From Action Recognition to Language

Part I
Schemas and Cooperative Computation

Preshaping While Reaching to Grasp

Hypothetical coordinated control program for reaching and grasping

Schemas are the "Programs" for Cooperative Computation

... based on the competition and cooperation of concurrently active schema instances.

Cooperation yields a pattern of "strengthened alliances" between mutually consistent schema instances that allows them to achieve high activity levels to constitute the overall solution of a problem

* perceptual schemas become part of the current short-term model of the environment;
* motor schemas contribute to the current course of action

As a result of Competition, instances which do not meet the evolving consensus lose activity, and thus are not part of this solution (though their continuing subthreshold activity may well affect later behavior).
Cooperative Computation
The HEARSAY Paradigm for Speech Understanding

Knowledge sources act to transform data between levels of the "blackboard". Hypotheses compete and cooperate to form part of the final interpretation of the utterance.

(This anticipates Jackendoff’s recent invocation of blackboards in his cognitive-linguistic architecture)

Arbib & Caplan (BBS, 1979) – Neurolinguistics Must be Computational – offered some speculations on how this computer model might be reconceptualized to exemplify Luria’s views on “distributed localization”.

The VISIONS system
Allen Hanson and Edward Riseman of UMass (a classic but not the state of the art)

The Example We Will Work With
(but the color is better in the original!)

A First Pass at Segmenting the Image
The Sky – Data-Driven First Approximation

The Roof – Hypothesis-Driven First Approximation

The Walls – Hypothesis-Driven First Approximation

Focusing the Resegmentation: Localizing the Conflict Between Sky and Wall Schemas
Resegmentation

The End Result

Part II
Visual Control of Grasping

“What” versus “How” in Human

DF: Jeannerod et al.
Lesion here: Inability to Preshape
(except for objects with size “in the semantics”)

AT: Goodale and Milner
Lesion here: Inability to verbalize or pantomime size or orientation

reach programming

Parietal Cortex

How (dorsal)

Visual Cortex

Inferotemporal Cortex

What (ventral)

grasp programming

Monkey Data:
Mishkin and Ungerleider on
“What” versus “Where”

The Sakata Protocol

A key theme of visuomotor coordination: parietal affordances (AIP) drive frontal motor schemas (F5)

F5 - grasp commands in premotor cortex
Giacomo Rizzolatti
Hideo Sakata

Grip Selectivity in a Single AIP Cell

A cell that is selective for side opposition (Sakata)
Differential Timing of Activity Peaks in Different AIP Neurons

Note the need for a broad database of many cells within each region to see that cells are not just “pattern recognizers” but also have a relationship to the time course of the ongoing behavior.

The Complete FARS Model

To code some variable $x$ lying in an interval $[a, b)$ we could take $n$ cells, with cell $i$ ($i = 0, \ldots, n-1$) firing if and only if the current value of $x$ lies in the $i$th subinterval

$$a + \frac{(b-a)j}{n} < x < a + \frac{(b-a)(j+1)}{n}$$

In coarse coding, we achieve much greater discrimination by taking into account the continuously varying firing level $f_i$ of each cell, and then we can decode values of $x$ actually varying across each interval, using some such formula as

$$\sum_{i=0}^{n-1} f_i \left( a + \frac{b-a}{n} \right)$$

Note: In the Georgopoulos study, we saw “negative votes” for firing below the neuron’s resting discharge rate.
The “Visual Front End” of the FARS Model

Visual Cortex

Parietal Cortex

VIP

(arm goal position)

VIP

(object/grasp transform)

How (dorsal)

VIP

(ventral)

F4

F5

AIP

Visual Cortex

What (ventral)

IT

(grasp type)

Note the use of coarse coding

In the paper we spoke of PIP where we now say cIPS.

A lens activates a precision grasp with a wide aperture.

A bottle cap activates a precision grasp with a narrow aperture.

The mapping from object identity in IT to maps directly to both the grasp type and the aperture of grasp in AIP when the nature of the object implies such data:

E.g., in the case of AT, the projection from IT can provide the necessary grasp type and parameters for a lipstick but not for a cylinder.

cIPS® → IT connections are hard-wired for a simple set of objects

cIPS® → AIP connections are hard-wired for a simple set of affordances

IT → AIP

A jar top maps to a precision grasp with a wide aperture.
F5 activity during execution of a precision grasp

The top two traces show the position of the thumb and index finger. 

**Left:** The next five traces represent the average firing rate of five F5 neurons (set-, extension-, flexion-, hold-, and release-related). The remaining five traces represent the various external (Ready, Go, Go2) and internal (SII) triggering signals. 

**Right:** Illustrating the temporally distributed coding of F5 cells.

---

**Hand-Arm Model; Modeling the Reach**

The aim of ILGM is to provide a wide range of hand object contact possibilities to mimic the infant’s variable reaches and grasp attempts, which mediate learning.

ILGM’s learning routine will exploit whatever successful grasps occur to learn appropriate grasps to use when the hand approaches the object in the manner specified by MG.

The reach component involves the computation of trajectories to achieve a desired position for the end effector (i.e. the inverse kinematics problem). The end effector used in the simulation was the tip of either the index or middle finger

The Jacobian transpose method is used for inverse kinematics.

---

**Infant grasping**

Vision is not fully utilized for grasp programming: no anticipatory pre-shaping before contact with the object.

So it is hard to accept that the mirror neuron system can compute the compatibility of another’s hand motion to the object in this period.

---

**Structure of ILGM**

The individual layers are trained based on somatosensory feedback

The affordance layer is used to convey the affordances available to the model.

In the present simulations, the affordance layer is engaged in encoding either the existence or the orientation or the position of the target object
Infants as young as 2 months of age play with their own hands, manipulate objects put in their hands, and play with rattles (Bayley 1936). When the hand contacts a glowing or sounding object in the dark, infants as young as 11 weeks try to grasp it (Clifton et al. 1993). The resulting tactile stimuli appear to motivate infants to engage in grasping and holding.

Our hypothesis: We propose that the sensory feedback arising from the stable grasp of an object is a uniquely positive, reward – *joy of grasping* – for motivating the infant to explore and learn actions that lead to grasp-like experiences.

ILGM is a probabilistic neural network architecture that uses ‘joy of grasping’ as the reward stimulus and employs adaptive mechanisms similar to those used in reinforcement learning. However, rather than maximize cumulative reward, ILGM seeks to discover the repertoire of rewarding actions that afford maximal adaptive flexibility.

**SE1b – Example Result**

A. After learning, ILGM planned and performed power grasps using different objects.

B: Two learned precision grips with a cube shaped object.

C: ILGM was able occasionally to generate two fingered precision grips.

**SE2a – Example Result**

**SE2b – Example Result**

---

**Part III**

The Mirror System for Grasping
Lecture 18.

The Mirror Neuron System Model (MNS)

Reading Assignment:

F5 Motor Neurons

F5 Motor Neurons include all F5 neurons whose firing is related to motor activity.
- We focus on grasp-related behavior. Other F5 motor neurons are related to oro-facial movements.
F5 Mirror Neurons form the subset of grasp-related F5 motor neurons of F5 which discharge when the monkey observes meaningful hand movements.
F5 Canonical Neurons form the subset of grasp-related F5 motor neurons of F5 which fire when the monkey sees an object with related affordances.

Mirror Neurons

Rizzolatti, Fadiga, Gallese, and Fogassi, 1995: Premotor cortex and the recognition of motor actions

Other
Self

The effective observed movement

AIP → F5 canonical: affordance recognition
PF → F5 mirror: action recognition

Computing the Mirror System Response

The FARS Model:
Recognize object affordances and determine appropriate grasp.

The Mirror Neuron System (MNS) Model:
We must add recognition of
* trajectory and
* hand preshape
to
* recognition of object affordances
and ensure that all three are congruent.

There are parietal systems other than AIP adapted to this task.

Modeling Challenges:
1) To have mirror neurons self-organize to learn to recognize grasps in the monkey’s motor repertoire
2) To learn to activate mirror neurons from smaller and smaller samples of a trajectory.
Hand State

We capture the hand and its relation to the target object information in the Hand State, a 7-dimensional trajectory $H(t)$ with the following components:

$H(t) = (d(t), v(t), a(t), o_1(t), o_2(t), o_3(t), o_4(t))$

- $d(t)$: distance to target at time $t$
- $v(t)$: tangential velocity of the wrist
- $a(t)$: Aperture of the virtual fingers involved in grasping at time $t$
- $o_1(t)$: Angle between the object axis and the (index finger tip – thumb tip) vector
- $o_2(t)$: Angle between the object axis and the (index finger knuckle – thumb tip) vector
- $o_3(t), o_4(t)$: The two angles defining how close the thumb is to the hand as measured relative to the side of the hand and to the inner surface of the palm.

Note that the whole history of $H(t)$ during a grasp is required to represent the grasp.

Key task

To determine whether the motion and preshape of a moving hand may be extrapolated to culminate in a grasp appropriate to one of the affordances of the observed object.

Reminder: Hand State components

For most components we need to know (3D) configuration of the hand.
• The Vision task is simplified using colored tapes on the joints and articulation points
• The First step of hand configuration analysis is to locate the color patches unambiguously (not easy!).

Use color segmentation. But we have to compensate for lighting, reflection, shading and wrinkling problems: Robust color detection

Thus we have 2 steps:
1. Extract the color marker positions
2. Estimate 3D pose

Training phase: A color expert is generated by training a feed-forward network to approximate human perception of color.

Actual processing: The hand image is fed to an augmented segmentation system. The color decision during segmentation is done by the consulting color expert.

Step 1 of hand shape recognition: system processes the color-coded hand image and generates a set of features to be used by the second step

Step 2: The feature vector generated by the first step is used to fit a 3D-kinematics model of the hand by the model matching module. The resulting hand configuration is sent to the classification module.
Virtual Hand/Arm and Reach/Grasp Simulator

A precision pinch

A power grasp and a side grasp

Power grasp time series data

*: aperture; #: angle 1; x: angle 2; •: 1-axisdisp1; ●: 1-axisdisp2; ◆: speed; ■: distance.

Core Mirror Circuit

Curve recognition system demonstrated for hand drawn numeral recognition (successful recognition examples for 2, 8 and 3).
What is to be learned?

- Connections from hand-state and object-affordance are adapted so that association neurons will respond when hand-state is congruent with object.
- Connections from association neurons are adapted so that their integrated activity will activate mirror neurons for the appropriate grasp.

**Power and precision grasp resolution**

Note that the modeling yields novel predictions for time course of activity across a population of mirror neurons.

**Did the Mirror System First Evolve for Social Understanding?**

My counter-hypothesis: Subdividing the emergence of the mirror system:

- a: The mirror system for grasping evolved originally to provide visual feedback for those hand movements requiring attention to object detail.
- b: Expectation: Exploiting this “self-ability” to map other individual’s actions into internal motor representation.

b: Recognize an action A; recall (perhaps implicitly) that action A is most likely to have consequences B or C; determine from the context that B is more likely to occur, and then use this expectation to speed the choice of a course of action appropriate to B.

But more is needed for imitation: Observing that a novel action A achieves goal B in context W, master the skill of performing A, and then use it next time B is one’s goal in a context related to W.
The child refines a crude map (superior colliculus) to make unstructured reach and "grab" movements at objects (Kuperstein et seq.)

The child develops a set of grasps which succeed by kinesthetic, somatosensory criteria. (ILGM*)

AIP develops as affordances of objects become learned in association with successful grasps. (Future modeling)

The (grasp) mirror neuron system develops driven by the visual stimuli relating hand and object generated by the actions (grasps) performed by the infant himself. (MNS1*)

Then the infant acquires the ability to create motor representations for novel actions observed.

*“Imitation”: Rudimentary in chimps; a crucial ability of humans

The Moral of the Story: The use of models of adaptive neural networks supports the view that the repertoire of the mirror system is acquired rather than innate.*
Placing action and language in a shared framework: Goal-Directed Action

A Motor “Sentential Form

An action such as this one is a “sentence” made up of “words” which are basic actions akin in complexity to “reach and grasp”.

A hierarchy expressed as sequence whose subsequences are not fixed in length but instead are conditioned on the achievement of goals and subgoals.

A “paragraph” or a “discourse” might then correspond to, e.g., an assembly task which involves a number of such “sentences”.

From Communicative Goal to Disambiguating Sentential Form

Communicative goal: Get a waiter to serve the intended customer:

Mixed strategy:
- Serve that man.
- Using deixis (pointing) to disambiguate which man.

Sentence planning strategy:
Repeat the “loop” \(<\text{add adjective or relative clause}>\) until (you think) ambiguity is resolved:
- Serve the man on the left.
- Still ambiguous?
- Serve the old man on the left.
- Still ambiguous? Apparently not, so say it to the waiter.
- But the waiter veers off in the wrong direction:
- No, no. The one who is reading a newspaper.

Adapting and “Unfurling” a nested hierarchical structure to extract a set of actions to reach a communicative goal.

A Bridge from Action to Syntax

ASL: American Sign Language

Classifier constructions:
The use of signing space to represent observed/imagined space

<table>
<thead>
<tr>
<th>HOUSE</th>
<th>BIKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>located here</td>
<td>Located here</td>
</tr>
</tbody>
</table>

The bike is near the house

The study of signed languages can create a unique bridge between the study of brachio-manual action and the study of language.

Adapted from slide supplied by Karen Emmorey

Neural systems underlying signed language

A direct contrast between speaking (purple) and signing (red)

Pantomime and signing dissociate with left hemisphere damage (Corina)
But there is no difference between “pantomimic” and non-pantomimic signs
The Mirror System Approach to the Evolution of Human Language

Monkey F5 (with its mirror system for grasping) is homologous to human Broca’s area. This Mirror System Hypothesis is that the evolutionary basis for language parity is provided by the mirror system for grasping, rooting speech in communication based on manual gesture. This provides a neural basis for the claim that hand movements grounded the evolution of language. From “praxis” to communication

Primate Evolution: Two Key Branch Points

Adapted from Clive Gamble: Timewalkers Figure 4.2

Deep Time

The divergence of the Romance languages took about one thousand years.
The divergence of the Indo-European languages with their immense diversity
• Hindi, German, Italian, English, ...
took about 6,000 years.
How can we imagine what has changed since the emergence of Homo sapiens some 200,000 years ago?
Or in 5,000,000 years of prior hominid evolution?
I will develop a series of “Just So” stories to establish the plausibility of a set of hypotheses which allows us to realize that Bickerton’s view – which seems widely shared – is by no means firmly established.
The debate must continue!

Language

From action-object frames to verb-argument structures:

Minimal Subscene

Action-Object Frame

Semantics

Verb-Argument Structure

Syntax

Sentence

But just as our attention to agent and action and so on can yield increasing detail of current relevance, so must language develop the tools to express these details in a coherent way …. And so we get to syntax and a compositional semantics…
The basis for the co-evolution of cognitive & linguistic complexity
distinguish the two meanings providing an "incentive" for coming up with an arbitrary signifying "grasping" from one meaning "a [graspable] raisin", thus in pantomime it might be hard to distinguish a grasping movement with meanings in more or less arbitrary fashion. Available as elements for the formation of compounds which can be paired divorced from their pragmatic origins (if such existed) and abstract gestures And here is a crucial transition: observed The imitator observes; the panto-mimer intends Pantomime is performed with the intention of getting the observer to think Imitation is the generic attempt to reproduce movements performed by in which similar actions are produced away from the goal object/pragmatic action directed towards a goal object pantomime in which similar actions are produced away from the goal object Imitation is the generic attempt to reproduce movements performed by another, whether to master a skill or simply as part of a social interaction. Pantomime is performed with the intention of getting the observer to think of a specific action or event. It is essentially communicative in its nature. The imitator observes; the panto-mimer intends to be observed

A Protosign Fossil® in ASL: Noun/Verb Pairs Differentiated by Movement

[*ASL is a fully formed human language, not a protolanguage] A change in the movement can change the meaning of a given handsform.
Two keys to human brain evolution to support language:

- The capacity to give names to yield the **lexicon** which underlies the ability to refer to objects or events in the external world.
- **Syntax** arose to express regularities in the ways in which different elements are combined to form utterances.

Selective pressure for the capacity to learn complex vocalizations through imitation and repeated practice was a key aspect in establishing a **phonological working memory system** that allowed temporary storage of phonological representations in order to rehearse them internally.

**Syntax** arose to express regularities in the ways in which different elements are combined to form utterances.

- Support: differentiation of an incipient Broca's area with its connections to the incipient Wernicke's region developing as a **phonological rehearsal device** that differentiated into the language areas.
  - Broca's Area (44+45(h) 44+45(m) is connected with 39+40(h) 7b+7ip(m))
  - The coordinated operation of networks involving granular frontal cortex and the semantic system represented in the temporo-parietal lobes, together with the phonological rehearsal loop to generate higher levels of syntax and discourse.
A Multiplicity of Working Memories: Aboitiz & Garcia (1997)

Linguistic working memory
- the phonological-rehearsal loop
39,40 ← BA44-47
Working memory for spatial location of objects
7 ← 46 & 8
Working memory for object characteristics
TE ← 45 & 12
Coordinated activity of working memory circuits
FGC, frontal granular cortex: areas 9 and especially 46
- distribution of attention
- handling of cognitive (semantic) information relevant for language processing
- linking object/feature information with visuospatial information

Towards a Mirror Neurolinguistics

Working Memory for Perception and Production of Utterances

Protopsych (basis for speech)
From Imitation to Describing Actions, Objects, Relationships

A Brief Comparison

Aboitiz and Garcia offer a retrospective theory
- They assume that the human brain evolved (in part) to support language.
- They look at the features of the human brain, seek the homologous areas of the macaque brain, note what has changed (some areas enlarge, some connections are strengthened) and then suggest how these changes could support a lexicon of spoken words and a syntax to bind them into sentences.

Rizzolatti and Arbib offer a prospective theory
- They start from an analysis of the monkey’s capabilities, especially the fact that species-specific vocalizations have their cortical outpost in the anterior cingulate but that a different area, involved in hand movements, is homologous to Broca’s area. This then sets us off to hypothesize how intermediate stages from the mirror system for grasping → action recognition → imitation → protosign → protospeech yield a brain rich enough to support the incremental invention and learning of languages.
- Arbib (2002) argues that humans evolved to have a language-ready brain, with the richness of human languages a “post-biological accumulation of inventions.”
- However, Rizzolatti and Arbib are relatively silent on the phonological loop and other working memories

Selected References