Bio-inspired Approach for the Recognition of Goal-Directed Hand Actions

Conference Paper · September 2009

DOI: 10.1007/978-3-642-03767-2_87 · Source: DBLP

3 authors, including:

Antonino Casile
Harvard University
57 PUBLICATIONS **1,460** CITATIONS
SEE PROFILE

Martin A. Giese
University of Tuebingen
374 PUBLICATIONS **5,241** CITATIONS
SEE PROFILE

Bio-inspired approach for the recognition of goal-directed hand actions

Falk Fleischer, Antonino Casile, and Martin A. Giese

Dept. of Cognitive Neurology, Hertie Institute for Clinical Brain Research,
Tübingen, Germany

Abstract. The recognition of transitive, goal-directed actions requires a sensible balance between the representation of specific shape details of effector and goal object and robustness with respect to image transformations. We present a biologically-inspired architecture for the recognition of transitive actions from video sequences that integrates an appearance-based recognition approach with a simple neural mechanism for the representation of the effector-object relationship. A large degree of position invariance is obtained by nonlinear pooling in combination with an explicit representation of the relative positions of object and effector using neural population codes. The approach was tested on real videos, demonstrating successful invariant recognition of grip types on unsegmented video sequences. In addition, the algorithm reproduces and predicts the behavior action-selective neurons in parietal and prefrontal cortex.

1 Introduction

The recognition of transitive actions requires additional computational mechanisms, compared to the recognition of human actions without goal objects. Recognition has to be invariant against changes in low-level image features, shifts in position, and shape transformations over time. At the same time, the distinction of different grip types (e.g. precision or power grip) requires a remarkable accuracy with respect to the detection of shape details (e.g. finger positions or their relationship to the grasped object).

This paper presents a physiologically-inspired model for the recognition of goal-directed hand actions. The model accomplishes the recognition of goal-directed hand actions from unsegmented gray-level videos. At the same time, it reproduces several biological findings in the mammal visual system, such as tuning properties of action-selective neurons in premotor cortex [1], view-dependence of recognition [2], and selectivity for the relationship between effector and object [3].

Related models have been discussed in robotics in the context of imitation learning. Many existing models in this domain are based on explicit three-dimensional shape models of effector and object (see [4] for an overview). Opposed to this work, we propose here an example-based approach that extends biologically-inspired models for the recognition of objects and actions [5–8]. Similar appearance-based approaches have been quite successful in object detection and recognition [9–11]. Opposed to previous work that has focused on the recognition of effector and body shapes from silhouettes (e.g. [12–14]), our system recognizes effector and object without previous segmentation. In contrast to other recent systems for action recognition, our system does not rely on combined space-time features (e.g. [15, 13]) or motion features (e.g. [16, 17]. Instead, spatio-temporal order is explicitly modeled by a dynamical interaction between shape representations using neural fields [18, 5]). In contrast to many existing models for shape recognition that are characterized by complete position invariance, the proposed system exploits partially position-invariant detectors for the reconstruction of the spatial relationship between effector and goal object. This relationship is crucial for the detection of functional and dysfunctional grips.

In the following, we first present the architecture and its components (Section 2). We then show results of evaluating the different components of the system in Section 3. Finally, in Section 4 implications and further extensions of the approach are discussed.

2 Architecture for the recognition of transitive actions

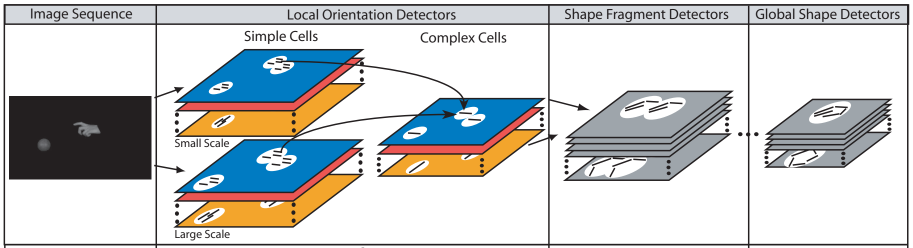
The architecture consists of three major components that correspond to cortical structures that seem to play a central role in visual action recognition: (1) hierarchical neural system for the view-dependent recognition of object and effector shapes, (2) circuit that is selective for temporal sequences of detector shapes, (3) a level that integrates the information about effector, object and their spatial relationship.

2.1 Neural hierarchy for shape recognition

The first levels of the developed system are formed by a hierarchical neural architecture for shape recognition. Each layer of this hierarchy consists of a set of neural feature detectors that are inspired by the properties of real physiological neurons. Levels with neurons that are selective for individual features alternate with levels that increase invariance by pooling over detectors with different spatial and scale preference using a maximum operation [6, 7]. The sequence of computations within each of the five layers of this hierarchy is given by: (i) feature detection through template matching, (i) maximum computation over detectors at neighboring spatial positions, (iii) application of a linear threshold function, and (iv) down-sampling by a factor of two. The parameters of the operations within each layer are summarized in Figure 1.

Layer V1/V2 - Local Orientation Detectors Local orientations are 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 θ and two different spatial frequencies ξ, as summarized in Figure 1.

Complex cells in the following layer integrate responses from simple cells with same orientation preference over position, scale and phase. Let $(x_1^{\mathrm{even}_{\theta,\xi}}, \ldots,$



Layer	1	2	3	4	5
Number Feature Maps	12	160	120	80	21
Filter Size	30x30; 40x40	8x8x12	8x8x160	8x8x120	7x7x12
Range of Max-Pooling	12x12x2	12x12	12x12	10x10	
Activity Threshold	0.4	0.4	0.5	0.6	

Fig. 1. Overview of the shape-recognition hierarchy

$x_m^{\text{even}_{\theta,\xi}}$) and $(x_1^{\text{odd}_{\theta,\xi}}, \ldots, x_m^{\text{odd}_{\theta,\xi}})$ denote the responses of the even and odd Gabor filters from the same local neighborhood S of size m and scale ξ. Then the response of a complex cell is given by $r^\theta = \max_{j \in S,\xi}\{(x_j^{\text{even}_{\theta,\xi}})^2 + (x_j^{\text{odd}_{\theta,\xi}})^2\}$. Above the second layer no distinction of different spatial frequency regimes was realized.

Layers V4/IT - Detectors for Shape Fragments The neurons in the three intermediate layers represent detectors that extract features of increasing complexity. The feature detectors on the intermediate layer i were defined by Gaussian Radial Basis Functions (RBFs) with the form

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

The centers $\mathbf{p}$ of the RBF functions were tuned to local combinations of input features from the previous layer $i-1$ that were specified by training patterns, and we chose $\beta = 0.5$.

During training, on each layer novel intermediate features $\mathbf{p}$ were extracted from the responses of the previous layer within a limited spatial region. Training images show individual hand configurations or objects. Over the training set, for dimensionality reduction, features were centered around the training mean $\mathbf{m}$ and their dimensionality was reduced by the mapping $\tilde{\mathbf{p}} = \mathbf{A}(\mathbf{p} - \mathbf{m})$, retaining only the PCA components that were necesary for explaining 99% of the variance. The transformed features $\tilde{\mathbf{p}}$ were then clustered based on their correlations, and the average feature of each cluster was retained. The number of remaining feature detectors on each intermediate layer is summarized in Figure 1.

Outputs again were thresholded, and responses within a local spatial neighborhood were pooled with a maximum operation, followed by a spatial downsampling with factor 2.

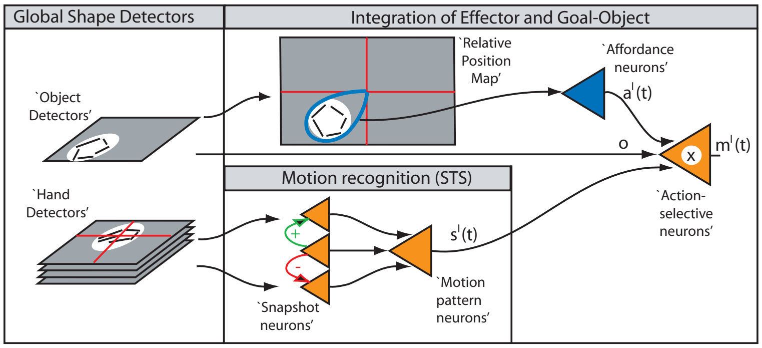


Fig. 2. Integration of hand and object information

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 is computed using a RBF as described before, while responses were not pooled and down-sampled. The responses of this level varied still partially with the object position, making it possible to read out the positions of object and effector by a simple population code.

2.2 Temporal Sequence Selectivity exploiting Neural Fields

The outputs of the detectors for the effector shapes that correspond to a specific grip type l , signified by $z_k^l(t)$, provide input to *snapshot neurons* that are selective for the temporal order with which these shapes occur. This temporal order selectivity was implemented using a simple recurrent neural network, which can be interpreted as a direction-selective neural field [5, 18]:

$$\tau_r \dot{r}_k^l(t) = -r_k^l(t) + \left(\sum_m w(k-m)\, [r_m^l(t)]_+ \right) + z_k^l(t) - h_r$$

where w is an asymmetric interaction kernel, h_r determines the resting level, and where τ_r is the time constant of the dynamics.

The responses of all snapshot neurons encoding the same action were integrated by *motion pattern neurons*, which smooth the activity over time. Their response depends on the maximum of the activities $r_k^l(t)$ of the corresponding snapshot neurons:

$$\tau_s \dot{s}^l(t) = -s^l(t) + \max_k \, [r_k^l(t)]_+ - h_s \qquad (2)$$

The motion pattern neurons are active for individual grip sequences, independent of the presence of a goal object.

2.3 Integration of Object and Effector

The recognition of functional transitive action 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 at the side of 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 $a_{\mathrm{E}}(u, v)$ and $a_{\mathrm{O}}(u, v)$ of the effector, respectively the object. In these maps, which corresponds to the highest layer of the shape recognition hierarchy described before, object and goal positions correspond to activity peaks. A simple neural network that can be described by the relationship

$$a_{RP}(u, v) = \int a_{\mathrm{O}}(u', v')\, a_{\mathrm{E}}(u' - u, v' - v)\, \mathrm{d}u'\, \mathrm{d}v'. \tag{3}$$

realizes a coordinate transformation that results in an activity map, whose peak position corresponds to the 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 $g_l(u, v)$ that are positive for all object positions relative to the effector (in this coordinate system) for which effector shape and position would result in an effective grip, and which are zero otherwise (cf. blue region indicated in Figure 1). The response of these detectors was given by:

$$a^l = \int a_{RP}(u, v)\, g_l(u, v)\, \mathrm{d}u\, \mathrm{d}v. \tag{4}$$

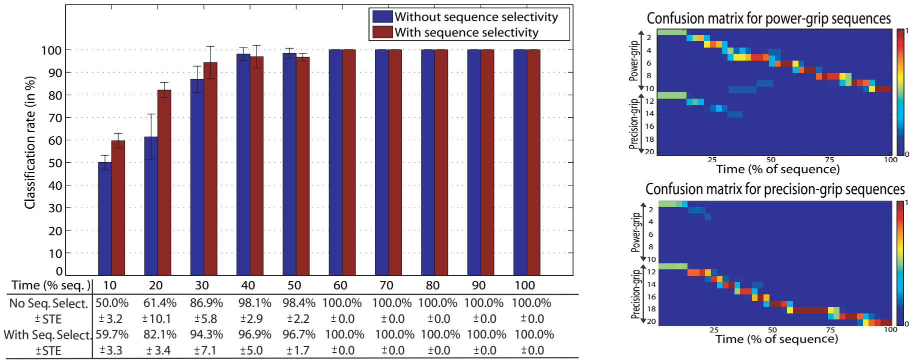
Finally, the information about this spatial congruency between effector and object can be integrated with the information about the grip type that is indicated by the motion pattern neurons. The response of the neural detectors at the highest level of the hierarchy was simply given by the product of the responses on the previous layers:

$$m^l(t) = s^l(t) \cdot a^l(t) \tag{5}$$

In consistency with action-selective cortical neurons (e.g. [1, 3]), these top-level detectors show strong activity only if the grip type and effector position and orientation matches the grasped object.

3 Results

We tested the model on unsegmented video sequences (640x480 pixels, RGB, 30 frames/sec, 30 to 40 frames) showing a side view of a hand grasping a ball (8cm diameter, 30cm starting distance) either with a power or a precision grip. We evaluated the performance by leave-one-out cross-validation on 10 sequences per grip type. For training of the feature detectors, images (120x120 pixels) containing either the hand or the object were extracted from the training sequences. The video frames were converted to grayscale and preprocessed by removing background noise, performing local contrast normalization and image whitening.



Time (% seq.)	10	20	30	40	50	60	70	80	90	100
No Seq. Select.	50.0%	61.4%	86.9%	98.1%	98.4%	100.0%	100.0%	100.0%	100.0%	100.0%
± STE	± 3.2	±10.1	±5.8	±2.9	±2.2	± 0.0	± 0.0	±0.0	±0.0	±0.0
With Seq. Select.	59.7%	82.1%	94.3%	96.9%	96.7%	100.0%	100.0%	100.0%	100.0%	100.0%
±STE	±3.3	± 3.4	±7.1	±5.0	±1.7	±0.0	±0.0	±0.0	±0.0	±0.0

(a) Performance of grip type recognition over stimulus time

(b) Confusion matrix

Fig. 3. (a) Recognition performance over time for the classification of power versus precision grips, with and without sequence selectivity; (b) corresp. confusion matrix

3.1 Recognition of grip type

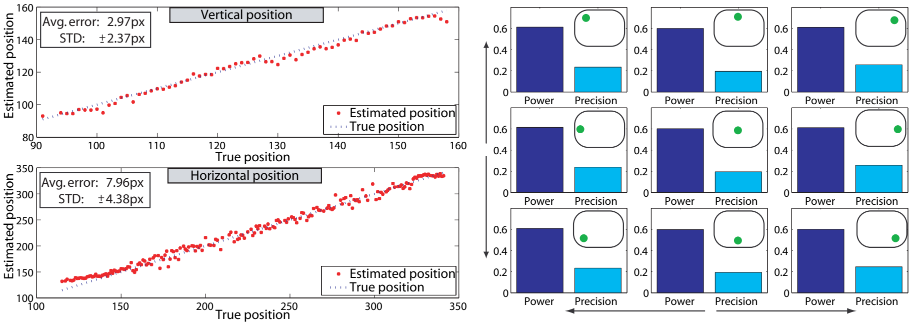
The performance of the hand-shape recognition was evaluated based on the output of the highest layer of the shape recognition hierarchy. Image frames were classified as representing either power or precision grip, according to the learned feature map with the maximum activity. Figure 3a shows the corresponding classification performance (percent correct) over time (blue bars). Averaged over the cross-validation test set, we achieve a perfect classification after approximately half of the sequence length. Figure 3b depicts the corresponding confusion matrix for power and precision grips. It is apparent that shapes usually are more likely to be confused with other shapes from the same grip type. Confusions are reduced by the sequence selection mechanism that results in an overall increase in recognition performance (see Figure 3a, red bars).

3.2 Position estimation

Figure 4a shows that the object position can be reconstructed with high accuracy from the activity maps of the highest hierarchy layer (error: 8% in horizontal and 3% in vertical direction). This high accuracy form the basis of the reliable relative position estimation realized by Eq. (3). Figure 4b demonstrates the efficiency of this mechanism, showing the very small variation of responses if action stimuli are presented at different positions of the visual field (standard deviation ±0.61% of the response to the prefered and ±2.65% to the non-prefered stimulus).

3.3 Recognition of functional vs. dysfunctional actions

The proposed system not only recognizes grip type, but also is suitable for distinguishing functional and dysfunctional grips, consistent with the properties of action-selective cortical neurons. Figure 5 depicts the average activity of the power grip detectors on the top-level of the system (Eq. 5) over the set of cross-validation stimuli. Strong responses arise only for the correct grip type in presence of the object and if the object is placed correctly relative to the effector.



(a) Estimation of average shape position

(b) Invariance to position shifts of action scene

Fig. 4. (a) Accuracy of position estimation using neural population code; (b) test of position invariance presenting the same action in different positions of the visual field; responses of output neuron trained with power grip

'Mimicked actions' where the object occurs next to the effector do not result in a significant response. This illustrates that the matching of grip affordances can be realized in an appearance-based framework without the assumption of three-dimensional representations. In addition, the behavior of the model closely resembles the ones of action-selective neurons in the superior temporal sulcus of monkeys (see inset).

4 Conclusions

We have presented a biologically inspired architecture for the recognition of transitive actions. The system explicitly models the interaction between an effector and a goal-object without a detailed reconstruction of 3D structure. The system successfully classifies different grip types based on unsegmented video stimuli and provides estimates for the 2D positions of effector and object. Recognition was highly invariant against position changes, at the same time being quite selective against the small image changes that characterize the differences between

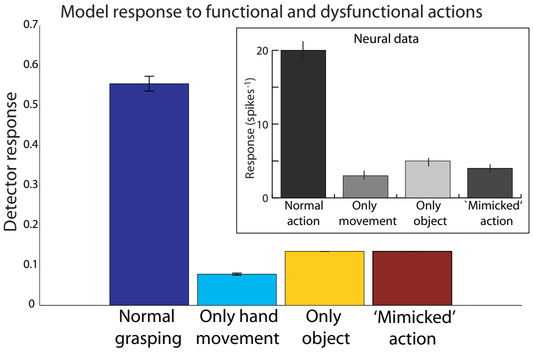


Fig. 5. Sensitivity of action-selective detectors (power grip) on the test set

precision and power grip. Ongoing work focuses on extending and testing the architecture on view-independent recognition tasks using an extended video data basis including a variety of object shapes.

References

1. Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G.: Action recognition in the premotor cortex. Brain **119 (Pt 2)** (1996) 593–609
2. Logothetis, N.K., Pauls, J., Poggio, T.: Shape representation in the inferior temporal cortex of monkeys. Curr Biol **5**(5) (1995) 552–63
3. Perrett, D.I., Harries, M.H., Bevan, R., Thomas, S., Benson, P.J., Mistlin, A.J., Chitty, A.J., Hietanen, J.K., Ortega, J.E.: Frameworks of analysis for the neural representation of animate objects and actions. J Exp Biol **146** (1989) 87–113
4. Oztop, E., Kawato, M., Arbib, M.: Mirror neurons and imitation: a computationally guided review. Neural Netw **19**(3) (2006) 254–71
5. Giese, M.A., Poggio, T.: Neural mechanisms for the recognition of biological movements. Nat Rev Neurosci **4**(3) (2003) 179–92
6. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nat Neurosci **2**(11) (1999) 1019–25
7. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. IEEE Trans Pattern Anal Mach Intell **29**(3) (2007) 411–26
8. Mutch, J., Lowe, D.G.: Multiclass object recognition with sparse, localized features. IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR) **1** (2006) 11–18
9. Weber, M., Welling, W., Perona, P.: Unsupervised learning of models of recognition. Proc. Europ. Conference on Computer Vision (ECCV) **12** (2000) 1001–1108
10. Heisele, B., Serre, T., Pontil, M., Vetter, T., Poggio, T.: Categorization by learning and combining object parts. Advances in Neural Information Processing Systems **14** (2002) 1239–1245
11. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: Proceedings of the Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic (May 2004)
12. Bobick, A.F., Davis, J.W., Society, I.C., Society, I.C.: The recognition of human movement using temporal templates. IEEE Trans on Pattern Anal and Mach Intell **23** (2001) 257–267
13. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Int. Conference on Computer Vision (ICCV) (2005) 1395–1402
14. Prevete, R., Tessitore, G., Santoro, M., Catanzariti, E.: A connectionist architecture for view-independent grip-aperture computation. Brain Research **1225** (Aug 2008) 133–145
15. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. Proc. IEEE Int. Conf. on Comp. Vision (ICCV) **1** (2007) 1–18
16. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR) **2** (2001) 123
17. Escobar, M.J., Masson, G.S., Vieville, T., Kornprobst, P.: Action recognition using a bio-inspired feedforward spiking network. Int. J. Comput. Vision **82**(3) (2009) 284–301
18. Zhang, K.: Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. J Neurosci **16**(6) (1996) 2112–26