COSC451: Artificial Intelligence

Lecture 5: Action recognition

Alistair Knott

Dept. of Computer Science, University of Otago

The purpose of action recognition

Why does an agent need to recognise the actions of other agents?

Ultimately, because the actions of others are important in determining his *own* actions.

- If agent A attacks me, I should attack back.
- If agent A picks up something I want, I should react.
- If agent A runs away, I should follow him. (Or chase him.)
- If agent A grooms me, I should groom him in return.
- If agent A grooms agent B—and A is dominant—I should treat B with respect.

Action recognition is particularly important in social animals.

Action recognition: some definitions

Assume we're working from visual input.

- The action recognition system takes retinal input for a period of time $T_0 - T_n$ depicting an action, and returns a representation of an action, including (i) the type of the action; (ii) its participants; (iii) its success.
- The action classification system takes similar retinal input, but just returns the type of the action.

Action participants can include AGENT, PATIENT etc.

Action types can include 'walk', 'grab', 'punch' etc.

The mirror system hypothesis

Current work in action recognition is dominated by the mirror system hypothesis.

> The basic idea is: the neural mechanisms which subserve action recognition partially overlap with those which subserve action execution.

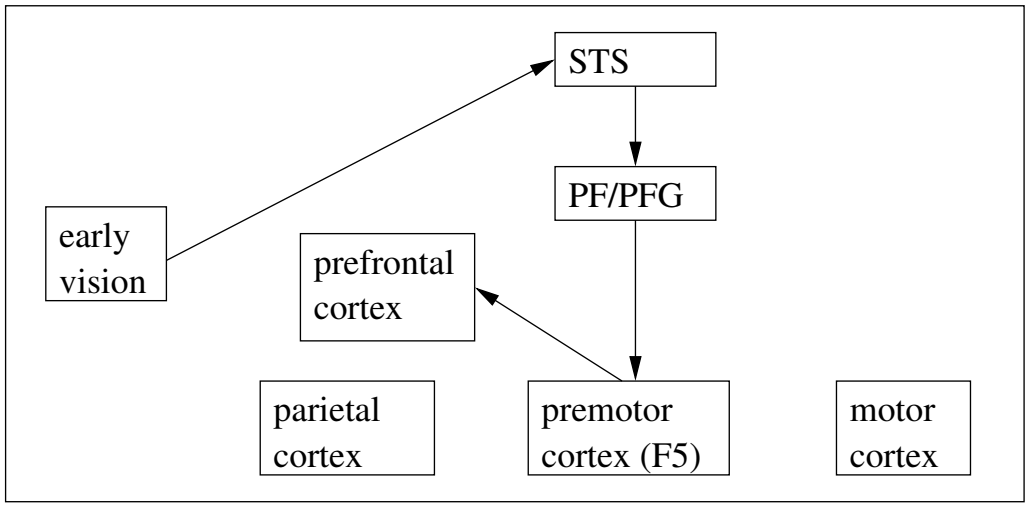
Evidence for this comes from lots of sources, including single-cell recordings in monkeys, and imaging/behavioural experiments on humans.

I'll review this evidence later in the lecture.

The action recognition circuit: overview

There are four key areas in a model of (visual) action recognition.

- Early visual processing (same as always).
- Superior temporal sulcus: biological motion and joint attention
- Premotor cortex: motor representations which are activated when an agent observes *another agent's actions*
- Prefrontal cortex: the highest level of action recognition is when the agent recognises the observed agent's *intentions*.



Outline of the lecture

1. STS: joint attention and biological motion perception
2. Mirror neurons and the mirror system
3. PFC: intention recognition during action observation
4. Visual perception of contact

Outline of today's lecture

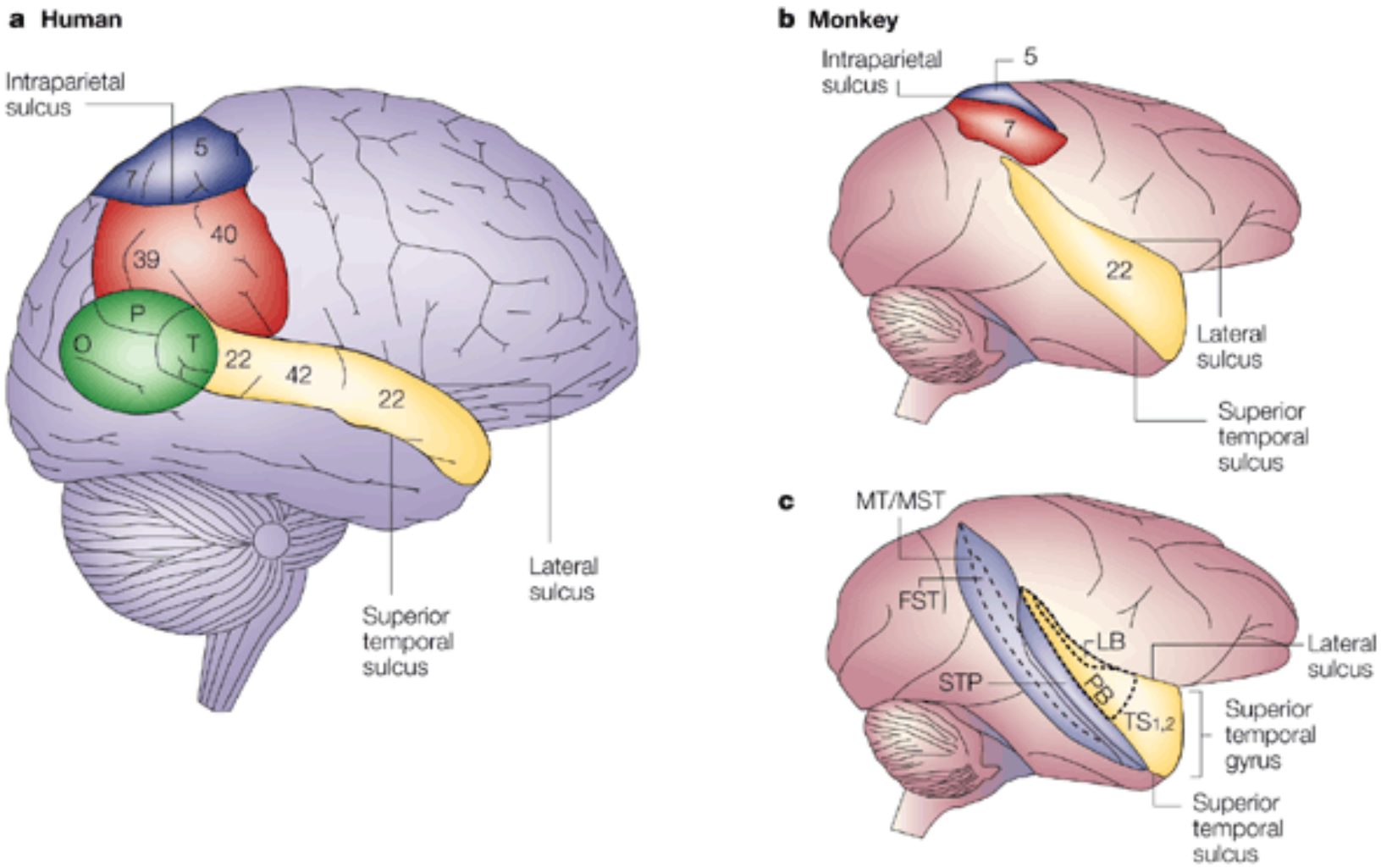
The start of the action recognition pathway

The action recognition pathway begins in the early visual system, just like action execution.

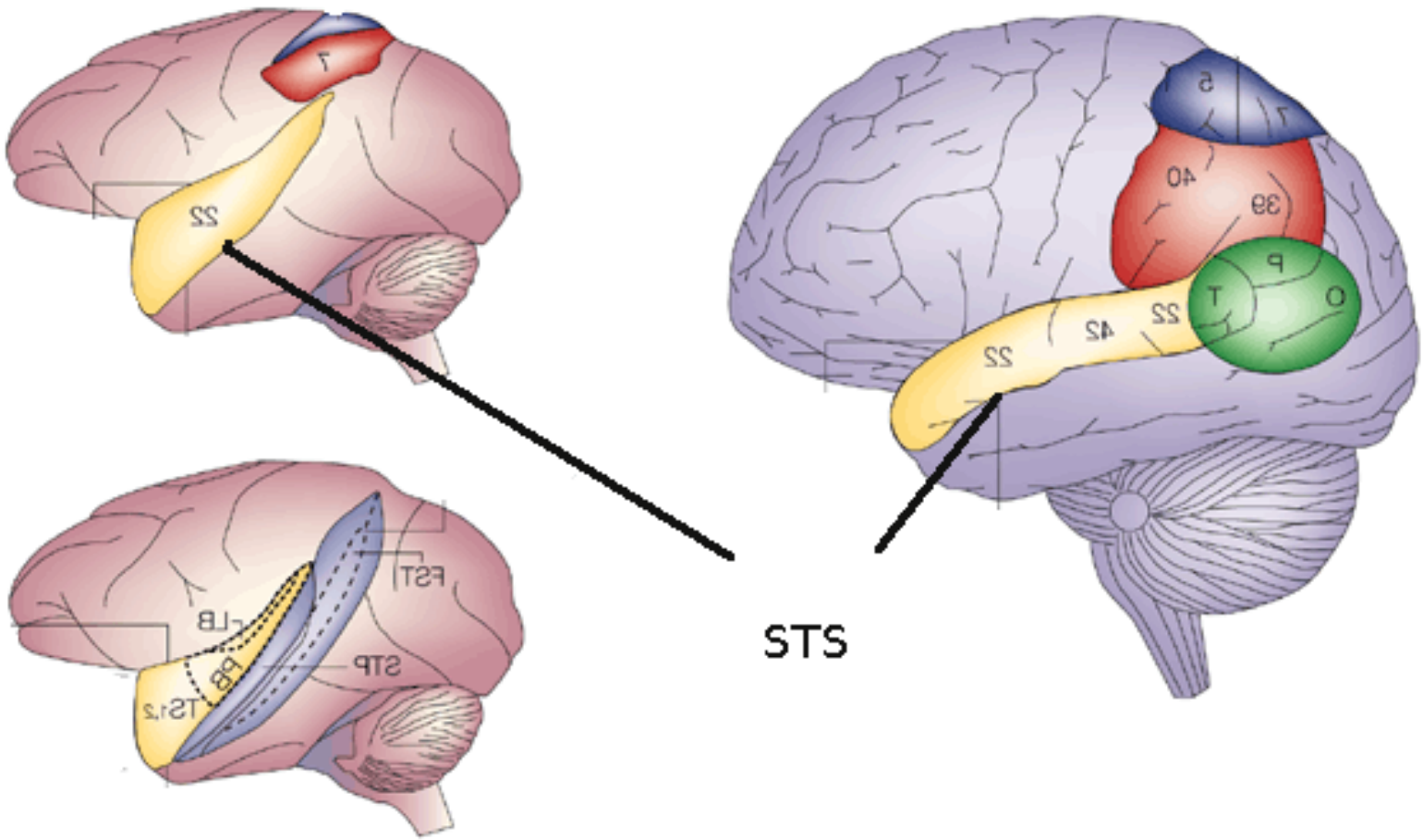
- However, it appears that action recognition employs some specialised visual machinery.
- Much of this is found within the superior temporal sulcus (STS).

The STS is involved in two separate processes which are important in action perception: biological motion perception and joint attention.

STS and related areas in humans and monkeys



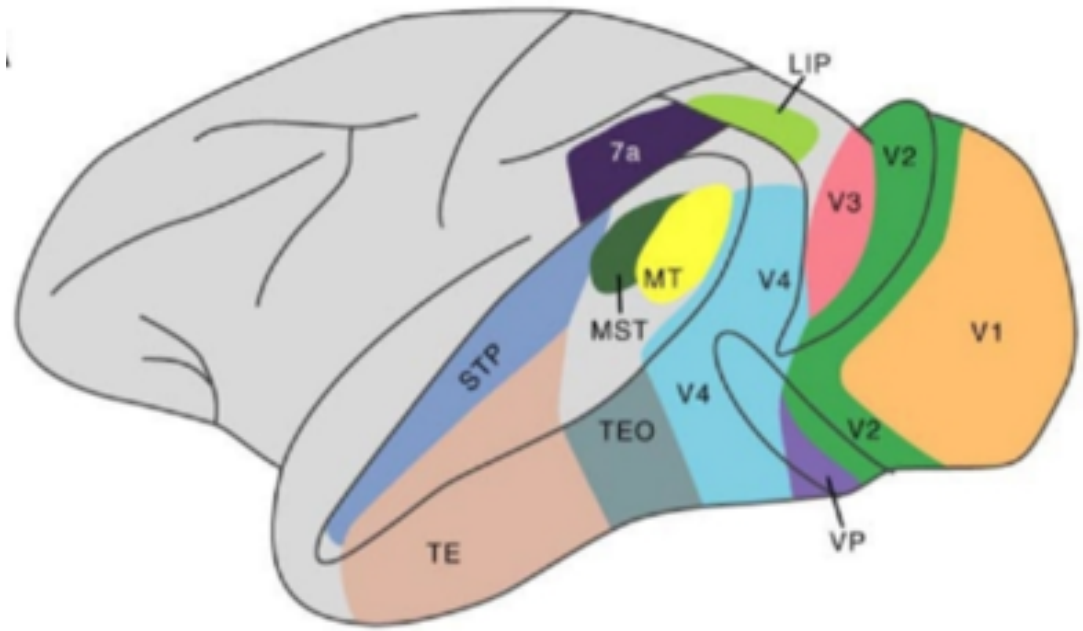
STS and related areas (flipped)



Visual pathways leading to STS

STS receives input from two main sources:

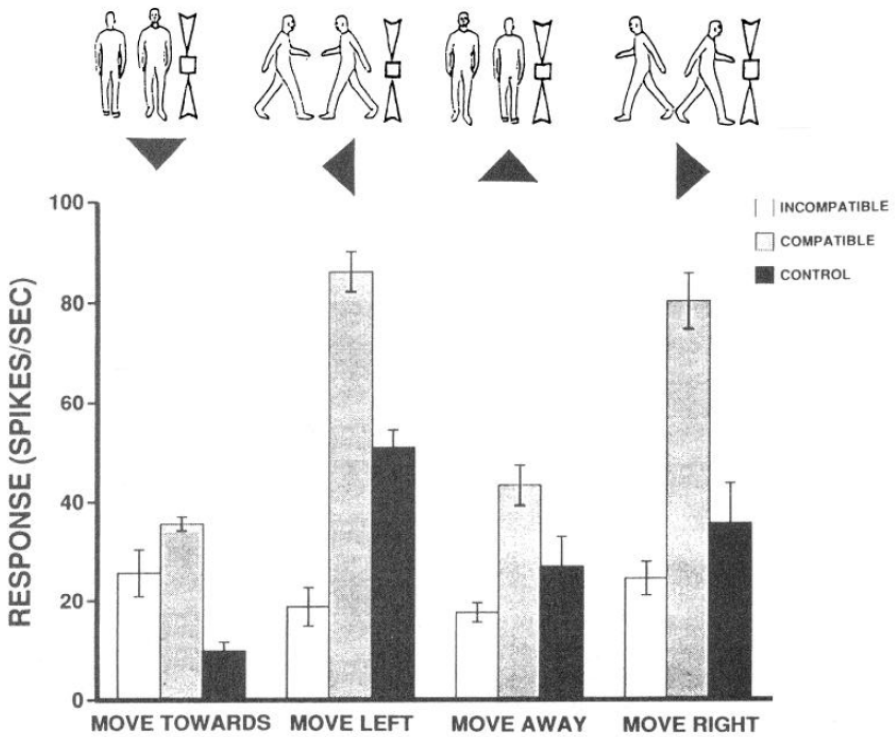
- The object classification pathway (culminating in TE)
- The 'visual motion' areas MT and MST.



It thus receives input from both the 'what' and 'where' pathways (Oram and Perrett, 1996).

STS and integration of form and motion cues

Many STS cells respond to particular combinations of form and motion.



This STP cell responds selectively to a figure walking left or right, but only if it's moving in the right direction (Oram and Perrett, 1996).

STS and biological motion recognition

An early discovery about action recognition is that observers are able to recognise actions from point-light displays (Johansson, 1973).

See biomotion lab demo.

STS and biological motion recognition

In 'biological motion' displays:

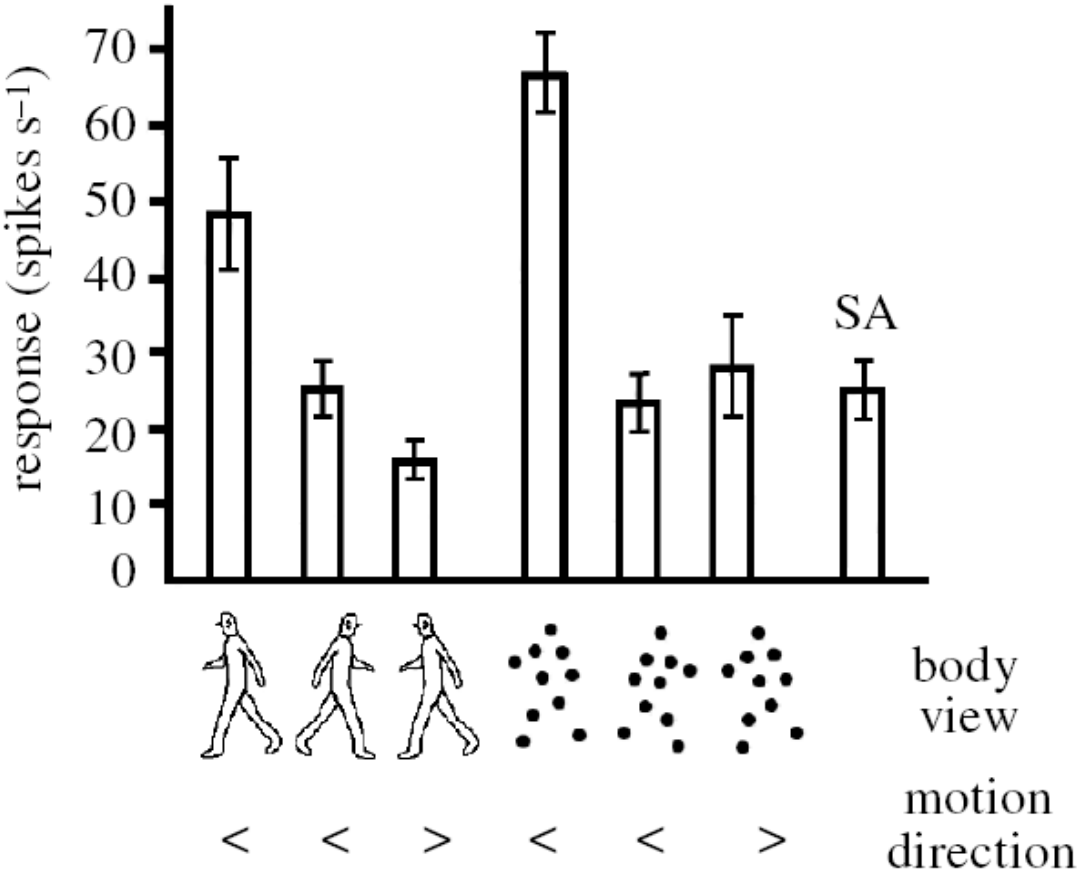
- Observers view films of an actor, in which the only visible elements are lights attached to the actor's arm and leg joints.
- If a single frame is displayed, observers recognise nothing. If a sequence of frames is played, observers immediately see an agent performing an action.

STS is involved in biological motion perception.

- Single cells in STS respond to specific actions, whether presented under point-light or full-view conditions (Oram and Perrett, 1994)

STS and biological motion recognition

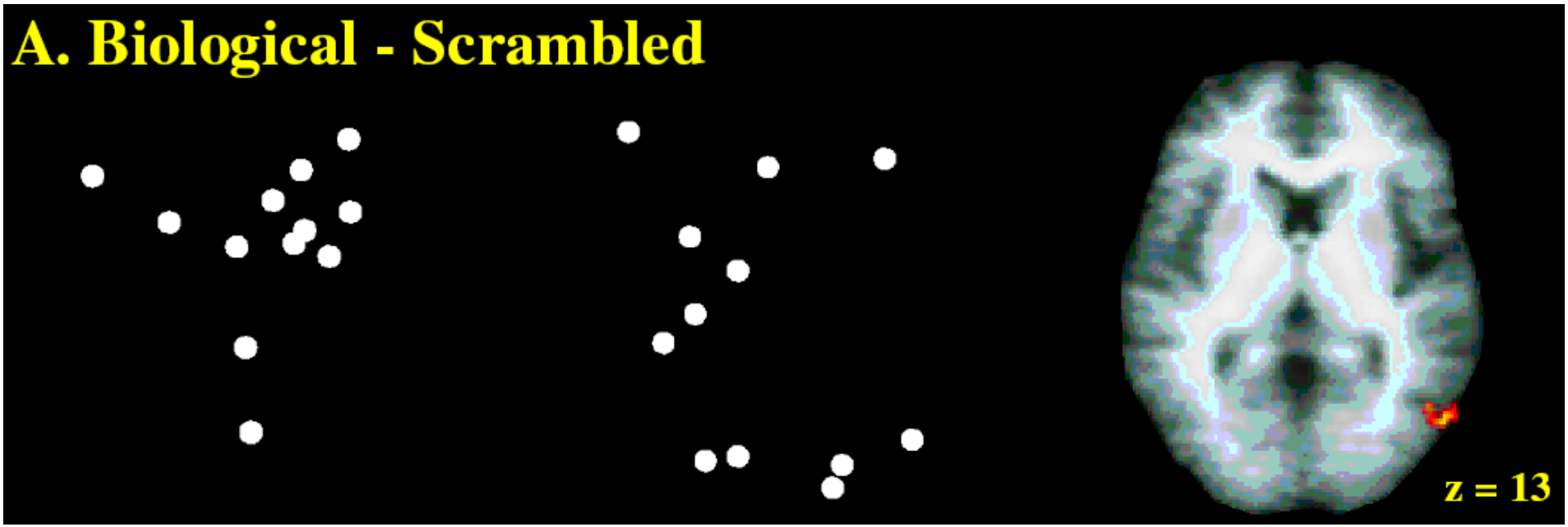
Here's an STP cell which responds to 'leftwards' walking, and which responds more strongly to a point-light stimulus than to a 'natural' one.



(Oram and Perrett, 1994)

STS and biological motion processing in humans

There are fMRI studies which suggest that STS is involved in processing biological motion in humans as well (see e.g. Grossman *et al.*, 2000).



The data are a little less clearcut here, though.

Does STS really process form as well as motion?

We might think that STS is just recognising actions based on motion information:

- relative motion of body parts;
- absolute motion of the agent in the environment.

However:

- STS receives inputs from IT/TE, which represents forms.
- There are many STS cells which respond only to static body shapes (see e.g. Oram and Perrett, 1996).
- We're better at recognising upside-down actions in video displays than in point-light displays.

Action recognition and joint attention

An observer watching an agent grasp a target needs to *attend to the target* to derive a full representation of the observed action.

Humans have a good capacity for joint attention (i.e. gaze following).

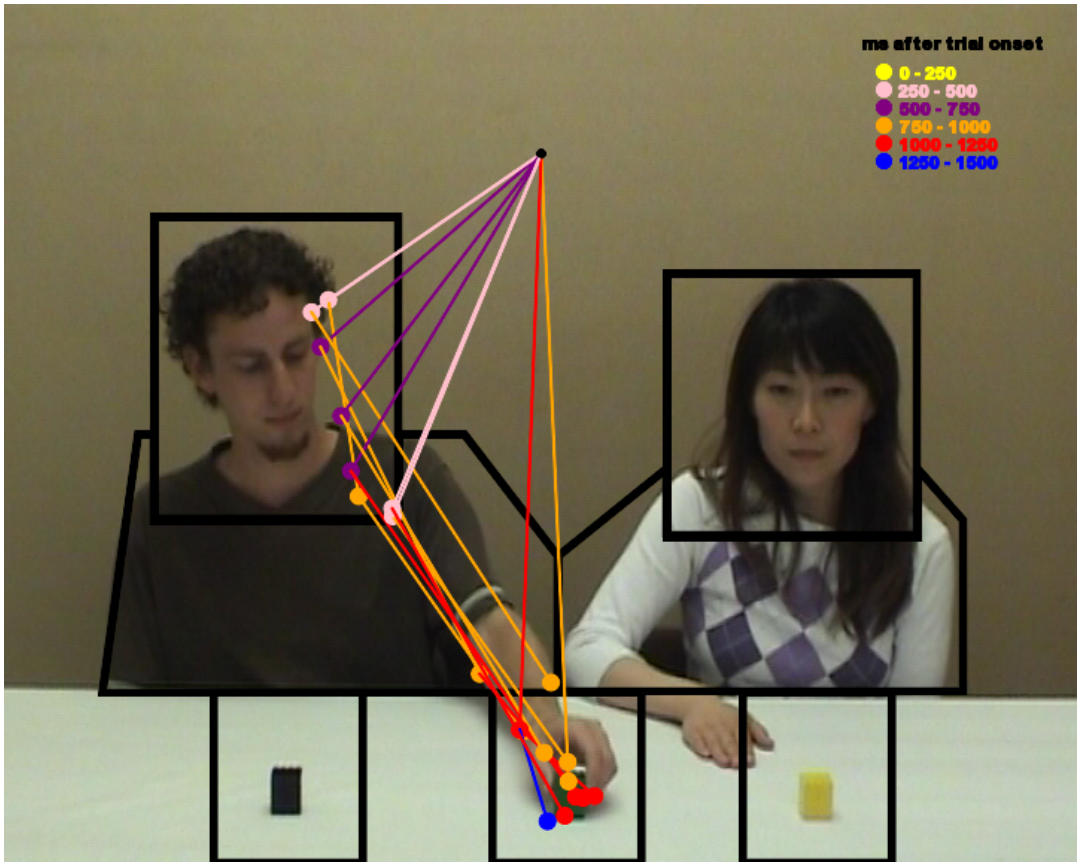
- Gaze following *captures attention* (Driver *et al.*, 1999)
- Early visual processing is improved at the point the observed agent is attending to (Schuller and Rossion, 2004).

Interestingly, observers watching an agent reach for a target object saccade to the target *in anticipation* of the agent's hand reaching the target (Flanagan and Johansson, 2003).

- This parallels the agent's own early saccade to the target object.

Action recognition and joint attention

Here's some typical saccade data for observers watching a reach-to-grasp action:

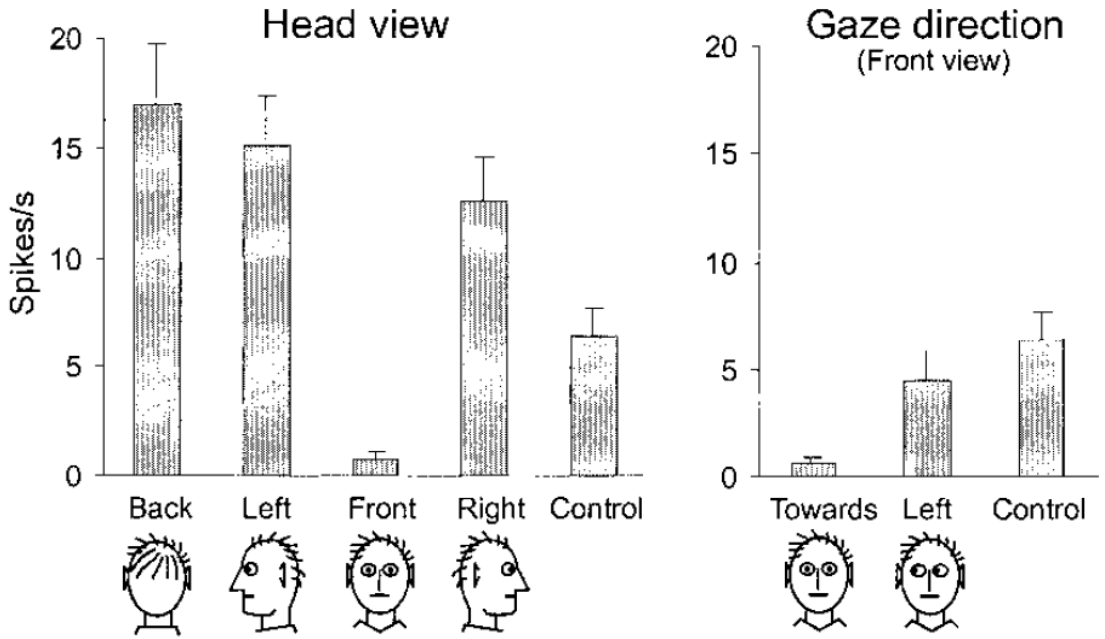


Note that saccades tend to reach the target object before the agent's hand does.

STS and gaze monitoring

STS is involved in gaze monitoring.

- Many STS cells are responsive to particular views of the head, and particular gaze directions (see e.g. the STS cell below, which encodes gaze to the left (from Jellema *et al.*, 2000).

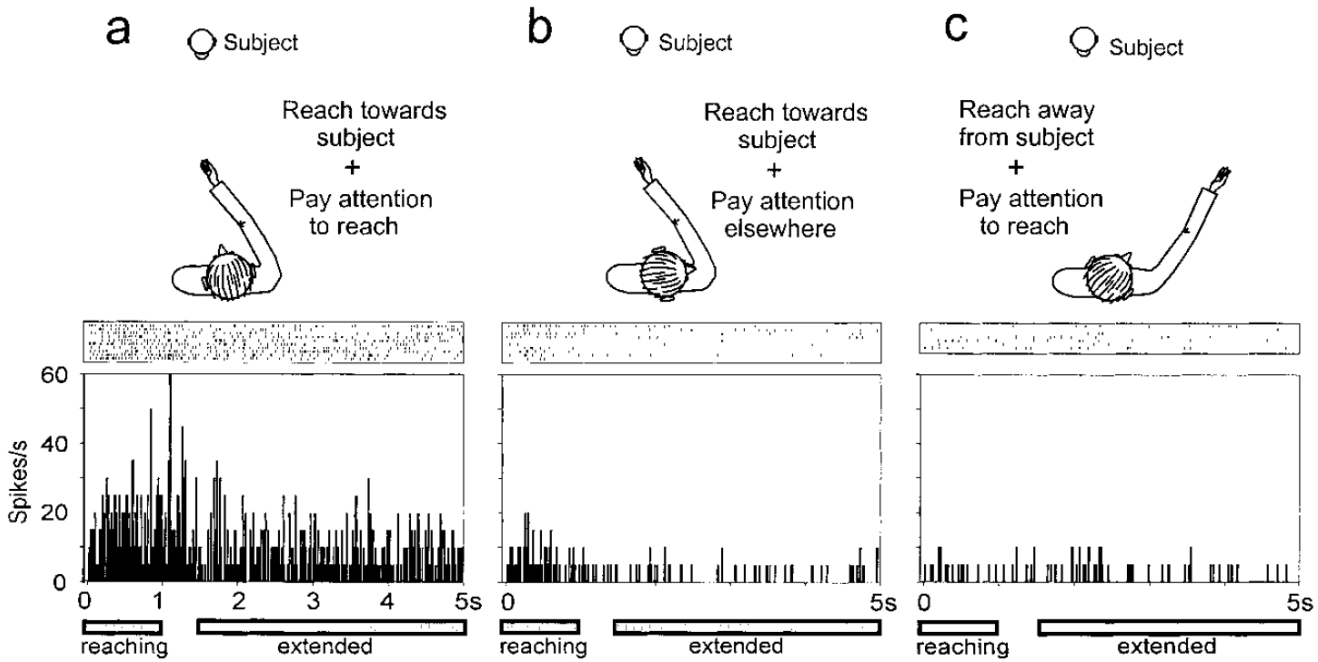


Note: STS projects to parietal areas involved in generating reflexive attentional shifts (Maioli *et al.*, 1998; Kingstone *et al.* 2000).

Integration of attentional and biological motion processing in STS

STS also integrates information about biological motion and joint attention.

- Jellema *et al.* (2000) found STS neurons which responded to an agent moving his hand towards a target, *but only if the agent was fixating the target*.



When do we start classifying an action?

Do we start classifying an action before we establish joint attention or after?

- Recall: observers saccade to the target *early* in action observation.
- Recall: many STS cells only encode actions if the agent is attending to the target.
- Note also: computational models of grasp perception (Oztop and Arbib, 2002; Oztop *et al.*, 2005) tend to assume knowledge of target location.

It's likely that observers identify an *intended target* before they classify the action they're watching.

Biological motion processing and object classification

The biological motion system generates the perception of an individual as well as of an action.

- STS identifies an individual as an animate agent, rather than as an object.

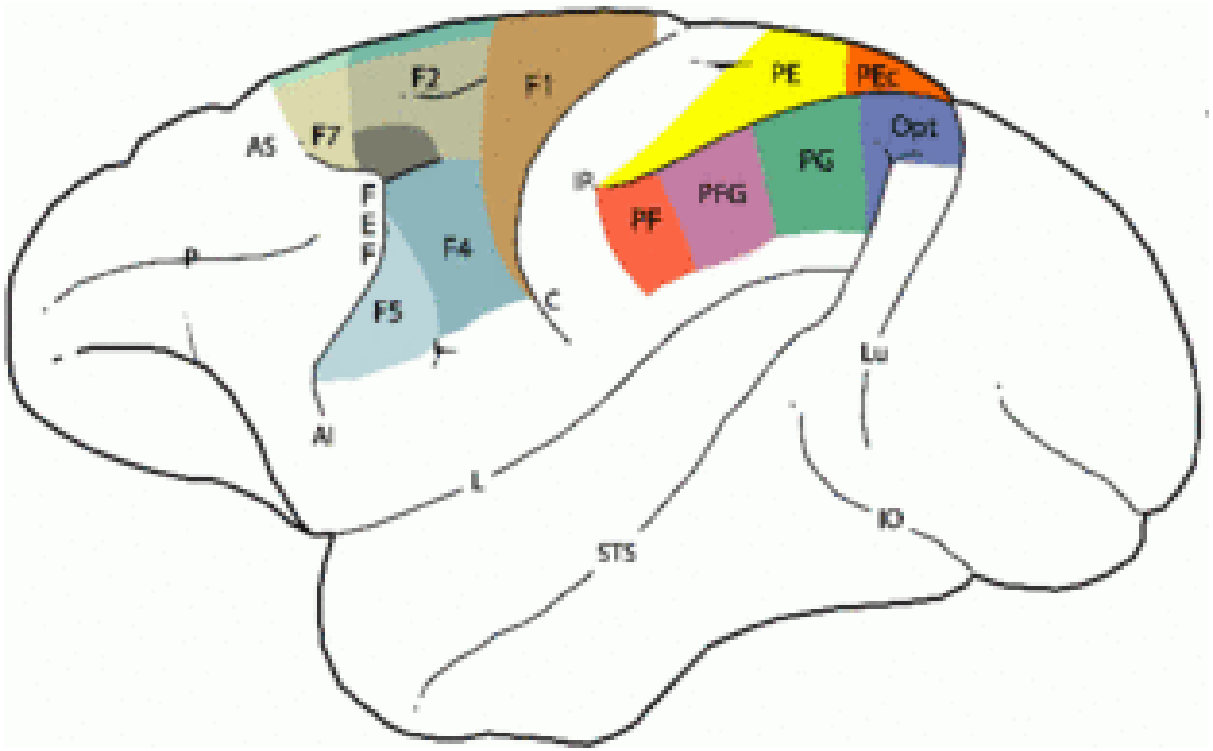
Action monitoring provides another opportunity for the integration of information across modalities:

- 'The individual you attended to before as an object is the same as the individual you are now *reattending to* as an animate entity'.

Outline of today's lecture

Mirror neurons: the original findings

'Mirror neurons' were first found in an area of macaque premotor cortex called F5 (di Pellegrino *et al.*, 1992).

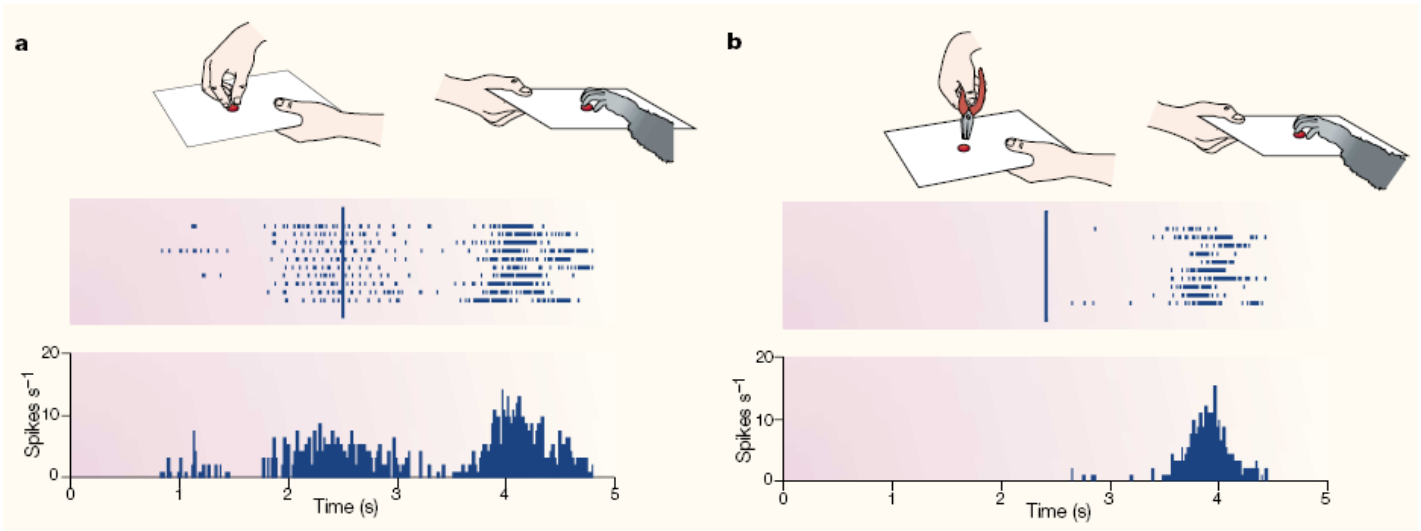


F5 is part of the grasp pathway.

Mirror neurons: the original findings

Most neurons in F5 respond to a specific type of grasp. (E.g. precision grip or whole hand grasp.)

- Canonical neurons in F5 respond to this grasp type only when it is performed by the monkey.
- Mirror neurons in F5 respond to this grasp type both when it is *performed* and when it is *observed* being done by another agent.



NB: mirror neurons must also fire when the monkey executes an action *without seeing its own hand*.

Other findings about mirror neurons

Many F5 mirror neurons are sensitive to combinations of *arm and finger movements*.

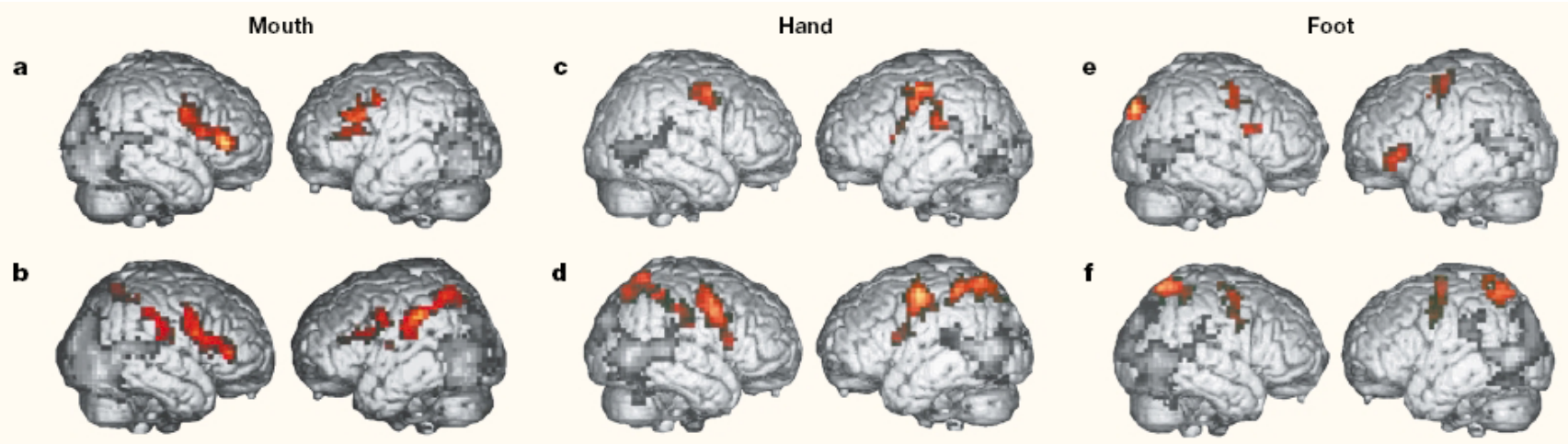
- F5 is part of the 'grasp' pathway (see Lecture 3).
- But recall that are *interactions* between the 'reach' and 'grasp' pathways.

Many F5 mirror neurons are sensitive to particular combinations of hand/arm actions and the *agents* or *objects* of these actions.

- Many mirror neurons fire when a particular action is done towards a particular target, but not to the target or the action by itself (Gallese *et al.*, 1996).
- Mirror neurons in F5c only respond to a hand action if the whole agent is in view (Nelissen *et al.*, 2005).

Evidence for the mirror system in humans

- Perception of an action results in sub-threshold activation of the observer's motor system (Fadiga *et al.*, 1995; Gangitano *et al.*, 2001)
- Evidence from imaging studies is less clearcut (see Turella *et al.*, 2007). However, there is some pretty good evidence that there are overlapping premotor regions (Bucino *et al.*, 2001).



From STS to F5

The proposal: during action recognition, STS provides input to F5/premotor regions.

- STS is linked to an inferior parietal region called PF/PFG. Some cells in this region also have mirror properties (Gallese *et al.*, 2002).
- PF/PFG is linked to F5—the premotor mirror neuron area.

Several people have suggested that activity flows from STS via PF/PFG to F5 during action recognition.

- Keysers and Perrett (2004): Hebbian learning
- Kilner *et al.* (2007): Bayesian inference.

Activation flowing from STS to F5 represents an *inference to the best explanation* of the observed agent's actions.

A simple model of the mirror neuron circuit

Assume a sequence of connected regions: STS↔PF/PFG↔F5.

The system is trained *during the agent's own actions*.

- The agent generates motor representations in F5 and PF/PFG.
- *The agent is watching his own action*, so representations are also produced in STS.
- A Hebbian/Bayesian rule creates associations between these representations.

In recognition mode, these associations allow STS representations to activate representations in PF/PFG, then in F5.

Outline of today's lecture

Intention recognition during action observation

F5 mirror neurons appear to fire when a grasp action is *inferred*, as well as when it's actually observed (Umiltà *et al.*, 2001).

Many PF/PFG cells seem to fire in anticipation of an observed agent's action (Gallese *et al.*, 2002).

- E.g. a PF/PFC cell may fire when the monkey picks up an object *and then brings it to his mouth*, but not when the monkey picks it up *and puts it in a box*.
- Many sequence-encoding cells also have mirror properties. These often fire after the first action in an expected sequence occurs.

It seems that F5 and PF/PFC areas encode the inferred intentions of an observed agent.

From F5 to PFC?

How are the observed agent's intentions recognised?

Maybe the predictions found in PF/PFG come from PFC.

- PFC encodes intentions during action execution...
- F5 is bidirectionally linked to PFC.

In action execution, PFC influences the forthcoming action.

- In action perception, maybe it generates *top-down expectations* about the forthcoming action.

Extending the circuit to PFC

Assume a sequence of connected regions: STS↔PF/PFG↔F5↔PFC.

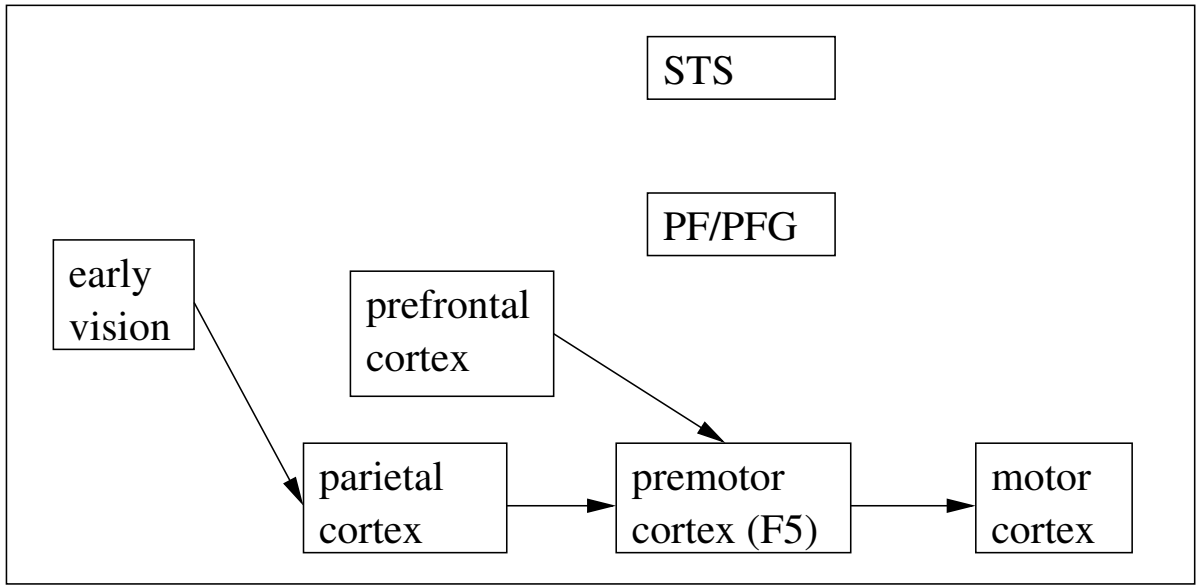
The system is trained *during the agent's own actions*.

- The agent's intentions in PFC generate action representations in F5 and PF/PFG.
- *The agent is watching his own action*, so representations are also produced in STS.
- A Hebbian/Bayesian rule creates associations between these representations.

In recognition mode, these associations allow STS representations to activate representations in PF/PFG, then in F5, then in PFC.

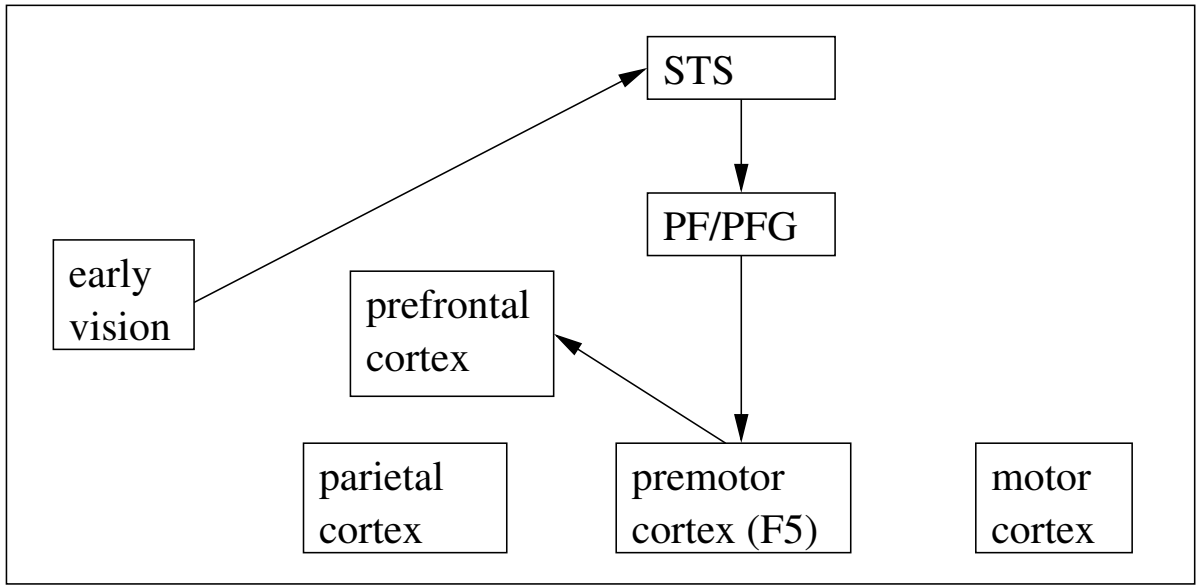
- But only if the observer *anticipates the target* when watching an action!

The action recognition circuit



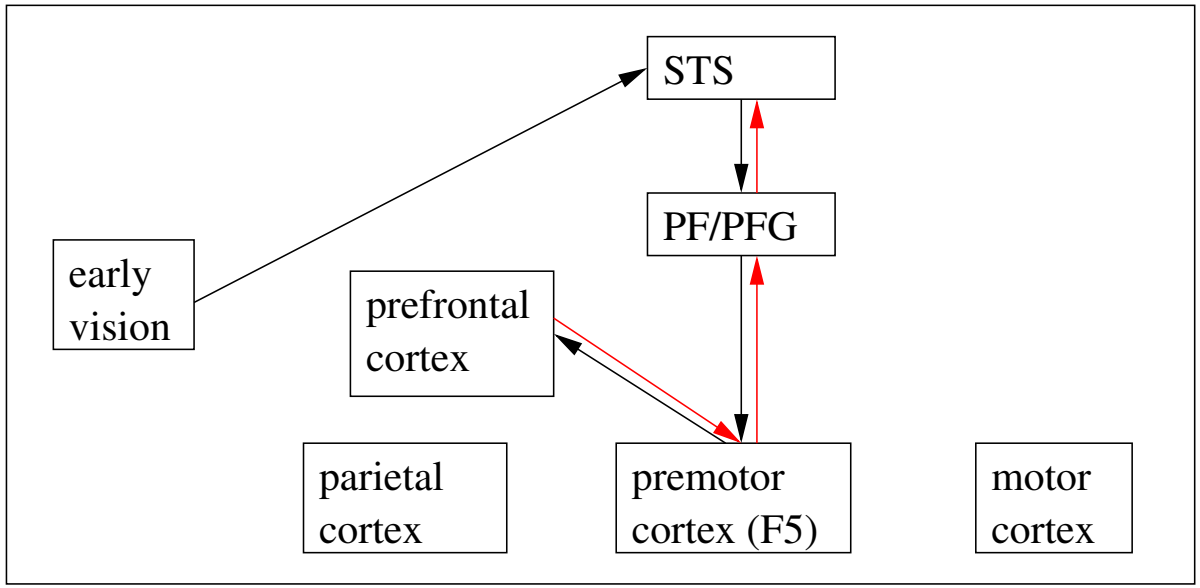
This is how action execution looks.

The action recognition circuit



Action recognition involves *some of* the same areas.
But with different connectivity.

The action recognition circuit



Action recognition involves *some of* the same areas.
But with different connectivity.

Outline of today's lecture

Visual perception of contact

An observer must be able to recognise that a perceived grasp action has been successfully completed. Note that haptic information will not be available in this case!

However, there is good evidence for a visual mechanism which makes contact with the relevant haptic circuits.

- When a human agent observes another person being touched, this generates the same kind of activity in secondary somatosensory cortex as is generated when they are touched themselves (Keysers *et al.*, 2002)
- Again, these cross-modal connections could be learned through Hebbian mechanisms, when an agent watches events which generate touch sensations in himself (Keysers and Perrett, 2004).

Summary

- Action recognition is very important in social animals.
- There is good evidence that humans (and other primates) use representations in premotor cortex to encode the actions of other agents they observe.
- The visual action recognition pathway: V1 $\rightarrow$ STS $\rightarrow$ PF/PFG $\rightarrow$ F5 $\rightarrow$ PFC.
- This pathway is trained while an agent observes *its own movements*.
- Recognition of a grasp action has sequential structure: we first identify the agent's intended target (and fixate it); then we use the mirror system to classify his action.