Module 2

Lesson 01: Neural Mechanisms of Vision: Low Level Processing

@May 3, 2023

What is vision?

Vision is not simply representing what is outside, it's interpreting what is outside, in the environment. Vision is about recognizing what are the objects in the environment and where they are located - What and Where question. It belongs to the Sensory System. The agent therefore can respond and adapt → It is necessary for survival, in particular, for humans and primates (monkeys' visual system has many similarities with the human one and has been therefore studied in monkeys). We will focus on the neural properties of the single vision neuron cell.

In humans and primates - visual brains - Vision is the dominant sense, studied vastly and many regions of the brain respond to visual stimuli/input. We typically have many sensory information from different modalities (visual, auditory,…) and many objects originate different sensory modalities at once → conflict between these information obtained simultaneously (e.g. ventriloquism). When there is a conflict, usually even when you can localize the source of the auditory input, you believe what you see. Two different uses of vision:

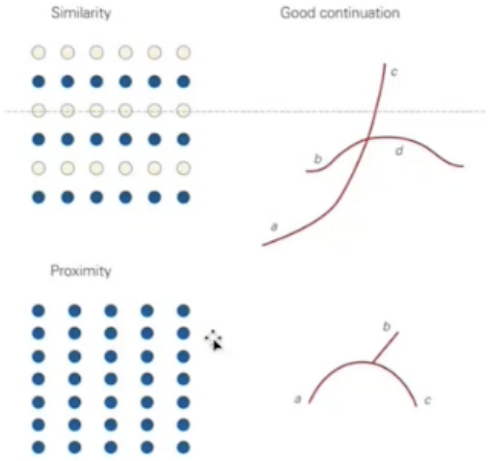
- Recognize and categorize objects
- Interact with objects → provide input to the motor system to interact/avoid/respond and act accordingly to the visual input

These are closely inter-related functions. In the brain they're parallel systems, so the functions can be processed separately and efficiently (respectively in the dorsal - action - and ventral - perception - visual pathways of the brain) and specialized.

In a camera, you just copy/replicate the outside world. In the Visual System, you don't merely copy the outside environment, but you interpret, understand it (e.g. the visual system is not good at identifying the amount of light in an image, but can recognize the wavelength, color), subdivide a scene into object and background, identify what is relevant through all the scene. Evolutionary pressure on the Visual System to be very fast in identifying only the

relevant details (Visual Attention). We have an emphasis on edges and contour and moving objects → recognize the shape and therefore the object to be safe.

Vision is not passive, it is an active and bidirectional process. We don't simply receive and analyze information (not a feed-forward system, but many bidirectional processes in vision). Not a tabula rasa in which information is added, but we have predictions on what could be in the environment. It is a constructive process, an inferential process. We have rules hard-wired already in the brain help us to parse the information. Rules of similarity and proximity, Good Continuation information → Gestalt (German for Form, Shape) rules for parsing the visual system information. We don't know if they are hard-wired or simply learned quickly in our infancy.

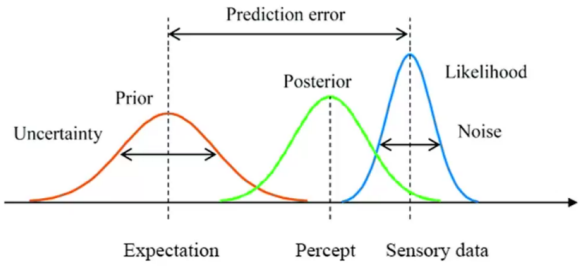


Representation of the Gestalt Rules

These rules are also important for attention. It is in fact much more easy to identify smoother lines because of contours. We like object, we prioritize whatever is object-like. Visual Priming → higher-order representations of shape (in memory) guide lower-order processes of surface segmentation (in other words, you're influenced by what you see first, one image primes the interpretation of the other). Semantic Priming → I give information about something and create an expectation to see that concept.

Bayesian Theories of Vision → based on Bayes's Theorem. They treat the visual system as an ideal observer that uses prior knowledge about visual scenes and information in the image to infer its most probable interpretation. The posterior probability of a possible real-world stimulus S (i.e. percept) is proportional to the product of the prior probability of S (that is, the probability of S before receiving the stimulus I, e.g. expectation) and the likelihood (the probability of I given S, i.e. sensory data)

$$p(S|I) = \frac{p(I|S)p(S)}{p(I)}$$



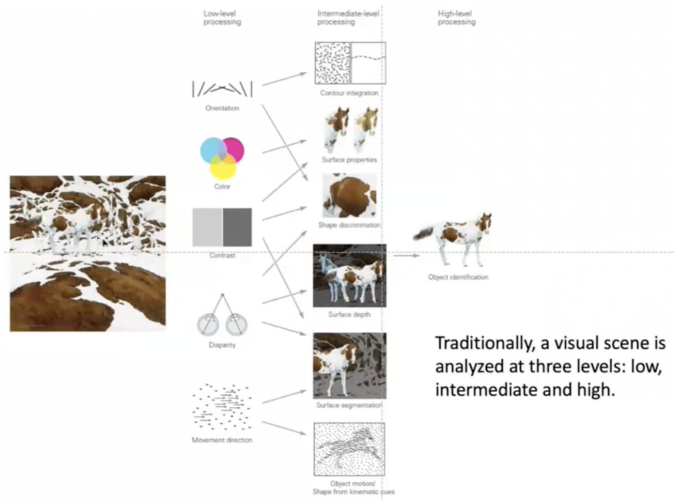
The posterior curve is obtained by integrating together Prior and Likelihood (red and blue)

As we're seeing, Vision is a constructive phenomenon.

Vision is subdivided in stages (also for didactic reasons):

- Early processing stage → we decompose the input in its elementary features and properties (visual primitives), e.g. color, lines and their orientation, contrast regions (white vs black - a jump between light intensities), depth (we recover it from ocular disparity) and motion.
- Intermediate Level Processing → orientations and surfaces patches are put together to obtain contours and so are color and contrast to obtain surfaces. Surfaces and contours are put together to obtain a shape
- High-Level Processing → a name is given to the object, we attribute it a category, a name. This is the true recognition phase to which follows a response action.

Approach of "Divide et Impera".

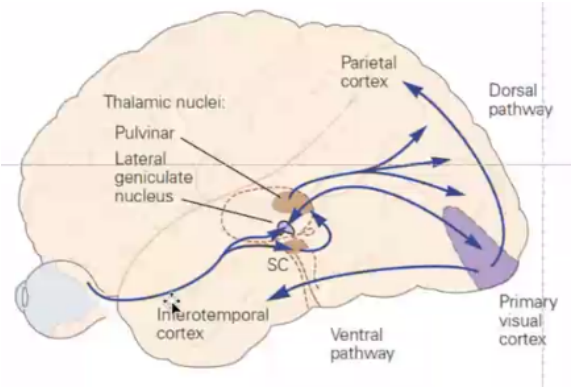


Every information need to pass through the retina.

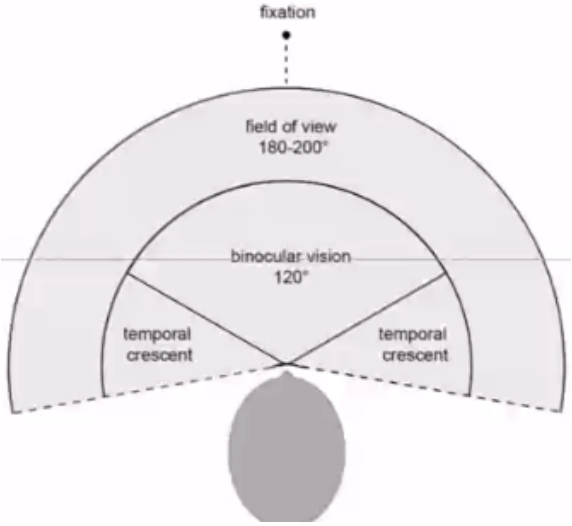
Neural Structure of the Visual System: General

The eye receives all the information and

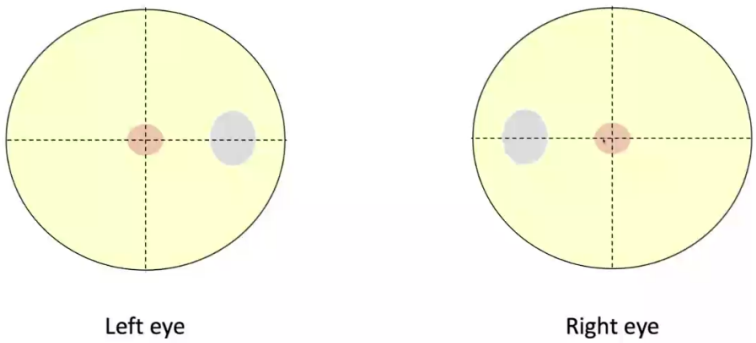
sets the limit of vision. You can't see better than what your eye allows you. What you lose in the eye is lost forever. From here the information is sent to the thalamus (Nucleus = neurons group together in a gray assembly- in the Thalamus Lateral Geniculate Nucleus LGN) and from here relayed to the Primary Visual Cortex (V1), the first cortical station in the brain. Then, information is broadcasted to all other visual areas in the brain



Imagine you're staring exactly in front of you, fixating in a point straight ahead. What you see is your Field of Vision: head directed straight ahead in front of you, fixed position of the eyes (Left and Right hemifields). The visual field is measured in Degree of Visual Angle (ideally more than 180° - but as our eyes are in an horizontal plane, we can't see more than 180°)



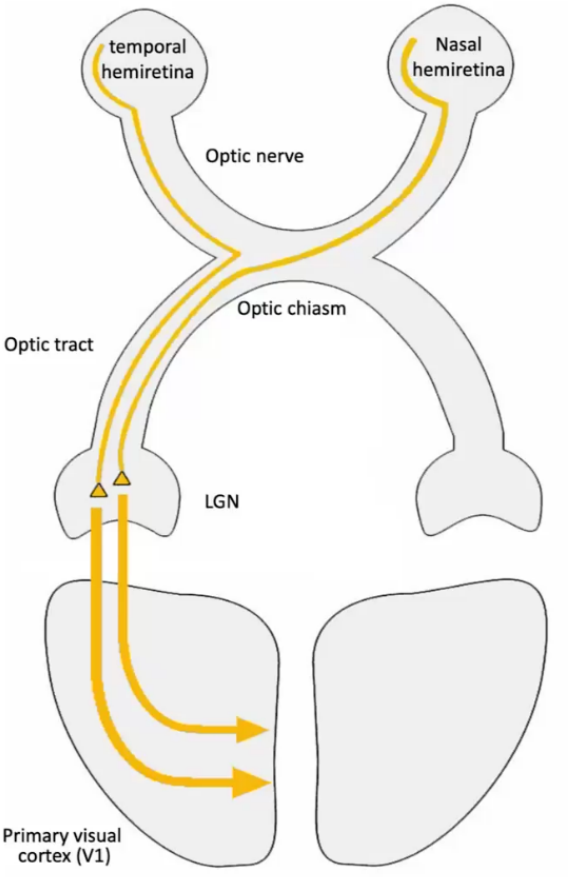
These are the eyes, left and right. In the center, this brownish area is the fovea, central part of the retina and the most sensitive, with its high number of photoreceptors. When you're fixating, you're locating the object on the fovea.



Left eye

Right eye

Consider the left part of each eye (the left hemiretinas in both eyes): they see exactly the same thing (the right visual field). Each hemiretina sees only half of the visual field.



Schematic Representation of the Visual System, with pathway of the right visual hemifield

There is a neuron going out from each one of the eyes, one to the LGN and one from LGN to V1. From V1 the information is sent everywhere (difficult to trace).

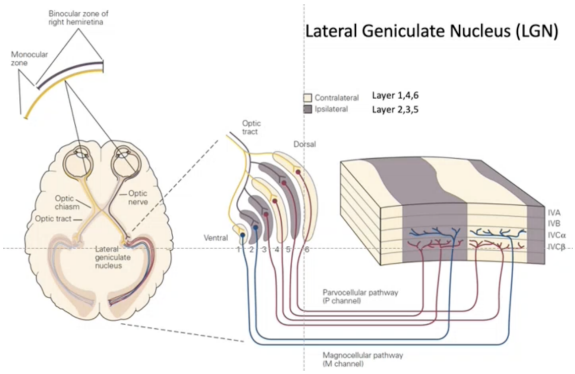
The hemiretinas both see the same thing simultaneously, so we have a complete full copy of the field from the two eyes. One copy remains on the same side - ipsilateral, the other crosses the midline (decussation) and goes on the other side. At the end, the LGN has a complete representation, as it receives the same information from both sides. Only one LGN sees one hemifield (left sees right), but in two copies.

Information is kept segregated - what comes from one eye is kept separated from the other - and continues to V1 (receives still separated information). Then eventually they will be integrated for visual processing analysis.

Destroying one LGN means losing the ability to see any information on the opposite visual hemifield (anopia).

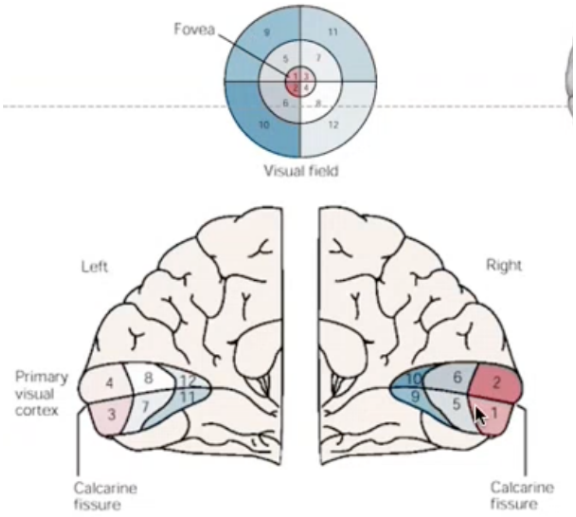
LGN is a layered structured (six laminas) inside the Thalamus. Alternation between dark - gray matter -and white matter (contains the axons). Two properties: information from the two hemifields is kept separated and don’t mix with each other. The controlateral sight is processed in

layers (1,4,6) different from the ipsilateral sight (2,3,5). Different wirings for (1,2) and (3,4,5,6). Layers 1,2 are the magno-cellular (large cells) pathway, the others are parvo-cellular (small cells) pathway. So there's a functional subdivision and, inside it, a separation of eyesights. This separation is kept also in V1 (Visual System is the largest sensory system of the brain). Up to layer 4, which is very huge in V1, we receive the bulk of information still segregated. Magno and Parvo cellular systems arrive in separate regions.



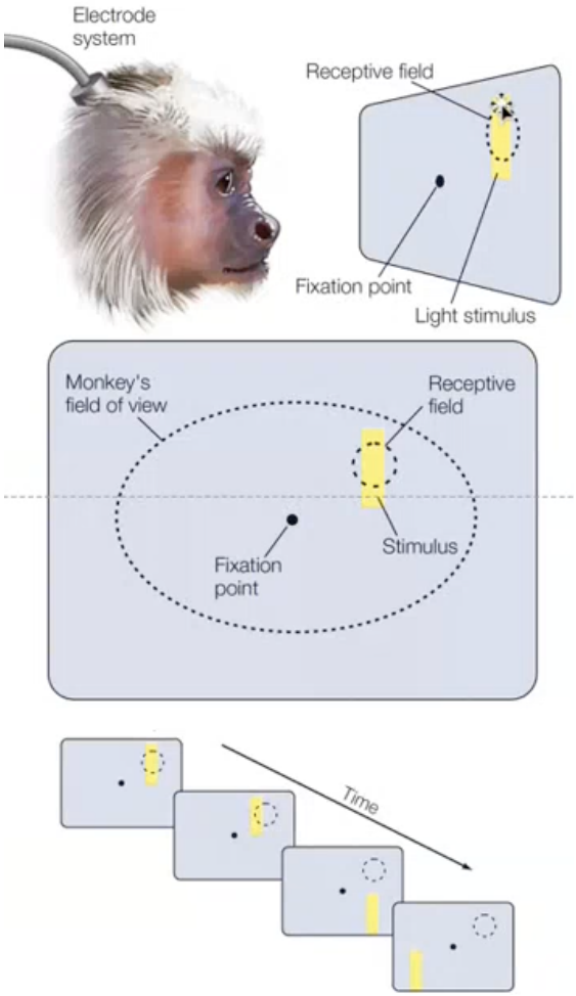
Right LGN sees the Left Visual Hemifield

Then we arrive in V1 in humans (Occipital portion of the brain). We have two visual cortices, linked with neurons that allow connection between fields. We have darker colors in the left visual field (right hemisphere) and brighter colors to the left hemisphere. The central section in red is the Fovea region, the region you're fixating in (the center). Everything else is the peripheral part. The visual field is topographically ordered (topographic map)→ closer in the visual field = closer in the brain. The brain preserves spatial relations.



To understand how this representation is obtained, we will focus on Single Neurons. We lower one of those fine-tipped, metal electrodes in the visual cortex and then the animal is presented a visual stimulus (a vertical bar of light) while fixated on the center of the screen. Crucial concept of receptive field of a neuron (= one area, location in the visual field in which when you put a stimulus inside you activate the

neuron whose receptive field is there). So you need the stimulus to be in the neuron's receptive field to activate the neuron. Visual field is what you see, inside it, each neuron has its own receptive field.

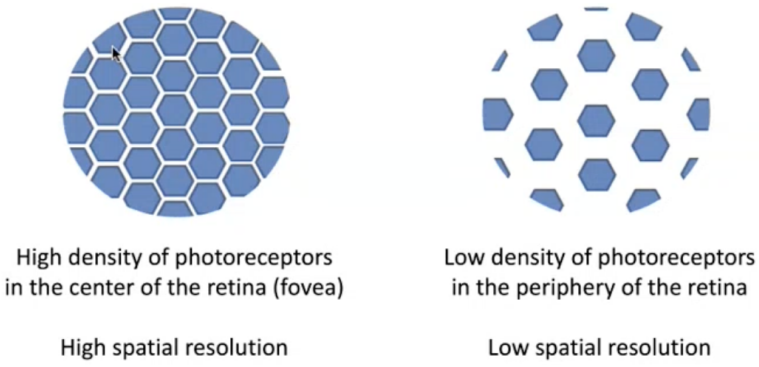


You can record the activity of a neuron → record the variations in action potentials, filter and amplify them, and also see the activity through an oscilloscope to study the properties of said neuron. Single-Unit is the most elementary meaningful entity. You can study the functional properties: which is the characteristic of the stimulus which is able to activate the neuron? Each neuron has its own receptive field. On average 30-40 spikes per second. When the cue is presented, the neuron fires. It cannot go less than 0 (lower bound: no action potential), the action potential is a code to which the neuron responds, with a fixed length. The highest possible is around 1000 (there are 1000 ms in a second), but it's uncommon, if not impossible. Stimulus has also a direction.

The more the stimulus is at the center of the receptive field, the more high is the probability of firing (gaussian distribution). The smaller the sampling area (receptive field), the higher is the resolution you have. Resolution is determined by the size of the receptive field, and it is the ability to see two stimuli as two separate dots (spatial resolution). This resolution is set in the retina, as the retina is the system where the photoreceptors have the smallest size.

The receptive field (important resource to see details) varies for location: each neuron has its own, to cover the entire visual field. Also, the size depends on two factors:

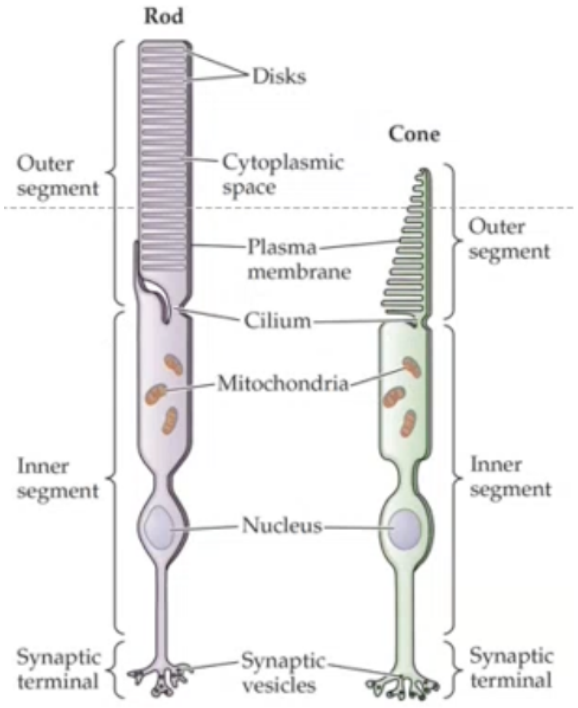
- Remember that Vision is a hierarchical system starting from the retina. The more you go up in the image hierarchy, the larger becomes the receptive field.
- Eccentricity → we are fixating the center of the scene, and what we're fixating on goes directly to the fovea (higher resolution). The more you move far away from the fovea where the eyes are fixed, the larger becomes the receptive field and more coarse the analysis becomes. And this increases with the levels of the hierarchy.



The same two stimuli can activate two different neurons in the fovea (which is more dense of neurons) and one single neuron in the periphery.

The Retina

The retina is not a single layer of neurons, it is an extroflexion of the brain, made of at least three layers. Photoreceptors have a graded response (it is graded based on the light intensity), they do not fire an action potential (either activated or not). We have two types of photoreceptors in the retina: cones and rods (from their shapes). The cones are all packed in the fovea, the center of the fovea region. Outside the fovea, there are almost no cones (this is why we move our eyes, so the image is sampled with the cones). The rods are outside the fovea and scarcely in the fovea.



During the day, we use both - the electromagnetic energy is transformed into neural code. Inside the photoreceptors there are many disks, inside these disks there are proteins associated to Retinol (Vitamin A) in two isomeric forms. When the light hits, it changes the sterical configuration and this triggers a cascade of transformations. The two photoreceptors are very different. Cones have low sensitivity for light (not active during the night as they need a lot of light to be activated) as they have few disks. During the night (scotopic vision) you use the Rods, that have a lot of disks (they can be enabled even by a single photon). When there is a lot of light, Cones. For nocturnal vision, Rods. During the day (photopic vision), Rods are so sensitive that are quickly saturated and do not contribute.

Also, we have a trichromatic vision. We have three types of cones, each sensitive to a different wavelength (visible spectrum - light we can see). Within the Visible Spectrum, the rods are all the same (we cannot see colors during the night, we see differences in intensity). Cones, on the other hand, are of three types, each associated to a different wavelength (long=red, medium=green, short=blue). You see colors by comparing the activity of the three cones to one another. This means that losing some of the types of cones makes you see black and white. Color is a construct.

The fovea of a normal subject has a high concentration of Long Cones, followed by Medium and a small quantity of Small. Protanopic retina only has Medium and Small (daltonism - cannot distinguish orange-red colors).

In the retina, we have one after the other (circuit) → photoreceptor → bipolar cells → Retinal-Ganglial Cell (RGC). Important, because they are the cells that exit the retina. It is important to understand what is the firing of this output neuron. We have 100 million of photoreceptors in our eye (majority of rods and 5 million cones), which are separated signals. These are compressed in 1 million axons which exit (RGC) → huge reduction in numbers, hundred-fold. There is a lot of editing and compression that happens.

The photoreceptors have the smallest receptive field possible. They’re in two states: active when there is light or inactive in the dark (there’s also the possibility of being in between).

The RGC are in three types, of which two are more common:

- P (Parvo) Cells → 70% of all.
- M (Magno) Cells
- K Ganglion Cells → function is still unknown

They are called Parvo and Magno as they supply the respective Lateral Geniculate Nucleus layers. Functional separation. What are the functional properties of the RGC?

Receptive field is made of two concentric circles, called Center and Surround. We have, regardless of the cell being Magno or Parvo, two types of RGC:

- ON-Center → at the time T1 we present a stimulus exactly to the center of the receptive field. The neuron is active.
- OFF-Center → at the time T1 we present a stimulus exactly to the center of the receptive field. The neuron is inactive (inhibited actually).

Presenting the dark shows the opposite behavior. Retina can represent situations which are brighter than the mean luminosity and darker than the mean luminosity. We have two separate channels, which make the response to stimuli faster and more precise.

The ON-Center is activated by light on the center. When I take out the darkness from the ON-Center, the cell responds. So the cell is dynamic: it signals when the light is on, but also when the light stops being off. And the same goes for the OFF-Center.

The Center and Surround are antagonistic, they are in competition with each other. If I illuminate the center, the neuron fires. If I illuminate the periphery, it is completely inhibited. When light is outside, it does not make any contribution. I can instead distinguish the inhibition when looking at a signal in the periphery. This emerges because if I simultaneously activate both the center and the periphery, I don't have any activity. The best stimulus is a stimulus which is relative to the neutral situation lighter at the center and darker in the surround (for the ON-Center). These cells are signaling the margins (the edges). The retina emphasizes the contrast between things in the environment (where the object begins and the background ends).

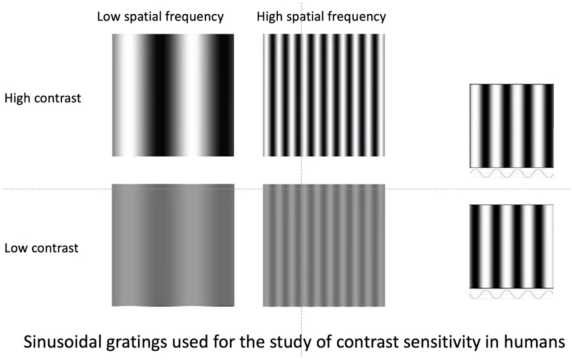
Other differences:

- Sustained response → the cell continues to respond at the same rate while the stimulus is inside its receptive field. In-depth analysis of a stimulus
- Transient response → the cell fires only at the very beginning, then it immediately stops responding. Important for dynamic changes in the environment

We have also a category of RGC (Parvo cells) that also respond to color oppositions (Red-Green and Blue-Yellow) and Achromatic Cells (Magno) that respond regardless of the color.

As a conclusion: RGCs are much different than photoreceptors, they do not respond to uniform stimulation, they do not respond to all-light/all-dark scenarios, they work better in opposition and emphasize the contrasts. They emphasize the object changing its movement along the retina. They don't care about the absolute brightness of the object, as it depends on the color of the object and its illuminance. We're interested in comparing, as the light in the environment changes continuously anyway.

Types of stimuli we can use to study the properties of RGCs in humans and animals.



Sinusoidal gratings used for the study of contrast sensitivity in humans

These are stimuli in which we can vary many characteristics of the stimuli (e.g. the main orientation, the spatial frequency, the contrast, polarity…)

We are not equally sensitive to all spatial frequencies, and this can be shown through some experiments made using sinusoidal gratings. Our sensitivity is higher for medium spatial frequencies, but we are bad for extremes. I can create a Contrast Sensitivity Function, an upside-down curve. We explain it through the antagonism of ON and OFF-Center in RGCs. We can transform their response into two gaussians each, one with a narrow base which is the Center, and one with a wider base which is the Surround. The difference between the two (the "Mexican Hat") is the neuron's response.

To conclude, we have two morphological types of cells, with complementary functions:

- Parvo cells → dense in the fovea, mostly receive input from the cones. They are many, small and so, they don't receive many convergence. Small receptive fields and dendritic trees. Slow (not necessarily) and sustained response. They also show color opponency, so they convey color. They though need light, as they have low contrast sensitivity. They respond to high spatial frequencies from small areas of the visual field (good for details)
- Magno cells → large receptive fields, common in the periphery. Completely achromatic, fast in recepting and with transitory responses. High contrast sensitivity (protect us in the dark)

The LGN cells are similar to RGCs.

The Primary Visual Cortex

The properties of neurons change substantially. Studies made by Hubel and Wiesel. Two general types of cells:

- Simple Cells → receptive fields become elongated in one dimension (asymmetric - still small, or even smaller). They become then sensitive to the orientation of the stimulus (as found out through sinusoidal gratings). It is an emergent property (couldn't foresee it from retina or LGN's cells, it originates in V1) - orientation selectivity. Tuning Curve tells us a cell is selective for one specific orientation - preferred orientation. Can also

have frequency and polarity preferences. Everything else is similar to RGC. We have a Center and a Surround antagonistic to each other

- Complex Cells → the stimulus is elongated, so there's a preference for orientation. Now, the polarity (bright vs dark) doesn't really matter anymore. We're in a higher level. We don't care about the location of the border → invariant way of encoding the border: we don't want the exact location. Step up in complexity: positional invariance.

According to Hubel and Wiesel, V1 cells are built starting from input at the previous stages: LGN cells convergence generate Simple Cells, and the convergence of those generates a Complex Cell. Hierarchical model of the receptive field.

Lesson 02: Neural Mechanisms of Vision: the Primary Visual Cortex and the Ventral Pathway

@May 10, 2023

The Primary Visual Cortex (Reprise)

Another property that emerges at the V1 Level is spatial disparity (depth).

In Complex Cells, the ON and OFF regions are not fixed, they're blended with each other and you can't isolate them. You still have a preference for orientation, but the stimulus' position can be changed and the cell is still activated (being inside the receptive field and respect the preferred orientation is enough) → positional invariance.

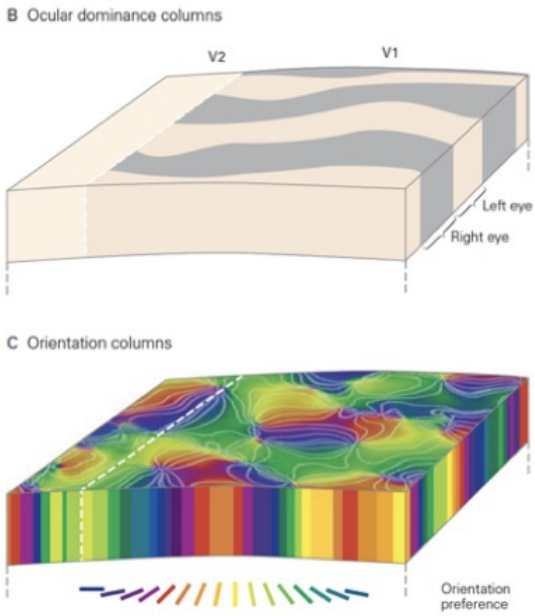
Image convoluted with the receptive field then rectified for simple cells (most common model). For complex cells there's more discussion: they could come from the integration of more simple cells (hierarchical model is still debated). Complex cells are important for the early stages of motion processes.

V1 are selective for a number of attributes: orientation, spatial frequency, direction, temporal frequency, disparity, color.

General Organization of V1

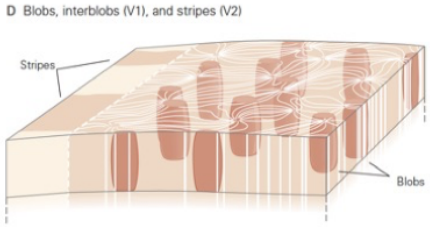
It is said that V1 is organized in columns (if you enter V1 with an electrode perpendicularly, you find they have the same preferred orientation). If you enter tangentially, you find that in 12-10 columns

you have a period (cover field of vision with steps of 15°). They are organized like pinwheels. Orientation changes around the center. An hypercolumn is a set of columns that cover all possible orientation = same position of the visual field.

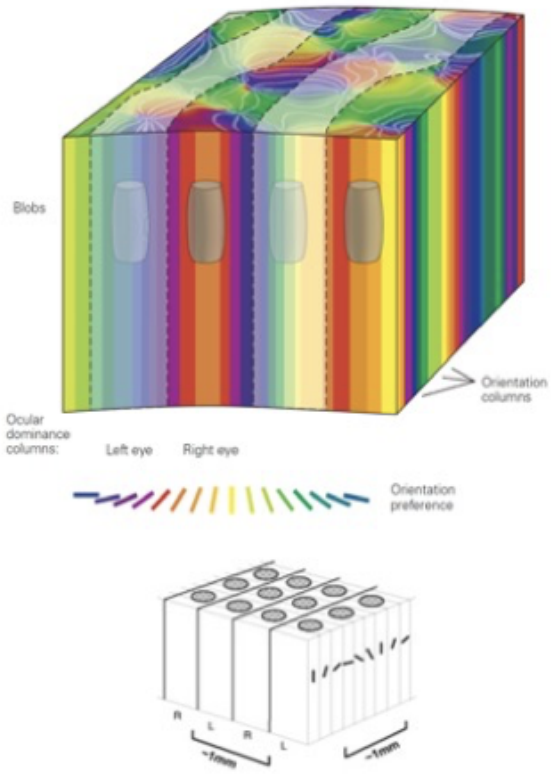


Ocular dominance column → in one column you may find that neurons have a preferred orientation, but fire more if the stimulus come from a “preferred eye” (there’s not an absolute all-or-nothing response).

Further subdivision based on color (necessary to perceive the difference between objects with the same intensity). In V1 we have clusters, called Blobs, with circular fields, that check for colors. Interblobs orientation specific.

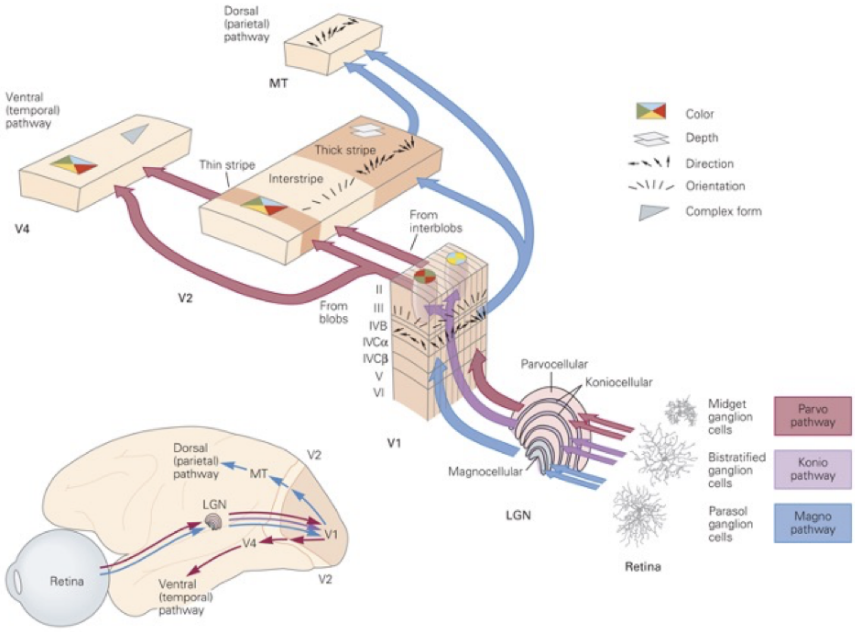


Eye Cube model of Hubel and Wiesel. Size of the cube is slightly less than 1mm. Covers the entire visual field like tiles. Fovea is magnified (Cortical Magnification). Remember, longer axons are costly, so areas that communicate often have to be close. This ordered structure is in Human and Primates.



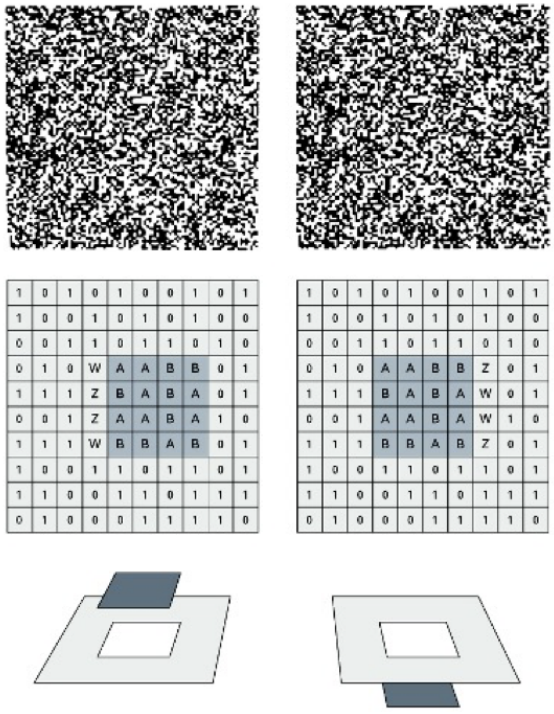
Organization of V2 and subsequent levels

Also in immediately subsequent areas (V2) the organization is maintained: areas associated with color, orientation (inter-stripe) and direction (thick stripe). V2 is interesting for its response to illusory contours (activated by something invisible, but perceived nonetheless). The organization then becomes more difficult to follow in subsequent areas, where signal is broadcasted to dorsal pathway (from V1 to parietal-frontal WHERE/HOW - faster) and ventral (to Infero-Temporal - WHAT - sluggish). Recognition and action are separated (there’s still a lot of integration, but there’s a slight desynchronization - you first know WHERE, so you can react accordingly (e.g. run from a predator), then after a while - order of ms, but measurable, you know WHAT was it). Dorsal is associated with the Magno Channel, Ventral with Parvo Channels.

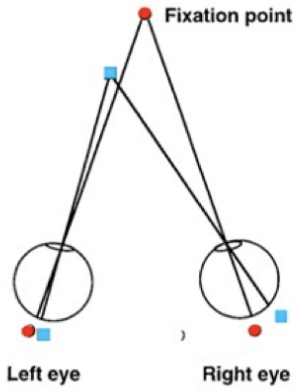


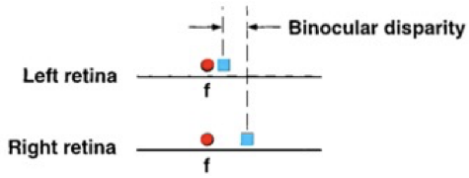
Stereopsis and Disparity

We have monocular information (e.g. shadows, interposition (objects that occlude others), objects changing size fast, movement - things that are closer move faster) → we have clues to extract a sense of distance and one eye is enough. In the case of humans, you have the same thing as two different cameras, where you don't see exactly the same thing, but more or less, and you integrate the two visions. Bela Julesz used random dots stereograms, one a copy of the other, where the central portion is switched, and when you observe, you perceive a square emerging.



Imagine you have two eyes and you're fixating a point, which falls on the fovea. Anything that is in front or behind the fixation plane won't fall on the same place in the left and right retinas. The two eyes

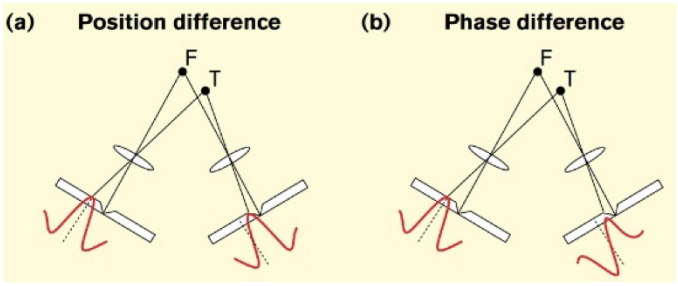




are recording a binocular disparity, as they have different vantage points. Disparity has to be computed at the level of V1 (first stage in which you can compare the information received from the two eyes - until fourth level of V1 they are separated). Until then, you are monocular. Stereopsis and Disparity cannot be computed no earlier than V1 (and beyond) - still, we have ocular dominance.

Two ways to compute this disparity, both likely to occur in the brain:

- Position Difference → neurons receive information about position of the stimulus in both eyes and calculate the coordinate difference between the two eyes.
- Phase Difference → the two receptive fields have exactly the same location, but the structure inside is different (phase difference). For simple cells in particular, we have a fixed structure, so it senses the slight difference (evidence that this is more likely)



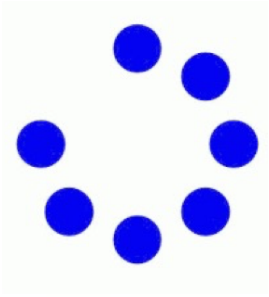
Is disparity equivalent to say depth? No, a neuron can respond to disparity without having a depth perception, at the level of V1. Disparity is used in V1 to adjust the eye movement (convergence and divergence to better observe).

Evidence that particularly V5, which is an area important for visual motion , is used for extracting depth from stimulus. V5 also contributes to Visual Motion, which is extremely important, especially for transient cells. Stimulus can move, we have direction selectivity in our neurons. Specialization for motion, we have an area associated. Visual Motion is important to segregate objects from the background (from the movement of the object we can

recognize the object itself). Also used to identify the depth/distance of an object (from its speed) - e.g. I want to grasp a moving object. Visual Motion grabs our attention - it is something hard-wired, our eyes go there. We are distracted by things that move, especially if they get larger (they're looming) → dangerous.

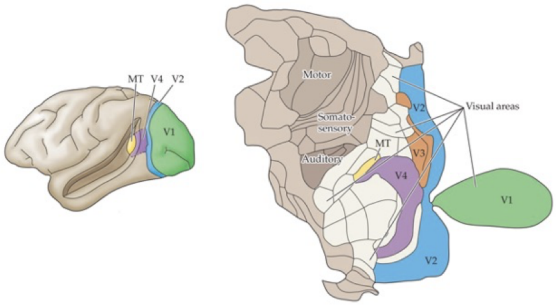
We can also appreciate illusory movement → alternation of illuminance (beta movement) or phi phenomenon. We have a mechanism. We perceive also implied movements (extract motion anticipations from some physical cues) and they activate the same areas used for motion perception.

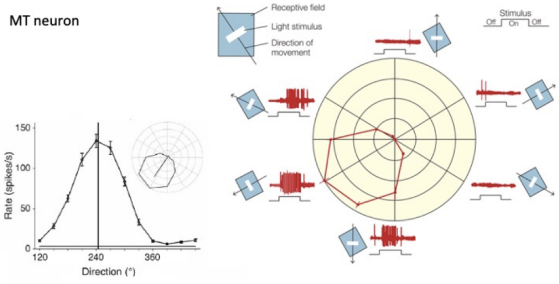
beta movement



phi phenomenon

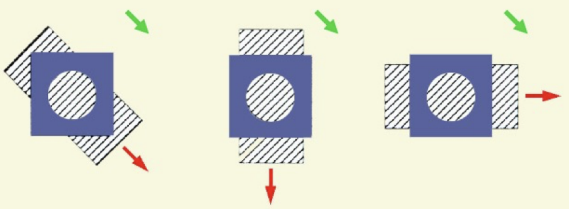
We have areas dedicated to motion, among which there is the crucial MT (V5 in humans) - Middle Temporal Area. What MT does differently from V1? In MT there are neurons similar to V1, in the sense that they respond to orientation (e.g. sinusoidal gratings) and direction of motion. Also selective to certain speeds. They have a columnar organization. They respond also to the non-preferred responses, with the smaller the standard deviation, the more selective the neuron (as can be seen in the Tuning curve). Random Dot Pattern → use black dots on white background which can move randomly (don't perceive movement) or a number can move coherently (there you see a motion). Can test the neuron with different types of stimuli → MT strongest

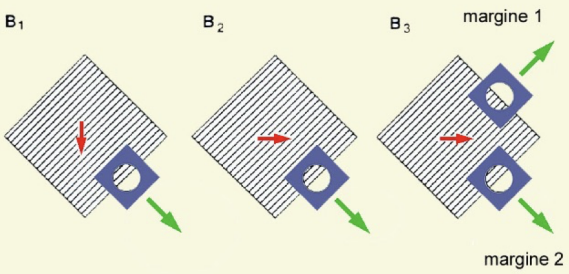




response for motion in general, and
specifically for the moving dot.

Problem: Aperture problem → occurs when you see things from a small opening (e.g. a hole). In V1 you make mistakes, due to the small receptive field. You only see movements that are orthogonal to the main margin of the bar. How do I solve? I read the activity of at least two receptive fields and integrate. So V1 would be fooled, V5 can integrate multiple cells from V1, so it can solve the issue.

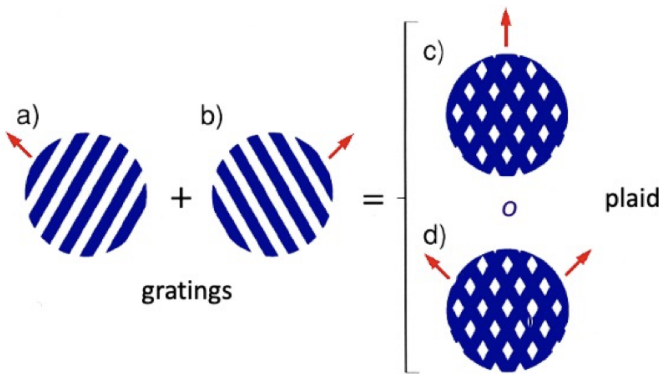




Visual motion processing occurs in two stages:

1. Single components are analyzed (Extract)
2. Integrate the single components into a pattern motion (Integrate)

Demonstrated through the use of Black and White Sinusoidal Gratings directed in the opposite directions. If we put them together one above the other, you obtain a pattern (PLAID) and you feel like there is one pattern moving up. V1 responds to the components, V5 to the whole pattern (and also to the components). The two-stage processing solves the aperture problem.



We have also neural responses to illusory contours. The illusions stop when the contours are weakened, e.g. added borders. V1 is immune to illusion (impenetrable by cognition). The more you go towards the higher level, the more your expectations influence what you see.

Ventral Visual Pathway

This is the pathway to Object Recognition. The behavior of Neurons is easily recognizable in V1, but reading single neuron behavior in AIT is more obscure. For color and shape, you deal with this in V4. Object Recognition happens in the Infero-Temporal Area

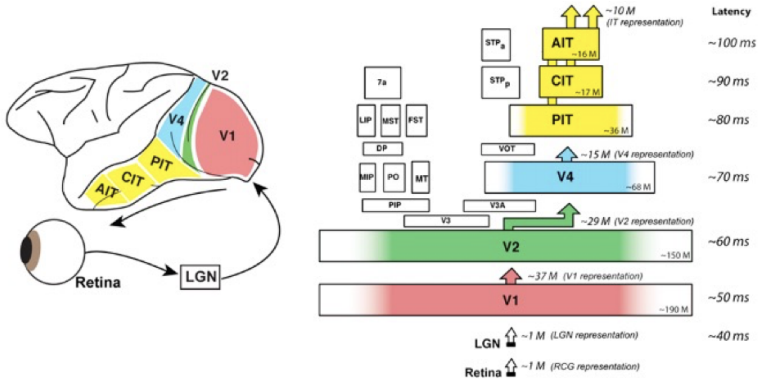
Visual objects are a set of visual characteristics joined together by their mutual respect of the Gestalt rules (principle by which we parse objects). Two tasks:

- Identification → I recognize one specific instance (easier to lose)
- Categorization → I classify

Major difficulty as we have opposite requirements:

- Selectivity → we want to recognize that specific object (e.g. my car from other cars). The object as it is seen by the retina, has an infinite number of projections on my retina. So I need to distinguish the different appearances.
- Invariance → I need to recognize the object regardless of shape, lighting, occlusion, details in general.

Vision is for both Perception and Action. We are very fast on Core Object Recognition → objects presented in a simple background, different perspectives or scale, illumination, take no more than 200ms. Recognition of conflict (incoherent presentation) slow the pipeline. We are also able to express a feeling in a slow time (e.g. recognize friends from enemies).

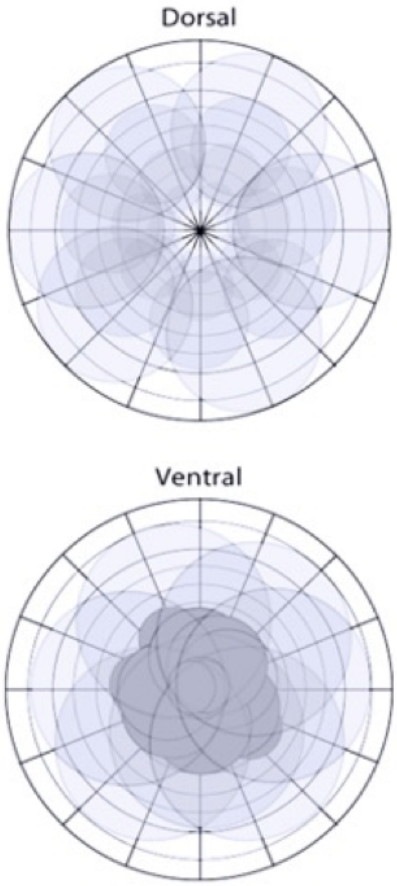


Infero-Temporal Cortex

Inferior Portion of the Temporal Lobe

The size of the receptive field of the cells in the ITC is much larger, more or less coincident with the entire visual field. Usually includes the fovea and it is more focused on it than the Dorsal Visual Pathway (which is

dedicated to exploration). They provide different information. As we move along the Visual Ventral Pathway, we have an increasing complexity of the stimulus (more complex reach of information to activate the neuron): e.g. a light is enough to activate V1, but not for AIT. The activation of the neuron is not guaranteed by the simple presentation of a simple stimulus: the stimulus must be complex. We have a column organization → similar properties activate the neuron. There's also a certain degree of invariance: if you change the size, position or presentation (e.g. polarity), you get more or less the same activation (similar Tuning Curves).



In the early '80s, a group of researches found neurons that responded selectively to faces. Tanaka goes for selective reduction: cut progressively some parts from an object to see if the firing is still there: decipher which parts of the stimulus are important to be present. What is represented in the cells is any meaningful stimulus (be it Jennifer Aniston or Halle Berry) - grandmother cell → Brain can represent gnostic units. Locally, neurons represent an entire concept. Probably there are no sufficient neurons to represent all concepts. Hypothesis seems unlikely as many neurons are involved in the representation of a concept. 10% of the neurons are really invariant to the perspective. More likely idea of a local, distributed representation, or a sparse representation (compromise).

Paper → idea that you can represent each object as a vector, whose dimensionality is the number of neurons that are activated (number of cells participating). You create an object manifold, as each different point of view creates its own vector. It is difficult for us to decode V1 as the manifold is completely tangled. Progressively, through selective analysis, you arrive to the ITC where you untangle separable representations and decode. Imagine ITC not as the last stage, but as a stop in a larger relay (another area capable to read what the ITC is sending). The researcher used a classifier that processed the input from ITC. Experiment: he presented to monkeys 8 categories of images on a grey background (for a total of 78 images). The animal is only required to fixate, it is passive (if the animal participates - tries to

recognize the object, the effect is much larger). 5 stimulus per second, but is not doing anything. Recordings across months from 2 monkeys, 300 hundred penetrations in the ITC. We have around 350 multi units, we sample a certain number of sites, you take out the response of the neurons and use to create some vectors, which will be the training and test of the classifier (Linear Regularized Classifier - 1 vs all). Used different bins with different latencies → 100ms after the stimulus presentation we have the most significant activity. Classification performance was much higher (easier task), we can change randomly the number of sites and see that the performance is linear (more sites = more performance).

Testing section → presented stimuli not presented at all during training or presented at different position/size. Overall performance very good. Impossible to see this effect in V1.

Lesson 03: Neural Mechanisms of Selective Visual Attention

@May 17, 2023

One of the central topics in Cognition, influence the memory, control, language.

Selective attention is actually everywhere: Visual, because for us is the most dominant sense, we trust it more than other senses. We, of the many visual information we receive in the retina, we don’t process all, but just a portion of it, the one relevant to our behavior. We pay attention to only a portion of all the possible information we receive. Why? The brain has billions of neurons, but a limited capacity for processing: the ability to process information is limited, we need to choose; we can’t represent completely and perfectly all the information that hit the retina. Since activating neurons is quite a costly problem (requires energy), we activate only the relevant neurons. Attention is a vital function for our survival.

One typical example is e.g. Change Blindness: we are blind to change. Example of the two photos changing and repeated in a loop: because the blank occurs almost simultaneously with the change of image, we do not immediately perceive the change due to the disturbance - so we see the leaf the second image adds only after a careful observation. The blank is a distractor, it steals attention from the image, so our resources are not sufficient to deal with the change.

We can process only one object at a time (completely) in our visual field. Most common theory: Biased Competition Model of Selective Visual Attention: at each moment, every object that hits the retina (the neural representations of the objects) competes in order to get resources, because only the object that “wins” the most resources is processed fully, else it will not arrive at the level of awareness. So we have competitive interactions between

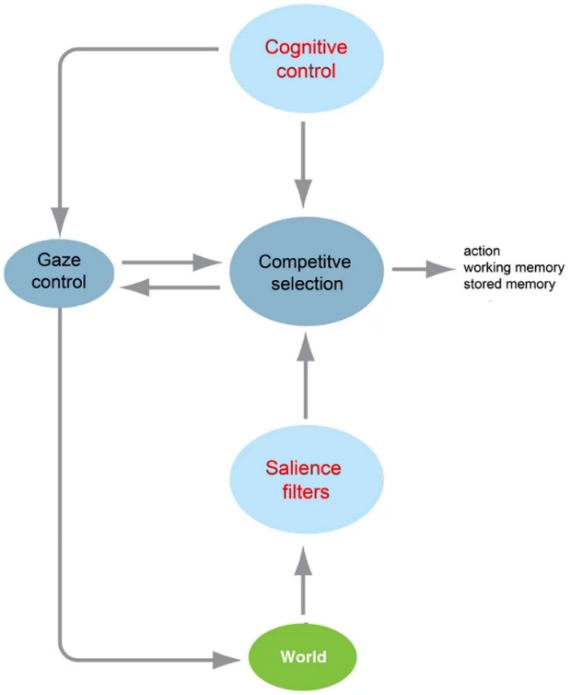
neurons (the representations of the objects). In order to alleviate the competition you need attention, as it biases: the objects that are relevant to our behavior get a winning advantage.

The competition is not always the same and mainly involves the higher-level areas. In Vision: in all areas there is competition, the degree of it may vary.

This competition is biased by attention, and this bias can arrive from:

- Below → bottom-up biases: obtained by saliency filters (e.g. they are brighter, moving, different…). Objects become relevant because they capture attention. Evolution has seen that these objects receive more attention as they might be important. Just depends on the environment
- Up → Cognitive Control. This depends on the agent (strategy, motivations…): the agent can select voluntarily a behavior. This is regardless of the saliency.

The two biases can work together or antagonistically depending on the situation.



Once I select the object I can move towards the object, it is more likely to be remembered and it will guide our gaze (move eyes), gain awareness about the environment.

We have dichotomies in Selective Attention:

- Top-Down and Bottom-Up: these are the two ways to select information.
- Space-Based (where they are located) vs Object-Based (non-spatial properties of the object, e.g. color, shape)
- Covert (attention can be moved without moving the gaze) vs Overt (attention and eye movement are closely linked) → one can perceive or not the presence of the attention

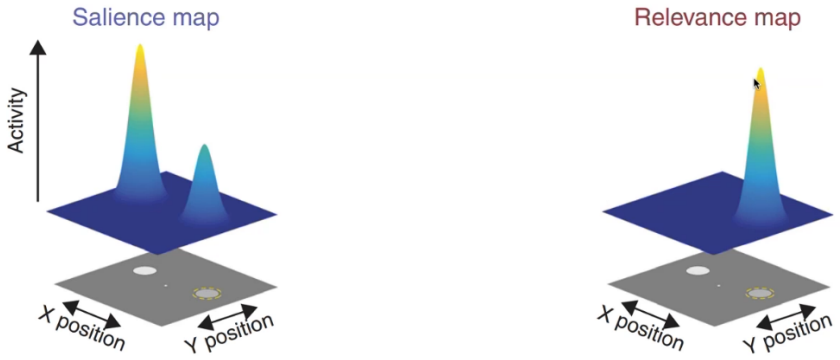
Top-Down vs Bottom-Up Selective Attention

Top-Down or Voluntary → I deliberately want to select that particular object, because it is important for my behavior.

Bottom-Up (automatic, given by the stimulus and not by intention) → something strikes to me because it is completely different from the entire background (e.g. the red bird in a green background). The object is relevant for me because it is salient, so I will pay attention even if it's not relevant to my current behavior. Driven by the characteristics of the stimulus.

Saliency Map → shows the saliency of the objects (which will grab more attention), regardless of what the agent wants to do

Relevance Map → shows which object is relevant for my behavior: signals what is relevant for the agent.



The two ways Attention can be assessed

Attentional Template → working memory is a temporary/short-term memory (it's a temporary storage for what's a relevant information in a current behavior, for dealing with the environment), kinda like a pc's RAM. Working memory is closely associated with voluntary attention (top-down). We create in our mind a representation of the object (the attentional template) which we use to give a competitive advantage to matching objects in the environment, even if they are not salient.

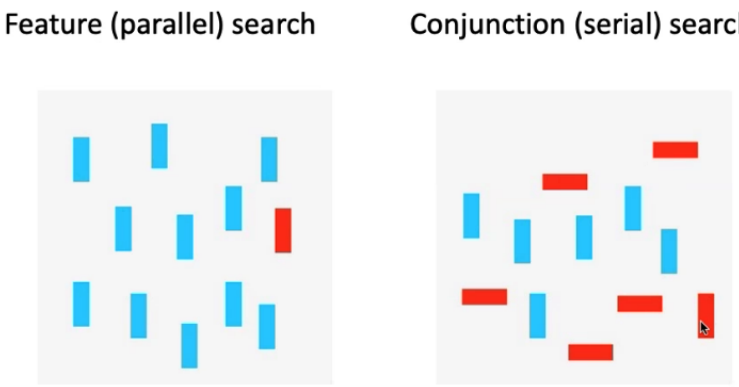
Once you have the attentional template in the working memory, you need to send it to all visual areas back in the brain; you need to send feedback, so that the relevant neurons can become aware (they are pre-activated by this feedback). Through the Attentional Template we create the bias in the Attention Competition.

Two tasks we can use to examine attentions (which have numerous variations):

- Cuing paradigm → way to study top-down attention for spatial information. You've been asked to fixate the center of the screen. After a while, over a cross appears an arrow (pointing right or left) - the cue. After a moment, a target will appear (oriented sinusoidal grating - gabor stimulus) and one has to respond by pressing a key as soon as the gabor appears. Reaction time is recorded. The arrow prepares us to look where the gabor will arrive. The cue is informative (likely that the cue will appear on the side it is pointing to - 80% of the cases), so if it is like that, one will be faster in pressing the key (one is faster

in recognizing something they anticipate). Gives us information on where the stimulus will appear (space-based attention)

- Visual Search Paradigm → Based on Visual Features. In a moment, you will see a green screen and a stimulus will appear. Press a bar if you see a red vertical bar, else you tell me it is absent. The stimulus is present in 50% of the cases, the manipulation is about distractors (crowded visual field or there are few). Now the reaction time does not depend on the number of distractors, as the bar is salient and so it shines, you are able to detect it immediately regardless of the number of distractors. It is an automatic task.

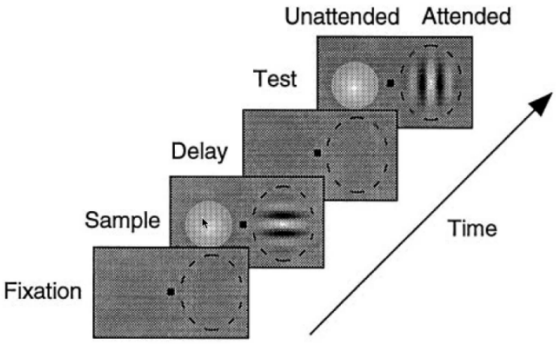


In the second case, the distractors (which can still vary in number) share their color with the target; as a consequence, the number of distractors becomes important as you need to stop and discriminate, because you need to integrate the color information and the orientation information.

As a conclusion, in the conjunction search, the efficiency depends on the number of distractors, as they are competing with our target for attention. Serial, as you proceed one at a time to find the target. In the parallel search/feature search, the target has a property that makes it salient w.r.t. the distraction.

Difference between attended and unattended objects: an experiment

Recording of single neurons from either V1 or V4 (ventral pathway area) across many months. The animal is doing a cuing experiment: it is fixating the screen on a central dot (the gaze cannot be free), and we have a receptive field indicated with a dashed line. I'm gonna put inside the receptive field an oriented gabor. On the other side there's another object, a blob of a



certain color. The animal is supposed to look at the screen and pay attention to only one of the objects, as it seems like it received a cue that looking right will be important. The animal has to remember the object and he has to respond whether the gabor has changed orientation or not after some time of delay and re-presentation of the object. The change on the left is irrelevant, as it is supposed to be non-attended. Then, after an interval, we switch attention to the other side. The gabor is still on the receptive field, but it should be not attended. Does the neuron response change?

The animal is motivated as it is gaining a reward (drink) for each correct answer. So, as soon as it starts making errors, it simply switches attention. The error is a cue to switch sides. The neurons in V1 and V4 are selective for orientation, so one can create a tuning curve and compare the one where the animal pays attention and the one where it doesn't. The selectivity doesn't change, but the peak is higher for "attended". The more you move away from the retina, the more the effect of attention increases (is magnified). The tuning curve for attended stimulus shifts upwards. The more you get deeper, the more variability is a factor. Here, the modulation is top-down, space-based.

Another experiment: this time we go in V5 (the Middle Temporal → direction of motion). I can use RDPs (Random Dot Patterns). They create movement, as a section of those moves either coherently or not, simulating movement.

Experiment 1: the animal is fixating the center of the screen and we send a random dot pattern inside the receptive field of the neuron going in a specific direction. What's important is that the attention is inside the receptive field or on the opposite side. Regardless, we record the cell. We obtain a tuning curve, that shows stronger response when the animal is attending the receptive field.

Experiment 2 (object-based attention): same structure as before. The stimulus is an upward pointing arrow. Consider a case in which the attention is always outside, so the animal attends the other dot pattern (opposite random dot patterns A or B). If the attention were just spatial, the movement direction of RDP really should not matter. If the directions of motions are opposite (A and C) the response decreases. So the feature of movement influences the firing. So, both location and pattern matter and are summable (both contribute to create the Attentional Template), with a multiplicative effect on neural responses.

Humans (and in general observers) respond more when the stimulus has higher contrast (distance between the whitest part of the scene and the darkest part of the scene). Three images that differ for contrast → the response is a sigmoid w.r.t. contrast (CRF - Contrast Response Function). This tells me that response of the neuron depends on the contrast up to a point (saturation). Mid-point of the curve = point of maximum change. This is a form of bottom-up attention (response to something that becomes more salient). But also voluntary attention changes the response of the neuron.

How top-down attention works (how does it affect the CRF):

$$\text{response} = \frac{R_{max} * C^n}{C^n + C50^n} + M$$

Attention can either increase R_{max} (maximum firing rate) - TDAttention is a response gain model - or the sensitivity of the curve (Contrast Gain Model) → a smaller contrast is sufficient to go towards the higher part of the curve. $C50$ is the point of mid-contrast.

Experiment: fixation point and a receptive field I can either attend or not. Top-down, space-based. I vary the contrast of the stimulus to see which is the model (RG or CG). The effect is much larger when the stimulus is of medium contrast. If there's high contrast, the attention cannot do more (so I don't change R_{max}). Endogenous attention is able to change when the stimulus is weak. So, the model is a Contrast Gain Model, changes the sensitivity.

Objects compete with one another. The competition is resolved using biases relevant to the behavior. One crucial resource over which objects compete is the receptive field, so that one or the other drives the neuron (only one object drives the neuron response, otherwise it would all become ambiguous). Attention may alter the shape of the receptive field.

Consider a neuron from V4 (one of the most studied areas for attention). The neurons from V4 respond to a combination of orientation and color. Inside the receptive field of this neuron we put either a horizontal green bar on the top left or a vertical red bar on the bottom right and present it for 200ms. We see that there is a slight increase of response (spiking rate) and rapid decrease for the green (it is a poor stimulus). The opposite should be good, and we see that the activity goes up, stays high and continues being high and eventually goes down once we turn off the stimulus. It is a strong stimulus. What happens if we apply both together? The response is an average of the two: it is an ambiguous signal.

Attention biases the competition, so that the capacity only goes to the relevant objects. We are recording from V4 neurons, which have relatively large receptive fields, so that we can include two stimuli, one green and one red. Now we have two conditions for the animal: we ask to pay attention to the left stimulus (spatial attention to cue the animal). We obtain a strong response. If we then ask to focus on the green stimulus, the response of the neuron is

low. We restore the clear signaling thanks to the attention. Stimuli that are unattended are filtered out. Attention resizes the shape of the receptive field (it is like it shrinks to include only the attended stimulus). We can produce the same effect in other areas.

Qualitative way to explain: Imagine we're recording from a neuron which receives two inputs from lower areas, one of the two is mostly excitatory and weakly inhibiting - the preferred input - and the other is mostly inhibiting - a non-preferred. When they are both present, there's an average response. Attention is no more than increasing the connectivity between the neurons.

Lesson 04: Selective Attention

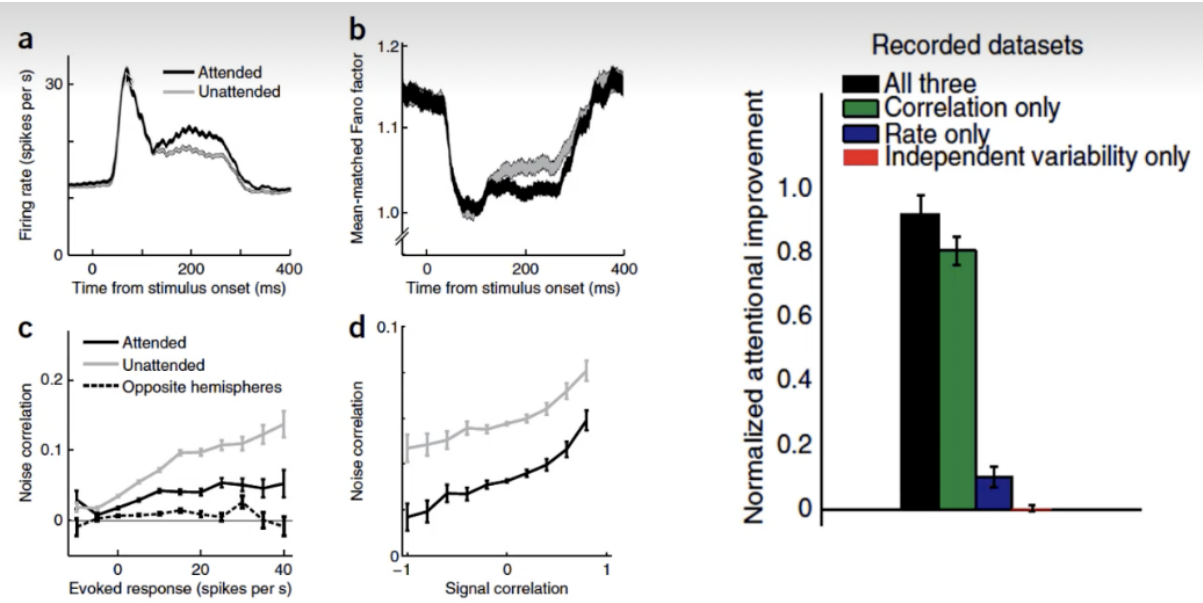
@May 24, 2023

When we experiment in humans or animals, we have a task (stimulus presented - response recorded). We need to repeat those trials many times, because individual responses have high variability (this "noise" is both a blessing and a curse). In each single trial you need to understand what is in the environment. If there's correlation between noise in individual neurons, averaging will not eliminate it, therefore it will stay. Attention can reduce noise correlation among neurons, so that each neuron behaves independently from the others.

We need to use a multipole electrode (array of electrodes). Cohen and Maunsell (2009) implanted a 6x8 array of microelectrodes in the brain of animals, measure and record the correlation among a pair of two neurons while doing a task. See whether the attention has a positive effect or not.

Task is a cueing experiment, using oriented gratings to the left or to the right, with one side attended and one not attended during a block of trials. The objective is to detect a change in the orientation of each stimulus. Results say that performance improves as the change increases. When the change occurs on the unattended side, performance is worse and to reach quality it requires a much larger shift.

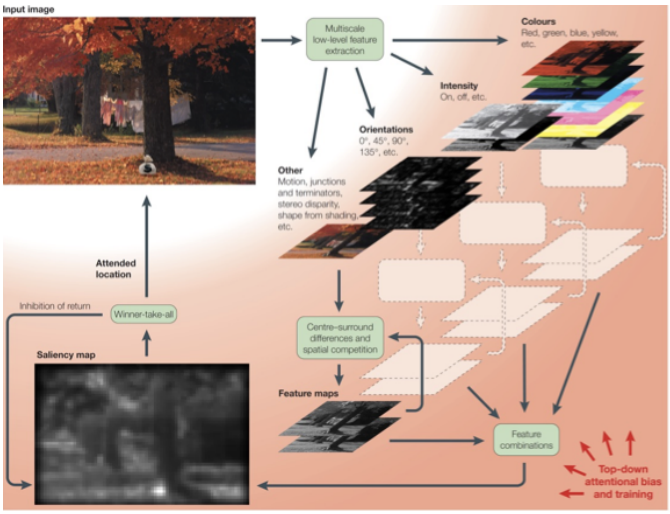
Results:



When the stimulus is attended, it produces a larger response (higher firing rate). Fano factor = ratio between the standard deviation and the mean: we see that when we pay attention, the variability across trials reduces, so the Fano Factor for the attended condition is lower. The major point is the correlation between neurons. Correlation is lower when the animal pays attention. Attention = more independence (Reduction of noise correlation), then Signal stronger, because of averaging. Conclusion: reduction of correlated noise when animal pays attention. This way, by averaging, the signal appears.

Bottom-Up Attention

Certain information are more salient than other (e.g. we tend to pick something that moves towards us, brighter, larger, novelty).



At the very early level, when we analyze the single features, this is performed in parallel (also without attention, it does not require it). Although attention is not required, one could use attention to amplify the operation of a single saliency map for a specific feature. All these feature maps form a saliency map that encodes what's salient for orientation, depth… an abstract map that says what's relevant across all the maps.

Visual Search kind of tasks. Either the target is salient (Popout Screen) because it is different from all others (odd man out) or in the Conjunction Screen, it is not easy to reject the distractor, so one scans one at a time the entire screen (use Top-Down Attention)

Present to an animal who fixates the center of the screen. The stimulus (fast - 200ms) is presented and compare the difference between popout and conjunction. Make the difference and get the so-called popout index (positive = more response from popout).

Causal Control of Visual Attention

According to certain theories, we shift attention using eye movement preparation (or gaze control). Analyze the relationship between eye movement and attention. Two definitions:

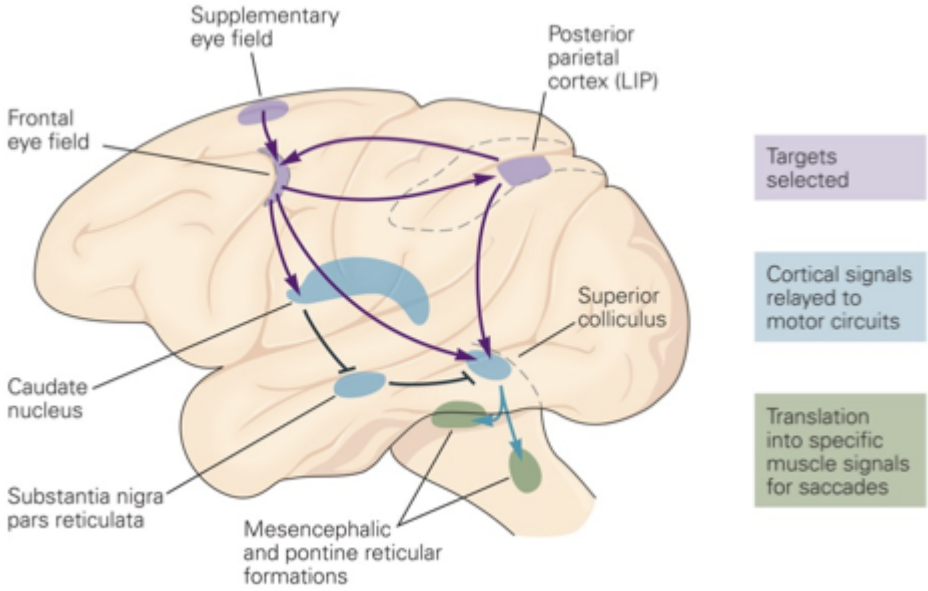
- Covert Attention → attend to objects of interest in the visual scene without shifting the gaze
- Overt Attention → gaze direction and attention focus are spatially aligned (this is the most common situation)

Majority of experiments we do is in a covert situation to not move the fovea (the retinotopic analysis changes completely).

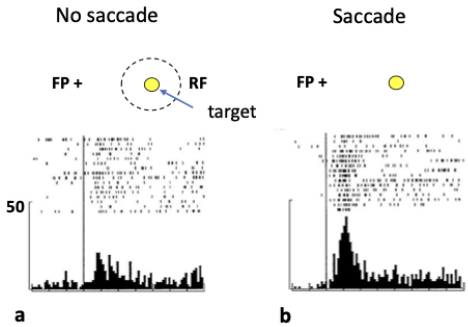
So, one can move attention without moving eyes, but one cannot move the eyes without moving attention: asymmetric relationship.

Frontal Eye Field (field of the Frontal Lobe in the Prefrontal Cortex) area related to voluntary eye movement (concept known since 1880s). Studied frontal eye field lesions in animals. 1980s: Premotor Theory of Attention (when you prepare eye movement, you shift attention there). It is a visual motor area, which contains both visual and movement-related neurons. Saccades = eye movements (they are slow in preparation, but when it's ready, the movement

is very fast). Influence activation of FEF with electrical microstimulation to obtain saccades of a particular vector (I can make the animal go in a specific portion of the visual scene).

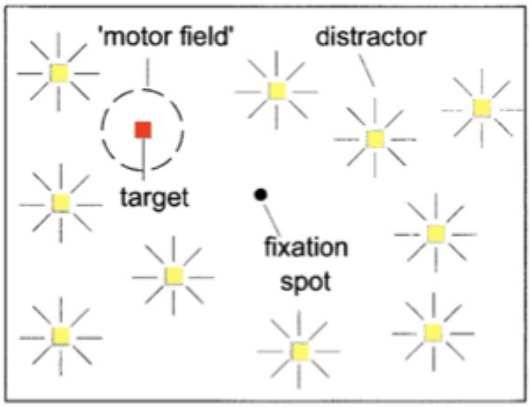


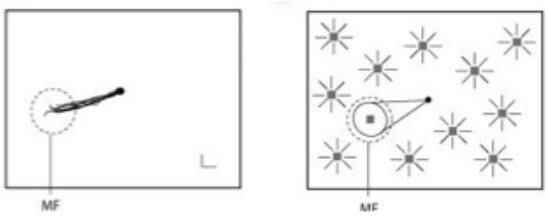
We also know from psychology and neuroscience that when an animal is preparing the eye movement, the attention is already there (pre-saccadic enhancement). Experiment: shows the relation between the visual response and eye movement. Stimulus presented in the receptive field of the neuron, with the animal keeping fixation on the center of the screen (fixation is more important than the target). Then I make the animal perform a saccade after presenting the stimulus. Now it becomes relevant because it has to provide the eye movement → response is stronger. Vision response and saccadic preparation are linked.



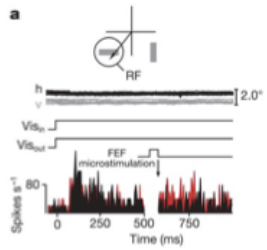
Experiment: the animal is fixating the center of the screen. A stimulus appears anywhere in the visual field (red target) and then decrease in luminance (signal that he has to move the eyes towards the location). This is made difficult by presenting some

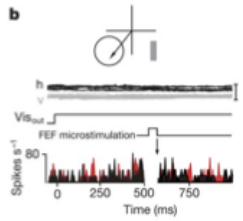
flashes (distractors), once the animal has “committed” to respond only to one target. If the animal fails, it is not rewarded, so it is motivated to avoid distractions.
Manipulation in the Frontal Eye Field: we perform microstimulation and we see which saccade we’re able to evoke (the trace of the eye movement the animal makes), which is the motor field (where I go once I stimulate the frontal eye field), then I present a stimulus inside the motor field. I’m not stimulating above threshold (I don’t want to evoke the saccade). It becomes more sensitive to the dimming of the target. This means that I can increase the performance of the animal by stimulating with a sub-threshold the FEF: there’s a relationship.





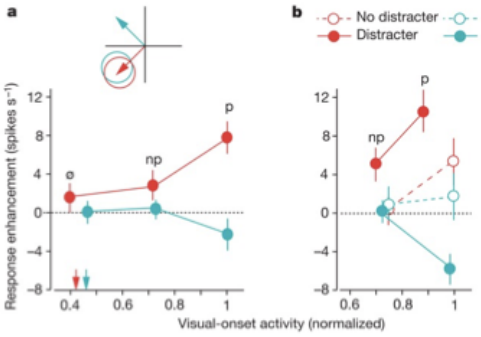
Another confirmation is the following experiment: the animal is fixating the center of the screen while a target is presented in the visual field. The animal is not required to do anything except fixating. As before, we microstimulate the FEF where the eye movement goes (motor field). Various trials and recording of simple neurons in the V4 area. The enhancement effect was significantly increased by the presence of distracting stimuli (attentional competition). Furthermore, stimulation of FEF sites with motor field not corresponding to the RF location of V4 neurons suppressed V4 responses. Results suggest that the gain of visual signals is modified according to the strength of spatially corresponding eye movement commands





a, V4 neuron's mean response during control (black) and stimulation (red) conditions. Shown after the onset of RF (Vis in) and non-RF (Vis out) stimuli and after a 50-ms subthreshold (20 μA) stimulation.

b, As in a, except that histograms show the neuron's response when the RF stimulus was not presented.



a, The mean response difference between stimulation and control conditions (stimulation - control) when the evoked saccades (arrows) shifted the monkey's gaze to a point within (red) or outside (blue) the RF (circles) of the recorded V4 neuron. 0 (no stimulus), np (a non-preferred stimulus) or p (a preferred stimulus) in the RF.

b, conditions with and without a non-RF stimulus (distracter) are shown separately.

Moore & Armstrong, Nature 2003