

Introduction

• Neurons specialized for the recognition of transitive (goal-directed) hand actions have been described in the STSa, ventral premotor cortex (area F5) and parietal areas PF/PFG [1-6]. Strong connections exist between these areas [7].

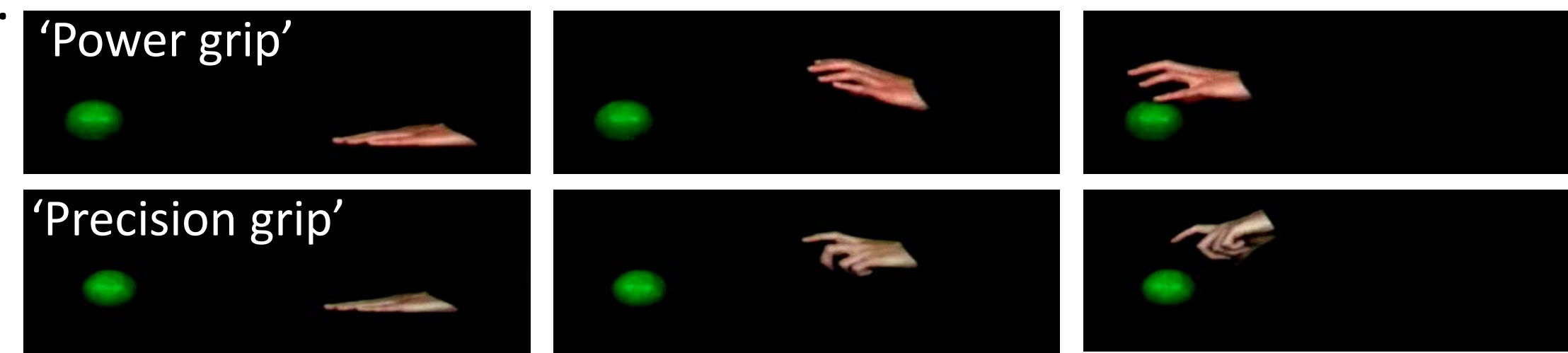


• One class of such neurons are **mirror neurons (MNs)**, found in areas F5 and PF/PFG. MNs respond both when the monkey performs an action, and when the monkey sees the same action performed by others [4-6,8].



Existing models

• Many models assume that recognition is based on internal simulation in 3D motor coordinates, based on a full reconstruction of the effector and object geometry [10-15]; this assumption contradicts the fact that many mirror neurons are view-dependent [9].



Question

How can the visual recognition of transitive actions from video sequences be accomplished using simple, physiologically plausible mechanisms?

Model architecture - Details

Shape recognition

• **Hierarchy of neural feature detectors** [16-21,31,32]

Lower level

- V1/2: Gabor filters: 12 orientations, 7 spatial scales
- Partial invariance by maximum pooling over spatial neighbourhood.

Higher levels

- Detector p on level l tuned to combinations of local orientations f modelled by Gaussian basis functions N .
- Response normalization by divisive normalization.

$$f_l^i(\vec{x}, t) = \frac{\pi_i N(f^{l-1}(\vec{x}, t) | \vec{p}_i, \Lambda_i)}{\sum_n \pi_n N(f^{l-1}(\vec{x}, t) | \vec{p}_n, \Lambda_n)}$$

Feature learning

- Unsupervised learning of detectors p by clustering of local image fragments sampled from training examples. Weights π estimated from the number of feature occurrences in the training data.
- Supervised learning of grip-phase specific hand shape detector units on highest hierarchy level by regularized regression.

Position reconstruction

- High-level shape-selective detectors only partially position-invariant (opposed to typical object recognition models) (cf. [22]).
- Computation of 2D retinal **position** $a(x, y, t)$ by population code based on marginalization over responses of these detectors.

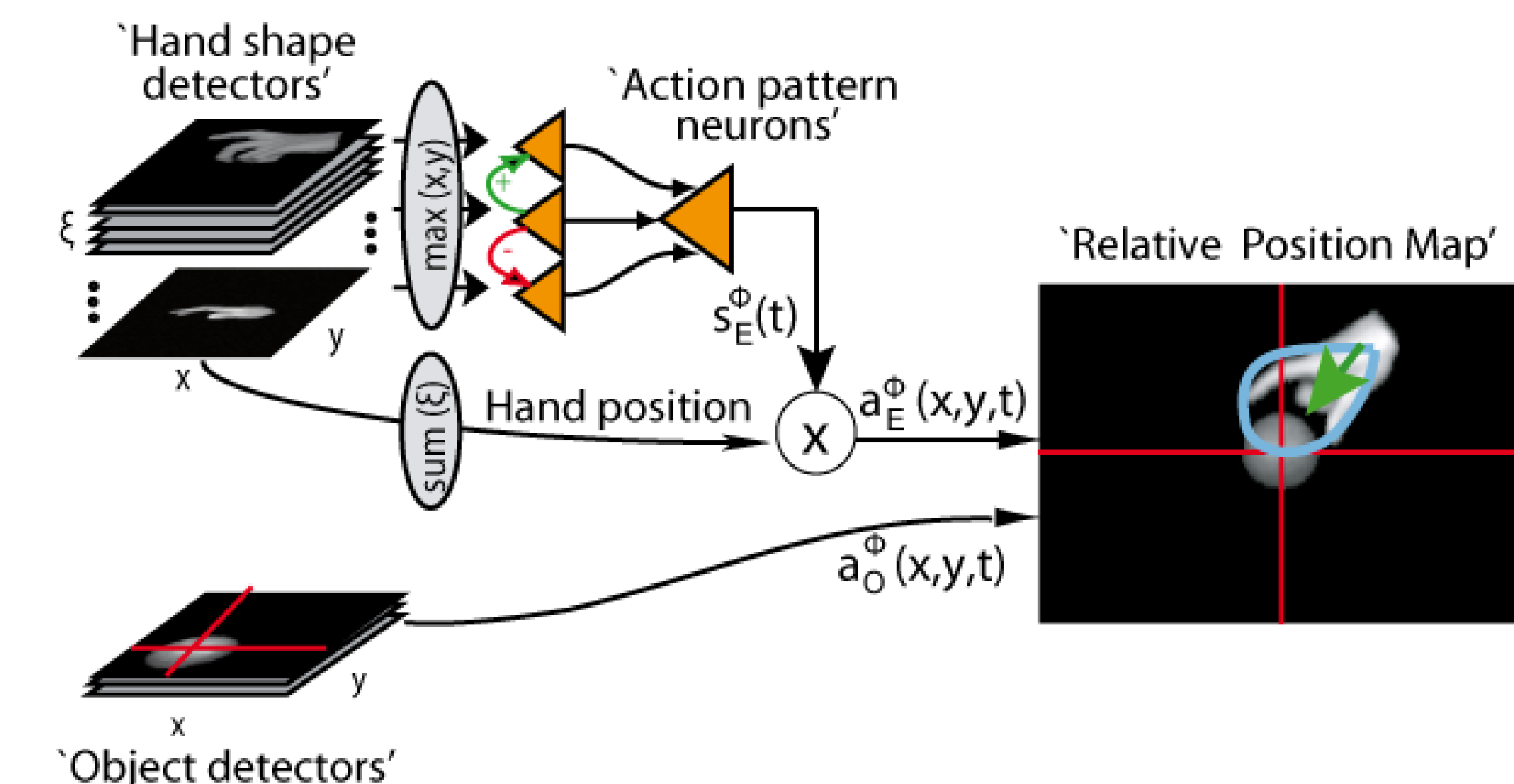
Selectivity for temporal order of shapes

- Neural field with asymmetric lateral connections integrates hand shape information over time [23-25].
- Receives position-invariant input from marginalization of hand shape detector responses over positions.
- Output of neural field integrated by action pattern neurons (activity vector s_e) that encode specific hand actions E .

$$\tau \dot{r}(x, t) = -r(x, t) + \int w(x - x') [r(x', t)] dx' + z(x, t) - h$$

$$\tau_s \dot{s}_E^\phi(t) = -s_E^\phi(t) + \max_k [r_{E_k}^\phi(t)] - h_s$$

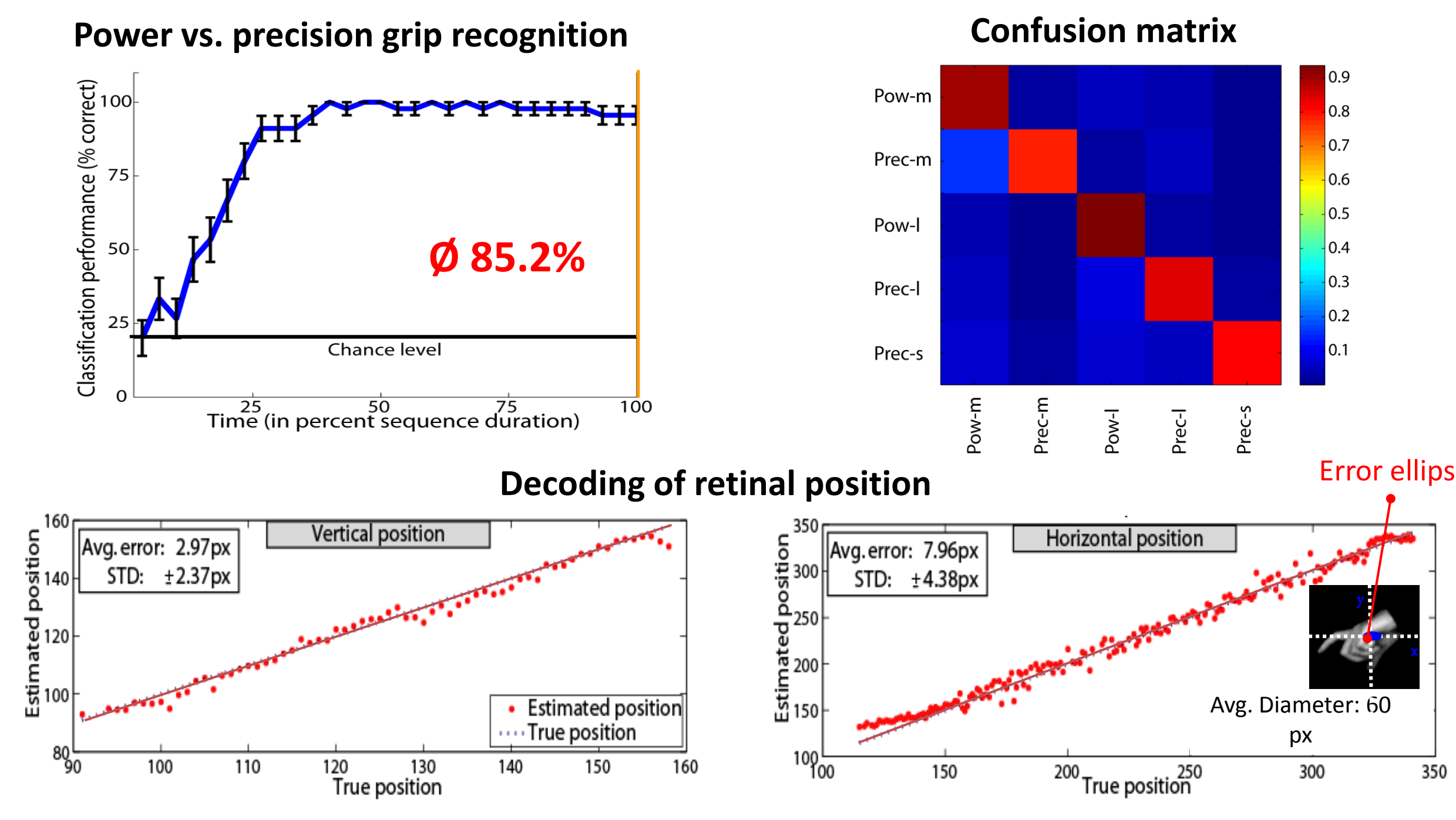
- The output of position-selective neurons at the hand location is modulated by the action pattern response s_E . This results in a position-dependent population encoding $a_E^\phi(x, y, t)$ of hand actions.



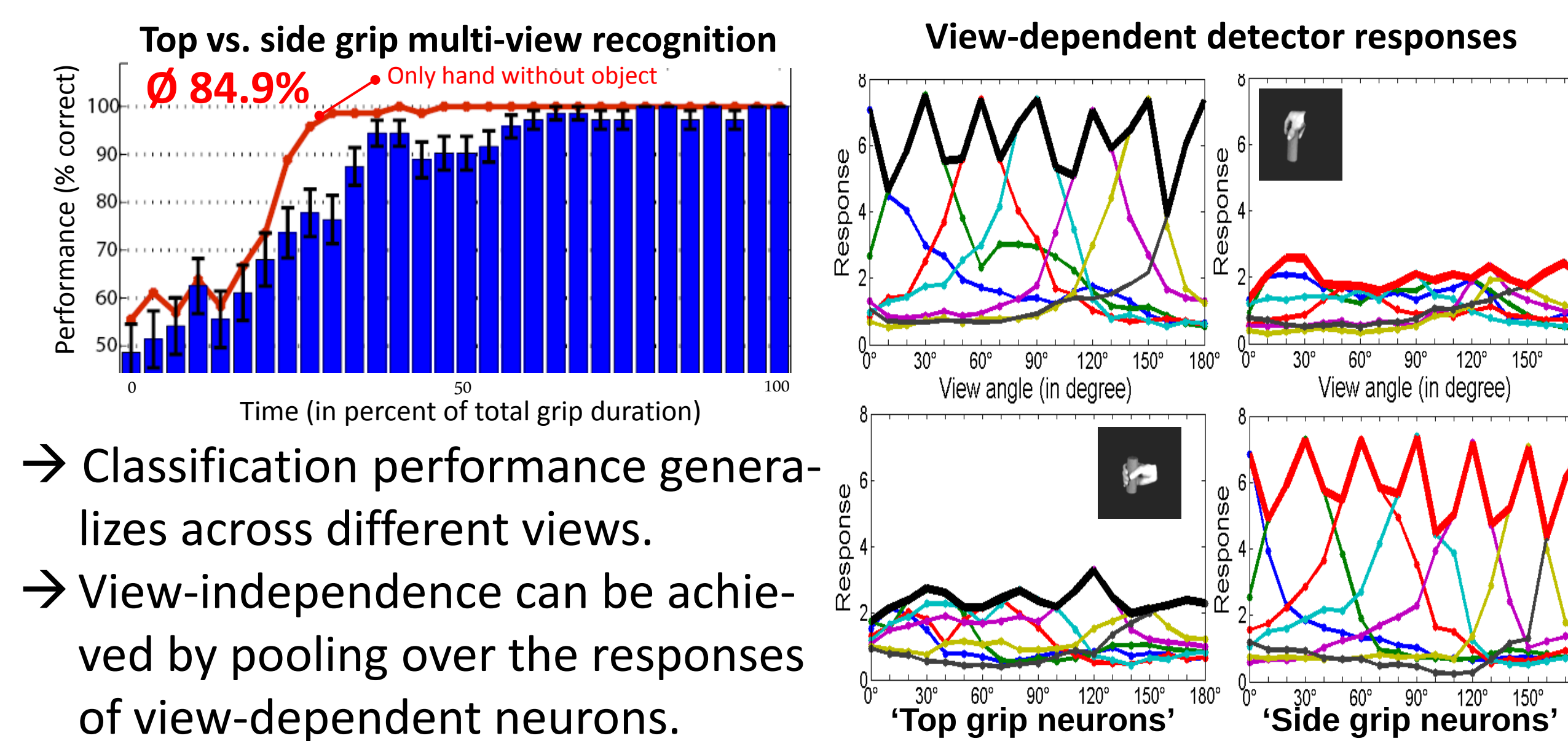
Performance: hand shape recognition

• **Grip classification on max. activity on highest model level:**

- Winner-takes-all classification over position and view.
- 3-fold leave-one-out cross validation on ten sequences per class.



• **Performance across views:**



→ Classification performance generalizes across different views.

→ View-independence can be achieved by pooling over the responses of view-dependent neurons.

Integration of hand and object

1. Transformation of retinal to **object-centered frame of reference** by gainfield-like mechanism [26,27].

→ **Relative position map:**

$$RPM_{EO}^\phi(x, y, t) = \int a_O(x', y', t) a_E^\phi(x' - x, y' - y, t) dx' dy'$$

2. Relative signal is gated by postulated **affordance neurons** with receptive fields (RF) that correspond to functional grips [28-30].

$$a_{EO}^\phi(t) = \max_{x, y} (RPM_{EO}^\phi(x, y, t)) \text{ if } (x, y) \in RF_{EO}^\phi$$

3. Simple motion energy detectors signal motion of hand → **Relative motion signal** $e(t)$ by neurons that integrate hand motion towards object (on the relative position map).

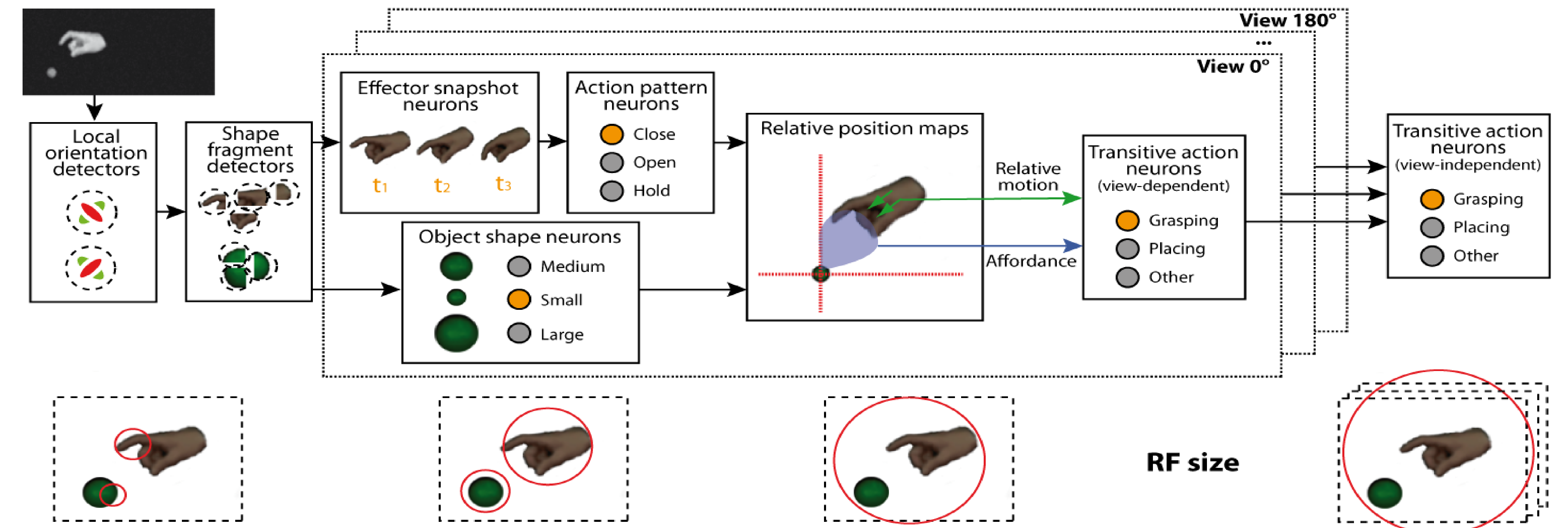
4. **Action-selective neurons** combine outputs of relative motion detectors and affordance neurons multiplicatively:

$$m_{EO}^\phi(t) = e(t) \cdot a_{EO}^\phi(t)$$

→ Response only if object matches grip type and hand moves correctly to interact with the object!

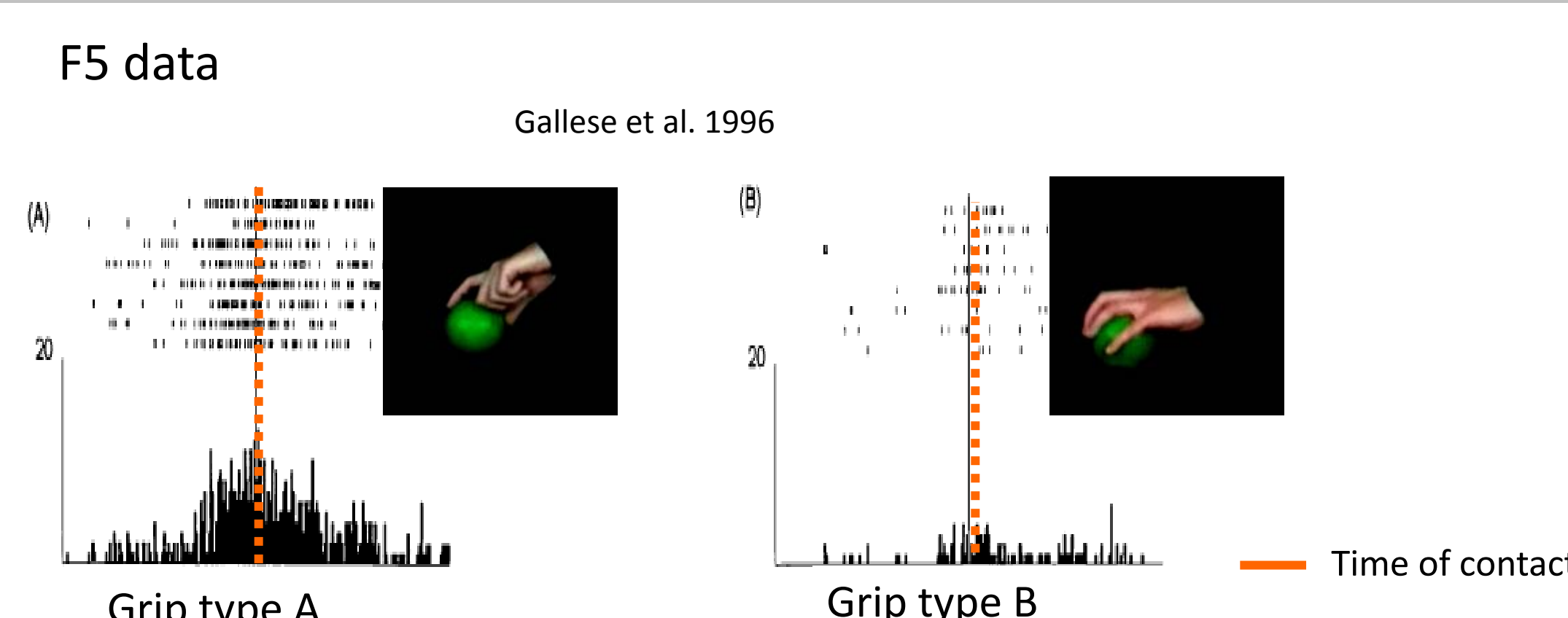
→ **View invariance by pooling over responses of view-tuned cells!**

Model architecture - Overview



Simulation of the visual tuning properties of action-selective neurons

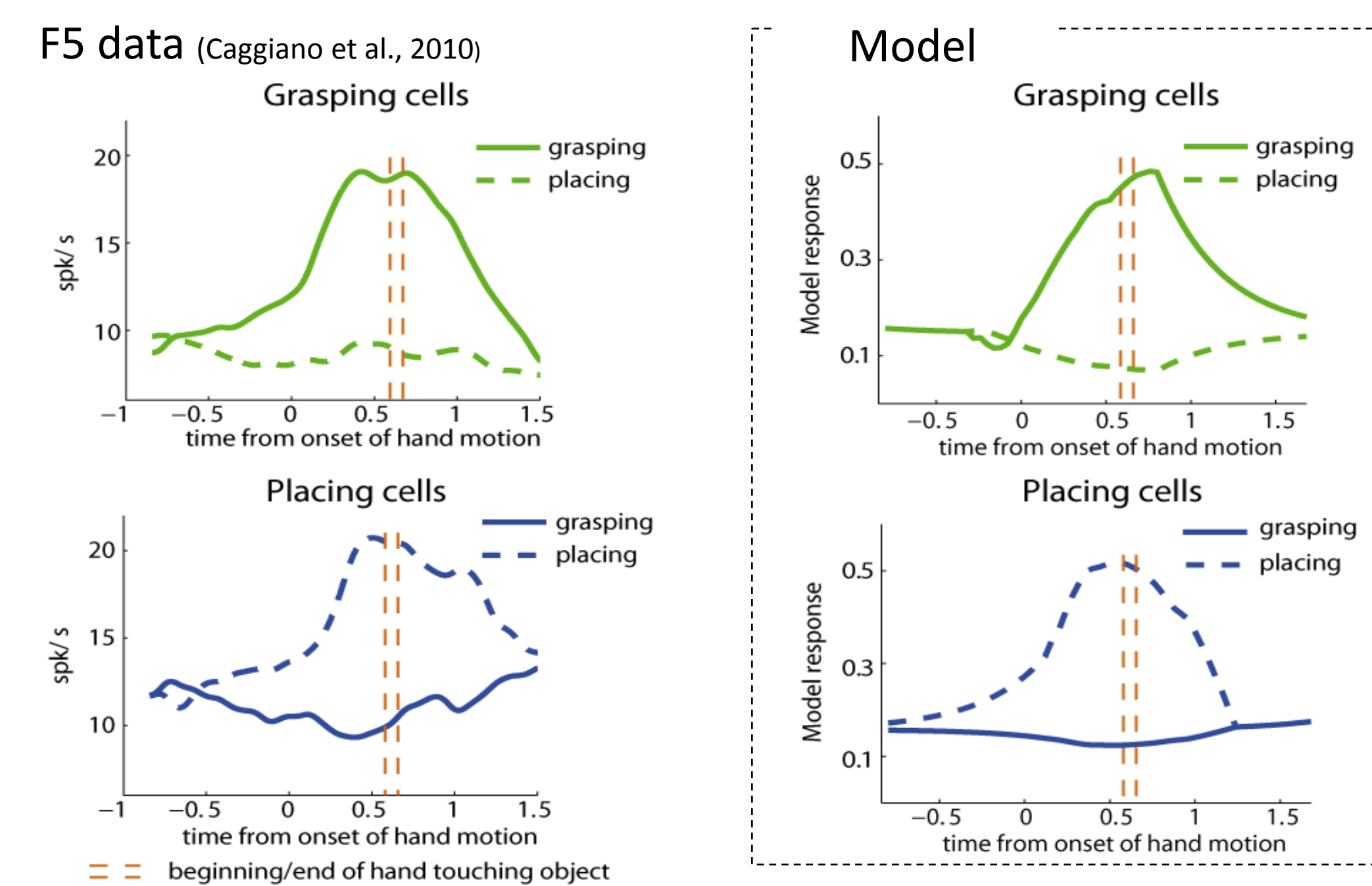
1. Grip tuning



- The model neurons are selective for individual grip types; almost perfect class. performance after 40% of stimulus duration.
- **Prediction:** Strong selectivity of response for correct temporal order (small response to scrambled or reversed sequences).

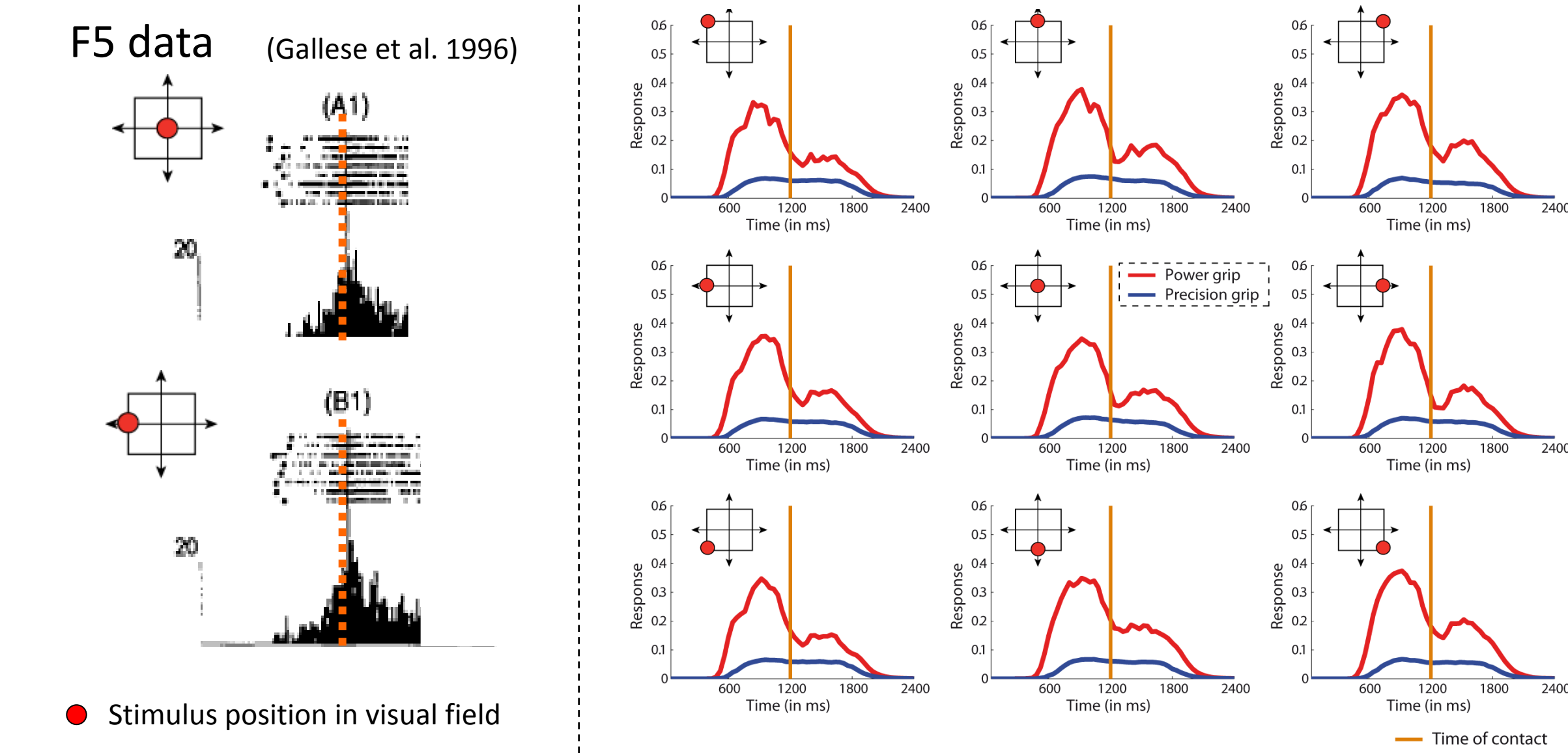
2. Sequence selectivity

- Model predicts tuning of F5 MNs to sequential order; this prediction was **confirmed experimentally**.



4. Position invariance

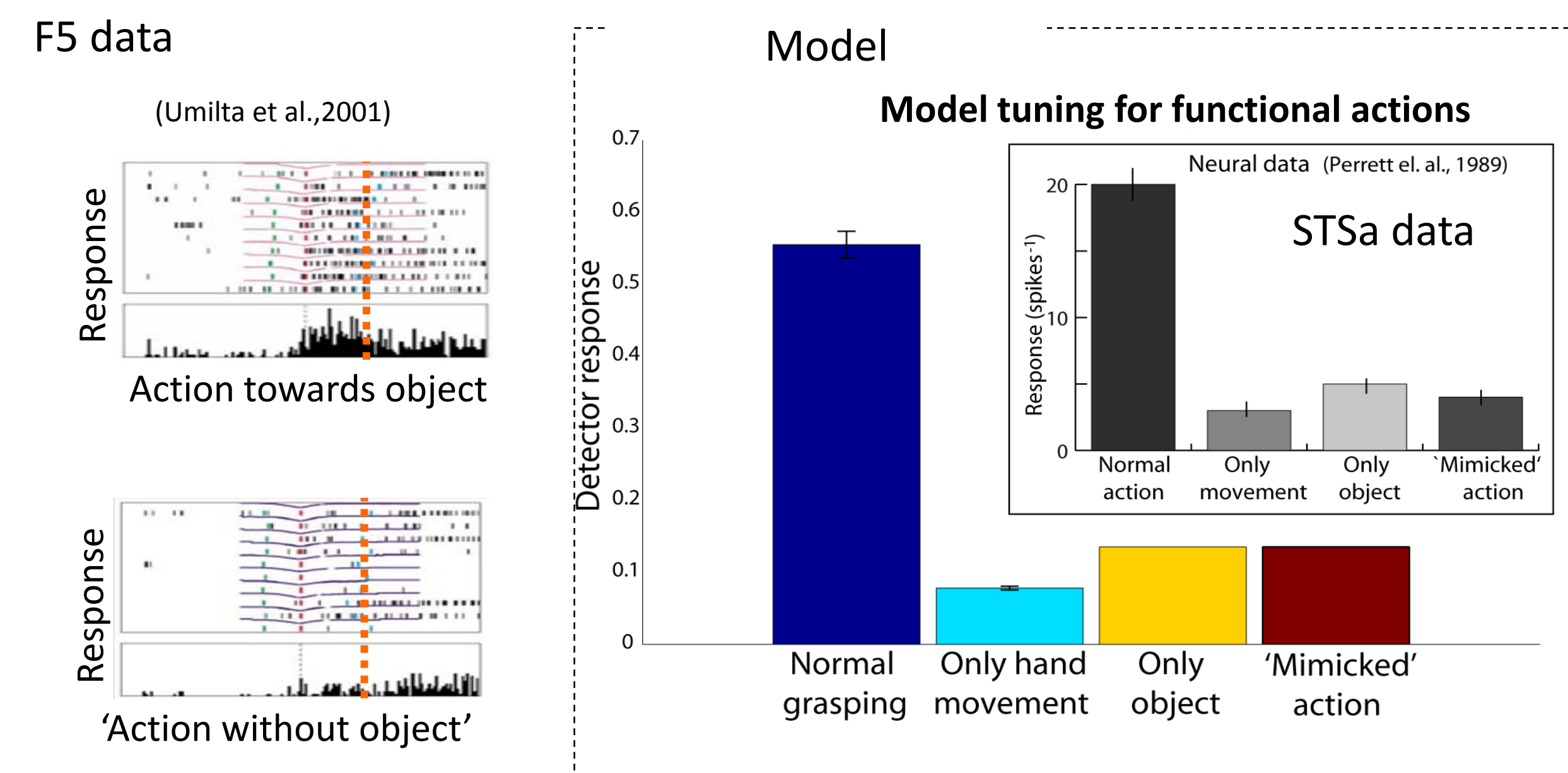
- Responses invariant against substantial shifts of stimulus within the visual field.



• Stimulus position in visual field

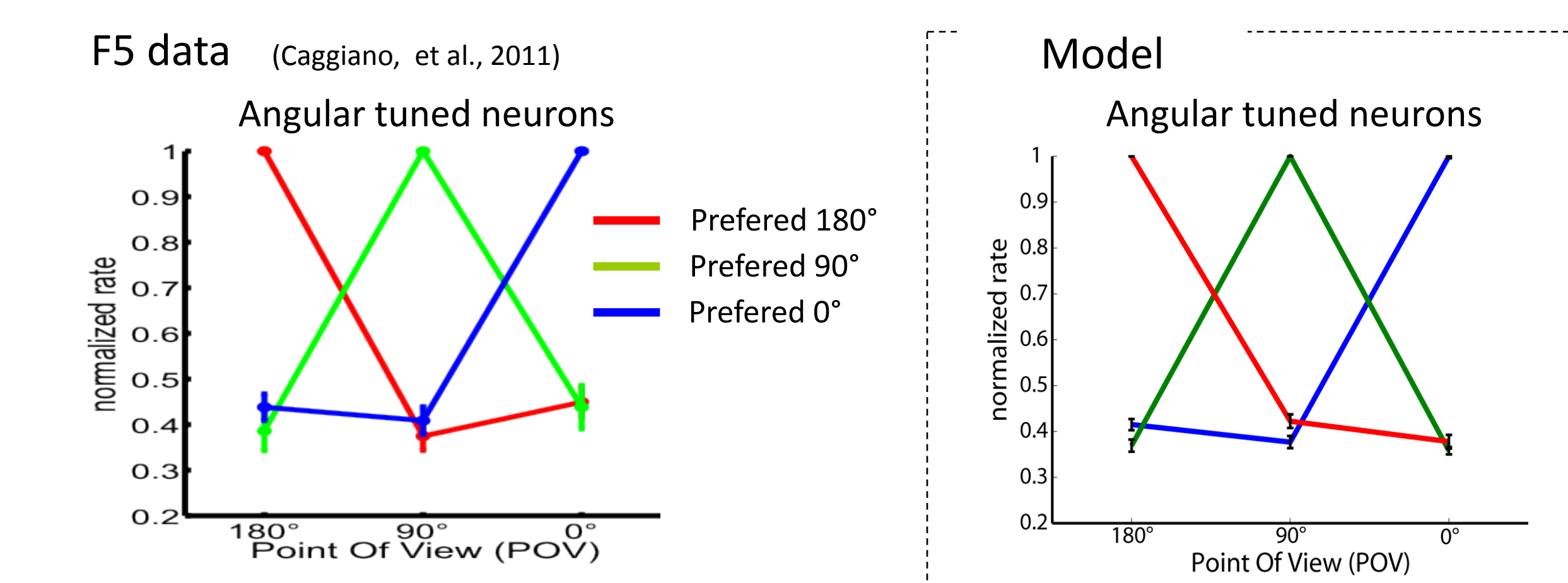
3. Tuning for functional actions

- Weak response of action-selective neurons to non-functional actions, or if goal object or effector are missing.



5. View-dependence

- Physiological recordings: 74% of the MNs are view dependent, 26% view-independent.
- **Both properties are reproduced by the model.**



Conclusions

1. Physiologically inspired model that reproduces a variety of electrophysiological results on MNs (especially view-dependence).
2. Implemented with simple physiologically plausible circuits like hierarchical shape recognition, view dependence, gain modulation, motion energy detectors and relative spatial encoding.
3. No explicit reconstruction of 3D-shape was necessary.
4. Model makes predictions that are being tested and partially confirmed in experiments, e.g. sequence selectivity of MNs.
5. The model emphasizes visual recognition processes as being sufficient for action recognition.