4. Computational Mechanisms of the Visual Processing of Action Stimuli

Falk Fleischer & Martin A. Giese

Section for Computational Sensomotorics, Dept. of Cognitive Neurology, Hertie Institute for Clinical Brain Research & Center for Integrative Neuroscience, University Clinic Tübingen, Frondsbergstr. 23, 72070 Tübingen, Germany;
Email: falk.fleischer@medizin.uni-tuebingen.de, martin.giese@uni-tuebingen.de

Abstract:

The recognition of body motion is a central function of the visual system that has stimulated substantial interest in neuroscience. At the same time, the recognition of body shapes, movements and actions from videos represents a complex computational problem, whose difficulty is sometimes bypassed by popular explanations of motion recognition in neuroscience. Only a serious interaction between neuroscience and computational theory will help to

identify the important computational steps of action recognition in the brain, and might contribute a clarification of their neural implementation. Computational models can specifically help to test the computational feasibility of possible explanations of the processing of body movements, and they help to derive theoretically well-defined predictions that can be tested experimentally. Such theoretical work helps to derive critical constraints for explanations of the processing of body motion, since some intuitive theories might be computationally not feasible or not robust enough to deal with real-world stimuli. This chapter reviews a class of neural theories for the recognition of body motion, which was originally developed in order to account for the processing of biological motion stimuli and the recognition of *non-transitive* body movements, that is non goal-directed movements such as walking. We show how these theories can be extended to models for the processing of goal-directed *transitive* actions, that is actions with a goal object such as grasping. We show that such an extension is possible by addition of a few simple physiologically plausible neural mechanisms. The resulting model accounts for the view-independent recognition of hand actions from real videos with an accuracy that is sufficient even for the detection of subtle differences between grips. In addition, the resulting model reproduces a number of key properties of the visual tuning of action-selective neurons in visual, parietal and premotor cortex. The relationship between this new model and other computational approaches for the visual processing of goal-directed actions is discussed.

4.1 INTRODUCTION

The recognition of biological motion and actions is a core function of the visual system with crucial importance for survival and social communication. Motion recognition addresses the processing of body movements of humans and other species. One class of such movements, also called ‘non-transitive actions’, is not primarily directed towards goal objects. Examples are locomotion, such as walking and running, or many communicative gestures like waving. Another important class of movements is goal-directed actions, also called ‘transitive movements’. These actions are directed towards specific goal objects. Examples are grasping, holding, pushing or pointing towards objects. In neuroscience, these two subfunctions of motion recognition have been investigated largely independently by different research communities. One group of researchers, rather coming from vision research, has focused on the visual processing of biological motion stimuli and other body movements, often focusing on non-transitive actions. Another community, stressing specifically potential links between representations for action perception and execution, have often focused on goal-directed actions and stressed the dependency of action goals. Theoretical approaches accounting for the processing of these two types of body motion stimuli have remained largely unrelated, and it is not really clear which processes might be shared between the processing of non-transitive and transitive actions. Many details about the neural mechanisms of motion and action recognition are

reviewed in Chapters 2, 3, and 17, so that we focus here on the aspects that are central for the modeling.

The systematic study of the visual recognition of *non-transitive actions* (without goal objects) has strongly been influenced by the classical work of Johansson (1973). His famous experiments have shown that body movements can be recognized from strongly impoverished stimuli, such as point-light walkers. Subsequent studies have demonstrated that perception from point-light stimuli is amazingly robust, e.g. against displacements of the dots along the skeleton of the moving figure (e.g., Dittrich, 1993; Beintema & Lappe, 2002), or against masking with substantial numbers of moving noise dots (Cutting, Moore & Morrison, 1988; Bertenthal & Pinto, 1994; Thornton, Pinto & Shiffrar, 1998). In addition, point-light stimuli can convey subtle details about motions style, conveying information about gender, identity or the emotion of walkers (e.g. Cutting & Kozlowski, 1977; Beardsworth & Buckner, 1981; Dittrich et al., 1996; Pollick et al., 2002; Chouchourelou et al., 2006). A detailed discussion on the recognition of bodily expressions is given in Chapter 3.

A variety of models have been developed for the recognition of biological motion and actions without goal objects. Early approaches have tried to exploit geometrical invariants of body motion, such as the fact that for the side view the distance between dots on the same limb remains approximately constant over the course of the motion (e.g. Hoffman & Flinchbaugh, 1982; Webb & Aggarwal, 1982). Another set of approaches has

tried to account for body motion recognition by the fitting of three-dimensional models of body shapes (e.g. Marr & Vaina, 1982). This approach has later been extensively extended in computer vision, combining it with probabilistic predictive models (e.g. Blake & Isard, 1999; and many others), but typically without claiming that the developed mechanisms are relevant for the brain.

A third class of computational approaches is coarsely inspired by cortical mechanisms and tries to account for motion recognition by the extraction of form and motion features, or spatio-temporal image features from video sequences. Early methods have compared such features with templates that were either constructed by hand, or which had been learned using sequence recognition methods such as Hidden Markov Models (HMMs) (e.g. Niyogi & Adelson, 1993; Bobick, 1997). In general, the most robust state-of-the-art body-motion recognition approaches in computer vision are based on the extraction of spatio-temporal image features combined with appropriate methods for the learning of optimized feature dictionaries and powerful classification methods (e.g. Efros et al. 2003; Laptev & Lindberg, 2003; Dollar et al. 2005; Gorelick et al 2007). (Much more detailed reviews about computer vision methods for action recognition are given, for example, in Gavrila, 1999; Moeslund & Granum, 2001; a more detailed review stressing the relationship between such methods and biological models see Giese, 2004).

Based on the idea of an extraction of motion and form features a

number of biologically-inspired models for the recognition of non-transitive actions have been developed that reproduce basic features from experiments in electrophysiology, psychophysics and functional imaging (Giese & Poggio, 2003; Casile & Giese 2005; Lange & Lappe, 2006). These models are based on hierarchies of neural detectors for form and / or motion features that mimic properties of cortical neurons that are involved in the recognition of motion patterns. In addition, these models assume mechanisms for the learning of motion and shape templates, which potentially determine the tuning of neurons in higher visual areas that are selective for moving bodies, such as the superior temporal sulcus (STS).

While these hierarchical neural models had originally been developed to model biological data without any technical relevance, new interest in such architectures has recently emerged in computer vision, where it has been shown that such architectures can achieve performance levels in motion classification that are competitive with non-biological state-of-the-art approaches in computer vision (Serre et al. 2007; Jhuang et al. 2007; Escobar et al. 2008; Schindler & van Gool, 2008).

Research on the perception of *transitive* actions, that is actions that are directly directed towards a goal object, has recently become a very popular topic in neuroscience. A vast number of studies have investigated the perception of goal-directed actions, such as grasping, predominantly using behavioral and imaging methods (reviews see for example Rizzolatti et al. 2001; Rizzolatti & Craighero, 2004; Ferrari et al. 2009). This research interest

has been stimulated by the discovery of so called *mirror neurons* in monkey premotor and parietal cortex. These neurons show selective tuning during the visual observation of actions as well as during action execution (di Pellegrino et al. 1992; Gallese et al. 1996; Rizzolatti et al. 2001; Fogassi et al. 2005), specifically if such actions are directed towards a goal object. Imaging studies in humans suggest the existence of a mirror neuron system also in the human cortex (e.g. Iacoboni et al. 2005; Kilner et al. 2009). However, the significance of such observations in human fMRI experiments is still under dispute (e.g., Dinstein et al. 2007). At the single-cell level, neurons with visual selectivity for goal-directed actions and for action-related properties of the goal-object have been found in the *superior temporal sulcus* (STS) (Perrett et al. 1989; Nelissen et al. 2006), and in the parietal cortex, e.g. in the *anterior intraparietal sulcus*, AIP) (Sakata et al. 1997; Murata et al. 2000; Gardner et al. 2007; Baumann et al. 2009). A more extensive review of the mirror neuron system in monkeys and humans is given in Chapter 2.

The observation of joint motor and visual tuning for actions in the same neurons, or the same areas in imaging studies, has been interpreted as evidence for recognition by *resonance.* This signifies the hypothesis, that actions are recognized by an internal simulation of the visually observed actions in motor representations, which are also involved in the control of the execution of the same action (Rizzolatti et al. 2001). In fact, behavioral research has provided an overwhelming amount of evidence (reviews see e.g.

Wilson & Knoblich, 2005; Blakemore & Frith, 2005; Schütz-Bosbach & Prinz, 2007) that neural representations of action execution and action vision are functionally tightly coupled, consistent with the theory of a *common coding* (Prinz, 1997) of actions in perception and execution (an extensive review of the principles of common coding is provided in Chapter 20). However, it remains an unresolved question how exactly the mirror neuron system contributes to the coupling of action execution and perception. A simple mechanism for recognition by resonance would, for example, predict that the tuning of mirror neurons for executed and perceived actions should be highly similar. However, such similarity is far from clearly present in many mirror neurons in area F5 (Gallese et al. 1996; Caggiano et al., personal commun.).

Data from behavioral and fMRI studies in humans on action recognition have also been used as support for much further-reaching speculations. It has been suggested that the mirror neuron system might play a fundamental role in the imitation learning of actions and also for the development of language (see Arbib, 2008, for a review). In addition, the involvement of the mirror neuron system in a variety of other higher cognitive functions has been claimed, such as action understanding, theory of mind, empathy, and even the perception of esthetic qualities (e.g. Gallese & Goldman, 1998; Rizzolatti & Fabbri-Destro, 2008; Frith & Singer, 2008; Gallese & Freedberg, 2007). We think that such extensions of the theory of 'recognition by resonance' provide food for many interesting discussions,

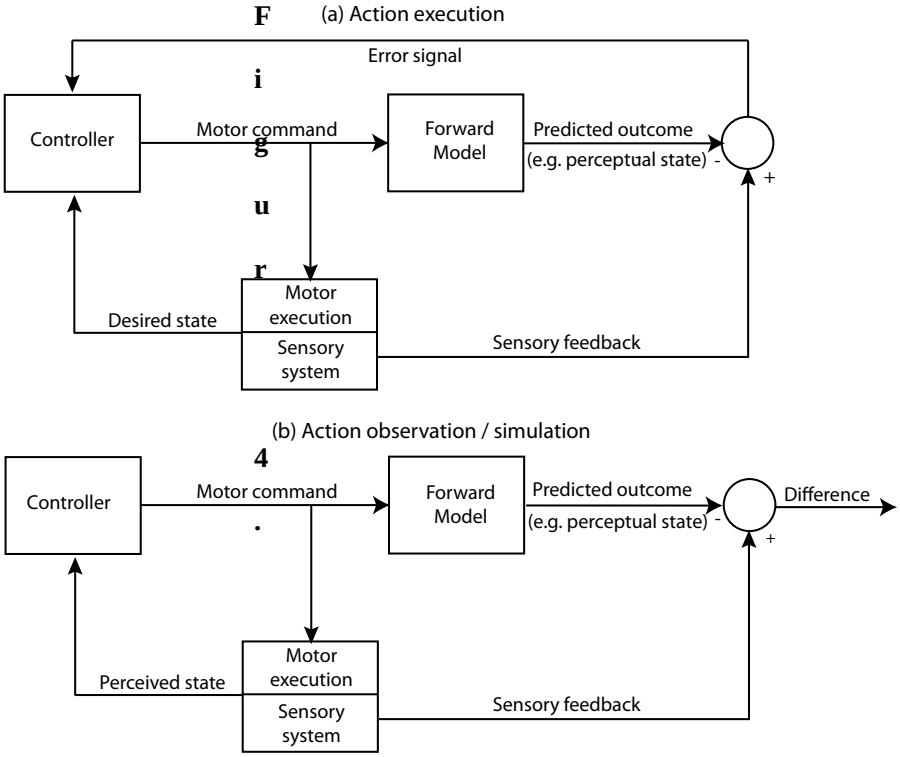
potentially even including important philosophical implications. However, we do not treat such aspects in the remainder of this chapter since many of the underlying concepts cannot be easily formulated with sufficient accuracy and strict quantitative links to empirical data, so that a treatment in the context of mathematical modeling is very difficult.

Corresponding to the strong interest in the relationship between action perception and action execution in the context of transitive actions, most biologically motivated computational models in this area have focused on the potential role of motor representations for action recognition. One of the first neural network models with relevance for the biological system (Oztop & Arbib, 2002; Fagg & Arbib, 1998) demonstrated that the recognition of goal-directed actions from video stimuli is possible by comparison of internally simulated sequences of the hand and arm configurations with the visual stimulus. This model has later been extended, linking it to the MOSAIC model for the selection of motor controllers (Haruno et al. 2001; Oztop et al 2006) and by inclusion of mechanisms that account for audio-visual properties of mirror neurons (Bonaiuto et al. 2007). In general, cortical mechanisms of motor control are assumed to rely on forward models that compute predictions of the sensory signals dependent on the controller output. This prediction can then be compared with the actual sensory feedback. By comparing the predictions of different controller models, which realize different possible actions, with the actual sensory input it is possible to select the most appropriate controller which results in the best prediction of the

actual sensory input. Once this controller has been found the underlying action has been recognized. (See for example Wolpert & Ghahramani, 2000; Wolpert et al., 2003, for further details.)

The underlying principles are illustrated in Figure 4.1. During action execution (Panel a) the motor controller generates a motor command that is mapped through a predictive forward model into predicted sensory signals. These signals are compared to the true sensory feedback, and the prediction error is used to update the controller. In the context of action observation (Panel b) the controller runs without this error input and generates predictions based on the actual sensory inputs. After optimization of the controller the sensory input activates the motor command which would be compatible with the actual observed action.

Figure4.1



Biologically-inspired dynamical controller architectures for action recognition that are based on these ideas have been successfully applied in the context of robot systems (e.g. Schaal et al. 2003; Demiris and Simmons 2006; Sauser and Billard 2006). Several models in this context have bypassed largely the aspect of visual processing by making strongly simplifying assumptions about the processing up to parietal areas and STS. These models assume, for example, that the three dimensional structure of effector and goal object and their metric relationship is given by the visual system in terms of low-dimensional variables, such as joint angles, which then serve as input for the dynamic controller architecture (e.g. Wolpert et al. 2003; Tani et al. 2004; Erlhagen et al. 2006). However, these approaches do not answer the question how a sufficiently accurate estimation of such low-dimensional parameters is possible. Such estimation is a difficult computational problem especially from monocular image sequences, which do not provide depth information through disparity cues. Yet, humans and animals are very good in action recognition from stimuli without such depth cues.

Summarizing, only few of the existing computational approaches for the recognition of goal-directed actions work on real video stimuli at all (Billard and Mataric, 2001; Oztop and Arbib, 2002; Demiris and Simmons, 2006; Metta et al. 2006; Kjellström et al. 2008). None of these models exploits mechanisms that approximate the functions of neurons in visual cortex, except for a single model for the estimation of grip aperture from

video sequences (Prevete et al. 2008).

The previous overview of the existing work raises two questions: 1) Which physiologically plausible mechanisms are computationally powerful enough to accomplish the visual recognition of transitive, goal-directed body movements from real video stimuli? Only architectures based on such mechanisms seems ultimately suitable as basis for the development of more detailed neural models of visual action recognition. 2) How are the principles for the recognition of transitive and non-transitive actions related? Can a part of the architecture for the recognition of non-transitive actions be exploited as well for the recognition of transitive actions? Which additional computational and neural steps are required for the recognition of transitive actions with goal objects?

We will try to provide answers for these questions in this chapter in the following steps: Section 4.2 presents a short overview of a, meanwhile established, basic architecture for the recognition of non-transitive actions that accounts for a variety of experimental data, and which has recently also been applied in biologically-inspired computer vision reaching competitive performance levels. In Section 4.3 we introduce an extension of this model architecture that makes it suitable for the recognition of transitive actions. The required additional computational mechanisms are explained in detail. In section 4.4 we show a few simulation results, illustrating that the proposed models is not only powerful enough to recognize actions from real videos, but also reproduces key properties of action-selective cortical neurons.

Furthermore, we show some predictions that can be validated easily at the level of single neurons. Section 4.5 discusses several limitations of the proposed model and presents concluding remarks.

4.2 BASIC MODEL FOR THE VISUAL RECOGNITION OF NON-TRANSITIVE ACTIONS

In the following, we present a basic model architecture for the recogniton of body movements that are not directed towards specific goal objects, like walking or waving. The presented model has been compared in detail with a variety of experimental results (Giese & Poggio, 2003; Casile & Giese, 2005), and it has motivated several new experiments that have partially confirmed predictions made by the model (e.g. Peuskens at al. 2005; Thurman & Grossman, 2008; Vangeneugden et al. 2008; Jastorff et al. 2009; Singer & Sheinberg, 2010). Furthermore, the basic architecture of the model has given rise to computationally more efficient implementations that have demonstrated the feasibility of the proposed approach even for technical motion recognition systems, reaching state-of-the-art performance levels (e.g. Jhuang et al. 2007).

The proposed motion recognition model is based on a number of principles that are well-established for the visual cortex. Some of them are shared with the processing of static shapes (e.g Riesenhuber & Poggio, 1999). An overview of the model architecture is shown in Figure 4.2.

4.2.1 Two hierarchical neural pathways

Consistent with the basic anatomical architecture of the visual cortex, the model is partitioned into two hierarchical neural pathways that model the ventral and the dorsal processing stream (Ungerleider & Mishkin, 1982; Felleman & van Essen, 1991; Goodale & Milner, 1992). The first pathway is specialized for the processing of form information, while the second pathway processes local motion information. Consistent with electrophysiological data (Saleem et al., 2000), these two pathways converge at a level that corresponds to the superior temporal sulcus (STS). While in real visual cortex these two processing streams likely are connected at multiple levels, the model makes the over-simplifying assumption that they remain separate until the level of the STS, where they are integrated.

Both pathways consist of hierarchies of neural detectors. Consistent with real neurons in the visual pathway, the complexity of the extracted features and the sizes of the receptive fields of the neural detectors increase along the hierarchy. The sizes of the receptive fields at different hierarchy levels coarsely match the receptive field sizes of corresponding neurons in the visual pathway. Invariance against translation and scaling of features is accomplished by nonlinear pooling along the hierarchy using a maximum operation (Fukushima, 1980; Riesenhuber & Poggio, 1999).

The neural detectors in the form pathway process form features with

different levels of complexity. The first hierarchy level is formed by *local orientation detectors,* mimicking the properties of simple cells in area V1 (Hubel & Wiesel, 1962). The response properties of these cells were modeled by Gabor filters. These are local filters that respond maximally to grating stimuli of particular orientation and spatial frequency. The detectors that form the next hierarchy level mimic the behavior of complex cells, e.g. in areas V2 or V4. Their responses were determined by computing the maximum of the responses of groups of Gabor filters with the same orientation preference, but slightly different receptive field centers and different spatial scales. It has been shown that such 'maximum pooling' results in (partial) position and scale invariance and increases the robustness of the responses of the orientation detectors against background clutter (Riesenhuber and Poggio, 1999). Related models for the recognition of static shapes have added additional layers at this level, forming successively more complex form features by combination of the features from previous layers (Riesenhuber & Poggio, 1999; Serre et al. 2007). The resulting detectors extract form features of intermediate complexity and result in tuning properties that match quite accurately the ones of neurons in area V4 (Cadieu et al. 2007). However, for the recognition of body motion from simple stimuli the inclusion of more complex intermediate level form features was not necessary (Giese & Poggio, 2003).

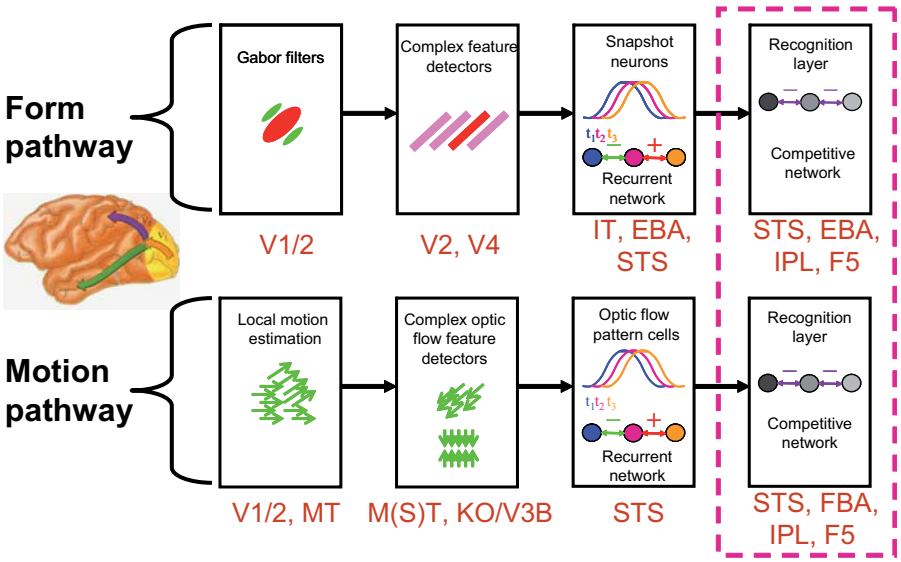


Figure 4.2

The next-higher level of the form recognition hierarchy consists of detectors that are selective for complex shapes, corresponding to body configurations that are characteristic for "snapshots" from movement sequences. It was assumed that these shape-selective neurons are similar to view-tuned neurons as described in area IT of monkeys (Logothetis & Sheinberg, 1996). These detectors were modeled by Gaussian radial basis functions (RBFs). These are model neurons with a Gaussian tuning function, typically within a multi-dimensional feature space. Their maximum response arises for a template feature vector (called 'center' of the RBF), and the response decays monotonically with the distance between the input vector and this template vector. Physiologically plausible circuits for the implementation

of RBFs have been proposed in Khou & Poggio (2008). The templates (RBF centers) were determined by learning from training image sequences, setting them to the output vectors from the previous hierarchy level that resulted for keyframes of the motion pattern, which were obtained by sampling with equidistant time steps. Neurons with properties similar to such snapshot neurons have been observed in the STS in monkey cortex (e.g. Perrett et al. 1985; Oram & Perrett, 1996; Perrett et al. 2009), and in fact, recently strong electrophysiological evidence has been provided for the existence of such neurons (Vangeneugden et al. 2009; Singer & Sheinberg, 2010).

The motion pathway has, in principle, the same architecture as the form pathway. In this case, the extracted features depend on the local motion in the stimulus. The first hierarchy layer of this pathway contains *local motion energy detectors* that are selective for different local motion directions and speeds. These detectors model motion-selective neurons in primary visual cortex and the mediotemporal area (MT) (Smith & Snowden, 1994). Extensions of the original model have realized this level using different types of motion detectors that are suitable for the processing of real video sequences (Jhuang et al. 2007; Escobar et al 2008). The second hierarchy level of the motion pathway consists of detectors for more complex local optic flow patterns, that is motion patterns that integrate different directions and speeds with specific local spatial configurations. In the original models these pattern were pre-defined (translation, opponent motion in different directions). In more recent implementations ‘dictionaries’ of optimized intermediate-level

optic flow features have been learned from example videos (Sigala et al. 2005; Jhuang et al. 2007). The resulting detectors correspond to motion-selective neurons, potentially in areas MT and MSTl, that are selective for complex intermediate optic flow features (e.g. Allman et al. 1985; Xiao et al. 1995; Eifuku & Wurtz, 1998; Born, 2000).

The next-higher level of the motion pathway consists of neural detectors for complex optic flow patterns that arise temporarily during action stimuli. These detectors are equivalent to the snapshot neurons in the form pathway. They are modeled by radial basis functions whose centers were determined by the feature vectors from the previous level, derived from training videos, just like the RBFs of the snapshot neurons in the form pathway. Neurons with selectivity for such highly complex motion patterns have been found in the STS (e.g. Perrett et al. 1985; Vangeneugden et al. 2008).

It is important to notice that the described optic flow pattern detectors are selective for patterns of local motion information with complex spatial organization. This spatial organization is holistic and covers the whole action stimulus. (Such holistic recognition mechanisms are sometimes also termed ‘configural’ in the psychophysical literature). It is important to notice here that holistic recognition can be accomplished as well based on form features (as for the snapshot neurons), as based on local motion features, since *local* motion also carries spatial information. The argumentation that holistic or configural recognition mechanisms automatically imply form-based

processing is thus a conceptual confound, which sometimes can be found in the related literature.

Recent work on similar models for object and motion recognition in computer vision has shown that for accomplishing good performance on real images and video sequences it is crucial to optimize the tuning properties of the neural detectors at the intermediate levels of the hierarchy by learning from image data (Serre et al. 2007; Jhuang et al. 2007). This approach is also taken in section 4.3.1.

4.2.2 Selectivity for temporal order

The recognition of action patterns is critically dependent on temporal order, i.e. the temporal sequence with which body shapes arise during the stimulus. To account for this effect, the model assumes that the snapshot and optic flow pattern neurons are embedded in dynamic recurrent neural networks (i.e. nonlinear neural networks with lateral connections) that make their responses dependent on the sequential temporal order. The details of this network are described in Section 4.3.2. As consequence of the lateral connections between them, the individual snapshot neurons fire strongly only if the corresponding body shapes occur in the right temporal context. Showing the same stimulus sequence in reverse or scrambled temporal order results in a substantial decay of their neural responses (Giese & Poggio, 2003). Likewise, due to this network dynamics the optic flow pattern neurons in the motion pathway

respond strongly only if the relevant optic flow patterns arise in the right temporal sequence. The form of these lateral connections can be easily established by Hebbian learning (Jastorff & Giese, 2004), a physiologically plausible learning rule that makes synaptic changes dependent on the correlations between pre- and postsynaptic signals and their relative timing.

The highest hierarchy level of the form and the motion pathway is given by *motion pattern detectors* that temporally smooth and summate the activity of all snapshot neurons and the optic flow pattern neurons. These neural detectors respond during the presence of a particular action (like "walking", "marching", "boxing", etc.). Their response is strongly sequence-selective so that they do not become activated, e.g. by actions shown in reverse temporal order (Giese & Poggio, 2003). Neurons with similar properties have been found in the STS (Perrett et al. 1985; Oram & Perrett, 1996; Jellema & Perrett, 2006; Vangeneugden et al. 2008).

4.2.3 Integration of form and motion

In the cortex the two visual pathways converge at the level of the STS (Saleem et al., 2000). This convergence of form and motion has been a central feature of the model by Giese & Poggio (2003) and can be simply modeled by summing the responses of the motion pattern neurons in the form and the motion pathway. Alternatively, one also could assume common motion pattern neurons for both pathways. The detailed mechanisms of the fusion of

form and motion in body motion recognition in the visual cortex remain unknown and have to be decided based on appropriate data from single cells. However, experimental evidence suggests that both pathways interact already at earlier levels than the STS and that there are top-down influences from motion pattern recognition into earlier levels of the motion and form pathway (e.g. Peuskens et al. 2005). Such top-down influences are not captured by the existing neural models for motion pattern recognition. A detailed discussion of top-down processes is given in Chapter 18. In addition, Chapter 17 gives a detailed discussion about the integration of form and motion information in biological motion processing based on human imaging data.

Deviating from the idea of a fusion of form and motion cues, e.g. at the level of the STS, as proposed in our model, in the field of biological motion processing a vivid discussion has emerged that tried to address the question whether the recognition of point-light walkers is based *exclusively* on form or *exclusively* on motion information. The starting point of this discussion was a novel point-light stimulus (Beintema & Lappe, 2002) that presented the dots at randomized positions on the skeleton in every frame, and which elicited the percept of biological motion in human observers, while it reduces the amount of available local motion information. This result has motivated the hypothesis that biological motion recognition might be based *exclusively* on form templates, and basically independent of local motion except for 'segmentation' of the figure from the background. As further support for this hypothesis Lange and Lappe (2006) have proposed a neural

form template fitting model, which is very similar to the form pathway of the model discussed in Section 4.2.1, but lacks mechanisms for scale and position invariance. In detailed simulations they showed that, assuming an appropriate pre-positioning of the template over the stimulus, the model resulted in good curve fits of several psychophysical results. However, more detailed simulations with our model including both, a motion and a form pathway suggested that spontaneous generalization between normal and point-light stimuli might be much easier to obtain in the motion pathway. In addition, these simulations showed that the stimulus by Beintema & Lappe (2002) contains substantial amounts of local motion information, and is thus also suitable for recognition by a purely motion-based model (Casile & Giese, 2005)..

The question if form or motion features contribute to the perception of body motion and actions seems interesting and has stimulated a large number of novel studies (e.g. Casile & Giese, 2005; Beintema et al. 2006; Thurman & Grossman, 2008; Thirkettle et al. 2009). From our point of view, however, extreme positions like ‘only form’ or ‘only motion’ being relevant for the processing of body motion are not particularly helpful for developing a deeper understanding of brain function. From the point of computational efficiency it seems extremely unlikely that the brain discards robust features, like local motion, from the recognition process. Also some of the arguments in favor of an exclusively form-based recognition process seem problematic: 1) The stimulus by Beintema & Lappe (2002) contains substantial local

motion (Casile & Giese, 2005) and thus does not provide an example of a local motion-free biological motion recognition. 2) The idea of the fitting of form templates (Lange & Lappe, 2006) bypasses completely how such templates are positioned on the stimulus. This problem is specifically critical in the presence of motion clutter. For dynamic backgrounds, clearly a 'segmentation' of the figure from the background 'by motion' as basis for a subsequent simple fitting of a form template is not possible. In this case, segmentation and recognition of the stimulus cannot be decoupled, and a simple testing of all possible template positions (and scales) is computationally not tractable. However, subjects are easily able to recognize point-light stimuli with arbitrary size and position in motion clutter. 3) Predictions from the hypothesis of exclusively form-based processing of biological motion are contradicted by experimental data (e.g. Thurman & Grossman, 2008). Furthermore, there is accumulating evidence from behavioral and imaging studies that support an essential involvement of both, form and motion information, in the recognition of biological motion. Likewise, it would be easy come up with a similar list of arguments against 'purely motion-based processing' of biological motion stimuli. In our view, a cue fusion account seems thus not only computationally more efficient, but also much more plausible in terms of cortical processing.

4.2.4 Limitations of the basic architecture

The described basic model architecture reproduced a range of empirical facts from electrophysiology, psychophysics, functional imaging, and even lesion studies in patients (e.g. Giese & Poggio, 2003). It specifically reproduced the recognition of point-light stimuli in substantial amounts of motion clutter without a priori knowledge of the position of the moving figure, and even from stimuli with degraded motion information such as the one proposed by Beintema and Lappe (Casile & Giese, 2005). Also the model reproduced the view-dependence of neurons that are selective for biological motion stimuli (Perrett et al. 1985, Oram & Perrett, 1996).

However, the original model was tested only with simplified artificial stimuli that had been used in neuroscience experiments. Recent extensions of such architectures have included neural models for the estimation of optic flow from gray-level videos. In addition, some of these studies have included learning of optimized mid-level features from example image sequences. A validation on bench-mark data bases from computer vision showed that such architectures can reach performance levels that are competitive with the state-of-the-art action recognition systems in computer vision that are not inspired by biology (Huang et al., 2007; Escobar et al. 2008; Schindler & van Gool, 2008).

In a biological context, however, a number of serious limitations of the proposed basic model architecture have to be addressed in future work.

For example, the described models have predominantly a feed-forward architecture and neglect the influence of attentional modulation and context information that is present in motion recognition in biological systems. (See Chapter 18 for a more detailed discussion of these issues.) In addition, this biologically-motivated work on action recognition has solely focused on non-transitive actions, which are not directed towards goals or objects. The major purpose of this chapter is to propose an extension of the original architecture that can account for the recognition of transitive actions by inclusion of a small set of additional neural mechanisms that are consistent with facts known from neurophysiology.

4.3 NEURAL ARCHITECTURE FOR THE RECOGNITION OF TRANSITIVE ACTIONS

The architecture described so far reproduces a number of properties of the neural mechanisms of the recognition of *non-transitive actions*, such as walking or waving, which do not involve a direct interaction with a goal object. In the following section we describe several extensions that make the general architecture described in Section 4.2 suitable for the recognition of *transitive actions* that are directed towards goal objects and specify interactions with them. More specifically, the proposed novel model accounts for the visual recognition of grasping acts from natural video sequences. In

comparison with the model described above, this novel architecture is distinguished by the following features:

1. In order to limit the computational complexity of the first implementation, the present version of the model contains only a form pathway. It turned out that form-based recognition was sufficient to accomplish relatively robust recognition form grasping videos. However, it seems possible that in the brain also the recognition of grasping acts combines form and motion features. A later extension of the model by inclusion of a motion pathway seems easily possible.
2. To accomplish robust performance on real video sequences the model was extended by an appropriate algorithm for the learning of optimized mid-level form features at the middle hierarchy levels of the form recognition module by unsupervised learning.
3. The model includes novel neural circuits, modelling computational mechanisms likely located in STS, the parietal and premotor cortex, which combine the information about the effector (hand) movement and the shape and location of the goal object, i.e. the object that is grasped.

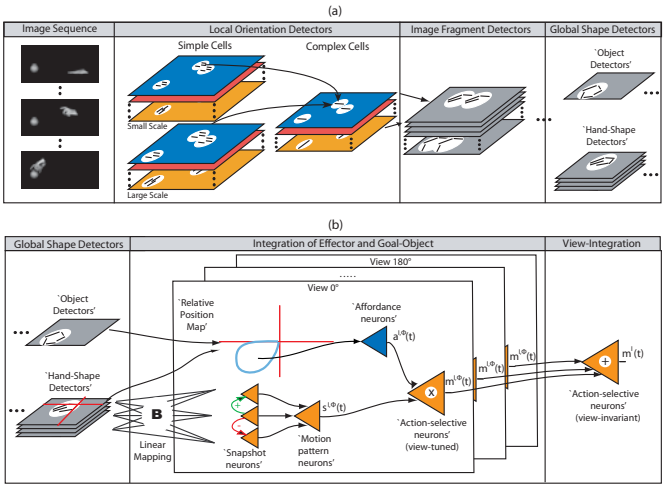


Figure 4.3

Figure 4.3 presnets an overview of the extended architecture. Panel (a) shows the neural hierarchy for the recognition of effector and object shapes. Its architecture closely resembles the form pathway of the original model presented in Figure 4.3. Panel (b) shows the additional proposed neural mechanisms that integrate the information about the goal object and the effector, and which are required to realize a view-independent recognition of hand actions. These additional mechanisms are sketched below in more detail.

A more extensive description of the model can be found in Fleischer et al. (2008, 2009a, 2009b).

4.3.1 Shape recognition hierarchy

The recognition of effector (hand) and object shape is based on the hierarchical neural framework introduced in Section 4.2.1, which was derived from well-established object recognition models (Perrett & Oram, 1993; Riesenhuber & Poggio, 1999; Mel & Fiser, 2000; Rolls & Milward, 2000; Serre et al., 2007). An overview of the shape recognition hierarchy is shown in Figure 4.3(a).

The first hierarchy level consist of orientation detectors that were again modeled by Gabor filters, with 12 different preferred orientations and two different spatial scales. 'Complex cell' responses were computed from the output signals of these detectors by pooling of responses of detectors with the same orientation preference and spatial scales, but slightly different spatial preferences, using a maximum operation (see Section 4.2.1.). The spatial receptive fields of these 'complex cells' would correspond to a size of about 1.0 deg, matching approximately the parameters derived from electrophysiological experiments. The output signals of all levels of the neural hierarchy were thresholded using linear threshold functions.

The next-higher level extracts form features of intermediate complexity. Opposed to the original model described before, the selectivity of these mid-level form detectors was optimized by learning. The detectors were

modeled by Gaussian radial basis functions (RBFs, see above) whose centers were defined by input signals from the previous hierarchy layer that were arising for training videos showing grasping actions. For the selection of a limited set (dictionary) of such intermediate form features we used a simple greedy clustering procedure that preserves frequently occurring features, while it tends to eliminate novel features that are redundant based on a similarity measure. The learned feature representation thus reflects the feature statistics that is present in the training data (Serre et al., 2007; Mutch & Lowe, 2006; Fidler et al., 2008). The thresholded responses of the learned feature detectors were pooled over a local spatial neighborhood diameter of 1.7 deg again using a maximum operation, generating detector responses with higher position invariance. The same procedure can be cascaded in order to generate several intermediate layers that extract more and more complex form features. For the given implementation we tested between two and four intermediate layers, dependent on the required recognition accuracy.

As for the model architecture discussed in section 4.2.1, the neural detectors at the highest level of the shape recognition hierarchy were given by Gaussian RBFs whose centers correspond to keyframes from training image sequences. These feature detectors are selective for individual views of objects and hands, and they correspond to the body shape snapshot detectors described in Section 4.2.1. Opposed to equivalent detectors in most standard object recognition models (e.g. Riesenhuber & Poggio, 1999), these highest-level shape detectors are *not* completely position-invariant. Instead, there

exists a small ensemble of detectors with different spatial preferences (tuning width corresponding to about 3.7 deg) for each learned shape. This makes it possible to read out the two-dimensional retinal position of the recognized shapes from this detector population using a simple population vector approach. (An example is an estimate of the stimulus position that is just given by a weighted average of the preferred positions of the individual detectors, where the weights are given by their normalized responses.) It will be shown below that the extraction of position information is important for the recognition of effective goal-directed actions, since it is crucial for the computation of the spatial relationship between the effector and goal object.

Several results from electrophysiology support the assumption that shape-selective neurons in ventral areas (such as IT) are characterized by a limited degree of position invariance (Baker et al., 2000; DiCarlo & Maunsell, 2003; Lehky et al., 2008, for a review see Kravitz et al., 2008), supporting this assumption of our model.

4.3.2 Selectivity for temporal order

For the recognition of the effector movements (like the closing of the grasping hand) the responses of the snapshot neurons were associated over time using a dynamic neural network (see section 4.2.2). For this purpose, the responses of hand shape detectors with different spatial preferences were pooled using a maximum operation, in order to realize position-independent detectors for the

recognition of individual hand shapes. In addition, we learned a linear remapping of the response of these detectors onto a one-dimensional activity map that provides input for the dynamic neural network. This mapping compensates for the fact that natural grips are characterized by strongly non-uniform speeds of the shape variations of the hand. (See Fleischer et al. (2009b) for further details.) The resulting input distribution can be characterized by a time-dependent vector with the elements r_k(t), where the index k indicates the position of the activity variable in the map. If the trained grasping action is presented, an activity peak arises in this map that propagates with approximately constant speed.

A dynamic neural network that results in sequence-selective responses can be derived from dynamic neural fields with asymmetric lateral connections (Zhang, 1996; Giese, 1999; Xiao & Giese, 2002). (A neural field is an idealized model for a network of neurons that encodes continuous parameters. In this case, the network can be approximated by a continuous neural medium with a continuous instead of a discrete index for the individual neurons. This approximation can make the mathematical treatment substantially easier (e.g. Amari, 1977; Giese, 1999)). To define the dynamic network, we assume the existence of dynamic *snapshot neurons* whose activity u_k(t) obeys the differential equation:

$$\tau \, \dot{u}_k(t) = -u_k(t) + \sum_l w(k-l)\left[u_l(t)\right]_+ + r_k(t) - h \qquad (1)$$

The function $[.]_+$ defines a *linear threshold function*, which is define as $[x]_+ = \max(x, 0)$. The time constant of the dynamics τ was about 160 ms. The positive threshold parameter h determines the resting level of the neural network without input. The interaction kernel $w(k)$ was chosen as a smooth asymmetric function that was adjusted in order to maximize the amplitude of the activity peak that emerged for the training stimulus sequence presented in the correct temporal order, and to minimize the response for the sequence presented in reverse temporal order. The snapshot neurons in this representation become activated strongly only if the presented hand shape matches the corresponding training shape, and if it occurs in the right temporal context. (See Giese & Poggio (2003) for further details.) The appropriate connection strength for the lateral connections $w(k)$ could be learned as well using physiologically plausible mechanisms (see Section 4.2.2).

The outputs of the snapshot neurons belonging to the same hand action were then temporally integrated by *motion pattern neurons* that obeyed the dynamical equation:

$$\tau_s \dot{s}\ (t) = -s\ (t) + \max_k \left[u_k(t)\right]_+ - h_s. \qquad (2)$$

The time constant of this integrator dynamics was given by $\tau_s = 200$ ms and h_s

is a positive threshold parameter. The motion pattern neurons respond for a particular view of a particular hand action, but continuously during the whole action sequence. If the corresponding hand shapes are presented in the wrong temporal order the response of these neurons is strongly reduced (Giese & Poggio, 2003).

4.3.3 Integration of the information about effector and object

The recognition of goal-directed actions requires not only the recognition of the effector movement but also the analysis of the relationship between the effector and the goal object. A grip that does not reach the goal object would be an unsuccessful action, and the same would be true if the hand shape is inappropriate for the realization of an efficient grip of the object. Neurophysiological data shows that many action-selective neurons show a critical dependence of their activity on the fact whether the relationship between effector and object is appropriate. For example, action-selective neurons in the STS or in area F5 in premotor cortex show strongly reduced responses if the grasping hand does not touch the goal object (‘mimicked action’) (Perrett et al. 1989; Umilita et al. 2001). A model for the visual recognition of goal-directed action thus has to account for this dependence of neural activity on the relationship between effector and object.

The proposed model explains this dependence by an additional circuit that receives its input from the hand shape and the object shape detectors

(Figure 4.3(b)). The proposed mechanism is critically dependent on the fact that, due to the incomplete position invariance of the shape detectors, the retinal positions of goal object and effector can be estimated from the responses of the shape detectors. The core of the proposed neural circuit is a neural map that represents the relative position of effector and object. In this two-dimensional *relative position map* the relative position of object and effector is encoded by a neural activity peak (Figure 4.3(b)). The map can be constructed by simple neural operations from the outputs of the shape detectors. Let $a_E(u,v,t)$ signify the retinotopic spatial activity map that represents the current effector position for a particular grip type, and let $a_O(u,v,t)$ correspond to the activity map that is defined by the object shape detectors of an object that can be efficiently grasped with this grip type. Then the activity in the relative position map can be computed according to the relationship:

$$a_{RP}(u,v,t) = \int a_O(u',v',t)\, a_E(u'-u,v'-v,t)\, du'\, dv'. \qquad (3)$$

This convolution can be computed by a simple neural architecture that is similar to a *gain field* (Salinas & Abbott, 1995; Pouget & Seijnowski, 1997), just by summing product terms from the two input maps. Gain fields have been a fundamental architecture for the realization of coordinate transformation, e.g. in parietal cortex. Due to the multiplication, a non-zero output signal in the map can arise only if both, effector and object are present

in the stimulus, and if the object shape is suitable for a grip that is encoded by the corresponding hand shape detectors. In this way the model matches the hand shape and the grip affordance of the object. In our present implementation the center of the relative position map corresponds to the position of the effector, and the object position is represented relative to the effector. Similar implementations of coordinate transformations by gain fields have been found in multiple regions in the parietal cortex. Examples are the change from an eye-centered to a head-centered frame of reference (Batista et al., 1999; Buneo et al., 2002), or the representation of the relative positions of object parts (Chafee et al., 2007).

A grip will only be successful if the effector is located within a certain range of spatial positions relative to a goal object. This range corresponds to a well-defined spatial region in the relative position map (Figure 4.3(b), region marked by the cyan curve). A further postulate of our theory is the existence of a class of neurons, called *affordance neurons* in the following, which sum the activity in the relative position map over these spatial regions. As a consequence, these neurons are activated only if both, effector and object are present and if they have a spatial relationship that is appropriate for a successful grip. We assume that the receptive fields of the affordance neurons are established by learning.

As final step of the integration of the information about effector and object we assume a multiplicative combination of the output signals of the affordance neurons and the motion pattern neurons that are detecting the same

grip. This multiplication is computed by the *view-dependent action-selective detectors* in our model (Figure 4.3(b)). These detectors respond only in the presence of the appropriate effector motion and of the right spatial relationship between effector and object.. Neurons with similar properties have been described in area STSa (Perrett et al., 1989) and area F5 (Rizzolatti et al., 2001). Motor neurons with specific tuning for object shapes and the relationship between object and effector have also been found in the parietal area AIP (Murata et al., 2000; Gardner et al., 2007; Baumann et al., 2009).

4.3.4 Integration of different views

The initial stages of our model realize recognition of learned views of the object and the effector. Correspondingly, the action-selective neural detectors described before are view-dependent, i.e. they depend on the visual perspective in which the action is observed. Their response thus decays if an action is presented with views that deviate increasingly from the training view. Such view-dependence is in accordance with electrophysiological data. View-dependent action-selective neurons have been observed in the STS (Perrett et al., 1985; Oram & Perrett, 1996), and the majority of F5 mirror neurons are also view-dependent (Caggiano et al., subm). This clearly argues in favor of a view-based representation of actions rather than of a full reconstruction of the three-dimensional geometry of effector and object. Such a full three-dimensional reconstruction is implicitly assumed by many other

models for the recognition of goal-directed actions (See Section 4.1).

In our model, view-independent action recognition is accomplished by pooling of the output signals of a limited number of view-specific modules using a maximum operation (Figure 4.3(b), right part). This realization of view-invariant recognition of actions closely matches a widely accepted principle for view-invariant object recognition in the ventral stream (e.g. Poggio & Edelman, 1990; Logothetis, Pauls & Poggio, 1995). Our simulations show that a very limited number of view-dependent modules is sufficient to accomplish fully view-independent recognition of actions.

4.4 SIMULATION RESULTS AND PREDICTIONS

In the following section a number of simulation results are presented that illustrate that the model is computationally powerful enough for the robust recognition of hand actions from real video sequences. In addition, the simulations demonstrate that the model reproduces a number of key properties of action-selective neurons in the cortex that have been observed in electrophysiological experiments. Further simulation results can be found in Fleischer et al. (2009a, 2009b).

4.4.1 Robust recognition from real video sequences

The recognition of hand actions from real videos is a challenging computer vision problem (e.g. Athitsos & Scarloff, 2003; Stenger et al., 2006; Kjellström et al., 2008). The distinction of different types of grips requires

highly accurate recognition of shape details. The difference between a precision grip (grasping with the index finger and the thumb) and a power grip (grasping with the full hand) might, depending on the view angle, be defined only by a relatively small number of pixels in a video stimulus. At the same time, recognition must be accomplished independent of the view of the action and of the position of the action stimulus within the visual field. Electrophysiological recordings from area F5 in macaque cortex show that action-selective neurons (mirror neurons) are highly selective for differences between different grip types. But also, they often show strong invariance against the position of the stimulus within the visual field (Gallese et al. 1996).

In order to investigate the behavior of our model with real-world stimuli, we recorded video sequences with four different hand actions filmed from 19 view angles using a CANON XL1-S camera with a frame rate of 25 Hz. The videos show a human hand grasping an object with different grip types. Videos were converted to gray-level and had a frame size of 350 times 315 pixels. The hand shape detectors were trained using example shapes derived from the original videos by color segmentation of hand an object. However, testing was based on unsegmented gray-level videos.

The performance of the snapshot neurons of the model is shown for the distinction between precision and power grip in Figure 4.4(a), assuming the same view for both grips. In addition, this simulation compared two variants of the model, one with sequence selectivity (red line) and one without

sequence selectivity (blue line). Classifications were based on the responses of the snapshot neurons, assigning the grip type as response that corresponded to the more strongly activated snapshot neuron for the same time step. Performance (percent of correct classifications) is plotted as function of the normalized time as fraction of the duration of the whole grasping action. The model was tested with novel grasping sequences that were not used for the training of the model.

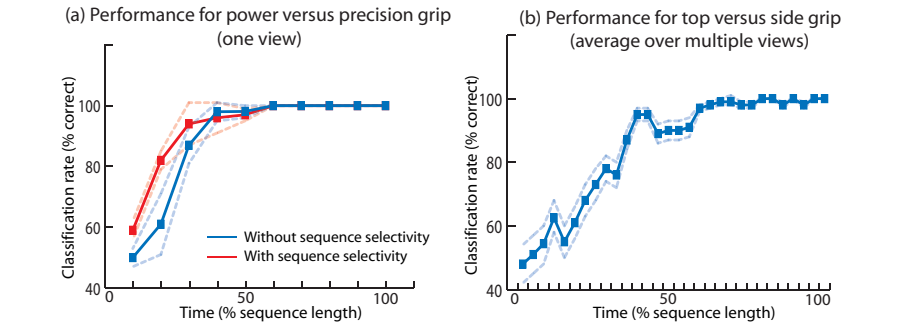


Figure 4.4

Figure 4.4(a) shows that almost perfect recognition performance is achieved already after about half of the overall duration of the action. In addition, the comparison between the two model variants shows that sequence selectivity results in a slight improvement for the classification in early stages of the grips, where the hand shapes of the two grips are still very similar. In this case sequential order information can disambiguate information arising from intermediate hand shapes. Even though the hand shape detectors were

trained with example images that did not contain the goal object the model efficiently generalizes to the stimuli including also goal objects.

A validation of the performance of the model with a variety of views is shown in Figure 4.4(b). In this case, the model was trained with 7 different views (in steps of 30 deg) of two actions (power grip from above and from the side of the object). Again the performance is shown for testing with videos that show the same action from novel views that were not presented during training. The average performance over all test views of the classification by the snapshot neurons is shown as function of the normalized duration of the action. Almost perfect performance is accomplished after less than 60 % of the overall time of the action, thus before the hand has reached its final configuration. This implies that the model accomplishes robust view-invariant recognition from real video stimuli, requiring only a relatively limited number of view-specific modules for each action.

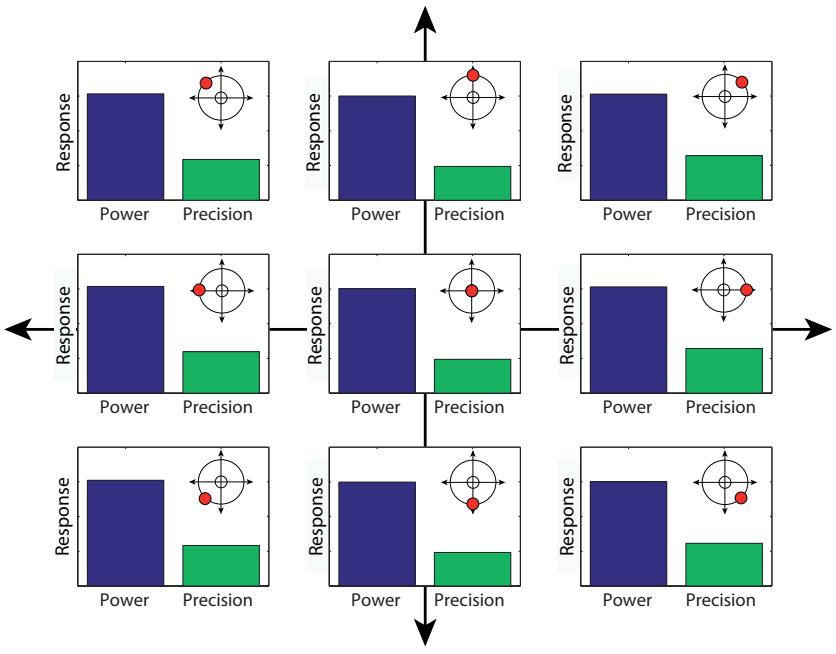


Figure 4.5

4.4.2 Position invariance

Many action-selective neurons, e.g. in premotor cortex, show only a weak variation of their responses with displacements of the stimulus in the visual field (Gallese et al., 1996). The model reproduces this strong position invariance.

In order to test position invariance, two grips (precision vs. power grip) were shown at nine different positions within the visual field of the model. As shown in Figure 4.5, the responses of action-selective neurons that are selective for precision or power grip almost do not change with the position of the stimulus within the visual field. This strong position invariance is achieved while the model shows high selectivity for the relative position of object and effector, as shown in Section 4.4.4.

Summarizing, the last two sections show that the proposed neural architecture even though it is based on very elementary neural operations, all of which in principle could be realized with cortical neurons, is sufficiently powerful to solve the hard computational problem of goal directed action recognition for real-world stimuli. The challenge of this problem is that high accuracy for the recognition of finger positions and the relationship between effector and object have to be realized together with substantial invariance against stimulus position and view.

4.4.3 View-dependence

View dependence is a natural consequence of the fact that the model is based on view-dependent representations of hand and object shapes. Such view dependence matches electrophysiological data, since view-dependent action-selective neurons have been observed in the STS (Perrett et al., 1985; Oram & Perrett, 1996) as well as in the premotor cortex of monkeys (Caggiano et al., subm.). View-tuning is also a common observation for neurons in ventral shape-selective areas, such as area IT (Logothetis, Pauls & Poggio, 1995; Tarr & Bülthoff, 1998).

The view tuning of action detectors at the highest level of the view-dependent modules (Figure 4.3(b)) is illustrated in Figure 4.6. In this case, seven view-specific modules have been trained that are selective for views that differ by 30 degrees. The training actions were power grips of a rod-like object from the side and from the top. The thin curves indicate the activities for the view-dependent action detectors, different colors indicating different view-dependent modules. All test views were different from the training views. Panels (a) and (c) show the response for a grip from the top, and panels (b) and (d) the responses for a grip from the side. The 'top grip neurons' (panels (a) and (b)) show strong responses only for the top grip, for views that are sufficiently close to their training view. In addition, they show a gradual decay of the tuning curve, which corresponds to a tuning width of about 60 deg. This view dependence and the tuning width match quantitatively

electrophysiological results obtained by studying the view dependence of mirror neurons in area F5 using similar stimuli (Caggiano et al., subm.). For the grip from the side, the 'top grip neurons' show only relatively weak responses and no clear view tuning. The behavior of the 'side grip neurons' from the view-dependent modules trained with side-grip stimuli is complementary: Strong responses arise only for side-grip stimuli, if their view is similar to the training view. Again, one finds smooth view-tuning curves with widths around 60 deg.

The thick lines in Figure 4.6 indicate the responses at the highest level of the model, that is formed by the view-invariant action detectors. These detectors show strong responses for all views of the trained action, and much smaller responses for the alternative action. Based on the responses of these detectors it is trivial to classify the actions (just selecting the action as recognized which corresponds to the action-selective detector with the higher activity). Most importantly, the direction of the activity difference between the two actions has the same sign for all views, and even for the untrained ones. This makes it possible to classify all views with only small number of trained view-dependent modules. View-independent action-selective neurons have been found in area F5 of the macaque (Caggiano et al. subm.), and in the STS (Perrett et al., 1989; Jellema & Perrett, 2006).

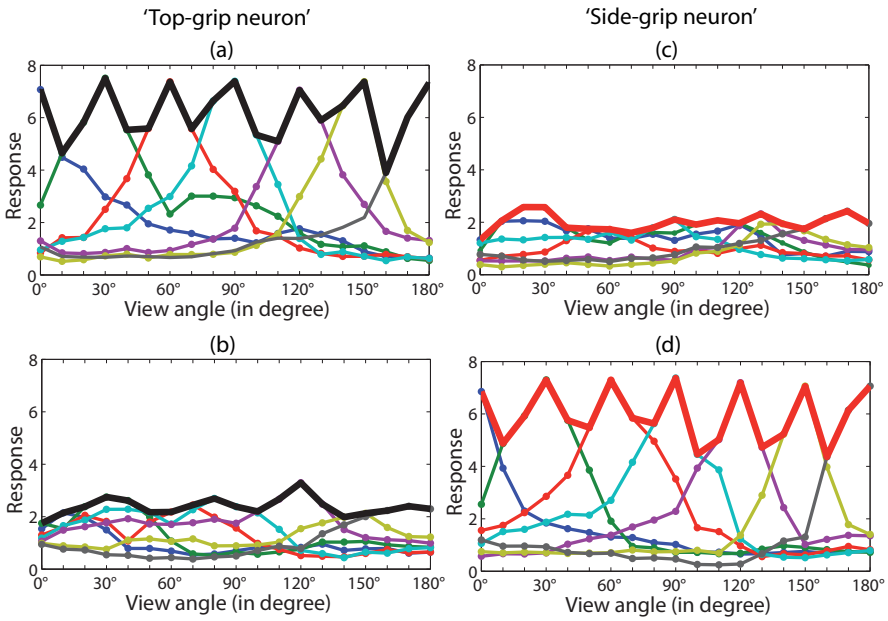


Figure 4.6

4.4.4 Selectivity for the relationship between effector and object

Many action-selective neurons show high selectivity of their responses for the relationship between effector and goal object. Mirror neurons in area F5 have been reported to fail to respond when either the effector or the goal object are missing in the stimulus (Umilta et al. 2001). In addition, many mirror neurons fail to respond for 'mimicked actions', where both effector and object are present, but where the effector misses the goal object, the experimenter grasping next to it. Similar observations have been made for action-selective neurons in the STS (Perrett et al., 1989).

The model reproduces this selectivity for the relationship between

effector motion and goal object even in quantitative detail. This is shown in Figure 4.7 that shows the activity of an action-selective neuron that has been trained with a power grip for stimuli that show effector and object with the correct spatial relationship, and with incorrect spatial relationship (‘mimicked action’). In addition, stimuli were tested which contained only the object or only the effector. The inset replots data from an electrophysiological study that has investigated the responses of neurons in the anterior STS using the same type of stimuli (Perrett et al. 1989). Clearly, the response of the action-selective detectors decays substantially if either the object or the effector is missing in the stimulus, or if a mimicked action is presented. The decay is even quantitatively similar to the response profile that was observed in the electrophysiological study.

We conclude that the proposed neural mechanism accounts, at least qualitatively, for this neurophysiological data. From the computational point of view, it seems not trivial to accomplish this selectivity for the spatial relationship between effector and object, at the same time guaranteeing strong position invariance for the action recognition.

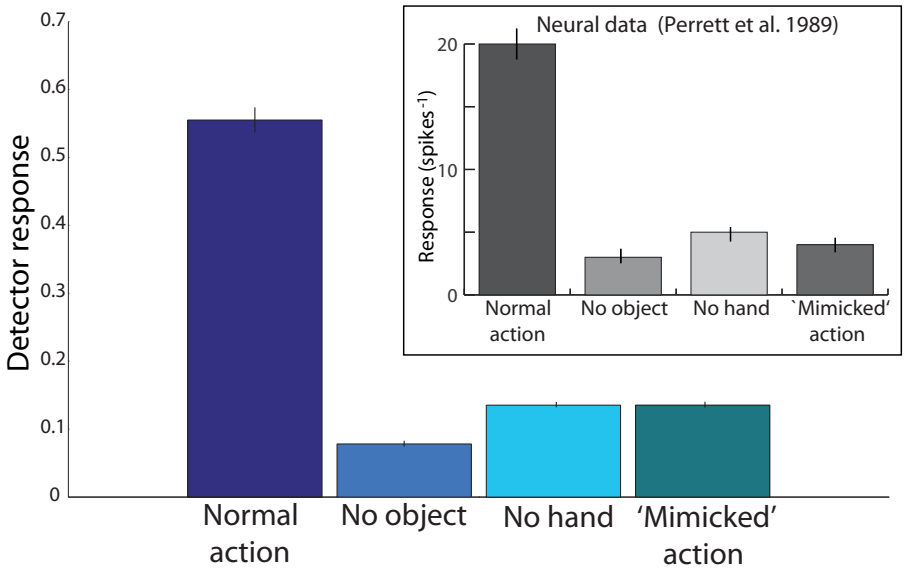


Figure 4.7

4.4.5 Predictions

The model is formulated largely in terms of mechanisms that could, in principle, be implemented in a relatively obvious way by real cortical neurons. This makes it possible to derive a number of predictions that can be tested immediately in experiments. Here are only a few examples:

- Action selective neurons should show *sequence selectivity*. This implies that the presentation of the same stimulus frames in forward and reverse order should result in different neural responses. In fact, initial data seems to confirm this prediction for action-selective neurons (mirror neurons) in area F5 of the macaque, obtaining a good

agreement between the behaviour of individual neurons in this area and the model (V. Caggiano, pers. comm.).

- Changing the relative positions of effector and object should result in well-defined gradual tuning curves for the dependence on relative position. This dependence should be invariant against position changes of the whole stimulus in the visual field. This prediction could be confirmed by recording of neurons, for example, in relevant areas in the parietal cortex or in area F5.
- The existence of the postulated neuron classes (affordance neurons, motion pattern neurons, view-dependent and view-invariant action-selective neurons) can be verified in electrophysiological studies. A coarse anatomical localization of the different postulated computational steps (relative position map, affordance neurons, etc.) might also be possible in carefully controlled fMRI experiments.

4.5 CONCLUSIONS AND OUTLOOK

We have reviewed in this chapter biologically inspired models for the visual recognition of body movements and actions. In Section 4.2 we have provided an overview of a class of architectures for the recognition of non-transitive actions that are not goal-directed, which meanwhile are relatively well established as models for brain functions and as basic architectures for

computer vision systems for action recognition. In the second part of this Chapter (Sections 4.3 and 4.4), we have presented an extension of this basic class of models making them suitable to account also for the recognition of transitive actions. We have shown that a model that is based on the proposed extensions is computationally powerful enough to realize recognition of transitive actions from real videos. In addition, we have shown that the model reproduces several neurophysiological results about action-selective cortical neurons. While it is still somewhat preliminary, this makes the proposed architecture interesting as starting point for the development of more elaborated models that can be fitted in much more detail to experimental data.

Obviously, the proposed model has a number of limitations, which at the same time define topics for future research. We list here only a few of the major points:

- The proposed model completely ignores the influence of disparity features, which would provide depth information obtained by a comparison of the retinal images from both eyes. It is know that many neurons, specifically in parietal regions, show disparity tuning (Tsutsui et al., 2005; Orban et al., 2006). It remains thus an important topic for future research to explore the role of disparity features in action recognition. To our knowledge, no neural model so far has addressed this topic.
- There has been an extensive discussion how the perception and

execution of actions are linked in the context of research on the 'mirror neuron system' (e.g. Prinz, 1997; Rizzolatti et al. 2001). Doubtless, there is strong empirical evidence for a tight interaction between neural representations for action execution and action perception as reviewed in Chapter 2 and Chapter 20. With respect to this discussion, the proposed model provides the insight that many visual tuning properties of action-selective visual neurons can be explained with relatively standard mechanisms that are also common to visual processes outside of action recognition. A direct coupling to motor representations or even the time-synchronous internal resimulation of the observed motor behavior within the motor system ('motor resonance') was not necessary to account for these results. However, the tight coupling of motor execution and visual recognition of body motion raises the question how exactly motor and visual representations are linked to each other and how the proposed model has to be modified to take this link into account. An interesting idea in this context is the existence of predictive dynamic mechanisms at multiple levels within a hierarchical system that can propagate predictions in a bottom-up and also a top-down fashion (Kiebel et al., 2008).

- The computational limits of the proposed architecture need to be investigated much more thoroughly, using more extended data sets that include many objects and more types of grips. Only in this way it

will be possible to judge how the proposed solution scales up for bigger problems and different tasks, such as the recognition of emotional expressions (Schindler et al 2008; See also Chapter 3). Another step that will be critical to make the system interesting for applications in computer vision and robotics is to improve the computational mechanisms at the different levels of the hierarchy in order to make the system applicable for action stimuli with background clutter.

- Another important problem that has almost not been addressed in the context of action vision is the influence of attention on the processing of complex motion stimuli (see e.g. Rodriguez-Sanchez et al., 2007). It seems likely that an integration of attentional selection will play a key role to make the proposed architecture applicable for more complex problems, like for visual scenes where multiple objects or effectors are present. Conversely, it also remains a question for future research of how action perception influences the control of attention, e.g. by directing attention towards action goals. (See also the discussion of top-down and bottom-up processes in Chapter 18).

Summarizing, we think that the proposed skeleton architecture might provide a first step toward more quantitative models for the visual recognition of goal-directed actions which make well-defined predictions that can be verified or falsified at the level of the behavior of single cells in relevant higher cortical

areas. Likely, only this approach will finally help to unravel the real neural circuits that underlie the visual processing of action stimuli and body motion.

Acknowledgments

We are grateful to M. Shiffrar for the invitation to write this chapter, to V. Caggiano for sharing his electrophysiological data and to L. Omlor for help with the video stimuli. We thank P. Thier and A. Casile, L. Fogassi and G. Rizzolatti for interesting discussion in the context of a project on 'mirror neurons'. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) SFB 550, EC FP6 project COBOL and FP7 project SEARISE. Further support from the Hermann Lilly Schilling foundation is gratefully acknowledged.

REFERENCES

Allman, J., Miezin, F., & Mc Guinness, E. (1985) Direction- and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT). *Perception* 14, 105-126.

Arbib, M.A. (2008) From grasp to language: embodied concepts and the challenge of abstraction. *J Physiol* 102, 4-20.

Athitsos, V. & Sclaroff, S. (2003) Estimating 3d hand pose from a cluttered image. *Proc. IEEE Int. Conf on Computer Vision and Pattern Recognition* 2, 432.

Baker, C.I., Keysers, C., Jellema, T Wicker, B., & Perrett, D.I. (2000) Coding of spatial position in the superior temporal sulcus of the macaque. *Curr Psychol Lett Behav Brain Cogn* 1, 71–87.

Batista, A.P., Buneo, C.A., Snyder, L.H., & Andersen, R.A. (1999) Reach plans in eye-centered coordinates. *Science* 285, 257.

Baumann, M.A., Fluet, M-C., & Scherberger, H. (2009) Context-specific grasp movement representation in the macaque anterior intraparietal area. *J Neurosci.* 29, 6436-48.

Beardsworth, T. & Buckner, T. (1981) The ability to recognize oneself from a video recording of one's movements without seeing one's body'. Bull. Psychon. Soc. 18, 19-22.

Beintema, J.P., & Lappe M. (2002) Perception of biological motion without local image motion. *Poc. Nat.l Acad. Sci. U S A*. 99, 5661-5663.

Beintema, J.A., Georg, K., & Lappe, M. (2006) Perception of biological motion from limited-lifetime stimuli. *Percept Psychophys*. 68, 613-24.

Bertenthal, BI, & Pinto, J. (1994). Global processing of biological motions. *Psychol. Science*, 5, 221-225.

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

Blake, A. & Isard, M. (1998) *Active Contours*. Springer, Berlin.

Blakemore, S.J. & Frith, C. The role of motor contagion in the prediction of action. *Neuropsychologia* 43, 260-67.

Bobick, A. (1997) Movement, activity, and action: The role of knowledge in the perception of motion. *Phil. Trans. Royal Society London B* 352, 1257-1265.

Bonaiuto, J., Rosta, E., & Arbib, M. (2007) Extending the mirror neuron system model, I.

audible actions and invisible grasps. *Biol Cybern* 96, 9–38.

Born, R.T. (2000) Center-surround interactions in the middle temporal visual area of the owl monkey. *J Neurophysiol*. 84, 2658-2669.

Buneo, C.A. Jarvis, M.R., Batista, A.P., & Andersen, R.A. (2002) *Nature* 416, 632-36.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C.E., Riesenhuber, M. & Poggio, T. (2007) A model of V4 shape selectivity and invariance. *J Neurophysiol.* 98, 1733-50.

Calvo-Merino, B., Grèzes, J., Glaser, D.E., Passingham, R.E., & Haggard P. (2006) Seeing or doing? Influence of visual and motor familiarity in action observation. *Curr Biol*. 16, 1905-10.

Casile, A. & Giese M.A. (2005) Critical features for the recognition of biological motion. *J Vis.* 5, 348-360.

Chafee, M.V., Averbeck, B.B., & Crowe, D.A. (2007) Representing spatial relationships in posterior parietal cortex: single neurons code object-referenced position. *Cerebral Cortex* 17, 2914-32.

Chouchourelou, A., Matsuka, T., Harber, K. & Shiffrar, M. (2006) The visual analysis of emotional actions. *Soc Neurosci.* 1, 63-74.

Cutting, J.E., & Kozlowski, L.T. (1977) Recognizing friends by their walk: Gait perception without familiarity cues. *Bull. Psychon. Soc.* 9, 353-356.

Cutting, J.E., Moore, C., & Morrison, R. (1988) Masking the motions of human gait. *Percept. Psychop*. 44, 339-347.

Demiris, Y., & Simmons, G. (2006) Perceiving the unusual: temporal properties of hierarchical motor representations for action perception. *Neural Netw* 19, 272–84.

DiCarlo, J.J. & Maunsell, J.H.R (2003) Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object position. *J Neurophysiol* 89, 3246-78.

Dinstein, I., Hasson, U., Rubin, N., & Heeger, D.J. (2007) Brain areas selective for both observed and executed movements. *J Neurophysiol.* 98, 1415-27.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992): Exp. Brain Res. 91, 176-180.

Dittrich, W.H. (1993) Action categories and the perception of biological motion.

Perception. 22, 15-22.

Dittrich, W.H., Troscianko, T., Lea, S.E., & Morgan, D. (1996) Perception of emotion from dynamic point-light displays represented in dance. *Perception* 25, 727-738

Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. *Proc. Intl. Conf. on Computer Communications and Networks,* 65-72.

Efros, A., Berg, A., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Proc. IEEE Intl. Conf. on Computer Vision,* Vol. 2, pp. 726–734.

Eifuku, S., & Wurtz, R.H. (1998) Response to motion in extrastiate area MSTl: center-surround interactions. *J. Neurophysiol*. 80, 282-296.

Erlhagen,W., Mukovskiy, A., & Bicho, E. (2006) A dynamic model for action understanding and goal-directed imitation. *Brain Research* 1083, 174–88.

Escobar, M., Masson, G., Vieville, T., & Kornprobst, P. (2009) Action Recognition Using a Bio-Inspired Feedforward Spiking Network. *Int J Comput Vis* 82, 284–301.

Fagg, A.H. & Arbib, M.A. (1998) Modeling parietal-premotor interactions in primate control of grasping. *Neural Netw.* 11, 1277-1303.

Felleman D.J., & van Essen, D.C. (1991) Distributed hierarchical processing in the primate visual cortex. *Cereb. Cortex* 1, 1-49.

Ferrari, P.F., Bonini, L., & Fogassi, L. (2009) From monkey mirror neurons to primate behaviours: possible 'direct' and 'indirect' pathways. *Philos Trans R Soc Lond B Biol Sci.* 364, 2311-23.

Filder, S., Boben, M., & Leonardis, A. (2008) Similarity-based cross-layered hierarchical representation for object categorization. *Proc. IEEE Conf on Comp Vision and Pattern Recognition* 1.

Fleischer, F., Casile, A., & Giese, M.A. (2008) Neural model for thevisual recognition of goal-directed movements. *In: Kurkova V, Neruda R , Koutnik J (eds.): Intl Conf Artificial Neural Networks, Part II* , *LNCS* 5164, 939-948.

Fleischer, F., Casile, A., & Giese, M.A. (2009a) Bio-inspired approach for the recognition of goal-directed hand actions. *In: Jiang, X. & Petkov (eds.): Intl. Conf on Comp Anal of Images and Patterns, LNCS* 5702, 714-722.

Fleischer, F., Casile, A., & Giese, M.A. (2009b) View-independent recognition of grasping actions with a cortex-inspired model. *Proc. IEEE Intl Conf on Humaniod Robots,* Paris, (in press).

Fogassi, L., Ferrari, P.F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005) Parietal lobe: from action organization to intention understanding. *Science* 29, 662-67.

Freedberg, D. & Gallese, V. (2007) Motion, emotion and empathy in esthetic experience. *Trends Cogn Sci.* 11, 197-203.

Frith, C.D. & Singer, T. (2008) The role of social cognition in decision making. *Philos Trans R Soc Lond B Biol Sci.* 363, 3875-86.

Fukushima, K. (1980) Neocognitron. A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern*. 36, 193-202.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996) Action recognition in the premotor cortex. *Brain* 119, 593-609.

Gallese, V. & Goldman, A. (1998) Mirror neurons and the simulation theory of mindreading, *Trends Cogn. Sci.* 2, 493–501.

Gardner, E.P., Babu, K.S., Reitzen, S.D., Ghosh, S., Brown, A.S., Chen, J., Hall, A.L., Herzlinger, M.D., Kohlenstein, J.B., & Ro, J.Y. (2007) Neurophysiology of prehension. III. Representation of object features in posterior parietal cortex of the macaque monkey. *J Neurophysiol.* 98, 3708-3730.

Gavrila, D.M. (1999) The visual analysis of human movement: a survey. *Comp. Vis. Image Underst*. 73, 82-98.

Giese, M.A. (1999) *Dynamic neural field theory for motion perception.* Kluwer Academic Publishers, Dordrecht, Netherland.

Giese. M.A. (2004) Neural model for biological movement recognition. In: L.M. Vaina, S. A. Beardsley, S. Rushton (eds.): *Optic Flow and Beyond.*Kluwer, Dordrecht.

Giese, M.A., & Poggio, T. (2003) Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* 4, 179-192.

Goodale, M.A., & Milner, A.D. (1992) Separate visual pathways for perception and action. *Trends Neurosci.* 15, 97–112.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as Space-Time Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 2247-2253.

Haruno, M., Wolpert, D. M., & Kawato, M. (2001). Mosaic model for sensorimotor learning and control. *Neural Comp.* 13, 2201–2220.

Hoffman, D.D. & Flinchbaugh, B.E. (1982) *The interpretation* of biological motion. *Biological Cybernetics*, 42,195-204.

Hubel, D.H., & Wiesel, T.N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol (Lond.)* 160, 106-154.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J.C., & Rizzolatti, G. (2005) Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol.* 3, e79.

Jastorff, J., Giese, M.A. (2004) Time-dependent hebbian rules for the learning of templates for motion recognition. In: *Ilg, U., Bülthoff, H.H., Mallot, H.A.M. (eds.): Dynamic Perception Infix, Berlin 5*, 151-156

Jastorff, J. & Orban, G.A. (2009) Human functional magnetic resonance imaging reveals separation and integration of shape and motion cues in biological motion processing. *J Neurosci.* 29, 7315-29.

Jellema, T. & Perrett, D.I. (2006) Neural representations of perceived bodily actions using a categorical frame of reference. *Neuropsychologia* 44, 1535–1546.

Jhuang, H., Serre, T., Wolf, L., & Poggio, T. (2007). A biologically inspired system for action recognition. In *Proc. IEEE Intl.Conf. on Computer Vision,* 1–8.

Johansson, G. (1973) Visual perception of biological motion and a model for its analysis. *Perc. Psychophys*. 14, 201-211.

Kiebel, S.J., Daunizeau, J., & Friston, K.J. (2008) A hierarchy of time-scales and the brain. *PLoS Comput Biol* 4, e1000209.

Kilner, J.M., Paulignan, Y., & Blakemore, S.J. (2003) An interference effect of observed biological movement on action. *Curr Biol* 13, 522-25.

Kilner, J.M., Neal, A., Weiskopf, N., Friston, K.J., & Frith, C.D. (2009) Evidence of mirror neurons in human inferior frontal gyrus. *J Neurosci*. 29, 10153-9.

Kjellström, H., Romero, J., Martínez, D., & Kragić, D. (2008) Simultaneous visual recognition of manipulation actions and manipulated objects. *Proc. IEEE Europ Conf on Computer Vision* , 336–349.

Kouh, M. & Poggio, T. (2008) A canonical neuronal circuit for cortical nonlinear operations. *Neural Comp.* 20, 1427-1451.

Kravitz, D.J., Vinson, L.D., & Baker, C.I. (2008) How position dependent is visual object recognition? *Trends Cogn Sci.* 12, 114-22.

Lange, J. & Lappe, M. (2006). A model of biological motion perception from configural form cues. *J. Neurosci.*, 26, 2894-2906.

Lange, J., Georg, K., & Lappe, M. (2006) Visual perception of biological motion by form: a template-matching analysis. *J Vis* 6, 836-49.

Laptev, I. & Lindeberg, T. (2003) Space-time interest points. *Proc. IEEE Intl. Conf on Computer Vision*, 432–439.

Lehky, S.R., Peng, X., McAdams, C.J., & Sereno, A.B. (2008) Spatial Modulation of Primate Inferotemporal Responses by Eye Position. *PLoS ONE* 3, e3492.

Logothetis, N.K., Pauls, J. & Poggio, T. (1995) Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol*. 5, 552-563.

Logothetis, N.K., & Sheinberg, D.L. (1996) Visual object vision. *Ann. Rev. Neurosc*. 19, 577-621.

Marr, D. & Vaina, L.M.V (1982) Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London B*, 214, 501-524.

Mel, B, & Fiser, J. (2000) Minimizing binding errors using learned conjunctive features. *Neural Comp*. 9, 779-796.

Metta, G., Sandini, G., Natale, L., Craighero, L., & Fadiga, L. (2006) Understanding mirror neurons: a bio-robotic approach. *Interaction Studies, special issue on Epigenetic Robotica* 7, 197–232.

Moeslund, T. B. & Granum, G. (2001) A survey of computer vision-based human motion capture. *Comp. Vis. Image Underst*., 81, 231-268.

Murata, A., Gallese, V., Luppino, G., Kaseda, M., & Sakata, H. (2000). Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. *J. Neurophysiol.* 83, 2580-2601.

Mutch, J. & Lowe, D.G. (2006) Multi-class object recognition with sparse, localized features. *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition* 1, 11-18.

Nelissen, K., Vanduffel, W. & Orban, G.A. (2006) Charting the Lower Superior Temporal Region, a New Motion-Sensitive Region in Monkey Superior Temporal Sulcus. *J. Neurosci*. 26, 5929 –5947.

Niyogi, S.A & Adelson, E.H. (1994) Analyzing and recognizing walking figures in XYT. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1994, 469–474.

Oram, M.W., & Perrett, D.I. (1996) Integration of form and motion in the anterior temporal polysensory area (STPa) of the macaque monkey. *J. Neurophys*. 76, 109-129.

Orban, G.A., Janssen, P., & Vogels, R. (2006) Extracting 3D structure from disparity. *Trends Neurosci* 29, 466-473.

Oztop, E. & Arbib, M.A. (2002) Schema design and implementation of the grasp-related mirror neuron system. *Biol Cybern*. 87, 116-140.

Oztop, E., Kawato, M., & Arbib, M. (2006) Mirror neurons and imitation: computationally guided review. *Neural Netw* 19, 254–71.

Perrett, D.I., Smith, P.A., Mistlin, A.J., Chitty, A.J., Head, A.S., Potter, D.D., Broennimann, R., Milner, A.D., & Jeeves, M.A. (1985) Visual Analysis of body movements by neurones in the temporal cortex in the macaque monkey: a preliminary report. *Behav. Brain Res.* 16, 153-170.

Perrett, D.I., Harries, M.H., Bevan, R., Thomas, S., Benson, P.J., Mistlin, A.J., Chitty, A.J., Hietanen, J.K., & Ortega, J.E. (1989) Frameworks of analysis for the neural representation of animate objects and actions. *J Exp Biol* 146, 87–113.

Perrett, D.I. & Oram, M.W. (1993) Neurophysiology of shape processing. *Img. Vis. Comput*. 11, 317-333.

Perrett, D.I., Xiao, D., Barraclough, N.E., Keysers, C. & Oram, M. (2009) Seeing the future: Natural image sequences produce "anticipatory" neuronal activity and bias perceptual report. *Quart. J. Exp. Psych.* 62, 2081-2104.

Peuskens, H., Vanrie, J., Verfaillie, K., & Orban, G.A. (2005) Specificity of regions processing biological motion. *Eur J Neurosci.,* 2864-75.

Poggio, T. & Edelmann, S. (1990) A network that learns to recognize three-dimensional objects. *Nature* 343, 263-266.

Pollick, F.E., Lestou, V., Ryu, J., & Cho, S.B. (2002) Estimating the efficiency of recognizing gender and affect from biological motion. *Vision Res*. 42, 2345-55.

Pouget, A. & Sejnowski, T.J. (1997) Spatial transformations in the parietal cortex using basis functions. *J Cogn Neurosci* 9, 223-37.

Prevete, R., Tessitore, G., Santoro, M., & Catanzariti, E. (2008) A connectionist architecture for view-independent grip-aperture computation. *Brain Research* 1225, 133–145.

Prinz, W. (1997) Perception and action planning. Europ. *J. Cogn. Psychol*. 9, 129-154.

Riesenhuber, M., & Poggio, T. (1999) Hierarchical models of object recognition. *Nat. Neurosc*. 2, 1019-1025.

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

Rizzolatti, G. & Craighero, L. (2004) The mirror-neuron system. *Annu. Rev. Neurosci.* 2004. 27, 169–92

Rizzolatti, G. & Fabbri-Destro, M. (2008) The mirror system and its role in social cognition. *Curr Opin Neurobiol* 18, 179-84.

Rodriguez-Sanchez, A.J., Simine, E., & Tsotsos, J.K. (2007) Attention and visual search. *Int J Neural Syst*. 17, 275-88.

Rolls, E.T., & Milward, T. (2000) A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures" *Neural Comput.* 12, 2547-2572.

Sakata, H., Taira, M., Kusunoki, M., Murata, A., Tanaka, Y. (1997) The TINS Lecture: The parietal association cortex in depth perception and visual control of hand action. *Trends Neurosci* 20, 350-357.

Saleem, K.S., Suzuki, W., Tanaka, K., & Hashikawa, T. (2000) Connections between anterior inferotemporal cortex and superior temporal sulcus regions in the macaque monkey. *J. Neurosci.* 20, 5083-5101.

Salinas, E. & Abbott, L.F. (1995) Transfer of coded information from sensory to motor networks. *J Neurosci* 75, 6461-74.

Sauser, E.L. & Billard, A.G. (2006) Parallel and distributed neural models of the ideomotor principle: an investigation of imitative cortical pathways. *Neural Netw* 19, 285–98.

Schaal, S., Ijspeert, A., & Billard, A. (2003). Computational approaches to motor learning by imitation. *Phil Trans R Soc London. Series B, Biol Sci*, 358, 537–547.

Schindler, K. & van Gool, L. (2008). Combining densely sampled form and motion for human action recognition. In: Rigoll, Gerhard (ed.): *Proc. DAGM symposium, LNCS* 5096, 122-131.

Schütz-Bosbach, S. & Prinz, W. (2007) Perceptual resonance: action-induced modulation of perception. *Trends in Cogn. Sci.* 11, 349-55.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007) Robust object recognition with cortex-like mechanisms. *IEEE Trans Pattern Anal Mach Intell* 29 (2007) 411-26.

Sigala, R., Serre, T., Poggio, T., & Giese, M.A. (2005): Learning features of intermediate complexity for the recognition of biological motion. *Proc. Intl Conf on Artificial Neural Networks*, *LNCS* 3696, 241-246.

Singer, J.M. & Sheinberg, D.L. (2010): Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. *J. Neurosci.* 30, 3133–3145.

Smith, A.T., & Snowden, R.J. (1994) *Visual Detection of Motion*. London: Academic Press, 1994.

Stenger, B. Thayananthan, A., Torr, P., & Cipolla, R. (2006) Model-based hand tracking using a hierarchical bayesian filter. *IEEE Trans Pattern Anal and Mach Intell* 28, 1372–1384.

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 Netw* 17, 1273–89.

Tarr M.J., & Bülthoff, H.H. (1998) Image-based object recognition in man, monkey and machine. *Cognition* 67, 1-20.

Thirkettle, M., Benton, C.E. & Scott-Samuel, N.E. (2009) Contributions of form, motion and task to biological motion perception. *J Vision* 9, 28.1-11.

Thornton, I. M., Pinto J., & Shiffrar, M. (1998). The visual perception of human locomotion. Cognitive Neuropsychology, 15, 535-552.

Thurman, S.M. & Grossman, E.D. (2008) Temporal "Bubbles" reveal key features for point-light biological motion perception. *J Vis* 8, 28.1-11.

Tsutsui, K-I., Taira, M., & Sakata, H. (2005) Neural mechanisms of three-dimensional vision. *Neurosci Res* 51, 221-29.

Umiltà, M.A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001) I know what you are doing: a neurophysiological study. *Neuron* 31, 155-65.

Ungerleider, L. G., & Mishkin, M. (1982) Two cortical visual systems. In: D.J. Ingle, M.A. Goodale, & R.J.W. Mansfield, R.J.W. (Eds.), *Analysis of Visual Behavior* (pp. 549-586) Cambridge: MIT Press.

Vangeneugden, J., Pollick, F., & Vogels, R. (2009) Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. *Cereb Cortex.* 19, 593-611.

Webb, J.A. & Aggarwal, J.K. (1982) Structure from motion of rigid and jointed objects. *Artif. Intell*. 19, 107-130.

Wilson, M. & Knoblich, G. (2005) The case of motor involvement in perceiving conspecifics. *Psychol. Bull*. 131, 460-73.

Wolpert, D.M. & Ghahramani, Z. (2000) Computational principles of movement neuroscience. *Nat Neurosci.* 3, 1212-1217.

Wolpert D.M., Doya K. & Kawato M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society*, 358. 593–602.

Xiao, D,K., Raiguel, S., Marcar, V., Koenderink, J., & Orban, G.A. (1995) Spatial heterogeneity of inhibitory surrounds in the middle temporal visual area. *Proc. Natl. Acad. Sci. USA* 92, 11303-11306.

Xie, X., & Giese, M.A. (2002) Nonlinear dynamics of direction-selective nonlinear neural media. *Phys Rev E Stat Nonlin Soft Matter Phys.* 65, 051904.

Zhang, K. (1996) Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J Neurosci* 16, 2112-26.

Figure Captions

Figure 4.1 Recognition of actions based on predictions generated by forward models. (a) During action execution a controller generates a motor command that depends on the desired motor state and an error signal, which results from a comparison of the true sensory input and the sensory input that is predicted from the motor command by an internal forward model. In absence of any perturbations the predicted perceptual state matches exactly the sensory feedback, so that the error signal disappears. (b) During action observation no real motor output is generated. However, in the context of an internal simulation, motor commands might be generated that are mapped onto predicted sensory outcomes by forward models. By computation of the difference between the predicted sensory outcomes and the true sensory information it is possible to determine the dynamic state that would correspond to the actually observed movement, and to determine the type of the observed action by comparing the prediction errors between multiple available controller models. (Modified from Wolpert et al. 2003).

Figure 4.2 Overview of the basic architecture for the recognition of non-goal-directed body movements with two pathways for the processing of form and optic flow information. Abbreviations indicate potentially corresponding

areas in monkey and human cortex (V: *visual cortex*, M(S)T: *medial (superior) temporal cortex*, KO*: kinetic-occipital area*, IT: *inferior temporal cortex,* EBA: *extrastriate body area*, FBA: *fusiform body area*, IPL: *inferior parietal lobule*, F5: *premotor cortex*).

Figure 4.3 Overview of the extended architecture for the recognition of goal-directed actions. (a) Neural hierarchy for the recognition of effector and object shapes. (b) Mechanisms for the integration of the information about the goal object and the effector (hand).

Figure 4.4 Classification performance of the model for real video stimuli. (a) Classification performance of the snapshot neurons for precision vs. power grip presented with the same view. A model including the neural mechanism for sequence selectivity (red curve) was tested against a version of the model without sequence selectivity (blue curve). (b) Testing with multiple views. Classification performance for a power grip from the top vs. a power grip from the side presented with 12 different views that were disjoint from the training views. Dashed curves signify standard error over repeated simulations.

Figure 4.5*:* Position invariance for precision vs. power grip. The responses of neurons at the highest hierarchy level selective for power grips are shown during presentation of power grip stimuli (blue bars) and of precision grip

stimuli (green bars) for nine different positions of the stimuli within the visual field (indicated by red dots in the insets).

Figure 4.6 Activity of the action-selective neurons (see Figure 4.3). Thin lines indicate the activity of the *view-dependent* detectors and thick lines the one of the corresponding *view-independent* detectors. Panels (a) and (c) show responses to a stimulus showing a power grip from the top, and panels (b) and (d) the responses for a power grip from the side. Panels (a) and (b) show the response for the neural modules that have been trained with a top grip ('top grip neurons'), and panels (c) and (d) the responses of the neurons from the model trained with the side grip ('side grip neurons'). Thick lines indicate the responses of the corresponding view-independent action detectors. Test views were different from the views that were used to train the model.

Figure 4.7 Selectivity for the correct relationship between effector and object. Responses are shown for an action-selective detector (selective for power grip) for a normal grasping stimulus, a mimicked action (the hand not reaching the object), and stimuli where either the hand or the object was missing. Error bars indicate standard deviation over ten independent simulations. The inset shows corresponding neurophysiological data from an electrophysiological experiment by Perrett et al. (1989).