Object recognition in ventral visual cortex

Giuseppe di Pellegrino

Department of Psychology, University of Bologna

g.dipellegrino@unibo.it

Cognition and Neuroscience

Second cycle Degree in Artificial Intelligence – 2024/25

Introduction: What is Vision?

Vision is a central cognitive function that allows us to interpret and navigate the world. At its core, vision is the process by which organisms interpret visual stimuli from the external environment. As David Marr (1982) succinctly described, to see is "to know what is where by looking." According to Marr and Nishihara (1978), vision constructs a description of the external world that is useful to the observer and stripped of irrelevant information. This highlights a key insight: vision does not aim to capture every detail but rather to filter and structure information to support behavior and cognition..

The Computational Nature of Vision

Vision is fundamentally an information-processing task, transforming images into meaningful representations to guide action and thought. The study of vision must therefore include not only the study of how to extract from images the various aspects of the world that are useful to us (i.e, processes), but also an inquiry into the nature of the internal representations by which we capture this information and thus make it available to support behavior and cognition.

Marr's Tri-Level Hypothesis and Computational Approaches

Marr proposed that visual systems should be understood at three levels: computational (the goal of the system), algorithmic (the process and representations), and implementational (the physical realization). While Marr emphasized independence between these levels, contemporary neuroscience highlights their interdependence. Neurobiological constraints play a crucial role in shaping viable computational theories, narrowing the vast space of possible solutions.

Vision as Bayesian Inference

Vision can also be conceptualized as a form of Bayesian inference. In this framework, the brain combines prior knowledge (priors) with sensory evidence (likelihoods) to infer the most probable interpretation of a visual scene (percepts). The visual system is thus treated as an ideal observer that estimates the posterior probability of a given percept. This probabilistic interpretation explains how perception remains robust despite noisy or ambiguous sensory input and accounts for phenomena such as perceptual illusions and rapid recognition in cluttered scenes.

Mechanistic Framework of Vision

Vision is not akin to a passive recording (like a camera) but involves active interpretation. It is hierarchical and multi-layered, encompassing systems-level pathways (dorsal and ventral streams), neural circuits, single neurons, and molecular mechanisms. Object recognition arises from integrating all these levels into a mechanistic understanding.

Visual Pathways and the Structure of Visual Processing

Visual information travels from the retina through the lateral geniculate nucleus (LGN) to the primary visual cortex (V1), continuing into 30 or more extrastriate visual areas. This path, known as the retino-geniculo-striate pathway, forms the basis of conscious visual perception of the visual environment. From V1, two main processing streams emerge:

- The dorsal stream ("where/how"), involved in spatial processing and action guidance.
- The ventral stream ("what"), critical for object recognition.

Vision is Studied at Three Levels.

Traditionally, a visual scene is analyzed at three levels: low, intermediate and high. At the lowest level, visual attributes such as local contrast, orientation, color, depth and motion are processed. Intermediate-level processing: low-level features are used to parse the visual scene. Local orientation is integrated into global contours (contour integration); local visual features are assembled into surfaces, objects are segregated from background (surface segmentation), surface shape is identified from depth, shading and kinematic cues. The highest level concerns object recognition. Once a scene has been analyzed by the brain and the objects have been recognized, the objects can be associated with memories of shapes and their meanings.

Single-unit Recording and the Receptive Field Concept

We will base much of our functional description of the visual system on the activity of cells found in the brains of nonhuman primates. Why use the nonhuman primate model? First, it possesses perceptual capabilities comparable to those of humans. Second, it enables high spatial and temporal resolution in neural measurements, as well as targeted causal perturbations to investigate brain circuits—methods that are not feasible in humans. Third, nonhuman primates share close evolutionary proximity to humans, with well-established homologies across brain areas.

The concept of the receptive field (RF; Charles Sherrington, 1906) is central to understanding how visual neurons respond to stimuli. A receptive field refers to the specific region of the visual field where a stimulus must appear to evoke a response from a given neuron. A given stimulus typically elicits the strongest response from the centre of the RF, with the response gradually declining as the stimulus is presented further away from the centre of the RF. Thus, the RF can be well described by a two-dimensional Gaussian distribution

RF properties are typically determined using single-unit recording techniques in awake, behaving animals. In these experiments, fine-tipped electrodes are inserted into the cortex to record extracellular action potentials from individual neurons while visual stimuli are presented at controlled locations and times. This technique allows precise mapping of RF location, size, and tuning properties.

Receptive fields become progressively larger and more complex along the visual hierarchy. At each level of the hierarchy, receptive fields (RFs) encoding foveal stimuli are smaller and more numerous than those encoding peripheral stimuli. This phenomenon is known as eccentricity dependence.

In the primary visual cortex (V1), neurons form a retinotopic map, where adjacent neurons correspond to adjacent regions in the visual field. Cortical magnification further enhances processing in the fovea, where visual resolution is highest.

Columnar Organization and the Ice Cube Model

The primary visual cortex (V1) exhibits a remarkable columnar organization, where neurons with similar functional properties are grouped into columns that span the cortical depth. Orientation columns consist of neurons that respond preferentially to the same stimulus orientation, while ocular dominance columns alternate inputs from the left and right eyes. These two systems intersect orthogonally, forming repeating functional units known as hypercolumns. Hubel and Wiesel proposed the “ice cube model” to describe this architectural layout. Each hypercolumn, approximately 1 mm^2 in size, contains a full set of orientation columns for both eyes, along with blobs—peg-like regions prominent in superficial layers 2 and 3 of V1—that are specialized for processing color information.

In this model, the visual cortex is conceptualized as a grid of hypercolumns, each corresponding to a specific region of the visual field. These hypercolumns function like individual tiles that, when assembled, create a complete representation of the visual scene, with each tile encoding a distinct spatial location. This modular organization enables parallel and efficient processing of visual features such as orientation, spatial frequency, and color.

From Retina to V1: Early Visual Processing

Phototransduction in the retina initiates visual processing, as photoreceptors convert light into neural signals. At the level of the photoreceptors, the visual scene is encoded as a mosaic of light intensities—essentially a pixel-like representation. However, this raw format is not directly useful for encoding behaviorally relevant information and must therefore be progressively reformatted along the visual system. Retinal ganglion cells (RGCs)—the output neurons of the retina whose axons form the optic nerve—perform the first major transformation by encoding local contrast through center-surround circular receptive fields. These RGCs project to the lateral geniculate nucleus (LGN), which maintains distinct magnocellular, parvocellular, and koniocellular pathways.

LGN neurons preserve the concentric center-surround organization and also have circular receptive fields, in contrast to the elongated receptive fields characteristic of cortical neurons in V1. It's important to note that ganglion cells respond to small spots of light and are not selective for the orientation of lines or edges—a feature that first emerges in V1.

Contrast Sensitivity Function

The Contrast Sensitivity Function (CSF) characterizes an observer's sensitivity to sinusoidal gratings across a range of spatial frequencies. It is typically measured through a contrast detection experiment, where the minimum (threshold) contrast required to detect gratings at different spatial frequencies is determined.

In humans, visual sensitivity to sinusoidal gratings peaks at spatial frequencies around 5–8 cycles per visual degree, where contrast detection is most effective. Sensitivity declines for both higher spatial frequencies—approaching the limits of visual acuity at around 30–50 cycles per degree—and for lower frequencies below 1 cycle per degree.

This tuning to spatial frequency can be explained by the properties of neurons at early stages of visual processing, such as retinal ganglion cells (RGCs) and lateral geniculate nucleus (LGN) neurons. These cells have center-surround receptive fields that act as band-pass filters, making them naturally responsive to specific ranges of spatial frequency and explaining the shape of the behavioral sensitivity curve.

The response of an individual neuron to a grating can be modeled by multiplying the grating’s spatial intensity profile with the receptive field's spatial sensitivity profile and integrating over space—a process equivalent to convolution. This determines how well the spatial structure of the stimulus matches the filter properties of the cell, and thereby predicts the strength of the neural response.

Primary visual cortex (V1)

Signals from the lateral geniculate nucleus (LGN) project to the primary visual cortex (V1), where neurons are organized into layers and columns (see above) and exhibit specialized response properties. Simple cells have elongated receptive fields and respond selectively to specific orientations and phases of edges or bars, with clearly defined ON and OFF subregions. Complex cells pool inputs from simple cells, integrating across position and phase to provide positional invariance—responding to oriented features regardless of their exact location within the receptive field. End-stopped cells (also known as hypercomplex cells) respond to specific lengths or curvatures of stimuli, making them well-suited for detecting corners, endpoints, and more complex edge structures.

The current model for interpreting the function of simple and complex cells suggests that they behave like Gabor filters (Carandini, 2006) —a mathematical model widely used in image processing and computer vision. A Gabor filter consists of a sinusoidal wave (representing spatial frequency) modulated by a Gaussian envelope (which localizes the response in space). This makes Gabor filters well-suited to detect localized spatial features such as edges, bars, and gratings. In computational terms, the response of a simple cell can be modeled as the convolution of the input image with a Gabor filter. Convolution is a mathematical operation that slides the filter (kernel) across the input image and computes a weighted sum of overlapping values at each position, effectively highlighting specific patterns or features. This operation closely mirrors the way simple cells detect oriented features at specific locations in the visual field. Complex cells, in turn, can be modeled as combining the outputs of multiple Gabor filters in a non-linear fashion, producing responses that are invariant to small changes in position or phase—key for robust visual perception.

Feedforward Hierarchical Model

According to the hierarchical model proposed by Hubel and Wiesel (1962), the receptive fields of simple cells in the primary visual cortex are formed through the convergent input from LGN neurons whose receptive fields are spatially aligned in the visual field. These aligned inputs give rise to the elongated, orientation-selective receptive fields characteristic of simple cells. In turn, complex cells are thought to emerge from the convergence of multiple simple cells that share similar orientation preferences but vary slightly in position. This arrangement allows complex cells to maintain orientation selectivity while achieving invariance to the exact position or phase of the stimulus—an essential step in building more abstract representations of visual features.

Cross-Orientation Suppression Reveals Limits of the Classic Feedforward Model

The classic feedforward model proposed by Hubel and Wiesel provides a foundational framework for understanding visual processing in V1. In this model, simple cells receive aligned inputs from LGN neurons, and complex cells pool inputs from simple cells with similar orientation preferences. According to this view, a complex cell’s response is largely driven by the presence of its preferred orientation, and should be unaffected by orthogonal stimuli, which do not activate the relevant feedforward inputs.

However, this model is incomplete. Experimental evidence shows that a V1 neuron—especially a complex cell—responds strongly to a grating at its preferred orientation, but its response can be suppressed when a second, orthogonal grating is superimposed. Notably, the orthogonal grating alone does not excite the neuron. This phenomenon, known as cross-orientation suppression, cannot be explained by the feedforward model alone. Instead, it points to the role of additional, nonlinear mechanisms, such as divisive normalization. In this process, the response of a neuron is divided by the total activity within a local neural population. Even if the orthogonal stimulus does not directly drive the neuron, it increases the overall activity in the surrounding network, leading to inhibitory effects. This type of computation likely involves lateral and/or feedback inhibition and reflects recurrent network dynamics.

Thus, cross-orientation suppression highlights that V1 neurons are: influenced by contextual and global stimulus properties; embedded in recurrent cortical circuits; shaped by nonlinear and inhibitory interactions. This challenges the notion of V1 as a simple feedforward stage and underscores its role as part of a dynamic, context-sensitive processing system.

Beyond V1

In area V2, neurons begin to integrate local features into more complex visual representations, extending beyond the basic edge and orientation selectivity seen in V1. Notably, many V2 neurons respond to illusory contours—boundaries that are perceived despite the absence of a corresponding luminance change—suggesting that visual perception involves inferred structure, not just direct input.

V2 is also involved in border ownership coding—the process by which the brain determines which side of an edge belongs to the figure and which to the background. This is crucial for figure-ground segregation, enabling the perception of coherent objects within a scene.

Despite its clear involvement in more abstract visual processing, there is currently no single, widely accepted functional theory of V2. One reason is that no simple response property reliably and consistently distinguishes V2 neurons from those in V1 using traditional stimuli like gratings or bars.

However, recent findings have begun to reveal more distinct functional differences. In particular, studies using naturalistic texture stimuli—designed to replicate the higher-order statistical dependencies present in real-world images—have shown that V2 neurons respond more strongly and selectively to these patterns compared to V1 neurons. These responses highlight V2's role in encoding texture structure and suggest that V2 is more tuned to statistical regularities in natural scenes, potentially supporting mid-level vision tasks such as surface segmentation and material perception.

Visual area V4 serves as an intermediate stage in the ventral stream. It contributes to figure-ground segmentation, and its neurons are selective for curvature and color. V4 also contains spatially organized domains (globs and interglobs) that differentiate hue and form processing. Some V4 neurons respond robustly to partially occluded shapes, revealing integration of bottom-up sensory data with top-down expectations.

What is an Object? Object Recognition and its Relevance

The visual experience of the world is fundamentally object-centered. An object is typically defined as a bounded, coherent unit in space and time, comprised of visual features such as shape, color, and texture, integrated into a perceptual whole. This feature integration is governed by Gestalt principles such as proximity, similarity, good continuation, and closure.

Object recognition is essential for survival. It allows organisms to distinguish food from predators, tools from obstacles, and friends from foes—guiding decisions and behavior.

Selectivity and Invariance in Object Recognition

A central computational challenge in object recognition is achieving a balance between two seemingly opposing demands:

Selectivity: the ability to generate distinct neural responses to different objects, allowing the visual system to discriminate among them.

Invariance: the ability to maintain stable representations of the same object despite variations in viewpoint, position, scale, and illumination.

Effective object recognition requires both—fine discrimination between different items and robust generalization across visual transformations.

This balance is accomplished through hierarchical processing along the ventral visual stream. As signals move from early visual areas (like V1 and V2) toward higher-order regions (like V4 and IT cortex), neurons exhibit progressively larger receptive fields, greater tolerance to transformations, and increased selectivity for complex, behaviorally relevant features.

This progression allows the brain to build abstract, invariant object representations from simple visual features, enabling reliable recognition in the dynamic, cluttered environments we navigate daily.

Object recognition at different levels:

- Categorization: assigning an object to a general class (e.g., animal, car, tree).
- Identification: recognizing a specific instance within a category (e.g., your own cat).

Human vision excels at categorization, which is often automatic and rapid, while identification may be more effortful. In contrast, computer vision systems traditionally find identification simpler due to reliance on template matching.

Object Recognition in the Inferotemporal Cortex

While early stages of the ventral stream—such as V1—are relatively well understood in terms of their encoding of low-level features (e.g., orientation, spatial frequency), the computations performed in higher-order areas like V4 and especially inferotemporal (IT) cortex remain less well defined. Nevertheless, a growing body of research underscores the central role of IT cortex in high-level object recognition.

IT neurons are characterized by large receptive fields that always include the fovea, the region of the retina responsible for fine visual discrimination. These receptive fields often extend across both visual hemifields along the vertical midline, and their size allows neurons to generalize across object location. Unlike earlier areas that encode simple visual features, IT neurons respond to complex visual patterns—such as faces, hands, tools, or animals—demonstrating selectivity for semantic categories rather than mere physical attributes.

Importantly, these neurons exhibit a degree of invariance: their responses are robust to changes in object size, position, pose, and illumination. This invariance emerges through hierarchical convergence from lower visual areas (V1 → V2 → V4 → IT), where representations become increasingly abstract and tolerant. Within IT, neurons in anterior regions (AIT) tend to show greater selectivity and invariance than those in posterior IT (PIT), reflecting a gradient of abstraction along the posterior-anterior axis.

Neuronal activity in IT peaks approximately 100–150 milliseconds after stimulus onset, consistent with a feedforward processing cascade through the ventral stream. This timing suggests that core object recognition can be accomplished rapidly, using the first spikes that reach IT.

Furthermore, IT responses are experience-dependent. Neurons can develop selectivity for novel categories or refine their tuning with training and exposure, supporting category learning, perceptual expertise, and conceptual abstraction. This plasticity reflects IT's critical role not only in recognizing known objects but also in learning and encoding new ones.

Notably, some IT neurons show correlations with higher-order cognitive functions such as decision-making, attention, and working memory—particularly in tasks like delayed match-to-sample. This blurs the line between perception and cognition, suggesting that IT representations are not limited to passive encoding of visual stimuli, but are meaningful, task-relevant signals integrated with broader cognitive processes.

Coding Strategies in Object Recognition: Local vs. Distributed Representations

Two major hypotheses have been proposed to explain how the brain encodes object identity:

Local coding (or the “grandmother cell” hypothesis): suggests that individual neurons represent specific objects or concepts. Under this view, a single neuron might fire only when you see a very specific and meaningful object, such as your grandmother—or Jennifer Aniston. Some neurons in the IT cortex are so selective that they appear to function as "gnostic units", responding only to highly specific stimuli, such as a particular face or object.

Distributed coding: proposes that object identity is represented by the pattern of activity across a population of neurons, rather than by any single cell. Each neuron contributes to multiple object representations, and each object is encoded by many neurons.

While striking examples of highly selective neurons have captured attention, the weight of empirical evidence supports the distributed coding model. In practice, the brain uses sparse but distributed representations—where a relatively small subset of neurons responds strongly to any given stimulus. This coding scheme offers several advantages such as flexibility in representing novel combinations of features, robustness to the loss or variability of individual neurons, and generalization across similar inputs.

Thus, although some neurons may exhibit extremely high selectivity, object recognition in the brain is best explained as an emergent property of distributed population activity.

Core Object Recognition and the Computational Role of the Ventral Visual Stream

Core object recognition refers to the brain’s ability to rapidly (~200 ms) and accurately identify objects, even when they undergo identity-preserving transformations—such as changes in position, scale, viewpoint, or background context. Primates excel at this task, reflecting the brain's remarkable capacity to extract stable object representations from highly variable visual inputs.

This capacity is thought to emerge from the hierarchical architecture of the ventral visual stream (V1 → V2 → V4 → IT), which progressively transforms raw visual signals into more abstract and invariant representations. According to DiCarlo et al. (2012), this hierarchy "untangles" object identity from confounding image variability through a cascade of linear-nonlinear (LN) transformations.

At each stage, each object view elicits a unique pattern of responses, or a response vector, in a high-dimensional neural space. When an object transforms (e.g., rotates or scales), the neural population shifts to a new response vector. Together, these vectors trace a low-dimensional surface in this space—a structure known as an object identity manifold.

The computational goal of the ventral stream is to reshape these manifolds such that they are compact (minimizing variability for a single object), and well-separated from manifolds of other objects

While early visual areas (e.g., retina, LGN) act as pointwise spatial sensors, their representations are highly tangled—akin to raw pixels. Primary visual cortex (V1) begins to untangle these inputs via nonlinear transformations and a roughly 30-fold increase in representational dimensionality. This creates an overcomplete population code, where object representations become more spread out, yet V1 alone remains insufficient for solving real-world recognition problems.

What improves recognition beyond V1 is the progressive application of local computation through successive cortical areas (V2, V4, posterior IT, anterior IT). As proposed by DiCarlo and colleagues (2012), the ventral stream can be conceptualized as an assembly line of modular processing units, each performing a localized version of the untangling task—a process they term “cortically local subspace untangling.” Although no individual subpopulation has access to the full untangling problem, the parallel and coordinated efforts of many such localized units collectively give rise to the powerful, general-purpose recognition capabilities observed at the top of the ventral stream hierarchy.

Neural Correlates of Object Recognition

A landmark study by Hung et al. (2005) demonstrated that object identity and category can be reliably decoded from the spiking activity of a small population (~300) of IT neurons, within a very short time window (<150 ms). Using classifier-based readouts (a regularized SVM classifier) they showed that these neurons carried rich, linearly decodable information about object identity—even for novel objects presented at different positions and scales.

This study is significant for several reasons:

- It provides evidence for rapid and explicit object coding in IT, in line with core object recosgnition behavior.
- It shows that IT representations are stable, accessible, and usable for guiding behavior without requiring additional complex transformations.
- It suggests that complex downstream processing is not necessary for basic object readout.

Such linear decodability implies that other cortical areas—such as the prefrontal cortex or hippocampus— can access and use object representations in IT directly for memory formation, attention, or decision-making. IT thus serves as a final, high-fidelity stage in the ventral stream, broadcasting object-related signals to the wider brain network.

Activity in mokey IT predicts object recognition in humans

In a follow-up Majaj et al. (2015) tested whether simple, learned weighted sums of neural activity from monkey IT cortex could predict human object recognition, and contrasted this with neural activity in earlier visual area V4. They recorded from ~128 neurons in V4 and ~168 IT, and evaluated how well neural responses could explain human performance across several (n=64) object recognition tasks. They explored mutiple hypotheses varying neural regions (V1, V4, IT), number of neurons, timing (temporal windows), and decoding methods. IT neurons provided much better predictions of human performance than V4 or V1 neurons.

Simple linear decoders of IT activity could match both the accuracy and pattern (consistency) of human recognition when extrapolated to the full IT population (~60,000 neurons). Consistency (model-human alignment) and performance were highest for IT-based models. The best temporal window for decoding IT activity was shortly after image onset (specific 100-

ms windows gave optimal results). V4 showed diminishing returns—its information was not as linearly separable for complex tasks.

Conclusion

Object recognition is a complex yet astonishingly efficient function of the visual system. It requires transforming raw, low-level visual features into abstract, invariant object representations through a cascade of neural computations. This transformation is carried out by a hierarchical, distributed, and largely feedforward architecture along the ventral visual stream, culminating in robust, linearly decodable object representations in the inferotemporal cortex.

Experimental and theoretical findings led DiCarlo et al (2012) to propose the parsimonious hypothesis that core object recognition—the rapid identification of objects within ~150–200 milliseconds—can be accomplished predominantly through a reflexive, feedforward cascade. In this framework, visual information is progressively untangled and re-encoded through stages of the visual hierarchy, and conveyed via a firing rate code across distributed neural populations. These representations are structured in such a way that simple weighted readouts can support rapid categorization, without the need for extensive feedback or iterative computation.

Crucially, this does not preclude the role of fast, local recurrent processing, such as divisive normalization, which likely contributes within each cortical stage. However, behavioral and neurophysiological data indicate that re-entrant, long-range feedback may not be necessary for the initial solution to core recognition tasks.

Understanding these mechanisms deepens our insight into biological vision and provides a powerful framework for advancing artificial intelligence. By emulating the brain's ability to balance selectivity and invariance through hierarchical processing and distributed coding, machine vision systems may approach the efficiency, speed, and flexibility of the primate visual system. A pivotal moment came in 2012, when the breakthrough performance of convolutional neural networks (CNNs) in the ImageNet competition demonstrated that machine vision systems could rival—and in some cases surpass—human-level accuracy in large-scale object recognition tasks.