rich3/zns-neusci/zns-neusci/zns01513/zns3601p13z | xppws | S=3 | 3/26/13 | 6:10 | Art: 4129-12 | Input-MH

Behavioral/Cognitive

Physiologically Inspired Model for the Visual Recognition of Transitive Hand Actions

AQ: au Falk Fleischer,[1,2] Vittorio Caggiano,[1] Peter Thier,[1] and Martin A. Giese[1,2]

[1]Hertie Institute for Clinical Brain Research and [2]Werner Reichardt Centre for Integrative Neuroscience, University Clinic Tübingen, 72076 Tübingen, AQ: A Germany

AQ: B The visual recognition of actions is an important visual function that is critical for motor learning and social communication. Action-selective neurons have been found in different cortical regions, including the superior temporal sulcus, parietal and premotor cortex. Among those are mirror neurons, which link visual and motor representations of body movements. While numerous theoretical models for the mirror neuron system have been proposed, the computational basis of the visual processing of goal-directed actions remains largely unclear. While most existing models focus on the possible role of motor representations in action recognition, we propose a model showing that many critical properties of action-selective visual neurons can be accounted for by well-established visual mechanisms. Our model accomplishes the recognition of hand actions from real video stimuli, exploiting exclusively mechanisms that can be implemented in a biologically plausible way by cortical neurons. We show that the model provides a unifying quantitatively consistent account of a variety of electrophysiological results from action-selective visual neurons. In addition, it makes a number of predictions, some of which could be confirmed in recent electrophysiological experiments.

AQ: C

Introduction

Fn1 Motor actions are often directed toward goal objects, such as grasping of a piece of food. The recognition of such transitive goal-directed actions is an important function of the visual system with high importance for motor learning and the interpretation of the actions of others. The neural basis of this visual capability is only partially understood. Neurons with visual selectivity for goal-directed hand actions have been found in multiple regions of monkey cortex, including the superior temporal sulcus (STS) (Perrett et al., 1989; Jellema and Perrett, 2006; Barraclough et al., 2009), parietal cortex (Fogassi et al., 2005; Rozzi et al., 2008; Bonini et al., 2010), and premotor cortex (for a review, see Rizzolatti and Sinigaglia, 2010). A subgroup of these neurons that has received enormous interest in cognitive neuroscience is the "mirror neurons," which combine visual selectivity for observed actions with selective motor tuning during action execution. (See Materials and Methods, Transitive action-selective neurons and view-independence.)

Most existing computational models for goal-directed action recognition have focused on the possible role of motor representations, and the "mirror neuron system" for action understanding (Wolpert et al., 2003; Oztop et al., 2006) (see Materials and Methods, Relationship to other models, for a more detailed review). Most of these models assume, implicitly, that action recognition occurs by a matching of observed and internally simulated motor behavior within a body-centered frame of reference, e.g., using joint angle representations. First, this computational approach predicts view-independence of the relevant neural representations. Second, this computational approach requires a relatively accurate reconstruction of the three-dimensional effector geometry, even from monocular action stimuli. AQ: D

The first point seems difficult to reconcile with the observation that many action-selective neurons in monkey cortex show view dependence, e.g., in the STS (Perrett et al., 1985; Oram and Perrett, 1996; Jellema and Perrett, 2003; Barraclough et al., 2009), and recently also area F5 in premotor cortex (Caggiano et al., 2011). A transformation in a body-centered frame might thus not occur until very late in the cortical processing hierarchy. View-dependent mechanisms are meanwhile accepted as a standard explanation for the recognition of three-dimensional shapes in the ventral stream. (See Materials and Methods, Relationship to other models, for a more detailed discussion.) With respect to the second point, it is known from computer vision that the estimation of three-dimensional joint angles from monocular image sequences is a very challenging computational vision problem (Weinland et al., 2011), and one might ask whether the brain really solves this problem if action recognition can be accomplished by computationally less costly strategies, e.g., bypassing the three-dimensional reconstruction of the effector configuration.

We present in the following a physiologically plausible model that reproduces visual properties of action-selective neurons in

Received Aug. 28, 2012; revised Feb. 5, 2013; accepted Feb. 14, 2013.

Author contributions: M.A.G. and F.F. designed research; F.F. performed research; F.F. analyzed data; F.F., V.C., P.T., and M.A.G. wrote the paper.

This research was supported by the Deutsche Forschungsgemeinschaft (SFB 550-C10, GI 305/4-1), EU projects FP7-ICT-215866 SEARISE, FP7-249858-TP3 TANGO, FP7-ICT-248311 AMARSi. We thank L. Fogassi, G. Rizzolatti, D. Endres, and J. Pomper for helpful discussions. We are grateful to A. Christensen for help with the simulations, and to M. Angelovska for help with the editing of the figures.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. Martin Giese, Section for Computational Sensomotorics, Hertie Institute for Clinical Brain Research and Werner Reichardt Centre for Integrative Neuroscience, Otfried-Müller-Strasse 25, D-72076 Tübingen, Germany. E-mail: martin.giese@uni-tuebingen.de.

DOI:10.1523/JNEUROSCI.4129-12.2013

 0270-6474/13/330001-18$15.00/0

rich3/zns-neusci/zns-neusci/zns01513/zns3601p13z | xppws | S=3 | 3/26/13 | 6:10 | Art: 4129-12 | Input-MH

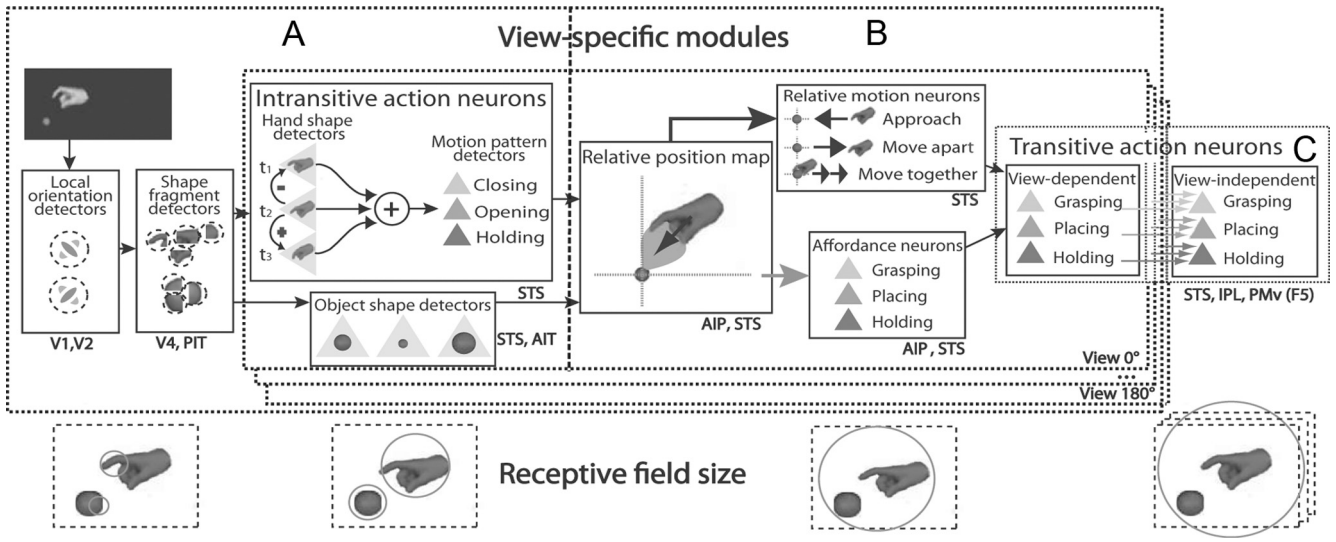

Figure 1. Overview of the model architecture for transitive action recognition. Model components: ***A***, Detector hierarchy for recognizing effector actions and shapes of goal objects. ***B***, Module that integrates information about effector movement and the goal-object. The central elements of this module is an RPM that represents the effector in image coordinates relative to the position of the goal object. From this map the responses of neural detectors for the matching of grip affordance and for the type for the relative motion between hand and object are derived. ***C***, Transient action-selective neurons combine the information from the previous modules. The recognition of goal-directed grasping acts is first accomplished in a view-dependent manner by view-specific modules. Only the highest hierarchy level pools the information over different views (view-independent transitive action neurons). The approximate receptive field sizes for the different levels compared with the stimuli are indicated by the insets below.

higher cortical areas of monkey cortex. The model accomplishes action recognition without an explicit reconstruction of the three-dimensional effector geometry, relying on well-established simple neural principles. The model is computationally powerful enough to recognize actions from real video sequences, accomplishing position- and view-invariance. It provides a unifying account for a variety of electrophysiological and imaging results from monkey cortex.

Materials and Methods

Overview of the model architecture

An overview of the model architecture is shown in Figure 1. The model consists of three main components: (1) A neural shape processing hierarchy that recognizes the moving effector (e.g., the hand) and goal objects, (2) a module that integrates the information about the relationship between effector and object, and (3) a module containing neurons that are selective for transitive actions and that establishes view-invariance of recognition.

The first model component follows closely well-known neural models for shape recognition in the ventral stream (Oram and Perrett, 1994; Riesenhuber and Poggio, 1999b; Rolls and Milward, 2000; Cadieu et al., 2007). Its core part is view-specific detectors for shapes. Invariance and feature complexity increase along the hierarchy, where position and scale invariance are achieved by maximum-pooling. Contrasting with the mentioned object recognition models, the shape-selective neurons in our model show only incomplete position-invariance. These neural units have spatially localized receptive fields with a diameter approximately 4° visual angle, corresponding to electrophysiological results from area IT (Op De Beeck and Vogels, 2000; DiCarlo and Maunsell, 2003; Aggelopoulos and Rolls, 2005). This makes it possible to decode the two-dimensional retinal positions of recognized goal objects and effectors from the population activity of such shape detectors.

A second modification compared with standard object recognition models is that the neural detectors for effector shapes, such as hand postures, are selective for the temporal order with which such shapes occur in the stimulus. Such temporal sequence selectivity is compatible with neural data, e.g., from the superior temporal sulcus (Jellema and Perrett, 2003; Vangeneugden et al., 2009; Singer and Sheinberg, 2010), and it can be accounted for by recurrent connections between shape-selective neurons (Giese and Poggio, 2003).

The second model component is substantially extending existing previous architectures and implements a physiologically plausible mechanism for the integration of the information about effector and goal object. This computational function is potentially associated with neurons in parietal cortex, and potentially also in the STS. The central component is a neural representation of the relative positions of effector and goal object, and of the matching between object type and grip [relative position map (RPM)].

The third model component contains neural detectors that are selective for goal-directed action stimuli. This component integrates the information from the previous modules. In addition, this component is critical for accomplishing view invariance of recognition, by pooling of the responses of a number of view-specific modules. The neural detectors in this model component reproduce properties of action-selective neurons in the STS and premotor cortex (e.g., area F5).

Relationship to other models

Many other biologically-relevant computational models for goal-directed action recognition have focused on the role of motor representations (Haruno et al., 2001; Wolpert et al., 2003) and specifically of the mirror neuron system (Oztop and Arbib, 2002; Demiris and Simmons, 2006; Oztop et al., 2006; Kilner et al., 2007). These models assume typically a matching of visual input to internal representations of motor programs that are represented in terms of variables relevant for motor control, such as joint angles. Only very few implementations have presented how such variables could be extracted from real image sequences (Oztop and Arbib, 2002; Metta et al., 2006; Tessitore et al., 2010). In this sense, the model presented here is complementary to approaches that mainly treat the relationship between visual and motor representations (Erlhagen et al., 2006; Kiebel et al., 2008; Bonaiuto and Arbib, 2010; Chersi et al., 2011).

The model presented in this paper represents actions in terms of learned sequences of learned example views of action stimuli. View-independence is accomplished by pooling over the output signals of neural classifiers that are specific for individual views. Such approaches are very common in computer vision (Weinland et al., 2011), proving their computational feasibility. In addition, the representation of three-dimensional structures in terms of learned example views is meanwhile accepted as a fundamental mechanism for the cortical representations of object shape in the ventral stream (Poggio and Edelman, 1990; Oram and Perrett, 1994; Logothetis et al., 1995; Tarr and Bülthoff, 1998; Riesenhu-

ber and Poggio, 1999a). This hypothesis seems consistent with electrophysiological data showing view-dependent and view-independent shape-selective neurons, and an experience-dependent modulation of tuning properties of neurons in area IT (Kobatake et al., 1998; Sigala and Logothetis, 2002; Freedman et al., 2006; Suzuki and Tanaka, 2011). Also, biologically inspired computational and neural models, based on learned example views, have successfully reproduced a variety of properties of the recognition of non-transitive actions, sometimes even reaching benchmark performance compared with computer vision algorithms (Giese and Poggio, 2003; Lange and Lappe, 2006; Jhuang et al., 2007; Prevete et al., 2008; Escobar et al., 2009; Jhuang et al., 2010). However, many action-selective neurons in monkey cortex show a critical dependence of their response properties on the presence of goal objects and their spatial relationship to the moving effector (like the grasping hand) (Perrett et al., 1989; Gallese et al., 1996; Umiltà et al., 2001; Barraclough et al., 2009). These previous models do not account for this property of action-selective neurons, which is likely essential for the decoding of the meaning of observed transitive actions. Our model proposes simple neural circuits that account for these neurophysiological observations, at the same time proposing a neural implementation of a computational step that might be essential for the realization of higher forms of action categorization. The following sections give a more detailed description of the individual components of the model. In parallel, we discuss different experimental results that support the core assumptions of the proposed architecture.

Shape recognition pathway

The recognition of effector and object shapes is accomplished by a hierarchical neural pathway whose structure is compatible with well-known models for visual object recognition (Perrett and Oram, 1993; Riesenhuber and Poggio, 1999b; Mel and Fiser, 2000; Rolls and Milward, 2000). It has been shown in previous work that such hierarchies can support action recognition by the recognition of shape sequences. For example, a body movement can be represented as a temporal sequence of body shapes (Giese and Poggio, 2003; Lange and Lappe, 2006; Prevete et al., 2008). Recent work in computer vision shows that neutrally inspired hierarchical architectures that recognize sequences of body shapes, or optic flow patterns, can be computationally quite powerful, reaching state-of-the-art performance in computer vision (Jhuang et al., 2007; Serre et al., 2007b; Schindler and van Gool, 2008; Escobar et al., 2009).

The shape recognition pathway consists of a hierarchy of layers, where the complexity of the extracted features increases along the pathway. The tuning properties of these detectors are predefined at the lowest hierarchy level (Gabor filters) and learned at higher hierarchy levels. Following previous shape recognition models (Fukushima, 1980; Riesenhuber and Poggio, 1999b), the pathway is organized in terms of layers that correspond functionally to "simple" and "complex cells." Assuming that the stimuli for the simulated experiments were typically foveated, we did not model the modulation of receptive field properties with the eccentricity within the visual field. The simple cells increase feature complexity, while the complex cells pool responses of simple cells of the same type over neighboring spatial positions and scales, resulting in an increase of position and scale invariance along the hierarchy (cf. Rust and diCarlo, 2010). The spatial resolution was down-sampled by a factor of two at each complex cell level. The output nonlinearity of the neural detectors was given by a linear threshold function. This nonlinearity provides a coarse approximation of the output nonlinearity of real cortical neurons (Movshon et al., 1978; Carandini et al., 1997) and results in a suppression of responses of suboptimally stimulated neural detectors. The parameters of the model neurons were, as far as possible, constrained by physiological parameters, partially taking over results from related models in the literature (Serre and Riesenhuber, 2004; Serre et al., 2007b). If no experimental evidence was available, parameter values were optimized for shape recognition performance in a separate cross-validation experiment (see Video stimuli and simulation procedures). Figure 1*A* shows a coarse overview of the shape recognition pathway, where the approximate receptive field sizes of the neural detectors are indicated by the insets below. A more detailed description of the different hierarchy levels is given in the following.

Shape recognition hierarchy. The first hierarchy level that models simple cells in primary visual cortex consists of local orientation detectors that are modeled by quadrature phase pairs of Gabor filters with eight different preferred orientations and seven different spatial scales (Jones and Palmer, 1987). Receptive fields sizes ranged from 0.35° to 0.99°; matching approximately the values observed in electrophysiological experiments (cf. Serre and Riesenhuber, 2004). The output signals of the Gabor filters were rectified and normalized (Heeger, 1993).

From the output signals of the Gabor filters, "complex cell" responses were computed by pooling of the responses of orientation detectors with the same orientation preference and spatial scale using a maximum operation. The spatial receptive fields of these complex cells had diameters between 0.63° to 1.37°, consistent with data from monkey cortex (Schiller et al., 1976; De Valois et al., 1982).

The model neurons at intermediate hierarchy levels extract shape features of intermediate complexity, similar to neurons in area V4 (Gallant et al., 1993; Pasupathy and Connor, 1999). The responses of the simple cells at the intermediate layers were given by Gaussian radial basis functions (RBFs) with divisive lateral inhibition (Heeger, 1993). The responses of detector type κ at hierarchy level l with receptive-field center $\mathbf{x}$ were given by the function:

$$f_k^l(\mathbf{x}, t) = \frac{\pi_\kappa^l N(\mathbf{h}^{l-1}(\mathbf{x}, t) \mid \mathbf{d}_\kappa^l, \lambda_k^l)}{c + \sum_n \pi_n^l N(\mathbf{h}^{l-1}(\mathbf{x}, t) \mid \mathbf{d}_n^l, \lambda_n^l)}, \quad (1)$$

where $N(\mathbf{x}|\mu, \Sigma)$ is the functional form of the multidimensional normal distribution with mean μ and covariance matrix Σ. The vector $\mathbf{h}^{l-1}(\mathbf{x}, t)$ signifies the outputs from a local neighborhood of complex cells at the previous hierarchy layer that feed into the simple cell of type κ with receptive field center $\mathbf{x}$, and the parameters π_κ [illegible] input. The small constant $c > 0$ prevents the d[illegible]ing. Mathematically, the last equation approxi[illegible] model. The centers $\mathbf{d}_k^l$ of the individual Gaus[illegible] unsupervised learning, using k-means clusteri[illegible] tors of complex cells sampled at random positions on the previous layer, computed from a set of training sequences. For each extracted cluster also the covariance matrix λ_k^l was estimated from the training data, and the weights π_k^l were set to values proportional to the size of the cluster.

A fixed number of 200 Gaussian RBFs were learned from training data and shared for all subsequent computations. The responses of the learned feature detectors were pooled over local spatial neighborhoods (diameter 1.59 to 2.35°) using a maximum operation, defining the responses of the corresponding complex cells that were characterized by an increased level of position invariance. The same procedure was replicated to generate a further intermediate layer that extracts even more complex form features (spatial pooling ranges: 2.16 to 3.19°). These two intermediate layers turned out to be sufficient to accomplish robust performance for the simulations presented in this paper. More complex visual tasks, e.g., including massive clutter or substantial variations in size, might require the introduction of additional intermediate hierarchy layers (Serre et al., 2007a; Fidler et al., 2008).

The highest level of the shape recognition hierarchy is formed by neural detectors that are selective for complete views of objects and effectors. These detectors were also modeled as radial basis functions with the functional form $N(\mathbf{x}|\mathbf{d}_\kappa, 0.1 \cdot \mathbf{I})$, where the RBF centers were sampled equidistantly in time from example frame sequences, and where $\mathbf{I}$ indicates the unit matrix. The receptive fields have diameters of approximately 3.9°, covering an area that contains whole object shapes (see insets Fig. 1*A*).

Only a subset of these shape detectors on the highest hierarchy level generalized robustly to novel instances of the same shape class. More robust detectors for the individual shape classes were constructed by learning of linear neural networks that map the response vectors $f_\kappa(\mathbf{x}, t)$ of the shape detectors at position x onto a shape class-specific activation $a_\gamma(\mathbf{x}, t)$. These linear networks were given by the equation

$$a_\gamma(\mathbf{x}, t) = \omega_\gamma^T \mathbf{f}(\mathbf{x}, t), \quad (2)$$

where the weights ω_γ were learned by linear regression with sparsification [Lasso method (Tibshirani, 1994)]. For the recognition of static

AQ: E

add space before and after |

large greek lambda; same as in denominator

add space before f

shapes (i.e., the goal object) this linear network was trained using the actual input vectors $f_\kappa(\mathbf{x}, t)$ and corresponding idealized binary class activities as training data (i.e., $a_\mu(\mathbf{x}, t) = 1$ if the stimulus belonged to pattern class γ and $a_\mu(\mathbf{x}, t) = 0$ otherwise). For the recognition of dynamic shapes the linear network was trained with the actual input vectors and idealized moving output distributions (moving activity peaks), where the details are given in Selectivity for the temporal order of effector shapes, below.

In the spatial continuum limit the function $a_\gamma(\mathbf{x}, t)$ can be interpreted as a two-dimensional activation field with a peak that is located at the retinal position of the recognized shape. Opposed to typical object recognition models, this representation at the highest level of the shape recognition hierarchy in our model is not completely position-invariant and retains coarse position information about the retinal coordinates of the recognized shapes. For each shape the model contains multiple replica of the shape detectors with different preferred positions and highly overlapping receptive fields with a diameter of approximately 3.9°. This representation of shape position is crucial for the subsequent levels of the model that determine the relative locations of effector and goal object (see below).

Selectivity for the temporal order of effector shapes. The recognition of hand actions depends strongly on the temporal order of the occurrence of hand shapes in the visual stimulus. This is immediately apparent if one observes a movie showing a hand action with random temporal order of the frames, or if such a movie is played in reverse temporal order. Reversing temporal order can sometimes even result in the perception of a completely different action (e.g., grasping vs placing).

There are multiple physiologically plausible mechanisms that can account for such temporal sequence selectivity. We used a mechanism that was proposed before in the context of neural action recognition models (Giese and Poggio, 2003). The network mechanism consists of a single network layer with asymmetric lateral connections between neurons that encode individual snapshots from the hand motion sequence. The resulting network dynamics can be described by a neural field (Wilson and Cowan, 1972; Amari, 1977; Ben-Yishai et al., 1997; Giese, 1999; Erlhagen et al., 2006) with an asymmetric interaction kernel (Zhang, 1996; Xie and Giese, 2002). It has been shown that this type of network, if activated by a moving localized input distribution, supports a form-stable output activation distribution that propagates along the network with the same speed as the input. Using a proceeding described in the study by Zhang (1996), the functional form of this traveling pulse was adjusted, by learning of the shape of the lateral interaction kernel, to fit reported average firing rates of body action-selective neurons in the STS (Oram and Perrett, 1996). The moving activity pulse is only a stable solution of the network within a limited range of speeds for the input distribution. If the input pulse moves in the opposite direction along the fields or with inadequate speed the stable solution of the network dynamics breaks down, and the output amplitude of the network is very small (Xie and Giese, 2002). Likewise, the activation of the inputs of the network with random temporal order results in outputs with very small amplitude (Giese and Poggio, 2003). In addition, previous work shows that the lateral connections of such networks can be learned easily by time-dependent Hebbian plasticity (Brody and Hopfield, 2003; Jastorff and Giese, 2004).

The sequence-selective networks that encode the time course of individual hand actions (e.g., closing for grasping and opening for placing) in this model consist of 20 coupled neurons per action type. We signify by $s_\nu(\xi, t)$ the input current of the neuron ξ encoding action ν, where ξ can be interpreted as the position of the neuron in a one-dimensional neural field. The network dynamics is specified by the differential equations:

$$\tau_u \dot{u}_\nu(\xi, t) = -u_\nu(\xi, t) + \sum_{\xi'} w_u(\xi' - \xi)\, g(u_\nu(\xi', t)) + s_\nu(\xi, t) - q_\nu(t) - h_u. \quad (3)$$

In this equation, $g(x)$ is a sigmoid activation function that behaves approximately linear in the relevant input range for $x > 0$ and decays exponentially for $x < 0$ (Zhang, 1996), $h_u = 0.15$ is a constant that specifies the resting activity level of the network, and τ_u (= 20 ms) is the time constant of the neural field. As consequence of the asymmetric lateral interaction kernel $w_u(\xi)$, an active neuron in the field preactivates neurons that encode temporally subsequent hand shapes, while it inhibits the other neurons.

The term $q_\nu(t) > 0$ specifies lateral inhibitory feedback from other neural fields that encode different action patterns. Such inhibition turned out to be critical for accomplishing robust behavior, especially for the discrimination between action patterns with different temporal order of the frames. Defining by $O_\nu(t) = \max_{\xi'} u_\nu(\xi', t)$ the maximum of the output activity of the field encoding pattern ν, the strength of this nonlinear feedback was given by the equation:

$$q_\nu(t) = q_o \max_{\nu' \neq \nu} 1\left(\frac{O_{\nu'}(t)}{O_\nu(t)} - \theta\right), \quad (4)$$

with $1(x) = 1$ for $x > 0$ and zero otherwise and with $\theta = 0.7$. This equation specifies an inhibition of fixed s[illegible] that at least one other field ν' is significa[illegible] ulus as field ν.

decrease space between 1 and (to show functional form 1(x)

The input distribution of the individu[illegible] from the output vector $\mathbf{f}(\mathbf{x}, t)$ of the s[illegible] learning of linear mappings similar to E[illegible] position-selective input vectors $\mathbf{s}_\nu(\mathbf{x}, t) = [s_\nu(1, \mathbf{x}, t), s_\nu(2, \mathbf{x}, t), \ldots]$, discretely sampling the position coordinate ξ of the neural field. These input vectors were approximated by the linear mapping $\mathbf{s}_\nu(\mathbf{x}, t) = \mathbf{\Omega}_\nu \mathbf{f}(\mathbf{x}, t)$, which was trained by pairs of input vectors $\mathbf{f}$ derived from (spatially centered) training patterns and corresponding idealized input peaks of the neural field that were given by Gaussian functions with maximum amplitude 0.2 and a width (variance) $\sigma_s^2 = 4$ that moved with an appropriate speed over the neural field. While in our model these linear mappings were constructed directly by supervised learning, other work shows that it is possible to exploit Hebbian plasticity mechanisms to learn such mappings by association of time varying inputs with stable solutions in dynamic neural networks that represent the time course of actions (Zhang, 1996; Markram et al., 1997; Song et al., 2000). However, such unsupervised learning mechanisms were not the focus of the work presented in this paper.

The input distribution of the neural field defined by Equation (3) was given by the position-specific inputs with the maximum amplitude, effectively realizing a competition between the inputs with different retinal position specificities. Discrete sampling of the function $s_\nu(\xi, \mathbf{x}, t)$ with respect to the variable ξ defines the position-selective input vectors $\mathbf{s}_\nu(\mathbf{x}, t)$. The input of the field encoding hand action type ν was then given by $\mathbf{s}_\nu(t) = \mathbf{s}_\nu(\mathbf{x}^*, t)$ with $\mathbf{x}^* = \mathrm{argmax}_\mathbf{x}\, \mathbf{s}_\nu(\mathbf{x}, t)$. The model assumes thus a complete position-invariance of the encoding of sequences of hand shapes. This assumption has the advantage that it avoids a combinatorial explosion of neurons due to a replication of the competitive set of neural fields for each represented spatial position. However, it seems likely that in the brain this assumed perfect decoupling of a (position invariant) encoding of hand shape sequences, and of the relative position of hand and object is much less strict and potentially not clearly separated. Further quantitative physiological data will be necessary to clarify this point.

Consistent with previous models for the recognition of non-transitive actions (Giese and Poggio, 2003), the highest level of the sequence-selectivity circuit is given by neural detectors that integrate the output signals of the individual neural fields over time. These motion pattern neurons become activated during the occurrence of particular hand actions, but only if the corresponding image frames appear in the correct temporal order and specify an approximately natural speed of the action. However, their activity is strongly reduced if the corresponding hand shapes occur in wrong temporal order or with unnatural speeds. The motion pattern neurons are defined by the differential equation:

$$\tau_m \dot{m}_\nu(t) = -m_\nu(t) + \max_\xi g(u_\nu(\xi, t)) - h_m. \quad (5)$$

The constant h_m determines the resting activity level, and $\tau_m = 40$ ms is the time constant of the leaky integrator dynamics.

We assume further the existence of position-selective motion pattern neurons (Baker et al., 2000; Jellema et al., 2004; Vangeneugden et al., 2009) that integrate the outputs of the motion pattern neurons and of the corresponding position-selective hand shape neurons multiplicatively. These neurons form an activation map that shows an activity peak at the retinal position of the hand only if the hand shapes arise in the right temporal sequence. Mathematically, the activation of the corresponding detectors was defined by the equation

$$m_\nu(\mathbf{x}, t) = m_\nu(t) \cdot (\mathbf{1}^T \mathbf{s}_\nu(\mathbf{x}, t)), \quad (6)$$

where $\mathbf{s}_\nu(\mathbf{x}, t)$ signifies the position-dependent input distribution for hand action type ν, evaluated at receptive field center $\mathbf{x}$ (see above). Again, a variety of mechanisms is suitable for the implementation of the same computational function, some of them assuming a less strict separation of position-selective and sequence-selective neural representations.

In general, it seems possible that local motion features, such as extracted by neurons in the medial temporal area (MT), might also contribute to the recognition of the effector motion. Our present model does not contain a motion pathway that is suitable for the analysis of the complex optic flow patterns that are associated with hand deformations. It seems plausible that such patters are exploited by the visual system, and it remains an interesting experimental question whether this is the case. In the domain of biological motion recognition (of non-transitive actions) from point-light displays a vivid discussion has emerged about the question how form and local motion features are integrated, where recent evidence points to a flexible integration of both cues (Giese and Poggio, 2003; Casile and Giese, 2005; Lange and Lappe, 2006; Vangeneugden et al., 2009; Thurman et al., 2010).

Representation of the hand-object interaction. The recognition of goal-directed actions requires an association of the extracted information about the effector (hand) movement and the shape and position of the goal object. Our model proposes a simple physiologically-inspired mechanism that accounts for this association. Opposed to other models that solve the same problem by an analysis of the three-dimensional structure of object and effector (Oztop and Arbib, 2002; Bonaiuto and Arbib, 2010), a computationally quite challenging problem especially for monocular stimuli, our model shows that action recognition can also be accomplished by view-specific mechanisms without an explicit reconstruction of three-dimensional structure.

As central mechanism for the analysis of the hand-object interaction we postulate a relative position map (RPM), a two-dimensional neural activation map that represents the two-dimensional position of the hand relative to the goal object in an image frame of reference by a localized activity peak (Fig. 1*B*). Exploiting the fact that the shape-selective neurons in our model are tuned to the retinal position of the recognized shapes, this activation map can be computed by a simple feed-forward network from the responses of the detectors on the highest level of the shape recognition hierarchy.

We present in the following a mathematical formulation of this step in the spatial continuum limit, while in the real implementation the network was based on a discretization of the two-dimensional spatial position $\mathbf{x}$ using 1500 model neurons whose receptive field centers were arranged within a rectangular grid. Let $a_\gamma(\mathbf{x}, t)$ signify the activation distribution of the shape recognition neurons for goal shape γ, and $m_\nu(\mathbf{x}, t)$ the activation map that corresponds to the position-selective motion pattern neurons for hand action ν (Eq. 6). Assuming that we have learned relevant combinations of objects and actions, the relative position map representing combinations of hand action ν and goal shape γ was defined by a simple feed-forward network that combines both variables multiplicatively:

$$r_{\nu\gamma}(\mathbf{d}, t) = \int m_\nu(\mathbf{x} - \mathbf{d}, t)^{\alpha_{\text{hand}}} a_\gamma(\mathbf{x}, t)^{1-\alpha_{\text{hand}}} \, d\mathbf{x}. \quad (7)$$

The multiplication corresponds to a generalized weighted geometric mean. The integral implements a summation over the whole two-dimensional visual field. The vector $\mathbf{d}$ signifies the two-dimensional position of the hand relative to the object in the RPM (Fig. 2*A*). The parameter α_{hand} determines in how far the model neurons defining the RPM are selective for the shape of the hand compared with the shape of the object. Differences in the input selectivity of action-selective neurons for effector and object shapes have been reported in the literature (cf. Perrett et al., 1989).

The proposed mechanism is equivalent to a gain field (Zipser and Andersen, 1988; Salinas and Abbott, 1995; Pouget and Sejnowski, 1997), and realizes a coordinate transformation from retinal to object-centered coordinates. Gain fields are an established model for the neural realization of coordinate transformations in parietal cortex and have been also discussed elsewhere in the context of invariant object representations (Deneve and Pouget, 2003; Crowe et al., 2008). Physiological data suggest the existence of goal-centered neural representations in parietal cortex (Fogassi et al., 2005; Bonini et al., 2011) and the STS (Perrett et al., 1989), comparable to the

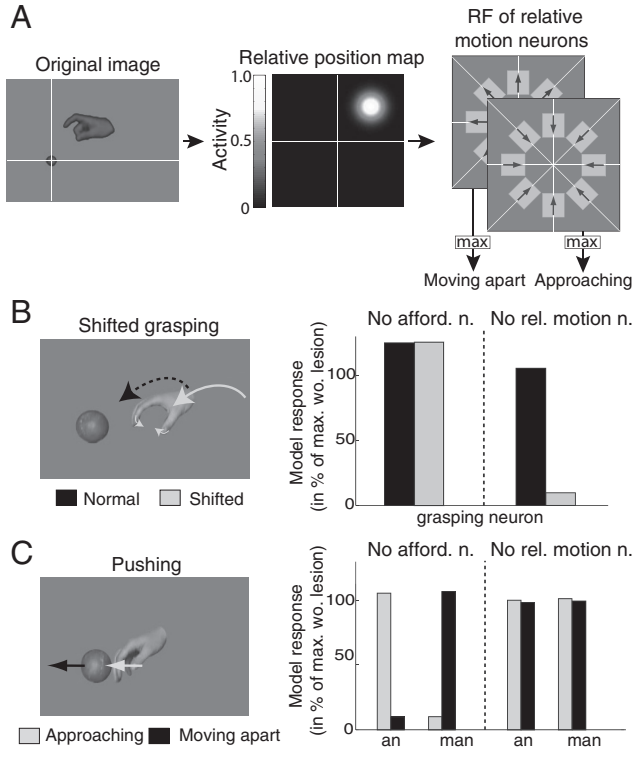

Figure 2. Integration of information about the relative position and motion of effector and object. ***A***, The position of the effector (hand) is encoded by an RPM as a localized activity peak within a two-dimensional coordinate system (white lines) that is centered on the image position of the goal object. Relative motion neurons pool the output responses of motion energy detectors (relative speed neurons) that analyze the motion of the activity peak in the RPM. Different weights for the pooling result in different types of relative motion neurons: Pooling motion responses toward the center of the coordinate system results in a detector that responds if the hand approaches the goal object. Pooling of motion responses away from the center defines a relative motion neuron that the hand moves away from the goal object (moving apart). ***B***, Computational necessity of the affordance neurons: If a normal grasping stimulus (black arrow and bars) is compared with a grasping action for which the hand does not reach the goal object (light gray arrow and bars) a model with only affordance neurons can distinguish the two stimulus types. This is not the case for a model with only relative motion neurons. The plot shows the activity of a neuron trained with normal grasping at the highest level of the recognition hierarchy. ***C***, Computational necessity of the relative motion neurons: A model without relative motion neurons cannot distinguish the two different phases of a pushing stimulus (approaching: hand approaches the stationary goal object (gray), and moving apart: the hand stops and the object moves away from it (black)). Neurons at the highest level of the model were trained to recognize these two events (an: approaching phase; man: moving apart phase). Only the model with relative motion neurons, but not the version with only affordance neurons reliably distinguishes between the two phases of pushing. (Black and gray bars show the activity of neurons at the highest level of the hierarchy, trained with the two events. Activities are normalized relative to the activity for the same stimulus in the normal model with both information channels.)

F2

object-centered representations in the ventral stream (Jellema and Perrett, 2006; Connor et al., 2007). However, some data points also to the existence of effector centered representations (Buneo et al., 2002; Ochiai et al., 2005; Pesaran et al., 2006). We have tested successfully versions of the model with both types of transformations, showing that both relative position representations result in similar computational performance.

The RPM represents a useful neural representation that permits to verify if the spatial relationship between hand and object is compatible with a functional, successful grasping action, exploiting simple neural circuits. In our model we assume neural detectors for two types of features that can be easily computed from the RPM: The first feature is the position of the hand relative to the goal object. We assume the existence of affordance neurons whose receptive fields include all (relative) hand positions that are consistent with successful grips. Their spatial receptive fields were learned from training stimuli, and they were defined mathematically by regions $G_{\nu\gamma}$ that included all hand positions that elicited at least 75% of the maximum activity in the RPM for a given combination of action and object. In addition, we assume in the model that affordance neurons integrate information over time. Their response dynamics is described by the following differential equati

greek epsilon not bold

$$\tau_A \dot{A}_{\nu\gamma}(t) = -A_{\nu\gamma}(t) + \max_{\mathbf{d}\epsilon G_{\nu\gamma}} r_{\nu\gamma}(\mathbf{d}, t), \quad (8)$$

where ν signifies the action, γ the goal shape, and where $\tau_A = 160$ ms is the time constant of the temporal integration. The affordance neurons respond only if the stimulus shows the right combination of hand shape and object shape, combined with their correct spatial arrangement. Neurons with similar selectivity for interactions have been found in the STS and the ventral premotor cortex (Perrett et al., 1989; Gallese et al., 1996).

The second feature that we computed from the RPM is the relative motion of the hand in relationship to the object. This feature turned out to increase substantially the robustness of our recognition results. Local motion at each location of the RPM was computed by simple correlation-based detectors (Adelson and Bergen, 1985; cf. Bayerl and Neumann, 2004), referred to as relative speed neurons in the following. Our model contains 49 detectors (per position) detecting different combinations of horizontal and vertical speed components, covering a speed regime of approximately ± 2.4° per second in both directions. In particular, our model includes detectors for zero relative speed (Bayerl and Neumann, 2004), which were important to detect actions without relative movement between hand and effector (e.g., placing of an object with the hand).

Detectors for meaningful relative motion events in the context of actions were constructed from these local detectors responses by weighted summation, where we assumed four classes of relative motion neurons (detectors for "moving apart," "approaching," and "moving together"). Figure 2*A* shows schematically how the detectors for abstract relative motion events can be constructed from the responses of the relative speed neurons. [Similar circuits have been proposed as models for optic-flow-selective neurons in area MST (Koenderink, 1986; Saito et al., 1986; Zemel and Sejnowski, 1998; Beardsley and Vaina, 2001)].

More precisely, the response of the relative motion neuron for motion type β was obtained by weighting the responses of the relative speed neurons $e_{\nu\gamma}(\phi, \nu, \mathbf{d}, t)$ by a function w^β and pooling them over the positions $\mathbf{d}$ in the RPM and relative motion direction ϕ (using summation), and over relative motion speed ν and position-tuning directions of the relative motion neurons ρ_p (by maximum computation). The resulting sum activity is smoothed over time by a leaky integrator with time constant $\tau_M = 160$ ms, resulting in the equation:

$$\tau_M \dot{M}^\beta_{\nu\gamma}(t) = -M^\beta_{\nu\gamma}(t) + \max_{\nu,\, \rho_p\, \phi,\, \mathbf{d}} \sum w^\beta(\phi, \nu, \rho_p, \mathbf{d})\, e_{\nu\gamma}(\phi, \nu, \mathbf{d}, t). \quad (9)$$

For the approaching and moving apart detectors the weight function $w^\beta(\phi, \nu, \phi_p, \mathbf{d})$ was proportional to the expression

$$\exp\left(-\frac{\mathbf{u}_p\mathbf{u}_\mathbf{d} - 1}{\sigma_\mathbf{d}^2}\right) \exp\left(-\frac{\eta\mathbf{u}_p\mathbf{u}_\phi - 1}{\sigma_\phi^2}\right)(1 - \delta_\nu), \quad (10)$$

(where $\mathbf{u}_d$, $\mathbf{u}_\phi$, and $\mathbf{u}_p$ are unit vectors in the directions of the vector $\mathbf{d}$, the preferred direction ϕ of the relative speed neuron, and a direction that defines the position selectivity of the relative motion neuron. The last term suppresses input signals from relative motion neurons with relative speed $\nu = 0$, δ_y signifying the Kronecker function that is one for $y = 0$ and zero otherwise.) The function w^β specifies direction templates (compare Fig. 2*A*), where $\eta = 1$ specifies a detector for expanding motion and $\eta = -1$ a detector for contracting motion. (Tuning width parameters: $\sigma_\mathrm{d} = 0.25$ and $\sigma_\phi = \pi/2$).

For the moving together detectors the function $w^\beta(\phi, \nu, \phi_p, \mathbf{d})$ was proportional to the term

$$\exp\left(-\frac{\mathbf{u}_p^T\mathbf{u}_\mathbf{u} - 1}{\sigma_\mathbf{d}^2}\right) \exp\left(-\frac{\nu^2}{2\sigma_\nu^2}\right), \quad (11)$$

that specifies the same position selectivity, combined with a speed-dependent term that decays gradually for increasing speeds (with $\sigma_\nu = 1$).

The information of the affordance neurons and the relative motion neurons is finally combined by neural detectors for transitive action that are described in Transitive action-selective neurons and view-independence.

The position- and shape-based information processed by the affordance neurons, and the relative motion information encoded by the relative motion neurons provide two separate channels that represent critical information about goal-directed action stimuli. Which of these features contributes more reliable information depends on the stimulus class.

More detailed simulations show that both pathways are computational beneficial for the processing of natural action stimuli. To demonstrate this we created two additional versions of the model, one which contains only the channel realizing action analysis with affordance neurons, and another one that includes only the channel realizing relative motion analysis. Figure 2*B* shows that a model with only the affordance neuron pathway, opposed to the model with only relative motion analysis, can distinguish successful and non-successful grasping actions, where the hand either correctly touches the object, or where it grasps next to the object ("mimicked action"). Clearly, the distinction of these two action types is critical for the correct recognition of normal grasping. By contrast, the model version with only relative motion processing and no affordance neurons can distinguish different phases during pushing actions, such as the approach of the object by the hand, or the movement of the object after the pushing (Fig. 2*C*). The same distinction is not possible with the model that contains only the processing channel with the affordance neurons. Likewise, we have shown elsewhere that such a model can be used to derive judgments of "perceived causality" from abstract motion displays (Fleischer et al., 2012). This provides a demonstration that both pathways fulfill important computational functions.

In some sense this relevance of form and motion features parallels the integration of form versus local motion features for the recognition of non-transitive body motion, which has been extensively discussed in the field of biological motion processing (Giese and Poggio, 2003; Casile and Giese, 2005; Lange and Lappe, 2006). However, opposed to this discussion the relevant motion here is the relative motion between effector and object, not the local motion in the image.

Transitive action-selective neurons and view-independence. Neurons with selectivity for transitive actions, whose responses are modulated by the exact relationship between the effector movement and goal objects, have been found in multiple regions of the monkey cortex, including the STS (Perrett et al., 1989; Jellema and Perrett, 2006; Barraclough et al., 2009), parietal areas (Fogassi et al., 2005; Rozzi et al., 2008; Bonini et al., 2010), and the premotor cortex (Rizzolatti and Sinigaglia, 2010). One subgroup of these neurons that has recently received particular interest in cognitive neuroscience are the mirror neurons, which also show selective motor tuning during action execution (Di Pellegrino et al., 1992; Gallese et al., 1996; Umiltà et al., 2001; Bonini et al., 2010; Caggiano et al., 2009; Kraskov et al., 2009). Functional imaging studies have suggested that

action-selective regions exist also in human cortex (Iacoboni et al., 1999, 2005; Buccino et al., 2004; Chong et al., 2008; Kilner et al., 2009). While fMRI adaptation studies have revealed partially inconclusive results about the presence of mirror neurons in human cortex (Chong et al., 2008; Dinstein et al., 2008; Lingnau et al., 2009), single-cell recordings in humans demonstrate the existence of action-selective and mirror neurons in various areas in the human cortex, including the supplementary motor area (SMA) (Mukamel et al., 2010). Detailed fMRI studies on action recognition that compare human and monkey cortex suggest a partial homology between the relevant areas in both species (Buccino et al., 2004; Nelissen et al., 2005, 2006; Jastorff et al., 2011).

In our model such neurons are modeled by detectors for transitive actions that integrate the information from the previous processing levels. We assume that this integration is first accomplished in a view-specific manner, and finally view-invariance is accomplished by pooling at the highest level of the model. The second-highest layer of our model hierarchy is formed by view-dependent transitive action neurons that integrate the responses from the affordance neurons and the relative motion neurons in a multiplicative way. We assume a multiplicative integration according to the equation (where we assume that the relative motion type β is chosen in accordance with the recognized action-object combination, so that this index can be dropped in the output variables):

$$T_{\nu\gamma}(t) = A_{\nu\gamma}(t)^{\frac{1}{2}} \cdot M^{\beta}_{\nu\gamma}(t)^{\frac{1}{2}}. \quad (12)$$

The whole architecture described up to this level is based on learned example views of shapes. Correspondingly, the activity of the transitive action neurons is selective for the view from which a particular action has been observed during the training of the system. Many classical theories have assumed that visual parameters, such as the view, are not relevant on cortical processing levels that represent the relationship between effectors and objects for action. Recent electrophysiological data, however, shows that the view angle of observed actions has a strong influence on the responses of the majority of the tested mirror neurons in premotor cortex (area F5) while only a minority is view-invariant (Caggiano et al., 2011). This strongly suggests that view parameters are cortically represented even in premotor cortex and by neurons that have well-defined motor properties. Consistent with this physiological result, our model assumes an organization in terms of view-based modules whose outputs are integrated only at the highest level of the processing hierarchy (compare Fig. 1C). The responses of the view-independent transitive action detectors are obtained by pooling (again by maximum computation) of the outputs of the view-dependent action detectors whose output signals are given by Equation 12. The view-independent transitive action detectors show responses to transitive actions independent of the point of view. Properties consistent with the transitive action detectors in the model have been observed in neurons in the STS and area F5 in macaque cortex (Perrett et al., 1989; Jellema and Perrett, 2006; Caggiano et al., 2011).

AQ: F

Video stimuli and simulation procedures

Datasets. For the evaluation of the model we recorded sets of video stimuli showing a hand grasping objects. Videos were recorded using a CANON XL1-S camera with a frame rate of 25 Hz. A subset of these stimuli was also used in physiological experiments with monkeys, partially testing hypotheses derived from the proposed model (Caggiano et al., 2011). All video frames were converted to gray-scale and preprocessed by removing low-intensity background noise using intensity thresholds. Typical example frames are shown in Figure 1.

The first data set (dataset A) consisted of 270 videos with a resolution of 360 × 176 pixels, depicting side views of grasps (view direction 90° relative to the facing direction of the actor, all actions being executed by the same actor). Videos showed a hand grasping balls with different diameters (4, 8, and 12 cm) with either a power or a precision grip. The stimulus set was derived from 50 original movies by video manipulation, where the original videos included power grips of large and middle-sized balls and precision grips of all tested ball sizes. For the original movies the hand started at a resting position 30 cm in front of the object on the table and moved naturally, grasping the object. The manipulated videos were generated by color segmentation of the hand, the object and the background. The manipulated set included movies showing only the hand, or only the object. Another set of movies showed spatially shifted versions of the action scenes (9 different positions displaced by maximum ± 4°, again for precision and power grip). Testing was based on tenfold cross-validation using a leave-one-out strategy: always, the data from nine repeated conditions was used for the fitting of the model parameters, and the remaining additional trial was used for validation. Data was averaged over all possible 10 partitions of the data in training and test set. (Repetition refers to an independent execution of the same action by the same actor).

The second set (dataset B) contained 150 videos (resolution 405 × 364 pixels), showing different views of power grips, performed either from the top or to the side of a cylindrical goal object (height 10 cm, diameter 4 cm). This action was recorded from 19 different view angles, differing by ~10°, and the grips being executed by the same actor. This angle set included specifically the first person perspective (0°) and the opposite view ("third person perspective"; corresponding to 180°). Each grip was repeated three times. An additional data set contained examples of the same action shown with three view angles (0, 90, and 180°) by two additional actors, again with three repetitions. Evaluation was based on leave-one-out cross-validation over the repeated trials.

A third dataset (dataset C), created by video edition, was a subset of the videos from dataset A. These data set contained videos showing grasping and placing actions, similar to the stimuli used in the studies by Barraclough et al. (2009) and Nelissen et al. (2005). In these movies, the hand entered the scene, grasped a small ball with a precision grip, and moved out (grasping). A second set of sequences was generated by reversing the order of the frames of the original videos, so that the hand entered the scene with the ball and left it after releasing the ball (placing). Additional control stimuli showed only the hand (pantomimed action) and only the object. Additional views for the test of view dependence were generated by mirror-reflecting the grasping and placing stimuli along the vertical axis, resulting in movies showing the opposite hand interacting with the object from the opposite side (cf. Barraclough et al., 2009). This data set was based on nine repetitions of each condition, and cross-validation was based on training of the relevant model parameters with the data from eight repetitions, testing on the remaining one, and appropriate averaging over all partitions in training and test set.

Learning of the model parameters and simulation procedures. The parameters of the model were learned from a set of 137 training stimuli, and generalization to novel stimuli was tested on at least nine independent cross-validation runs. The training set consisted of 85 sequences from dataset A and 52 sequences from dataset B, including different view angles that differed by 30°. The remaining sequences and in particular all sequences from dataset C were used for testing. From each training sequence we extracted images containing only the hand or only the object using color segmentation, as well as frames with typical hand-object interactions, presented in the center of the images. These data was specifically used to estimate the parameters of the model in Equations 1 and 3, and for the learning of the linear mappings according to Equation 2.

The results in the following section are all based on cross-validation data sets that were disjoint from the training stimuli. Model parameters, estimated by the previously described procedure, were identical for all simulations presented in this paper. The model provides thus a unifying quantitative account for the experimental results shown in Results.

To account for the fact that some electrophysiological and fMRI studies present results that average cell classes with different computational properties (Nelissen et al., 2005; Barraclough et al., 2009; Caggiano et al., 2011), we specified two additional parameters in the simulation of those results that account for the fractions of the different cell populations in these studies. The first parameter α_{trans} determines the contributions of neurons that are selective for transitive (goal-directed) and non-transitive actions to the population activity, which were different in the simulated studies investigating neurons in areas F5 and STS. In the relevant simulations the population activity including both types of action-selective neurons, for hand action ν and goal object γ, was modeled by the linear combination:

$$T_{\mathrm{tr/intr},\,\gamma\nu} = \alpha_{\mathrm{trans}}\, T_{\nu\gamma} + (1 - \alpha_{\mathrm{trans}}) m_{\nu}. \quad (13)$$

Likewise, some reported experimental data mixed contributions of neurons with different degrees of selectivity for object shape and hand shape. For the simulation of these data we fitted the parameter α_{hand} from Equation 7 using a least-squares procedure. For the other simulations, the values of the parameters were $\alpha_{\mathrm{trans}} = 1$ and $\alpha_{\mathrm{hand}} = 0.5$.

The parameters of the model presented in this paper were largely determined by supervised learning from labeled example patterns. A biologically more plausible theory would require the acquisition of the relevant patterns by unsupervised or partially supervised learning. Unsupervised learning of hierarchies in object recognition, and recently also in action recognition, is an actual topic in computer vision and machine learning. A variety of approaches has successfully realized learning of recognition hierarchies by applying sparseness constraints to multi-layer architectures, such as convolutional hierarchies (Kavukcuoglu et al., 2010) or compositional representations (Fidler et al., 2009). Unsupervised learning of spatiotemporal feature hierarchies has also been realized by independent subspace analysis (Le et al., 2011) and slow feature analysis (Nater et al., 2011). Another dominant approach has been the learning of deep architectures (Hinton, 2007; Bengio, 2009), e.g., using deep belief nets with successful application to the recognition and modeling of human gait trajectories (Taylor et al., 2010). Another approach for the learning of hierarchical dynamical models that has been applied for the modeling and recognition of actions and other complex spatiotemporal patterns uses recurrent neural networks as generative models in combination with a variational Bayesian framework (dynamic expectation maximization) (Yildiz and Kiebel, 2011; Bitzer and Kiebel, 2012). For many of these approaches it is largely unclear how they relate to plastic mechanisms in real neural systems. However, it has been discussed that unsupervised learning algorithms, such as PCA, independent component analysis (ICA), or sparse learning, might be realized in physiologically plausible ways by combining Hebbian and anti-Hebbian learning rules with intrinsic adaptation mechanisms within individual neurons (Földiák, 1990; Falconbridge et al., 2006; Gerhard et al., 2009).

AQ: G *Special simulation procedures for individual experiments.* For the simulation of the experimental data by Nelissen et al. (2005), we evaluated the response of the model to grasping stimuli and the corresponding control stimuli from dataset C. Following the experimental study, we also used static control stimuli, each presenting one single frame extracted from the middle of each test sequence (no hand-object contact). We approximated the changes of the BOLD signal in the relevant areas by summation of the activity of the model neurons on the corresponding hierarchy level over the whole stimulus duration. For comparison with the experimental data from the action-selective area F5a we used an average activity value that weighted contributions from neurons that are selective for

AQ: H transitive and for intransitive actions (compare Eq. 13).

To simulate the results of the study by Barraclough et al. (2009), we tested the model on the complete dataset C. The processing of the data follows closely the description in Barraclough et al. (2009). The sequence length was down-sampled to 800 ms (20 frames) to match the stimulus conditions in the study. We evaluated only model neurons showing a strong response to the action stimuli used in the experimental study, i.e., action detectors encoding precision grips of small objects. Responses of the model neurons were aligned with the time course of the stimulus following the alignment procedures described in Barraclough et al. (2009). For comparison with the experimental data the response of each model neuron was renormalized, setting the maximum response over all test sequences to one and the baseline activity to zero.

For comparison of the model's performance with the electrophysiological data reported by Perrett et al. (1989), we used the video stimuli from dataset A that showed power grips of a medium sized ball. We created additional similar control stimuli, following the physiological study, from the original videos using color segmentation (hand pantomiming the action, presentation of only the object, hand mimicking the action at a distance of 4 cm from the object). To model the results of the experiment, we evaluated only the responses of transitive action-selective model neurons that responded to power grip actions without motion of the hand relative to the object (zero relative speed).

Results

Neurons with visual selectivity for goal-directed actions have been described in the superior temporal sulcus (Perrett et al., 1989; Barraclough et al., 2009), the parietal cortex (Fogassi et al., 2005; Rozzi et al., 2008; Bonini et al., 2010, 2011), and in premotor cortex, especially in ventral area F5 (Di Pellegrino et al., 1992; Gallese et al., 1996; Umiltà et al., 2001; Caggiano et al., 2009), as well as in dorsal premotor cortex and even in area M1 (Tkach et al., 2007). It seems likely that computational levels of processing do not exactly map onto individual areas in cortex. Instead, it appears that neurons with quite similar computational properties sometimes exist in multiple cortical areas. For example, it has been described that neurons in the STS parallel a lot of properties of mirror neurons in area F5 (Keysers and Perrett, 2004). We thus define here neuron classes according to their functional properties, being aware that the same class of neurons might simultaneously exist at multiple levels of the cortical processing hierarchy, e.g., in the STS and in area F5. A well-established dissociation between STS and F5 is that the superior temporal sulcus does not contain motor neurons. Motor properties are not captured by the proposed model, which focuses on the visual processing mechanisms.

The proposed model provides a unifying account for a variety of visual properties of action-selective neurons that have been reported in single-cell recordings in the superior temporal sulcus and area F5 in macaques. We focus in the following on effects that highlight important computational properties.

Tuning for action type and critical relevance of the goal object

Many action-selective neurons in premotor cortex are selective for the type of the observed goal-directed action (e.g., precision vs power grip). This is illustrated in Figure 3 that shows data from F3 the study by Gallese et al. (1996) from mirror neurons in area F5 of the premotor cortex of macaque monkeys. These neurons show selective motor responses during the execution of hand actions (such as grasping, placing, or object manipulation) and, at the same time, they respond selectively to visually observed actions of other agents (monkeys or humans). Due to the simultaneous presence of visual and motor selectivity, these neurons have been termed mirror neurons. Approximately half of the mirror neurons in this area that were selective for grasping showed also selectivity for the grip type (precision vs power grip), as illustrated in the left and middle panel of Figure 3*A*. The neuron responds strongly to a precision grip and fails to respond almost completely to a power grip. The rightmost panel shows the response for a mimicked action without a goal object. While the hand performed the same movement the neuron remained silent. Such selectivity for the presence of a goal object is typical for many action-selective neurons in premotor cortex (Gallese et al., 1996).

The recognition of action type from real video stimuli is a challenging vision problem since the grip type depends on subtle variations of the finger position, corresponding to changes of only a few pixels in the images. In addition, varying object shapes cause clutter and occlusion for the recognition of the finger configuration. Despite these computational challenges the proposed model accomplishes this recognition task, reproducing the action-type selectivity of cortical neurons. This is shown in Figure 3*B* that shows the response of the view-independent transitive action detectors at the highest level of the model that have been trained with different types of grips of goal objects with different sizes (dataset A, see Materials and Methods). The different linestyles and colors indicate different types of grips. The gray thin

rich3/zns-neusci/zns-neusci/zns01513/zns3601p13z | xppws | S=3 | 3/26/13 | 6:10 | Art: 4129-12 | Input-MH

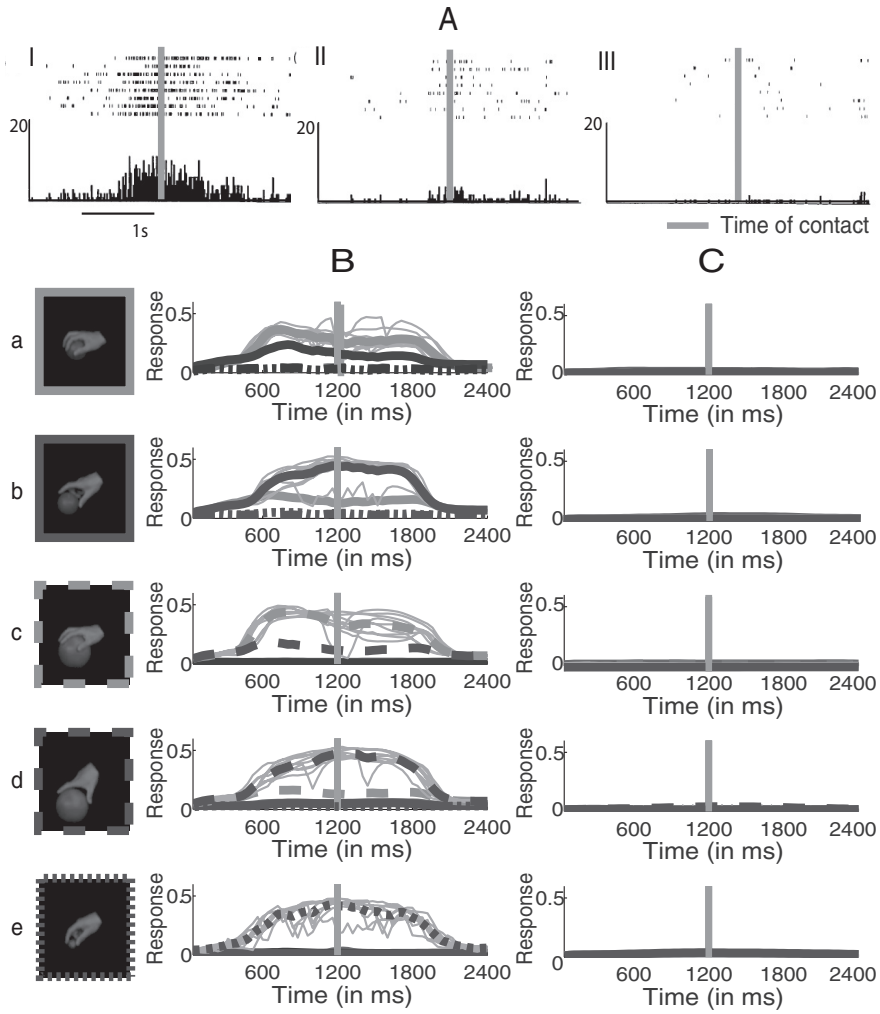

Figure 3. Tuning of real and model neurons for transitive actions. ***A***, Example response of a cortical neuron from area F5 in premotor cortex for (I) a precision grip, (II) a power grip, and (III) to a pantomimed precision grip in the absence of a goal object [from the study of Gallese et al. (1996)]. The vertical bars indicate the contact time of effector and object. ***B***, Response of the view-independent transitive-action detectors of the model for five different grip types with goal objects of different size (***a–e***) as a function of time for five model neurons that were trained with different grasping actions. Thick lines indicate the average responses over multiple trials of the same type, where the different colors and line types indicate different action stimuli. Thin gray lines indicate responses for different realizations of the optimal action stimulus for the individual detectors, which were contained in our test data set. ***C***, Response of model neurons for the same actions performed in the absence of the goal object ("pantomimed actions"). (Conventions as in ***B***).

curves indicate the neural activity for individual trials of the preferred action. The thick curves indicate the average activity over trials for different action types. The action-selective model neurons show a robust selectivity for the different grip types, and the model is able to discriminate robustly precision and power-grip stimuli. The time course of the activity is similar to the neural data, showing a weak initial response that increases when the hand approaches the goal object and when grip-specific hand shapes become distinguishable. Consistent with the neural data, the action-selective model neurons respond only weakly if the same stimuli are presented without goal object (Fig. 3*C*).

Similar selectivity for action type and the presence of the goal object has also been observed in monkey fMRI experiments (Nelissen et al., 2005). In general, the relationship between neural activity at the single-cell level (spikes and local field potentials) and the BOLD signal measured in fMRI experiments is quite involved and dependent on the specific brain area (for review, see Logothetis, 2002, 2008; Logothetis and Wandell, 2004; Nir et al., 2007; Ekstrom, 2010). However, some studies have successfully linked functional imaging results and behavior of groups of neurons at the level of single cells in higher visual areas (Op de Beeck et al., 2008) (for review, see Tsao and Livingstone, 2008). Here we make the strongly simplifying assumption that fMRI responses in the relevant areas might be modeled by the sum activity of the corresponding neural levels of the model. Consistent with the single cell data, visual selectivity for transitive action stimuli was found in area F5 of the premotor cortex. While selective activation during the observation of transitive actions in the caudal part of the premotor cortex (area F5c) was found only for stimuli showing the whole upper body of the acting agent, more anterior regions (areas F5a,p) were also selectively activated by stimuli showing only the hand and the goal object. Since our model focuses on the recognition of hand actions we modeled the activity in these subregions.

Figure 4*A* shows the BOLD activity relative to fixation baseline from two separate fMRI experiments. The first experiment contrasted the full action stimulus, a static picture of the action from the middle part of the stimulus sequence, and movies showing only the moving hand or the static object. High activation emerges only for the full stimulus. Substantially reduced activation was observed for moving and static object stimuli. For the static hand-alone stimuli almost no activation was observed at all. The second experiment contrasted dynamic and static stimuli, and stimuli showing the normal action with correct contact between hand and object and mimicked actions, where the hand executed the same movement in absence of a goal object. Compared with the normal action dynamic mimicked action stimuli induced a reduced response. The static stimuli (derived from normal and mimicked action movies) induced almost no response.

For the simulation of the BOLD responses in this study, we computed the sum activity over all neurons in the two highest levels of the recognition model (view-dependent and the view-invariant transitive-action detectors). For the simulations with the model we used a visual stimulus set that closely matched the properties of the stimuli in Nelissen et al. (2005) (see Materials and Methods, stimulus set C). Since the relevant premotor areas contain a mixture of neurons with selectivity for transitive and intransitive actions, we optimized the parameter α_{trans} that determines the relative influence of neurons with selectivity for transitive and non-transitive actions on the sum signal (choosing $\alpha_{trans} = 1/3$). Likewise, the parameter α_{hand} that determines the relative contributions of hand- and object-selective shape detectors to the activity of the neurons in the RPM was chosen to be $\alpha_{hand} = 4/5$ since this resulted in the best fit of the BOLD data. The sum activities derived from the model were normalized and

rich3/zns-neusci/zns-neusci/zns01513/zns3601p13z | xppws | S=3 | 3/26/13 | 6:10 | Art: 4129-12 | Input-MH

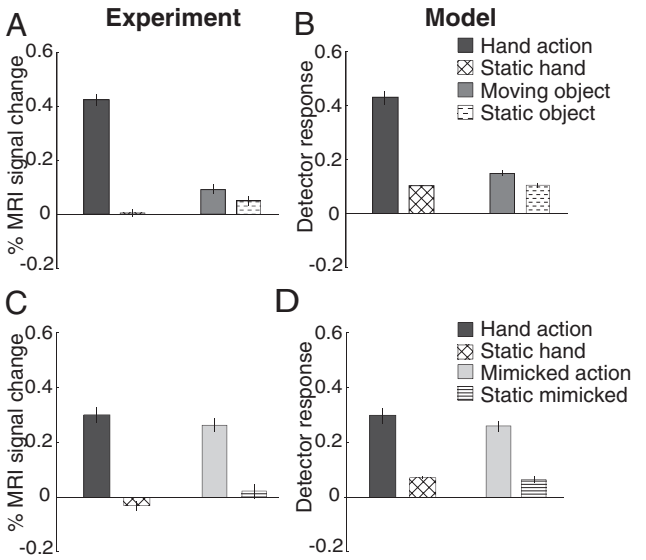

Figure 4. BOLD responses in area F5 measured by monkey fMRI. ***A***, Results are from the study by Nelissen et al. (2005). The experiment compared responses for normal actions and reduced stimuli that were either static, or showed the goal object only (moving or static). The figure shows the BOLD activity against fixation baseline. ***B***, Corresponding simulation results showing the normalized sum activation at the highest level of the model (view-dependent and view-invariant transient-action detectors). ***C***, Results from another experiment in this paper that compared the activity for normal actions and mimicked actions, again in comparison with static stimuli. ***D***, Corresponding simulation results. Error bars indicate SEs.

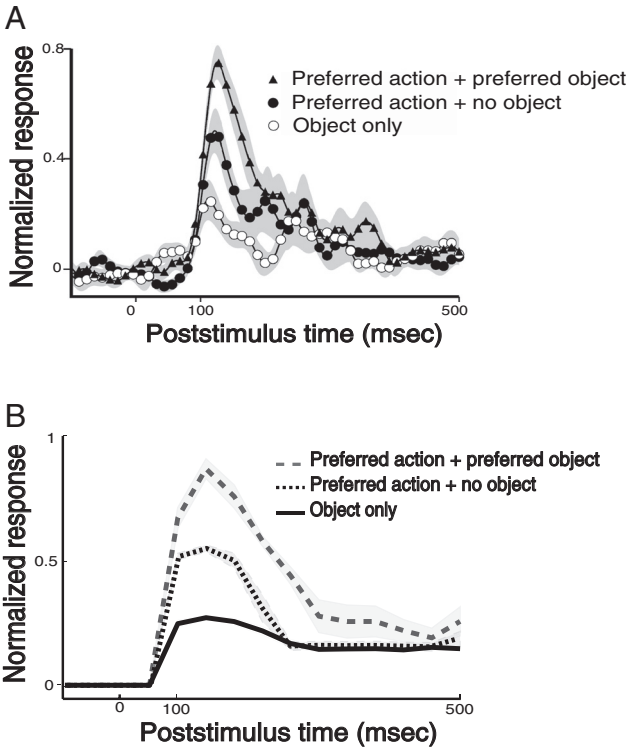

Figure 5. Tuning of action-selective neurons in the STS. ***A***, Average response of STS neurons from a study of Barraclough et al. (2009) that compares the responses of grasping and placing stimuli with control stimuli showing only the hand or only the goal object. Gray region indicates standard error. (Figure reproduced with permission of the authors and MIT Press from Journal of Cognitive Neuroscience.) ***B***, Corresponding average activity predicted from action-selective neurons in the model. Shaded areas indicate standard errors over nine independent simulations.

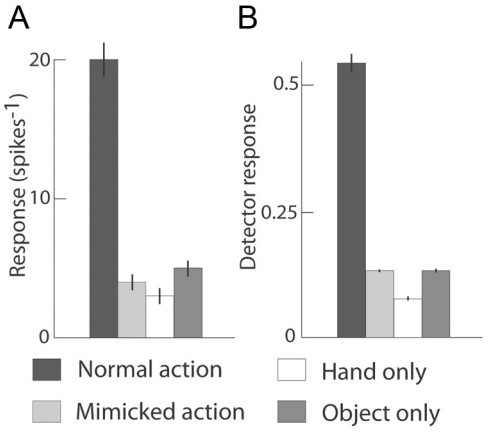

Figure 6. Selectivity for the correct relationship between effector and object. ***A***, Average cell response of a hand interacting with an object from the study by Perrett et al. (1989). The presented action was either natural, pantomimed (only the hand was visible), or mimicked, the hand reaching next to the object. In addition, the static object was presented alone. Error bars indicate SE. ***B***, Corresponding simulation results for a hand grasping a ball with a power grip, and corresponding control stimuli. Error bars indicate SE over 10 independent simulations.

scaled with a constant factor, to simplify comparison with the experimental data.

The resulting normalized sum activities, shown in Figure 4, *B* and *D*, are qualitatively highly similar to the experimentally observed BOLD activities (*A* and *C*). As for the real fMRI data, the activations for static stimuli are strongly reduced. Mimicked actions induce a substantial response, which however does not reach the level of the normal actions. Presentation of the object alone induces a weak response that is bigger for moving objects. The model reproduces thus, at least qualitatively, a variety of effects that have been observed in this fMRI experiment in monkeys.

Neurons that are selective for the visual observation of transitive actions have not only been found in the premotor cortex, but also at lower cortical levels, such as the STS. The STS, through parietal areas, projects to area F5 in the premotor cortex (Seltzer and Pandya, 1978; Matelli et al., 1986; Keysers and Perrett, 2004). The action-selective neurons in this area show a number of properties that resemble closely those of the neurons on area F5 (Perrett et al., 1989; Barraclough et al., 2009). We tried to reproduce data from a study by Barraclough et al. that compared the responses of single cells in the STS for grasping and placing with (transitive), and without goal object (non-transitive actions), and to stimuli presenting the goal object alone.

F5

Figure 5*A* shows the original data from the study by Barraclough et al. (2009), where neural responses (spike density functions) were temporally aligned by the response latencies for the individual stimuli and are displayed with a default latency of 100 ms. Normal transitive actions induced, on average, substantially higher activity in the recorded hand action-selective neurons than stimuli showing the hand action without a goal object.

However, also actions without goal object (intransitive) induced significant activity. Stimuli showing the goal object alone resulted in rather small activity, and clearly activity below the level that is induced by stimuli presenting only the moving hand.

The corresponding simulation results are shown in Figure 5*B* and provide a good qualitative match of the experimental data. In the experimental study, neurons with selectivity for transitive and non-transitive actions had not been distinguished. For the simu-

rich3/zns-neusci/zns-neusci/zns01513/zns3601p13z | xppws | S=3 | 3/26/13 | 6:10 | Art: 4129-12 | Input-MH

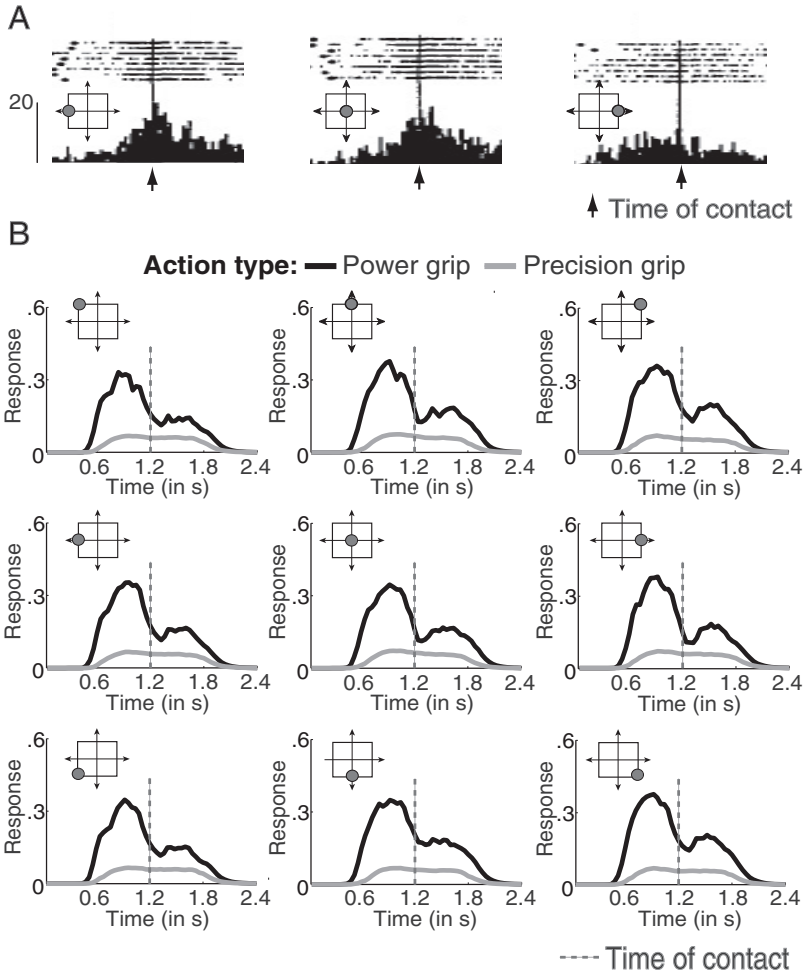

Figure 7. Position invariance of neurons with visual selectivity for transitive actions. ***A***, Response of an example neuron from area F5 from the study Gallese et al. (1996) to similar grasping actions, performed at different positions of the stimuli within the visual field (indicated by the gray discs in the insets). ***B***, Corresponding simulations showing the response of view-independent transitive action detectors with selectivity for a power grip for visual stimuli (power grips and precision grips) with different retinal positions. (The dashed lines indicate the time of hand-object contact).

lation of this STS data the sum activity was a weighted sum of the motion pattern neurons, which also respond to non-transitive actions, and of the transitive action detectors. The parameter α_{trans} that controls the contribution of these two detector populations on the sum activity was set to 2/3 since this matched the approximate ratio of non-transitive and transitive action-selective neurons in the electrophysiological study. In addition, for the simulation of this STS data we chose the value $\alpha_{\text{hand}} = 3/4$ for the parameter that determines the strength of the influences of object and hand shape on the RPM activity.

A similar result was obtained in a classical study on visual neurons with selectivity for hand actions in the lower bank of the STS (area TEa) (Perrett et al., 1989). This study not only tested normal transitive action stimuli and stimuli showing only hand or the object. It also included a condition showing mimicked actions, where the hand did not touch the object correctly, while it was moving in a very similar way as in the normal stimulus. Like before, the responses to stimuli showing only the hand or the object were strongly reduced (Fig. 6). The same applies to the condition with the mimicked stimuli, where the hand failed to touch the object, the distance between hand and object in the image frames being <0.55°. This implies a high spatial selectivity of the underlying neural mechanisms that detect the hand-object contact.

The model nicely reproduces this high selectivity for the relationship between effector and goal object (Fig. 6*B*). This selectivity is a consequence of the receptive field properties of the affordance neurons, which are selective of the retinal effector position relative to the object (Fig. 2). The fact that for this experiment the activity of the action-selective neurons for stimuli without goal object is lower than for the simulation results in Figure 5 is a consequence of the different fractions of transitive action-selective neurons included in the population averages, which we tried to match with the experimental data (see Materials and Methods).

Summarizing, the model reproduces the high selectivity for different hand action types and the precise tuning for the spatial relationship between effector and object, as observed for action-selective single cells in multiple cortical areas. The high selectivity for the action type is explained by the shape-selectivity of the hand detectors in the shape recognition pathway. The high selectivity for the relationship between effector and object is a consequence of the tuning properties of the affordance neurons, whose response depends on the relative positions of effector and object.

Position invariance

Despite the high selectivity of cortical action-selective neurons discussed in the last section, such neurons show a remarkable degree of invariance with respect to the position of transitive action stimuli in the image. This is illustrated in Figure 7*A* that shows the response of a mirror neuron in area F5 to grasping stimuli presented in the left hemifield, the center, or in the right hemifield [adapted from the study by Gallese et al. (1996)]. The red disc illustrates the stimulus position in the visual field. The response of the neuron is largely unaffected by the retinal position of the stimulus. Since this physiological study did not include a control of eye movements it seems likely that a substantial part of the observed invariance is due to the foveation of the stimulus by the monkey. However, substantial amounts of invariance with respect to stimulus position, even with a control of eye position, have been shown for shape-selective neurons in the inferotemporal cortex as well as for shape-selective neurons in the dorsal stream (Op De Beeck and Vogels, 2000; Janssen et al., 2008).

Our model is able to reproduce a high degree of position invariance. This is illustrated in Figure 7*B* that shows the responses of a view-invariant transitive motion detector (selective for power grip) in the model for nine different retinal positions of the action stimulus, where the distance between neighboring stimulus positions corresponds to 4° of visual angle. (The stimulus size was approximately 8°.) Responses for different retinal positions are almost identical, demonstrating almost perfect position invariance. The model is able to accomplish even much

rich3/zns-neusci/zns-neusci/zns01513/zns3601p13z | xppws | S=3 | 3/26/13 | 6:10 | Art: 4129-12 | Input-MH

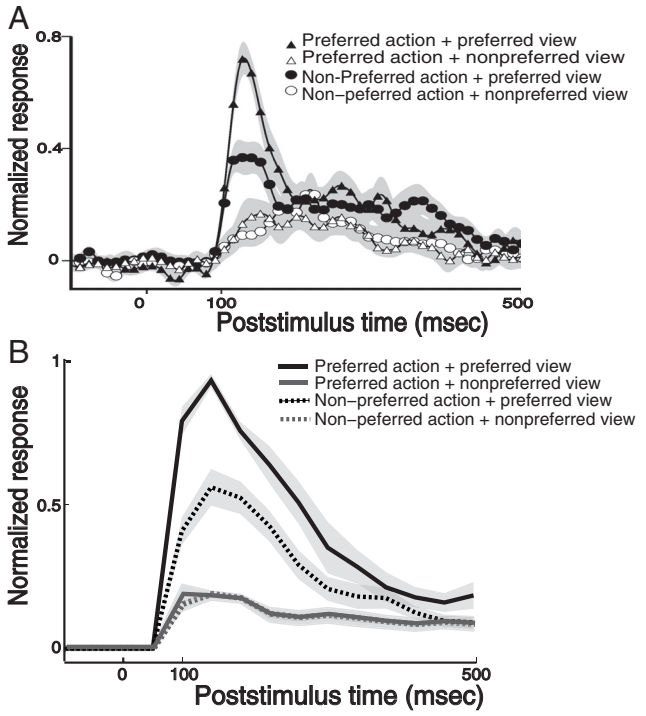

Figure 8. View dependence of neurons in the STS that are selective for transitive actions. ***A***, Average response of 23 neurons from the study Barraclough et al. (2009) (reproduced with permission of the authors and MIT Press from Journal of Cognitive Neuroscience). Two actions (grasping and placing) were shown either from the right from the monkey's point of view, or from the opposite view (rotated by 180° about the vertical). ***B***, Average response of the view-dependent transitive action detectors in the model that was trained with grasping and placing actions using the same action stimuli as in the experiment. Shaded areas indicate standard errors averaged over nine trials.

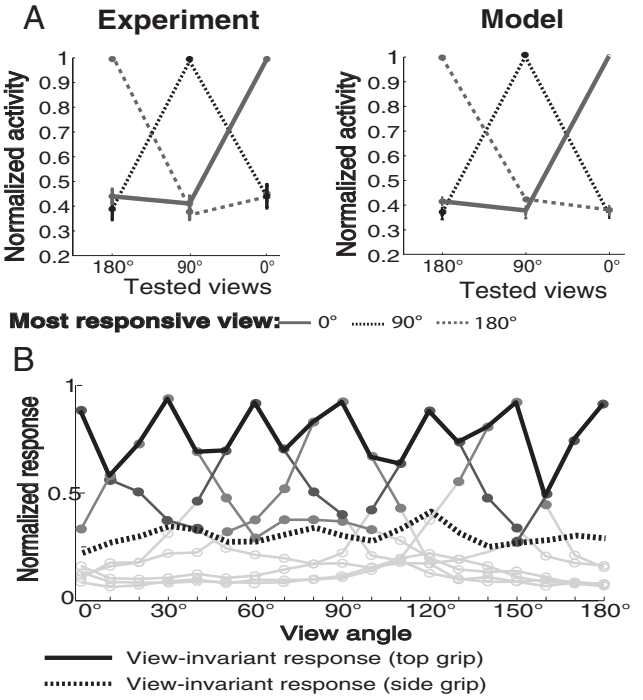

Figure 9. ***A***, View tuning of transitive action neurons in macaque premotor area area F5. Left, Average peak-normalized response of view-selective mirror neurons using the data from Caggiano et al. (2011), who presented the same grasping action from three different perspectives [third-person view (180°)), side view (90°), and first person perspective (0°)]. Right, Average normalized response of the view-dependent transitive-action detectors in the model using the same type of action stimuli. Error bars indicate SEs. ***B***, Realization of view-independence with a small number (7) of view-dependent modules for the distinction of top and side grips of a cylinder. Gray lines indicate the activity of the view-specific transitive action neurons that belong to different view-specific modules. The black lines show the activity of a view-invariant transitive action neuron at the highest hierarchy level that was trained with a top grip, and tested with real videos of top and side grips with 19 different view angles.

larger spatial invariance regimes, and we tested successfully with up to ±30°, as observed in physiological experiments.

In the model, position invariance is accounted for by the combination of two mechanisms: (1) The maximum pooling operations in the form processing pathway, which makes the shape detectors at the highest level of the form recognition pathway partially position invariant (Fukushima, 1980; Riesenhuber and Poggio, 1999b); and (2) the computation of the relative position of the effector and the goal object in the RPM, which explicitly computes a coordinate transformation.

View tuning

View-dependent coding is a well-known property of shape-selective neurons in the inferotemporal cortex (Logothetis et al., 1995; Tarr and Bülthoff, 1998), as well as of shape and action-selective neurons in the STS (Perrett et al., 1982, 1989; Oram and Perrett, 1996). In a recent study, Barraclough et al. (2009) have tested the view dependence of the responses of action-selective neurons. The tested neurons were selective for visually observed grasping and placing actions, and they were tested with views of hands interacting with an object either from the left or from the right side relative to the viewpoint of the monkey. The corresponding average responses, as function of time, are shown in Figure 8*A*. Neurons showed a strong selectivity for the preferred view (black symbols). The presentation of the non-preferred action from the preferred view resulted in higher responses than the presentation of the preferred action with the non-preferred view. View preference, thus, modulated the (average) responses of the neurons more than the type of the action.

Figure 8*B* illustrates the corresponding simulation results obtained with our model. The figure shows the average responses of the view-selective transitive action detectors at the second-highest processing level. Modeling results are qualitatively quite similar to the real data from the STS. Like for the other simulation of the STS data, the parameter we chose $\alpha_{\text{hand}} = 3/4$ for the parameter that determines the influence of hand versus object in the RPM. Since the population of cells underlying this study of STS neurons seemed not to be identical with the one underlying the data shown in Figure 5, we refitted the value of the parameters $\alpha_{\text{trans}} = 0.9$ (fraction of transitive action-selective neurons). However, fitting both simulations from Figures 5*B* and 8*B* with joint identical parameters leads to qualitatively very similar results.

Since our model, such as other biologically-inspired models of form and action recognition (Poggio and Edelman, 1990; Oram and Perrett, 1994; Riesenhuber and Poggio, 1999b; Giese and Poggio, 2003), is organized in terms of view-specific modules, it can reproduce the view-selectivity of action-selective cortical neurons. However, it is not necessarily expected that it also reproduces the fact that the stimulus view has a stronger influence on the tuning of these neurons than the action type. In the model, this behavior is explained by the fact that stimulus views are processed separately up to a very high level of the processing hierarchy, while different actions observed with the same views share many low and mid-level features. The presentation of a non-preferred action thus induces some rudimentary activity in

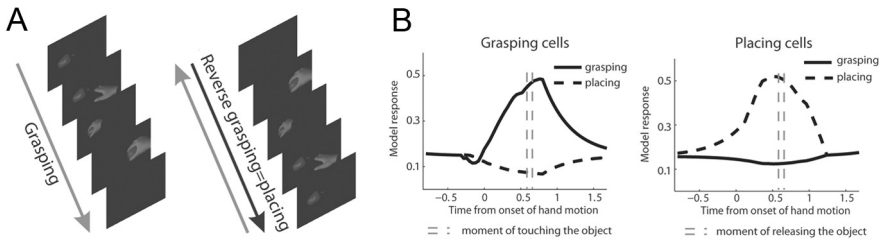

Figure 10. Sequence selectivity of transitive action neurons in the model. ***A***, Video stimuli, showing a hand grasping an object (pepper) and leaving the scene together with it. The same sequence shown in reverse order creates the impression of a "placing" action. ***B***, Average responses of the populations of view-dependent and view-independent transitive action neurons that are selective for grasping and placing (temporally reversed grasping). Responses are highly selective for temporal order (solid vs dashed curves). The two dashed vertical lines indicate the approximate interval during which the hand touched the object for the first time in the grasping stimulus (respectively last touched the object for the placing stimulus).

neurons that encode different actions that are observed with the same view.

View dependence of action-selective neurons has not only been observed in primarily visual structures, such as the STS, but also at higher representation levels. A recent study, which was partly motivated by this model, tested the view dependence of mirror neurons in area F5 of the premotor cortex, exploiting well-controlled video stimuli instead of real actions executed in front of the monkey (Caggiano et al., 2011). Since area F5 is functionally very close to motor cortex, we expected to find a large number of neurons that encode visually observed actions in a body-centered frame of reference independent of the stimulus view. However, presenting the same action from three different views, we found a quite large fraction (74%) of mirror neurons with clear view tuning. Only a smaller fraction (26%) showed view-independent responses. In addition, we failed to find a clear preference for the first person view, as might be expected if the monkey learned particularly well the relationship between own actions and the associated visual feedback signals. The left panel in Figure 9*A* shows the normalized activity of the measured F5 mirror neurons for the three tested views, different line types referring to the subsets of neurons that showed a significant preference for the individual views. The right panel shows the corresponding simulation result (average responses computed with the same normalization procedure as for the electrophysiological data) for the view-dependent transitive action detectors, which form the second-highest hierarchy level of our model. Clearly, the simulation result nicely matches the experimental data from the view-dependent subset of mirror neurons.

Due to the limited recording time, in the physiological experiment the number of stimulus views that could be tested was quite limited. In simulations with the model we could test, however, how many stimulus views are required to accomplish robust view-independent recognition at the highest layer of the model for real video stimuli. This is an important question, since a mechanism that requires a storage of huge numbers of stimulus views would be computationally inefficient or even infeasible.

Quantitative simulations showed that with as few as seven view-specific modules we could accomplish a robust view-independent recognition of goal-directed hand actions from real videos, at the same time achieving high selectivity for the distinction of different action types. This is illustrated in Figure 9*B*, which shows the responses of the view-dependent transitive-action detectors in gray and the resulting response of the corresponding view-invariant detector in black. The model was trained with seven views of one action (grasping a cylinder from the top) and was tested with 19 different views, differing by view angle steps of 10° between the views (dataset B, see Materials and Methods). The tuning width of the view-dependent detectors was approximately 50°, coarsely consistent with data about view-dependent neurons in area STS and IT (Perrett et al., 1991; Logothetis et al., 1995). (Precise data about the view dependence of transitive action-selective neurons is presently still unavailable.) For the trained action the responses of the view-dependent detectors degrade gradually with the distance between the training and the test view. For the distractor action the responses of the view-dependent detectors remain weak for all tested views. The response of the view-independent detectors remains high for all views. Even though their response still varies slightly with the stimulus view, it robustly discriminates for all views between the trained action (solid line) and the distractor action (dashed line).

Prediction: temporal sequence selectivity of transitive action-selective neurons

A central assumption for the proposed mechanism for the recognition of hand actions was the temporal sequence-selectivity of the motion pattern neurons, which form the basis for the association of information about hand postures over time. Reversing the temporal order of the stimulus frames substantially reduces the responses of the motion pattern neurons, and as consequence also the responses of action-selective neurons at higher processing levels. The temporal sequence selectivity of action-selective neurons at lower levels is consistent with recent electrophysiological data from neurons in the STS (Vangeneugden et al., 2009, 2011; Singer and Sheinberg, 2010). The model predicts that sequence selectivity should also be observed at the highest level of the neural processing hierarchy, for neurons that are selective for transitive actions. This prediction can be easily tested in an electrophysiological experiment by showing the same action movie in normal and reverse temporal order.

Figure 10*A* illustrates the relevant stimulus set, two movies showing the frames of a grasping actions in normal and reversed order. Reversely played grasping looks like the placing of an object (the hand coming in and leaving the scene without the object). Following the conventions by Barraclough et al. (2009), we refer to reverse grasping as "placing" in the following. The activation of the transitive action detectors in our model (pooled over view-dependent and view-independent model neurons), after training with grasping respectively placing, for both types of stimuli are shown in Figure 10*B*. Clearly, the transitive action-selective neurons in the model show a very strong degree of temporal sequence selectivity. This selectivity is not only a consequence of the neural field dynamics discussed before. It is further augmented by the fact that stimuli played in reverse temporal order also reverse the relative motion vectors between effector and goal object, which are detected by the relative motion neurons. Both influences are combined multiplicatively by the transitive action detectors of the model.

Motivated by this model prediction, these stimuli were really tested in an electrophysiological experiment, recording the activity of mirror neurons in area F5. Consistent with the prediction, a significant fraction of the measured neurons (63%) showed strong sequence selectivity, and the quantitative results look

AQ: K

strikingly similar to the simulations shown in Figure 10*B* (V. Caggiano, J. Pomper, F. Fleischer, M. A. Giese, and P. Thier, unpublished observations).

Performance limitations of the model

While the proposed model successfully accomplishes recognition on real videos of hand-object interactions, we want to stress here that the main purpose of the work was the reproduction of data from neurons and not the maximization of the computational performance in the sense of computer vision. We acknowledge that many not biologically-inspired algorithms have been developed in this field (for review, see Pavlovic et al., 1997; Mitra and Acharya, 2007; Weinland et al., 2011), which certainly would outperform our model on challenging data sets, which for example include substantial amounts of background clutter. The effective processing of complex scenes with complex clutter likely necessitates improved dictionaries of detectors for the intermediate-level features. The learning of such detector hierarchies has been a core problem of the fields of shape and action recognition in computer vision in the last decade (Moeslund et al., 2006; Serre et al., 2007b), and the principle architecture of our model would not change by inclusion of such improved hierarchies of shape detectors.

Another important major addition that seems necessary for the processing of complex realistic scenes with many objects and potentially multiple acting effectors (e.g., from multiple agents) is attentional control and the tracking of attended objects and effectors. Neural mechanisms supporting such computational functions have been extensively studied in the context of neural models for attention (Deco and Rolls, 2004; Hamker, 2006; Tsotsos, 2011). Such mechanisms could be integrated in our model by adding a network dynamics to all layers of the hierarchy and by introducing appropriate backward connections. In fact, first attempts to integrate such mechanisms in action processing models related to ours have been made (Layher et al., 2012).

Concluding, the present model clearly has strong computational limits, some of which might be mitigated by including other physiologically plausible mechanisms. However, the performance limits for the processing of complex real action scenes using such neural architectures will have to be explored after adding such extensions to the present architecture.

Discussion

In this paper we have presented a physiologically inspired neural model for the visual recognition of transitive hand actions, defined by interactions between a moving hand and a goal object. The model is based largely on well-established neural principles, all of which can be implemented by physiologically plausible circuits. The model provides a unifying account for a variety of physiological results about action-selective neurons at the single-cell level, as well as for results about the population activity in relevant areas in macaque cortex. To our knowledge, this is the first model for the visual recognition of transitive actions that provides such detailed comparisons with neural data.

The proposed model has been shown to be computationally powerful enough to recognize actions from real video sequences. This gives credibility to the computational feasibility of the postulated neural principles as basis of the processing of natural action stimuli. This also distinguishes our action recognition model from many others that assume abstract visual input signals, not specifying exactly how they can be derived from real images by physiologically plausible mechanisms. In addition, this property made it possible to test the model with original stimuli that have been used in physiological experiments. However, the model would need several substantial extensions to deal, for example, with substantial amounts of clutter, or scenes that include multiple possible goal objects or observed effectors. Some possible extensions and performance limitations of the model were discussed in Results, Performance limitations of the model.

Our model not only provides a unifying account for a number of physiological results from action-selective neurons in monkey cortex. It also leads to several important theoretical insights.

First, it shows that the recognition of goal-directed actions and visual tuning properties of action-selective neurons can be accounted for by established mechanisms, which are based on learned view-specific neural representations, and without the necessity of an accurate reconstruction of the three-dimensional structure of the effector and the object. Since the estimation of joint angles, especially from monocular images, is a challenging computer vision problem (for review, see Wu and Huang, 1999; Erol et al., 2005), our model suggests that the brain might bypass this computational step using, at least to a substantial degree, representations that are based on two-dimensional views. In addition, such a solution seems theoretically attractive since it postulates that the brain uses similar neuro-computational principles for the processing of static and dynamic three-dimensional stimuli (compare Materials and Methods, Relationship to other models). In addition, it is at least an interesting observation that the majority of robust algorithms for action detection and classification exploits example-based (view-specific) representations (Gavrila, 1999; Moeslund et al., 2006). The focus on visual processing mechanisms makes our model complementary to many other models for the visual recognition of hand actions that focus on the role of motor representations, making simplifying assumptions about the visual processing (see also Materials and Methods, Relationship to other models).

Second, the model proposes a set of concrete circuits for the integration of the information about objects and dynamic effectors that could be implemented with real cortical neurons. At the same time, the model makes precise predictions about the behavior of such neurons that can be validated by single-unit recordings. Because of space limitations, we discuss here only a few examples can be discussed: (1) The model postulates the existence of neurons that encode the relative position of effector and object (relative position map), and a multiplicative integration of the relevant input signals from shape-selective representations. Neurons with such properties might be found in the superior temporal sulcus (Perrett et al., 1989) or the inferior parietal lobule (Fogassi et al., 2005; Chafee et al., 2007; Crowe et al., 2008; Rozzi et al., 2008). (2) The existence of affordance neurons, e.g., in parietal areas, with spatially organized receptive fields can be tested. (3) The model assumes a hierarchical architecture, where information is first processed in view-specific modules and then integrated by pooling at the highest level of the hierarchy. This predicts specific connections between view-specific and view-invariant action-selective neurons, e.g., in the premotor cortex or in the STS. Recent electrophysiological results proof the existence of view-specific representation very high up in the processing stream, even in premotor cortex (Caggiano et al., 2011). (4) The model postulates neurons that are selective for the relative motion between effector and object (relative speed and motion neurons). Contrasting with regular motion detectors, e.g., in area MT, such neurons process motion in the RPM, and thus they should be characterized by a high degree of shape selectivity. Neurons of this type might be present in the STS or parietal areas.

Many other specific predictions follow from the proposed architecture. Some predictions, such as the view dependence and sequence-selectivity of mirror neurons, have been confirmed by electrophysiological experiments that were partially motivated by this theory (Caggiano et al., 2011). In addition, the model makes also predictions about the population activity in cortical areas that are associated with the different postulated computational modules. Such predictions seem ideally suited for comparisons with fMRI data. Additional simulations addressing such aspects are in progress and might help to develop a more complete theory that links corresponding mechanisms in the brains of human and non-human primates.

Undoubtedly, our model makes a number of very strong simplifications, some of which violate known facts about the modeled cortical structures. In addition, many fundamental aspects about the model have to be refined in future work. Again only a few fundamental limitations can be discussed here: (1) The model focuses purely on the visual processing of actions and lacks completely interactions with motor representations. Especially, it does not account for the motor properties of some action-selective neurons in parietal and premotor cortex, and especially of mirror neurons. A large body of literature suggests, in addition, interactions between visual and motor representations, and the mirror neuron system might play a central role in establishing such interactions (Rizzolatti and Craighero, 2004; Kilner et al., 2007; Schütz-Bosbach and Prinz, 2007). The existence of feedback connections from motor to visual representations (e.g., between premotor areas, area PFG and the STS) is strongly suggested by anatomical data (Rizzolatti and Craighero, 2004; Rizzolatti and Sinigaglia, 2010). An adequate theoretical framework to capture such feedback influences are hierarchies of predictive (neuro-)dynamical representations (Demiris and Simmons, 2006; Kiebel et al., 2008), such as neural fields. It seems straight forward, and has been successfully established in previous work in robotics, to couple such neural field representations for motor programs (Erlhagen and Schöner, 2002; Cisek, 2006) with ones for visual input sequences (Erlhagen et al., 2006). (2) Beyond the top-down connections from motor representations, the visual pathway is characterized by strong feedback connectivity (Felleman and Van Essen, 1991; Salin and Bullier, 1995) that is not captured by our model. In the context of action recognition, such connections might support especially the dynamic tracking of objects and effectors in the scene, and the attentional selection of individual objects in complex or cluttered scenes with multiple possible targets by attentional mechanisms. (See Results, Performance limitations of the model, for further details.) (3) As for previous models for the recognition of non-transitive actions (Giese and Poggio, 2003; Jhuang et al., 2007; Escobar et al., 2009) one might consider a second primary visual pathway that processes local motion and optic flow features instead of form features. In how far form versus motion features influence the visual recognition of goal-directed actions is, to our knowledge, largely unclear, and seems to define an interesting question for future research. (4) A further important shortcoming of the proposed model is the complete lack of disparity features. Many neurons in the dorsal as well as in the ventral stream are disparity-selective (Shikata et al., 1996; Janssen et al., 1999; Taira et al., 2000; Durand et al., 2007; Srivastava et al., 2009; Orban, 2011. Also, some recent evidence shows the existence of disparity-selective neurons in cortical areas that are involved in action processing, such as the premotor area F5 (Joly et al., 2009; Theys et al., 2012). It seems possible to extend the chosen example-based approach by inclusion of disparity-dependent features, such as relative disparity. Similar approaches have been proposed for object recognition from stereo images in computer vision (Helmer and Lowe, 2010). Such extensions might provide interesting insights in the computational role of disparity features in the perception and control of actions, and the internal representation of the geometry of the external space during action execution (La'davas, 2002).

References

Adelson EH, Bergen JR (1985) Spatiotemporal energy models for the perception of motion. J Opt Soc Am A 2:284–299.

Aggelopoulos NC, Rolls ET (2005) Scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. Eur J Neurosci 22:2903–2916.

Amari S (1977) Dynamics of pattern formation in lateral-inhibition type neural fields. Biol Cybern 27:77–87.

Baker CI, Keysers C, Jellema T, Wicker B, Perrett DI (2000) Coding of spatial position in the superior temporal sulcus of the macaque. Curr Psychol Lett 1:71–87.

Barraclough NE, Keith RH, Xiao D, Oram MW, Perrett DI (2009) Visual adaptation to goal-directed hand actions. J Cogn Neurosci 21:1806–1820.

Bayerl P, Neumann H (2004) Disambiguating visual motion through contextual feedback modulation. Neural Comput 16:2041–2066. AQ: L

Beardsley SA, Vaina LM (2001) A laterally interconnected neural architecture in mst accounts for psychophysical discrimination of complex motion patterns. J Comput Neurosci 10:255–280.

Bengio Y (2009) Learning deep architectures for AI. Found Trends Mach Learn 2:1–127.

Ben-Yishai R, Hansel D, Sompolinsky H (1997) Traveling waves and the processing of weakly tuned inputs in a cortical network module. J Comput Neurosci 4:57–77.

Bitzer S, Kiebel SJ (2012) Recognizing recurrent neural networks (rrnn): Bayesian inference for recurrent neural networks. Biol Cybern 106:201–217.

Bonaiuto J, Arbib MA (2010) Extending the mirror neuron system model, ii: what did i just do? a new role for mirror neurons. Biol Cybern 102:341–359.

Bonini L, Rozzi S, Serventi FU, Simone L, Ferrari PF, Fogassi L (2010) Ventral premotor and inferior parietal cortices make distinct contribution to action organization and intention understanding. Cereb Cortex 20:1372–1385. AQ: M

Bonini L, Serventi FU, Simone L, Rozzi S, Ferrari PF, Fogassi L (2011) Grasping neurons of monkey parietal and premotor cortices encode action goals at distinct levels of abstraction during complex action sequences. J Neurosci 31:5876–5886.

Brody CD, Hopfield JJ (2003) Simple networks for spike-timing-based computation, with application to olfactory processing. Neuron 37:843–852.

Buccino G, Lui F, Canessa N, Patteri I, Lagravinese G, Benuzzi F, Porro CA, Rizzolatti G (2004) Neural circuits involved in the recognition of actions performed by nonconspecifics: an fmri study. J Cogn Neurosci 16:114–126.

Buneo CA, Jarvis MR, Batista AP, Andersen RA (2002) Direct visuomotor transformations for reaching. Nature 416:632–636.

Cadieu C, Kouh M, Pasupathy A, Connor CE, Riesenhuber M, Poggio T (2007) A model of v4 shape selectivity and invariance. J Neurophysiol 98:1733–1750.

Caggiano V, Fogassi L, Rizzolatti G, Thier P, Casile A (2009) Mirror neurons differentially encode the peripersonal and extrapersonal space of monkeys. Science 324:403–406.

Caggiano V, Fogassi L, Rizzolatti G, Pomper JK, Thier P, Giese MA, Casile A (2011) View-based encoding of actions in mirror neurons of area f5 in macaque premotor cortex. Curr Biol 21:144–148. AQ: N

Carandini M, Heeger DJ, Movshon JA (1997) Linearity and normalization in simple cells of the macaque primary visual cortex. J Neurosci 17:8621–8644.

Casile A, Giese MA (2005) Critical features for the recognition of biological motion. J Vis 5:348–360.

Chafee MV, Averbeck BB, Crowe DA (2007) Representing spatial relationships in posterior parietal cortex: single neurons code object-referenced position. Cereb Cortex 17:2914–2932.

Chersi F, Ferrari PF, Fogassi L (2011) Neuronal chains for actions in the parietal lobe: a computational model. PLoS ONE 6:e27652.

Chong TT, Cunnington R, Williams MA, Kanwisher N, Mattingley JB (2008) fMRI adaptation reveals mirror neurons in human inferior parietal cortex. Curr Biol 18:1576–1580.

Cisek P (2006) Integrated neural processes for defining potential actions and deciding between them: a computational model. J Neurosci 26:9761–9770.

Connor CE, Brincat SL, Pasupathy A (2007) Transformation of shape information in the ventral pathway. Curr Opin Neurobiol 17:140–147.

Crowe DA, Averbeck BB, Chafee MV (2008) Neural ensemble decoding reveals a correlate of viewer-to object-centered spatial transformation in monkey parietal cortex. J Neurosci 28:5218–5228.

Deco G, Rolls ET (2004) A neurodynamical cortical model of visual attention and invariant object recognition. Vision Res 44:621–642.

Demiris Y, Simmons G (2006) Perceiving the unusual: temporal properties of hierarchical motor representations for action perception. Neural Netw 19:272–284.

Deneve S, Pouget A (2003) Basis functions for object-centered representations. Neuron 37:347–359.

De Valois RL, Albrecht DG, Thorell LG (1982) Spatial frequency selectivity of cells in macaque visual cortex. Vision Res 22:545–559.

DiCarlo JJ, Maunsell JH (2003) Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. J Neurophysiol 89:3264–3278.

Dinstein I, Thomas C, Behrmann M, Heeger DJ (2008) A mirror up to nature. Curr Biol 18:R13–R18.

di Pellegrino G, Fadiga L, Fogassi L, Gallese V, Rizzolatti G (1992) Understanding motor events: a neurophysiological study. Exp Brain Res 91: 176–180.

Durand JB, Nelissen K, Joly O, Wardak C, Todd JT, Norman JF, Janssen P, Vanduffel W, Orban GA (2007) Anterior regions of monkey parietal cortex process visual 3D shape. Neuron 55:493–505.

Ekstrom A (2010) How and when the fMRI BOLD signal relates to underlying neural activity: the danger in dissociation. Brain Res Rev 62:233–244.

Erlhagen W, Schöner G (2002) Dynamic field theory of movement preparation. Psychol Rev 109:545–572.

Erlhagen W, Mukovskiy A, Bicho E (2006) A dynamic model for action understanding and goal-directed imitation. Brain Res 1083:174–188.

Erol A, Bebis G, Nicolescu M, Boyle RD, Twombly X (2005) A review on vision-based full DOF hand motion estimation. Paper presented at IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, June.

AQ: O

Escobar MJ, Masson GS, Vieville T, Kornprobst P (2009) Action recognition using a bio-inspired feed-forward spiking network. Int J Comput Vision 82:284–301.

Falconbridge MS, Stamps RL, Badcock DR (2006) A simple hebbian/anti-hebbian network learns the sparse, independent components of natural images. Neural Comput 18:415–429.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47.

Fidler S, Boben M, Leonardis A (2008) Similarity-based cross-layered hierarchical representation for object categorization Paper presented at IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, June.

Fidler S, Boben M, Leonardis A (2009) Optimization framework for learning a hierarchical shape vocabulary for object class detection Paper presented at British Machine Vision Conference (BMVC), London, UK, September.

Fleischer F, Christensen A, Caggiano V, Thier P, Giese MA (2012) Neural theory for the perception of causal actions. Psychol Res 76:476–493.

AQ: P

Fogassi L, Ferrari PF, Gesierich B, Rozzi S, Chersi F, Rizzolatti G (2005) Parietal lobe: from action organization to intention understanding. Science 308:662–667.

Földiák P (1990) Forming sparse representations by local anti-hebbian learning. Biol Cybern 64:165–170.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2006) Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. Cereb Cortex 16:1631–1644.

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.

Gallant JL, Braun J, Van Essen DC (1993) Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. Science 259:100–103.

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

Gavrila D (1999) The visual analysis of human movement: A survey. Comput Vis Image Underst 73:82–98.

Gerhard F, Savin C, Triesch J (2009) A robust biologically plausible implementation of ica-like learning. Paper presented at the 17th European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, April.

Giese MA (1999) Neural field theory of motion perception. Dordrecht: Kluwer Academic.

Giese MA, Poggio T (2003) Neural mechanisms for the recognition of biological movements. Nat Rev Neurosci 4:179–192.

Hamker FH (2006) Modeling feature-based attention as an active top-down inference process. Biosystems 86:91–99.

Haruno M, Wolpert DM, Kawato M (2001) Mosaic model for sensorimotor learning and control. Neural Comput 13:2201–2220.

Heeger DJ (1993) Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. J Neurophysiol 70:1885–1898.

Helmer S, Lowe DG (2010) Using stereo for object recognition. Paper presented at IEEE International Conference on Robotics and Automation (ICRA), Anchorage, AL, May.

Hinton GE (2007) Learning multiple layers of representation. Trends Cogn Sci 11:428–434.

Iacoboni M, Woods RP, Brass M, Bekkering H, Mazziotta JC, Rizzolatti G (1999) Cortical mechanisms of human imitation. Science 286:2526–2528.

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

Janssen P, Vogels R, Orban GA (1999) Macaque inferior temporal neurons are selective for disparity-defined three-dimensional shapes. Proc Natl Acad Sci U S A 96:8217–8222.

Janssen P, Srivastava S, Ombelet S, Orban GA (2008) Coding of shape and position in macaque lateral intraparietal area. J Neurosci 28:6679–6690.

Jastorff J, Giese M (2004) Time-dependent hebbian learning rules for the learning of templates for visual motion recognition In: Dynamic perception (Ilg U, Bülthoff H, Mallot H, eds), Vol 151–156. Amsterdam: IOS.

AQ: Q

Jastorff J, Clavagnier S, Gergely G, Orban GA (2011) Neural mechanisms of understanding rational actions: Middle temporal gyrus activation by contextual violation. Cereb Cortex 21:318–329.

Jellema T, Perrett DI (2003) Perceptual history influences neural responses to face and body postures. J Cogn Neurosci 15:961–971.

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

Jellema T, Maassen G, Perrett DI (2004) Single cell integration of animate form, motion and location in the superior temporal cortex of the macaque monkey. Cereb Cortex 14:781–790.

Jhuang H, Serre T, Wolf L, Poggio T (2007) A biologically inspired system for action recognition. Paper presented at IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, October.

Jhuang H, Garrote E, Mutch J, Poggio T, Steele A, Serre T (2010) Automated home-cage behavioral phenotyping of mice. Nature Comm 1:1–9.

Joly O, Vanduffel W, Orban GA (2009) The monkey ventral premotor cortex processes 3D shape from disparity. Neuroimage 47:262–272.

Jones JP, Palmer LA (1987) An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58:1233–1258.

Kavukcuoglu K, Sermanet P, Boureau YL, Gregor K, Mathieu M, LeCun Y (2010) Learning convolutional feature hierarchies for visual recognition. In: Advances in neural information processing systems (NIPS) (Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, eds), pp1090–1098. Curran Associates.

Keysers C, Perrett DI (2004) Demystifying social cognition: a hebbian perspective. Trends Cogn Sci 8:501–507.

Kiebel SJ, Daunizeau J, Friston KJ (2008) A hierarchy of time-scales and the brain. PLoS Comput Biol 4:e1000209.

AQ: R

Kilner JM, Friston KJ, Frith CD (2007) Predictive coding: an account of the mirror neuron system. Cogn Process 8:159–166.

Kilner J, Neal A, Weiskopf N, Friston KJ, Frith CD (2009) Evidence of mirror neurons in human inferior frontal gyrus. J Neurosci 29:10153–10159.

Kobatake E, Wang G, Tanaka K (1998) Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. J Neurophysiol 80:324–330.

Koenderink JJ (1986) Optic flow. Vision Res 26:161–179.

Kraskov A, Dancause N, Quallo MM, Shepherd S, Lemon RN (2009) Corticospinal neurons in macaque ventral premotor cortex with mirror properties: a potential mechanism for action suppression? Neuron 64:922–930.

La'davas E (2002) Functional and dynamic properties of visual peripersonal space. Trends Cogn Sci 6:17–22.

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

Layher G, Giese MA, Neumann H (2012) Learning representations for animated motion sequence and implied motion mecognition. Paper presented at IEEE International Conference on Neural Networks (ICANN), Lausanne, Switzerland, September.

Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, June.

Lingnau A, Gesierich B, Caramazza A (2009) Asymmetric fMRI adaptation reveals no evidence for mirror neurons in humans. Proc Natl Acad Sci U S A 106:9925–9930.

Logothetis NK (2002) The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. Philos Trans R Soc Lond B Biol Sci 357:1003–1037.

Logothetis NK (2008) What we can do and what we cannot do with fMRI. Nature 453:869–878.

Logothetis NK, Wandell BA (2004) Interpreting the BOLD signal. Annu Rev Physiol 66:735–769.

Logothetis NK, Pauls J, Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. Curr Biol 5:552–563.

Markram H, Lübke J, Frotscher M, Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. Science 275: 213–215.

Matelli M, Camarda R, Glickstein M, Rizzolatti G (1986) Afferent and efferent projections of the inferior area 6 in the macaque monkey. J Comp Neurol 251:281–298.

Mel BW, Fiser J (2000) Minimizing binding errors using learned conjunctive features. Neural Comput 12:731–762.

Metta G, Sandini G, Natale L, Craighero L, Fadiga L (2006) Understanding
AQ: S mirror neurons: a bio-robotic approach. Epigen Robot 7:197–232.

Mitra S, Acharya T (2007) Gesture recognition: a survey. IEEE Trans Syst Man Cybern C Applicat Rev 37:311–324.

Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis 104:90–126.

Movshon JA, Thompson ID, Tolhurst DJ (1978) Spatial summation in the receptive fields of simple cells in the cat's striate cortex. J Physiol 283:53–77.

Mukamel R, Ekstrom AD, Kaplan J, Iacoboni M, Fried I (2010) Single-neuron responses in humans during execution and observation of actions. Curr Biol 20:750–756.

Nater F, Grabner H, Van Gool L (2011) Temporal relations in videos for unsupervised activity analysis. Paper presented at British Machine Vision Conference (BMVC), Dundee, Scotland, UK, August.

Nelissen K, Luppino G, Vanduffel W, Rizzolatti G, Orban GA (2005) Observing others: multiple action representation in the frontal lobe. Science 310:332–336.

Nelissen K, Vanduffel W, Orban GA (2006) Charting the lower superior temporal region, a new motion-sensitive region in monkey superior temporal sulcus. J Neurosci 26:5929–5947.

Nir Y, Fisch L, Mukamel R, Gelbard-Sagiv H, Arieli A, Fried I, Malach R (2007) Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. Curr Biol 17:1275–1285.

Ochiai T, Mushiake H, Tanji J (2005) Involvement of the ventral premotor cortex in controlling image motion of the hand during performance of a target-capturing task. Cereb Cortex 15:929–937.

Op De Beeck H, Vogels R (2000) Spatial sensitivity of macaque inferior temporal neurons. J Comp Neurol 426:505–518.

Op de Beeck HP, Dicarlo JJ, Goense JB, Grill-Spector K, Papanastassiou A, Tanifuji M, Tsao DY (2008) Fine-scale spatial organization of face and object selectivity in the temporal lobe: do functional magnetic resonance imaging, optical imaging, and electrophysiology agree? J Neurosci 28:11796–11801.

Oram MW, Perrett DI (1994) Modeling visual recognition from neurobiological constraints. Neural Netw 7:945–972.

Oram MW, Perrett DI (1996) Integration of form and motion in the anterior superior temporal polysensory area (stpa) of the macaque monkey. J Neurophysiol 76:109–129.

Orban G (2011) The extraction of 3D shape in the visual system of human and nonhuman primates. Annu Rev Neurosci 34:361–388.

Oztop E, Arbib MA (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: a computationally guided review. Neural Netw 19:254–271.

Pasupathy A, Connor CE (1999) Responses to contour features in macaque area v4. J Neurophysiol 82:2490–2502.

Pavlovic VI, Sharma R, Huang TS (1997) Visual interpretation of hand gestures for human-computer interaction: a review. IEEE Trans Pattern Anal Mach Intell 19:677–695.

Perrett D, Oram M (1993) Neurophysiology of shape processing. Image Vis Comput 11:317–333.

Perrett DI, Rolls ET, Caan W (1982) Visual neurons responsive to faces in the monkey temporal cortex. Exp Brain Res 47:329–342.

Perrett DI, Smith PA, Potter DD, Mistlin AJ, Head AS, Milner AD, Jeeves MA (1985) Visual cells in the temporal cortex sensitive to face view and gaze direction. Proc R Soc Lond B Biol Sci 223:293–317.

Perrett DI, Harries MH, Bevan R, Thomas S, Benson PJ, Mistlin AJ, Chitty AJ, Hietanen JK, Ortega JE (1989) Frameworks of analysis for the neural representation of animate objects and actions. J Exp Biol 146:87–113.

Perrett DI, Oram MW, Harries MH, Bevan R, Hietanen JK, Benson PJ, Thomas S (1991) Viewer-centred and object-centred coding of heads in the macaque temporal cortex. Exp Brain Res 86:159–173.

Pesaran B, Nelson MJ, Andersen RA (2006) Dorsal premotor neurons encode the relative position of the hand, eye, and goal during reach planning. Neuron 51:125–134.

Poggio T, Edelman S (1990) A network that learns to recognize three-dimensional objects. Nature 343:263–266.

Pouget A, Sejnowski T (1997) Spatial transformations in the parietal cortex using basis functions. J Cogn Neurosci 9:222–237.

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

Riesenhuber M, Poggio T (1999a) Are cortical models really bound by the "binding problem"? Neuron 24:87–93, 111–25.

Riesenhuber M, Poggio T (1999b) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025.

Rizzolatti G, Craighero L (2004) The mirror-neuron system. Annu Rev Neurosci 27:169–192.

Rizzolatti G, Sinigaglia C (2010) The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nat Rev Neurosci 11:264–274.

Rolls ET, 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.

Rozzi S, Ferrari PF, Bonini L, Rizzolatti G, Fogassi L (2008) Functional organization of inferior parietal lobule convexity in the macaque monkey: electrophysiological characterization of motor, sensory and mirror responses and their correlation with cytoarchitectonic areas. Eur J Neurosci 28:1569–1588.

Rust NC, Dicarlo JJ (2010) Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. J Neurosci 30:12978–12995.

Saito H, Yukie M, Tanaka K, Hikosaka K, Fukada Y, Iwai E (1986) Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. J Neurosci 6:145–157.

Salin PA, Bullier J (1995) Corticocortical connections in the visual system: Structure and function. Physiol Rev 75:107–154.

Salinas E, Abbott LF (1995) Transfer of coded information from sensory to motor networks. J Neurosci 15:6461–6474.

Schiller PH, Finlay BL, Volman SF (1976) Quantitative studies of single-cell

properties in monkey striate cortex. iii. spatial frequency. J Neurophysiol 39:1334–1351.

Schindler K, van Gool L (2008) Combining densely sampled form and motion for human action recognition. Paper presented at DAGM Symposium, Munich, Germany, June.

Schütz-Bosbach S, Prinz W (2007) Perceptual resonance: action-induced modulation of perception. Trends Cogn Sci 11:349–355.

Seltzer B, Pandya DN (1978) Afferent cortical connections and architectonics of superior temporal sulcus and surrounding cortex in rhesus-monkey. Brain Res 149:1–24.

Serre T, Riesenhuber M (2004) Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex. In: AI Memo 2004–017/CBCL Memo 23. Cambridge, MA: MIT.

AQ: T

Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T (2007a) A quantitative theory of immediate visual recognition. Prog Brain Res 165:33–56.

AQ: U

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007b) Robust object recognition with cortex-like mechanisms. IEEE Trans Pattern Anal Mach Intell 29:411–426.

Shikata E, Tanaka Y, Nakamura H, Taira M, Sakata H (1996) Selectivity of the parietal visual neurones in 3D orientation of surface of stereoscopic stimuli. Neuroreport 7:2389–2394.

Sigala N, Logothetis NK (2002) Visual categorization shapes feature selectivity in the primate temporal cortex. Nature 415:318–320.

Singer JM, Sheinberg DL (2010) Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. J Neurosci 30:3133–3145.

Song S, Miller KD, Abbott LF (2000) Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. Nat Neurosci 3:919–926.

Srivastava S, Orban GA, De Mazière PA, Janssen P (2009) A distinct representation of three-dimensional shape in macaque anterior intraparietal area: fast, metric, and coarse. J Neurosci 29:10613–10626.

Suzuki W, Tanaka K (2011) Development of monotonic neuronal tuning in the monkey inferotemporal cortex through long-term learning of fine shape discrimination. Eur J Neurosci 33:748–757.

Taira M, Tsutsui KI, Jiang M, Yara K, Sakata H (2000) Parietal neurons represent surface orientation from the gradient of binocular disparity. J Neurophysiol 83:3140–3146.

Tarr MJ, Bülthoff HH (1998) Image-based object recognition in man, monkey and machine. Cognition 67:1–20.

Taylor GW, Sigal L, Fleet DJ, Hinton GE (2010) Dynamical binary latent variable models for 3D human pose tracking. Paper presented at The 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, June.

Tessitore G, Donnarumma F, Prevete R (2010) An action-tuned neural network architecture for hand pose estimation. Paper presented at International Conference on Fuzzy Computation and International Conference on Neural Computation, Valencia, Spain, October.

Theys T, Srivastava S, van Loon J, Goffin J, Janssen P (2012) Selectivity for three-dimensional contours and surfaces in the anterior intraparietal area. J Neurophysiol 107:995–1008.

Thurman SM, Giese MA, Grossman ED (2010) Perceptual and computational analysis of critical features for biological motion. J Vis 10(12):15.

AQ: V

Tibshirani R (1994) Regression shrinkage and selection via the lasso. J Royal Statist Soc Ser B 58:267–288.

Tkach D, Reimer J, Hatsopoulos NG (2007) Congruent activity during action and action observation in motor cortex. J Neurosci 27:13241–13250.

Tsao DY, Livingstone MS (2008) Mechanisms of face perception. Annu Rev Neurosci 31:411–437.

Tsotsos JK (2011) A computational perspective on visual attention. Cambridge, MA: MIT.

Umilt à MA, Kohler E, Gallese V, Fogassi L, Fadiga L, Keysers C, Rizzolatti G (2001) I know what you are doing. a neurophysiological study. Neuron 31:155–165.

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.

Vangeneugden J, De Mazière PA, Van Hulle MM, Jaeggli T, Van Gool L, Vogels R (2011) Distinct mechanisms for coding of visual actions in macaque temporal cortex. J Neurosci 31:385–401.

Weinland D, Ronfard R, Boyer E (2011) A survey of vision-based methods for action representation, segmentation and recognition. Comp Vis Image Under 115:224–241.

Wilson HR, Cowan JD (1972) Excitatory and inhibitory interactions in localized populations of model. Biophysics 1–24.

Wolpert DM, Doya K, Kawato M (2003) A unifying computational framework for motor control and social interaction. Philos Trans R Soc Lond B Biol Sci 358:593–602.

Wu Y, Huang TS (1999) Vision-based gesture recognition: a review. In: Proceedings of the international gesture workshop on gesture-based communication in human-computer interaction (Braffort A, Gherbi R, Gibet S, Richardson J, Teil E, eds.), pp 103–115. London: Springer-Verlag.

AQ: W

Xie X, Giese MA (2002) Nonlinear dynamics of direction-selective recurrent neural media. Phys Rev E Stat Nonlin Soft Matter Phys 65:051904.

Yildiz IB, Kiebel SJ (2011) A hierarchical neuronal model for generation and online recognition of birdsongs. PLoS Comput Biol 7:e1002303.

Zemel RS, Sejnowski TJ (1998) A model for encoding multiple object motions and self-motion in area mst of primate visual cortex. J Neurosci 18:531–547.

AQ: X

Zhang K (1996) Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. J Neurosci 16:2112–2126.

Zipser D, Andersen RA (1988) A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature 331:679–684.