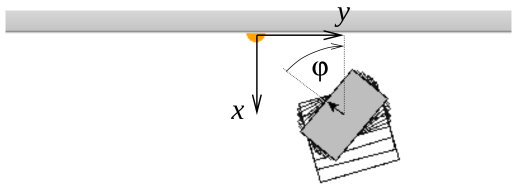

Fig. 4. The simulated environment containing the robot at coordinates x, y and rotation angle φ. The robot has performed ten movement steps and currently turns away from the wall in the learnt 'go' behaviour.

5.1 Single-layer and hierarchical architectures

Both imitation models used an associator network based on the Helmholtz machine approach [6]. The Helmholtz machine generates representations of data using unsupervised learning. Bottom-up weights W^{bu} generate a hidden representation $\boldsymbol{r}$ of some input data $\boldsymbol{z}$. Conversely, top-down weights W^{td} reconstruct an approximation of the data $\tilde{z}$ from the hidden representation. Both sets of weights are trained by the unsupervised wake-sleep algorithm which uses the local delta rule. Parameterised by a sparse coding approach the Helmholtz machine creates biologically realistic edge detectors from natural images [37] and unlike a pure bottom-up recognition model [18] produces also the generative model of the data via neural connections. This is used during testing when we regard either the language area or the motor area as the model's output.

These two models' multimodal inputs included the higher-level vision which represents the x and y coordinates and rotation angle φ of the teacher robot, a language input consisting of a 80-dimensional binary phoneme representation and the motor directives of the four motor units as input.

For the single-layer model all inputs are fed into the hidden layer at the same time during training. The hidden layer of the associator network in Fig. 5 that acted as the student robot's "computatioal cortex" had 16 by 48 units. The sparse coding paradigm of the wake-sleep algorithm leads to the extraction of independent components in the data which is not desired since many of these components would not span over multiple modalities. Therefore we augmented the sparsity toward a winner-take-all mechanism as used in Kohonen networks [18]. The drawback, however, of this winner coding is that the activation of just one unit must account for all input modalities' activations. So if there is a variation in just one modality, for example if an action can be described by two different words, then twice as many units are needed to represent this action. This inefficiency motivates the hierarchical model.

In the hierarchical model there is the association of the motor and high-level vision inputs using the first hidden layer, denoted HM area, which uses sparse but distributed population coding. 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 and the second hidden layer uses Kohonen's self-organising map learning algorithm. Such an architecture allows the features created on the Helmholtz machine hidden layer to relate a specific action to one of the three behaviours given the particular high-level visual information to "flexible" associations of pairs/patterns of activations on the hidden area.

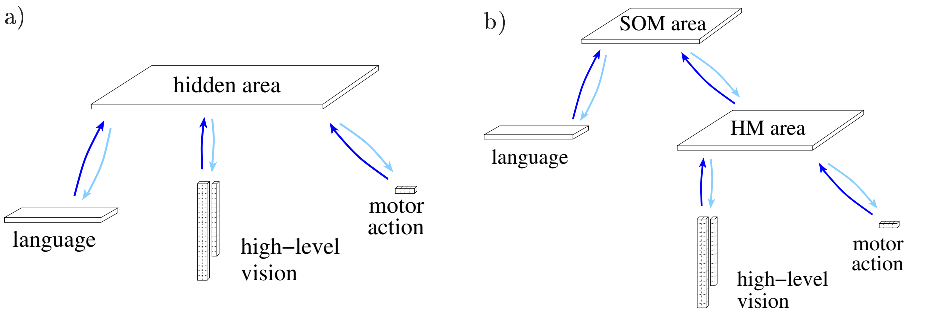

Fig. 5. a) A single-step (3-to-1) architecture. b) A two-layer hierarchical architecture. Bottom-up weights are depicted dark, top-down weights light.

5.2 Processing of data

On the language region representations of phonemes were presented. This approach used a feature description of 46 English phonemes based on the phonemes in the CELEX lexical databases (http://www.kun.nl/celex/). Each of the phonemes was represented by 20 phonetic features, which produced a different binary pattern of activation in the language input region for each phoneme. These features represent the phoneme sound properties for instance voiced or unvoiced, so similar phonemes have similar structures. The input patterns representing the three used words are depicted in Fig. 6.

The higher-level vision represents the x and y coordinates and rotation angle φ of the teacher 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. For a close distance of the robot to the nearest wall, the x position was a Gaussian of activation centred near the first unit while for a robot position near the middle of the arena the Gaussian was centred near the last unit of the first column of 36 units. The next column of 36 units represented the y coordinates so that a Gaussian centred near the middle unit represented the robot to be in the centre of the environment along the y axis. Rotation angles φ from -180^o to 180^o were represented along 24 units with the Gaussian centred on the centre unit if $\varphi = 0^o$.

As final part of the multimodal inputs the teacher robot's motor directives were presented on the 4 motor units (forward, backward, turn right and turn left) one for each of the possible actions with only one active at a time. The activation values in all three input areas were between 0 and 1.

During training the models received all the inputs, however when testing, either the language area or the motor inputs were omitted. The language input was omitted when the student network was required to take the other inputs that would be gained from observing the teacher robot and recognise the behaviour that was performed. Recognition was verified by comparing the units which are activated on the language area via the top-down weights W^{td} (Fig. 5) with the activation pattern belonging to the verbal description of the corresponding behaviour. The motor input was omitted when the student robot was required to perform the learnt behaviours based on a language instruction. 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.

The size of the HM hidden layer is 32 by 32 units and the SOM layer has 24 by 24 units. The number of training steps 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.

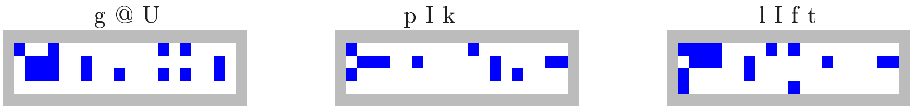

Fig. 6. The phonemes and the corresponding 4×20-dimensional vectors representing 'go', 'pick' and 'lift'.

5.3 Training algorithms

The algorithms used the Helmholtz machine [6] and the self-organising map (SOM) algorithm [18] generate internal representations of their training data using unsupervised learning. Bottom-up weights W^{bu} (Fig. 5) generate a hidden representation $\boldsymbol{r}$ of some input data $\boldsymbol{z}$. Conversely, top-down weights W^{td} an used to reconstruct an approximation $\tilde{z}$ of the data from the hidden representation.

The characteristic of the SOM is that each single data point is represented by a single active ("winning") unit on the hidden area, thus only one element of $\boldsymbol{r}$ is non-zero. The network approximates a data point by this unit's weights. In contrary, the canonical Helmholtz machine's internal representation $\boldsymbol{r}$ contains a varying number of inactive and active, binary stochastic units. A data point is thus reconstructed by a linear superposition of individual units' contributions. Their mean activation can be approximated using a continuous transfer function instead of binary activations. Furthermore, by changing transfer function parameters the characteristics can be manipulated such that units are predominantly inactive which leads to a sparse coding paradigm. At the extreme (that would involve lateral inhibition) one unit might only be allowed to become active at a time.

The learning algorithm for the single-layer model and the HM layer of the hierarchical model is described in the following and consists of alternating wake- and sleep phases to train the top-down and the bottom-up weights, respectively.

In the wake phase, a full data point $\boldsymbol{z}$ is presented which consists of the full motor, higher-level vision and in the case of the single-layer model also language. The linear hidden representation $\boldsymbol{r}^l = W^{bu}\boldsymbol{z}$ is obtained first. In the single-layer model, a competitive version $\boldsymbol{r}^c$ is obtained from this by taking the winning unit of $\boldsymbol{r}^l$ (given by the strongest active unit) and assigning activation values under a Gaussian envelope to the units around the winner. Thus, $\boldsymbol{r}^c$ is effectively a smoothed localist code. On the HM area of the hierarchical model, the linear activation is converted into a sparse representation $\boldsymbol{r}^s$ using the transfer function $r_j^s = e^{\beta x_j}/(e^{\beta r_j^l} + n)$, where $\beta = 2$ controls the slope and $n = 64$ the sparseness of firing. The reconstruction of the data is obtained by $\tilde{z} = W^{td}\boldsymbol{r}^{c/s}$ and the top-down weights from units j to units i are modified according to

$$\Delta w_{ij}^{td} \;=\; \eta\, r_j^{c/s} \cdot (z_i - \tilde{z}_i) \tag{1}$$

with an empirically determined learning rate $\eta = 0.001$. The learning rate was increased 5-fold whenever the active motor unit of the teacher changed. This was critical during the 'go' behaviour when the robot turned for a while in front of a wall until it would do its first step forward. Without emphasising the 'forward' step, the student would learn only the 'turn' command which dominates this situation. Behaviour changes are significant events [13] and neuroscience evidence supports that the brain has a network of neurons that detect novel or significant behaviour to aid learning [16, 19].

In the sleep phase, a random hidden code $\boldsymbol{r}^r$ is produced to initialise the activation flow. Binary activation values were assigned under a Gaussian envelope centred on a random position on the hidden layer. Its linear input representation $\boldsymbol{z}^r = W^{td}\boldsymbol{r}^r$ is obtained, and then the reconstructed linear hidden representation $\tilde{r}^r = W^{bu}\boldsymbol{z}^r$. From this, in the single-layer model we obtain a competitive version $\tilde{r}^c$ by assigning activation values under a Gaussian envelope centred around the winner. In the HM area of the hierarchical model, we obtain a sparse version $\tilde{r}^s$ using the above transfer function and parameters on the linear representation. The bottom-up weights from units i to units j are modified according to

$$\Delta w_{ji}^{bu} \;=\; \epsilon\,(w_{ji}^{bu} - z_i^r) \cdot \tilde{r}_j^{c/s} \tag{2}$$

with an empirically determined learning rate $\epsilon = 0.01$.

The learning rates η and ϵ were decreased linearly to zero during the last quarter of training in order to reduce noise. All weights W^{td} and W^{bu} were rectified to be non-negative at every learning step. In the single-layer model, the bottom-up weights W^{bu} of each hidden unit were normalised to unit length. In the HM area of the hierarchical model, to ensure that weights did not grow too large, a weight decay term of $-0.015 \cdot w_{ij}^{td}$ is added to Eq. 1 and $-0.015 \cdot w_{ji}^{bu}$ to Eq. 2.

Only the wake phases of training involved multimodal inputs from the motor, higher visual and language regions $\boldsymbol{z}$ based on observing the actions of the teacher robot performing the three behaviours. The sleep phases on the other hand use only random initial activations.

The SOM area of the hierarchical model was trained by the classical self-organising map algorithm [18]. The hidden representation $\boldsymbol{o}$ is in our model the activation vector on the SOM area while its input data $\boldsymbol{i}$ is the concatenated vector from the language input together with the HM area activation $\boldsymbol{r}$. Only the bottom-up weights, depicted dark in Fig. 5 b), are trained. Top-down weights are not modelled but can formally be obtained from the bottom-up weights by taking the transpose of the weight matrix. Training of the SOM area weights was done after the HM area weight learning was completed.

The representation o_k of unit k is established by determining the Euclidean distance of the weight vector to its inputs, given by: $o_k = \|\boldsymbol{w}_k - \boldsymbol{i}\|$. The weights are originally randomised and hence a unit of the network will react more strongly than others to a specific input representation. The winning unit is the unit $k^{'}$ where the distance $o_{k'}$ is smallest. The weight vector of this winning unit $k^{'}$ as well as the neighbouring units are altered based on the following equation which leads to the weight vectors resembling more the data:

$$\Delta w_{kj} = \alpha T_{kk'} \cdot (i_j - w_{kj}).$$

The learning rate α was set to 0.01. The neighbour function was a Gaussian: $T_{k,k'} = \exp(-d_{k,k'}^2/2\sigma^2)$, where $d_{k,k'}$ is the distance between unit k and the winning unit $k^{'}$ on the SOM area grid.

At the beginning of training, a larger neighbourhood ($\sigma = 12$) achieved broad topologic learning following a reduction during training to ($\sigma = 0.1$). Additional finer training was done with smaller neighbourhood interaction widths by reducing σ from 0.1 to 0.01.

6 Single-layer model results

The single-layer associator network imitation learning robot performed well when recognising the behaviour being performed by the teacher robot and performing the behaviour based on a language instruction. Recognition was tested by the network producing a phonological representation on the language area which was compared to the appropriate language instruction.

Furthermore, when considering if the trained student robot was able to produce a certain behaviour requested by a language input, the movement traces in Fig. 7 on the next page show that when positioned in the same location the robot performs these different behaviours successfully.

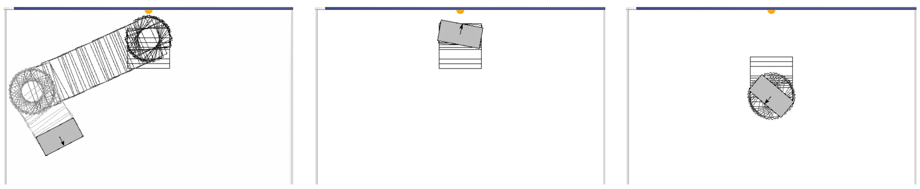

Fig. 7. The simulated trained student robot performance when positioned at the same point in the environment but instructed with different language input. The robot was initially placed in the top middle of the arena facing upward. In the 'go' behaviour it moves around the arena; during 'pick' it approaches the middle of the top wall (target position) and then alternates between left- and right turns; during 'lift' it moves back and then keeps turning.

Fig. 8 indicates that the three behaviours are represented at three separate regions on the hidden area. In contrast, the four motor outputs are represented each at more scattered patches on the hidden area (Fig. 9). This indicates that language has been more dominant in the clustering process.

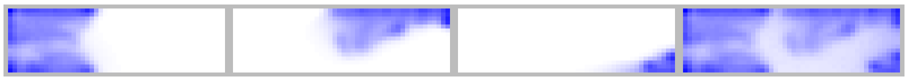

Fig. 8. Trained weights W^{td} to four selected language units of the student robot. Each rectangle denotes the hidden area, dark are strong connections from the corresponding regions. Each of the three left units is active only at one input which is denoted above. The rightmost unit is active at all language words.

7 Hierarchical model results

First, we have trained a HM area to perform a single behaviour, 'pick', without the use of a higher-level SOM area. The robot thereby self-imitates a behaviour it has previously learnt by reinforcement [38]. Example videos of its movements can be seen on-line at: `www.his.sunderland.ac.uk/supplements/AI04/`.

Fig. 10 a) shows the total incoming innervation originating from the motor units (left) and the high-level vision units (right) on the HM area. The figure has been obtained by activating all four motor units or all high-level vision units, respectively, with activation 1 and by displaying the resulting activation pattern on the HM area.

It can be seen that the patches of motor innervation avoid areas of high-density sensory innervation, and vice versa. This effect is due to competitive

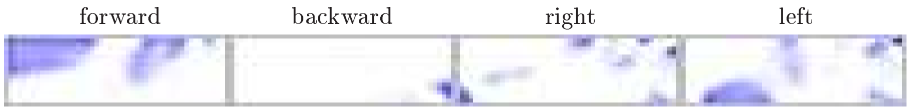

Fig. 9. The trained weights W^{td} to the four motor units of the student robot. As in Fig. 8 the regions from which strong connections originate in the hidden area are depicted dark.

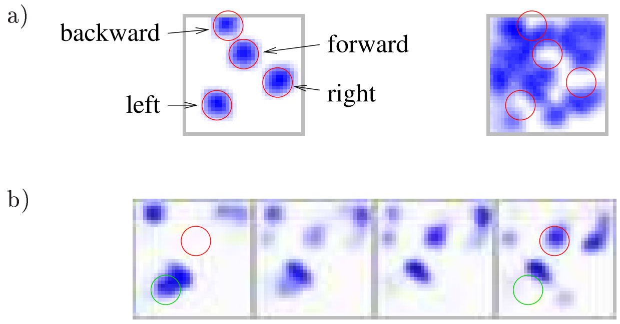

Fig. 10. a) Left, the projections of the four motor units onto the HM area. Right, the projections of all high-level vision inputs on to the HM area. b) Four neighbouring SOM units' RFs in the HM area. These selected units are active during the 'go' behaviour. Circles indicate that the leftmost units' RFs overlap with those of the 'left' motor unit while the rightmost unit's RF overlaps with the RF of the 'forward' motor unit.

effects between incoming innervation. This does not mean that motor activation is independent of sensory activation: Fig. 10 b) shows the innervation of SOM area units on the HM area which bind regions specialised on motor- and sensory input.

The leftmost of the four units binds the "left" motor action with some sensory input while the rightmost binds the "forward" motor action with partially different sensory input. In the cortex we would expect such binding not only to occur via another cortical area (such as the SOM area in our model) but also via horizontal lateral inner-area connections which we do not model.

The action patterns during recognition of the 'go' behaviour action sequence depicted in Fig. 4 and during its performance are shown in Figs. 11 and 12, respectively. At first glance, the activation patterns on the HM- and SOM areas are very similar between recognition and performance which suggests that most neurons display mirror neuron properties.

The largest difference can be seen within performance between the two activation steps of the HM area: in the first step it is activated from vision alone (top row of Fig. 12) in order to perceive the robot state and in the second step it is activated from the SOM area (bottom row of Fig. 12) in order to relay activation to the associated motor unit. The difference between these two steps comes

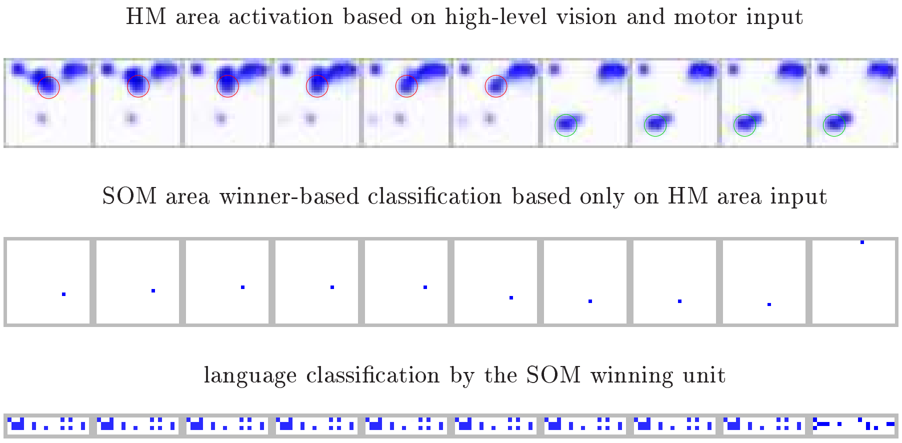

Fig. 11. Activation sequences during observation of a 'go' behaviour, without language input. Strong activations are depicted dark, and shown at ten time steps from left to right. Circles mark the bottom-up input of the active motor unit of the teacher which changes from 'forward' in the first 6 steps to 'turn left' during the last 4 steps (cf. Fig. 10 b)). Language classification is correct except for the last time step which is classified as 'pick' (cf. Fig. 6)).

from the lack of motor input in the first step and the completion of the pattern to include the motor induced activation as would come during full observation in the second step. Naturally, the second step's activation pattern resembles the pattern during recognition in the top row of Fig. 11, since patterns reconstructed from SOM units resemble the training data.

The differences in HM area unit activation patterns during recognition and performance are thus localised at the RF site of the active motor unit. If during training, the input differs only by the motor input (which happens if in the same situation a different action is performed according to a different behaviour) then the difference must be large enough to activate a different SOM unit, so that it can differentiate between behaviours. During performance, however, the absence of the motor input is not desired to have a too strong effect on the HM area representation, because the winner in the SOM area would become unpredictable and the performed action a random one.

The last row in Fig. 11 shows the activations of the language area as a result of the top-down influence from the winning SOM area unit during recognition. An error is made at the last time step which as far as the input is concerned (HM area activation in top row) is barely distinguishable from the second last time step. Note that the recognition error is in general difficult to quantify since large parts of some behaviours are ambiguous: for example, during 'go' and 'pick', a forward movement toward the front wall is made in large areas of the arena at certain orientation angles φ, or a 'turn' movement near the wall toward the centre might also be a result of either behaviour. Inclusion of additional information like the presence of a goal object or the action history could disambiguate many situations, if a more complex model was used.

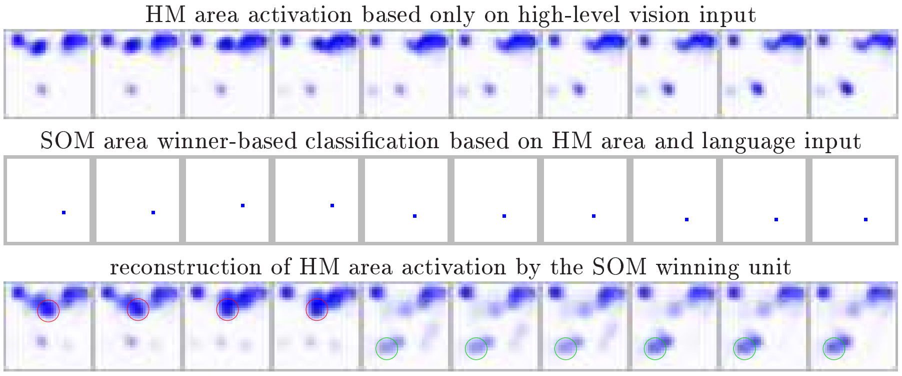

Fig. 12. Activation sequences during performance of a 'go' behaviour, i.e. without motor input. The performed sequence is visualised in Fig. 4. Circles mark the region on the HM area at each time step which has the decisive influence on the action being performed (cf. Fig. 10).

In the examples depicted in Figs. 11 and 12, the teacher and learner robots are initialised at the same position. Both then act similar during the first 4 time steps after which the learner decides to turn, while the teacher turns only after 6 time steps (see the circled areas in these figures).

8 Discussion

Each model recreates some of the neurocognitive evidence on word representation and the mirror neuron system. While for the single layer model a single unit is active at a specific time-step, for the hierarchical model multiple units are active. In terms of the neurocognitive evidence it can be argued that the hierarchical model is closer to the brain as it involves a distributed representation.

The ability of the single-layer and hierarchical model controlled robot to both recognise an observed behaviour and perform the behaviour that it has learnt by imitating a teacher shows the models were able to recreate one core concept of the mirror neuron system. For instance, in the single-layer model the student robot displays mirror neuron properties by producing similar regional unit activation patterns when observing the behaviour and performing it, as seen on some examples in Fig. 13. Furthermore, the achieved common "action understanding" between the teacher and student on the behaviour's meaning through language corresponds to the findings in the human mirror neuron system expressed by Arbib [2] whereby language would be allowed to emerge.

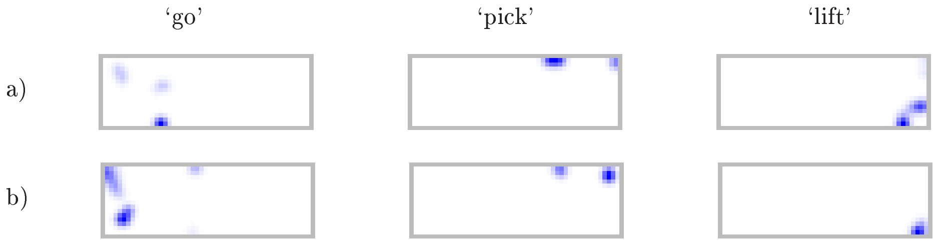

Fig. 13. Activations for the associator network summed up during short phases while the student robot **a)** correctly predicts the behaviours and **b)** performs them based on a language instruction.

With regard to the hierarchical model it is suggested 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 related HM units.

Mirror neurons have so far been reported in F5. By design, the hierarchical 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" [8]. 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. 10 a)). The units in these regions would correspond to the canonical motor neurons which make up one third of F5 neurons. These non-mirror neurons have only motor control function and are not activated by action observation alone.

A prediction of our model would then be that if the visually related mirror neurons alone are activated, e.g. by electrode stimulation, then neurons downstream would not be directly excited and no motor action would take place. It is, however, difficult to activate such a distinguished group of neurons since horizontal, lateral connections in the cortex are likely to link them to the canonical motor neurons.

9 Conclusion

We have developed both a single-layer and an hierarchical approach to robot learning by imitation. We considered an approach to ground language with vision and actions to learn three behaviours in a robot system. The single-layer model relies on a competitive winner-take-all coding scheme. However, the hierarchical approach combines a sparse, distributed coding scheme on the lower layer with winner-take-all coding on the top layer. Although both models offer neural based robot language grounding by recreating concepts of the mirror neuron system in region F5, the hierarchical model suggests analogies to the organisation of motor cortical areas F5 and F6 and to the properties of mirror neurons found in these areas. In doing so it provides insight to the organisation and activation of sensory-motor schemata from a computational modelling perspective. Considering functional processing logics it explains the position of mirror neurons connecting multiple modalities in the brain. This hierarchical architecture based on multi-modal inputs can be extended in the future to the inclusion of reward values that are also represented in cortical structures [28] to achieve goal driven teleological behaviour.

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

References

1. Billard A. and Hayes G. Drama, a connectionist architecture for control and learning in autonomous robots. *Behavior Journal*, 7(1):35–64, 1999.
2. M. Arbib. From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and Brain Science*, pages 1–9, 2004.
3. D. Bailey, N. Chang, J. Feldman, and S. Narayanan. Extending embodied lexical development. In *Proceedings of the Nineteenth Annual Meeting of the Cognitive Science Conference*, 1998.
4. G. Buccino, S. Vogt, A. Ritzi, G. Fink, K. Zilles, H.-J. Freund, and G. Rizzolatti. Neural circuits underlying imitation learning of hand actions: An event-related fMRI study. *Neuron*, 42:323–334, 2004.
5. W. Burgard, A.B. Cremers, D. Fox, D. Hahnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1-2), 2000.
6. P. Dayan. Helmholtz machines and wake-sleep learning. In M. Arbib, editor, *Handbook of Brain Theory and Neural Network*. MIT Press, Cambridge, MA, 2000.
7. Y. Demiris and G. Hayes. Imitation as a dual-route process featuring prediction and learning components: a biologically-plausible computational model. In K. Dautenhaln and C. Nehaniv, editors, *Imitation in animals and artifacts*. MIT Press, 2002.
8. V. Gallese and A. Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Science*, 2(12):493–501, 1998.
9. A. Glenberg and M. Kaschak. Grounding language in action. *Psychonomic Bulletin and Review*, 9:558–565, 2002.
10. S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

11. S. Harnad. The symbol grounding problem. In *Encyclopedia of Cognitive Science*, Dublin, 2003.
12. O. Hauk and F. Pulvermüller. Neurophysiological distinction of action words in the frontal lobe: An ERP study using minimum current estimates. *Human Brain Mapping*, pages 1–9, 2004.
13. G. Hayes and J. Demiris. A robot controller using learning by imitation. In *Proceedings of the 2nd International Symposium on Intelligent Robotic Systems, Greijcnnnoble, France*, 1994.
14. I. Infantino, A. Chella, H. Dzindo, and I. Macaluso. A posture sequence learning system for an anthropomorphic robotic hand. In *Proceedings of the IROS-2003 Workshop on Robot Programming by Demonstration*, 2003.
15. C. Keysers, E. Kohler, M.A. Umilt, L. Nanetti, L. Fogassi, and V. Gallese. Audiovisual mirror neurons and action recognition. *Exp. Brain Res.*, 153:628–636, 2003.
16. R. Knight. Contribution of human hippocampal region to novelty detection. *Computer Speech and Language*, 383(6597):256–259, 1996.
17. E. Kohler, C. Keysers, M. Umilta, L. Fogassi, V. Gallese, and G. Rizzolatti. Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, 297:846–848, 2002.
18. T. Kohonen. *Self-Organizing Maps*. Springer Verlag, Heidelberg, 1997.
19. B. Opitz, A. Mecklinger, A.D. Friederici, and D.Y. von Cramon. The functional neuroanatomy of novelty processing: Integrating erp and fMRI results. *Cerebral Cortex*, 9(4):379–391, 1999.
20. W. Penfield and T. Rasmussen. *The cerebral cortex of man*. Macmillan, Cambridge, MA, 1950.
21. F. Pulvermüller. Words in the brain's language. *Behavioral and Brain Sciences*, 22(2):253–336, 1999.
22. F. Pulvermüller, R. Assadollahi, and T. Elbert. Neuromagnetic evidence for early semantic access in word recognition. *European Journal of Neuroscience*, 13:201–205, 2001.
23. G. Rizzolatti and M. Arbib. Language within our grasp. *Trends in Neuroscience*, 21(5):188–194, 1998.
24. G. Rizzolatti, L. Fogassi, and V. Gallese. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Review*, 2:661–670, 2001.
25. G. Rizzolatti, L. Fogassi, and V. Gallese. Motor and cognitive functions of the ventral premotor cortex. *Current Opinion in Neurobiology*, 12:149–154, 2002.
26. G. Rizzolatti and G. Luppino. The cortical motor system. *Neuron*, 18(2):889–901, 1995.
27. G. Rizzolatti, G. Luppino, and M. Matelli. The organization of the cortical motor system: New concepts. *Electroencephalography and Clinical Neurophysiology*, 106:283–296, 1998.
28. E.T. Rolls. The orbitofrontal cortex and reward. *Cereb. Cortex*, 10(3):284–94, 2000.
29. D. Roy. Learning visually grounded words and syntax of natural language sk. *Computer Speech and Language*, 16(3), 2002.
30. Wermter S., Weber C., Elshaw M., Panchev C., Erwin H., and Pulvermüller F. Towards multimodal neural robot learning. *Robotics and Autonomous Systems Journal*, 47:171–175, 2004.
31. Wermter S. and Elshaw M. Learning robot actions based on self-organising language memory. *Lecture Notes in Artificial Intelligent*, 16(5-6):661–669, 2003.
32. S. Schaal. Is imitation learning the route to humanoid robots. *Trends in Cognitive Science*, 3(6):233–242, 1999.

33. S. Schaal, A. Ijspeert, and A. Billard. Computational approaches to motor learning by imitation. *Transaction of the Royal Society of London: Serial B, Biological Sciences*, 358:537–547, 2003.
34. S. Thrun, M. Bennewitz, W. Burgard, F. Dellaert, D. Fox, D. Haehnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz. Minerva: A second generation mobile tour-guide robot. In *Proceedings of the IEEE international conference on robotics and automation (ICRA'99)*, 1999.
35. M. Umilta, E. Kohler, V. Gallese, L. Fogassi, L. Fadiga, and G. Keysers, C.and Rizzolatti. I know what you are doing: A neurophysical study. *Neuron*, 31:155–165, 2001.
36. P. Vogt. Grounding language about actions: Mobile robots playing follow me games. In J. Meyer, A. Berthoz, D. Floreano, H. Roitblat, and S. Wilson, editors, *SAB00, Honolulu, Hawaii*. MIT Press, 2000.
37. C. Weber. Self-organization of orientation maps, lateral connections, and dynamic receptive fields in the primary visual cortex. In *Proceedings of the ICANN Conference*, 2001.
38. C. Weber, S. Wermter, and A. Zochios. Robot docking with neural vision and reinforcement. *Knowledge-Based Systems*, 17(2-4):165–72, 2004.