

# EXPLICIT OBJECT REPRESENTATION BY SPARSE NEURAL CODES: A COMPUTATIONAL MODEL

Stephen Waydo

Control & Dynamical Systems  
California Institute of Technology  
waydo@cds.caltech.edu

Christof Koch

Computation & Neural Systems  
California Institute of Technology

Highly sparse representations of objects in the visual environment in which individual neurons display a strong selectivity for only one or a few stimuli (such as familiar individuals or landmark buildings) out of perhaps 100 presented to a test subject have been observed in the human medial temporal lobe (MTL), a brain area believed to be crucial to the formation of new semantic memories. The process by which more distributed representations earlier in the visual pathway are transformed to produce such highly selective and invariant units results in information represented only *implicitly* by the pattern of light impinging on the retina and in the firing of neurons in early visual areas being made *explicit* at the level of MTL. This “sparsification” may be an important design principle underlying the structure of this brain region. We apply a model of sparsification in which a network of nonlinear neurons generates a sparse representation of its inputs through an unsupervised learning process to the outputs of a biologically realistic model of the human ventral visual pathway. Although the underlying constraint in the model is merely to produce a sparse representation of the input set, units emerge that respond selectively to specific image categories. The sparseness constraint thus facilitates the formation of explicit representations of image categories, despite the category information being represented only implicitly in the input images.