COSC451: Artificial Intelligence

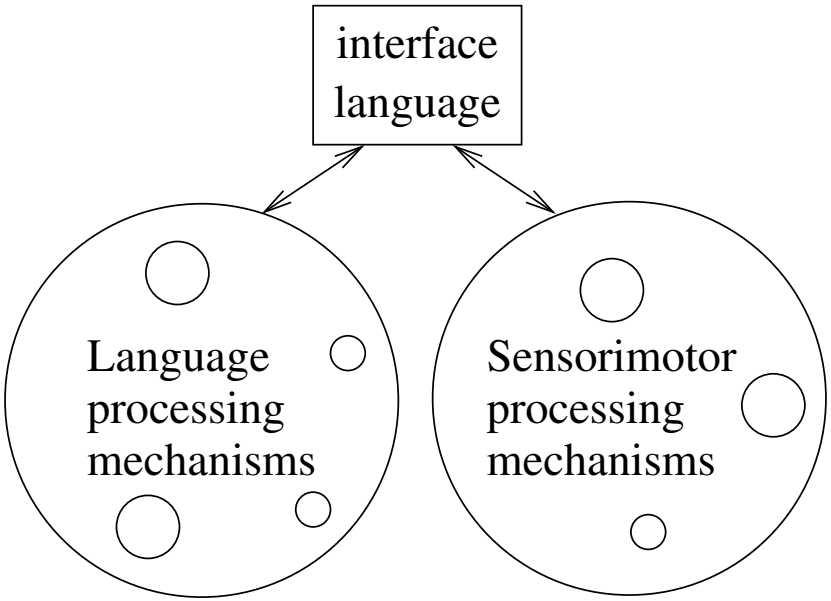
Lecture 2: Visual object classification and visual attention

Alistair Knott

Dept. of Computer Science, University of Otago

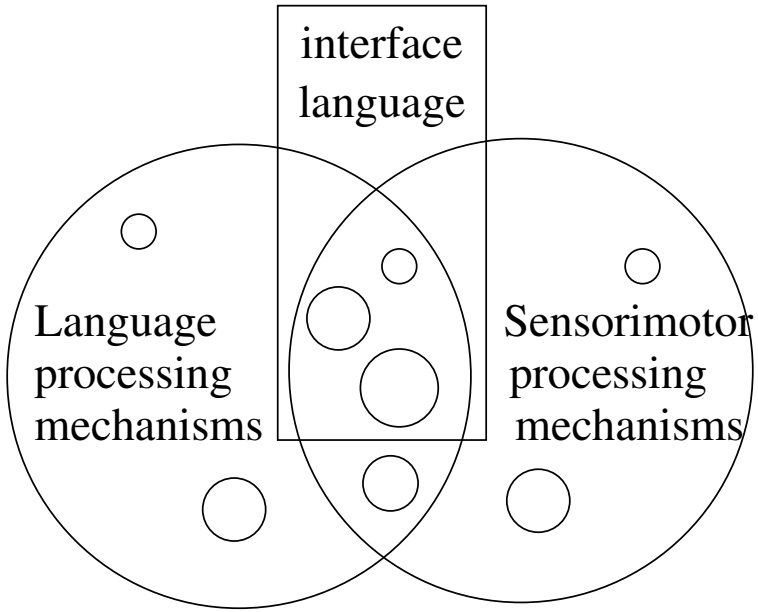
Recap: The shared mechanisms hypothesis

Language and SM processing are modules



Semantic representations abstract away from details of SM processing

Language and SM processing share mechanisms



Semantic representations retain details of SM processing

Recap: Structure of the course

There are three parts to the course:

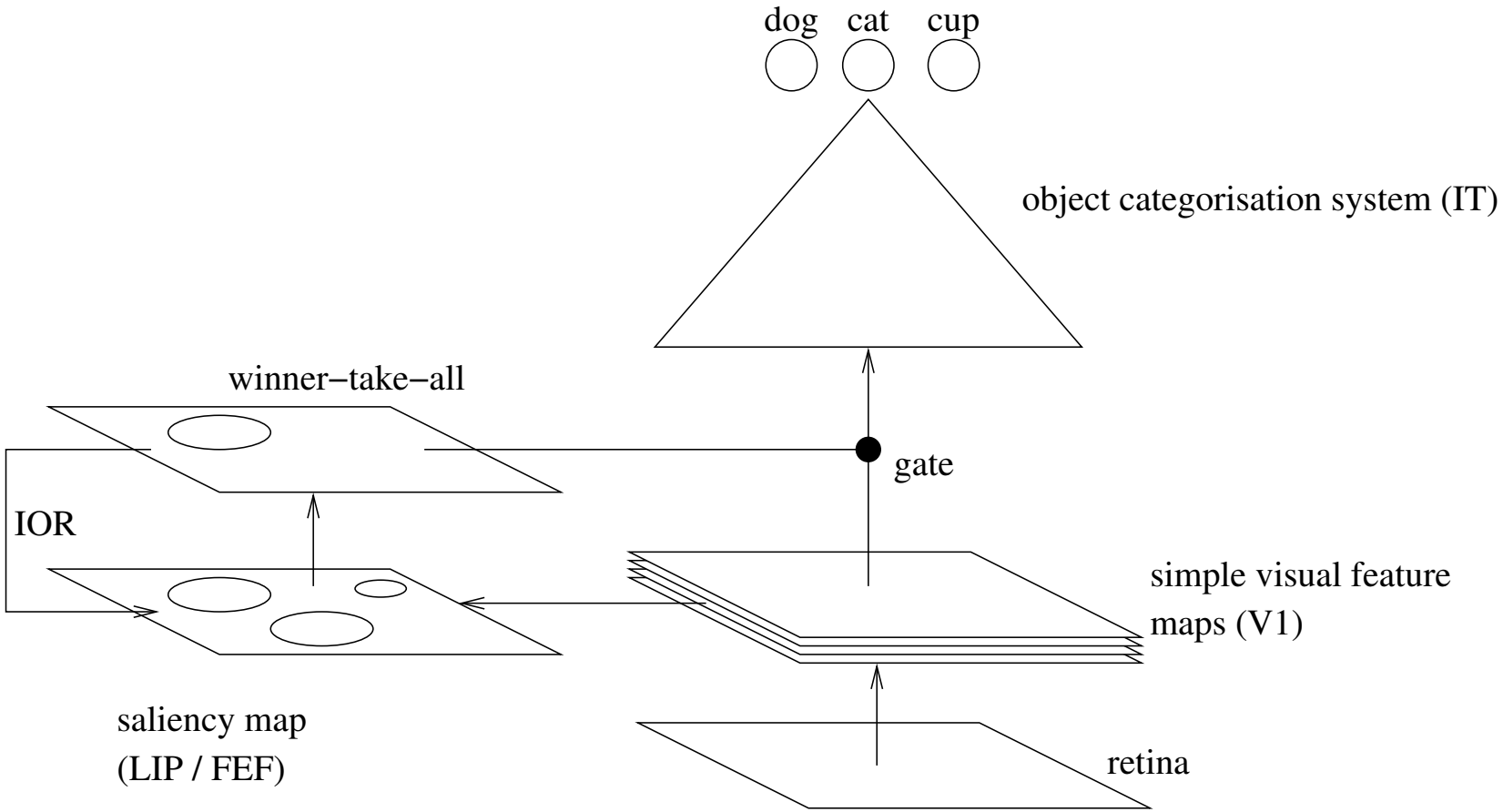
Part 1	Develop a SM model of the processes involved in grabbing a cup, or in observing a cup-grabbing event (motivated from SM psychology / neuroscience)
Part 2	Develop a syntactic model of the cup-grabbing sentence (motivated on purely linguistic grounds)
Part 3	Look for formal similarities between these models.

If there are nontrivial similarities, this is support for the shared mechanisms hypothesis.

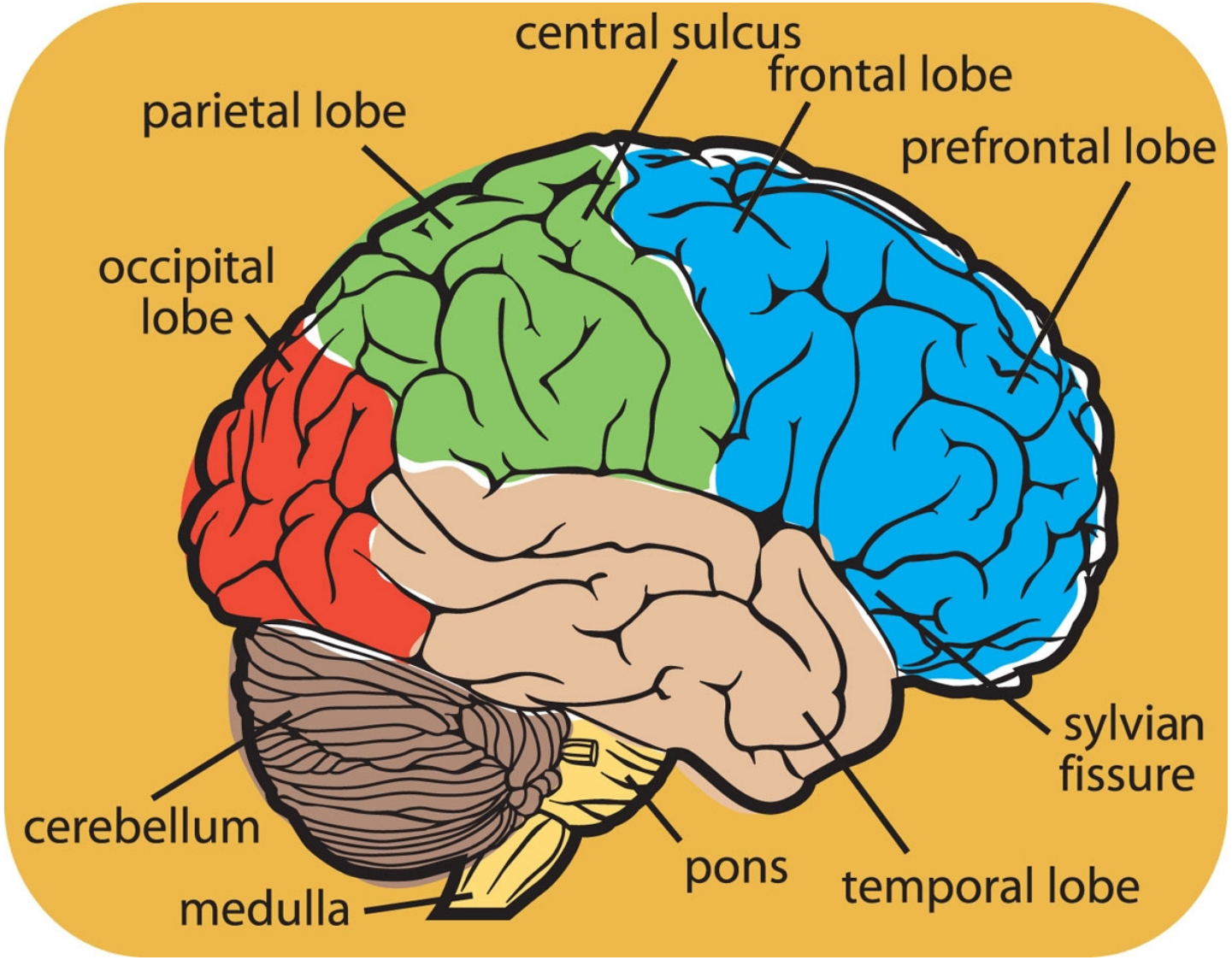
Part 1 overview

Week	Topic
2	Visual object classification and visual attention
3	Vision for action: the reach and grasp motor pathways
4	Planning actions: the prefrontal cortex
5	The action recognition pathway
6	The 'who' pathway: representations of self and other
7	Sequential structure in experience of a reach-to-grasp action
8	Working memory representation of a reach-to-grasp action
9	Episodic memory representation of a reach-to-grasp action

Neural pathways involved in object perception



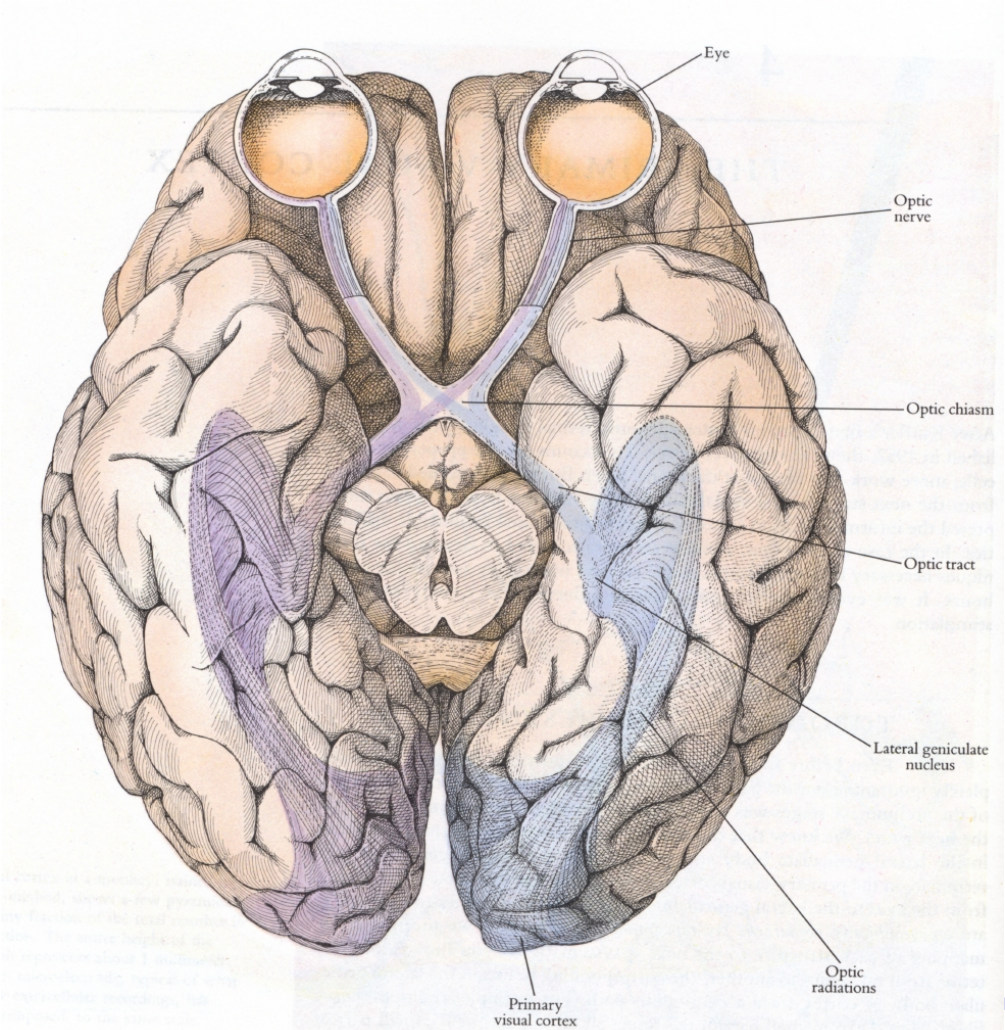
Areas of the brain



Outline of today's lecture

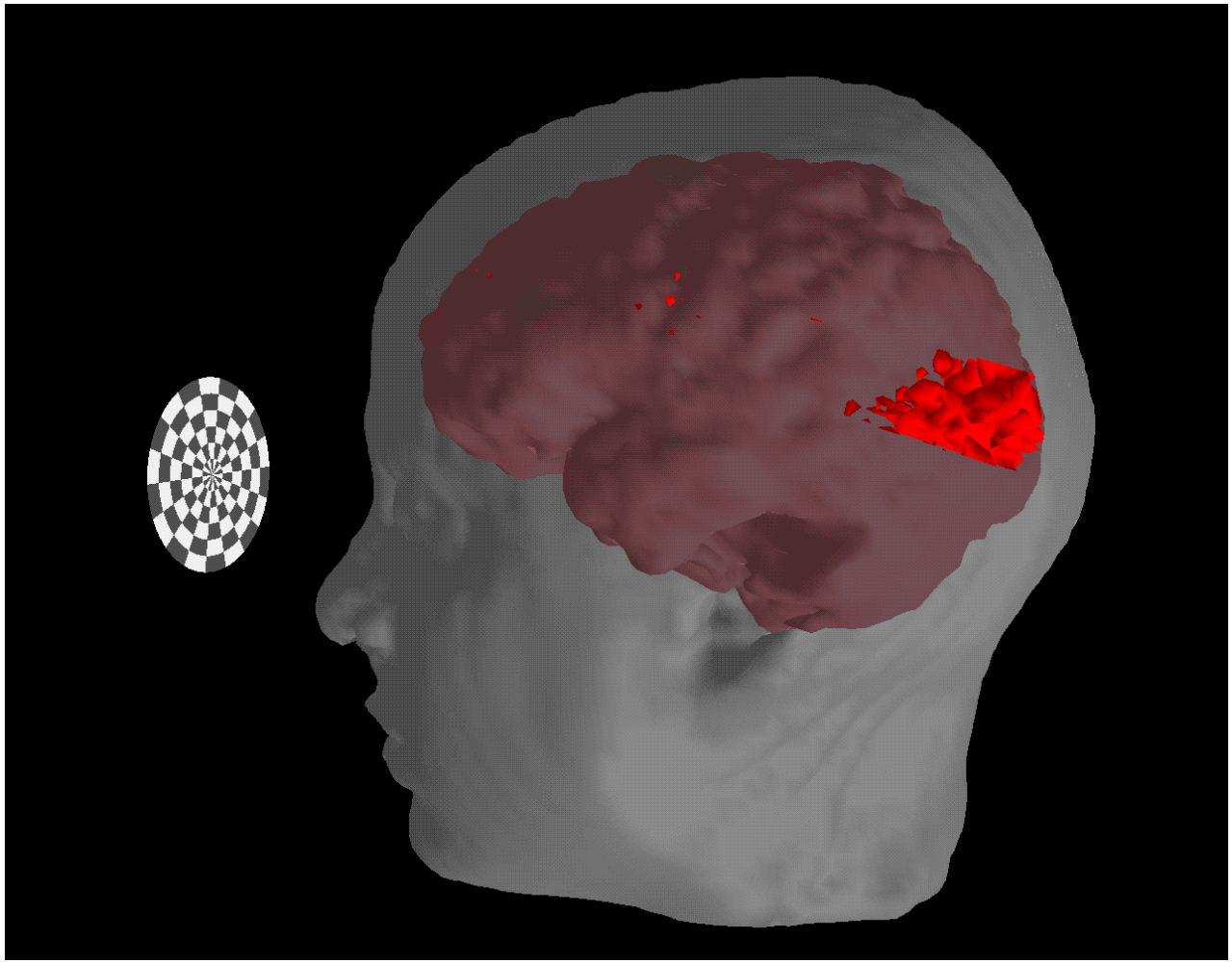
Early visual processing

Information from the retina is transmitted via the lateral geniculate nucleus to primary visual cortex (V1) in the occipital lobes.



V1 in humans

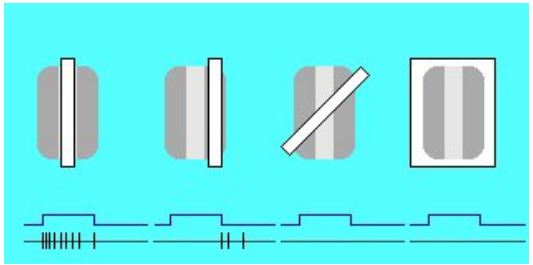
Here's activity from an fMRI scan generated while a (human) subject watches a simple visual pattern:



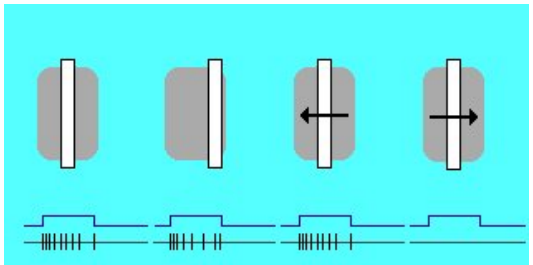
Simple and complex cells in V1

The structure of cells in V1 was discovered by Hubel and Wiesel (1968), using single-cell recordings.

- Cells in V1 are organised retinotopically. They compute a range of primitive feature maps over the retina.
- A simple cell responds best to a stimulus with a particular orientation, and a particular size, at a particular point on the retina.

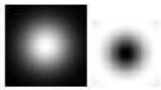


- A complex cell responds best to a stimulus with a particular orientation and size anywhere within a (small) area of retina.



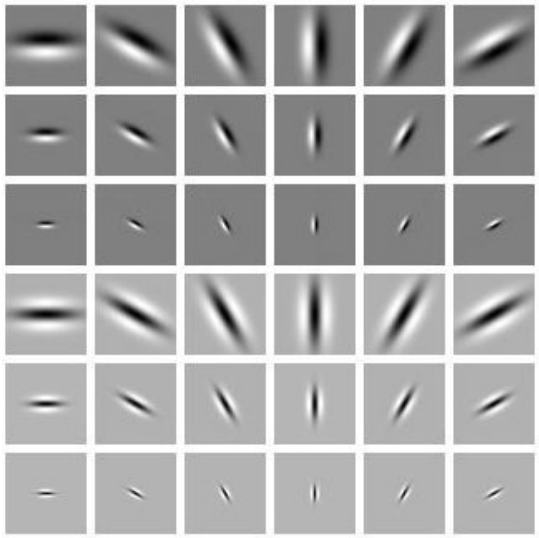
The kind of filters computed by V1

Laplacian-of-gaussian filters:



Difference-of-Gaussian filters ('blob detectors'):

Oriented Gaussian filters ('line detectors'):



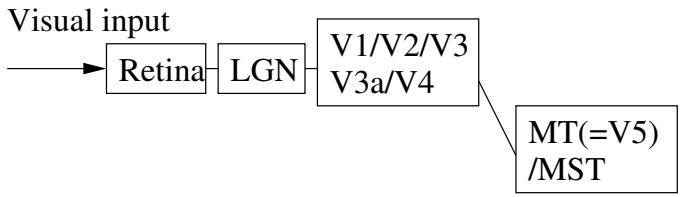
Projections from V1

V1 projects to V2, which is also retinotopically organised (though at a coarser granularity).

V1 and V2 project to several other more specialised retinotopic areas.

- V3 cells are sensitive to orientation and binocular disparity (Adams and Zeki, 2001) but not to colour (Baizer, 1982).
- V4 cells are sensitive to simple shapes (Cadieu *et al.*, 1998).
- MT (=V5) and MST cells are sensitive to motion (Maunsell and van Essen, 1983).

To give a very simple summary:

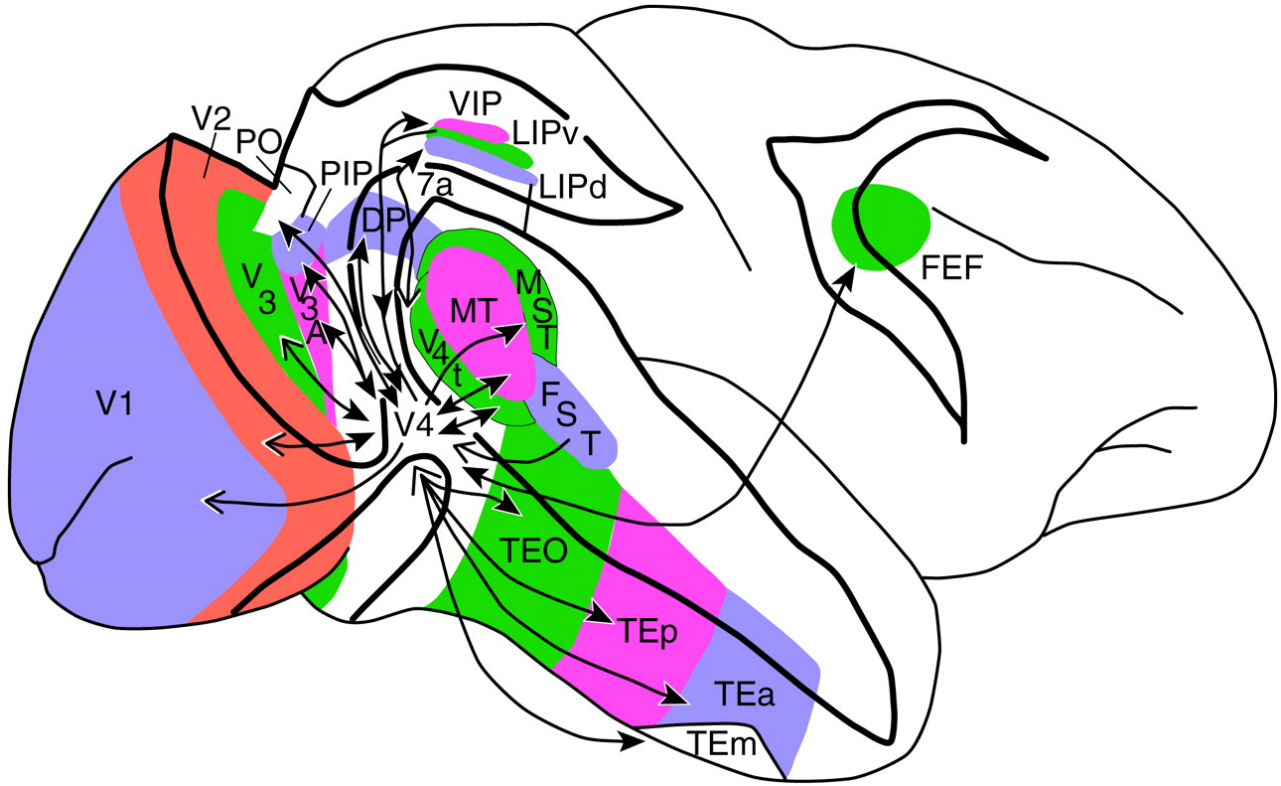


Outline of today's lecture

The object classification pathway

There is a specialised pathway for categorising *complex object shapes*: the kind of information which is necessary to identify the *type* of an object, or to recognise individual objects.

- The pathway receives input mainly from V4, and involves the inferotemporal (IT) areas TEO and TE.



Characteristics of cells in the classification pathway

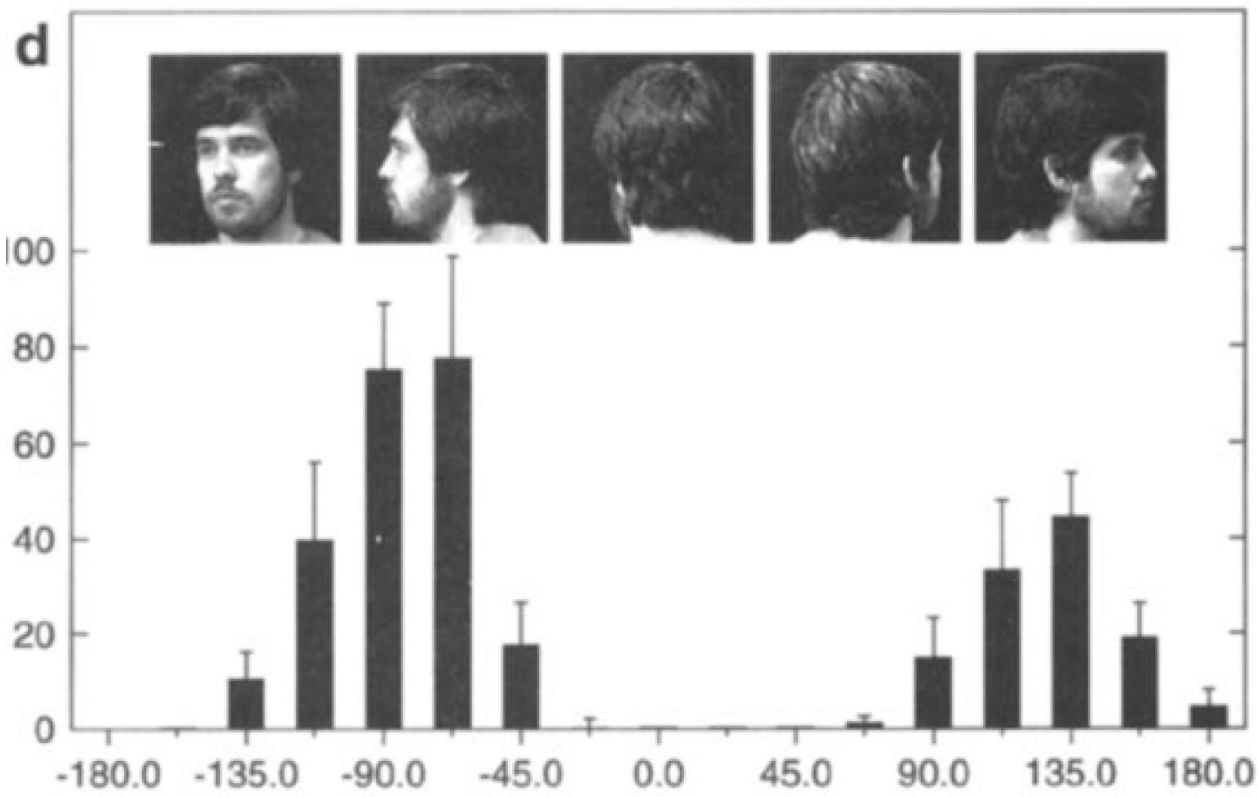
As we progress along the pathway from TEO to TE, cells respond to progressively more complex stimuli, and have progressively wider receptive fields (Tanaka, 1993).

The cells at the end of the pathway respond to fairly complex shapes. Logothetis *et al* (1995) trained monkeys to discriminate between 'paperclip' stimuli:

They found individual cells in TE which responded selectively to particular stimuli. (Most were selective to particular *views*, but some responded to multiple views.)

Responses to biological shapes in IT

Many IT cells respond selectively to ‘biological’ shapes, such as faces and hands.

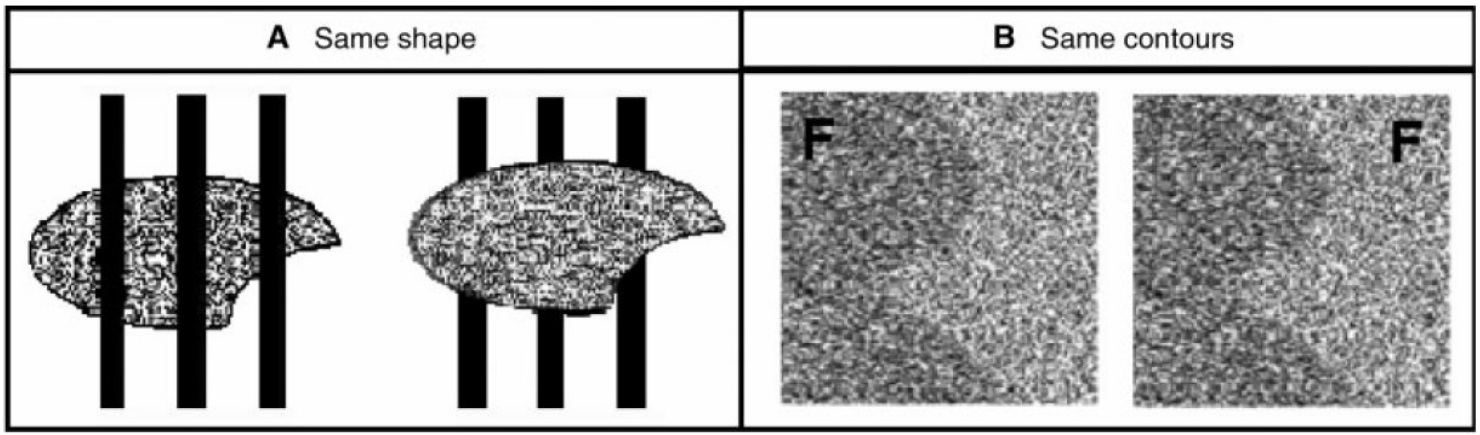


Here’s a cell that responds selectively to particular views of faces (Logothetis and Sheinberg, 1995).

The object classification pathway in humans

In humans, the shapes of objects are represented most strongly in lateral occipital cortex (LO) (Kourtzi and Kanwisher, 2001).
(Note: 'LO' extends into inferior temporal regions.)

The experiment used an *fMRI adaptation* paradigm: LO responded less if a shape was presented twice, even if low level features of the display changed.



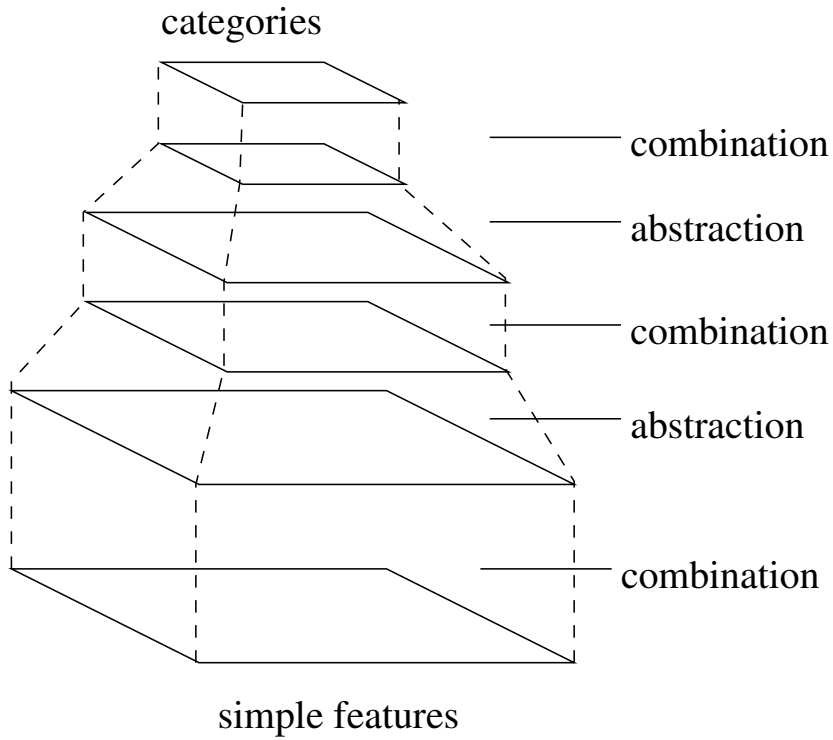
(I.e. adaptation was not just to 'contours'.)

A model of the visual categorisation system

The categorisation system is often modelled as a convolutional NN. Le Cun & Bengio (1995); Mozer & Sitton (1996); Riesenhuber & Poggio (1999)

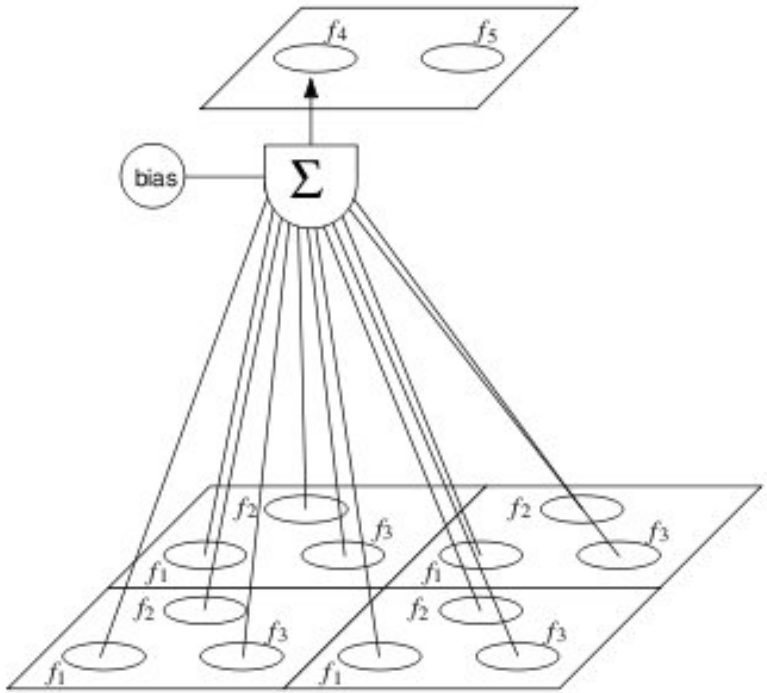
- Its input is a map of simple visual features.
- Each layer takes a map of features and returns a map of combined features.
- To avoid a combinatorial explosion, each layer also abstracts over space.

This is a reasonably good model of cells in the IT pathway.



A model of the visual categorisation system

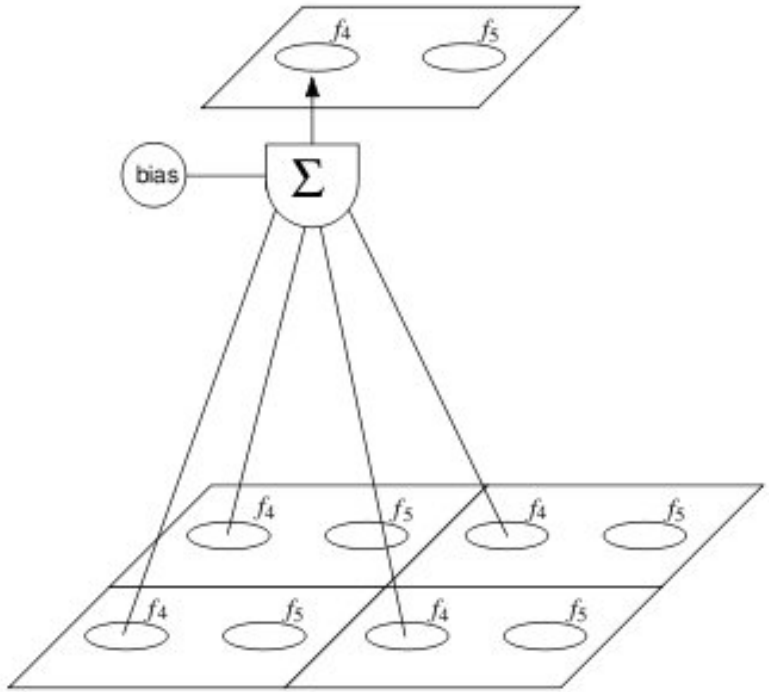
Combination layers look like this:



Shared weights are used, to simulate training at each retinal location.

A model of the visual categorisation system

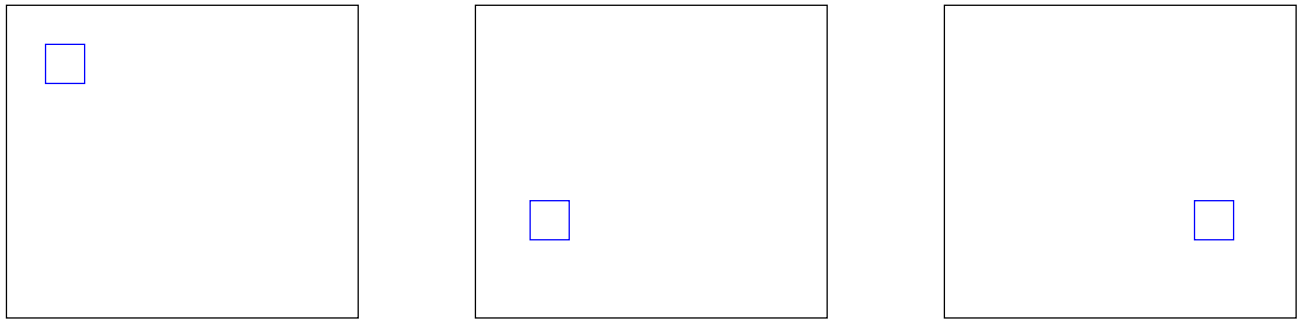
Abstraction layers look like this:



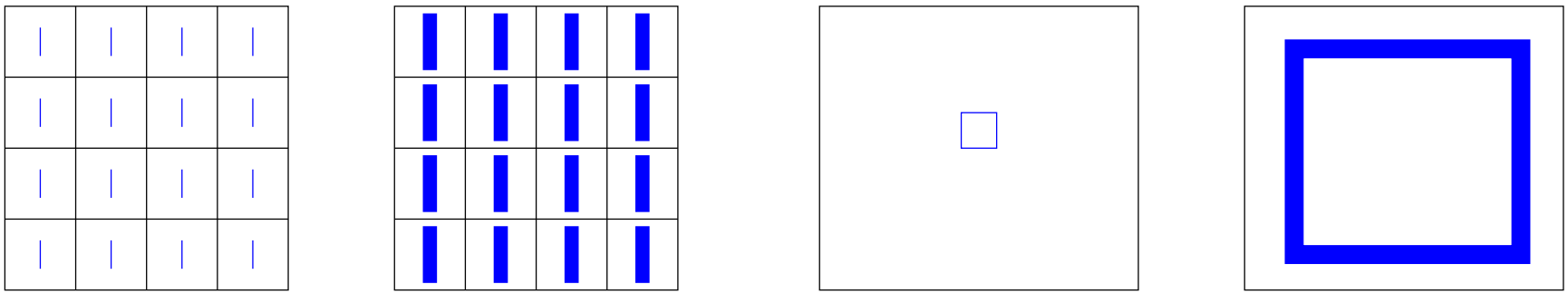
The weights in these layers are often fixed (i.e. not trainable).

Translation and scale invariance of a convolutional NN

The abstraction operations allow an object to be categorised anywhere on the retina.



Input feature maps of different spatial frequencies allow an object to be categorised at a range of sizes.

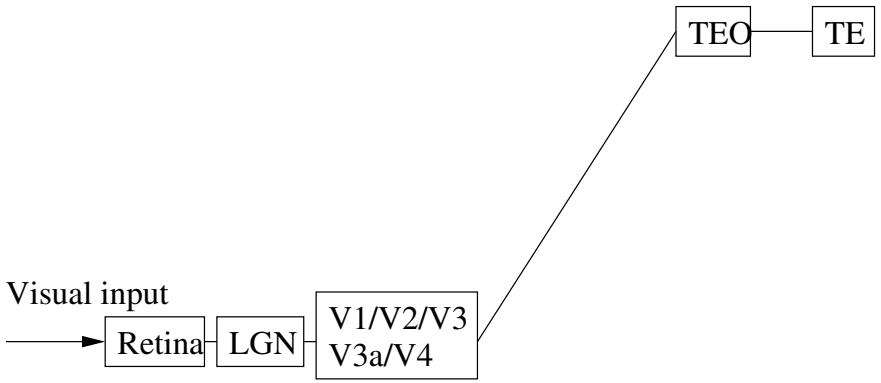


Top-down influences on object categorisation

Object categorisation is not only driven by perceptual information from the retina. There are top-down influences as well, which relate to the observer's *expectations*.

- Some expectations are general, and relate to the type of scene which the observer is looking at. These can be referred to under the umbrella term priming.
- Other expectations are specific, and relate to the particular objects which the observer has recently been looking at.

The object classification pathway



Outline of today's lecture

The attentional pathway

Cells at the end of the object classification pathway respond to specific shapes, but abstract away from retinal location. So how do we know *where* a classified object is?

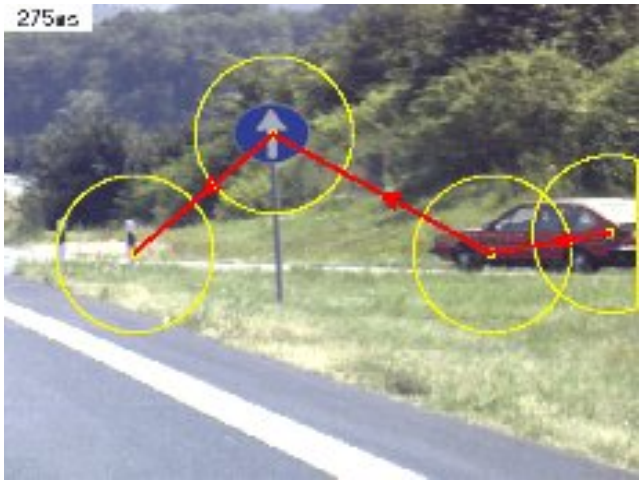
- There's a separate visual pathway in posterior parietal cortex which represents the location of objects.
  Posterior parietal cortex is involved in converting retinal locations into locations in a *motor coordinate system*. (See Lecture 3.)

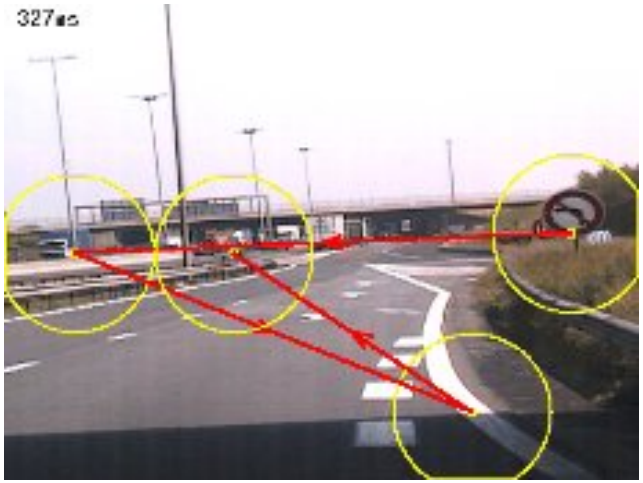
- There's another visual area called the frontal eye fields which is involved in controlling eye movements (and maybe other attentional operations), which also represents the location of objects on the retina.

Visual attention

The basic idea:

- The visual field is full of stimuli. We need to focus in on the most important ones.
- In the attentional pathway, we produce a *map* of the most important locations to attend to.
- Then we attend to them one by one.

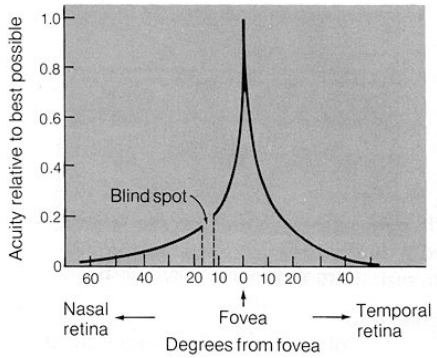




Overt attentional actions: saccades

Our eyes are designed to ‘focus’ on one location at a time.

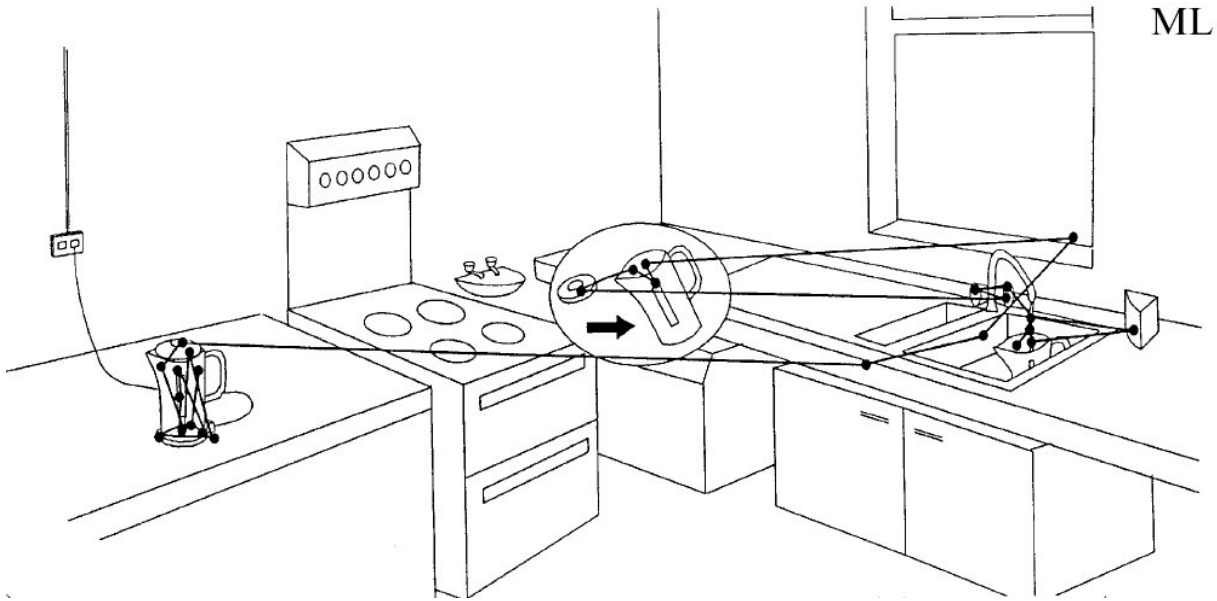
- The retina has a fovea at the centre, where visual acuity is hugely higher than in the periphery.



- The fovea ‘sees’ only 2 degrees of the visual field, but it contains about half the photoreceptors on the retina.
- We perceive the world by directing the fovea at a series of different locations.
- Eye movements are called saccades: we make around 3 a second, the whole of our waking lives.

Overt attentional actions: saccades

Here's a summary of the saccades recorded while an observer made a cup of tea (Land *et al.*, 1999):



Idea: the main job of peripheral vision is to build a map of important locations to attend to.

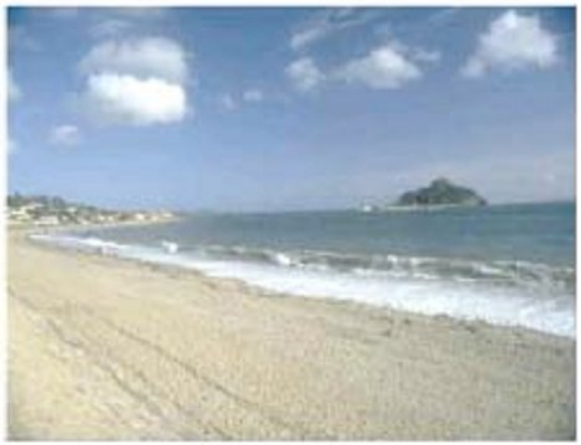
The saliency map

What counts as an ‘interesting location’?

- There are bottom-up things: e.g. *local contrast*, *movement*.
- There are also top-down things: e.g. low-level features of ‘something you are looking for’.

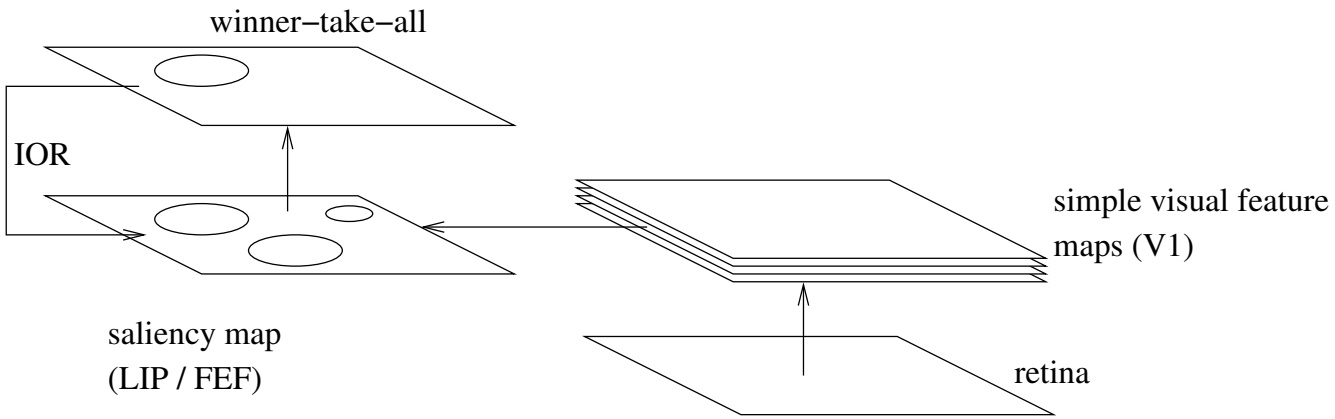
Itti *et al.* (1998) implemented a simple saliency map function.

- Its input comes from ‘early vision’: it’s a set of maps of simple visual features, at different orientations and scales.
- It computes local contrast in all these maps, and sums the results.



The saliency map

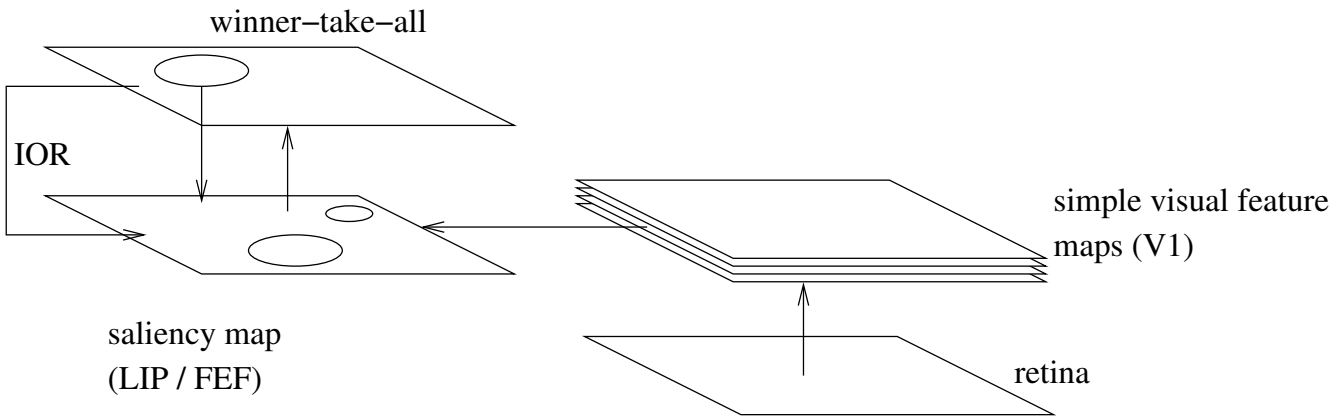
The saliency map projects to a winner-take-all (WTA) map, where active regions compete against one another, and the 'most salient region' is selected.



- After a time, an active region in the WTA map *inhibits* its corresponding region in the saliency map.
- This biases competition in the saliency map, and a new region becomes the winner.
- In this way, each of the salient regions is selected in turn.

The saliency map

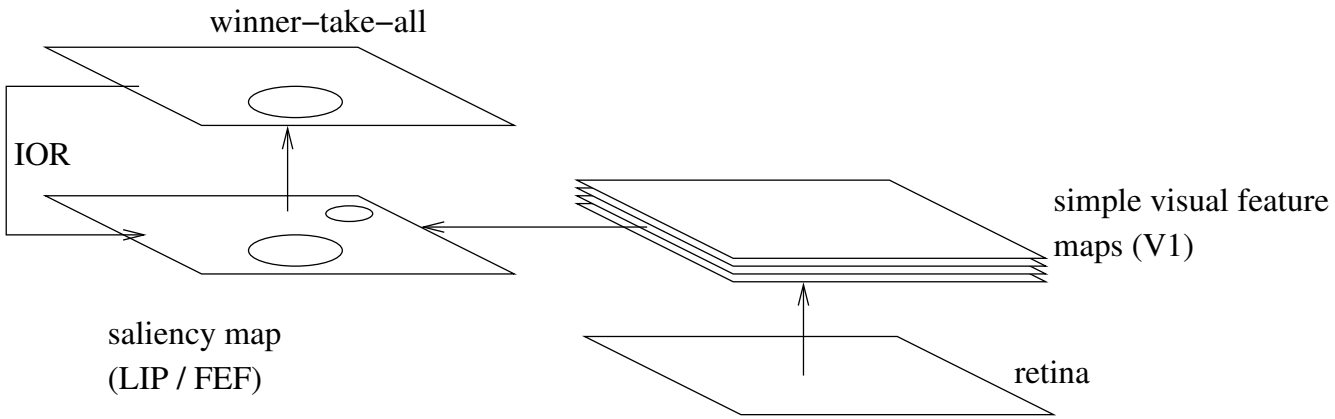
The saliency map projects to a winner-take-all (WTA) map, where active regions compete against one another, and the 'most salient region' is selected.



- After a time, an active region in the WTA map *inhibits* its corresponding region in the saliency map.
- This biases competition in the saliency map, and a new region becomes the winner.
- In this way, each of the salient regions is selected in turn.

The saliency map

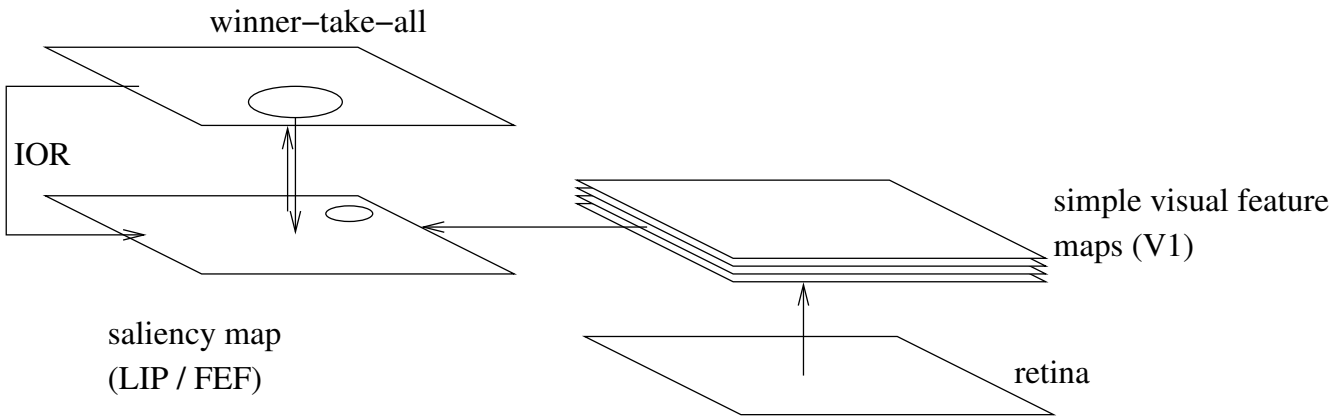
The saliency map projects to a winner-take-all (WTA) map, where active regions compete against one another, and the 'most salient region' is selected.



- After a time, an active region in the WTA map *inhibits* its corresponding region in the saliency map.
- This biases competition in the saliency map, and a new region becomes the winner.
- In this way, each of the salient regions is selected in turn.

The saliency map

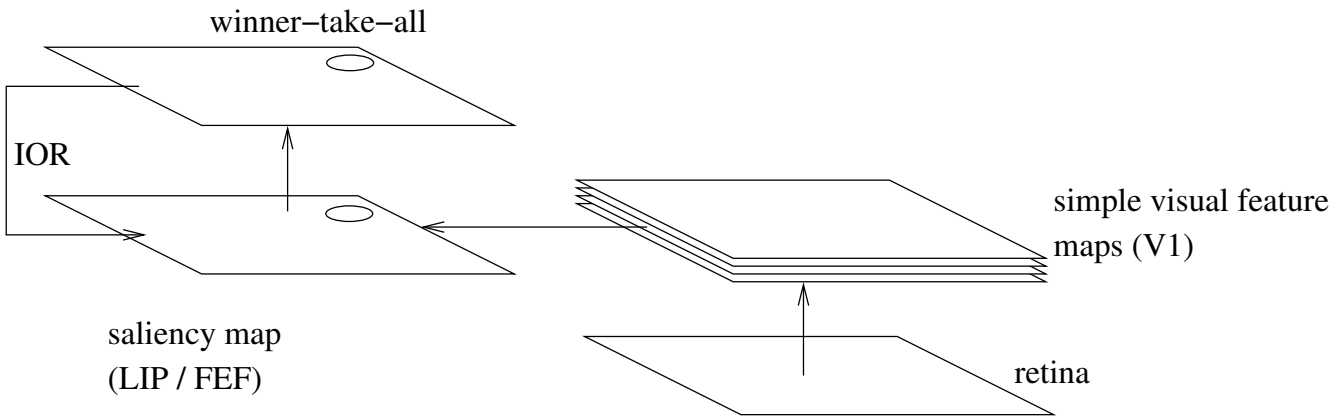
The saliency map projects to a winner-take-all (WTA) map, where active regions compete against one another, and the 'most salient region' is selected.



- After a time, an active region in the WTA map *inhibits* its corresponding region in the saliency map.
- This biases competition in the saliency map, and a new region becomes the winner.
- In this way, each of the salient regions is selected in turn.

The saliency map

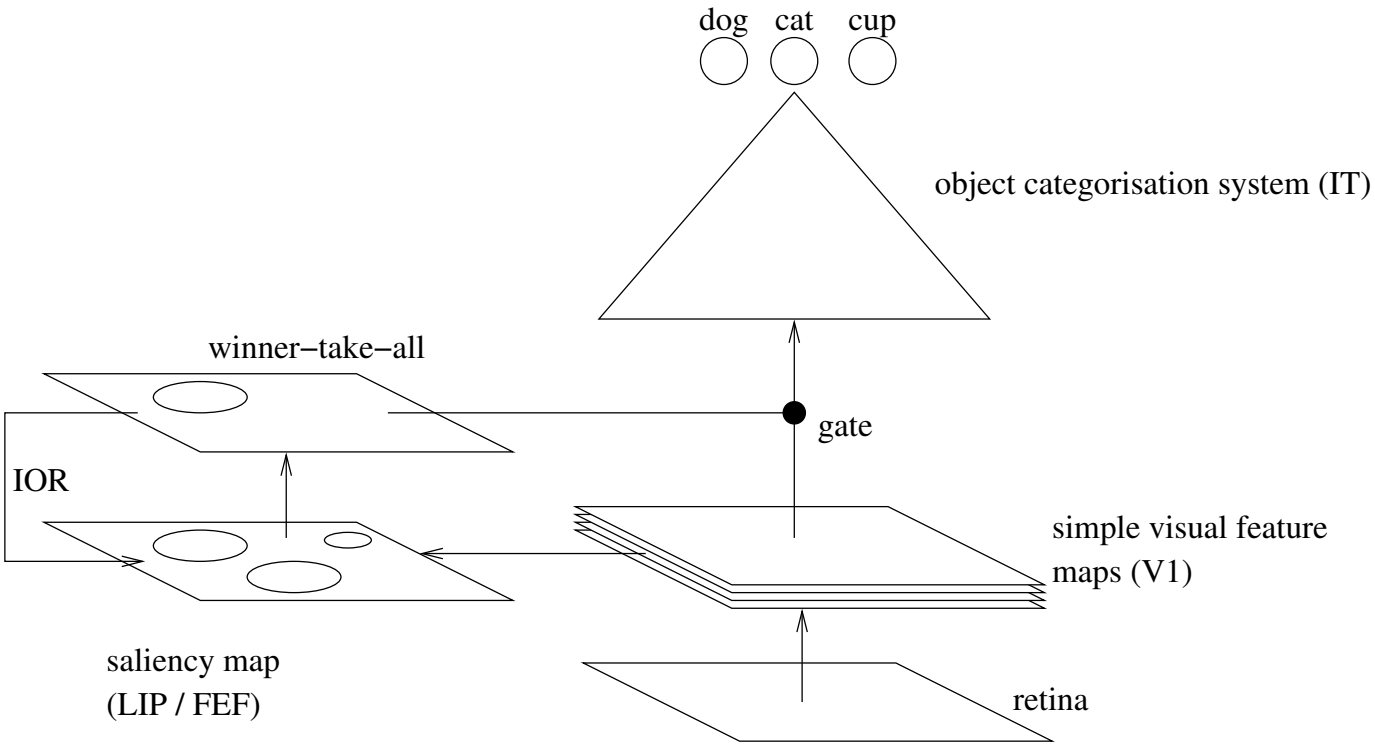
The saliency map projects to a winner-take-all (WTA) map, where active regions compete against one another, and the 'most salient region' is selected.



- After a time, an active region in the WTA map *inhibits* its corresponding region in the saliency map.
- This biases competition in the saliency map, and a new region becomes the winner.
- In this way, each of the salient regions is selected in turn.

The saliency map

Finally, we assume the most salient location *gates input to the classification system*.

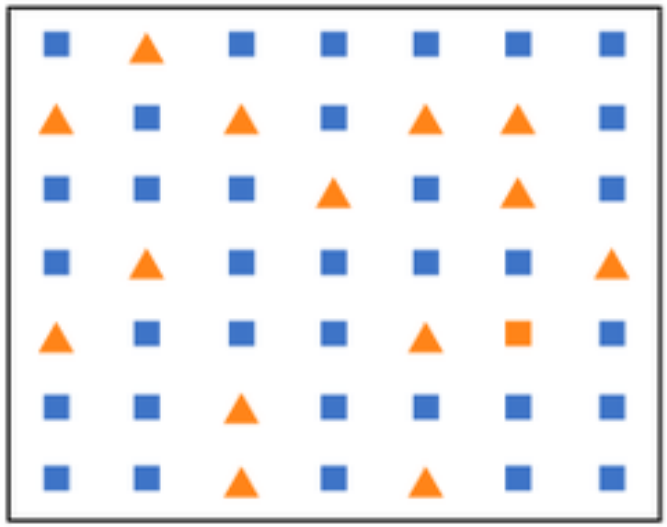


Now only the object which is attended to is classified.

Experimental evidence for visual attention

A lot of the evidence comes from visual search experiments.

- The subject hunts for a target in a field of distractors.
- If the target can be distinguished from the distractors by a ‘simple visual feature’, it can be identified immediately.
- If not, the time spent searching is proportional to the number of distractors.



Is there a saliency map in the brain?

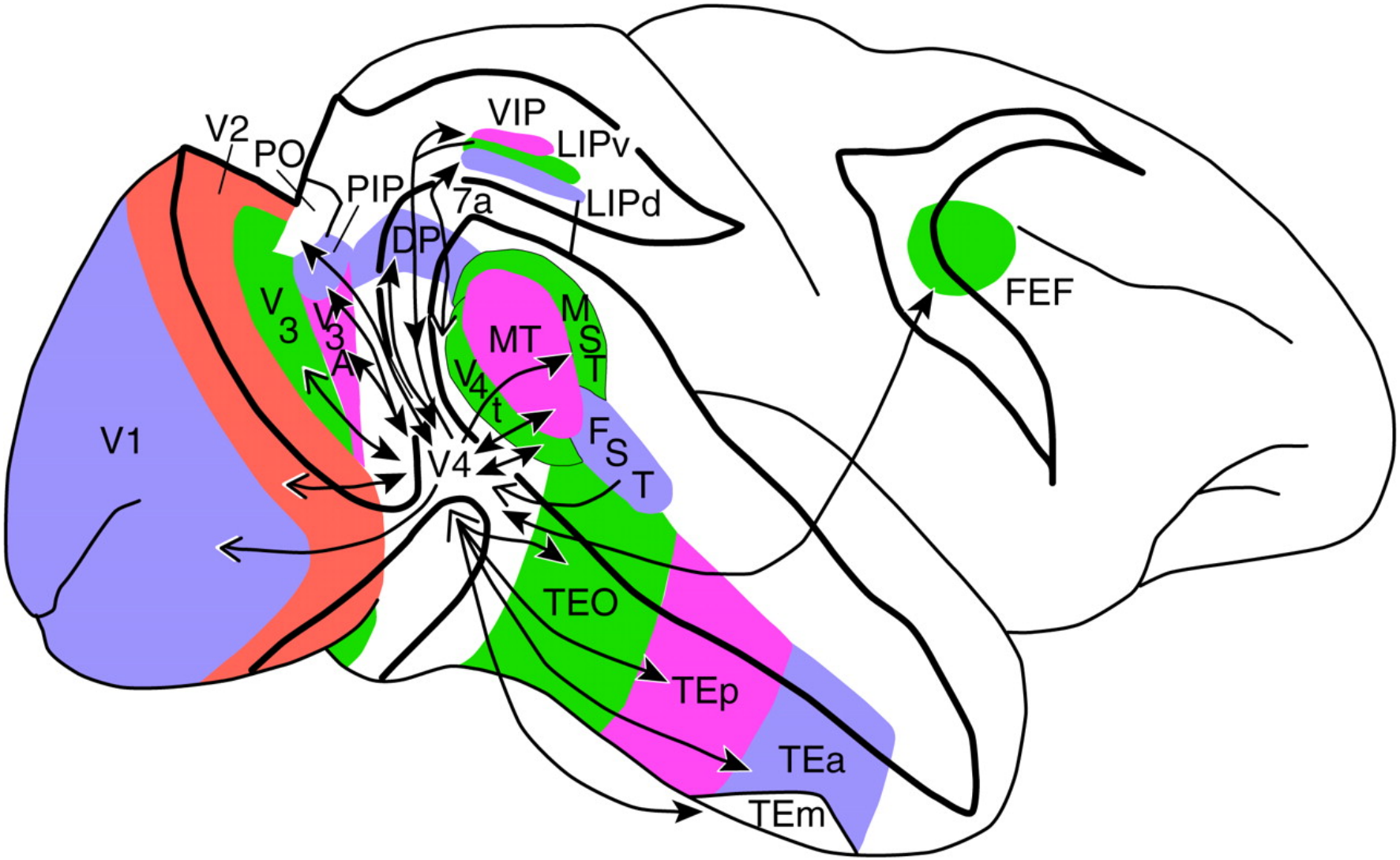
One area of parietal cortex—lateral intraparietal cortex (LIP)—seems a good candidate.

- LIP neurons respond more to 'newly appearing' stimuli (Gottlieb *et al.*, 1998).
- *Some* LIP neurons respond equally well to a salient stimulus whether the agent has to saccade *towards* it or *away from* it (Kusunoki *et al.*, 2000). This suggests they don't just encode saccades.

Another good candidate is the frontal eye fields.

- FEF cells respond better to stimuli which 'pop out' of visual displays (Bichot *et al.*, 2001).
- *Some* FEF cells encode salience, not saccades (Schall, 2004).

LIP and FEF


V2
PO
VIP
LIPv
PIP
7a
LIPd
DP
V3
V3A
MT
MST
V4t
FST
V1
V4
TEO
TEp
TEa
TEm
FEF

Overt and covert attention

LIP and FEF both influence eye movements.

- They project to the superior colliculus, which generates eye movements.

But we are also able to attend to peripheral objects *covertly*, without eye movements.

- Covert attention must involve gating retinal input to the classification system.
- Stimulating cells in FEF has been shown to modulate activity in the corresponding area of V4 (Moore and Armstrong, 2003).
- TMS of the human FEF improves detection of near-threshold stimuli (Grosbras and Paus, 2003).

Top-down influences on attention

At any given time, an agent has a particular cognitive set—one or more 'tasks' which s/he is actively pursuing.

- Cognitive sets are represented in prefrontal cortex (PFC).
- Some tasks are attentional—these are called search tasks.

During a search task, the target being searched for appears to be represented in PFC (Hagesawa *et al.*, 2000).

PFC also projects to FEF and LIP. And cells in these areas respond in a task-specific manner (Kusunoki *et al.* 2000; Bichot *et al.*, 2001).

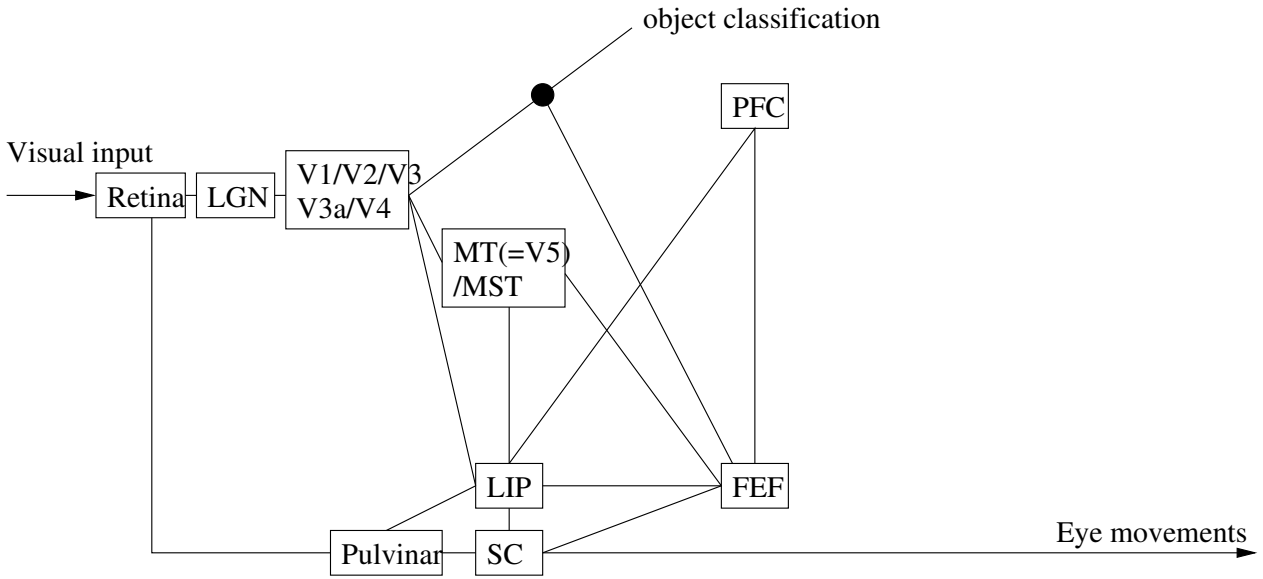
Top-down actions in visual search

Say we're searching for an object—a coke can. We can implement this by imposing two top-down biases:

- In one area of PFC, we can impose a bias on the saliency maps in FEF and LIP, to give preference to 'coke-can-like stimuli'.
- In another area of PFC, we can evoke a representation of the search target (a coke can), to be matched against incoming IT patterns.

Inhibition-of-return will cycle through the salient objects until a match is found, or until all items have been attended to.

The visual attention pathway



Summary

The model so far:

- Early vision represents simple visual features
- The object categorisation pathway combines visual features, and abstracts away from retinal location
- The attention pathway represents salient items in the visual field, and allows the agent to attend to these one by one (either overtly, or covertly).

Next week: visuomotor control (reaching and grasping).