View-independent recognition of grasping actions with a cortex-inspired model

Falk Fleischer, Antonino Casile, and Martin A. Giese*

April 15, 2012

Abstract
To recognize how people interact with objects is essential for humans and artificial systems like robots. However, this recognition task is difficult and requires the capturing of the details of effector and goal object under a wide range of image transformations, such as view or position changes. Here, we demonstrate how specific effector-object interactions can be efficiently recognized by a simple, biologically plausible neural model. In line with biological evidence, the model applies a view-based approach for the recognition of grasping sequences from videos. The model generalizes to untrained views by interpolation between stored example views. In addition, it presents a novel physiologically plausible mechanism to capture the spatial relationship between effector and object. The results support the view that where and how an object will be grasped by an agent can be predicted without estimation of the 3D structure of the scene.

1 Introduction
People do interact with objects in various ways, usually in order to achieve different kinds of goals. For example, to move a glass from one position to the next it might be grasped from the top, while for drinking from it, it likely will be grasped from the side. Recognizing these details of effector-object interaction will aid artificial systems like robots to infer the goal of the presented action. Furthermore, capturing the goal of an action as early as possible allows systems in real-time settings to react before the action is actually finished. Recognizing action goals has been in the focus of recent robotic research for tasks like learning by demonstration [1, 2, 3, 4, 5].

In this paper, we studied the generalization properties of a biologically plausible model for the view-independent recognition of grasping actions. The architecture was originally developed in order to model the tuning properties

---

*This work was supported by the DFG, EC, EU FP7 project SEARISE, and the Hermann and Lilly Schilling Foundation.
†All Authors are with the Hertie Institute for Clinical Brain Research and the Center for Integrative Neuroscience, Tübingen, Germany, falk.fleischer@medizin.uni-tuebingen.de, {antonino. casile, martin.giese}@uni-tuebingen.de
of action-selective neurons in monkey cortex [6, 7]. It builds upon a hierarchical model for form and motion recognition [8] using a view-based recognition approach that is consistent with principles in monkey and human cortex [9]. Specifically, the model recognizes the actions from video frames without a 3D reconstruction of object or effector. The current version of the model discriminates effector and object based solely on shape properties, without information on image motion. However, we explicitly model temporal sequences of shapes, not using a probabilistic framework as e.g. [10, 4, 11, 12], but instead dynamic neural representations that are based on neural fields with asymmetric lateral connections [13, 8]. In contrast to many existing models for shape recognition that are characterized by complete position invariance, the proposed system exploits partially position-invariant detectors that still retain coarse information of the retinal positions of effector and object. This allows to recover the position of hand and object from the scene. An overview over the architecture is given in Fig. 1.

We present an extension of the model that allows to cope with view-invariant recognition of grasping actions (Section 3). In Section 4, it is demonstrated that the model generalizes to previously unseen views and positions of a specific grasping action. It is shown that the mechanism can be applied to recognize different actions on a single object early in time by explicitly capturing the spatial relationship between effector (hand) and object. In the following, related approaches are discussed.

2 Related work
Existing approaches for view-independent recognition in imitation tasks and learning by demonstration usually assume the existence of a 3D-model of the object and the effector [14, 15, 16, 17, 5, 18], for example learned from 3D reconstruction or marker-based tracking (e.g. [19]). Other approaches have implemented an appearance-based framework to recognize at least the effector from the images of the scene [20, 21, 22]. Appearance-based approaches have been quite successful in computer vision in object detection and recognition tasks [23, 24, 25, 26, 27], as well as for action recognition [28, 29, 30, 31]. A further line of research addresses systems for gesture recognition and hand
tracking (e.g. [32, 21, 33]). However, these approaches usually do not treat goal-directed actions.

Prevete et al. discussed a similar biological framework as ours that allows for the distinction of hand aperture and hand view by interpolation between a few stored examples [34]. However, they do not address the spatial interaction between hand and object. Furthermore, in contrast to [34] we explicitly model sequences of hand shapes allowing us to predict successful grasping actions early in time.

Kjellström et al. [22] studied a variety of six grasping types using an appearance-based approach for recognizing the hand from two different points of view. The hand was segmented from the image, while the appearance and the position of the object was not extracted but rather assumed to be known. In a different contribution, these authors explicitly investigate the interaction of hand actions and objects recorded from one viewpoint using conditional random fields [12]. The focus of this work were joint recognition of hand actions and goal-objects in order to capture more general action classes like drinking from a cup. In contrast, our approach addresses the more fine-grained problem of modelling action primitives like the interaction of the hand with the different affordances of a single object, even before the actual grasping takes place.

Only few other related studies have claimed to be relevant for biological systems [14, 15, 34].

3 Model architecture

The architecture consists of three major components that correspond to cortical structures that play a central role in visual action recognition: (1) a hierarchical neural system for the view-dependent recognition of object and effector shapes, (2) a circuit that is selective for temporal sequences of detector shapes, (3) a level that integrates the information about effector, object and their spatial relationship. Further details of the implementation can also be found in [7].
3.1 View-dependent shape recognition

The first component of the developed system is formed by a hierarchical neural architecture for shape recognition similar to [35, 27]. On each hierarchy layer, features are detected through neural detectors that are selective for a template pattern. In addition, features at neighboring spatial locations are pooled using a maximum operation, increasing invariance against local shifts in position [35]. A linear threshold function was applied at each hierarchy level and the output was downsampled spatially by a factor of two.

3.1.1 Layer V1/V2 - Local Orientation Detectors

Local orientations were extracted by simple cells that are modeled by a set of Gabor filters. To cover the structure of the hand, we use Gabor filters with 12 different preferred orientations $\theta$ and three different spatial frequencies $\xi$. Complex cells in the following layer integrate responses from simple cells with same orientation preference over position, scale and phase. Let $(x_{\text{even}}^{\theta, \xi}, \ldots, x_{m}^{\text{even}, \xi})$ and $(x_{\text{odd}}^{\theta, \xi}, \ldots, x_{m}^{\text{odd}, \xi})$ denote the responses of the even and odd Gabor filters from the same local neighborhood $S$ of size $m$ and scale $\xi$. Then the response of a complex cell is given by

$$r^{\theta} = \max_{j \in S, \xi} \{ (x_{\text{even}}^{\theta, \xi})^{2} + (x_{\text{odd}}^{\theta, \xi})^{2} \}.$$  

After the second layer the different spatial frequency regimes were pooled.

3.1.2 Layers V4/IT - Detectors for Shape Fragments

The neurons in the intermediate layers represent detectors that extract features of intermediate complexity. The feature detectors on the intermediate layer $i$ were defined by Gaussian Radial Basis Functions (RBFs) with the form

$$r^{i} = \exp \left( -\beta \frac{\| \tilde{r}^{i-1} - \tilde{p} \|^2}{\| \tilde{p} \|^2} \right).$$  

(1)

The centers $p$ of the RBF functions were tuned to local combinations of input features from the previous layer $i - 1$ that were specified by training patterns, with $\beta = 0.5$. During training, on each layer novel intermediate features $p$ were extracted from the responses of the previous layer within a limited spatial region. Training images showed individual hand configurations or objects. For dimensionality reduction, features were centered around the mean $m$ over the training set and their dimensionality was reduced by the linear mapping $\tilde{p} = A(p - m)$, retaining only the PCA components that were necessary for explaining 99% of the variance. The transformed features $\tilde{p}$ were then clustered based on their correlations, and the average feature of each cluster was retained. Outputs, again, were thresholded, and responses within a local spatial neighborhood were pooled with a maximum operation, followed by a spatial down-sampling with factor 2.

3.1.3 Layer IT/STS - Shape Templates for Hand and Object

The feature detectors on the highest level of the recognition hierarchy respond selectively to views of objects and hands, being sensitive to configuration, orientation and size. The response function was computed using RBFs as described
before, while responses were not pooled and down-sampled. The output of this layer is defined by an activity vector \( r(t) \) over the detector responses.

The responses on this level varied still partially with the position of shapes, making it possible to read out the positions of object and effector by a simple population code, weighting the receptive field centers by the normalized detector responses.

3.2 Temporal Sequence Selectivity exploiting Neural Fields

Selectivity for sequential order can be efficiently modeled by neural fields with asymmetric lateral connections [13, 8]. Such dynamical systems are characterized by traveling pulse solutions whose amplitude is maximal for form-stable input distributions that propagate with a particular speed \( v_0 \). The shape detectors of the IT/STS layer, whose selectivity is determined by cluster centers in feature space, are not activated for equal time intervals during the stimulus sequences. This makes it necessary to transform the activation distribution of the shape detectors into an input signal with approximately constant propagation speed, which is suitable for eliciting solutions with high amplitude in the neural field.

A form-stable moving input peak, sampled along the spatial dimension with \( N \) steps, can be characterized mathematically by the relationship: \( y(t) = [y_1(t), ..., y_N(t)] \) with \( y_i(t) = s(i \Delta x - v_0 t) \). We established a linear mapping between the temporally inhomogeneous detector responses, given by the vector \( r(t) \), and the input distribution \( y(t) \) of the form \( y(t) \cong B \cdot r(t) \). To learn this mapping we defined training pairs \([y(t_i), r(t_i)]\) of the detector responses and the idealized form-stable input peak. The elements of the matrix \( B \) were estimated by linear regression from these training examples. Separate input signals were computed for different grip types \( l \) and training views \( \phi \).

These signals provide input to snapshot neurons that are selective for the temporal order of the shapes. These neurons are embedded in a recurrent neural network (neural field) [36] that is defined by the equations

\[
\tau_r \dot{g}_k^{l,\phi}(t) = -g_k^{l,\phi}(t) + \left( \sum_m w(k - m) [g_m^{l,\phi}(t)]_+ \right) + y_k^{l,\phi}(t) - h_r
\]

where \( w \) is an asymmetric interaction kernel, \( h_r \) determines the resting level and \( \tau_r \) the time constant of the dynamics.

The responses of all snapshot neurons encoding the same action and view were integrated by motion pattern neurons, which smooth the activity over time. Their response depends on the maximum of the activities \( g^{l,\phi}(t) \) of the corresponding snapshot neurons:

\[
\tau_s \dot{s}_k^{l,\phi}(t) = -s_k^{l,\phi}(t) + \max_k [g_k^{l,\phi}(t)]_+ - h_s
\]

The motion pattern neurons, therefore, code for a particular view of an individual grip sequence, independent of the presence of a goal object. Cells that pool the responses of the motion pattern neurons of one grip type over different views can respond approximately invariant to viewpoint changes.
3.3 Integration of Object and Effector

The recognition of successful transitive actions requires the detection of the correct match of object shape, effector configuration and relative position. For example, if a bottle is grasped from the side, the form of the bottle, the opening and orientation of the hand and the location of the hand next to the bottle need to be jointly recognized.

In order to compute the relative spatial positions of the effector and object, we computed a relative position map (RPM) from the activity maps $a_E(u,v)$ and $a_O(u,v)$ of the effector, respectively the object. In these maps, which are learned from the population activity on the highest layer of the shape recognition hierarchy, object and goal correspond to a local activity peak. The neural network that is described by the relationship

$$a_{l,\phi}^{t,\phi}(u,v) = \int a_{O}^{\phi}(u',v') a_{E}^{l,\phi}(u - u', v - v') \, du' \, dv'.$$

computes an activity map, whose peak position corresponds to the relative position of the goal object in a coordinate system that is centered in the (retinal) position of the effector. This allows the definition of tuning functions $f_l(u,v)$ that are positive for all object positions relative to the effector for which effector shape and position would result in an effective grip, and which are zero otherwise (cf. blue region in the RPM indicated in Fig. 1). The response of these “affordance neurons” is given by:

$$a_l^t = \int a_{l,\phi}^{t,\phi}(u,v) f_l(u,v) \, du \, dv.$$  \(\text{(5)}\)

This information about the spatial congruency between effector and object was integrated with the information about the grip type, indicated by the motion pattern neurons, by the neural detectors at the highest level of the model. Their response is simply given by the product of the responses of the motion pattern neurons and the affordance neurons:

$$m^{l,\phi}(t) = s^{l,\phi}(t) \cdot a_{l,\phi}^{t,\phi}(t)$$  \(\text{(6)}\)

In consistency with action-selective cortical neurons (e.g., [37, 38]), these top-level detectors show strong activity only if the grip type and effector position and orientation matches the grasped object. Finally, pooling over the responses of the detectors that are selective for different views makes it possible to recognize a hand independent of the view $\phi$.

4 Results

We evaluated the generalization performance of the model to unseen views of the right hand of one subject grasping a cylinder (10cm height, 4cm diameter, appr. 30cm starting distance). The dataset consisted of 114 videos (350x315 pixels) showing two grasping actions (grasping from the top and grasping from the side) from 19 different view angles (first person perspective (0°) to third person perspective (180°), three repetitions each). The system was trained on a fixed set of seven views per grip type differing by 30 degrees. For training of the feature detectors, images (120x120 pixels) containing either the hand or
Figure 3: Tuning over time for motion pattern neurons coding different grip types. (Upper row) responses of motion pattern neurons trained on top grips; left plot shows responses to sequences of top grips and right to sequences of side grips from different views; (Lower row): corresponding responses of motion pattern neurons trained on different views of side grips.

The correct recognition of the sequence of hand shapes is essential for the performance of the overall approach. A first indication of the generalization performance of the model is shown in Fig. 4. It depicts the tuning of the motion pattern neurons over time for a top and a side grip. The plots on the left side show the responses of the motion pattern neurons for example sequences of different views of grips from the top, while the one on the right show examples for sequences from side grips, respectively. It is evident that the motion pattern responses are clearly tuned to individual views of grasping from the top (upper row) and from the side (lower row). It should be noted that while the individual activity peaks for each view correspond to training views, the results on the intermediate views correspond to sequences from the disjoint testset.

We tested the shape-recognition performance on the untrained intermediate views in more detail by evaluating the classification performance for grip types. Image frames were classified as belonging to either top or side grip, independent of the actual hand-shape position relative to the object, according to the motion
pattern neuron with maximum activity. Fig. 4 shows the performance of the grip type classification (percent correct) as a function of sequence time for testing with novel views that were not included in the training images (blue bars). The system achieved a time-independent correct classification performance of 84.9% on the given dataset, with an almost perfect classification after two thirds of the sequence length. This performance level was obtained even though the system was trained on hand pictures without goal object. The system thus efficiently ignores the clutter created by the goal object. Similar performance levels were obtained for much more subtle hand-shape differences, e.g. the discrimination between precision and power grips [7]. For comparison, the red line shows the position-independent classification performance on a testset for hand shapes without goal object (see Fig. 2, lower panel. In this case, the model achieved a perfect classification after approximately 50% of the sequence length, independent of the view. This demonstrates that the grip type recognition works efficiently also without information about the position of the hand relative to the object. The high recognition performance early in the sequence makes it possible to predict the outcome of actions even before the hand actually touches the object.

4.1 Estimation of view angle

The view angle of the presented actions can be extracted by a linear function of the output of the shape detectors $r(t)$ of the recognition hierarchy. This function was learned by linear regression from training pairs $[\phi, r_{t_i}(t_i)]$ that relates the detector output vector corresponding to the training views $\phi$ with the view angle ($\phi \in \{0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ, 180^\circ\}$). Fig. 4.1 shows the result

Figure 4: Recognition performance over time for the classification of top versus side grips from untrained action views. Blue bars show the classification of unsegmented video frames containing hand and object. The red line shows the performance for test images containing hands only. Errorbars indicate standard errors.
of decoding the view of the test sequences over time (in red) for 19 sequences per grip type. The plot demonstrates that the correct view of grasping sequences originating from untrained intermediate views can be efficiently decoded from the population response of the shape detectors with an average error of 5.2° (standard deviation ±5.1°).

![Figure 5: Estimated view angle over 570 test frames from 19 action sequences per grip type of different view point (red line) compared to the correct view (blue line). The upper graph shows the result for top grip and the lower graph for side grip. The x-axis represents the linear frame index of all tested sequences. The bar structure indicates the frames from sequences of one view angle each. Training sequences are highlighted in blue.](image)

4.2 Estimation of position

The robust representation of the spatial relationship of effector and object, as realized by Eq. (4), requires a high accuracy in the estimation of the retinal
Figure 6: Accuracy of position estimation as a function of sequence length derived from a neural population code. The plot shows the average error in pixel over time of all grasping sequences from the testset (in blue) computed according to the euclidean distance from the center of mass of the hand shapes (true position in orange). Errorbars indicate standard deviations.

positions of the shapes. Fig. 4.2 shows the mean error of the estimation of the hand position over time for all sequences in the testset using a population code that is derived from the responses of the highest layer of the shape-recognition hierarchy (in blue). The error (in pixels) was computed as the average euclidean distance between the position estimate and the center of mass of the hand shapes in the original video frames. A value of 60 pixels indicates approximately the average diameter of the hand. Fig. 4.2 demonstrates that the position of shapes in the scene can be efficiently reconstructed from the activity maps of the shape-recognition hierarchy with an average error over sequence time of approximately 5 pixels (STD: ±4.28px).

5 Conclusions

We have presented a biologically inspired architecture for the view-invariant recognition of transitive actions. The applied neural approach demonstrates that an example-based shape representation can be extended by simple neural mechanisms in order to solve the problem of the recognition of goal-directed actions. The system successfully classifies different grip types independent of the point of view and generalizes efficiently to untrained views based on a very limited number of necessary training views. In particular, grip-type recognition was possible long time before the action was completed, even without the knowledge about the relative position of object and effector. In addition, the viewpoint could be estimated with relatively high accuracy from the neural population response of the hand-shape detectors. The proposed neural circuit for integrating the information about effector and object is highly sensitive for the spatial relation between effector shape and goal object, even without a detailed reconstruction of the 3D structure. [7].

While model-based approaches usually require a computationally expensive
fitting of the model parameters to the image sequence (e.g. [39]), the proposed architecture could be realized with relatively simple computational operations that are particularly well suited for a parallel implementation. Ongoing work focuses on testing of the architecture on video databases with a higher variability in terms of grip types, object shapes and view conditions, including examples with background clutter.

References


