A conjugate neural representation of visual objects in three dimensions

Kechen Zhang & Terrence J. Sejnowski

Howard Hughes Medical Institute
Computational Neurobiology Laboratory
The Salk Institute for Biological Studies
La Jolla, California 92037

Abstract

An arbitrary movement of a rigid object in 3-D space can always be decomposed instantaneously into a translation plus a rotation. How should the neural representation of a static view of a 3-D object be updated according to its instantaneous motion? Assuming that a motion-sensitive neuron detects the rate of change of an arbitrary function of the static view of an object, we predict that the generic response properties of such a neuron can always be specified by a preferred translation direction and a preferred rotation axis in 3-D space, with cosine directional or axial tuning and linear firing-rate modulation by speed or angular speed. This theoretical framework includes known properties of some motion-sensitive cells as special cases. From the activity of a population of these neurons, a feedforward network can readily extract the instantaneous rotation axis, the angular speed, and the translation velocity, and thereby completely determine the motion of the object. We propose that neurons tuned to 3-D motion (same view, different possible rotation axes) and neurons tuned to static views (same rotation axis, different possible views) can form a conjugate representation of a 3-D object. In this way, the internal neural representation can contain both static and dynamic information and effectively mimic how the static view of an object changes during arbitrary movement.

Introduction

There is a dichotomy in the visual system between the way static and motion information is represented for 3-D objects.

Static views. Given a fixed rotation axis, different static views of the same object are probably represented by the ventral visual pathway in primates. It is known that many cells in the inferior temporal cortex in monkey respond to specific objects, and the responses often drop off smoothly as the object is rotated away from the preferred view (Perrett, Oram, Harries, Bevan, Hietanen, Benson & Thomas, 1991; Logothetis & Pauls, 1995). The view-dependence of the representation is generally consistent with the recent optical imaging data (Wang, Tanaka & Tanifuji, 1996) and various psychophysical experiments (Edelman & Bülthoff, 1992; Sinha & Poggio, 1996). Some important existing theoretical models also belong to this class (Poggio & Edelman, 1990; Ullman & Basri, 1991).

Instantaneous motion. When the view of an object is fixed, the instantaneous motion of the object, including both translation and rotation, might be represented by motion-sensitive cells in the dorsal visual pathway, including areas MT, MST and parietal cortex.

A complete representation of the dynamic state of an object would require representations of both the static view and the instantaneous motion. The generic response properties for moving 3-D object are still unknown. The aim of this paper is to predict such properties by general theoretical considerations and propose how these cells could be used for 3-D object representation.

Basic assumption

Assumption

The generic response properties of a motion-sensitive neurons for a moving 3-D object can be derived from the following assumption: *The firing rate above baseline is proportional to the time derivative of an unknown smooth function of object position and orientation in 3-D space.* In other words,

$$f = f_0 + \frac{dA}{dt}, \tag{1}$$

where f is the firing rate, f_0 is the baseline rate, and $A = A(x, y, z, \theta, \phi, \psi)$ is an arbitrary function of object position and orientation. Here (x, y, z) describe the position of the object in space and the Euler angles (θ, ϕ, ψ) describe its orientation.

The basic idea is that function A is the most general formulation of a view-dependent representation of an object because it may include all visual features of the object, even shadows and shading caused by lighting. Since the time derivative is determined by the instantaneous motion of the object, problem of occlusion is avoided automatically. A motion-sensitive neuron is expected to respond to changes of the visual features. So the proportionality to time derivative is essentially a linear approximation.

Conjugate variables

In mechanics, for a particle of mass m and position q_i along some axis, the conjugate momentum is proportional to the time derivative:

$$p_i = m\frac{dq_i}{dt}. \tag{2}$$

Unique determination of the instantaneous state of an autonomous mechanical system requires both the positions and the conjugate momenta. Under the assumption (1), the mechanical example is analogous to our dichotomy between static view and instantaneous motion considered in the beginning of this paper.

Predictions and implications

Arbitrary motion of a rigid object

The instantaneous motion of the rigid object can be completely determined by its angular velocity $\mathbf{\Omega}$ and the invariant translation velocity $\mathbf{U} \equiv \mathbf{V} - \mathbf{\Omega} \times \mathbf{C}$, where $\mathbf{C}$ a reference

point through which the rotation axis passes, and $\mathbf{V}$ is the translation velocity which depends on the reference point $\mathbf{C}$ because parallel rotation axes are equally legitimate choices. Once $\mathbf{\Omega}$ and $\mathbf{U}$ are known, the instantaneous object motion is completely determined, and the velocity of any material point on the object is

$$\mathbf{v} = \mathbf{V} + \mathbf{\Omega} \times (\mathbf{r} - \mathbf{C}) = \mathbf{U} + \mathbf{\Omega} \times \mathbf{r}, \tag{3}$$

where vector $\mathbf{r}$ is the position of the point.

Predicted tuning rule

Given a 3-D object moving at instantaneous angular velocity $\mathbf{\Omega}$ and translation velocity $\mathbf{U}$, we predict that the firing rate of a generic neuron should be

$$f = f_0 + \mathcal{T} \cdot \mathbf{U} + \mathcal{R} \cdot \mathbf{\Omega}, \tag{4}$$

where f_0 is the background firing rate, vector $\mathcal{T}$ is the preferred translation direction and vector $\mathcal{R}$ is the preferred rotation axis. Both $\mathcal{T}$ and $\mathcal{R}$ may depend upon the object as well as the view of the object, but not upon the translation velocity and angular velocity. This tuning rule is a logical consequence of the assumption (1) and the geometric constraints of rigid object motion. The derivation is straightforward and is omitted here.

In other words, we predict that *the responses should be proportional to the cosine of the angle between the actual axis and the preferred axis for both rotation and translation; in addition, the firing rates should be modulated by angular speed and translation speed approximately linearly.*

Predicted displacement constraints: changing object position in space

For a given object, both the preferred translation direction $\mathcal{T}$ and the preferred rotation axis $\mathcal{R}$ in general depend on the exact object position in space, which determines the exact view. Thus $\mathcal{T}$ and $\mathcal{R}$ are both vector fields in real 3-D space. It can be shown that $\mathcal{T}$ must be curl-free, namely,

$$\oint \mathcal{T} \cdot d\mathbf{l} = 0 \tag{5}$$

along any closed curve in 3-D space; and $\mathcal{R}$ must be divergence-free, namely,

$$\iint \mathcal{R} \cdot d\mathbf{s} = 0 \tag{6}$$

on any closed surface in 3-D space. These conditions provide strong constraints on the distribution of the preferred translation directions and rotation axes in space.

Predicted existence of central axis: changing origin of rotation axis

The preferred translation axis $\mathcal{T}$ is independent of the choice of reference point $\mathbf{C}$ in Eq. (3), but the preferred rotation axis is. The exact amount of drift of the preferred rotation axis can be predicted. In particular, the reference point can be chosen such that the preferred rotation axis becomes parallel with the translational axis. This choice is unique, and we call this axis the *central axis.* The general tuning rule can be written as

$$f = f_0 + \mathcal{T} \cdot \mathbf{V} + (\mathcal{R} + \mathcal{T} \times \mathbf{C}) \cdot \mathbf{\Omega}. \tag{7}$$

It can be shown that the central axis can be specified by the reference point

$$\mathbf{C} = \frac{\mathcal{R} \times \mathcal{T}}{|\mathcal{T}|^2} + k\mathcal{T}, \tag{8}$$

where coefficient k is arbitrary because the axis is in parallel with $\mathcal{T}$. Now the new rotation axis is parallel with the preferred translation direction $\mathcal{T}$:

$$\mathcal{R}' \equiv \mathcal{R} + \mathcal{T} \times \mathbf{C} = \frac{\mathcal{R} \cdot \mathcal{T}}{|\mathcal{T}|^2}\mathcal{T}. \tag{9}$$

Explicit examples

Consider a local-motion detector responding to object motion projected orthogonally onto the frontoparallel plane. It has a preferred direction $\mathbf{p}$ so that its response to local velocity $\mathbf{v}$ is $\mathbf{p} \cdot \mathbf{v}$. If the motion is generated by a point with coordinates $\mathbf{r}$ on a rigid object in real 3-D space, then the preferred translation direction and rotation axis can be obtained explicitly:

$$\mathcal{T} = \mathbf{p}, \qquad \mathcal{R} = \mathbf{r} \times \mathbf{p}. \tag{10}$$

By adding up two local motion detectors with opposite preferred directions $\mathbf{p}$ and $-\mathbf{p}$ responding to two points $\mathbf{r}_1$ and $\mathbf{r}_2$ on an object, we can build a detector with

$$\mathcal{T} = \mathbf{0}, \qquad \mathcal{R} = (\mathbf{r}_1 - \mathbf{r}_2) \times \mathbf{p}, \tag{11}$$

which are invariant with respect to parallel shift of rotation axis.

Proposed experimental tests

Recordings should be made from a neuron in the dorsal visual pathway while slightly oscillating an object around a fixed axis. In theory, the oscillation should be infinitesimal. In practice, it just needs to be small so that occlusion problem can be avoided. Suppose the oscillation is sinusoidal with frequency F, then according to the tuning rule (4), the response should be

$$f(t) = f_0 + kF\cos(\alpha)\cos(2\pi Ft + \phi), \tag{12}$$

where α is the angle between the actual axis and the preferred axis, k is a constant coefficient and ϕ is a phase shift. Note that the modulated response is also proportional to the frequency F of the oscillation. Systematically changing the orientation of the axis in 3-D space while keeping the view fixed allows the preferred axis to be determined. Similarly, the response to translation in 3-D space could be tested by oscillating the whole object along a straight line.

Testing the assumption on known biological systems

The generic tuning rule (4) is derived from the basic assumption (1). Although the assumption looks reasonable, it validity depends ultimately on whether predictions based on it are verified experimentally. To test its validity, we applied the same argument to some known biological systems and found that it often leads to encouraging results.

Motor cortical directional tuning

Consider stereotyped reaching movement in which the configuration of the whole arm is determined completely by the hand position (x, y, z) in space. In this situation our basic assumption becomes that *the firing rate above baseline is proportional to the time derivative of an unknown smooth function of hand position in 3-D space.* This arbitrary function $A(x, y, z)$ includes any functions of arm configuration, such as muscle length, joint angles, and any combination of those (Mussa-Ivaldi, 1988). According to the assumption, we have

$$f - f_0 = \frac{dA}{dt} = \frac{\partial A}{\partial x}\frac{dx}{dt} + \frac{\partial A}{\partial y}\frac{dy}{dt} + \frac{\partial A}{\partial z}\frac{dz}{dt} = \mathcal{P} \cdot \mathbf{v}, \tag{13}$$

where vector $\mathcal{P} = \nabla A$ is the *preferred direction* and $\mathbf{v}$ is the reaching velocity, which is always pointing in the instantaneous reaching direction.

This formula captures two major effects: cosine directional tuning and linear speed modulation. The first effect is ubiquitous in the motor cortices (Georgopoulos, Schwartz & Kettner, 1986), while the second effect is implied in the fact that adding up the population vector head-to-tail approximately reproduces the hand trajectory, with the same scaling law for curvature speed trade-off (Schwartz, 1994).

Similar to Eq. (5), the vector field of preferred direction must have zero curl; that is,

$$\oint \mathcal{P} \cdot d\mathbf{l} = 0 \tag{14}$$

along any closed curve in 3-D space. It is known that the preferred direction of a motor cortical cell typically depends on the starting point of hand position (Caminiti, Johnson, Galli, Ferraina & Burnod, 1991). So the curl-free condition (14) limits the distribution of the preferred directions for different starting positions in 3-D space. The curl-free condition provides useful predictions. For example, it unequivocally rules out the possibility of any circular arrangement of preferred directions in space.

Place cell speed modulation

The firing of a hippocampal place cell is determined not only by the animal's spatial position but is also modulated approximately linearly by the running speed, at least for movement confined to a narrow track (McNaughton, Barnes & O'Keefe, 1983). In one-dimensional case, by similar argument as in the preceding section, we find

$$f(x, v) = f_0(x) + G(x)v, \tag{15}$$

where v is the running speed and

$$G(x) = \frac{dA(x)}{dx} \tag{16}$$

is the gradient of the arbitrary function A of spatial position x. Equation (15) captures linear speed modulation and allows arbitrary place tuning $G(x)$. In reality, the linear speed modulation is a very good approximation, with a correlation coefficient greater than 0.95 when averaged over all simultaneously recorded cells (Zhang, Ginzburg, McNaughton & Sejnowski, 1997).

Areas MT and MST

These motion-sensitive visual areas in monkey are a candidate for searching for cells tuned to 3-D object motion. Although the existing results are not sufficient to confirm our general predictions, we can verify that some known properties of some cells in these regions are indeed special cases of our generic tuning rule (4).

For the movement of a small dot, an approximate cosine directional tuning with linear speed modulation can be expected from the same argument shown in the preceding sections. This is a crude characterization of a typical MT cell (Rodman & Albright, 1987). Understandably the real data have nonlinear effects in both the directional tuning and the speed modulation.

Many MST cells are known to respond best to a large-field spiral motion (Graziano, Andersen & Snowden, 1994). Consider a large rigid object, which could be the environment itself. For this object, if the preferred rotation axis and preferred translation direction are both pointing towards the observer, then the optic flow generated by the optimal object motion is an expansion plus a rotation, which is a spiral field. For a pure expansion cell, the translation direction is pointing towards the observer whereas the rotation axis is missing. Although shift of focus for expansion and rotation fields has been studied (Duffy & Wurtz, 1995), general 3-D object motion has not been tested. The real question here is how to parameterize the visual stimulus so that it can be varied systematically. This is also a relevant consideration for surround effects. Our new results predict what to expect if moving 3-D objects are used as stimuli. Coding 3-D object motion in MST has recently been considered by Zemel & Sejnowski (1995).

Extracting instantaneous object motion from population activity

Suppose we have N cells obeying the predicted tuning rule, and $n_1, n_2, \ldots, n_N$ spikes are collected from these N cells within the time window τ. Then the instantaneous arbitrary motion of an object can be determined completely by using feedforward mechanisms, which are considered biologically plausible. Let the mean firing rate of cell i be

$$f_i = f_{0i} + \mathcal{T}_i \cdot \mathbf{U} + \mathcal{R}_i \cdot \mathbf{\Omega}. \tag{17}$$

The population vector is the simplest method:

$$\mathbf{U} \propto \sum_{i=1}^{N} n_i \mathcal{T}_i \qquad \text{and} \qquad \mathbf{\Omega} \propto \sum_{i=1}^{N} n_i \mathcal{R}_i \tag{18}$$

for large N. The validity of this method relies on the conditions that both matrices $\sum_i \mathcal{T}_i \mathcal{T}_i^T$ and $\sum_i \mathcal{R}_i \mathcal{R}_i^T$ are proportional to the 3×3 identity matrix, while matrix $\sum_i \mathcal{T}_i \mathcal{R}_i^T$ vanishes.

The probabilistic method based on Bayesian approach is more accurate:

$$\{\mathbf{U}, \mathbf{\Omega}\} = \arg \max_{\mathbf{U},\mathbf{\Omega}} \sum_{i=1}^{N} [n_i \log(\tau f_i) - \tau f_i], \tag{19}$$

where f_i is a function of $\mathbf{U}$ and $\mathbf{\Omega}$ as given by Eq. (17). Here the assumptions are Poisson spike distribution and independence of different cells. When these conditions

are met, the Bayesian method can achieve the Cramér-Rao lower bound. This method can also be implemented as a feedforward network. For more detailed discussions, see Zhang, Ginzburg, McNaughton & Sejnowski (1997).

Discussion

Moving rigid objects are natural stimuli that can be expected to have efficient representations in the visual cortex. It is surprising that a general argument with minimal assumption can fully predict the generic behaviors for neurons tuned to the motion of 3-D objects. The key argument presented in this paper, conceptually simple, is able to make interesting predictions and has the unifying power to capture some known generic properties of some neurons in the motor cortex, the hippocampus, and the visual cortex. This may help to explain the ubiquity of these properties. In all these cases, the responses properties are ultimately determined by the geometry of the spatial transformation. On the other hand, the real biological system is invariably more complex, and any prediction of cosine tuning curves is unlikely to be exactly true especially at the negative side lobes because of low firing rate cutoff.

To our knowledge, systematic experiment using rotating 3-D object as a stimulus while keeping the view fixed has not been performed. Thus our prediction of the generic behavior of neurons can be directly tested. If the predicted tuning rule turns out to be true, it would have important biological implications. The instantaneous arbitrary motion of an object can then be determined completely by using feedforward mechanisms considered in the preceding section.

A view-specific representation of the dynamic state of an object would require information about both the current view and the instantaneous motion; that is, how the view is changing. The dichotomy between static view and instantaneous motion representations is essentially valid because neurons with transient and sustained responses are segregated early on in the visual pathways, although the implementation using the first order time derivation is only an approximation. The instantaneous motion determines how the activity pattern for the static view in the temporal cortex should be updated so that a new cell population can be activated in accordance with the rotation axis. By coupling both representations, the internal neural representation of an object can contain both static and dynamic information and effectively mimic how the static view of an object changes during arbitrary movement.

References

Caminiti, R., Johnson, P. B., Galli, C., Ferraina, S., & Burnod, Y. (1991). Making arm movements within different parts of space: Premotor and motor cortical representation of a coordinate system for reaching to visual targets. *Journal of Neuroscience, 11*, 1182–1197.

Duffy, C. J. & Wurtz, R. H. (1995). Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. *Journal of Neuroscience, 15*, 5192–5208.

Edelman, S. & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research, 32*, 2385–2400.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233*, 1416–1419.

Graziano, M. S., Andersen, R. A., & Snowden, R. J. (1994). Tuning of MST neurons to spiral motions. *Journal of Neuroscience, 14*, 54–67.

Logothetis, N. K. & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representation in the primate. *Cerebral Cortex, 3*, 270–288.

McNaughton, B. L., Barnes, C. A., & O'Keefe, J. (1983). The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. *Experimental Brain Research, 52*, 41–49.

Mussa-Ivaldi, F. A. (1988). Do neurons in the motor cortex encode movement direction? An alternative hypothesis. *Neuroscience Letters, 91*, 106–111.

Perrett, D. I., Oram, M. W., Harries, M. H., Bevan, R., Hietanen, J. K., Benson, P. J., & Thomas, S. (1991). View-centered and object-centered coding of heads in the macaque temporal cortex. *Experimental Brain Research, 86*, 159–173.

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

Rodman, H. R. & Albright, T. D. (1987). Coding of visual stimulus velocity in area MT of the macaque. *Vision Research, 27*, 2035–2048.

Schwartz, A. B. (1994). Direct cortical representation of drawing. *Science, 265*, 540–542.

Sinha, P. & Poggio, T. (1996). Role of learning in three-dimensional form perception. *Nature, 384*, 460–463.

Ullman, S. & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*, 992–1005.

Wang, G., Tanaka, K., & Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. *Science, 272*, 1665–1668.

Zemel, R. S. & Sejnowski, T. J. (1995). Grouping components of three-dimensional moving objects in area MST of visual cortex. In G. Tesauro, D. Touretzky, & T. K. Leen (Eds.), *Advances in Neural Information Processing Systems*, volume 7, pages 165–172. Cambridge, MA: MIT Press.

Zhang, K., Ginzburg, I., McNaughton, B. L., & Sejnowski, T. J. (1997). Interpreting neuronal population activity by reconstruction: A unified framework with application to hippocampal place cells. *Journal of Neurophysiology*. Submitted.