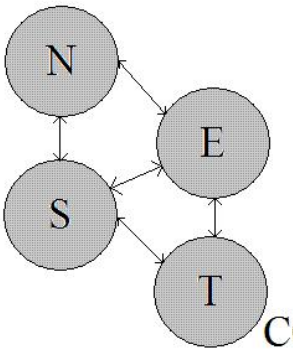

What it means to Communicate

NESTCOM

What it Means to Communicate

Project reference
Contract No: 043374 (NEST)

Neural Control of Actions Involving Different Coordinate Systems (WP3)

NESTCOM Report 5

Deliverable 3

Cornelius Weber, Mark Elshaw, Jochen Triesch and Stefan Wermter

Report Version: 1

Report Preparation Date: 31 August 2007

Classification: Public

Contract Start Date: 1st January 2007 Duration: Two Years

Project Co-ordinator: Professor Stefan Wermter

Project Co-ordinator Organisation: University of Sunderland

Partners: University of Sunderland, Medical Research Council, Universita degli Studi di Parma

Project funded by the European Community under the Sixth Framework Programme NEST -
New and emerging science and technology

Chapter n

Neural Control of Actions Involving Different Coordinate Systems

Cornelius Weber[a], Mark Elshaw[b], Jochen Triesch[a] & Stefan Wermter[b]
[a] *Frankfurt Institute for Advanced Studies, Johann Wolfgang Goethe University, 60438 Frankfurt am Main, Germany*
[b] *Hybrid Intelligent Systems, School of Computing and Technology, University of Sunderland, UK. www.his.sunderland.ac.uk*

1. Introduction

The human body has a complex shape requiring a control structure of matching complexity. This involves keeping track of several body parts that are best represented in different frames of reference (coordinate systems). In performing a complex action, representations in more than one system are active at a time, and switches from one set of coordinate systems to another are performed. During a simple act of grasping, for example, an object is represented in a purely visual, retina-centered coordinate system and is transformed into head- and body-centered representations. On the control side, 3-dimensional movement fields, found in the motor cortex, surround the body and determine the goal position of a reaching movement. A conceptual, object-centered coordinate space representing the difference between target object and hand position may be used for movement corrections near the end of grasping.

As a guideline for the development of more sophisticated robotic actions, we take inspiration from the brain. A cortical area represents information about an object or an actuator in a specific coordinate system. This view is generalized in the light of population coding and distributed object representations. In the motor system, neurons represent motor "affordances" which code for certain configurations of object- and effector positions, while mirror neurons code actions in an abstract fashion.

One challenge to the technological development of a robotic / humanoid action control system is — besides vision — its complexity, another is learning. One must explain the cortical mechanisms which support the several processing stages that transform retinal stimulation into the mirror neuron and motor neuron responses (Oztop et al., 2006). Recently, we have trained a frame of reference transformation network by unsupervised learning (Weber & Wermter, 2006). It transforms between representations in two reference frames which may dynamically change their position to each other. For example the mapping between retinal and body-centered coordinates while the eyes may move. We will briefly but concisely present this self-organizing network in the context of grasping. We will also discuss mechanisms required for unsupervised learning such as requested slowness of neuronal response changes in those frames of reference that tend to remain constant during a task. This book chapter shall guide and inspire the development of sensory-motor control strategies for humanoids.

This book chapter is organized as follows. Section 2 reviews neurobiological findings; Section 3 reviews robotic research. Then, after motivating learning in Section 4, we will carefully introduce neural frame of reference transformations in Section 5, and in Section 6 present a model for their unsupervised learning. Section 7 discusses the biological context, the model's solution for visual routing, and open questions for motor control.

2. Neurobiology

2.1 Cortical Areas Involved in Sensory Motor Control

This section addresses some individual cortical areas, based primarily on macaque data (Luppino & Rizzolatti, 2000), highlighting the variety of frames of reference that exist in the cortex. In lower visual areas such as V1 and V2, neurons are responsive only to visual stimuli shown at a defined region in the visual field, the receptive field. They code in a retinal (eye-centered) coordinate frame. In higher visual areas the receptive fields become larger and can comprise half of the visual field. IT (infero temporal cortex) responses are for example dominated by the presence of an object. The retinal coordinate frame is unimportant, neither is any other spatial frame of reference. MT/MST (middle temporal / medial superior temporal) neurons respond to motion stimuli.

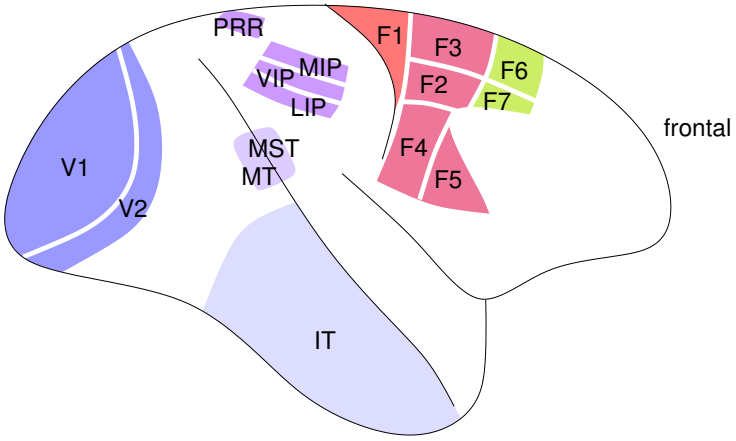

Fig. 1: Cortical areas involved in visuomotor computations. The figure is schematic rather than faithful (drawn after Luppino and Rizzolatti (2000) and Van Essen et al. (1992)).

Of specific interest to frame of reference transformations are the areas of the posterior parietal cortex (PPC):

- LIP (lateral intraparietal) neurons encode locations retinotopically, i.e. in eye-centered coordinates (Duhamel et al., 1997).
- VIP (ventral IP) neurons encode locations in eye- and also in head-centered coordinates (Duhamel et al., 1997). Some cells show response fields that shift only partway with the

eyes as gaze changes (Batista, 2002) ("intermediate reference frame"). Others have a head-centered response: the receptive field is fixed w.r.t. the head, but the response magnitude can be scaled depending on the position of the eyes. Because of this multiplicative interaction these are termed "gain fields".

Many parietal neurons, such as in LIP and VIP, respond when an eye movement brings the site of a previously flashed stimulus into the receptive field (Duhamel et al., 1992). Hence they predict reference frame changes caused by eye movement.

- MIP (medial IP) neurons represent reach plans in a retinal coordinate frame. For example, a neuron fires during reaching movements, but only when the eyes are centered at the reach target (Batista, 2002).
- PRR (parietal reach region) neurons code the next planned reaching movement. In the neighboring area 5 of the PPC, targets are coded w.r.t. both eye and hand (Buneo et al., 2002). Buneo et al. (2002) suggest that the transformation between these two reference frames may be achieved by subtracting hand location from target location, both coded in eye centered coordinates.

The motor cortex (for reviews see Rizzolatti et al. (2001); Luppino and Rizzolatti (2000); Graziano (2006)) is also called agranular cortex, because it misses the granular layer which in sensory areas receives the bottom-up sensory input. The motor areas can be subdivided into two groups. One group of motor areas connects to the spinal cord. These more caudally situated areas transform the sensory information into motor commands:

- F1 projections end directly on motor neurons. Neural activations are correlated with hand position, finger control, and velocity to perform actions such as grasping, with lesion studies showing that damage to this area prevents hand grasping (Fadiga & Craighero, 2003).

F2, F3 and parts of F4, F5 activate preformed medullar circuits and also send connections to F1 (Luppino & Rizzolatti, 2000).

- F2 encodes motor execution and motor preparation and is somatotopically organized. There are few visually responsive neurons ($\approx$16%), mostly within parts of the forelimb representation (Fogassi et al., 1999).
- F3 encodes complete body movements. Stimulation evokes proximal and axial muscle movements; typically a combination of different joints. There are frequent responses to somato-sensory stimulation (Luppino & Rizzolatti, 2000).
- F4 is active at simple movements, e.g. toward mouth or body. It responds to sensory stimulation: 50% of neurons to somato-sensory, 50% to somato-sensory and visual stimuli (Luppino & Rizzolatti, 2000). Visual receptive fields are 3-dimensional, located around the tactile receptive fields, such as on face, body or arms. Hence, there is are egocentric, body-part centered frames of reference.
- F5 controls hand and mouth. Neuronal firing is correlated with action execution, such as precision- or power grip, but not with its individual movements. Some neurons do not respond to vision, some respond to a sheer 3-dimensional presentation of objects (pragmatic representation of graspable objects), finally, "mirror neurons" respond to action observation. They will be described in more detail in Section 2.3.

The other group of motor areas do not have direct connections to the spinal cord, nor to F1. These more frontally situated areas are involved in controlling the more caudal motor areas:

- F6 neurons display long leading activity during preparation of movement. They are modulated by the possibility of grasping an object. Stimulation, only with high currents, leads to slow complex movements which are restricted to the arm. Visual responses are common. It receives input from cingulate cortical areas which are associated with working memory, temporal planning of movements, motivation (Rizzolatti & Luppino, 2001).
- F7 displays visual responses. It may be involved in coding object locations in space for orienting and coordinated arm-body movements (Luppino & Rizzolatti, 2000).

Because of different investigation methods, there are different classification schemes of cortical areas. The primary motor cortex F1 is also referred to as MI. The dorsal premotor area (PMd) comprises F2 and F7. F3 is referred to as supplementary motor area (SMA), and F6 as pre-SMA. The ventral premotor area (PMv) comprises F4 and F5 (Matsumoto et al., 2006).

2.2 Key Aspects of Functionality

We summarize the following general observations:

- There are many frames of reference, accounting for the dynamic complexity of the body. Some frame of reference transformations are likely to depend on the correct functioning of certain others, while others may function in parallel as independent systems.
- Some neurons code in "intermediate" reference frames, such as between eye- and head-centered coordinates. Likewise, neurons with a constant receptive field position (in a certain reference frame) may have their responses modulated by, e.g. eye position ("gain fields").
- There is convergence: For example, an object position can be computed in a body-centered frame from visual input; the hand position can be computed from somato-sensory signals; from these two positions, a motor error (the difference between target object and hand) can be computed. On the other hand, when both hand and object are in sight, this difference can be read directly in retinal coordinates (Buneo et al., 2002). This redundancy may be used to align different frames of reference, or supervise the learning of one representation by the representation of another.

Frame of reference transformations enable the understanding of actions performed by others, as observed in the mirror neurons in F5. This is a prerequisite for social communication, and because of its importance we will discuss the mirror neuron system in the following.

2.3 Mirror Neurons

Rizzolatti and Arbib (1998) and Umilta et al. (2001) describe neurons located in the rostral region of a primate's inferior area, the F5 area (see Fig. 1), which are activated by the movement of the hand, mouth, or both. These neurons fire as a result of the action, not of the movements that are the components of this action. The recognition of motor actions depends on the presence of a goal, so the motor system does not solely control movement (Gallese & Goldman, 1998; Rizzolatti et al., 2002). A seen tool also activates regions in the premotor cortex, an effect which increases when subjects name the tool use (Grafton et al., 1997). The F5 neurons are organized into action categories such as 'grasping', 'holding' and 'tearing' (Gallese & Goldman, 1998; Rizzolatti & Arbib, 1998). More generally, the motor cortex is partly organized around action-defined categories (Graziano, 2006).

Certain grasping-related neurons fire when grasping an object, whether the grasping is performed by the hand, mouth, or both. This supports the view that these neurons do not represent the motor action but the actual goal of performing the grasping task. Within area F5 the pure motor neurons only respond to the performing of the action. Visuomotor neurons also respond to the presentation of an object ("canonical neurons") or when a monkey sees the action performed (mirror neurons) (Kohler et al., 2002; Rizzolatti & Arbib, 1998; Rizzolatti et al., 2001). The mirror neuron system indicates that the motor cortex is not only involved in the production of actions but in the action understanding from visual and auditory information (Rizzolatti et al., 2002; Rizzolatti & Luppino, 2001; Rizzolatti & Arbib, 1998) and so the observer has the same internal representation of action as the actor (Umilta et al., 2001).

These mirror neurons are typically found in area F5c and do not fire in response to the presence of the object or mimicking of the action. Mirror neurons require the action to interact with the actual object. They respond not only to the aim of the action but also how the action is carried out (Umilta et al., 2001). Already an understanding that the object is there without being visible causes the activation of the mirror neurons if the hand reaches for the object in the appropriate manner. This is achieved when primates are first shown the action being performed completely visible and then with the hand-object interaction hidden behind an opaque sliding screen (Umilta et al., 2001). Since these areas are active during both performance and recognition, when simply observing the action, there must be a set of mechanisms that suppress the movements to perform the action.

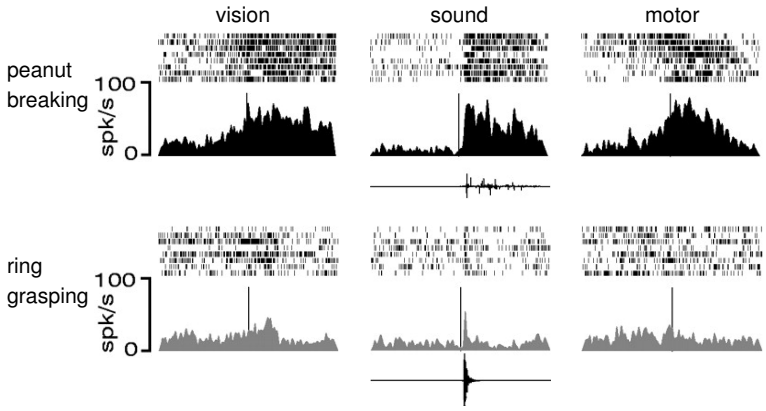

Fig. 2: Responses of a macaque F5 mirror neuron during action recognition and performance (from Kohler et al. (2002), with permission).

Fig. 2 provides mirror neuron responses. The individual pictures show at the top a rastergram that represents the spikes during 10 trials, below which is a histogram (the central line represents onset of stimulus). Recognition is achieved through visual stimuli, left, or auditory stimuli, middle, for which oscillograms are shown below. The motor readings, right, are recordings of the primate when performing the action. It can be seen that this mirror neuron

is active when the monkey breaks a peanut or observes someone performing this action, but not during a control action of ring grasping. These audiovisual mirror neurons have a role in the discrimination of different actions, and constitute together with Broca's area for language representation, a part of a "hearing-doing" system (Lahav et al., 2007).

2.4 Mirror Neurons and Imitation

One possible application for the mirror neuron system is imitation learning. According to Schaal et al. (2003) and Demiris and Hayes (2002) imitation learning is common to everyday life and is able to speed up the learning process. Imitation learning allows the observer to gain skills by creating an abstract representation of the teacher's behavior, understanding the aims of the teacher and creating the solution (Dillmann, 2003). Imitation requires the ability to understand the seen action and produce the appropriate motor primitives to recreate it (Buccino et al., 2004). The role of mirror neurons is to encode actions so they are understood or can be imitated, by gaining the reason for the action (Rizzolatti & Arbib, 1998; Sauser & Billard, 2005a).

A possible explanation for the ability to imitate is the internal vocabulary of actions that are recognized by the mirror neurons (Rizzolatti & Luppino, 2001). This ability to understand others' actions, beliefs, goals and expectations aids the inclusiveness of the group. This allows the observer to predict the actions and so determine if they are helpful, unhelpful, threatening, and to act accordingly (Gallese & Goldman, 1998; Gallese, 2005). It is argued by Demiris and Hayes (2002) that through the mirror neuron system when a primate or human watches an action they are to imitate they put themselves in the place of the demonstrator. Understanding the actions of the demonstrator comes from creating alternatives and choosing the most appropriate one. A requirement for imitation is to connect the sensory system with the motor system so that the multimodal inputs are linked to the appropriate actions.

2.5 Mirror Neurons and Frame of Reference Transformations

If mirror neurons are involved in imitation then they have to mediate between different coordinate systems in order to establish a mapping from the visual representation of another agent's action to an appropriate motor behavior. If the observed action is targeted at a specific object, then a visual representation of agent and object in retinal coordinates must be transferred into an object centered representation of the action. If the object-directed action is then to be imitated, it has to be transfered to the appropriate motor commands taking into account a potentially different relative position of the imitator to the object.

A very instructive example is that of gaze following. Gaze following is the skill of looking to an object because another agent turns to look at it. In this case the object-directed action is simply to look at it. Human infants seem to learn this behavior during their first two years of life presumably because they learn that whereever other people are looking there is typically something interesting to see — such as another person, an interesting object, or the other person's hands manipulating something (Triesch, Teuscher et al., 2006). Consider the situation of an infant and her mother facing each other. When the mother turns to her right to look at something, the infant will see the head of the mother turning. But how does the infant know that she needs to turn her own head to the left and how much she needs to turn to look at the same object? The question is not trivial and in fact infants seem to need more than 18 month

to fully master this skill. The infant needs to learn to associate different locations and head poses of the mother with potential locations in space where the mother may be looking. These locations comprise the line of sight of the mother, which is represented, presumably, in an ego-centric reference frame of the infant. Recently, Triesch, Jasso and Deáak (2006) presented a simple model that demonstrates how such a mapping can be learnt with generic reinforcement learning mechanisms. Specific head poses of the mother become associated with gaze shifts of the infant that have a high probability of matching the location where the mother is looking. This learning process is driven by rewards the infant receives for looking at interesting objects, whose locations are predicted by the looking direction of the mother. Interestingly, the model develops a mirror-neuron-like premotor representation with a population of model neurons that become activated when the infant plans to look at a certain location or when the infant sees the mother looking in the direction of that location. The existence of a similar representation in the brain is the major prediction of the model.

The example demonstrates that the three-dimensional pose of the agent and the imitator and their interrelation need not be fully computed to achieve imitation behavior. On the other hand, Sauser and Billard (2005b) present a model (cf. Section 3.2) in which imitation is performed, only by constructing several consecutive frame of reference transformations, in order to map from an agent- to an imitator coordinate system. How much geometry has to be calculated for imitation?

In another view, mirror neurons represent actions in an abstract fashion, such as when activated by sound (cf. Fig. 2), but are in general unresponsive to the individual movements that make up an action. Hence, they are invariant with respect to the exact geometrical constellations. The demand for invariances makes geometrical frame of reference computations even more challenging, and encourages an alternative view in which the mere presence of features, such as a sound, are detected by mirror neurons with little involvement of geometry. Both views can be reconciled by the notion of convergence (cf. Section 2.2): the result of a geometrical frame of reference computation may be supported by, e.g., the target being successfully focused, or grasped. The error- or reinforcement signal at the end of a successfully completed action can give feedback to the learning of the geometrical calculations.

3. Action in Robotics

The complex geometry of humanoid robots complicates the mapping between sensors and actuators. Additionally, sensors (head, eyes/camera, ears/microphone) may be moved independently of the body on which the actuators are mounted. This is demonstrated in Fig. 3. The robot must know not only the visually seen direction of the object (blue dotted arrow) in retinal coordinates, but also the gaze direction (yellow arrow) which defines the retinal reference frame, before acting in the body-centered reference frame (defined by the long red arrow). Geometrical calculations are complicated by the fact that axes do not coalign, nor do they lie in the centers of sensors.

3.1 Coordinate Transformations in Robotics

There has been a lot of traditional and current research on coordinate transformations in robotics since traditional robotic models rely extensively on converting perceptual input to internal representations and external actions. For instance, at the most recent Intelligent

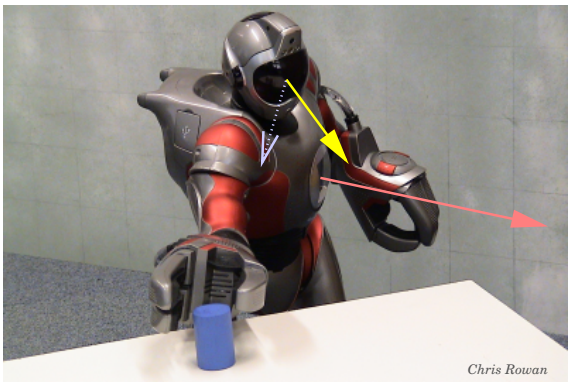

Fig. 3: Grasping by a humanoid robot. The object is localized in a visual reference frame (yellow arrow points into gaze direction) by the direction of the dotted, blue arrow. The grasping is performed in a body-centered reference frame (long red arrow). The relation between visual and body-centered systems varies when the head turns.

Robotics and Systems international conference proceedings, 2006, there are 62 papers addressing "coordinate transformations" in some form. Coordinate transformations are used in traditional control modeling for instance for motion generation, ball catching, sound localization, visual tracking, simultaneous localization and mapping, path following, motion planning, cooperating manipulators, balancing tasks, dance step selection, or pedestrian tracking.

If multiple audio visual and motor maps are all involved in controlling robot behavior these maps have to be coordinated. Most of the existing coordinate transformation approaches in robotics rely on traditional inverse dynamics / inverse kinematics models based on control approaches. However, for new intelligent robotics and in particular humanoid robots with many degrees of freedom researchers look for new and alternative solutions to coordinate transformations. For instance, one important cluster of work relates to the motion generation and motion interpretation for humanoid robots (Harada et al., 2006) and in order to grasp a cup with a handle from a table, the robotic head and the arm have to be coordinated (Tsay & Lai, 2006). For such tasks several different coordinate transformations are involved in a simple grasping action, e.g. including an object coordinate system, a palm coordinate system, gaze coordinate system, arm base coordinate system, and wrist coordinate system etc. Furthermore, if sound maps are involved, for instance different sizes of heads, ear shapes have an influence on the sound maps and coordinate transformations in reality are difficult to calibrate from sound to vision or motor maps (Hoernstein et al., 2006).

Particularly challenging are also coordinate transformations between different maps of different robots, the so-called multi robot map alignment problem. A new approach has been proposed to use the relative pose measurements of each of two robots to determine the coordinate transformation between two maps while assuming that the initial poses of the robots are not known (Zhou & Roumeliotis, 2006).

Besides these traditional approaches based on formal algebraic transformations there are more and more non-traditional approaches to motor control based on biomimetic or cognitive approaches. The general problem of movement control is to map from a cartesian task space to a solution in joint space. Especially for higher degrees of freedom robots with 30-40 degrees of freedom (similar to humans) the task of inverse dynamics of computing these coordinate transformations is very demanding. Therefore, different from traditional approaches some new bioinspired control models have been suggested to address computational efficiency with task independent motor control, only involving those joints which also participate in the task (Gu & Ballard, 2006). This approach by Gu and Ballard proposes that actions are planned in segments with equilibrium end points in joint space. Movements are done by moving between equilibrium points.

Another interesting novel approach has been proposed as a learning-based neurocontroller for a humanoid arm. This approach is based on a self-organizing neural network model and does not assume knowledge about the geometry of the manipulators. Essential is an action-perception cycle which helps to learn transformations from spatial movement to joint movements in the neural map (Asuni et al., 2006). Another new approach to biologically inspired cross modal mapping is further pursued for robotic eye-hand systems (Meng & Lee, 2006) where they incrementally construct mapping networks for eye hand coordination based on extended Kalman filters and radial basis function networks. The system converts the location of a target to an eye-centered coordinate system from which it is mapped via another network into a hand-based coordinate system. Then the difference between the actual and desired position can be computed to steer the motor commands.

3.2 Robot Mirror Neuron System Imitation Learning

The mirror neuron system's principles offer inspiration for developing robotic systems particularly with regards to imitation learning to allow robots to cope with complex environments by reducing the search space (Belpaeme et al., 2003; Triesch et al., 1999).

A mirror neuron-based approach for imitation that relies on learning the reference frame transformation is that of Sauser and Billard (2005b, 2005a). This approach for three-dimensional frames of reference transformations is based on a recurrent multi-layer neural network. This two layer neural network can create a non-linear composition from its inputs. The model includes an attractor network in the first layer with lateral weights, with the second layer containing neurons that receive inputs from a recurrent population but lack lateral connections. This network is able to represent two characteristics, direction and amplitude, in a population vector code. It is able to perform translation or rotation of vectorial activities. To perform non-linear transformation such as rotation there was the need for an intermediary population known as the gain field. A rotation around an axis the rotation is split into three transformations. When carrying out rotation for a reaching activity the target is observed using the visual system and represented using head-centered coordinate neurons. The angle between the head and body is represented in neurons that get proprioceptive information from the appropriate muscles receptors. To produce a movement to the target there is a need to take the head- centered coordinate representation and alter it so it is a body-centered frame of reference vector represented in a population of neurons.

However many of the approaches simplify the problem to get round the need to include a

learnt approach for reference transformation. Takahashi et al. (2006) have put forward a mirror neuron-based model for learning behaviors and others' intentions that does not depend on a precise model of the world or coordinate transforms. This approach relies on reinforcement learning to perform activities including navigation and ball passing with in a modular approach. Once the observer robot has developed the appropriate behaviors through reinforcement learning, the observer watchers the behavior and maps the sensory information from the observer's position to that of the performers based on state variables created during reinforcement. This model is based on a modular learning system made up of behavior modules.

An additional robotics approach that uses imitation learning based on mirror neurons is that of Demiris and Hayes (2002) and Demiris and Johnson (2003) through behavior and forward models. A reference transformation is made in that the observer and the actor robots face each other. The behavior model gives information on the current state and the goal and produces the required motor commands. The forward model then creates the expected next state based on the output from the behavior model. The predicted state is compared with the actual state of the demonstrator to produce an error signal. A confidence value is created from the error signal and used to establish the confidence by which a particular action or series of actions is identified by the observer robot. The architecture not only allows the observer robot to produce the appropriate behavior but also to recognize the behavior being performed by the demonstrator.

A further mirror neuron-based approach is that of Billard and Matarić (2001) who use a hierarchy of neural networks and provides an abstract and high level depiction of the neurological structure that is the basis of the visuo-motor pathways to examine the ability to reproduce human arm movements. The model consists of three parts for visual recognition, motor control and learning and uses seven modules. A module based on the temporal cortex processes visual information to identify the direction and orientation of the teacher's arms with reference to a point on the teacher's body. The temporal cortex model receives as input Cartesian coordinates for the limbs of the person demonstrating the action and this is transformed into the frame of reference of the observer. This transfer is supported by studies that have observed orientation-sensitive cells in the temporal cortex. The motor control is based on a hierarchical model with a spinal cord module at the lower level. Learning of movement occurs in the premotor cortex and cerebellum modules and learning creates links between the primary motor cortex, premotor cortex and the cerebellum and within the premotor cortex and the cerebellum. These modules use a dynamic recurrent associative memory architecture which is a fully connected recurrent network that enables time series and spatio-temporal data to be learnt using short-term memory. The model when tested on series of arm movements is found to reproduce all motions despite the noisy environment.

An additional mirror neuron system based approach for grounding is that of Tani et al. (2004). A recurrent neural network with parametric biases (RNNPB) learns to recognize and produce multiple behaviors with distributed coding using a self-organizing technique. The reference transformation takes the spatial coordinates of the actor's hands which are mapped to the robot's hands using centered cartesian coordinates without learning. In this approach, sections of spatio-temporal data of sensory-motor flow are depicted by using vectors of small dimensions. The nonlinear dynamical system is produced using a Jordan-type recurrent net-

work that has parametric biases (***PB***) incorporated in the input layer function. Learning is achieved through the self-organizing mapping of the ***PB*** and the behavior representation. To reproduce the characteristics of the mirror neuron system, the RNNPB creates the appropriate dynamic pattern from fixed ***PB*** to learn and perform recognition by producing the PB from a target pattern. Movement patterns are learnt using the forward model by producing the ***PB*** vectors and a synaptic weight matrix. Following learning it is possible to produce sensory-motor series by using the forward dynamics of the RNNPB with the parameter biases fixed. When the network produces a behavior it operates in a closed loop where the prediction of the next action is fed back as an input.

As part of the MirrorBot project a mirror neuron-based docking action was generated using a 4-step model. First, feature detectors from the visual input of neurons in a "what" area are learnt unsupervised. Second, associator weights within and between the "what" area and a "where" area are learnt supervised. After training, these two levels visually localize an object in a camera-centered coordinate system. Third, weights to a robot motor output are trained by reinforcement learning to drive the robot to a position to grasp the object (Weber et al., 2004). Since the position of the camera was held fixed, the pixel coordinates could be directly used for motor control, and no dynamic reference frame transformation was needed. In an equivalent approach, Martínez-Marín and Duckett (2005) make the camera always focus at a grasp target, which allows to use the camera gaze angle directly to control the robot. In (Weber et al., 2006) finally, a fourth layer observes the self-performed actions and learns to perform and to predict them based on only visual input. This loosely mimics mirror neurons, but as visual recognition of other robots was not done, neurons are active only when observing self-performance. Another, higher level with additional language input learnt to associate words to actions (Wermter et al., 2005). In simulation, the network could perform and recognize three behaviors, docking ('pick'), wander ('go'), and move away ('lift').

4. Importance of Learning

Maybe one of the most fundamental questions we can ask about coordinate transformations is how the brain comes to know how to compute them. More precisely, to what extent is the ability to compute proper coordinate transformations already built into our brains at birth and can be seen as a product of our evolutionary heritage? And to what extent is it the product of experience dependent learning mechanisms, i.e. a product of our lifetime experience? This set of questions is a particular example of what is often called the nature/nurture debate (Elman et al., 1996).

Several pieces of evidence point to an important role of evolution in setting up proper sensorimotor coordinate mappings in our brains. Most important, maybe, is the fact that even newborn infants are robustly capable of computing certain sensorimotor transformations. For example they will readily turn their head in the direction of a salient visual stimulus, requiring them to map the location of a visually perceived object represented in retinotopic coordinates to the appropriate motor commands for turning the head. This finding is important, because it shows that some sensorimotor transformations are already in place minutes after birth, before there was much time for any experience dependent processes to learn such a mapping. The sensorimotor abilities of other species are even more striking. In some precocious species like gazelles, newborns will be able to run at high speeds a few hours after birth, successfully

avoiding obstacles and navigating across uneven terrain.
The remarkable sensorimotor abilities of some species at birth suggest that much of the solution may be “hardwired” into our brains, but experience dependent learning processes must also play an important role in setting up and maintaining proper coordinate transformations. A chief reason for this is that the required coordinate transformations are not constant but change during the organism’s development as the body grows and the geometry of sense organs and limbs matures. In fact, our ability to constantly adapt our sensorimotor coordination seems to be ubiquitous and it even persists into adulthood. A well-studied example is that of prism adaptation (Harris, 1965). When subjects wear glasses with wedge prisms that produce a sidewards shift of the visual scene, they will initially be quite inaccurate at a task like throwing darts at a target. The prism glasses introduce a bias to throw the dart to one side. Within a brief period, however, subjects are able to adapt their sensorimotor mappings to compensate for the effect of the prism glasses and their throws become accurate again. If the glasses are now taken away again, the subjects will again make errors — but this time in the opposite direction. Thus, our sensorimotor mappings can be viewed as being the product of adaptation processes that change the required coordinate transformations whenever needed.
At this point it is clear that both nature and nurture play important roles in allowing us to become so adept at performing sophisticated coordinate transformations. The fact that our coordinate transformations are so adaptable suggests that humanoid robots may also benefit from the flexibility to learn and adapt their coordinate transformations during their "life-time".

5. Neural Frame of Reference Transformations

5.1 Neural Population Code

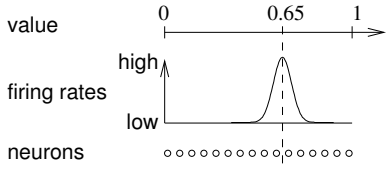

Fig. 4: Neural population code. A scalar value, e.g. 0.65, is encoded by the activation rates of an array of neurons, the *population vector*.

A continuous number is likely to be represented in the brain on an array of neurons, rather than in the firing rate of a single neuron. Fig. 4 visualizes how neurons may encode a continuous value by a neural population code. Each neuron codes for one value for which it will fire maximally, but it also responds to similar values. The set of neurons must cover all values that can be coded. The coded value can be read from the center of mass of the activation hill.
One reason for such a code in the brain may be that a cortical neuron’s firing is noisy and its firing rate is hard to adjust precisely. Many neurons for example decrease their firing rate within a few seconds of constant stimulus presentation, a phenomenon called firing rate adaptation (Blakemore & Campbell, 1969).

A second reason for population coding may be that sensory neurons are spatially distributed. A seen object activates a blob of neurons on the retina, so it is computationally straightforward to retain the positional information topographically in cortical visual areas.
Furthermore, there is more information in such a code than just the value coded in the maximally active neuron. A wider, or narrower hill of activation may account for uncertainty in the network's estimate of the encoded value. If two estimates, e.g. in the dark an uncertain visual position estimation and a more precise sound-based estimation are combined, networks that perform Bayesian cue integration can combine the information to obtain an estimate that has higher saliency than an estimate based on one cue alone (Battaglia et al., 2003).
A neural activation pattern not only encodes location information. A visual stimulus activates several hierarchically organized cortical areas, each analyzing a different aspect, such as color, shape, identity, etc. The pure location information of an object may be most beneficial in the motor cortex, but also there, information like object orientation, size and expected weight is relevant, e.g. if the object is to be grasped. A more complex shape of the neural activation pattern can represent such additional information. However, if the activation pattern becomes too complex, then information might interfere, and the coded variables may be read out incorrectly.

5.2 Principles of Frame of Reference Transformations

Coordinate transformations exist in several complexities. In the simple case, for example visual coordinates (x_1, x_2) representing a grasp target are transformed into arm angle coordinates (z_1, z_2) used for reaching the target. This can be described as a fixed (non-linear) mapping $\boldsymbol{f} : \vec{x} \mapsto \vec{z}$, and assumes that only the arm moves. The body of the agent, its head, eyes, and the object are statically fixed (Ghahramani et al., 1996). The mapping gets more involved if the arm angle coordinates have redundant degrees of freedom, as is the case in human(oid)s (Asuni et al., 2006).
In this book chapter we are focusing on the more complex, *dynamic* coordinate transformations. In the example, the visual-motor mapping would be altered by other influences, such as the gaze direction $\vec{y}$, which is determined by the posture of eyes and head. Since the mapping is now also influenced by the values of (y_1, y_2), we may express it as $\boldsymbol{f} : (\vec{x}, \vec{y}) \mapsto \vec{z}$. It is these dynamic transformations that are in the literature referred to as frame of reference transformations.
Our paramount example (for population variables $\vec{x}$ and $\vec{y}$ representing scalar variables) is that μ_x is the horizontal object position on the retina, and μ_y is the horizontal gaze angle (composed of eye- and head-angle). Then the body-centered horizontal position of the target is

$$\mu_z = \mu_x + \mu_y \,. \tag{1}$$

This is challenging, because there is no other sensory input supplying μ_z, and the computation is done with population codes.
For this scalar case, Fig. 5 shows how a neural frame of reference transformation can be performed. The scalar variables μ_x and μ_y define the centers of Gaussian hills of neural activations $\vec{x}$ and $\vec{y}$, each along one dimension. The outer product of these population codes is then represented on two dimensions, depicted as squares in Fig. 5. These two dimensions contain all information of $\vec{x}$ and $\vec{y}$, be it in a rather wasteful manner. The advantage is that for each

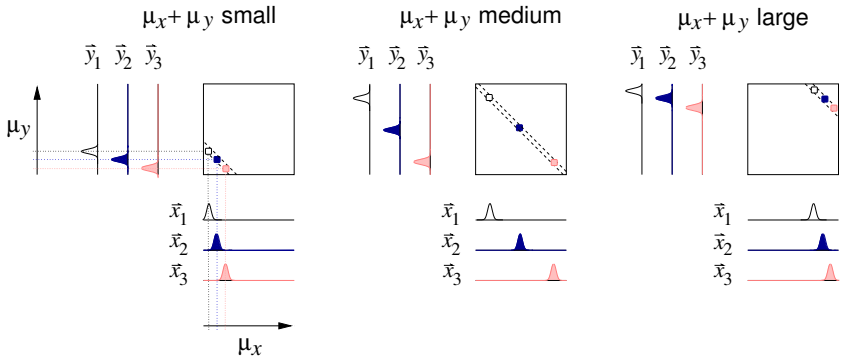

Fig. 5: Schematic of a frame of reference transformation. Three different variations of the inputs are shown for each of three different results of the transformation. i.e. resulting in small, medium and large sums. The input is a pair of Gaussian population vectors, e.g. $(\vec{x}_2, \vec{y}_2)$, representing μ_x and μ_y, the to be transformed variables. Different combinations of μ_x and μ_y are possible that lead to the same sum; these combinations lie on a diagonal on the depicted squares. A neuron that responds to a given sum will retrieve its input from one such diagonal.

constant μ_z, a diagonal line in that square represents all possible combinations of μ_x and μ_y for which Eq. 1 holds.

A straightforward method to get the result is to plaster the square with neurons which give input to a one-dimensional array of output neurons. Each output neuron receives its input from a diagonal line with an offset corresponding to the result μ_z that it represents (Rossum & Renart, 2004).

Such networks have been termed "basis function networks" (Deneve et al., 2001) referring to the units on the middle (square) layer, and partially shifting receptive fields of these units in response to a change of their input have been reported, akin to responses in several parietal and premotor areas. The term "gain field architecture" (Sauser & Billard, 2005a) refers to the multiplication of the two inputs (a component of $\vec{x}$ times a component of $\vec{y}$) which effects the middle layer responses, a behavior also observed in some cortical areas (see Section 2).

A problem in such a two-layer neural network is to find an unsupervised learning scheme. This is important, because, in the above example, μ_z is not directly available, hence there is no supervisor. We will present in Section 6 a network that replaces the middle layer by enhanced synapses of the output layer neurons, and which allows unsupervised training.

It is also worthwhile mentioning that the body geometry may considerably constrain, and therefore simplify, the possible transformations. For example, if the head turns far right, then no object on the left side of the body is visible. Hence, one input such as μ_y already constrains the result μ_z without any knowledge of the other input μ_x (assuming that only visible objects are of interest). A simple transformation network without any middle layer that was trained in a supervised fashion takes advantage of this simplification (Weber et al., 2005).

6. A Mapping with Sigma-Pi Units

A standard connectionist unit i is activated by the sum of input activations x_j weighted by its weights w_{ij}:

$$a_i = \sum_j w_{ij} x_j \tag{2}$$

This net input a_i is then usually passed through a transfer function.
A Sigma-Pi neuron i evaluates the weighted sum of multiplications

$$a_i = \sum_{j,k} w_{ijk} x_j y_k \tag{3}$$

As a specific case, the input vector $\vec{y}$ can be the same as $\vec{x}$, but in our example we have different input layers. In general, Sigma-Pi units may evaluate the product of more than just two terms.

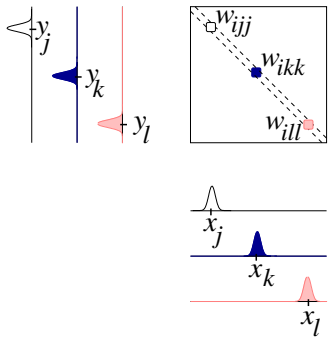

Fig. 6: A Sigma-Pi neuron with non-zero weights along the diagonal will respond only to selected input combinations, such as (x_j, y_j) or (x_k, y_k) or (x_l, y_l). This corresponds to Fig. 5, middle, where $\mu_x + \mu_y$ has medium value.

The advantage of the Sigma-Pi neuron for our problem can be seen in Fig. 6. Consider a neuron that shall be active if, and only if, $\mu_x + \mu_y$ leads to a medium sum, as in Fig. 5 middle. We can construct it by assigning non-zero weights to the according combinations in the x- and y-input layers, as depicted by the small blobs on the diagonal of the square in Fig. 6, and zero weights away from the diagonal. This neuron will be activated according to

$$a_i = w_{ijj} x_j y_j + w_{ikk} x_k y_k + w_{ill} x_l y_l \tag{4}$$

so it will be activated by the selected combinations of x- and y-inputs. It will not be activated by different combinations, such as e.g. (x_j, y_k), because w_{ijk} is zero. Such a selective response is not feasible with one connectionist neuron.

6.1 A Sigma-Pi SOM Learning Algorithm

1. Choose a random sum location $\boldsymbol{\mu_z}$.

   This sum location will be constant for the following two sample input pairs. If $\boldsymbol{\mu_z}$ is a body-centered object position, then the following two sample pairs correspond, e.g., to two different views of the object, such as two different eye positions and corresponding retinal positions of the object.

2. (a) Sample randomly input pair $(\boldsymbol{\mu_x}, \boldsymbol{\mu_y})$ with $\boldsymbol{\mu_x} + \boldsymbol{\mu_y} = \boldsymbol{\mu_z}$ and produce the corresponding population code $(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}})$

   E.g. eye position 1, and retinal object position 1.

   (b) Activate all map units $\{\boldsymbol{i}\}$: $\boldsymbol{a_i} = \sum_{\boldsymbol{j},\boldsymbol{k}} \boldsymbol{w_{ijk}} \boldsymbol{x_j} \boldsymbol{y_k}$

   Determine which map unit will be activated by co-occurrence of the given inputs, using the higher-order weights $\boldsymbol{w_{ijk}}$.

   (c) Find the winning unit: $\boldsymbol{m} = \text{argmax}_{\boldsymbol{i}} \boldsymbol{a_i}$

   This determines the center of the representation on the output layer which is constant for both views of the object.

   (d) Obtain activations: $\boldsymbol{z_i}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) = \boldsymbol{G_{xy}}(|\boldsymbol{i} - \boldsymbol{m}|, \boldsymbol{\sigma})$

   A Gaussian $\boldsymbol{G}$ around the winning unit makes the network forgive small assignment errors, and organizes the units topographically.

   (e) Decrease the neighborhood interaction width $\boldsymbol{\sigma}$

   The Gaussian is made more narrow, so after the units have learnt roughly, they can become more distinct.

3. Sample randomly input pair $(\bar{\boldsymbol{\mu}}_{\boldsymbol{x}}, \bar{\boldsymbol{\mu}}_{\boldsymbol{y}})$ with $\bar{\boldsymbol{\mu}}_{\boldsymbol{x}} + \bar{\boldsymbol{\mu}}_{\boldsymbol{y}} = \boldsymbol{\mu_z}$ and produce the corresponding population code $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$

   E.g. eye position 2, and retinal object position 2. We may also compute the activation on the output layer from this input pair, in order to check whether this leads to the same (or nearby) winning unit as in Step 2(c). However, the asymmetric learning rule needs the postsynaptic activation of only one of the two data pairs.

4. Change weights incrementally: $\boldsymbol{\Delta w_{ijk}} = \boldsymbol{\epsilon} \underbrace{\boldsymbol{z_i}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}})}_{\text{post}} \underbrace{(\bar{\boldsymbol{x}}_{\boldsymbol{j}} \bar{\boldsymbol{y}}_{\boldsymbol{k}} - \boldsymbol{w_{ijk}})}_{\text{pre}}$

   The post-synaptic learning term is the map activation resulting from the 1st data pair; only activated neurons will learn. The pre-synaptic term is the difference between the 2nd data pair and the current Sigma-Pi weight vector.

Repeat until the neighborhood interaction width $\boldsymbol{\sigma}$ is zero.

Fig. 7: One iteration of the Sigma-Pi SOM learning algorithm.

The main idea for an algorithm to learn frame of reference transformations exploits that a representation of an object remains constant over time in some coordinate system while it changes in other systems. When we move our eyes, a retinal object position will change with the positions of the eyes, while the head-centered, or body centered, position of the object remains constant. In the algorithm presented in Fig. 7 we exploit this by sampling two input pairs (e.g. retinal object position and position of the eyes, at two time instances), but we "connect" both time instances by learning with the output taken from one instance with the input taken from the other. We assume that neurons on the output (map) layer respond invariantly while the inputs are varied. This forces them to adopt, e.g. a body centered representation.

In unsupervised learning, one has to devise a scheme how to activate those neurons which do not see the data (the map neurons). Some form of competition is needed so that not all of these "hidden" neurons behave, and learn, the same. Winner-take-all is one of the simplest form of enforcing this competition without the use of a teacher. The algorithm uses this scheme (Fig. 7, step 2(c)) based on the assumption that exactly one object needs to be coded.

The corresponding winning unit to each data pair will have its weights modified so that they resemble these data more closely, as given by the difference term in the learning rule (Fig. 7, step 4). Other neurons will not see these data, as they cannot win any more, hence the competition. They will specialize on a different region in data space. The winning unit will also activate its neighbors by a Gaussian activation function placed over it (Fig. 7, step 2(d)). This causes neighbors to learn similarly, and hence organizes the units to form a topographic map.

Our Sigma-Pi SOM shares with the classical self-organizing map (SOM) (Kohonen, 2001) the concepts of winner-take-all, Gaussian activation, and a difference-based weight update. The algorithm is described in detail in Weber and Wermter (2006). Source code is available at the ModelDB data base: http://senselab.med.yale.edu/senselab/modeldb (Migliore et al., 2003).

6.2 Results of the Transformation with Sigma-Pi Units

We have run the algorithm with two-dimensional location vectors $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$, as relevant for example for a retinal object location and a gaze angle, since there are horizontal and vertical components. $\boldsymbol{\mu_z}$ then encodes a two-dimensional body-centered direction. The corresponding inputs in population code $\vec{x}$ and $\vec{y}$ are each represented by 15×15 units. Hence each of the 15 $\times 15$ units on the output layer has 15^4 = 50,625 Sigma-Pi connection parameters.

For an unsupervised learnt mapping, it cannot be determined in advance exactly which neurons of the output layer will react to a specific input. A successful frame of reference transformation, in the case of our prime example Eq. 1, is achieved, if for different combinations $(\boldsymbol{\mu_x}, \boldsymbol{\mu_y})$ that belong to a given $\boldsymbol{\mu_z}$ always the same output unit is activated, hence $\vec{z}$ will be constant. Fig. 8, left, displays this property for different $(\vec{x}, \vec{y})$ pairs. Further, different output units must be activated for a different sum $\boldsymbol{\mu'_z}$. Fig. 8, right, shows that different points on one layer, here together forming an "L"-shaped pattern, are mapped to different points on the output layer in a topographic fashion. Results are detailed in Weber and Wermter (2006).

The output $\vec{z}$ (or possibly, $\vec{a}$) is a suitable input to a reinforcement-learnt network. This is despite the fact that, before learning, $\vec{z}$ is unpredictable: the "L" shape of $\vec{a}$ in Fig. 8, right, might as well be oriented otherwise. However, after learning, the mapping is consistent. A reinforcement learner will learn to reach the goal region of the trained map (state space) based on a reward that is administered externally.

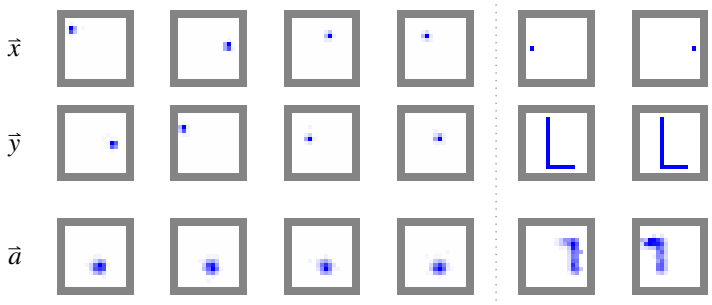

Fig. 8: Transformations of the two-dimensional Sigma-Pi network. Samples of inputs $\vec{x}$ and $\vec{y}$ given to the network are shown in the first two rows, and the corresponding network response $\vec{a}$, from which $\vec{z}$ is computed, in the third row. Leftmost four columns: random input pairs are given under the constraint that they belong to the same sum value $\vec{\mu}_z$. The network response $\vec{a}$ is almost identical in all four cases. Rightmost two columns: when a more complex "L"-shaped test activation pattern is given to one of the inputs, a similar activation pattern emerges on the sum area. It can be seen that the map polarity is rotated by 180^o.

6.3 An Approximate Cost Function

A cost function for the SOM algorithm does not strictly exist, but approximate ones can be stated, to gain an intuition of the algorithm. In analogy to Kaski (1997) we state (cf. Fig. 7):

$$E(\{W\},\{m(x,y)\}) = \frac{1}{2}\sum_{i,(x,y),j,k} G_{xy}(|i-m|,\sigma)\cdot(\bar{x}_j\bar{y}_k - w_{ijk})^2 \tag{5}$$

where the sum is over all units, data, and weight indices. The cost function is minimized by adjusting its two parameter sets in two alternating steps. The first step, winner-finding, is to minimize E w.r.t. the assignments $\{m(\vec{x},\vec{y})\}$ (cf. Fig. 7, Step 2 (c)), assuming fixed weights:

$$m(x,y) = \text{argmin}_i \sum_{j,k}(w_{ijk} - x_j y_k)^2 \;\approx\; \text{argmax}_i \sum_{j,k} w_{ijk}x_j y_k \tag{6}$$

Minimizing the difference term and maximizing the product term can be seen as equivalent if the weights and data are normalized to unit length. Since the data are Gaussian activations of uniform height, this is approximately the case in later learning stages when the weights approach a mean of the data. The second step, weight-learning (Fig. 7, Step 4), is to minimize E w.r.t. the weights $\{\vec{w}_i\}$, assuming given assignments. When convergend, $\Delta w_{ijk} = 0$ and

$$w_{ijk} = \frac{\sum_{(x,y)} z_i(\vec{x},\vec{y})\cdot\bar{x}_j\bar{y}_k}{\sum_{(x,y)} z_i(\vec{x},\vec{y})} \tag{7}$$

Hence, the weights of each unit reach the center of mass of the data assigned to it. Assignment uses $(\vec{x},\vec{y})$ while learning uses a pair $(\bar{x},\bar{y})$ of an "adjacent" time step, to create invariance. The many near-zero components of $\bar{x}$ and $\bar{y}$ keep the weights smaller than active data units.

7. Discussion

Sigma-Pi units lend themselves to the task of frame of reference transformations. Multiplicative attentional control can dynamically route information from a region of interest within the visual field to a higher area (Andersen et al., 2004). With an architecture involving Sigma-Pi weights activation patters can be dynamically routed, as we have shown in Fig. 8 b). In a model by Grimes and Rao (2005) the dynamic routing of information is combined with feature extraction. Since the number of hidden units to be activated depends on the inputs, they need an iterative procedure to obtain the hidden code. In our scenario only the position of a stereotyped activation hill is estimated. This allows us to use a simpler, SOM-like algorithm.

7.1 Are Sigma-Pi Units Biologically Realistic?

A real neuron is certainly more complex than a standard connectionist neuron which performs a weighted sum of its inputs. For example, there exists input, such as shunting inhibition (Borg-Graham et al., 1998; Mitchell & Silver, 2003), which has a multiplicative effect on the remaining input. However, such potentially multiplicative neural input often targets the cell soma or proximal dendrites (Kandel et al., 1991). Hence, multiplicative neural influence is rather about gain modulation than about individual synaptic modulation.

A Sigma-Pi unit model proposes that for each synapse from an input neuron, there is a further input from a third neuron (or even a further "receptive field" from within a third neural layer). There is a debate about potential multiplicative mutual influences between synapses, happening particularly when synapses gather in clusters at the postsynaptic dendrites (Mel, 2006). It is a challenge to implement the transformation of our Sigma-Pi network with more established neuron models, or with biologically faithful models.

A basis function network (Deneve et al., 2001) relates to the Sigma-Pi network in that each each Sigma-Pi connection is replaced by a connectionist basis function unit – the intermediate layer built from these units then has connections to connectionist output units. A problem of this architecture is that by using a middle layer, unsupervised learning is hard to implement: the middle layer units would not respond invariantly, when in our example, another view of an object is being taken. Hence, the connections to the middle layer units cannot be learnt by a slowness principle, because their responses change as much as the input activations do.

An alternative neural architecture is proposed by Poirazi et al. (2003). They found that the complex input-output function of one hippocampal pyramidal neuron can be well modelled by a two-stage hierarchy of connectionist neurons. This could pave a way toward a basis function network in which the middle layer is interpreted as part of the output neurons' dendritic trees. Being parts of one neuron would allow the middle layer units to communicate, so that certain learning rules using slowness might be feasible.

7.2 Learning Invariant Representations with Slowness

Our unsupervised learnt model of Section 6 maps two fast varying inputs (retinal object position $\vec{x}$ and gaze direction $\vec{y}$) into one representation (body-centered object position $\vec{z}$) which varies slowly in comparison to the inputs. This parallels a well known problem in the visual system: the input changes frequently via saccades while the environment remains relatively constant. In order to understand the environment, the visual system needs to transform the

"flickering" input into slowly changing neural representations – these encoding constant features of the environment.

Examples include complex cells in the lower visual system that respond invariantly to small shifts and which can be learnt with an "activity trace" that prevents fast activity changes (Földiák, 1991). With a 4-layer network reading visual input and exploiting slowness of response, Wyss et al. (2006) let a robot move around while turning a lot, and found place cells emerging on the highest level. These neurons responded when the robot was at a specific location in the room, no matter the robot's viewing direction.

How does our network relate to invariance in the visual system? The principle is very similar: in vision, certain complex combinations of pixel intensities denote an object, while each of the pixels themselves have no meaning. In our network, certain combinations of inputs $(\vec{x}, \vec{y})$ denote a $\vec{z}$, while $\vec{x}$ or $\vec{y}$ alone have no information. The set of inputs that lead to a given $\vec{z}$ is manageable, and a one-layer Sigma-Pi network can transform all possible input combinations to the appropriate output. In vision, the set of inputs that denotes one object is rather unmanageable; an object often needs to be recognized in novel view, such as a person with new clothes. Therefore, the visual system is multi-level hierarchical and uses strategies such as de-composition of objects into different parts.

Computations like our network does may be realized in parts of the visual system. Constellations of input pixel activities that are always the same can be detected by simple feature detectors made with connectionist neurons; there is no use for Sigma-Pi networks. It is different if constellations need to be detected when transformed, such as when the image is rotated. This requires the detector to be invariant over the transformation, while distinguishing from other constellations. Rotation invariant object recognition, reviewed in Bishop (1995), but also the routing of visual information (Van Essen et al., 1994), as we show in Fig. 8 b), can easily be done with second order neural networks, such as Sigma-Pi networks.

7.3 Learning Representations for Action

We have seen above how slowness can help unsupervised learning of stable sensory representations. Unsupervised learning ignores the motor aspect, i.e. the fact that the transformed sensory representations only make sense if used for motor action. Cortical representations in the motor system are likely to be influenced by motor action, and not merely by passive observation. Learning to catch a moving object is unlikely to be guided by a slowness principle.

Effects of action outcome that might guide learning are observed in the visual system. For example, neurons in V1 of rats can display reward contingent activity following presentation of a visual stimulus which predicts a reward (Shuler & Bear, 2006). In monkey V1, orientation tuning curves increased their slopes for those neurons that participated in a discrimination task, but not for other neurons that received comparable visual stimuli (Schoups et al., 2001). In the Attention-Gated Reinforcement Learning model, Roelfsema and Ooyen (2005) combine unsupervised learning with a global reinforcement signal and an "attentional" feedback signal that depends on the output units' activations. For 1-of-n choice tasks, these biologically plausible modifications render learning as powerful as supervised learning.

For frame of reference transformations that extend into the motor system, unsupervised learning algorithms may analogously be augmented by additional information obtained from movement.

8. Conclusion

The control of humanoid robots is challenging not only because vision is hard, but also because the complex body structure demands sophisticated sensory-motor control. Human and monkey data suggest that movements are coded in several coordinate frames which are centered at different sensors and limbs. Because these are variable against each other, dynamic frame of reference transformations are required, rather than fixed sensory-motor mappings, in order to retain a coherent representation of a position, or an object, in space. We have presented a solution for the unsupervised learning of such transformations for a dynamic case. Frame of reference transformations are at the interface between vision and motor control. Their understanding will advance together with an integrated view of sensation and action.

Acknowledgements

We thank Philipp Wolfrum for valuable discussions. This work has been funded partially by the EU project MirrorBot, IST-2001-35282, and NEST-043374 coordinated by SW. CW and JT are supported by the Hertie Foundation, and the EU project PLICON, MEXT-CT-2006-042484.

9. References

Andersen, C.; Essen, D. van & Olshausen, B. (2004). Directed Visual Attention and the Dynamic Control of Information Flow. In *Encyclopedia of visual attention*, L. Itti, G. Rees & J. Tsotsos (Eds.), Academic Press/Elsevier.

Asuni, G.; Teti, G.; Laschi, C.; Guglielmelli, E. & Dario, P. (2006). Extension to end-effector position and orientation control of a learning-based neurocontroller for a humanoid arm. In *Proceedings of IROS*, pp. 4151-4156.

Batista, A. (2002). Inner space: Reference frames. *Curr Biol*, *12*, 11, R380-R383.

Battaglia, P.; Jacobs, R. & Aslin, R. (2003). Bayesian integration of visual and auditory signals for spatial localization. *J Opt Soc Am A*, *20*, 7, 1391-1397.

Belpaeme, T.; Boer, B. de; Vylder, B. de & Jansen, B. (2003). The role of population dynamics in imitation. In *Proceedings of the 2nd international symposium on imitation in animals and artifacts*, pp. 171-175.

Billard, A. & Matarić, M. (2001). Learning human arm movements by imitation: Evaluation of a biologically inspired connectionist architecture. *Robotics and Autonomous Systems*, *941*, 1-16.

Bishop, C. (1995). *Neural networks for pattern recognition.* Oxford University Press.

Blakemore, C. & Campbell, F. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *J Physiol*, *203*, 237-260.

Borg-Graham, L.; Monier, C. & Fregnac, Y. (1998). Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature*, *393*, 369-373.

Buccino, G.; Vogt, S.; Ritzi, A.; Fink, G.; Zilles, K.; Freund, H.-J. & Rizzolatti, G. (2004). Neural circuits underlying imitation learning of hand actions: An event-related fMRI study. *Neuron*, *42*, 323-334.

Buneo, C.; Jarvis, M.; Batista, A. & Andersen, R. (2002). Direct visuomotor transformations for reaching. *Nature*, *416*, 632-636.

Demiris, Y. & Hayes, G. (2002). Imitation as a dual-route process featuring prediction and learning components: A biologically-plausible computational model. In *Imitation in animals and artifacts*, pp. 327-361. Cambridge, MA, USA, MIT Press.

Demiris, Y. & Johnson, M. (2003). Distributed, predictive perception of actions: A biologically inspired robotics architecture for imitation and learning. *Connection Science Journal*, *15*, 4, 231-243.

Deneve, S.; Latham, P. & Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature Neurosci*, *4*, 8, 826-831.

Dillmann, R. (2003). Teaching and learning of robot tasks via observation of human performance. In *Proceedings of the IROS workshop on programming by demonstration*, pp. 4-9.

Duhamel, J.; Bremmer, F.; Benhamed, S. & Graf, W. (1997). Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, *389*, 845-848.

Duhamel, J.; Colby, C. & Goldberg, M. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, *255*, 5040, 90-92.

Elman, J. L.; Bates, E.; Johnson, M.; Karmiloff-Smith, A.; Parisi, D. & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MIT Press.

Fadiga, L. & Craighero, L. (2003). New insights on sensorimotor integration: From hand action to speech perception. *Brain and Cognition*, *53*, 514-524.

Fogassi, L.; Raos, V.; Franchi, G.; Gallese, V.; Luppino, G. & Matelli, M. (1999). Visual responses in the dorsal premotor area F2 of the macaque monkey. *Exp Brain Res*, *128*, 1-2, 194-199.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neur Comp*, *3*, 194-200.

Gallese, V. (2005). The intentional attunement hypothesis. The mirror neuron system and its role in interpersonal relations. In *Biomimetic multimodal learning in a robotic systems*, pp. 19-30. Heidelberg, Germany, Springer-Verlag.

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

Ghahramani, Z.; Wolpert, D. & Jordan, M. (1996). Generalization to local remappings of the visuomotor coordinate transformation. *J Neurosci*, *16*, 21, 7085-7096.

Grafton, S.; Fadiga, L.; Arbib, M. & Rizzolatti, G. (1997). Premotor cortex activation during observation and naming of familiar tools. *Neuroimage*, *6*, 231-236.

Graziano, M. (2006). The organization of behavioral repertoire in motor cortex. *Annual Review Neuroscience*, *29*, 105-134.

Grimes, D. & Rao, R. (2005). Bilinear sparse coding for invariant vision. *Neur Comp*, *17*, 47-73.

Gu, X. & Ballard, D. (2006). Motor synergies for coordinated movements in humanoids. In *Proceedings of IROS*, pp. 3462-3467.

Harada, K.; Hauser, K.; Bretl, T. & Latombe, J. (2006). Natural motion generation for humanoid robots. In *Proceedings of IROS*, pp. 833-839.

Harris, C. (1965). Perceptual adaptation to inverted, reversed, and displaced vision. *Psychol Rev*, *72*, 6, 419-444.

Hoernstein, J.; Lopes, M. & Santos-Victor, J. (2006). Sound localisation for humanoid robots - building audio-motor maps based on the HRTF. In *Proceedings of IROS*, pp. 1170-1176.

Kandel, E.; Schwartz, J. & Jessell, T. (1991). *Principles of neural science*. Prentice-Hall.

Kaski, S. (1997). *Data exploration using self-organizing maps*. Doctoral dissertation, Helsinki University of Technology. Published by the Finnish Academy of Technology.

Kohler, E.; Keysers, C.; Umilta, M.; Fogassi, L.; Gallese, V. & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, *297*, 846-848.

Kohonen, T. (2001). *Self-organizing maps* (3. ed., Vol. 30). Springer, Berlin, Heidelberg, New York.

Lahav, A.; Saltzman, E. & Schlaug, G. (2007). Action representation of sound: Audiomotor recognition network while listening to newly acquired actions. *J Neurosci*, *27*, 2, 308-314.

Luppino, G. & Rizzolatti, G. (2000). The organization of the frontal motor cortex. *News Physiol Sci*, *15*, 219-224.

Martínez-Marín, T. & Duckett, T. (2005). Fast reinforcement learning for vision-guided mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2005).*

Matsumoto, R.; Nair, D.; LaPresto, E.; Bingaman, W.; Shibasaki, H. & Lüders, H. (2006). Functional connectivity in human cortical motor system: a cortico-cortical evoked potential study. *Brain*, *130*, 1, 181-197.

Mel, B. (2006). Biomimetic neural learning for intelligent robots. In *Dendrites*, G. Stuart, N. Spruston, M. Hausser & G. Stuart (Eds.), (in press). Springer.

Meng, Q. & Lee, M. (2006). Biologically inspired automatic construction of cross-modal mapping in robotic eye/hand systems. In *Proceedings of IROS*, pp. 4742-4747.

Migliore, M.; Morse, T.; Davison, A.; Marenco, L.; Shepherd, G. & Hines, M. (2003). ModelDB Making models publicly accessible to support computational neuroscience. *Neuroinformatics*, *1*, 135-139.

Mitchell, S. & Silver, R. (2003). Shunting inhibition modulates neuronal gain during synaptic excitation. *Neuron*, *38*, 433-445.

Oztop, E.; Kawato, M. & Arbib, M. (2006). Mirror neurons and imitation: A computationally guided review. *Neural Networks*, *19*, 254-271.

Poirazi, P.; Brannon, T. & Mel, B. (2003). Pyramidal neuron as two-layer neural network. *Neuron*, *37*, 989-999.

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

Rizzolatti, G.; Fogassi, L. & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Review*, *2*, 661-670.

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

Rizzolatti, G. & Luppino, G. (2001). The cortical motor system. *Neuron*, *31*, 889-901.

Roelfsema, P. & Ooyen, A. van. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neur Comp*, *17*, 2176-2214.

Rossum, A. van & Renart, A. (2004). Computation with populations codes in layered networks of integrate-and-fire neurons. *Neurocomputing*, *58-60*, 265-270.

Sauser, E. & Billard, A. (2005a). Three dimensional frames of references transformations using recurrent populations of neurons. *Neurocomputing*, *64*, 5-24.

Sauser, E. & Billard, A. (2005b). View sensitive cells as a neural basis for the representation of others in a self-centered frame of reference. In *Proceedings of the third international symposium on imitation in animals and artifacts, Hatfield, UK.*

Schaal, S.; Ijspeert, A. & Billard, A. (2003). Computational approaches to motor learning by imitation. *Transactions of the Royal Society of London: Series B*, *358*, 537-547.

Schoups, A.; Vogels, R.; Qian, N. & Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, *412*, 549-553.

Shuler, M. & Bear, M. (2006). Reward timing in the primary visual cortex. *Science*, *311*, 1606-1609.

Takahashi, Y.; Kawamata, T. & Asada, M. (2006). Learning utility for behavior acquisition and intention inference of other agents. In *Proceedings of the IEEE/RSJ IROS workshop on multi-objective robotics*, pp. 25-31.

Tani, J.; Ito, M. & Sugita, Y. (2004). Self-organization of distributedly represented multiple behavior schemata in a mirror system: Reviews of robot experiments using RNNPB. *Neural Networks*, *17*, 8-9, 1273-1289.

Triesch, J.; Jasso, H. & Deáak, G. (2006). Emergence of mirror neurons in a model of gaze following. In *Proceedings of the Int. Conf. on Development and Learning (ICDL 2006).*

Triesch, J.; Teuscher, C.; Deák, G. & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental Science*, *9*, 2, 125-147.

Triesch, J.; Wieghardt, J.; Mael, E. & Malsburg, C. von der. (1999). Towards imitation learning of grasping movements by an autonomous robot. In *Proceedings of the third gesture workshop (gw'97).* Springer, Lecture Notes in Artificial Intelligence.

Tsay, T. & Lai, C. (2006). Design and control of a humanoid robot. In *Proceedings of IROS*, pp. 2002-2007.

Umilta, M.; Kohler, E.; Gallese, V.; Fogassi, L.; Fadiga, L.; Keysers, C. et al. (2001). I know what you are doing: A neurophysiological study. *Neuron*, *31*, 155-165.

Van Essen, D.; Anderson, C. & Felleman, D. (1992). Information processing in the primate visual system: an integrated systems perspective. *Science*, *255*, 5043, 419-423.

Van Essen, D.; Anderson, C. & Olshausen, B. (1994). Dynamic Routing Strategies in Sensory, Motor, and Cognitive Processing. In *Large scale neuronal theories of the brain*, pp. 271-299. MIT Press.

Weber, C.; Karantzis, K. & Wermter, S. (2005). Grasping with flexible viewing-direction with a learned coordinate transformation network. In *Proceedings of Humanoids*, pp. 253-258.

Weber, C. & Wermter, S. (2007). A self-organizing map of Sigma-Pi units. *Neurocomputing*, (in press).

Weber, C.; Wermter, S. & Elshaw, M. (2006). A hybrid generative and predictive model of the motor cortex. *Neural Networks*, *19*, 4, 339-353.

Weber, C.; Wermter, S. & Zochios, A. (2004). Robot docking with neural vision and reinforcement. *Knowledge-Based Systems*, *17*, 2-4, 165-172.

Wermter, S.; Weber, C.; Elshaw, M.; Gallese, V. & Pulvermüller, F. (2005). A Mirror Neuron Inspired Hierarchical Network for Action Selection. In *Biomimetic neural learning for intelligent robots*, S. Wermter, G. Palm & M. Elshaw (Eds.), pp. 162-181. Springer.

Wyss, R.; König, P. & Verschure, P. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biology*, *4*, 5, e120.