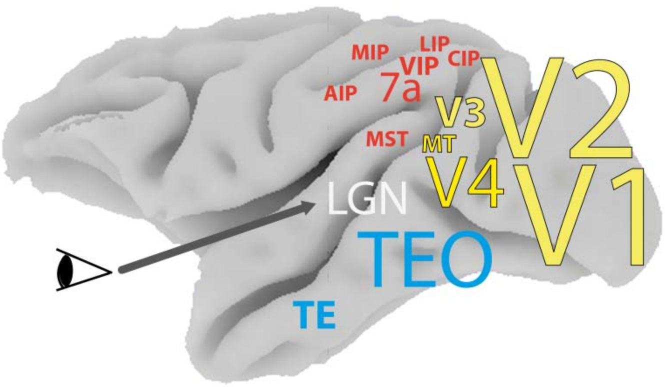


From Early Vision to Symbols

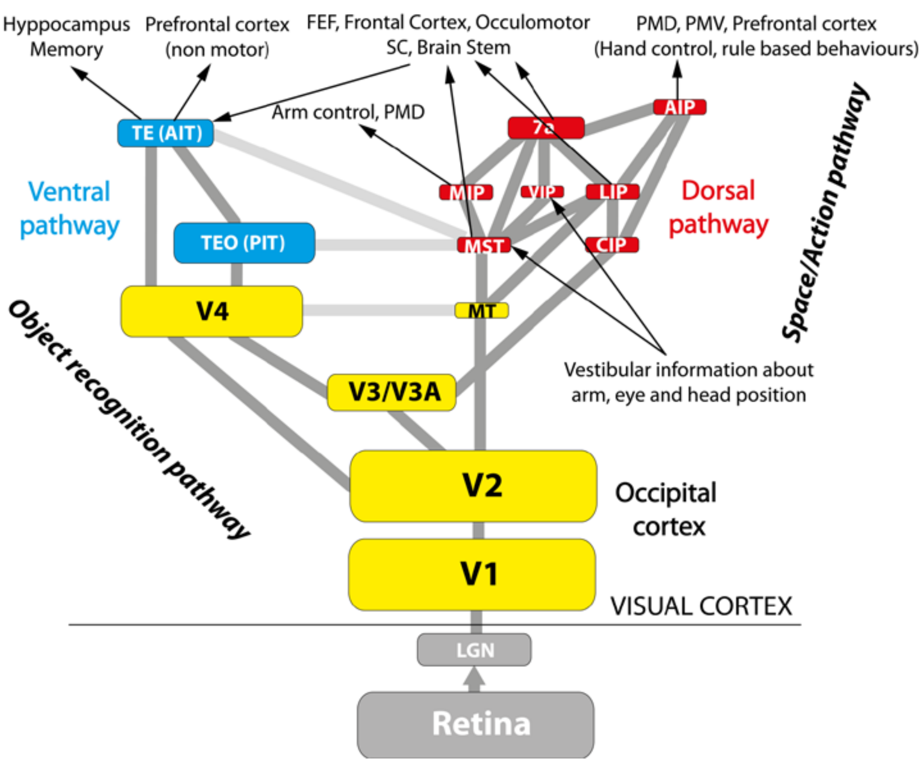
Norbert Kruger

University of Southern Denmark

Cognitive and Applied Robotics Group

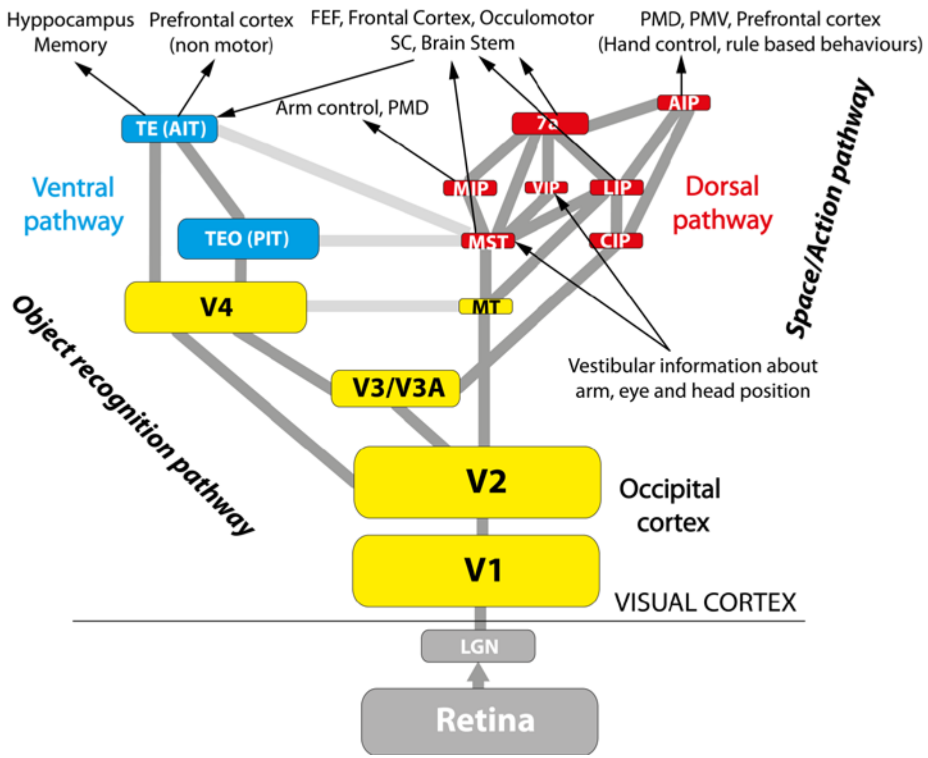
Overview

- **Background Information**
- The primate's vision system: A deep Hierarchy
- From Signals to Symbols: Birth of the Object and its affordances
- Reflections

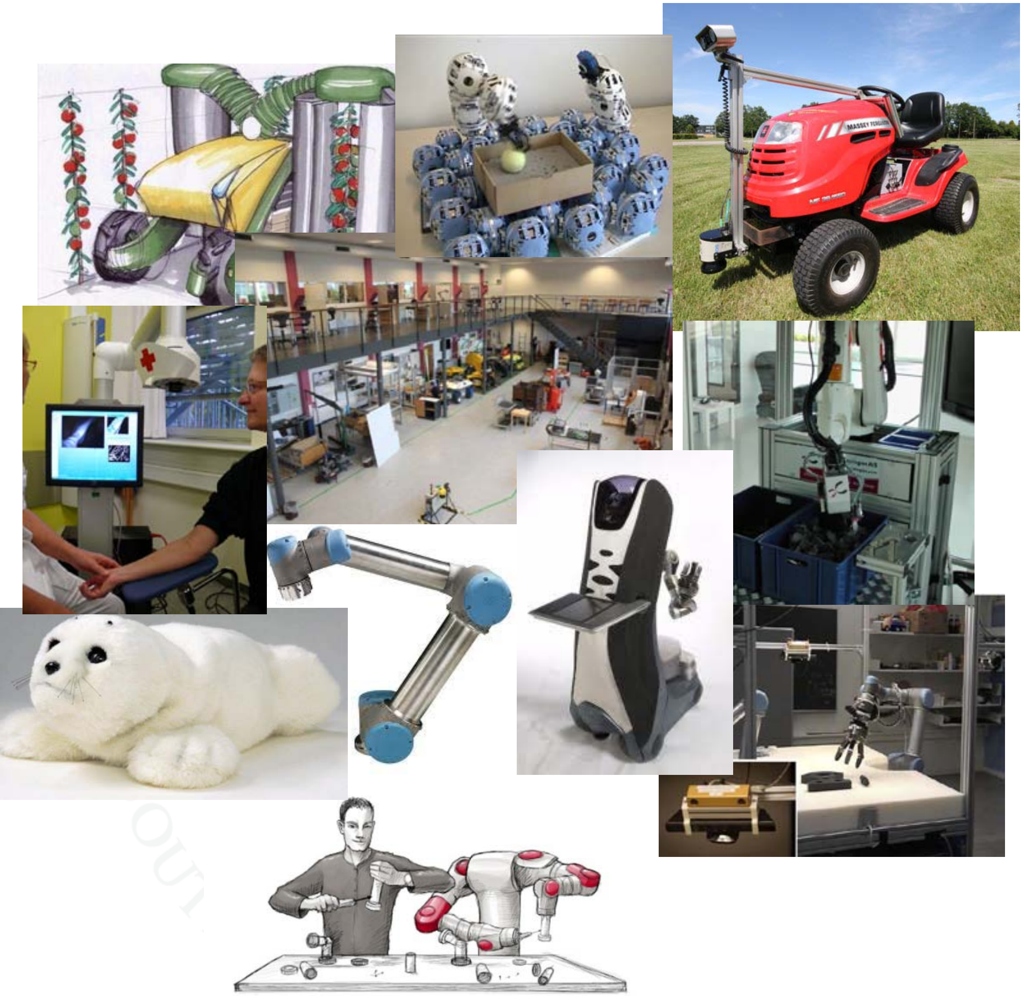


Overview

- **Background Information**
- The primate's vision system: A deep Hierarchy
- From Signals to Symbols: Birth of the Object and its affordances
- Reflections



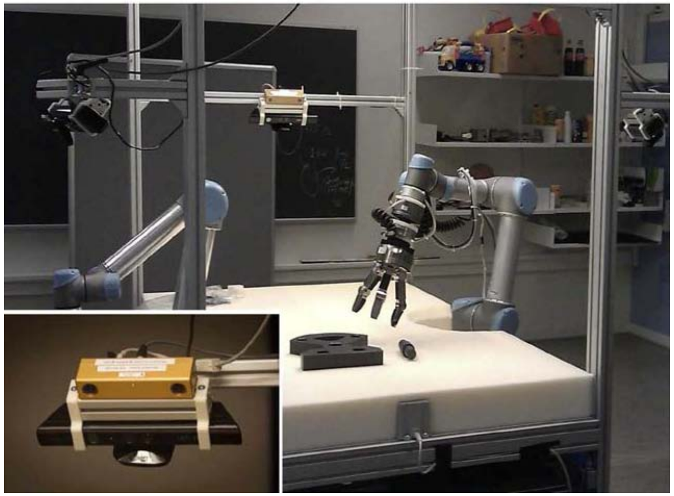
Robotics in Odense



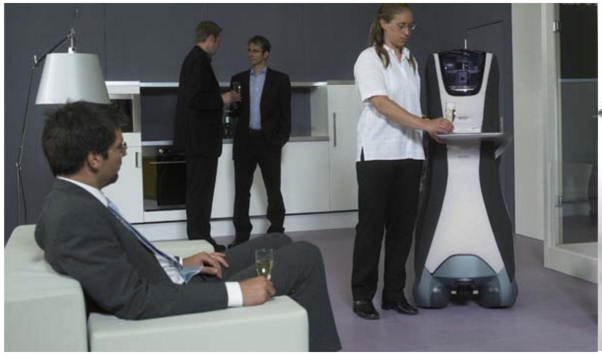
- **University of Southern Denmark (SDU) (ca. 60 members of staff in robotics)**
  - The Maersk Mc-Kinney Moller Institute
  - RoboLab
  - Robocluster
- **Danish Technological Institute (ca. 80 members of staff in robotics)**
  - Technology Transfer Institute
- **A number of robotic/vision companies**
  - Universal Robots
  - Scape
  - TriVision
  - …
- **Recent robotic events in Odense**
  - European Robotics Forum 2011 (DTI)
  - SAB 2012 (SDU)

Robot Platforms

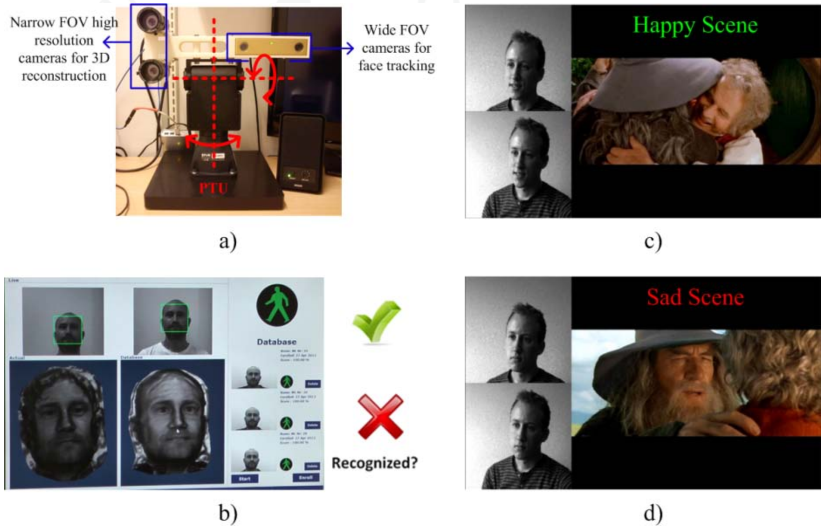
Intelligent Work Cell



Service Robotics

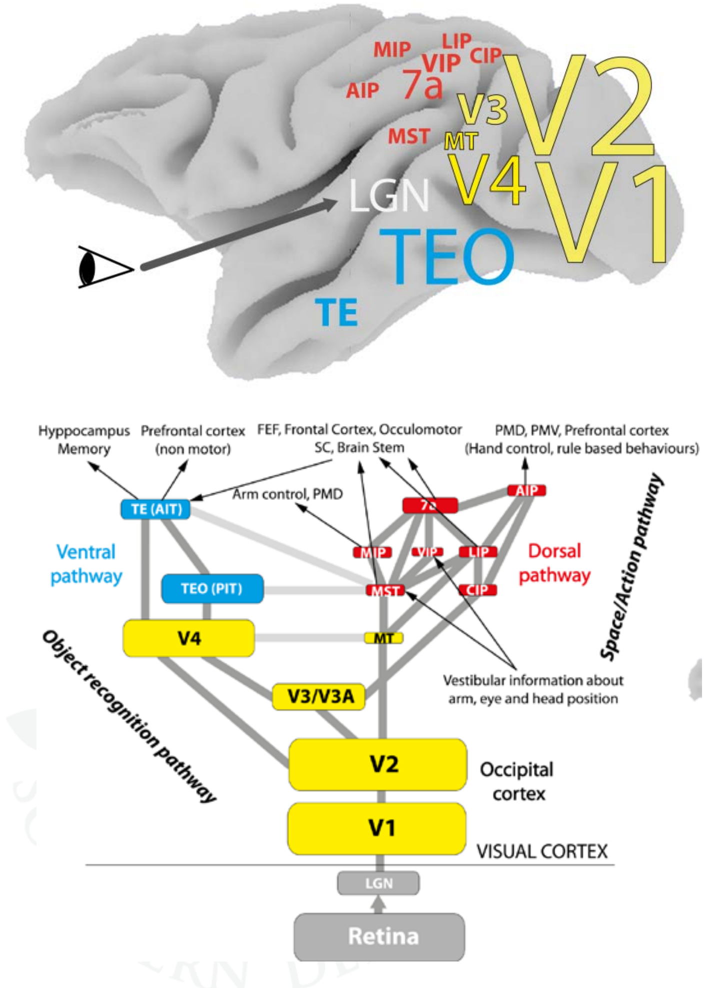


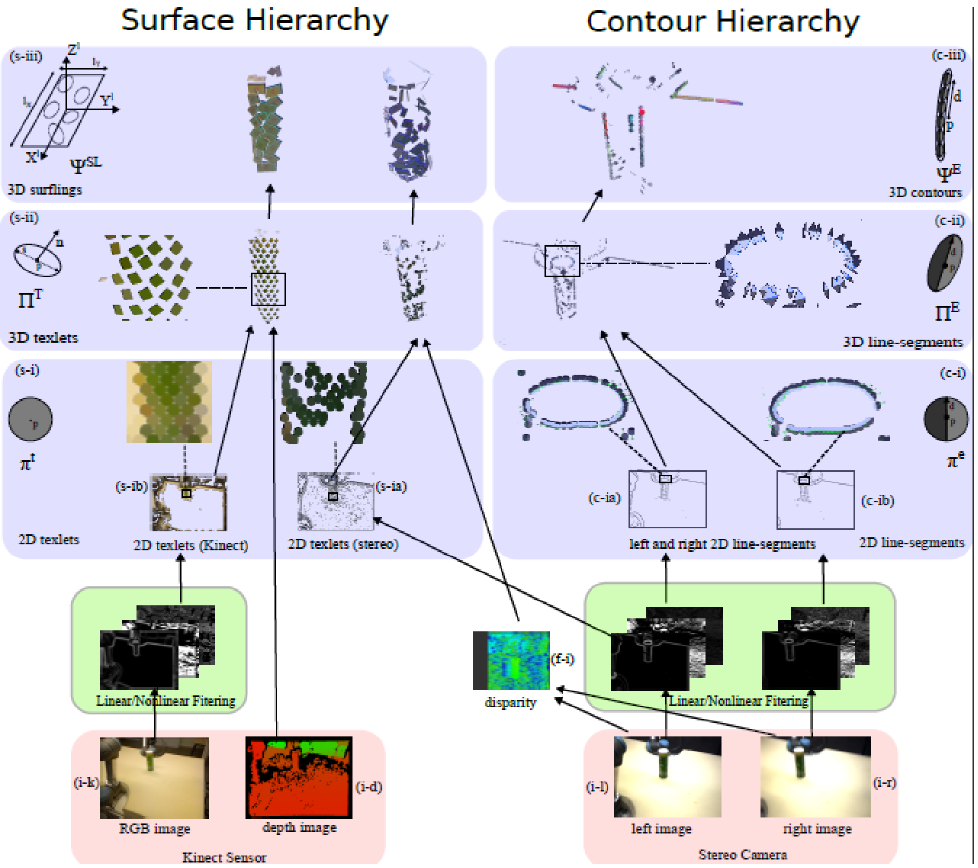
Active Vision System



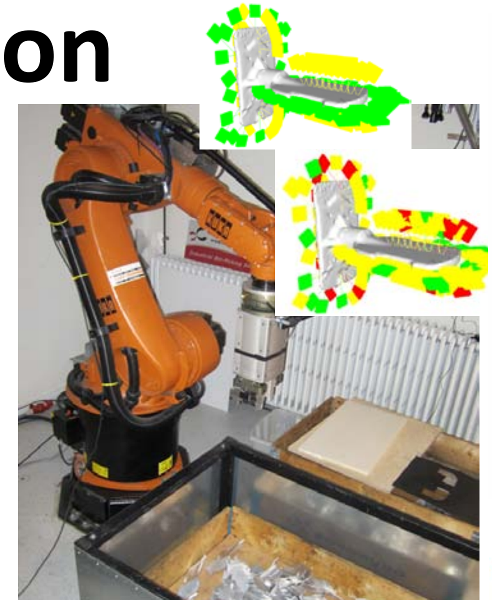
Milling Platform

Early Cognitive Vision System

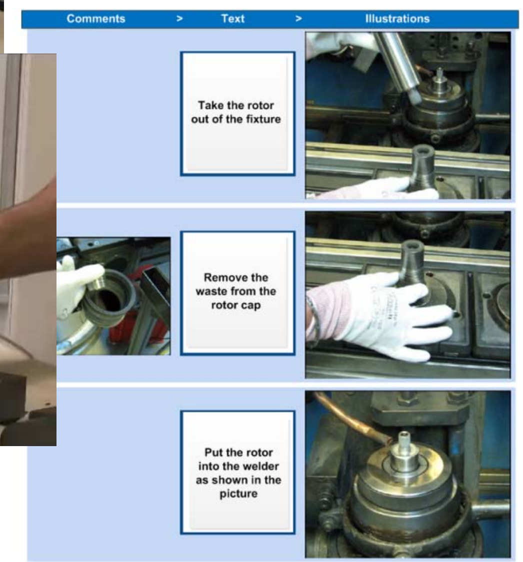




Mission: Bring Cognition into (in particular) Production



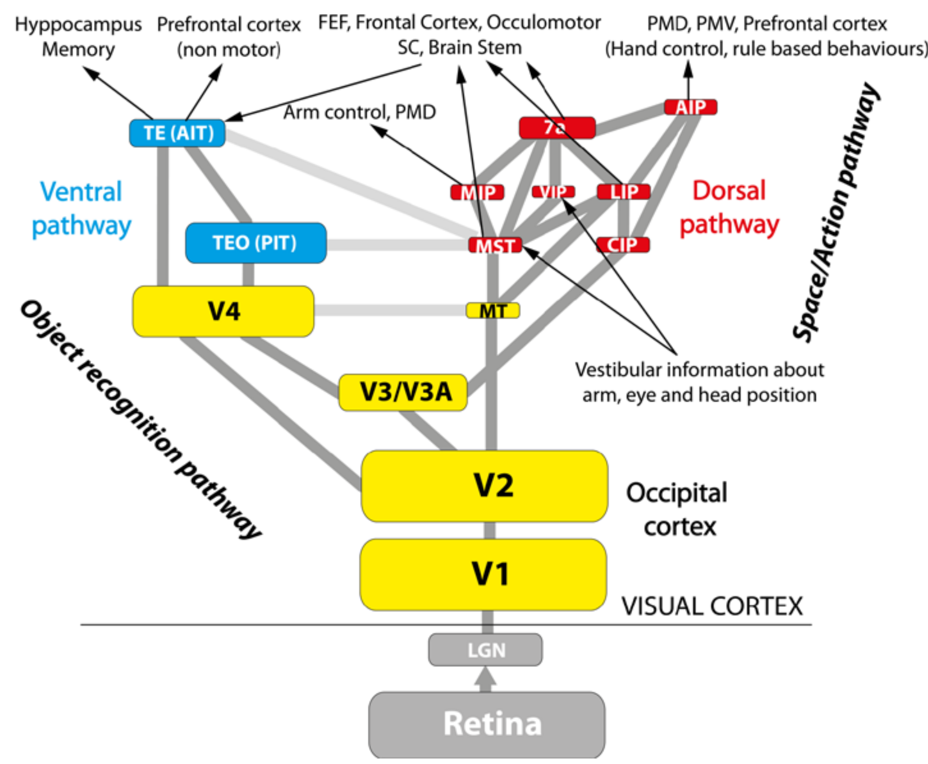
- **Main projects**
  - Xperience (2011-2016)
  - LearnBiP (2011-2012)
  - IntellAct (2011-2014)
  - ACAT (2013-2016)
  - CARMEN (2013-2017)
- **Others**
  - TailorCrete, FiberLab, patient@home



Our Communication with Robots

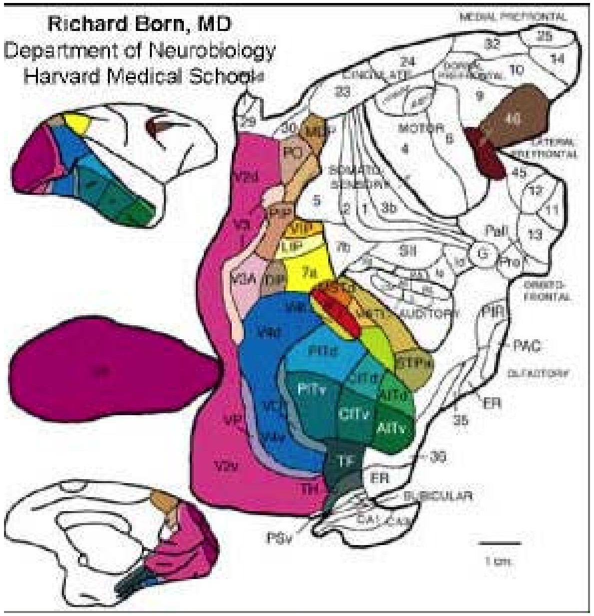
Overview

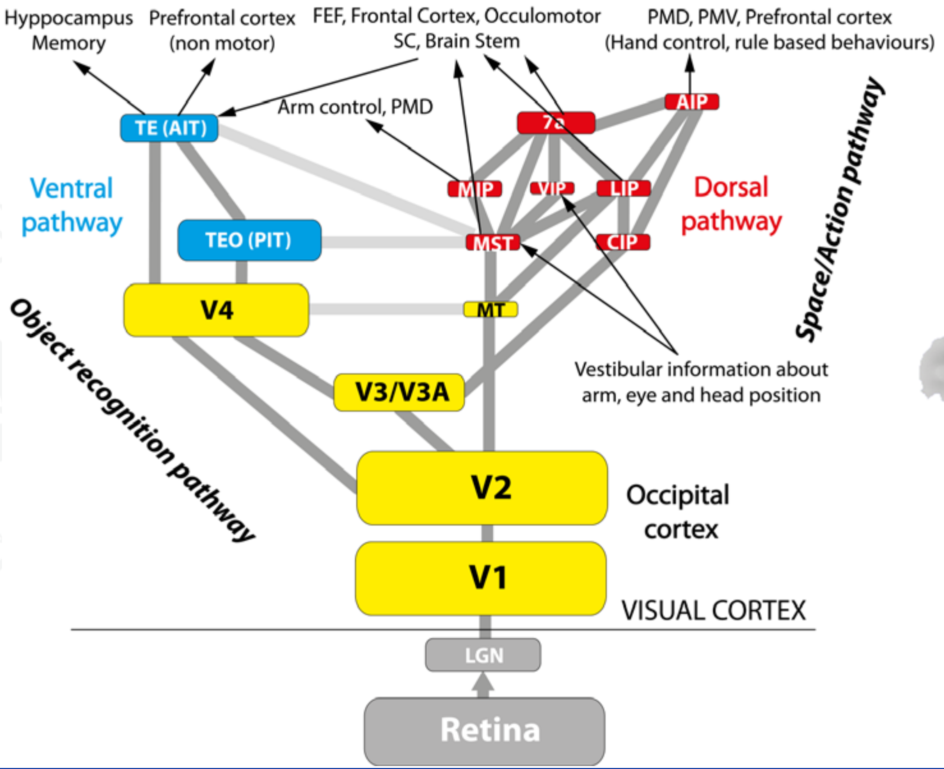
- Background Information
- **The primate's vision system: A deep Hierarchy**
  - Half of the brain in 15 minutes
  - N. Krüger , P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez and L. Wiskott (2013), Deep Hierarchies in the Primate Visual Cortex: What Can We Learn for Computer Vision?, IEEE PAMI 2013.
- From Signals to Symbols: Birth of the Object and its affordances
- Reflections

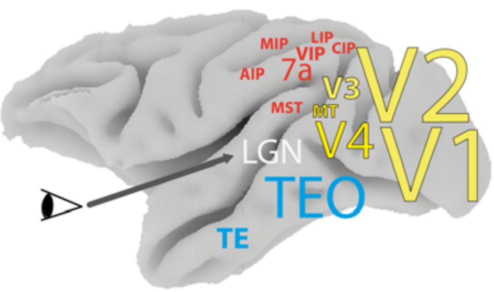


Basic facts

- **55% of the neo-cortex of the primate brain is concerned with vision**
- **Devision in**
  - Occipitel Cortex
  - Dorsal Pathway
  - Ventral Pathway







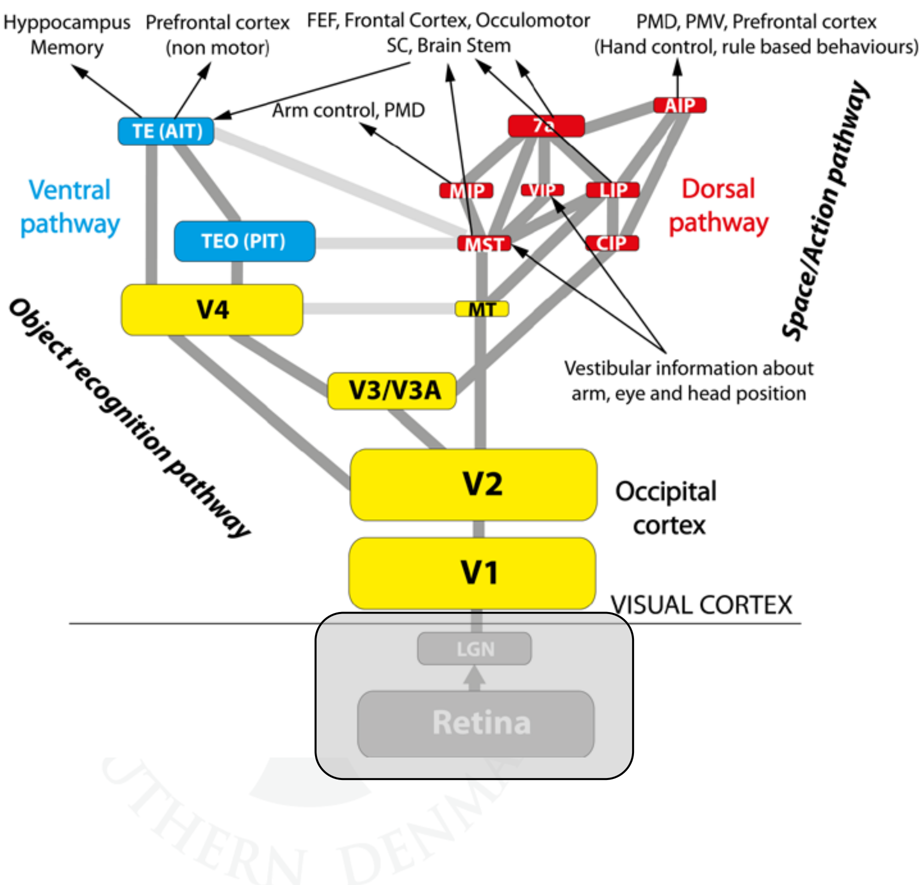
Basic Facts

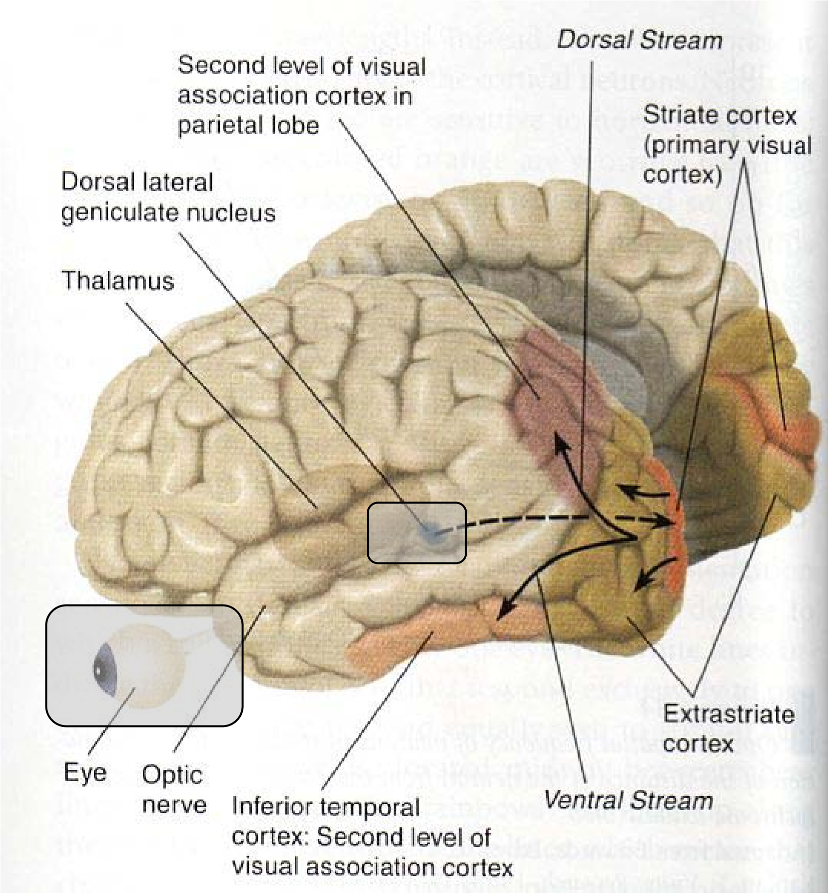
Area	Size (mm^2)	RFS	Latency (ms)	co/bi lat.	rt/st/cl/co	CI/SI/PI/OI	Function
	Sub–cortical processing						
Retina	1018	0.01	20-40	bl	+/-/-/-	-/-/-/-	sensory input, contrast computation
LGN		0.1	30-40	co	+/-/-/-	-/-/-/-	relay, gating
	Occipital / Early Vision						
V1	1120	3	30-40	co	+/-/-/+	-/-/-/-	generic feature processing
V2	1190	4	40	co	+/-/-/+	-/-/-/-	generic feature processing
V3/V3A/VP	325	6	50	co	+/-/-/+	-/-/-/-	generic feature processing
V4/VOT/V4t	650	8	70	co	+/-/-/+	+/-/-/-	generic feature processing / color
MT	55	7	50	co	+/-/-/+	+/+/-/+	motion
Sum	3340						
	Ventral Pathway / What (Object Recognition and Categorization)						
TEO	590	3-5	70	co	(+)/-/-/+	?/-/-/?	object recognition and
TE	180	10-20	80-90	bl	-/-/+/+	+/+/+/+(-)	categorization
Sum	770						
	Dorsal Pathway / Where and How (Coding of Action Relevant Information)						
MST	60	>30	60-70	bl	+/-/+/-	I	optic flow, self-motion, pursuit
CIP	?	?	?		+/-/?/?	+/?/?/?	3D orientation of surfaces
VIP	40	10-30	50-60	bl	-/+/-/-	I	optic flow, touch, near extra personal space
7a	115	>30	90	bl	(+)/-/-/-	?/?/+/?	Optic flow, heading
LIP	55	12-20	50	cl	+/-/-/-	?/-/-/-	salience, saccadic eye movements
AIP	35	5-7	60	bl	?/+/+/?	?/+/+/?	grasping
MIP	55	10-20	100	co	+/-/?/?	I	reaching
Sum	585						

TABLE 1

Basic facts on the different areas of the macaque visual cortex based on different sources [44], [28], [95], [141], [161] *First column:* Name of Area. *Second column:* Size of area in mm^2. '?' indicates that this information is not available. *Third column:* Average receptive field size in degrees at 5 degree of eccentricity. *Fourth column:* Latency in milliseconds. *Fifth Column:* Contra versus bilateral receptive fields. *Sixth Column:* Principles of organization: Retinotopic (rt), spatiotopic (st), clustered (cl) columnar (co) *Seventh Column:* Invariances in representation of shape: Cue-Invariance (CI), Size Invariance (SI), Position Invariance (PI), Occlusion Invariance (OI). 'I' indicates that this entry is irrelevant for the information coded in these areas. *Eighth Column:* Function associated to a particular area.

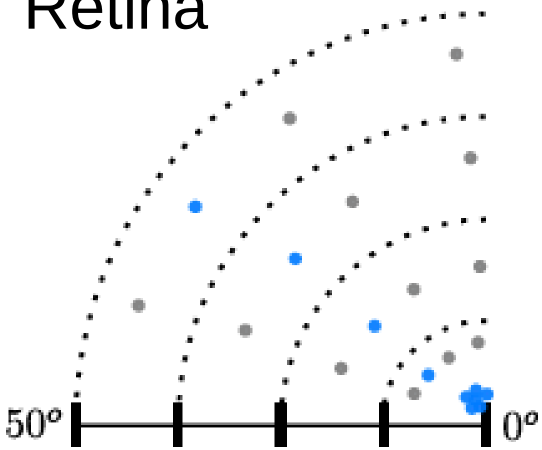
Pre-cortical Areas



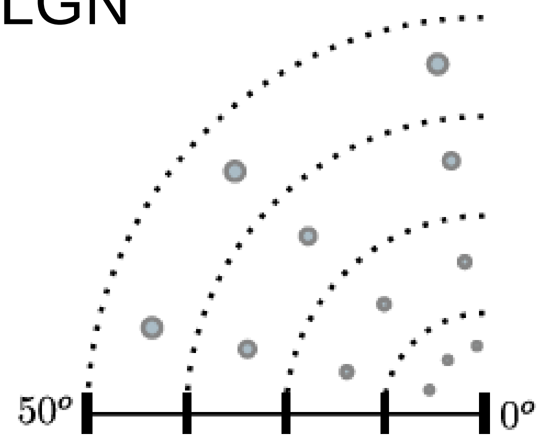


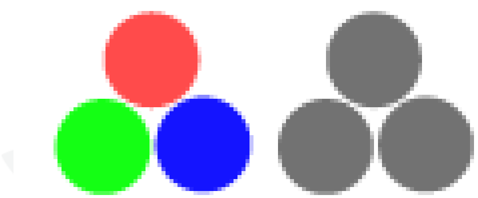
Precortical Areas

Retina

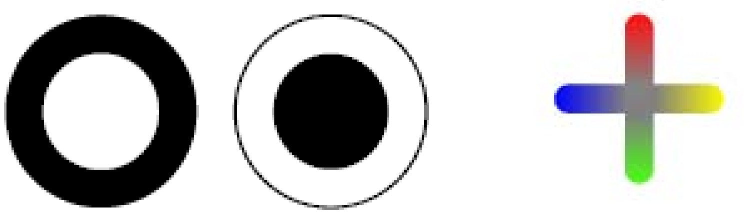


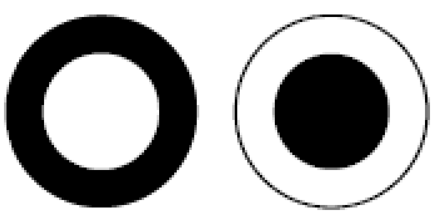
LGN



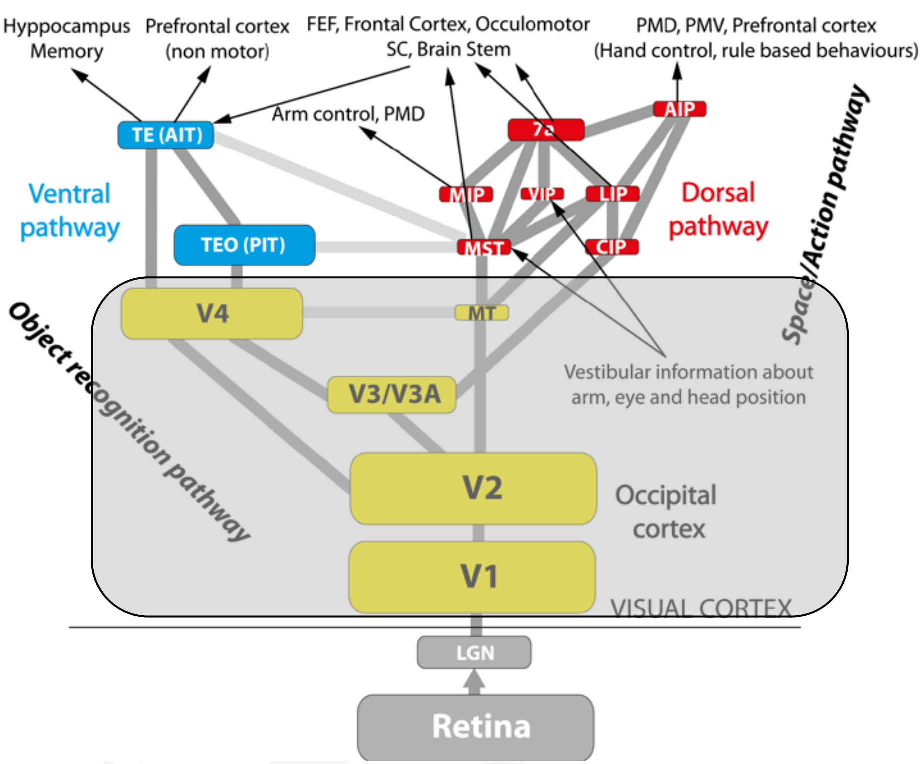


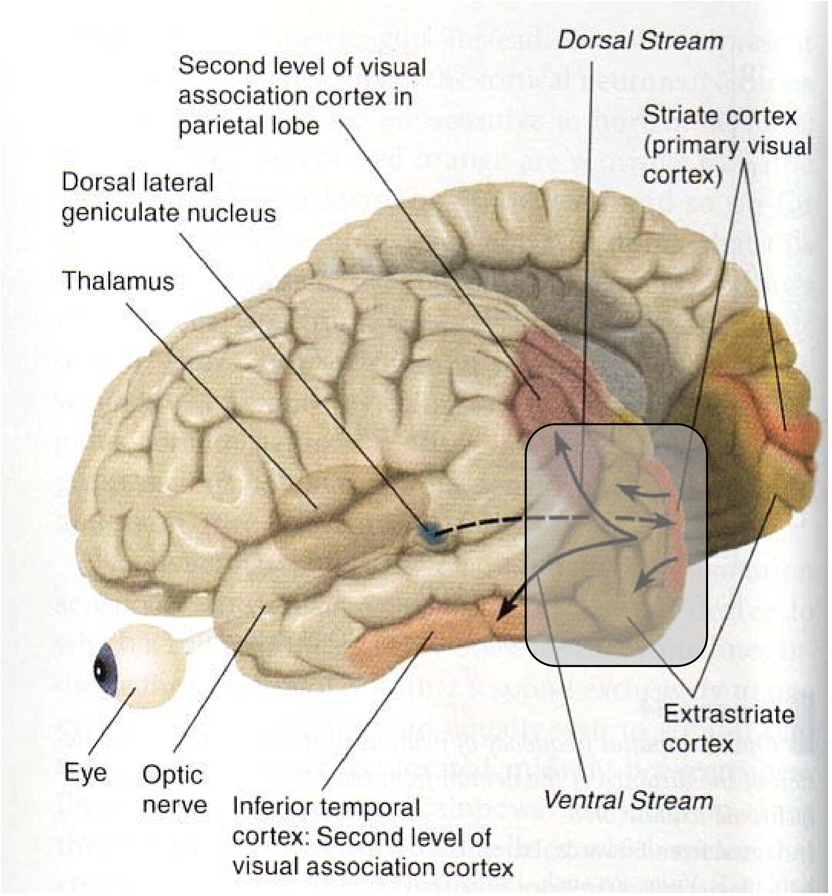
- No Feature Transformation
- Preparing for Stereo





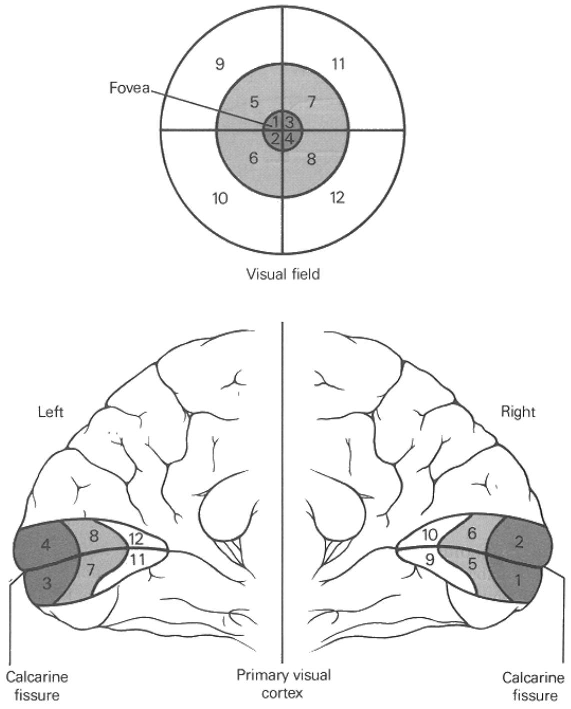
Occipital Cortex

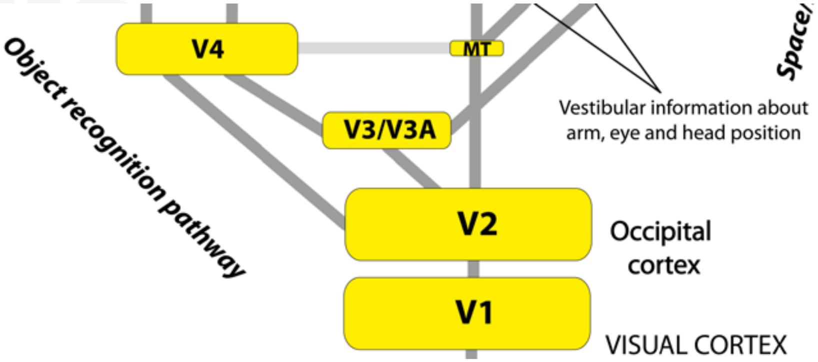




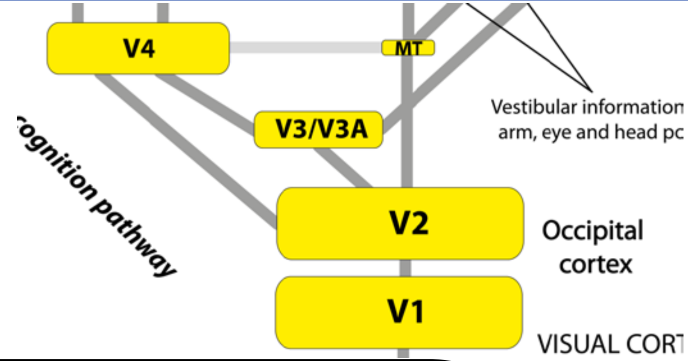
Occipital Cortex

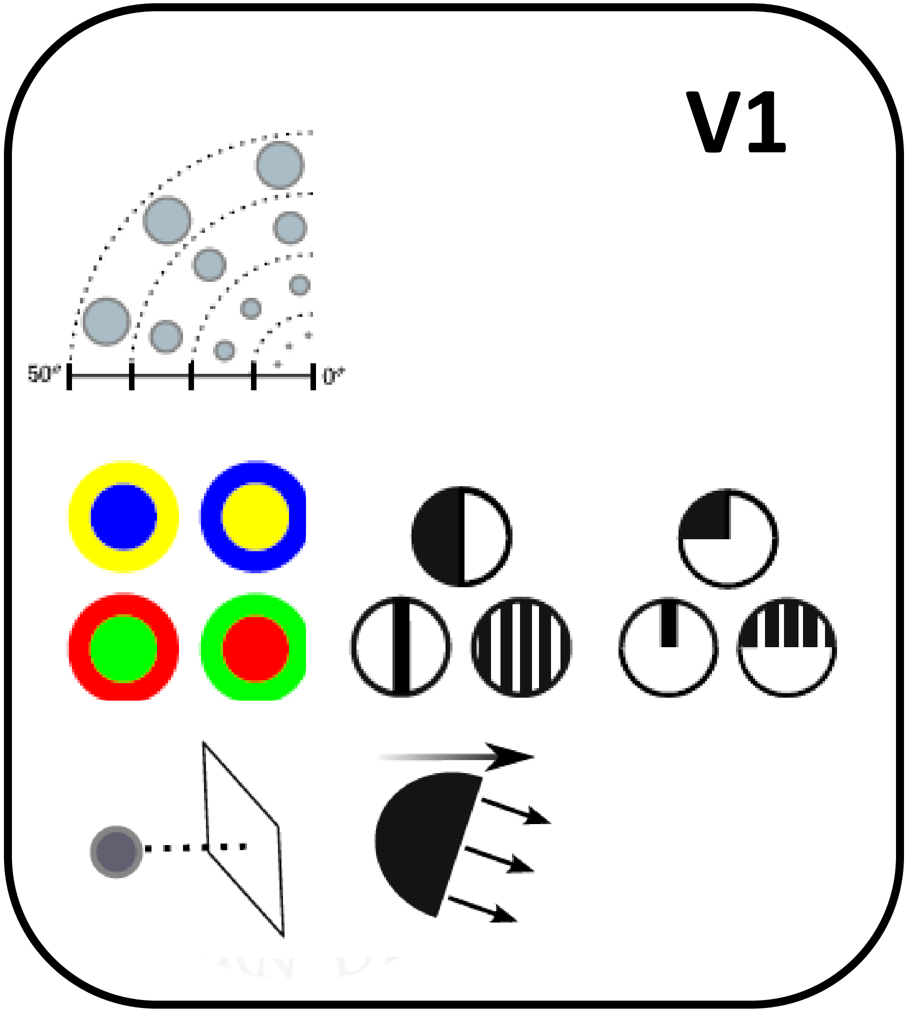
- **More than 70% of the visual cortex**
  - Occipital Cortex 3340mm^2
  - Ventral Pathway 770mm^2
  - Dorsal Pathway 585mm^2
- **Processing**
  - Task unspecific generic scene representation

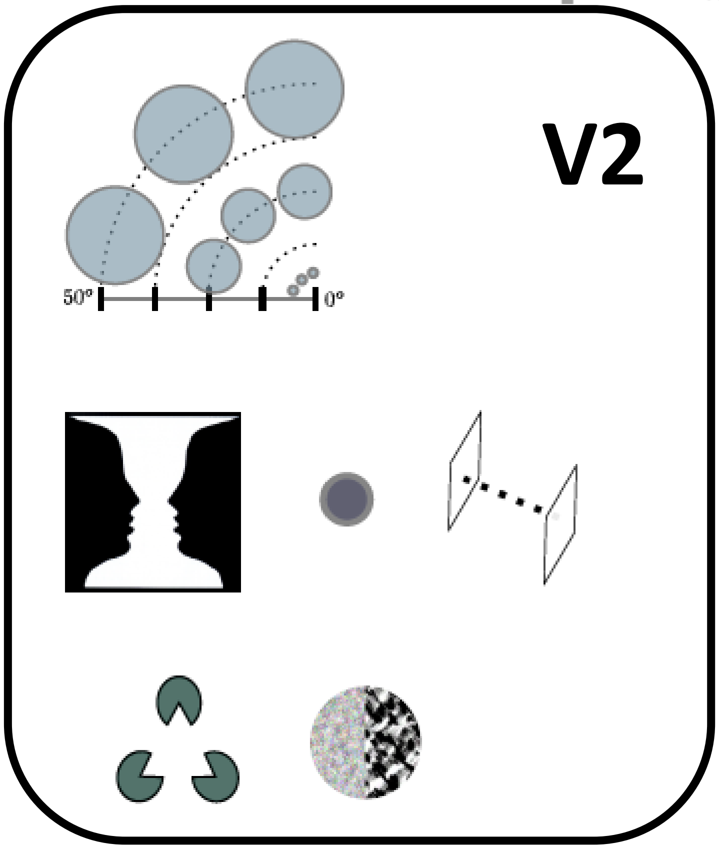




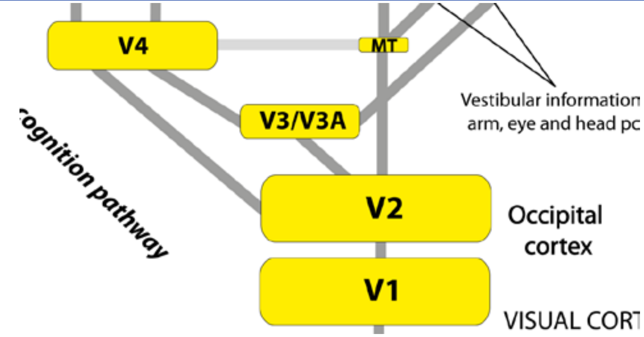
Occipital Cortex: V1 and V2



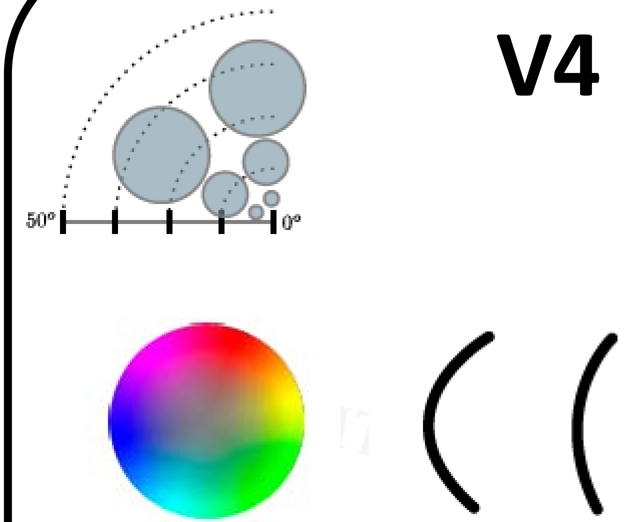




V4 and MT

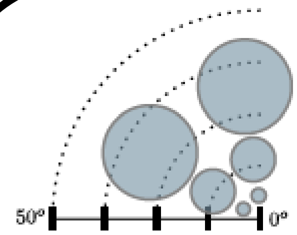


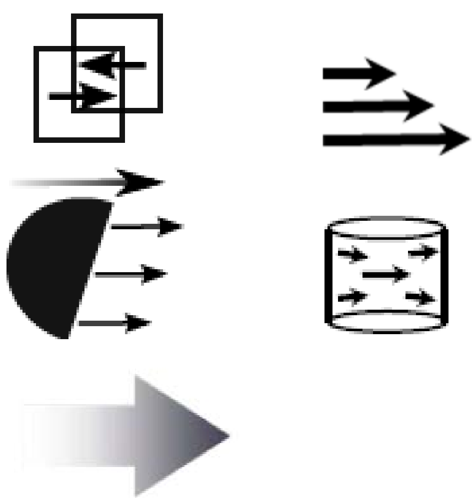
V4



Concept of Hue as Object Property
Linguistic Concept of ‘red’ or ‘blue’

MT

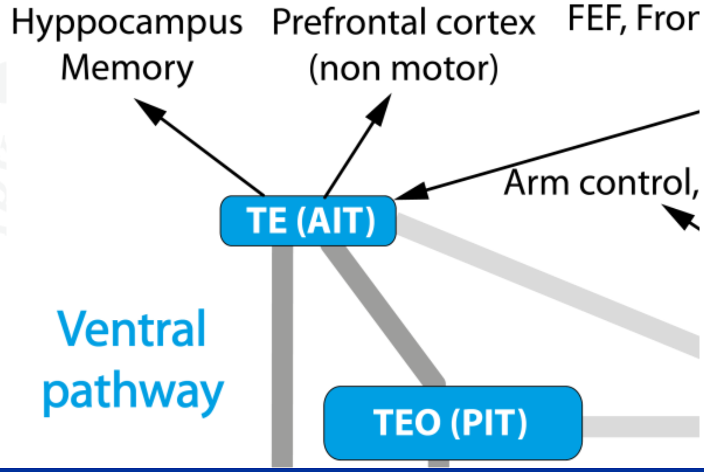




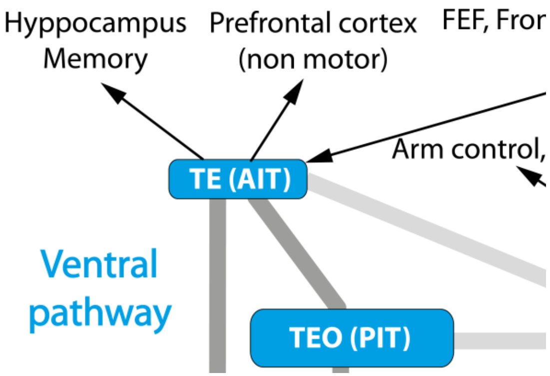
2D Motion **3D Motion**

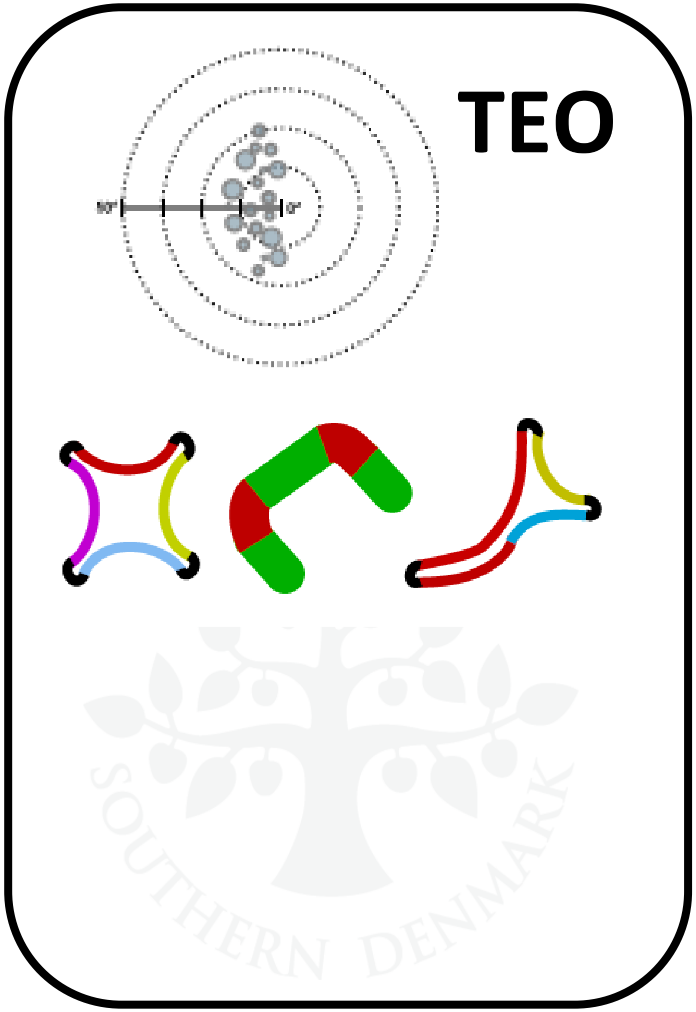
Ventral Pathway

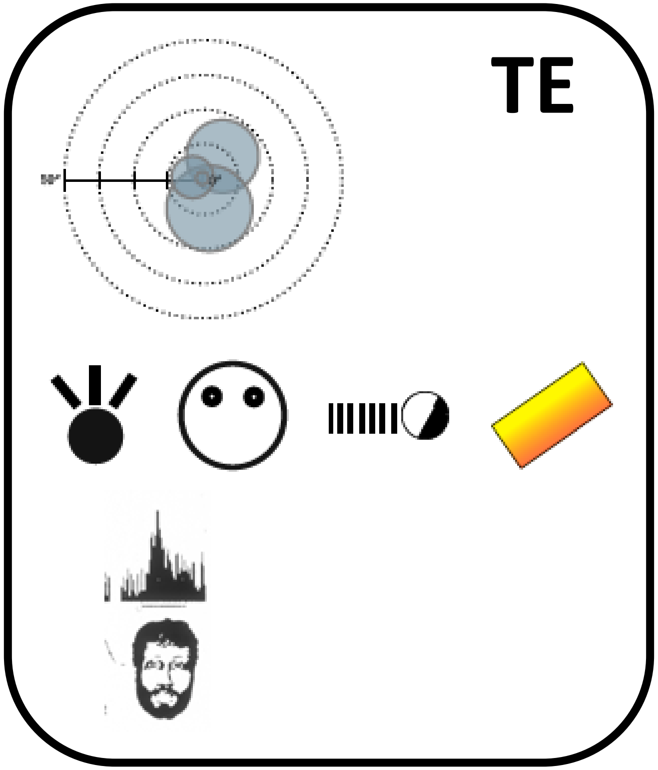
- **More than 70% of the visual cortex**
  - Occipital Cortex 3340mm^2
  - Ventral Pathway 770mm^2
  - Dorsal Pathway 585mm^2
- **Processing**
  - Object Recognition and Categorization
  - Many suggestions for how to divide into areas

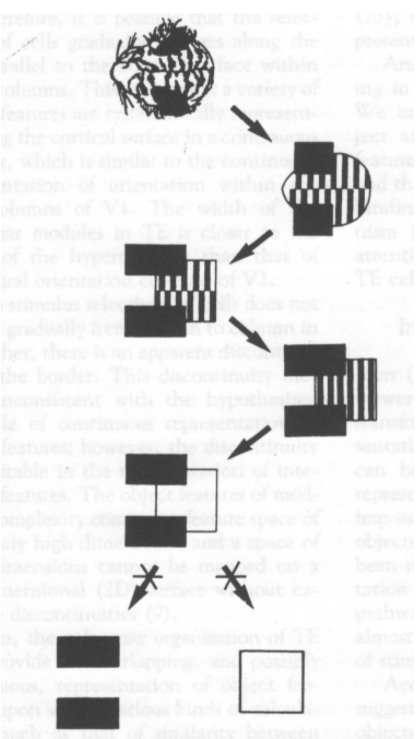


Ventral Pathway: TEO and TE





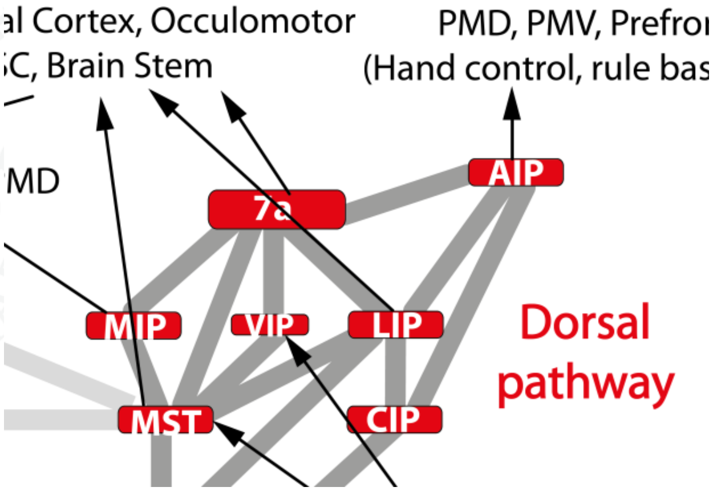




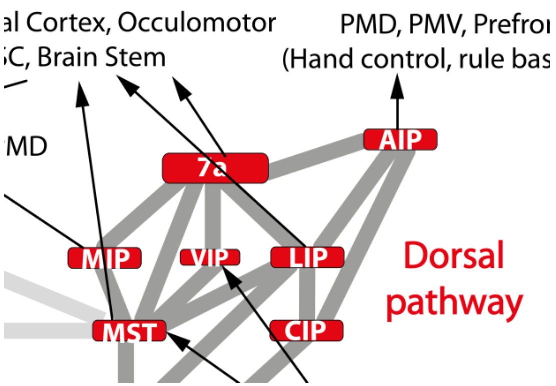
Tanaka

Dorsal Pathway

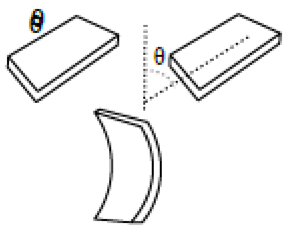
- **More than 70% of the visual cortex**
  - Occipital Cortex 3340mm^2
  - Ventral Pathway 770mm^2
  - Dorsal Pathway 585mm^2
- **Processing**
  - Much less known than Ventral Pathway
  - Many more distinguished areas
  - Coding visual information related to action and position in space



Dorsal Pathway

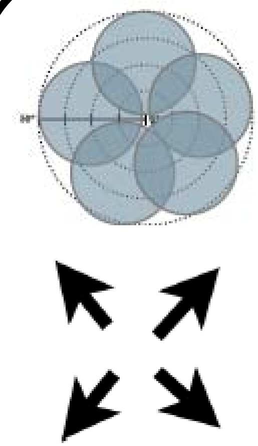


CIP



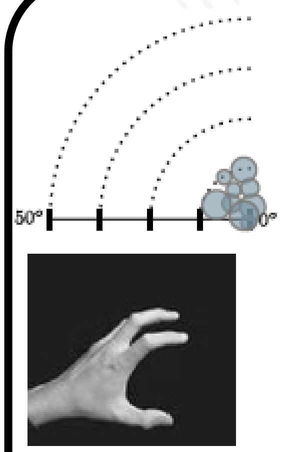
Cue invariant 3D shape

MST



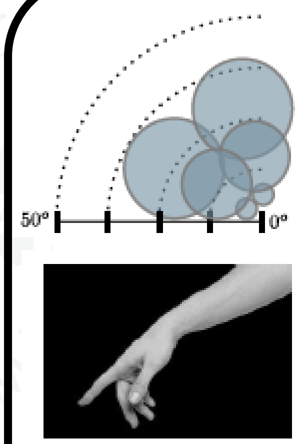
Ego-motion

AIP



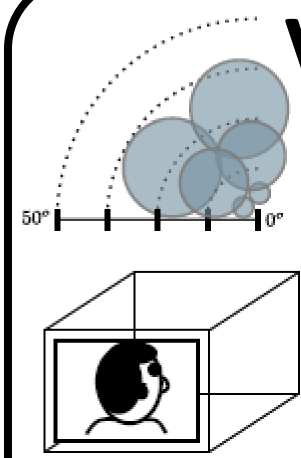
Hand shape and affordances

MIP



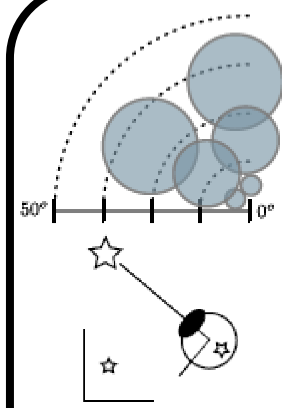
Reaching

VIP



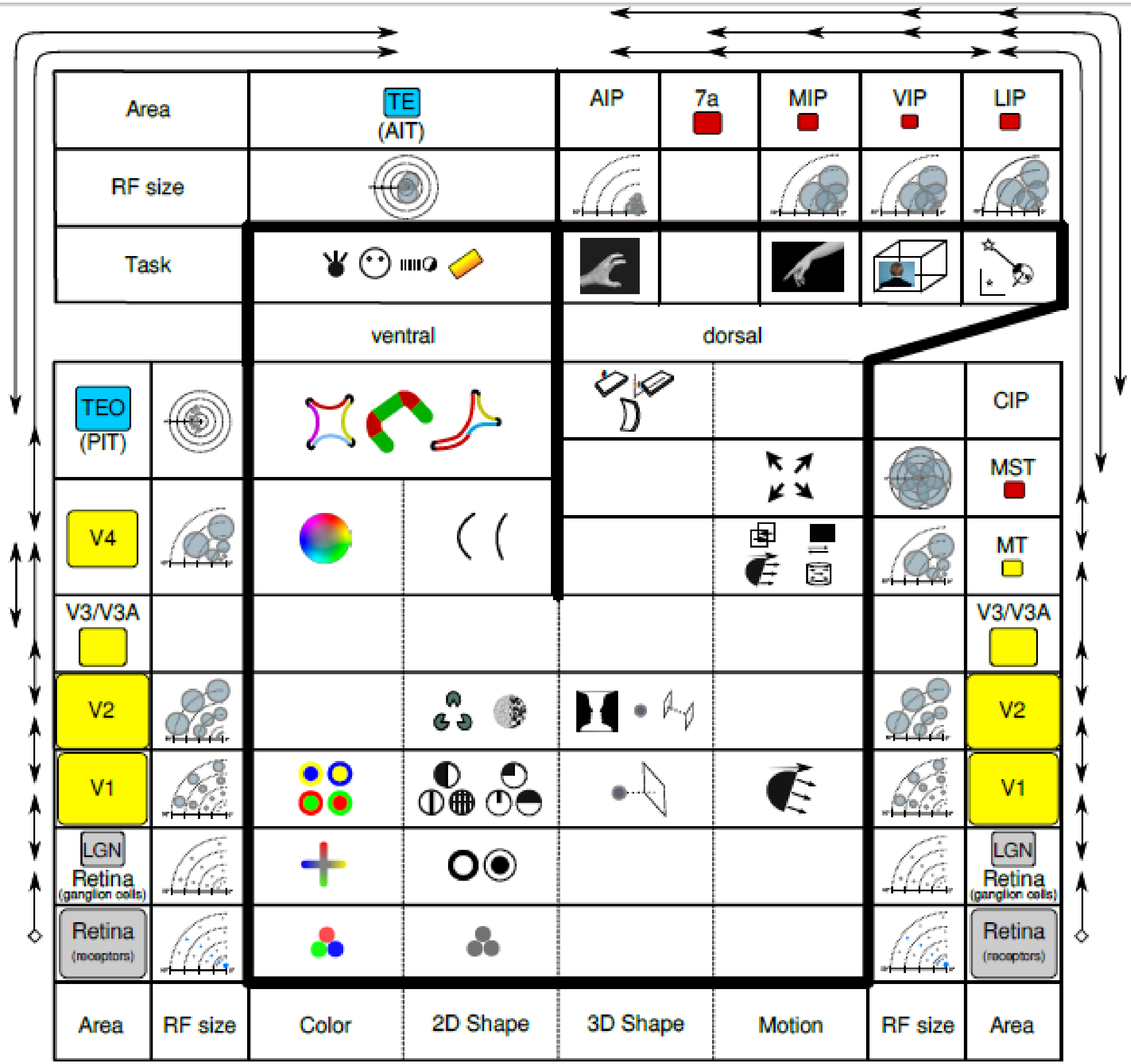
Ego-space

LIP



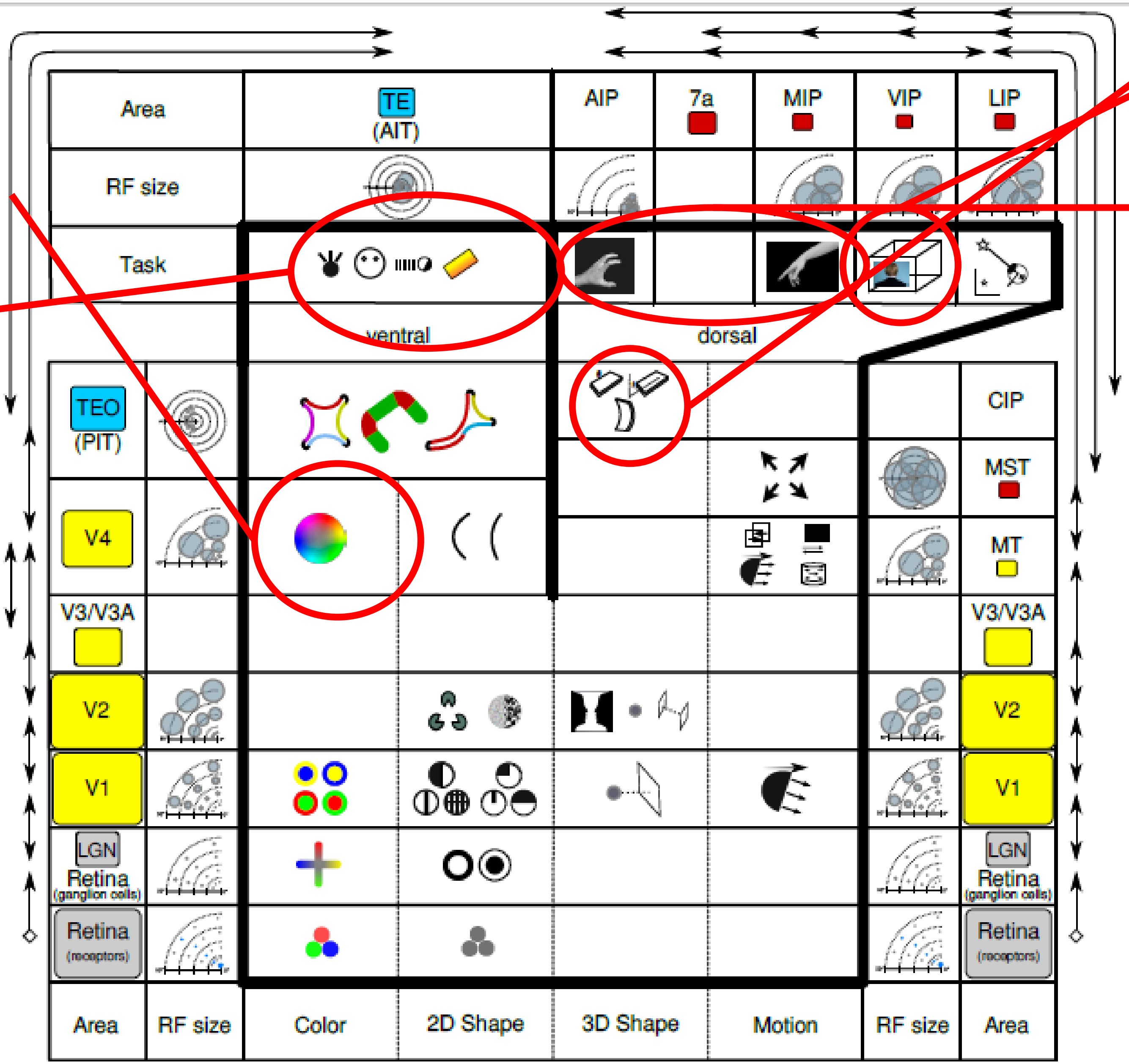
Saccadic related retinotopic repr.

Vertical View



Where does language come in?

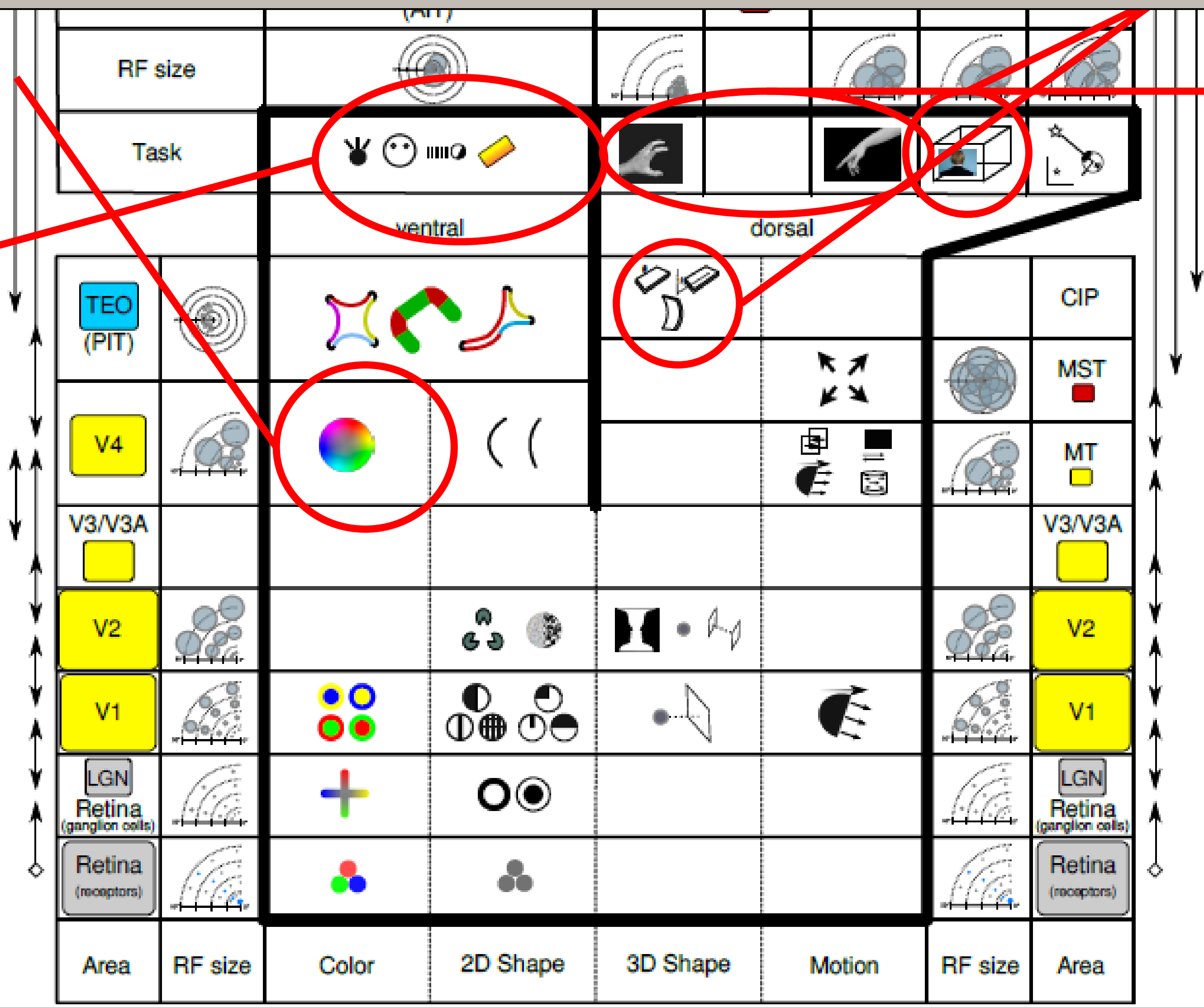
- Colour as object property
- Objects



- Preposit ons
- Actions/ verbs

A lot of visual information relevant for linguistic is in the temporal structure: e.g., SEC level

- **Colour as object property**
- **Objects**

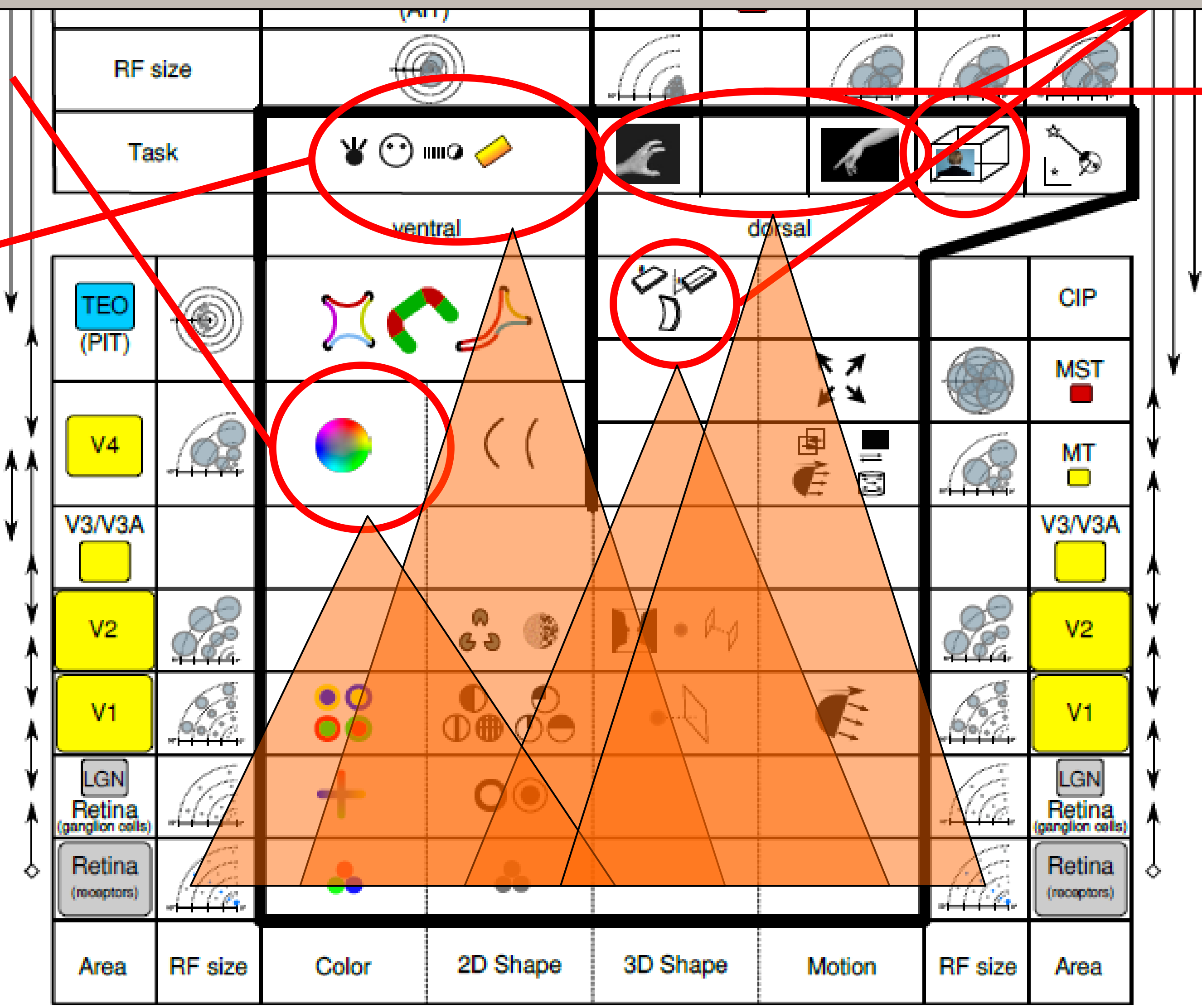


Prepositions

- **Actions/ verbs**

A lot of visual information relevant for linguistic is in the temporal structure: e.g., SEC level

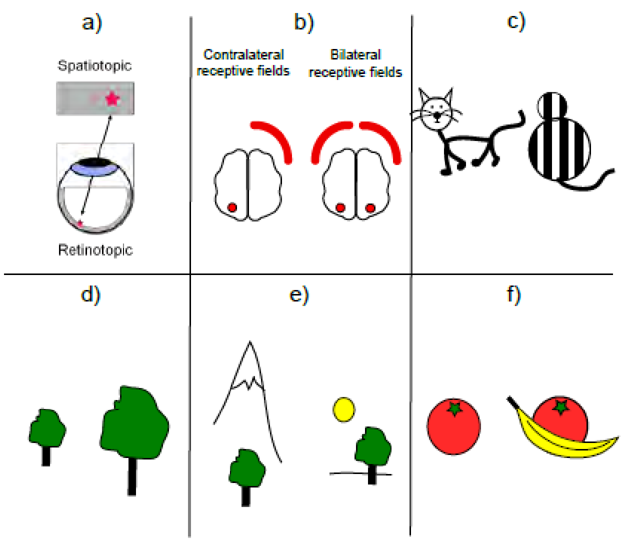
- Colour as object property
- Objects



Prepositions

- Actions/verbs

Basic Terms



- **Retinotopic/Spatiotopic**
- **Different kinds Of Invariances**
  - Cue Invariance
  - Size Invariance
  - Position Invariance
  - Occlusion Invariance

Area	co/bi lat.	rf/st/cl/co	CI/SI/PI/OI
	Sub–cortical processing		
Retina	bl	+/-/-/-	-/-/-/-
LGN	co	+/-/-/-	-/-/-/-
	Occipital / Early Vision		
V1	co	+/-/-/+	-/-/-/-
V2	co	+/-/-/+	-/-/-/-
V3/V3A/VP	co	+/-/-/+	-/-/-/-
V4/VOT/V4t	co	+/-/-/+	+/-/-/-
MT	co	+/-/-/+	+/+/-/+
Sum			
	athway / What (Object Recognition and		
TEO	co	(+)/-/-/+	?/-/-/?
TE	bl	-/-/+/+	+/+/+/+(-)
Sum			
	/ Where and How (Coding of Action R		
MST	bl	+/-/+/-	I
CIP		+/-/?/?	+/?/?/?
VIP	bl	-/+/-/-	I
7a	bl	(+)/-/-/-	?/?/+/?
LIP	cl	+/-/-/-	?/-/-/-
AIP	bl	?/+/+/?	?/+/+/?
MIP	co	+/-/?/?	I
Sum			

What do we know about primate's vision which is relevant for engineers and linguists?

- **Richness of representation**
- Deep Hierarchy versus flat Architectures
- Separation of information

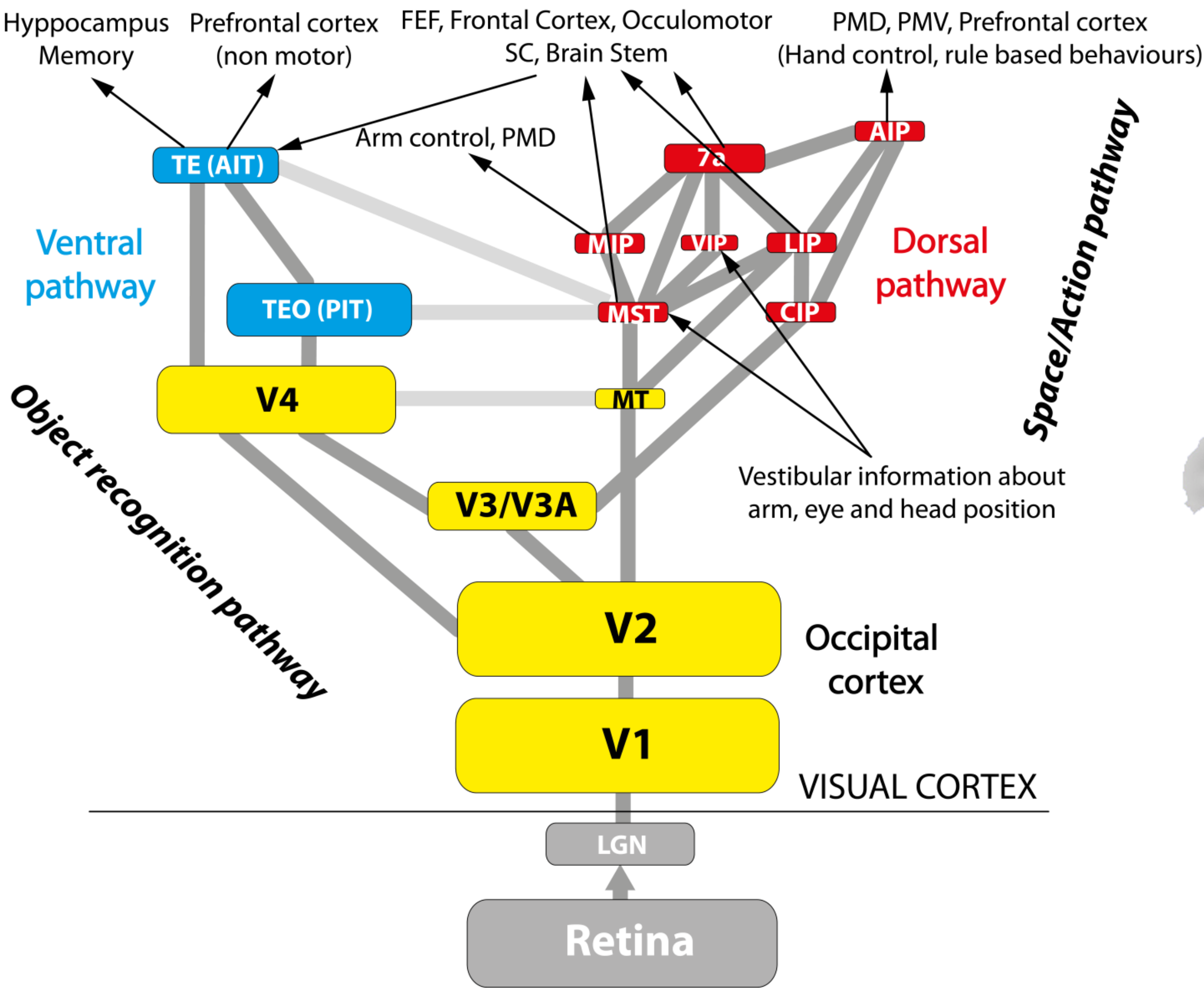
Richness of representation

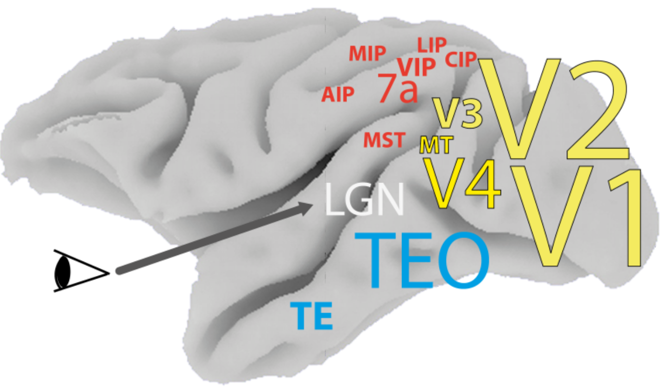
- **The occipital cortex provides a huge variety of visual aspects at different levels of granularity and different levels of abstractions**
  - Challenge: Designing/learning this hierarchy is difficult but maybe required
- **What is important for learning a certain task or category is unclear**
  - Challenge: Learning algorithms that are able to deal with such a huge and at the same time highly structured input space
    - Today there is done a lot of hardwiring of categories/planning operators
    - Relevant feature spaces are pre-selected or/and designed and probably much too simple
    - It is difficult for learning algorithms to utilize structure (e.g., SVM can not do that in a good way)

What do we know about primate's vision which is relevant for engineers and linguists?

- Richness of representation
- **Deep Hierarchy versus flat Architectures**
- Separation of information

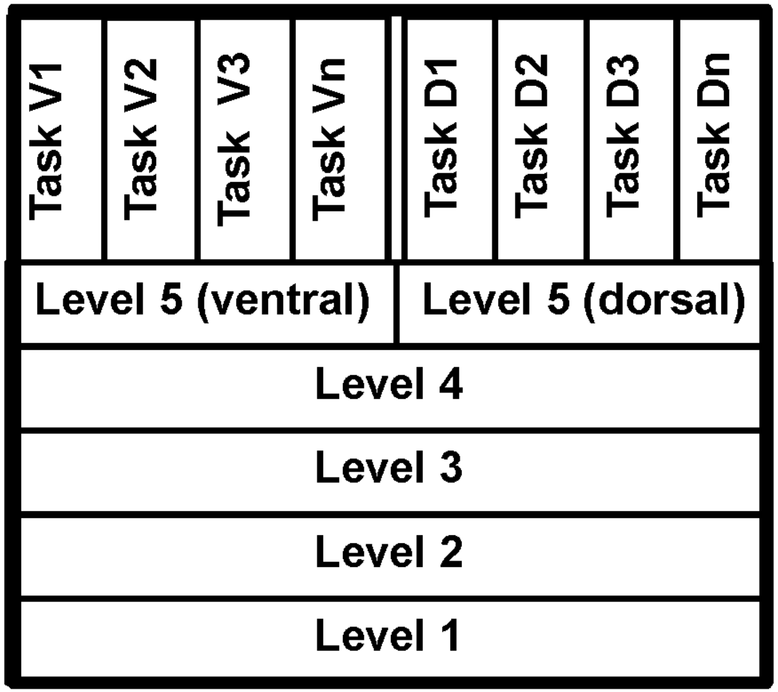
Deep Hierarchary



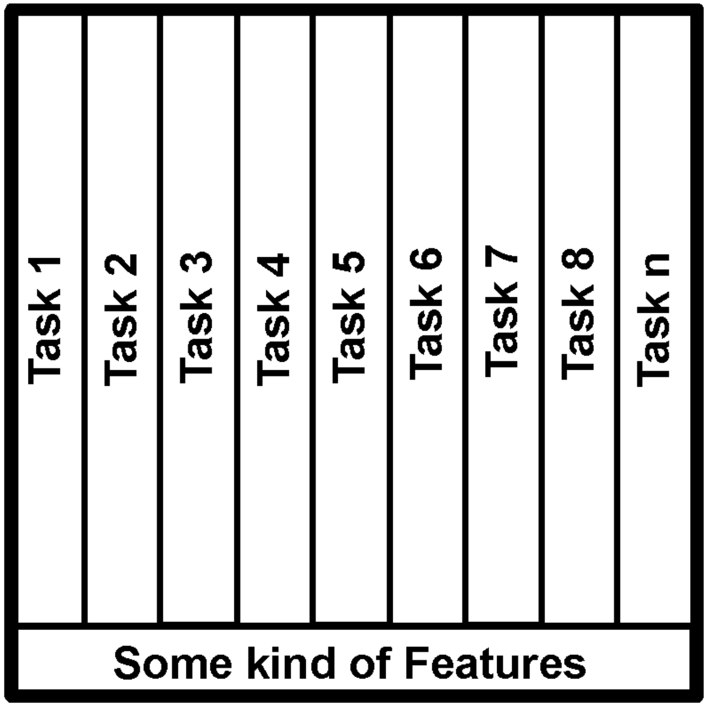


Flat versus deep Hierarchies

Deep Hierarchy



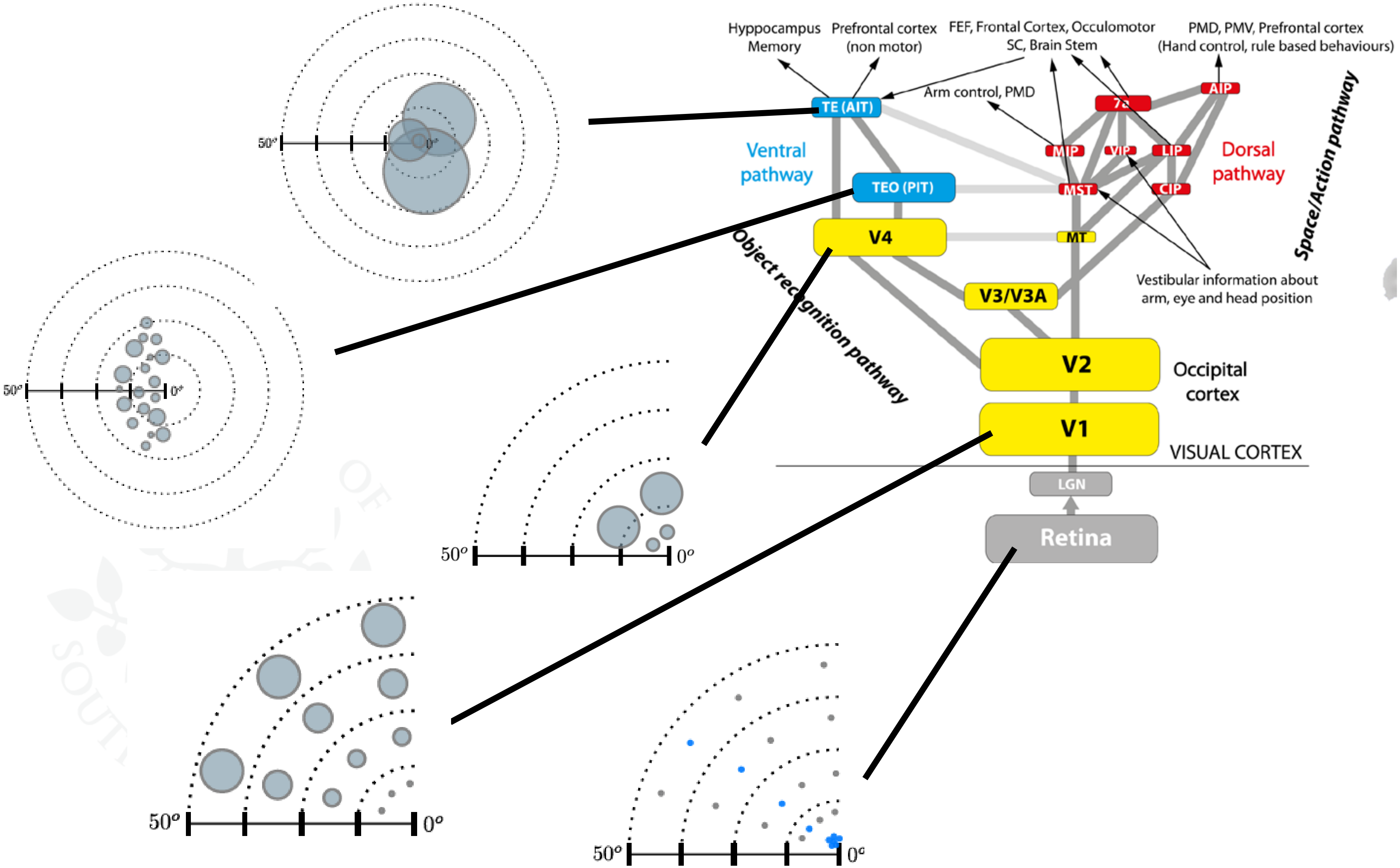
Flat Hierarchy



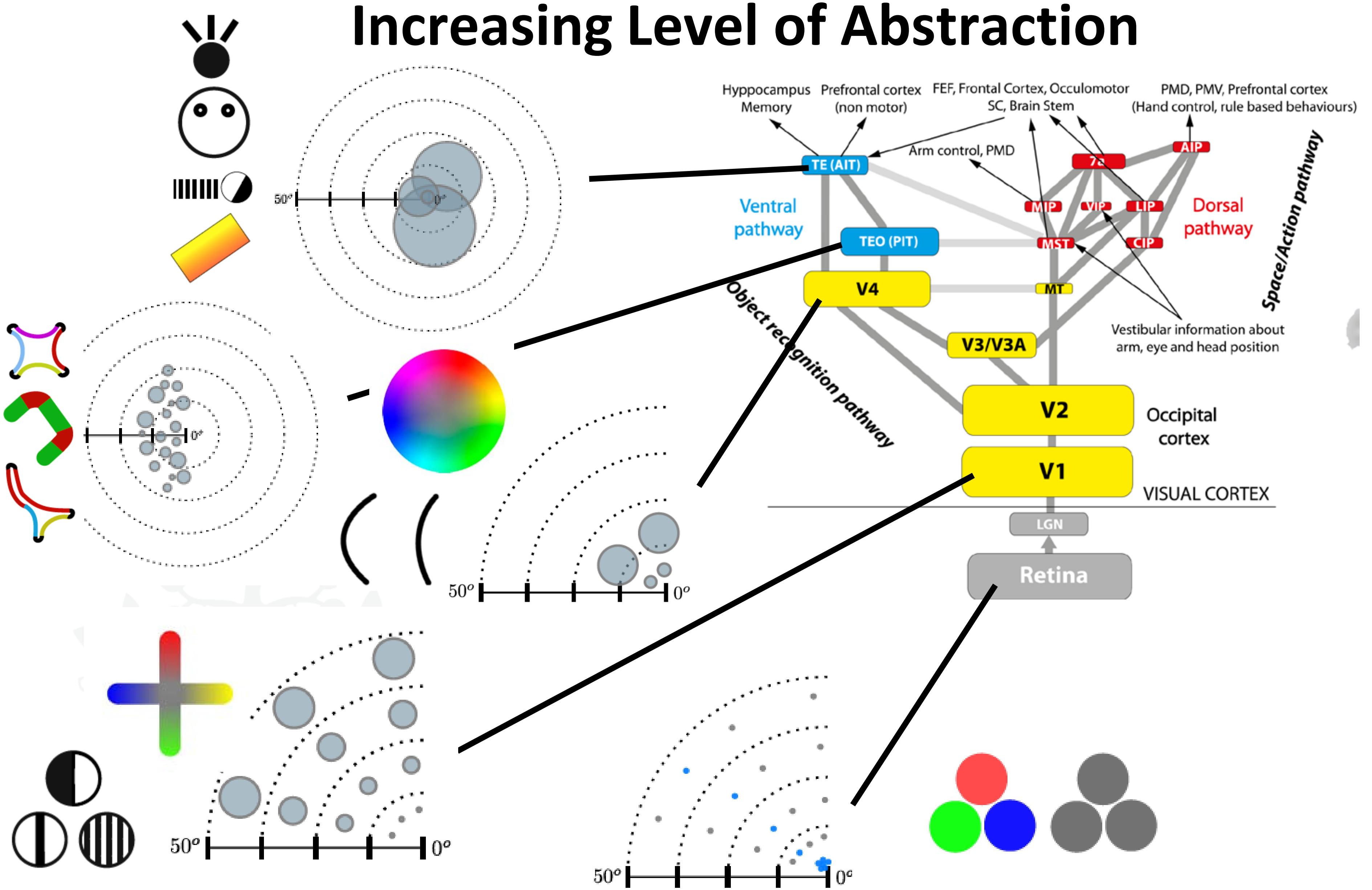
Example of a flat hierarchy

J. Y. Lettvin et al. (1959). What the frog's eye tells the frog's brain. Proceedings of the Institute of Radio Engineers

Increasing Level of Abstraction

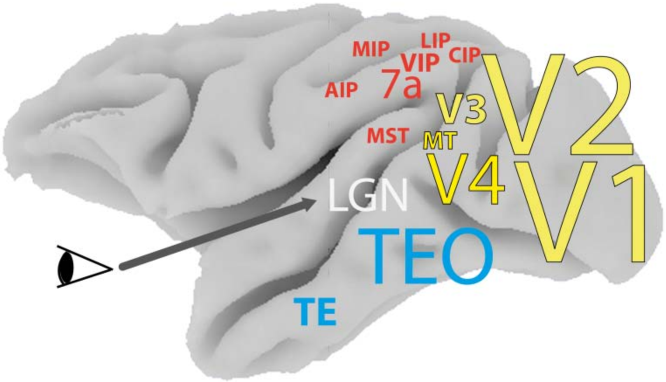


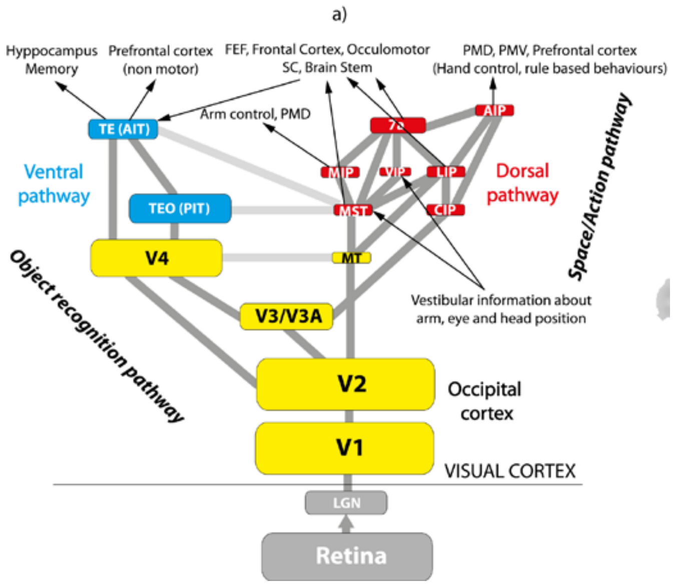
Increasing Level of Abstraction



Flat versus deep hierarchies

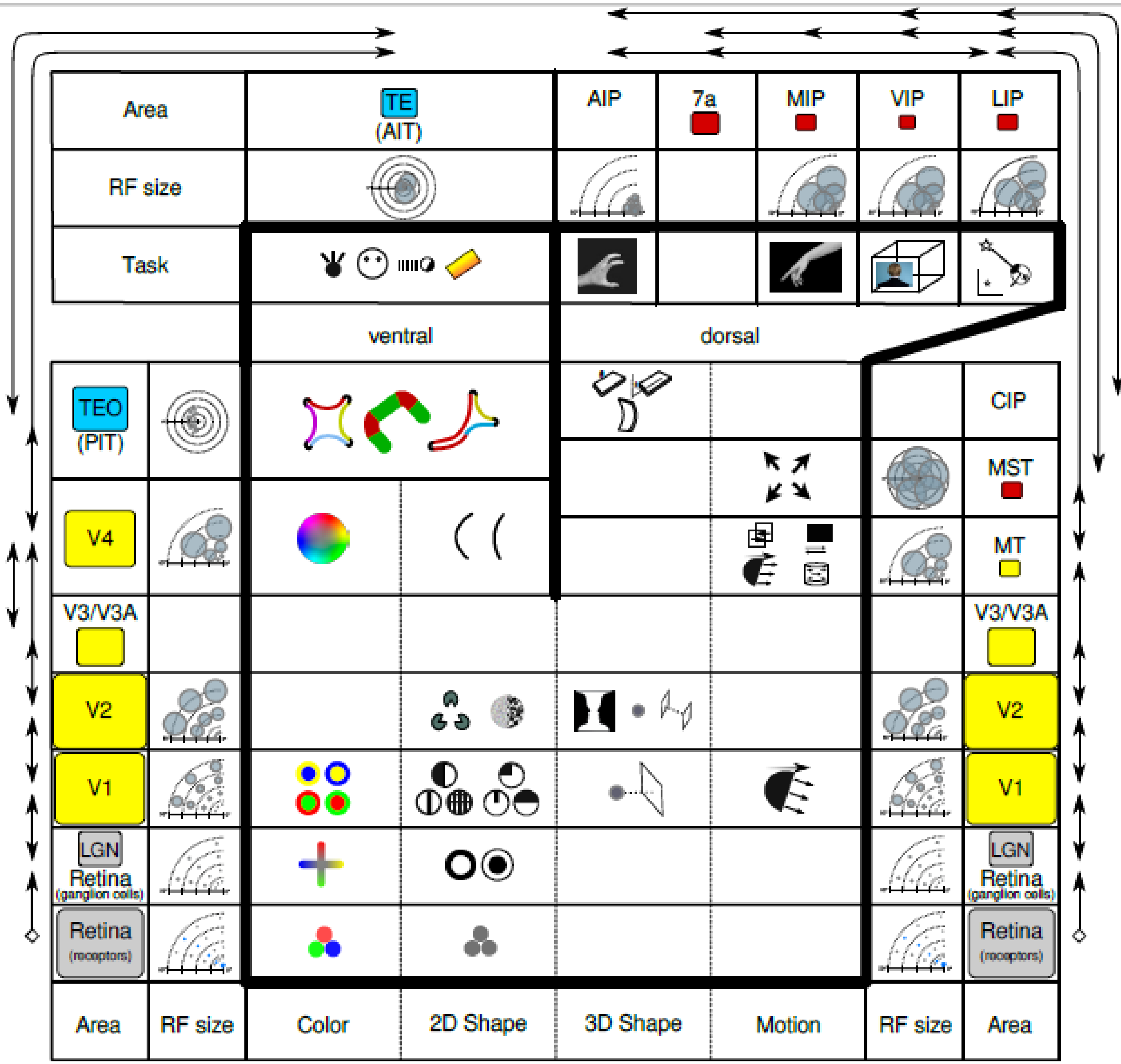
- Flat Hiererachies are inefficient
  - No sharing of computational recources
  - Transfer of experience across tasks is facilitated within the same representations
- Philipp Cimiano: ’Going beyond bag of words’





What do we know about primate's vision which is relevant for engineers and linguists?

- Richness of representation
- Deep Hierarchy versus flat Architectures
- **Separation of information**

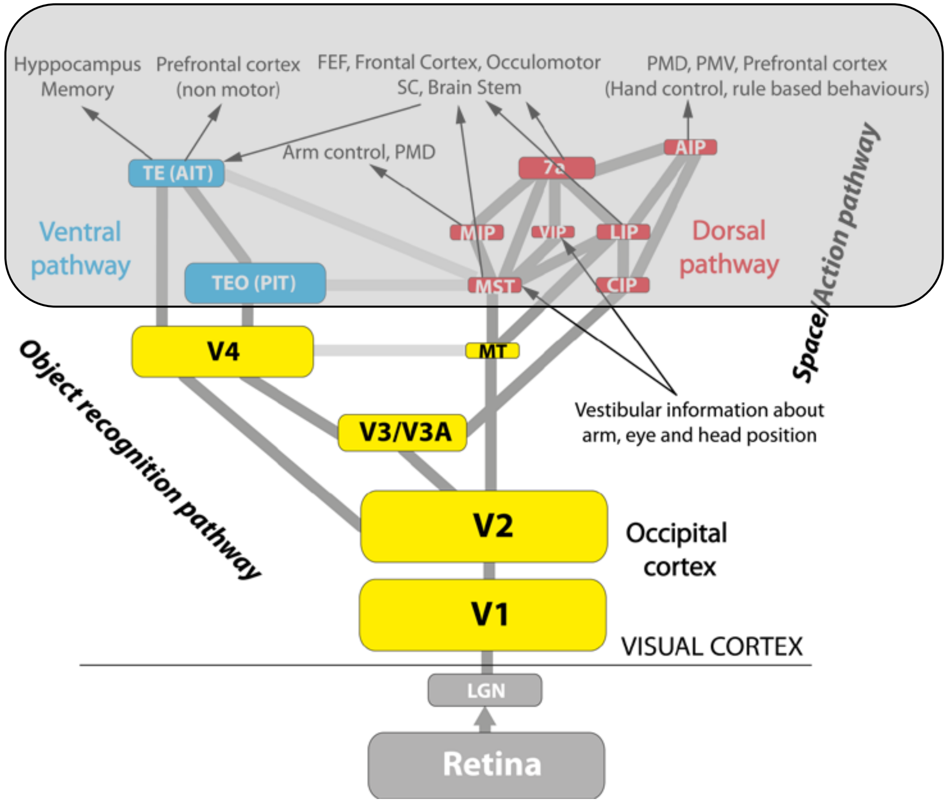

Area
TE
(AIT)
AIP
7a
MIP
VIP
LIP
RF size
Task
ventral
dorsal
TEO
(PIT)
CIP
MST
V4
MT
V3/V3A
V2
V1
LGN
Retina
(ganglion cells)
Retina
(receptors)
Area
RF size
Color
2D Shape
3D Shape
Motion
RF size
Area

Separation of Information

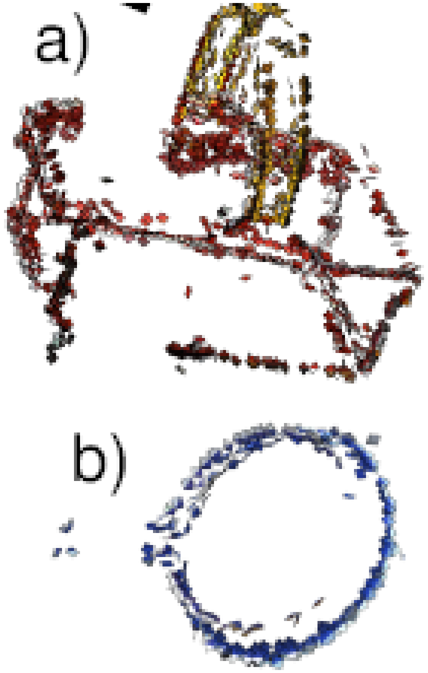
- Colour, 2D shape, 3D shape and motion become separated and are then up to a certain level of the hierarchy processed largely independently (while in the pixel domain these aspects are deeply intertwined)
- For learning problems this allows for cutting off non-relevant dimensions
- It allows also to discover relations between different aspects of visual information on a higher level (e.g., motion and 3D shape)

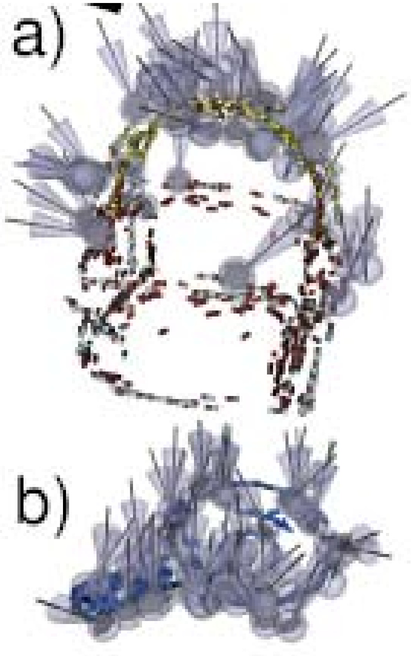
Overview

- Background Information
- The primate's vision system: A deep Hierarchy
  - Krüger et al. (2013), Deep Hierarchies in the Primate Visual Cortex: What Can We Learn for Computer Vision?, IEEE PAMI 2013.
- **From Signals to Symbols: Birth of the Object and its affordances**
  - Kraft et al. (2010), Development of Object and Grasping Knowledge by Robot Exploration. IEEE TAMD.
- Reflections

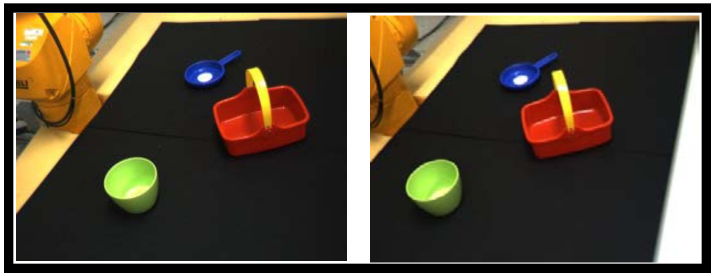


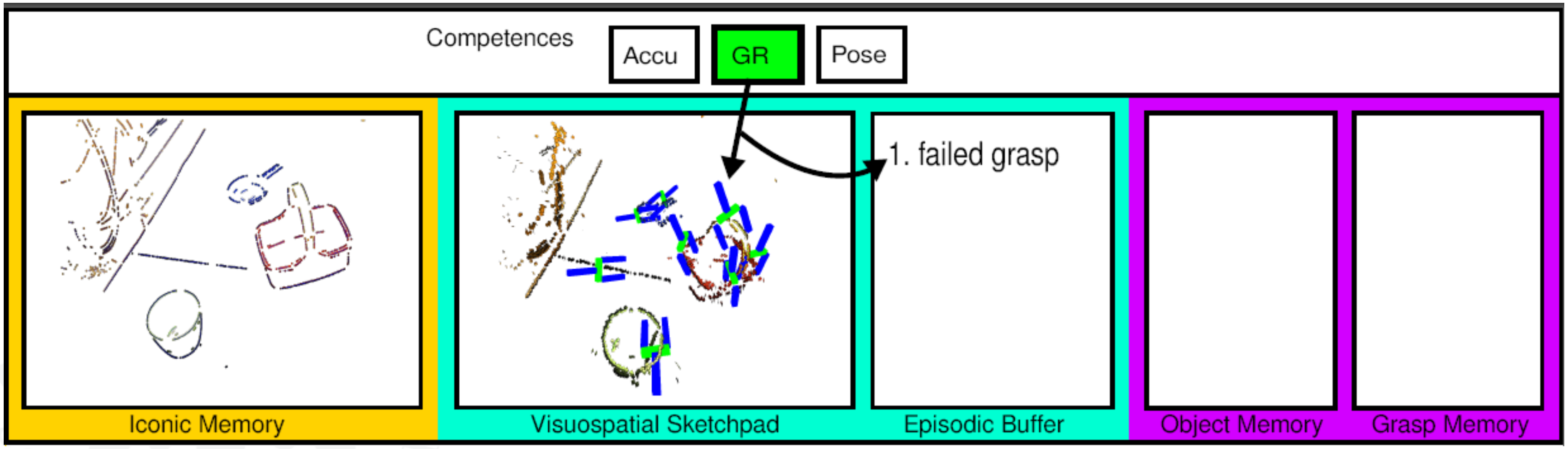
Boostrapping Robots: Grounding Objects and grasping affordances





World-view of the Innate System



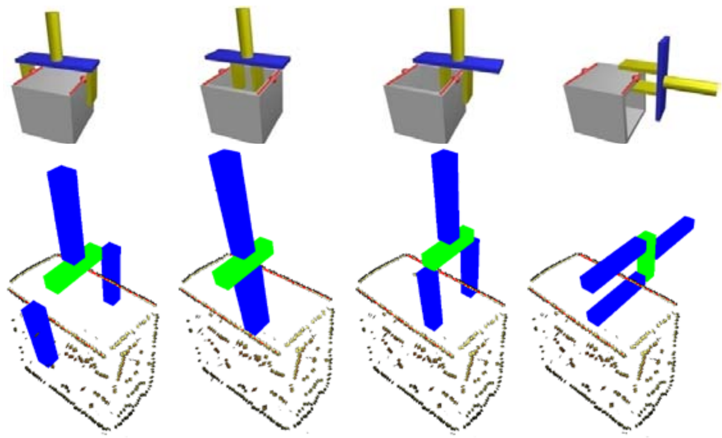


- Early Cognitive Vision (ECV) System
- GR: First 'reflex-like' behaviour

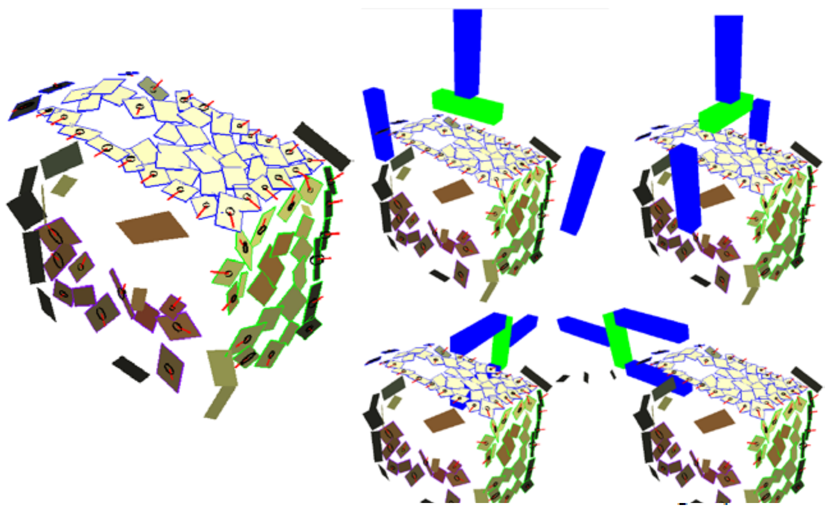
Edge and Surface based Grasp Affordances

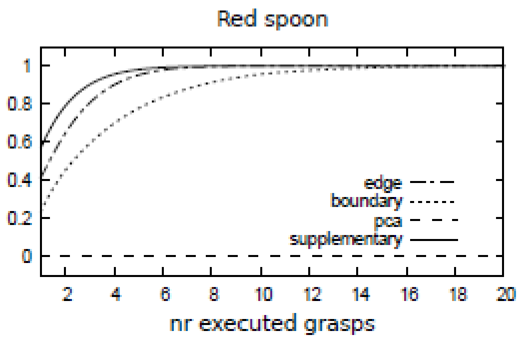
Video

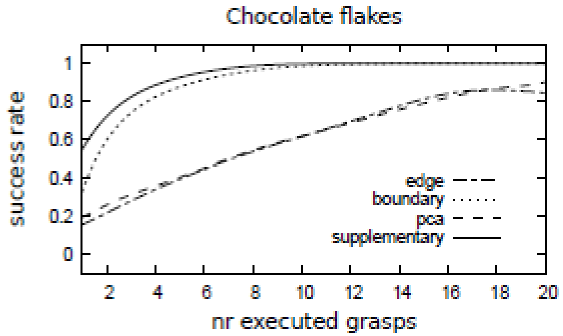
Edge based



Surface based



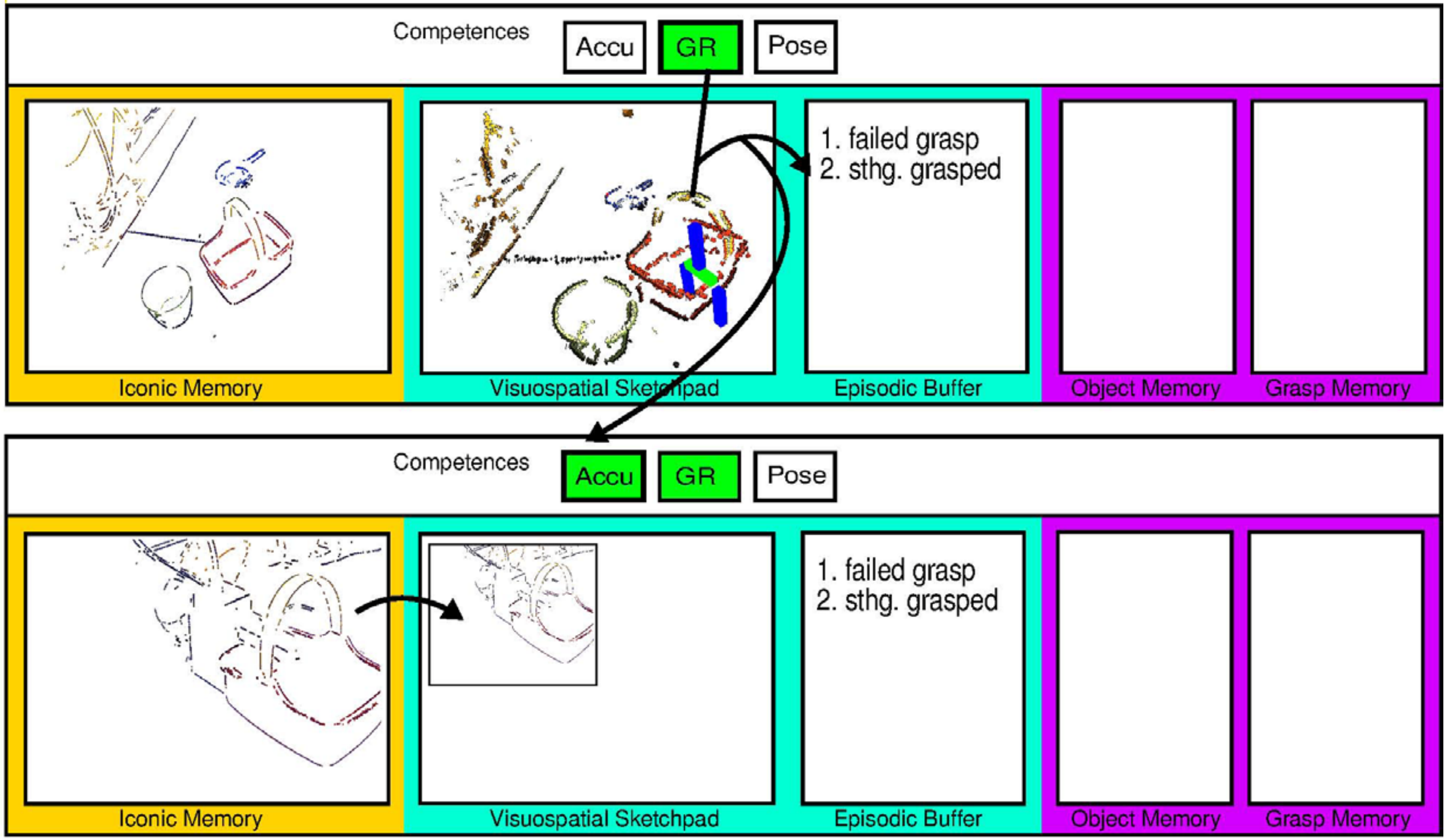




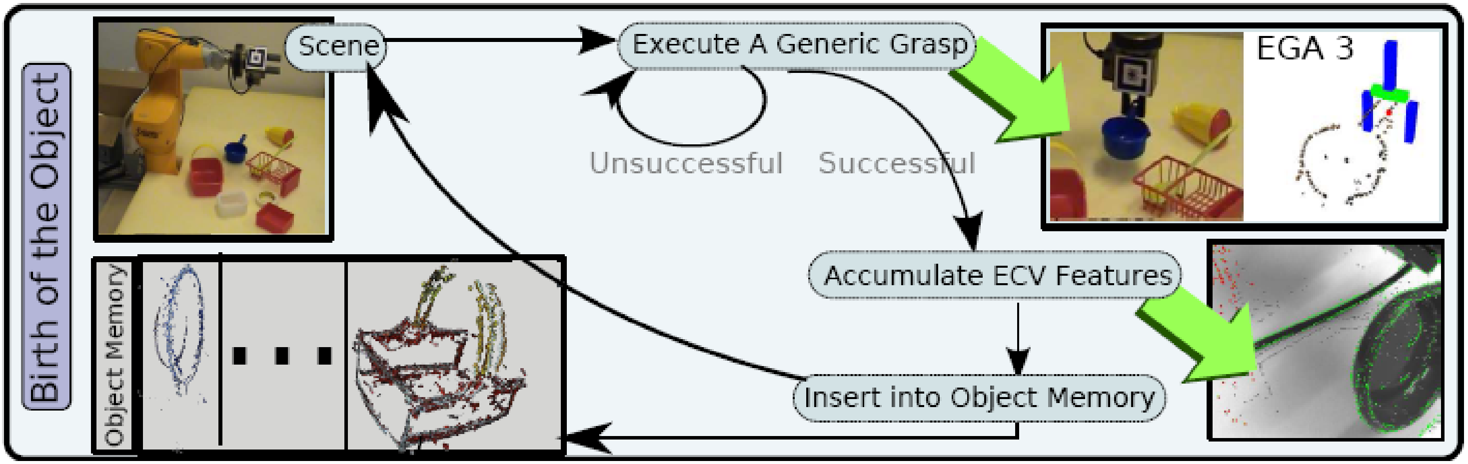
M. Popović, G. Kootstra, J. A. Jørgensen, D. Kragic and N. Krüger. Grasping Unknown Objects using an Early Cognitive Vision System for General Scene Understanding. IROS 2011 (nominated as one of the finalists for an IROS Awards)

G. Kootstra, M. Popovic, J. A. Jorgensen, K. Kuklinski, K. Miatliuk, D. Kragic and N. Krüger. Enabling grasping of unknown objects through a synergistic use of edge and surface information. International Journal of Robotics Research, vol. 31, no. 10, pp. 1190 - 1213, 2012.

World view of the Cognitive System



Birth of the Object

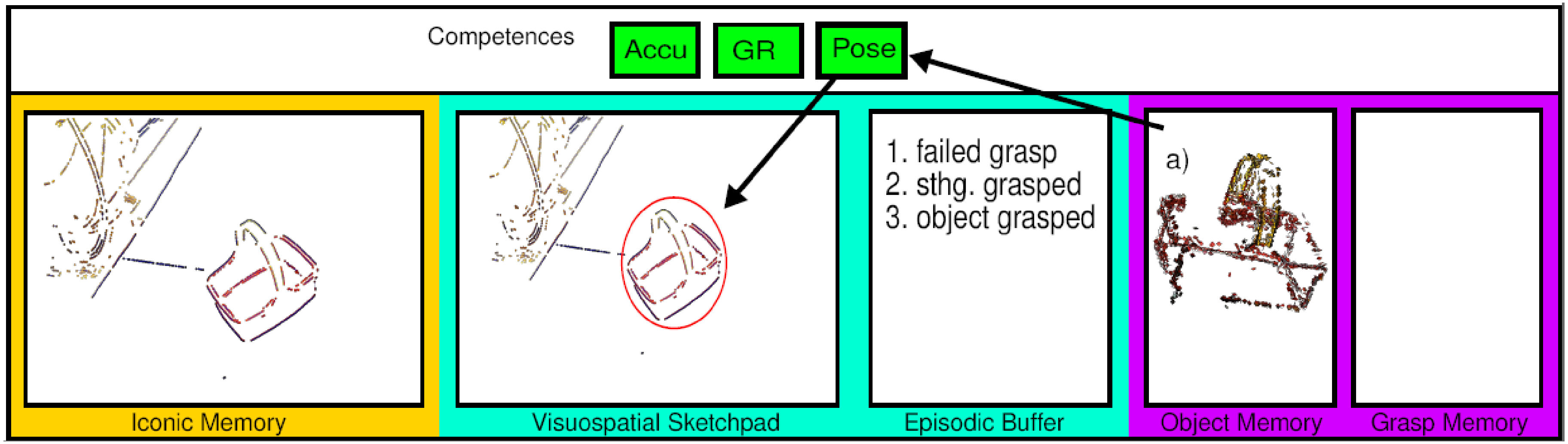


Object = Visual features changing according to robot motion

Video

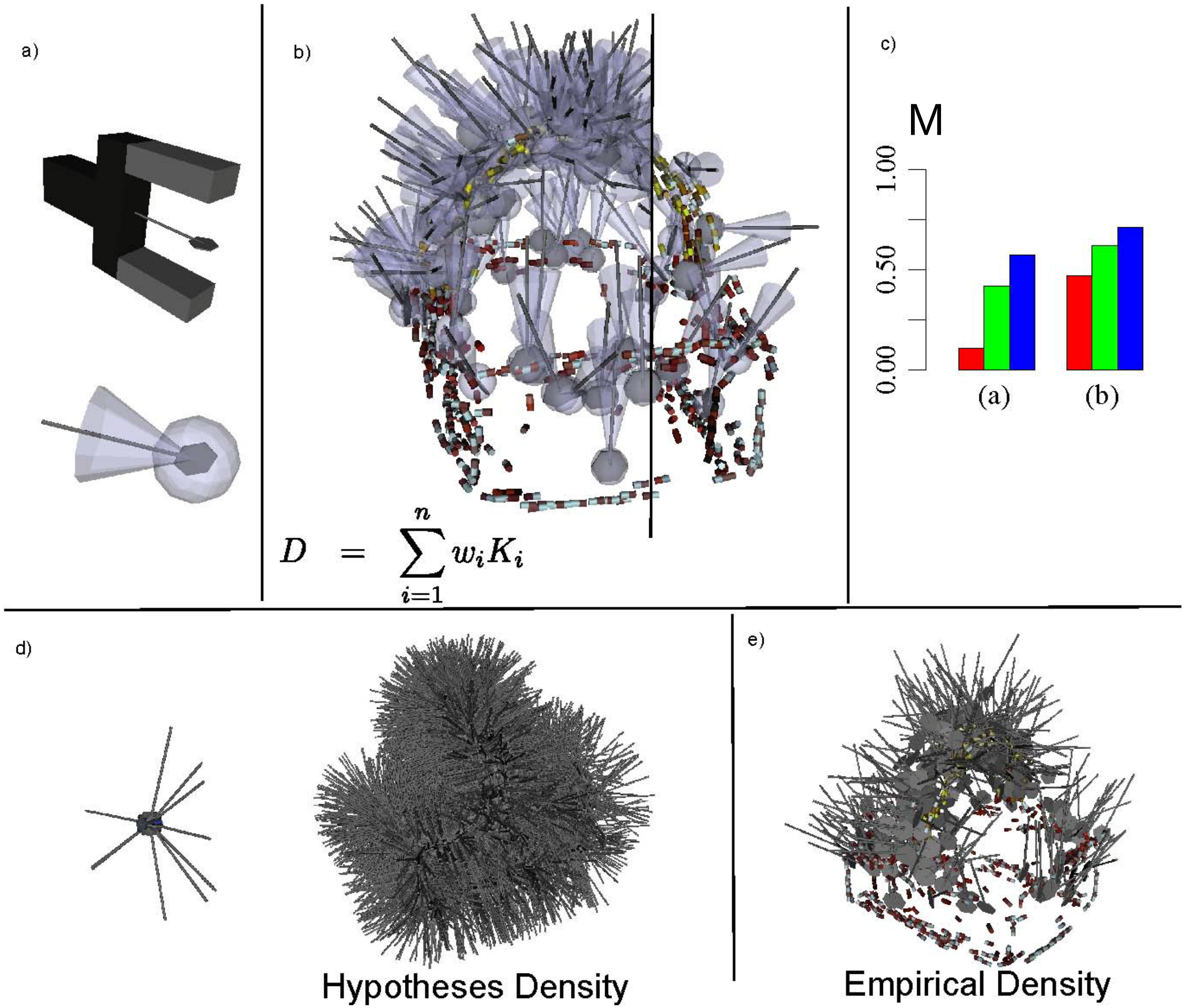
Kraft et al. (2008) Birth of the Object: Detection of Objectness and Extraction of Object Shape through OACs (IJHR).

World view of the Cognitive System



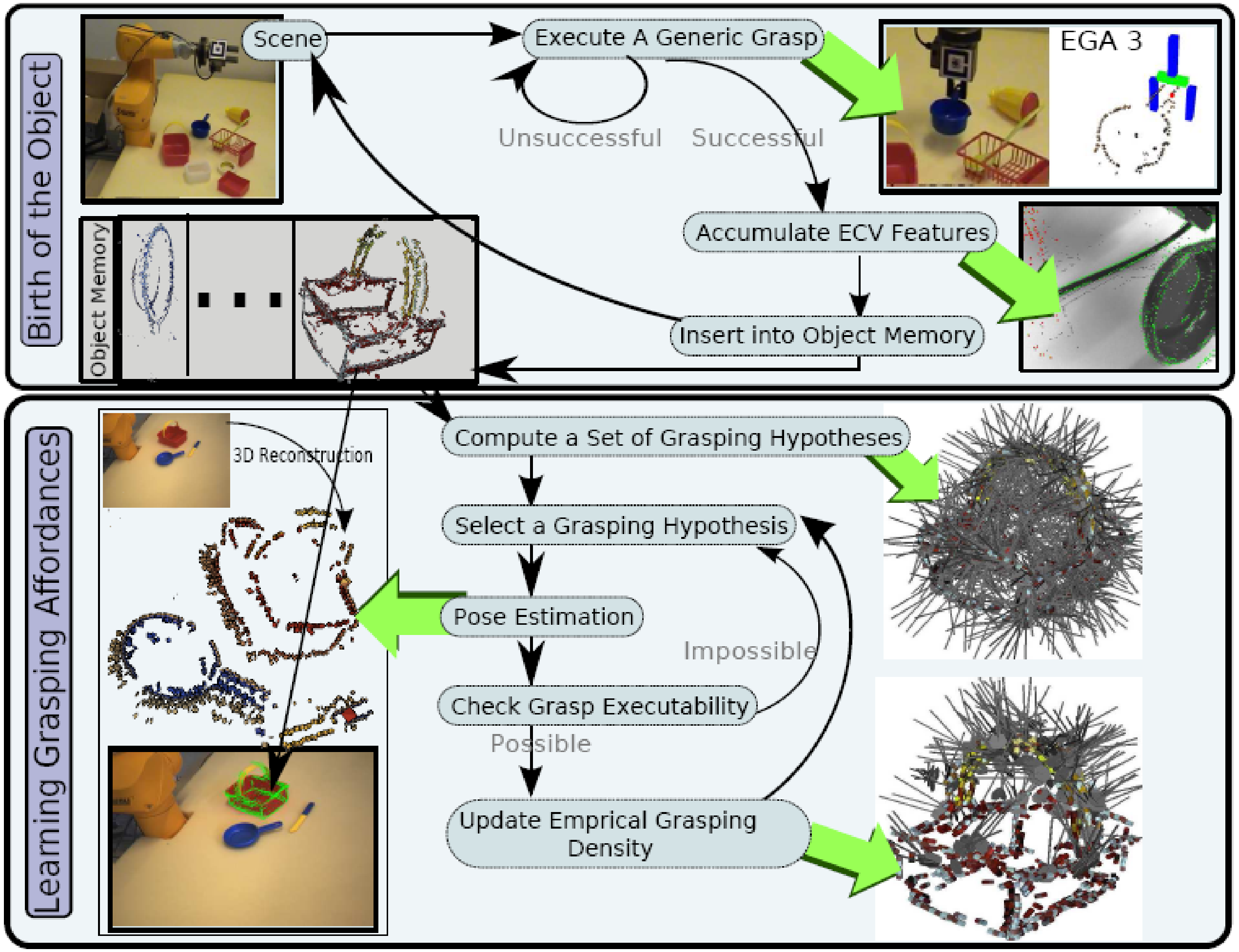
Pugeault et al. (2008), BMVC; Detry et al (2009), PAMI

Object specific Grasping: Grasp Densities



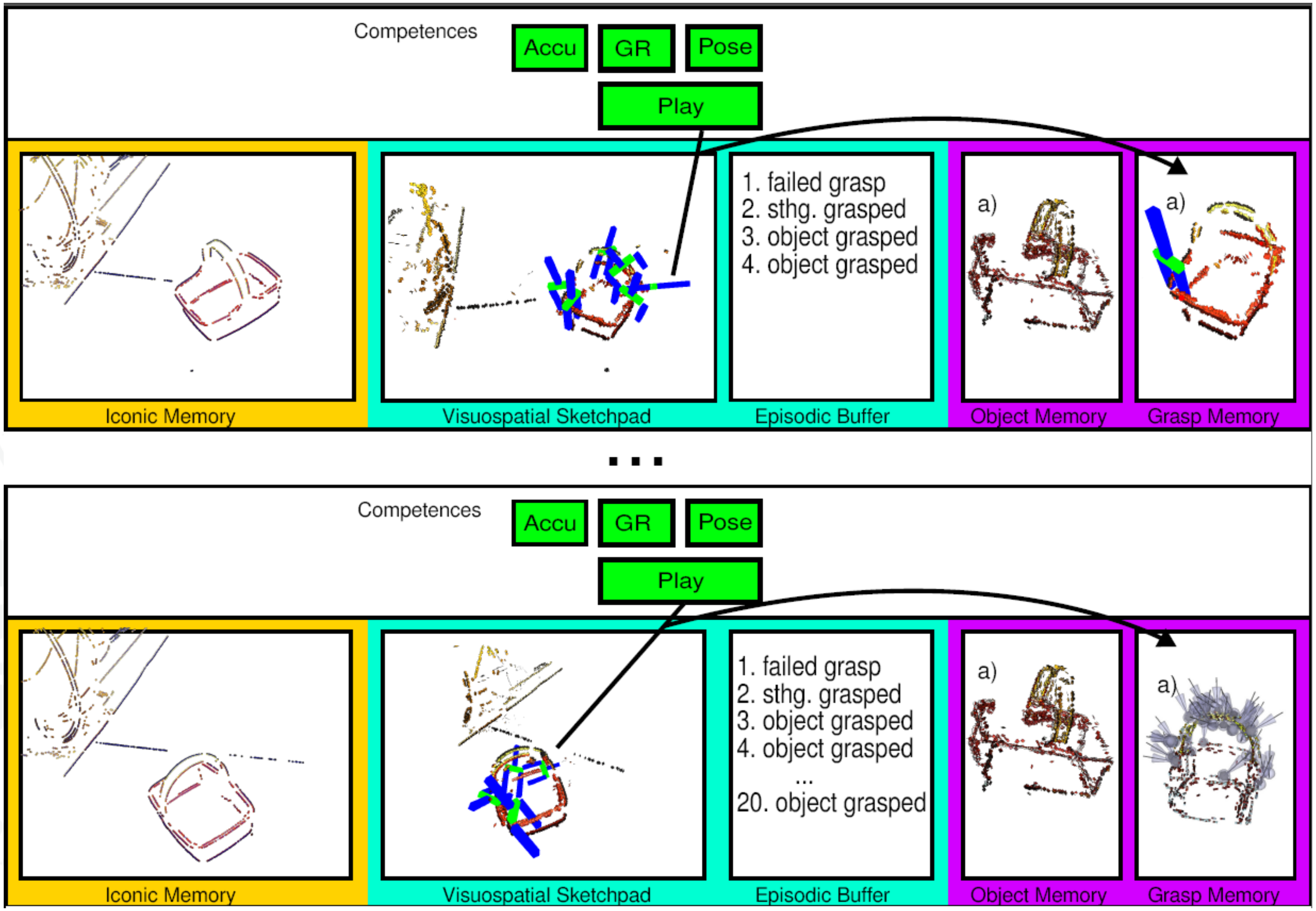
Detry et al. ICRA (2010)

Building up World Knowledge by Playing

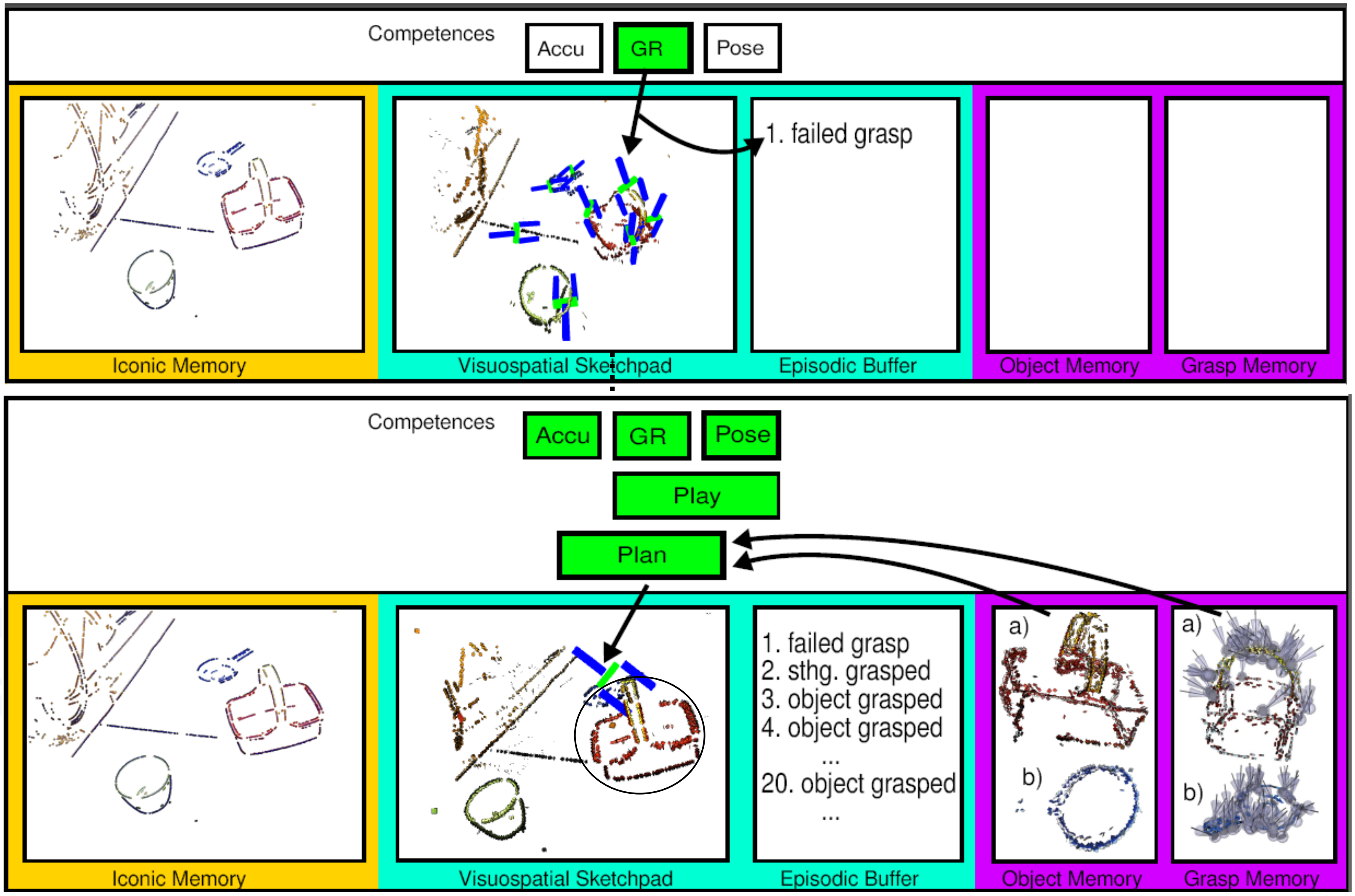


Video

World view of the Cognitive System

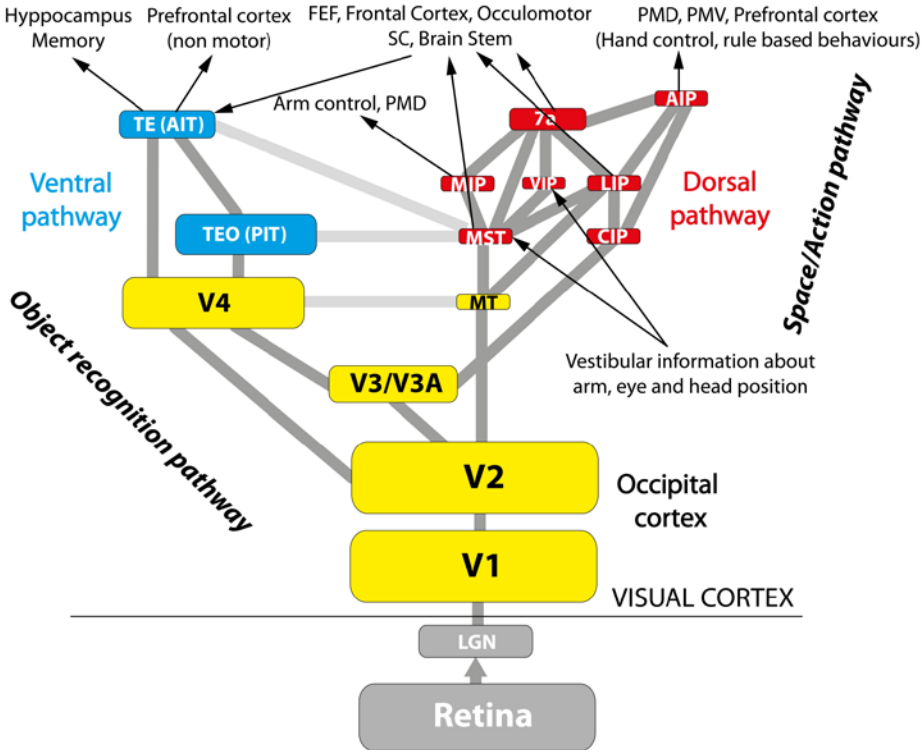


World view of the Cognitive System



Overview

- Background Information
- The primate's vision system: A deep Hierarchy
- From Signals to Symbols I: Feature Transformations
- **Reflections**

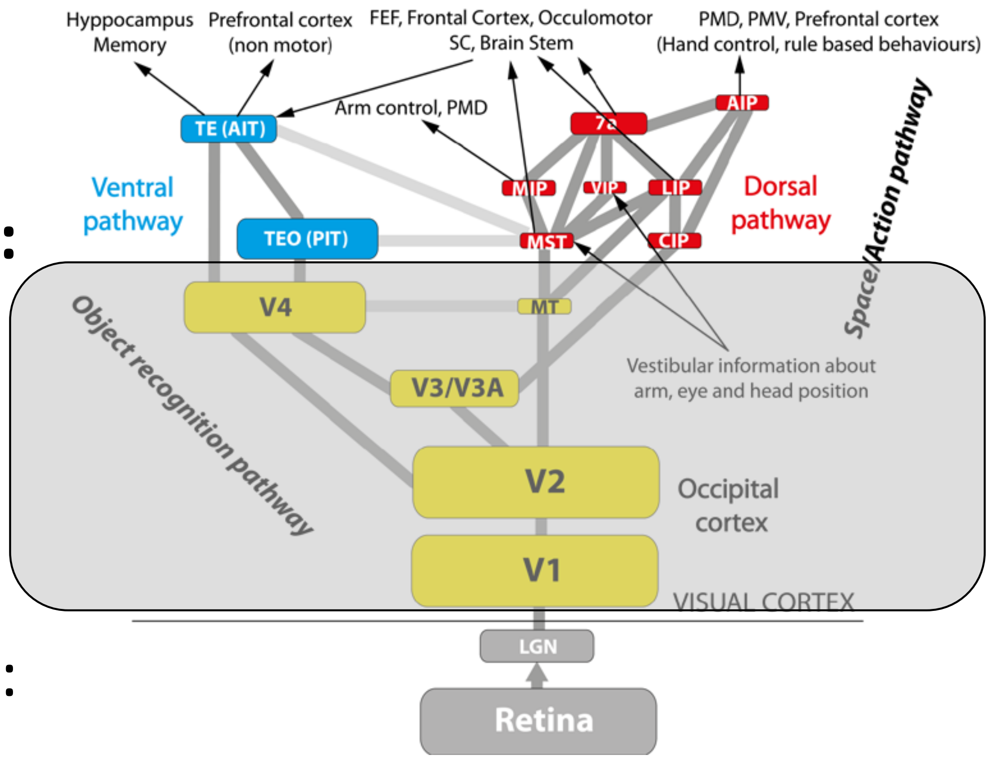


Some Reflections

- **Vision is probably a quite hard problem**
  - It uses resources occupying more than 50% of our brain
  - It is far from 'being solved'
- **Of that 70% is generic scene processing**
  - Deep hierarchy with increasing invariant representations
  - It spans a huge feature space as a basis for grounding processes
  - This space has a high degree of structure
    - Motion
    - Spatial Relations
- **We can learn from the human visual system (e.g., about what features to extract at what stage in the hierarchy)**
- **A crucial question is to learn/bootstrap/ground objects, linguistic categories and affordances making use of this huge space**
- **One example of the painstaking grounding process**

Overview

- Background Information
- The primate's vision system: A deep Hierarchy
- **From Signals to Symbols I: Feature Transformations**
  - P. Konig und N. Kruger 2003. Perspectives: Symbols as self-emergent entities in an optimization process of feature extraction and predictions. Biological Cybernetics.
- From Signals to Symbols II: Birth of the Object and its affordances
- Reflections



Symbols and multi-modal Primitives

- **Standard notion of a symbol**
  - (SE) symbols are condensed and discrete semantic representatives for certain pieces of knowledge (Expression)
  - (SS) on which operations can be performed that correspond to relevant functional relations in this framework (Syntax).
- **Multi modal primitives are**
  - (SE) condensed representations of local scene information
  - (SS) with which predictive relations can be formalized

Visual Primitives

Learning early visual Features

- **Early visual features have been learned from natural image statistics**
  - simple cells (Olshausen)
  - complex cells, disparity (Koenig, Hafner, Wiskott, ..)
- **The objectives used in unsupervised learning**
  - were based on criteria such as sparseness, slowness, reconstructability, ..
  - neither the relations of visual events nor the need to communicate these relations have been regarded
- **Proposal: Learn features within an early cognitive vision framework**
  - integrate need for predictions into unsupervised learning schemes

Four Hypotheses about extension of feature learning

- Hypothesis 1: **Unsupervised learning works also on other stages of (visual) processing**
- Hypothesis 2: **Predictions across visual events are a powerful approach to resolve ambiguities.**
- Hypothesis 3: Cortical Processing of sensorial information can be explained by a mutual optimization of condensation (CE) and predictions (CS).
  - Predictability (CS): A good feature gives rise to the prediction of other temporally and/or spatially distinct features and needs to be predictable from those.
  - Condensation (CE): Since predictive mechanisms work in a higher dimensional relational space for an efficient coding the local information has to be condensed.

Four Hypotheses about extension of feature learning

- Hypothesis 4: **In the process of mutual optimization of features and predictions symbols emerge as condensed entities on which predictions are performed.**
  - Predictability and Condensation correspond to the two properties if symbols (Expression (SE) and Syntax (SS))

Koenig, Krueger (submitted).Perspectives: Symbols as self-emergent entities in an optimization process of feature extraction and predictions