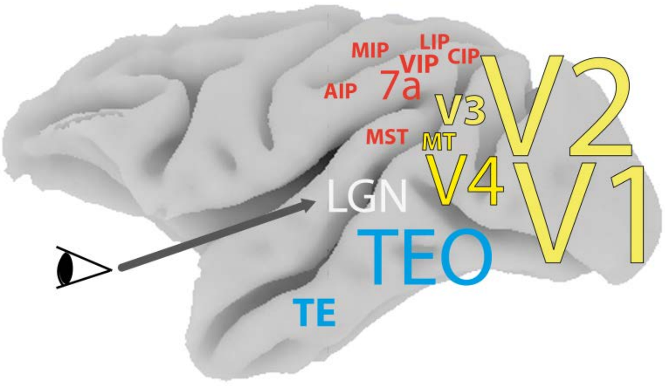


Deep Hierarchies in Human and Computer Vision

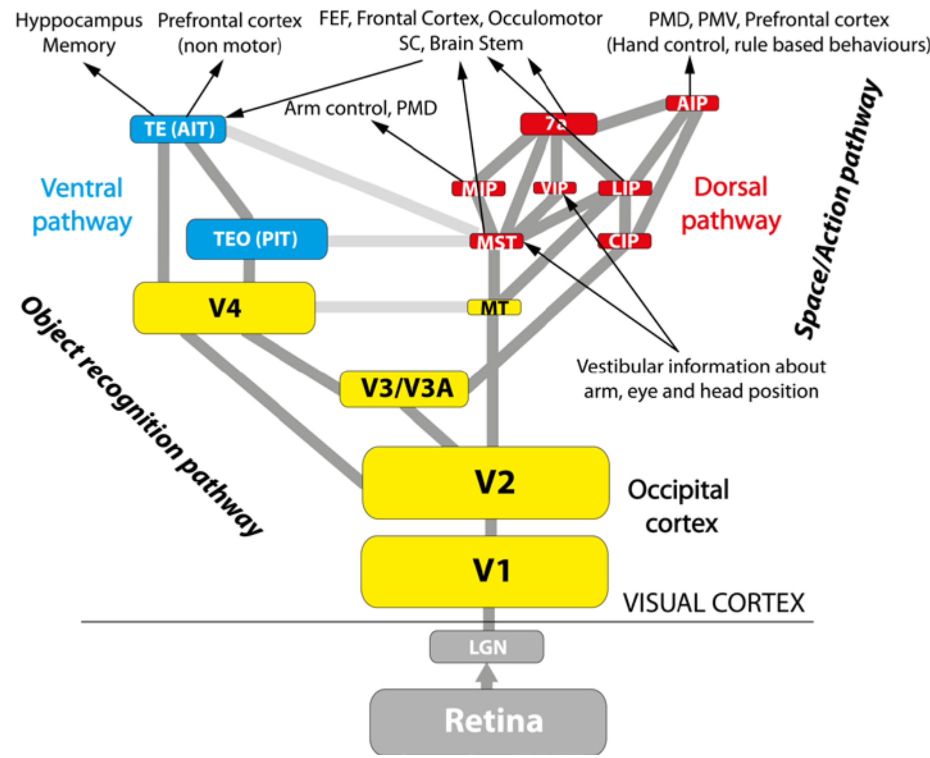
Norbert Kruger

University of Southern Denmark

Cognitive and Applied Robotics Group

Overview

- Some annoying prior remarks
- The primate's vision system: A deep Hierarchy
- SotA and Problems of research on deep hierarchical systems
- Reflections



IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, DOI: 10.1109/TPAMI.2012.272, AUTHOR FINAL DRAFT 1

IEEE Trans Pattern Anal Mach Intell. 2013 Aug;35(8):1847-71.

Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?

Norbert Krüger, Peter Janssen, Sinan Kalkan, Markus Lappe, Aleš Leonardis, Justus Piater, Antonio J. Rodríguez-Sánchez, Laurenz Wiskott

Abstract—Computational modeling of the primate visual system yields insights of potential relevance to some of the challenges that computer vision is facing, such as object recognition and categorization, motion detection and activity recognition or vision-based navigation and manipulation. This article reviews some functional principles and structures that are generally thought to underlie the primate visual cortex, and attempts to extract biological principles that could further advance computer vision research. Organized for a computer vision audience, we present *functional principles* of the *processing hierarchies* present in the primate visual system considering recent discoveries in neurophysiology. The hierarchal processing in the primate visual system is characterized by a sequence of different levels of processing (in the order of ten) that constitute a *deep hierarchy* in contrast to the *flat* vision architectures predominantly used in today's mainstream computer vision. We hope that the functional description of the deep hierarchies realized in the primate visual system provides valuable insights for the design of computer vision algorithms, fostering increasingly productive interaction between biological and computer vision research.

Index Terms—Computer Vision, Deep Hierarchies, Biological Modeling

1 Introduction

The history of computer vision now spans more than half a century. However, general, robust, complete satisfactory solutions to the major problems such as large-scale object, scene and activity recognition and categorization, as well as vision-based manipulation are still beyond reach of current machine vision systems. Biological visual systems, in particular those of primates, seem to accomplish these tasks almost effortlessly and have been, therefore, often used as an inspiration for computer vision researchers.

Interactions between the disciplines of "biological vision" and "computer vision" have varied in intensity throughout

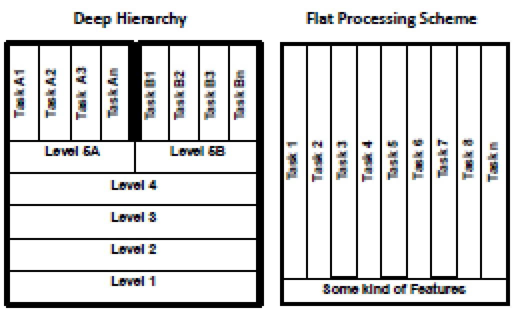
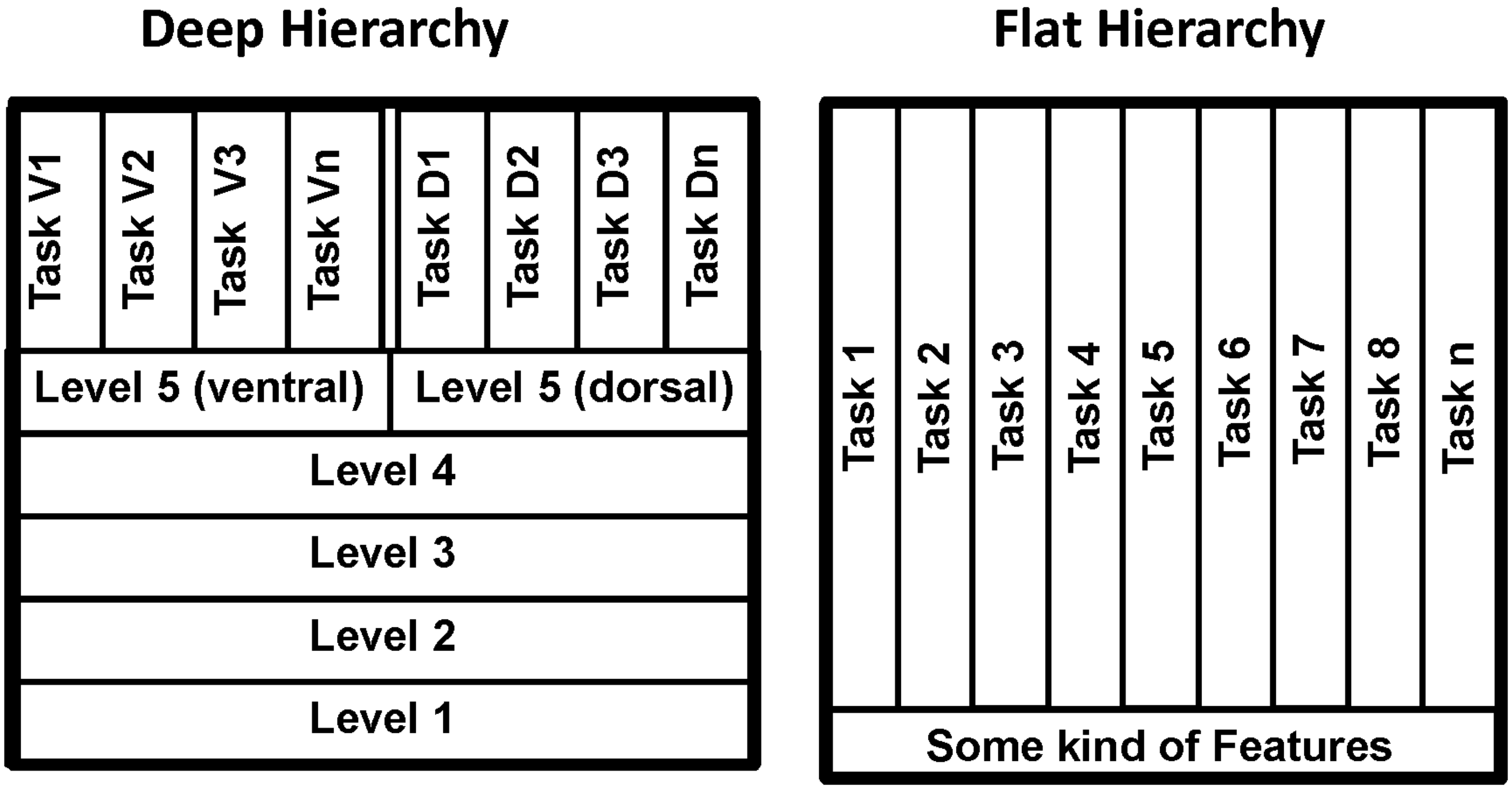


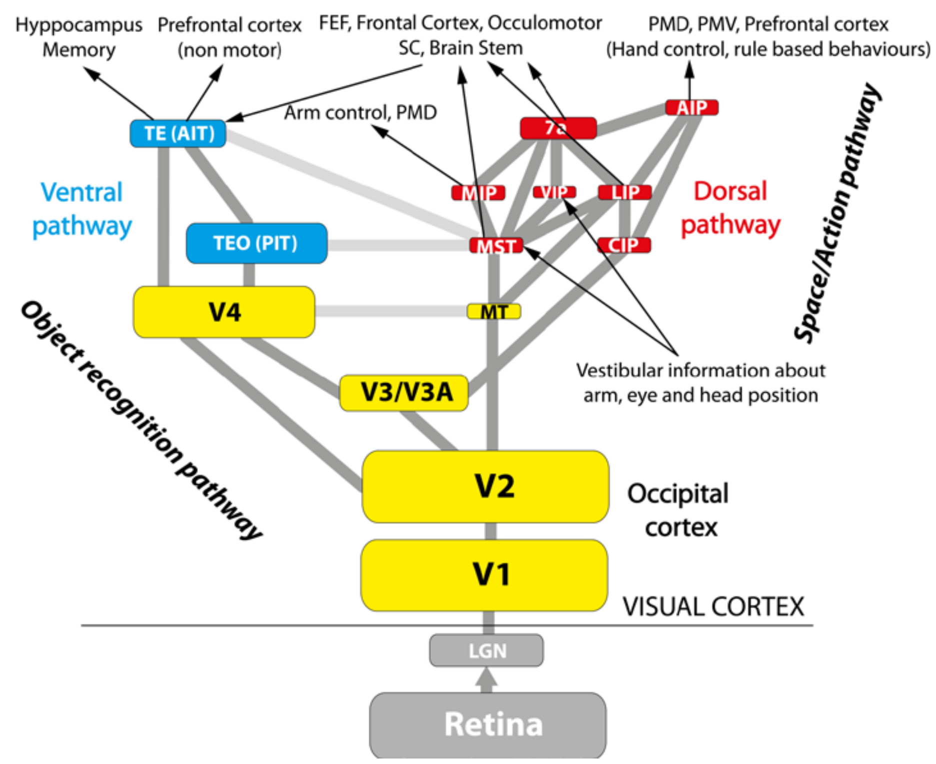
Fig. 1. Deep hierarchies and flat processing schemes

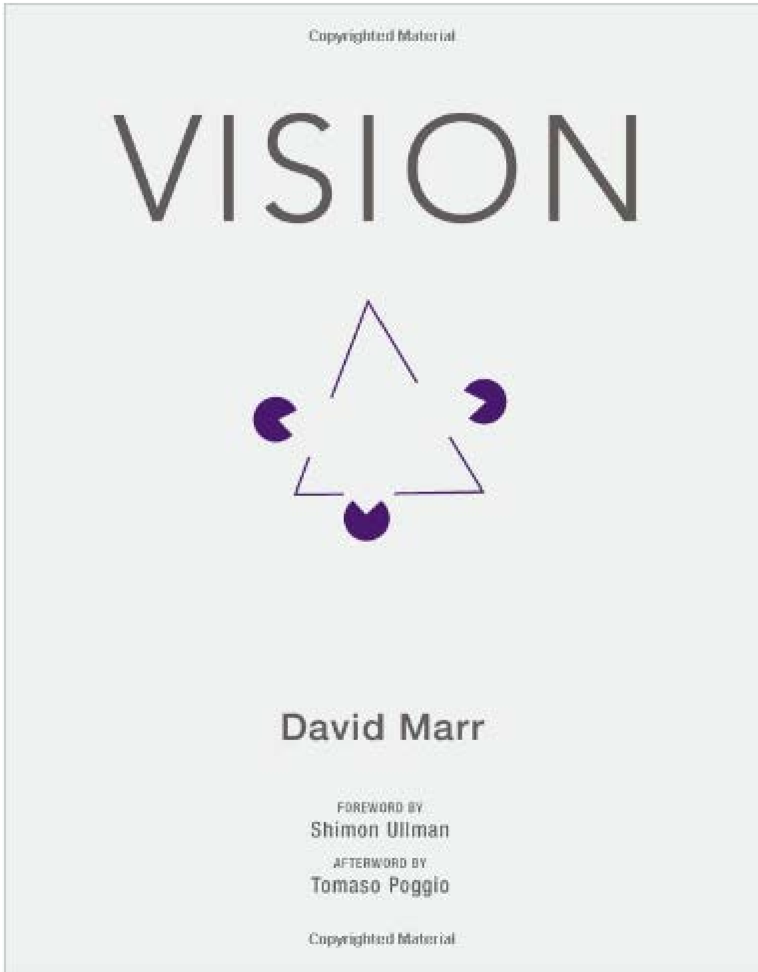
Flat versus deep Hierarchies



Overview

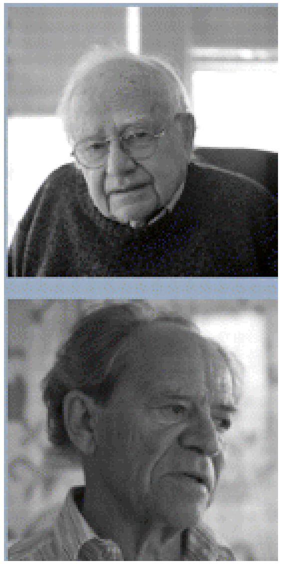
- **Some annoying prior remarks**
- The primate's vision system: A deep Hierarchy
- SotA and Problems of resaerch on deep hierarchical systems
- Reflections

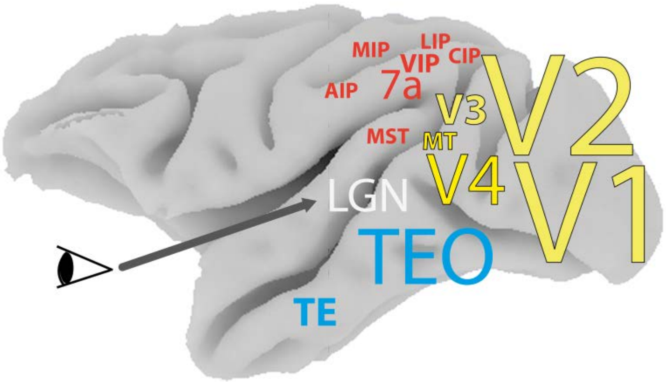


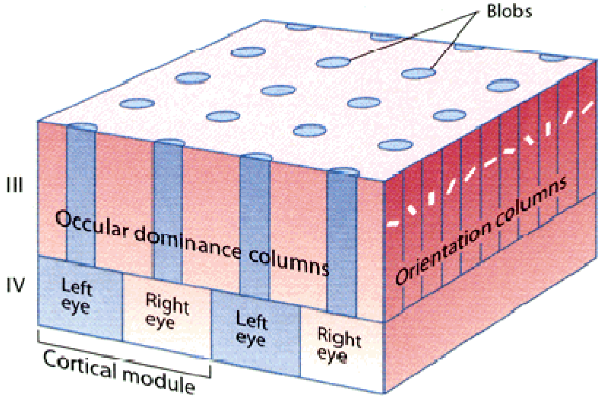


David Marr (1982): *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*.

David Hubel and Torsten Wiesel





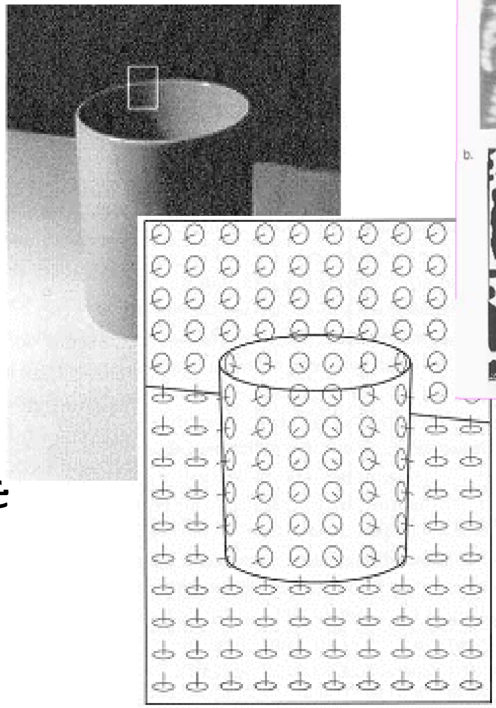


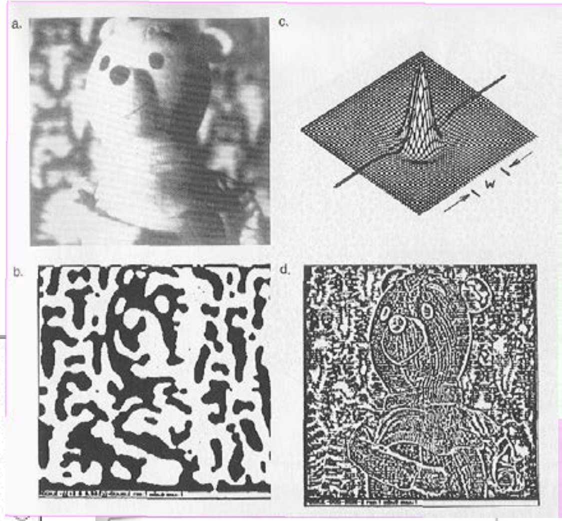
(Aus Gazzaniga et al., 1998)

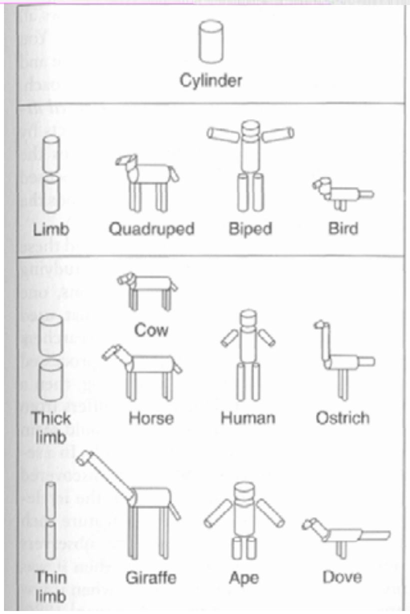
The Nobel Prize in Medicine 1981

Some remarks on the interaction of human vision research and computer vision

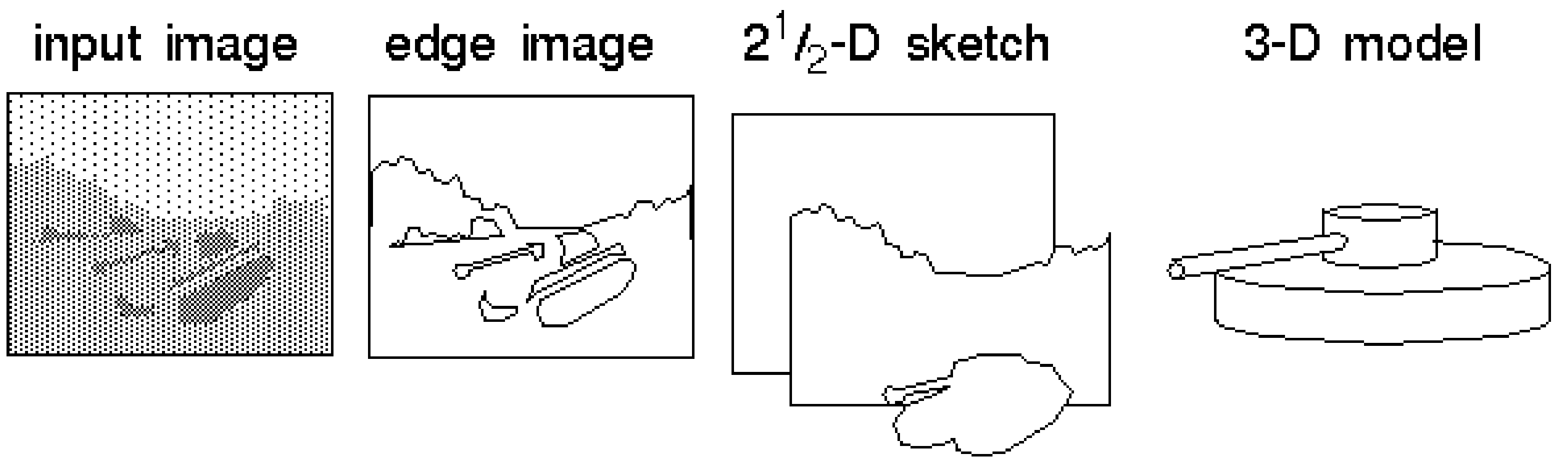
- **David Marr 1982: Vision: A computational investigation into the human representation and processing of visual information**
- **3 Stages**
  - Primal Sketch: Multi-scale Edge Detection
  - 2.5D Sketch: Viewer centered Scene Representation
  - 3D Sketch: Object Centered Representation







Viewer centred			*Object centred*
Input Image	Primal Sketch	2 1/2-D Sketch	3-D Model Representation
Perceived intensities	Zero crossings, blobs,edges, bars, ends, virtual lines, groups, curves boundaries.	Local surface orientation and discontinuities in depth and in surface orientation	3-D models hierarchically organised in terms of surface and volumetric primitives

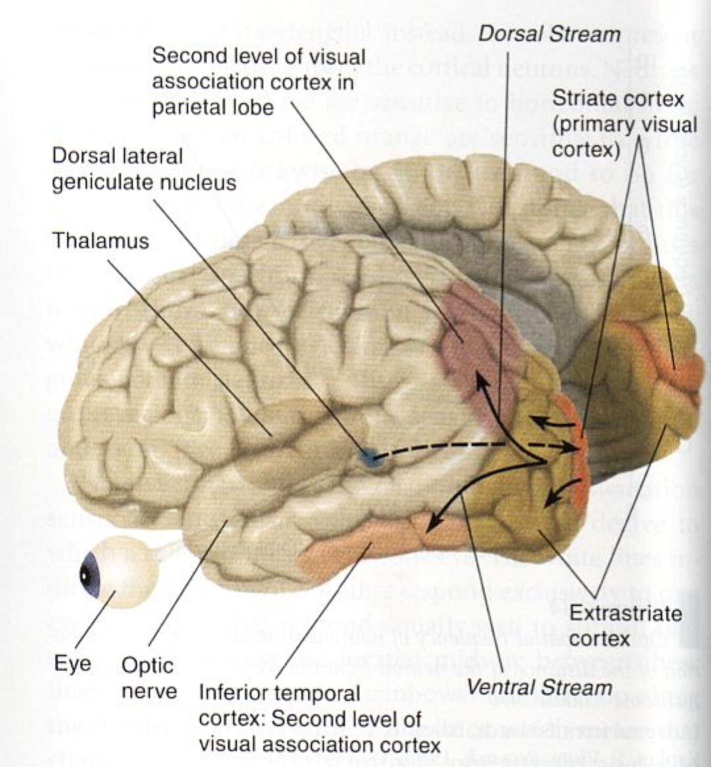


Why did that 'fail'? Two reasons

- **The project was too ambitious at Marr's time**
  - Lack of knowledge on low-level modalities
    - Optic flow
    - Edge detection
    - Stereo
    - Structure-from-Motion
- **Lack of computational resources**
  - Slow clock frequency
  - No GPUs

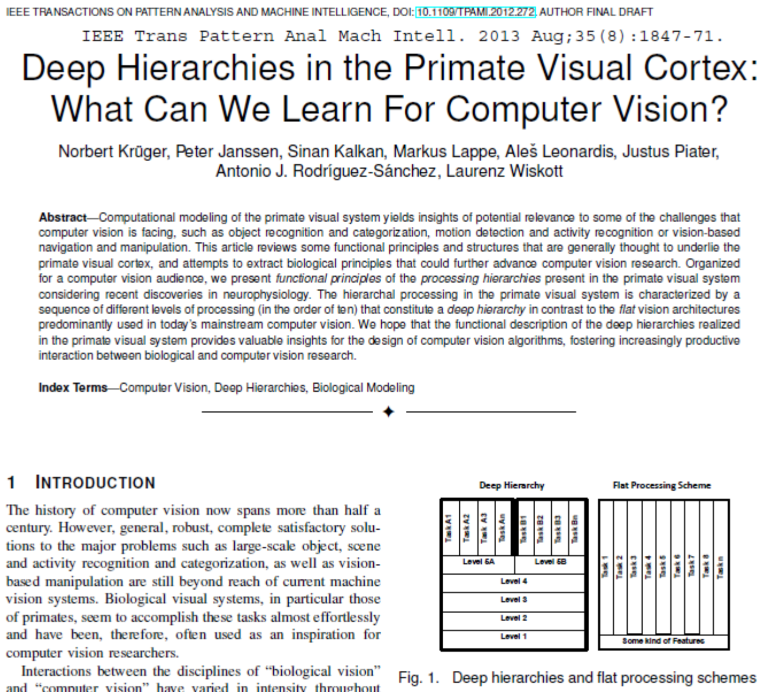
'Computer Vision' and 'Biological Vision'

- **In the 80th and 90th there was a strong link**
- **This link has been kind of diluted from 'both sides'**
  - Computer Vision became a sub-discipline of Machine Learning
  - Many neurophysiologists have given up on understanding the brain on a functional level
- **'Biologically inspired' got a somehow bad reputation**
  - Not efficient
  - Everything could somehow be biologically inspired



Maybe a restart is worthwhile

- **Much better understanding of early vision**
- **Significantly larger computational resources**
- **Still many unsolved problems in CV**
- **Aim of the paper**
  - Distill essential knowledge on the human visual system for Engineers



IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, DOI: 10.1109/TPAMI.2012.272 AUTHOR FINAL DRAFT 1

IEEE Trans Pattern Anal Mach Intell. 2013 Aug;35(8):1847-71.

Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?

Norbert Krüger, Peter Janssen, Sinan Kalkan, Markus Lappe, Aleš Leonardis, Justus Piater, Antonio J. Rodríguez-Sánchez, Laurenz Wiskott

Abstract—Computational modeling of the primate visual system yields insights of potential relevance to some of the challenges that computer vision is facing, such as object recognition and categorization, motion detection and activity recognition or vision-based navigation and manipulation. This article reviews some functional principles and structures that are generally thought to underlie the primate visual cortex, and attempts to extract biological principles that could further advance computer vision research. Organized for a computer vision audience, we present *functional principles* of the *processing hierarchies* present in the primate visual system considering recent discoveries in neurophysiology. The hierarchal processing in the primate visual system is characterized by a sequence of different levels of processing (in the order of ten) that constitute a *deep hierarchy* in contrast to the *flat* vision architectures predominantly used in today's mainstream computer vision. We hope that the functional description of the deep hierarchies realized in the primate visual system provides valuable insights for the design of computer vision algorithms, fostering increasingly productive interaction between biological and computer vision research.

Index Terms—Computer Vision, Deep Hierarchies, Biological Modeling

1 INTRODUCTION

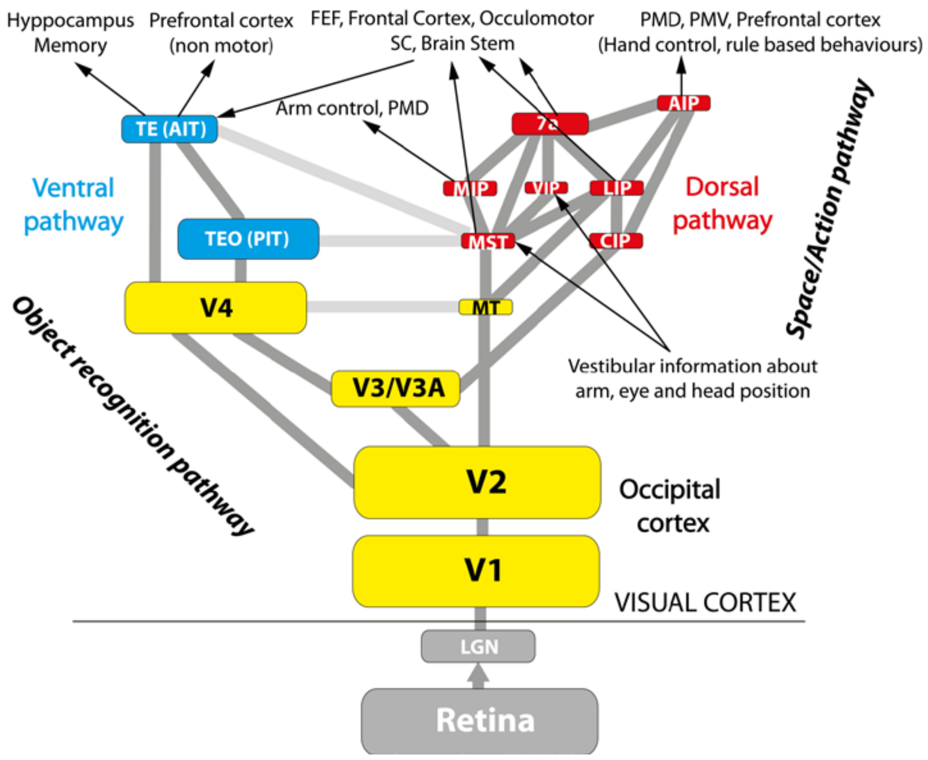
The history of computer vision now spans more than half a century. However, general, robust, complete satisfactory solutions to the major problems such as large-scale object, scene and activity recognition and categorization, as well as vision-based manipulation are still beyond reach of current machine vision systems. Biological visual systems, in particular those of primates, seem to accomplish these tasks almost effortlessly and have been, therefore, often used as an inspiration for computer vision researchers.

Interactions between the disciplines of "biological vision" and "computer vision" have varied in intensity throughout

Fig. 1. Deep hierarchies and flat processing schemes

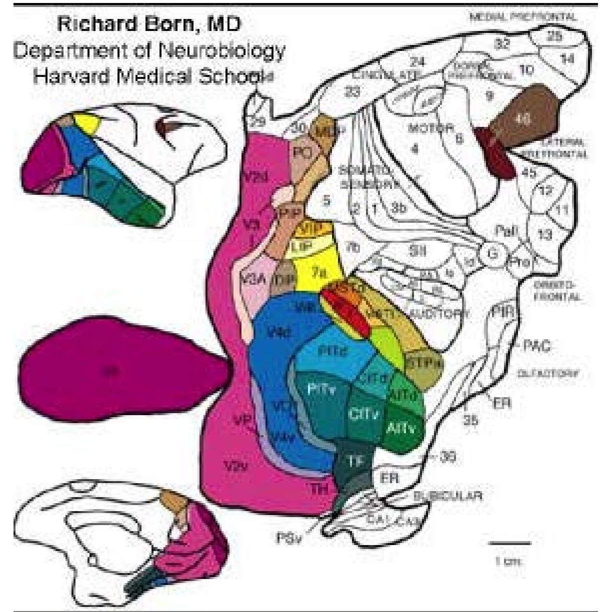
Overview

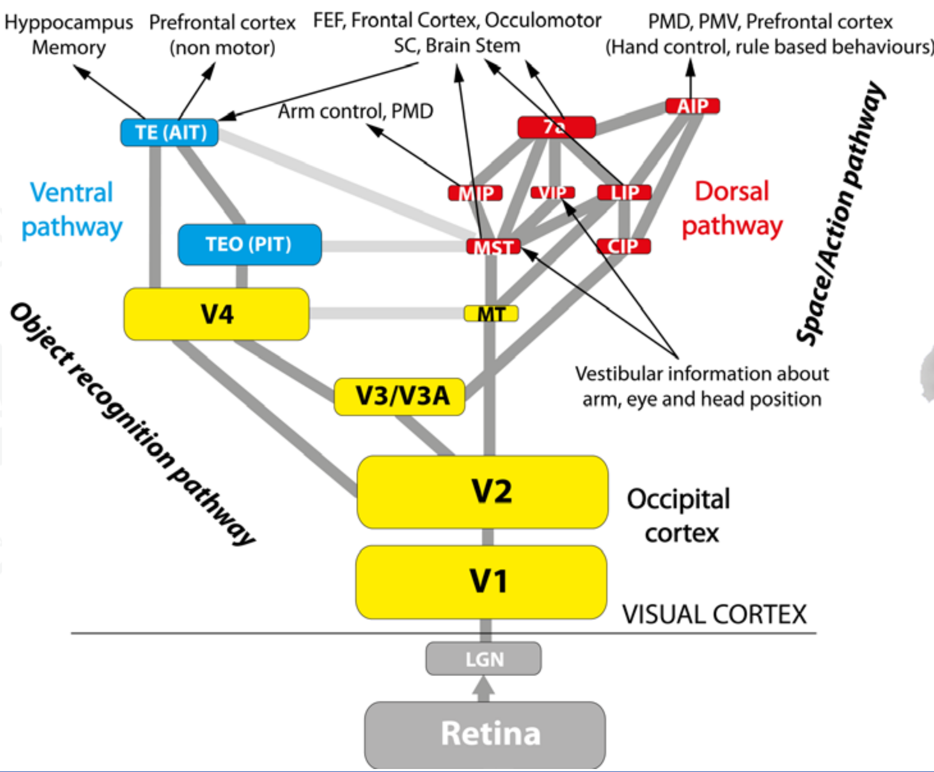
- Some annoying prior remarks
- **The primate's vision system: A deep Hierarchy**
- SotA and Problems of resaerch on deep hierarchical systems
- Reflections

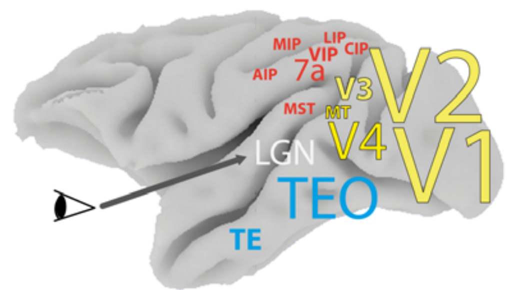


Basic facts

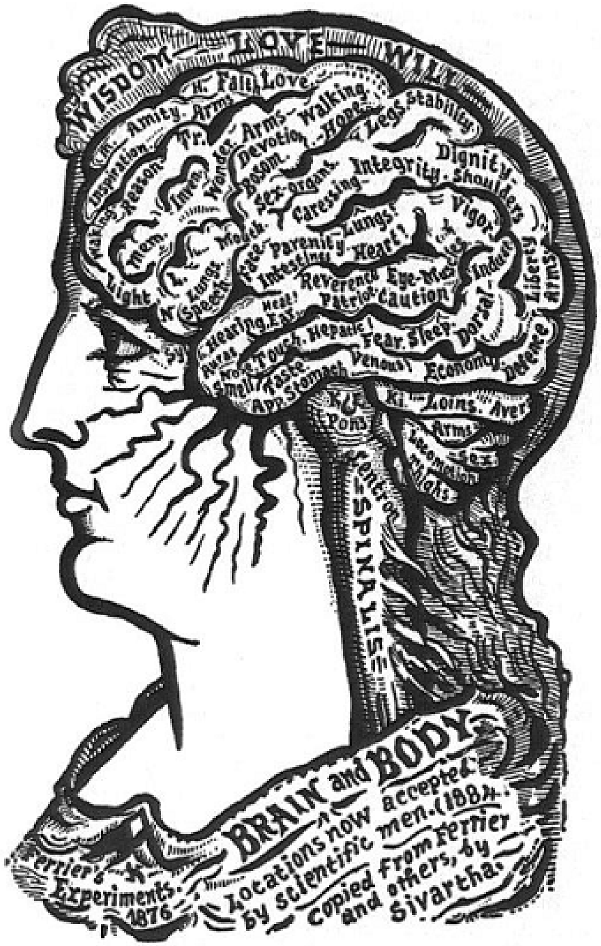
- **55% of the neo-cortex of the primate brain is concerned with vision**
- **Devision in**
  - Occipitel Cortex
  - Dorsal Pathway
  - Ventral Pathway



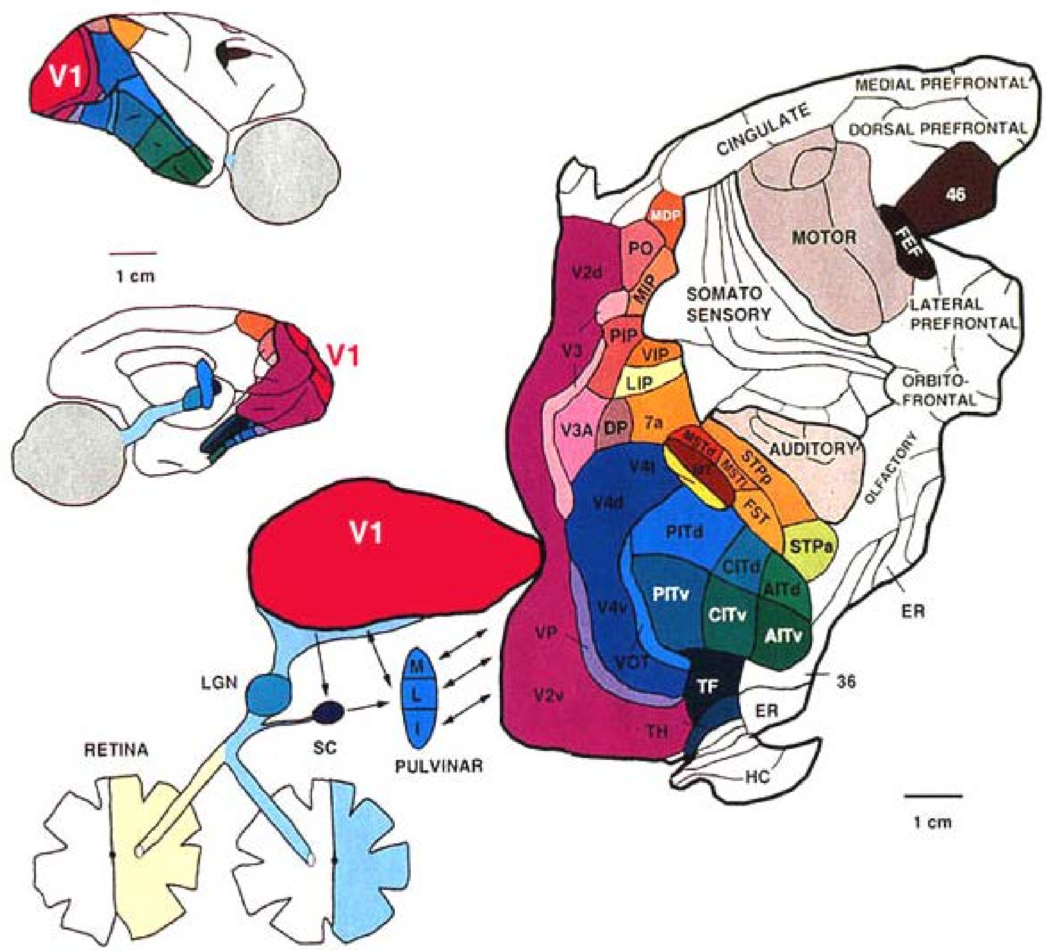




Brain Maps

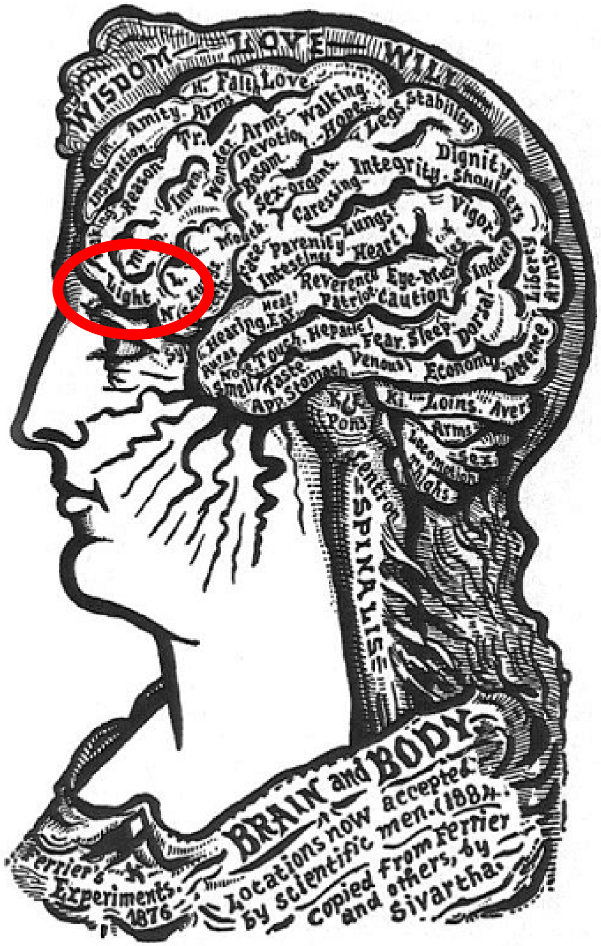


Dr. Alesha Sivartha in the late 1800s (published in his metaphysical book *The Book of Life: The Spiritual and Physical Constitution of Man*)'

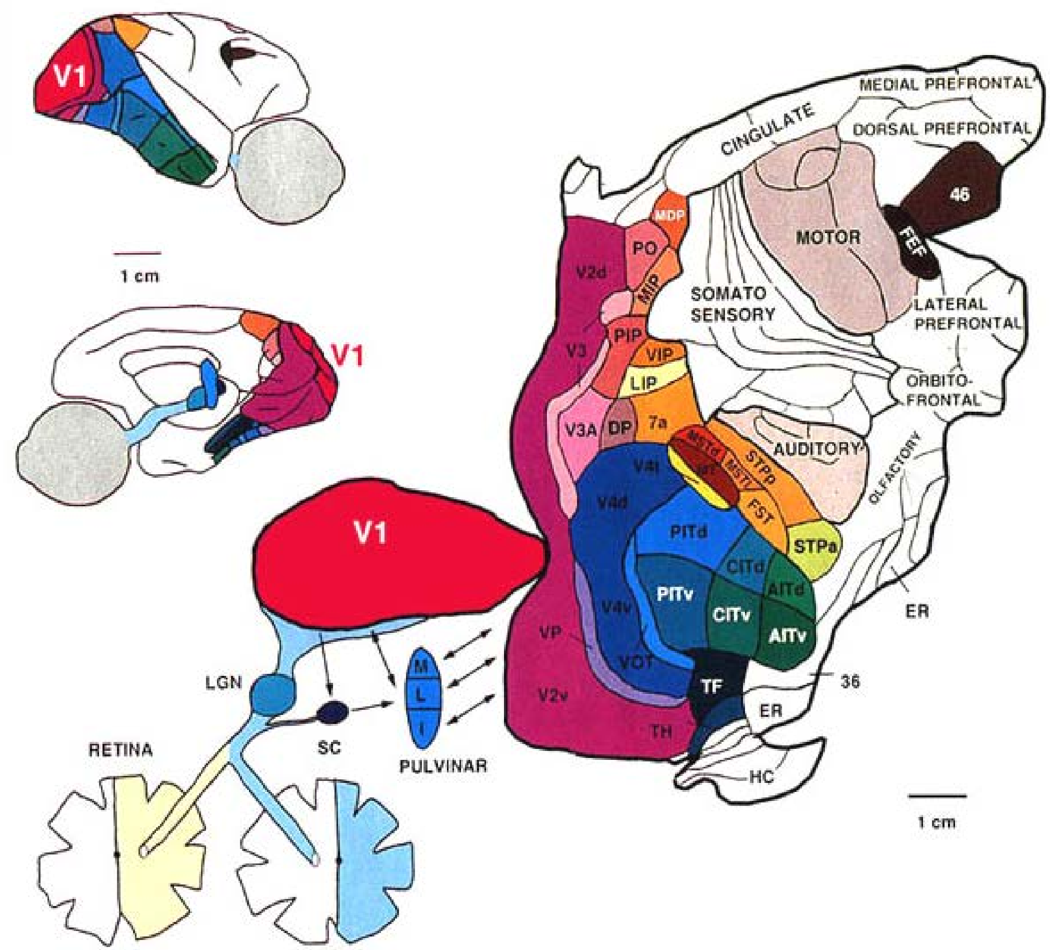


From: van Essen 1992

Brain Maps

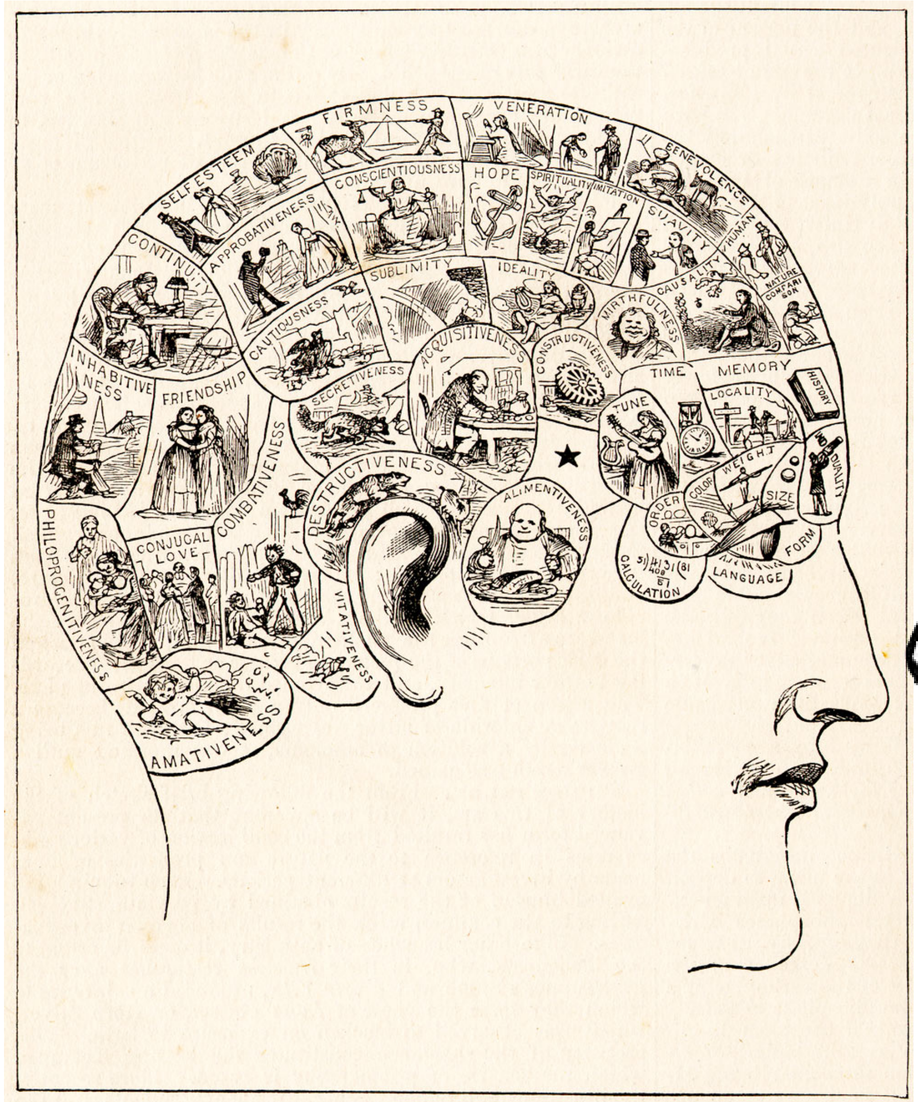


Dr. Alesha Sivartha in the late 1800s (published in his metaphysical book *The Book of Life: The Spiritual and Physical Constitution of Man*)'

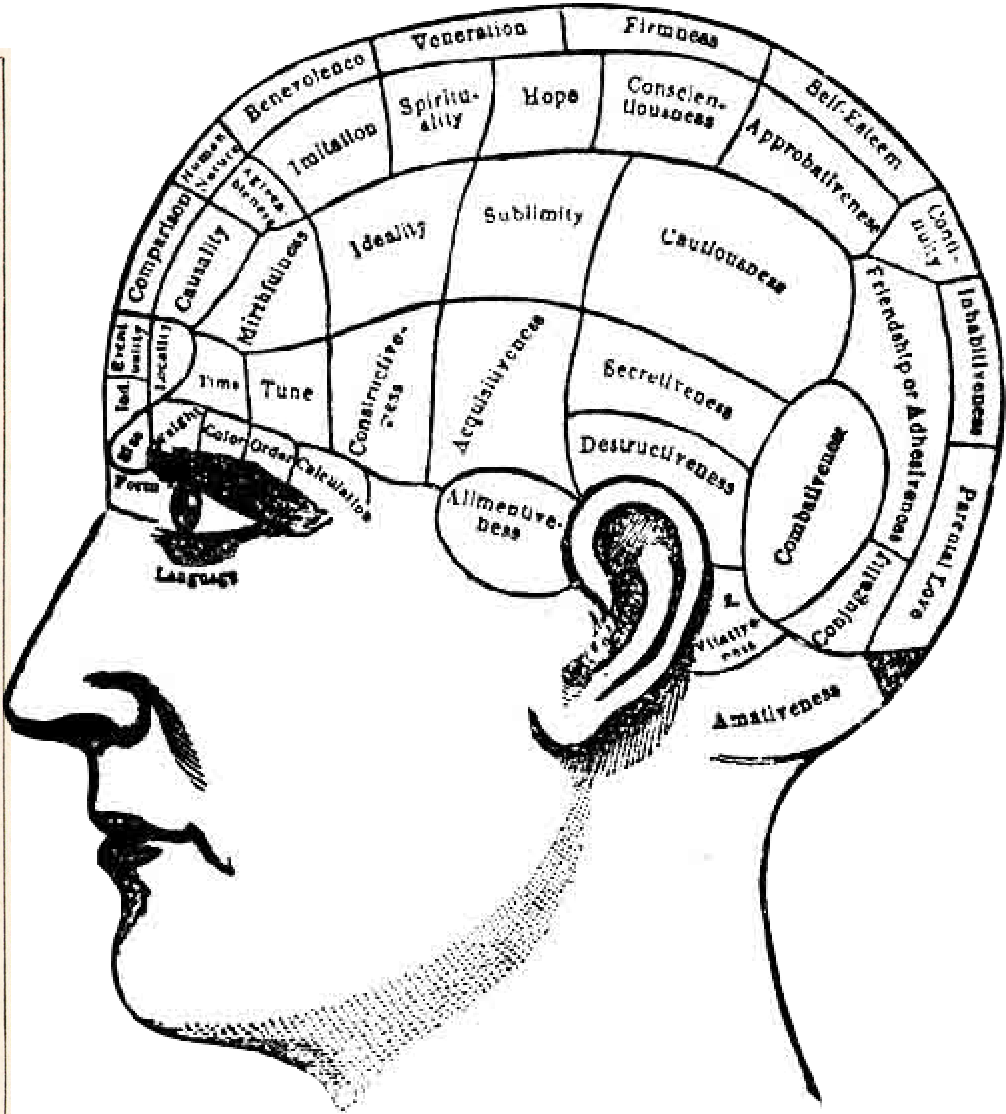


From: van Essen 1992

Gall (1758–1828): Phrenology



Phrenological Chart of the Faculties.



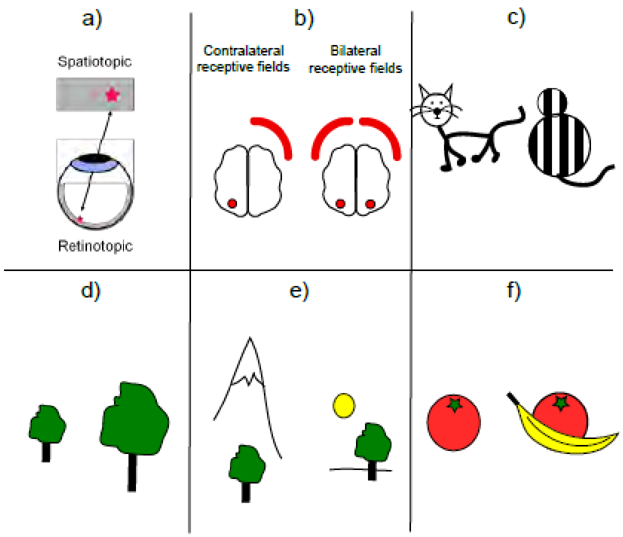
Basic Facts

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.

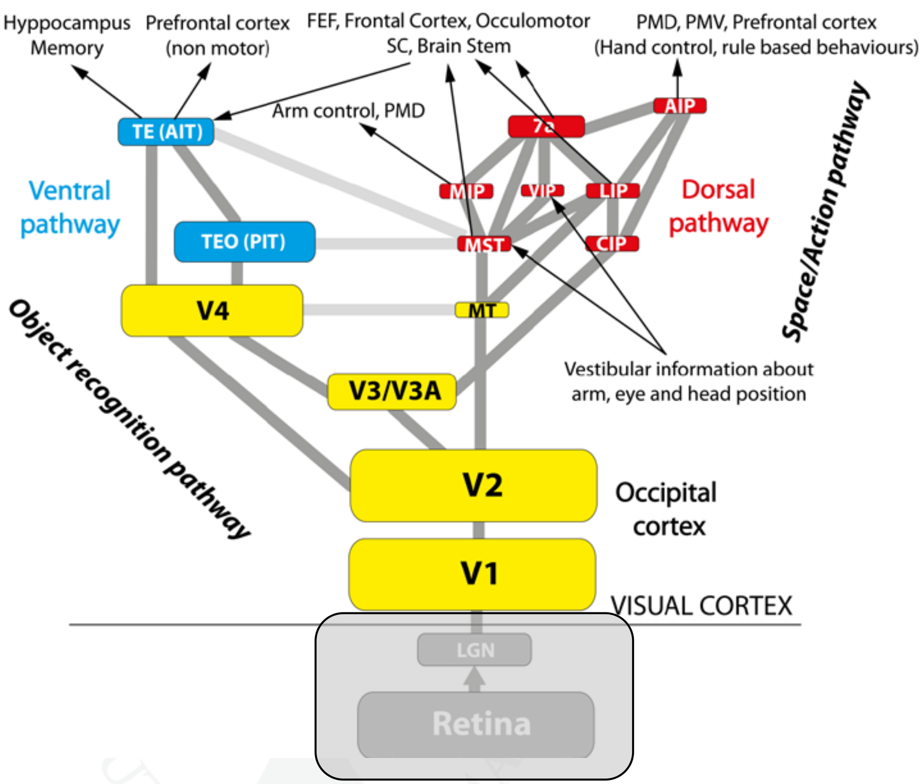
Basic Terms

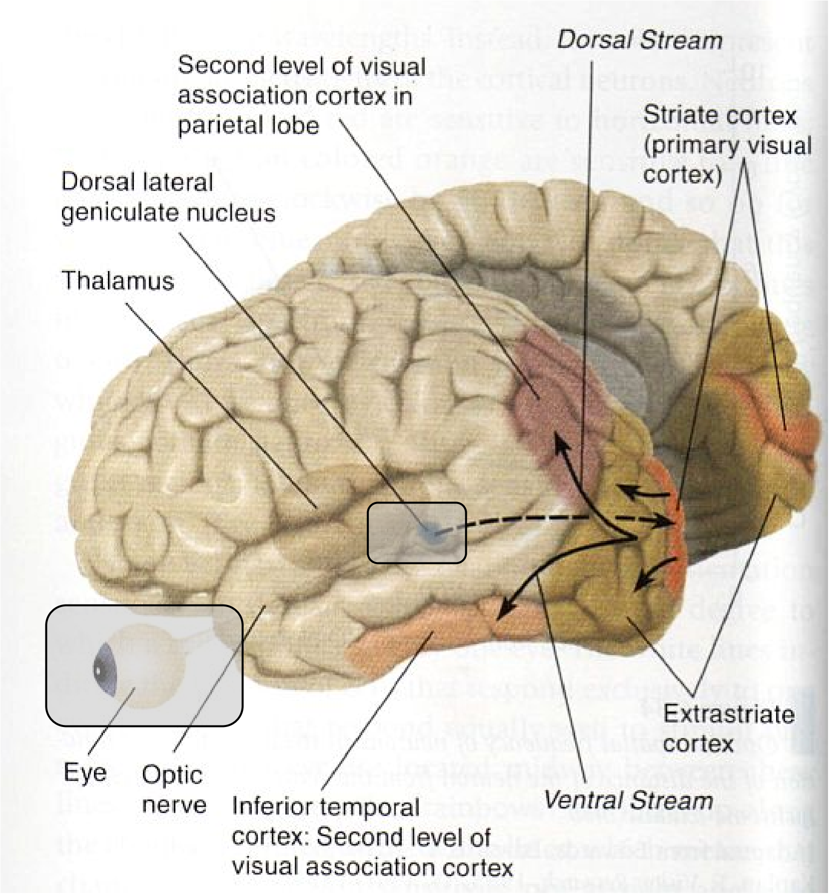


- **Retinotopic/Spatiotopic**
- **Different kinds Of Invariances**
  - Cue Invariance
  - Size Invariance
  - Position Invariance
  - Occlusion Invariance

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			

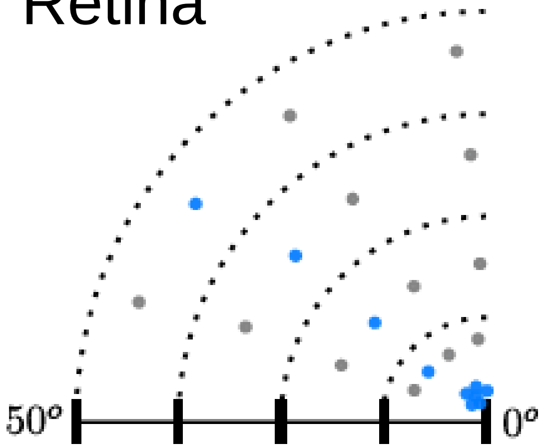
Pre-cortical Areas



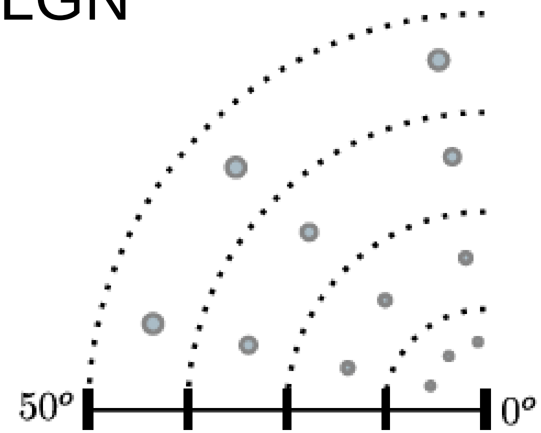


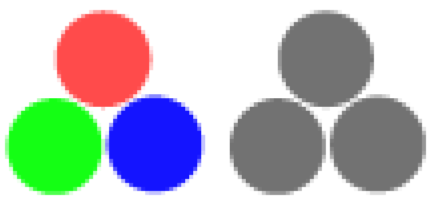
Precortical Areas

Retina

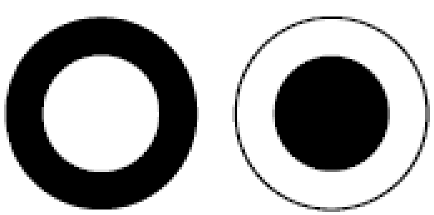


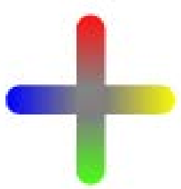
LGN

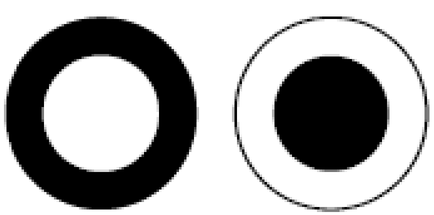




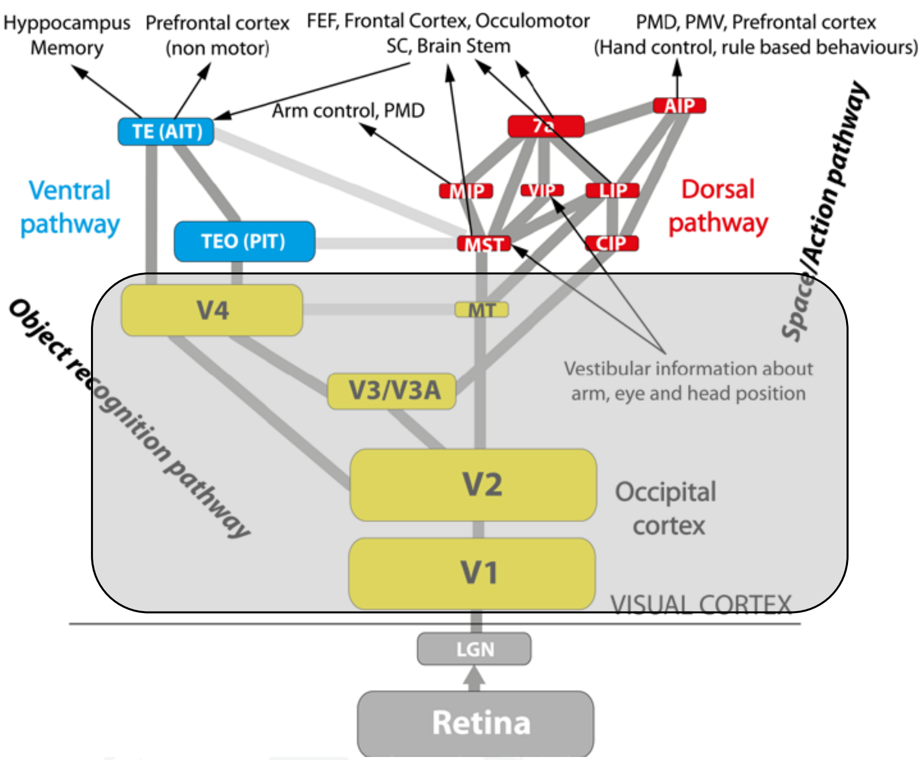
- No Feature Transformation
- Preparing for Stereo

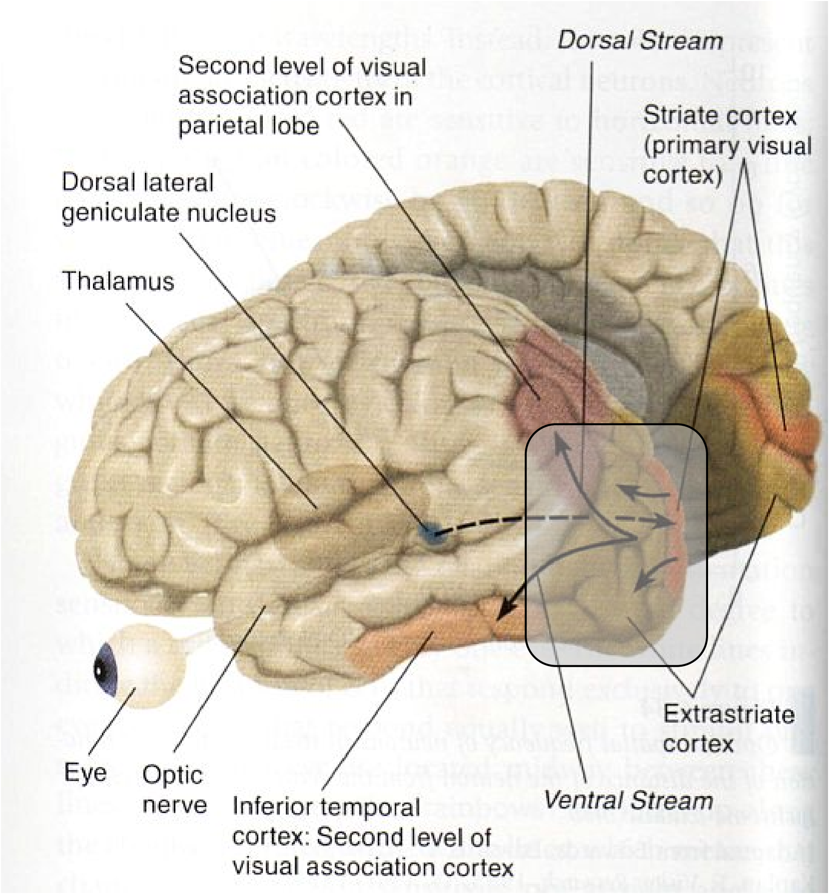






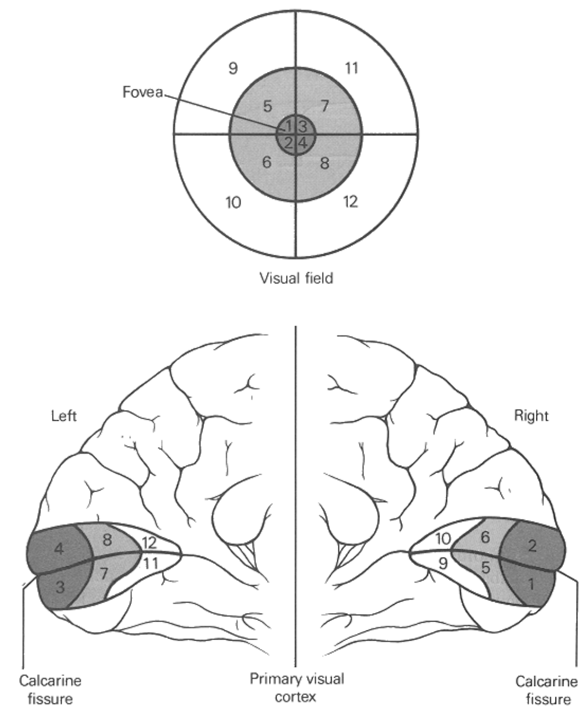
Occipital Cortex



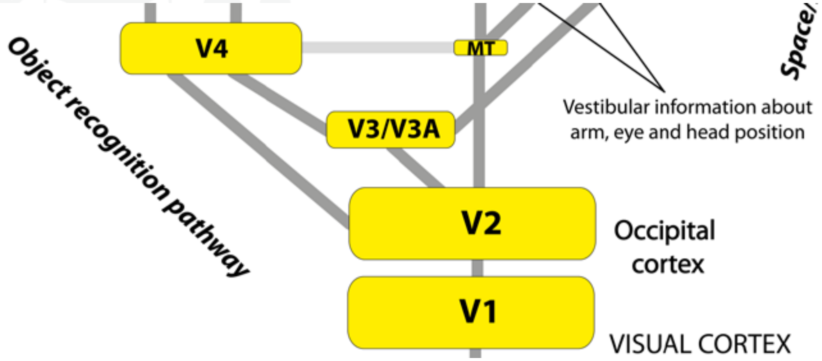


Occipital Cortex

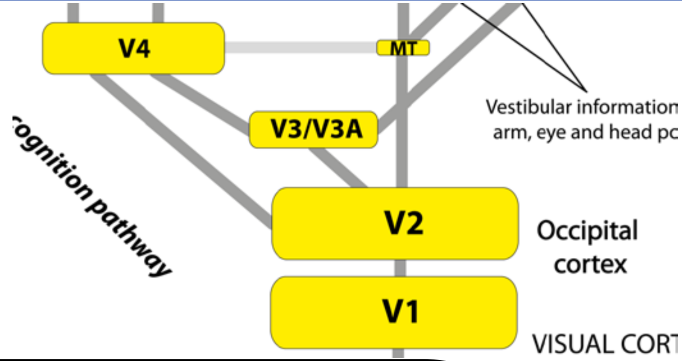
- **More than 70% of the visual cortex**
  - Occipital Cortex 3340mm^2
  - Ventral Pathway 770mm^2
  - Dorsal Pathway 585mm^2
- **Processing**
  - Task unspecific generic scene representation

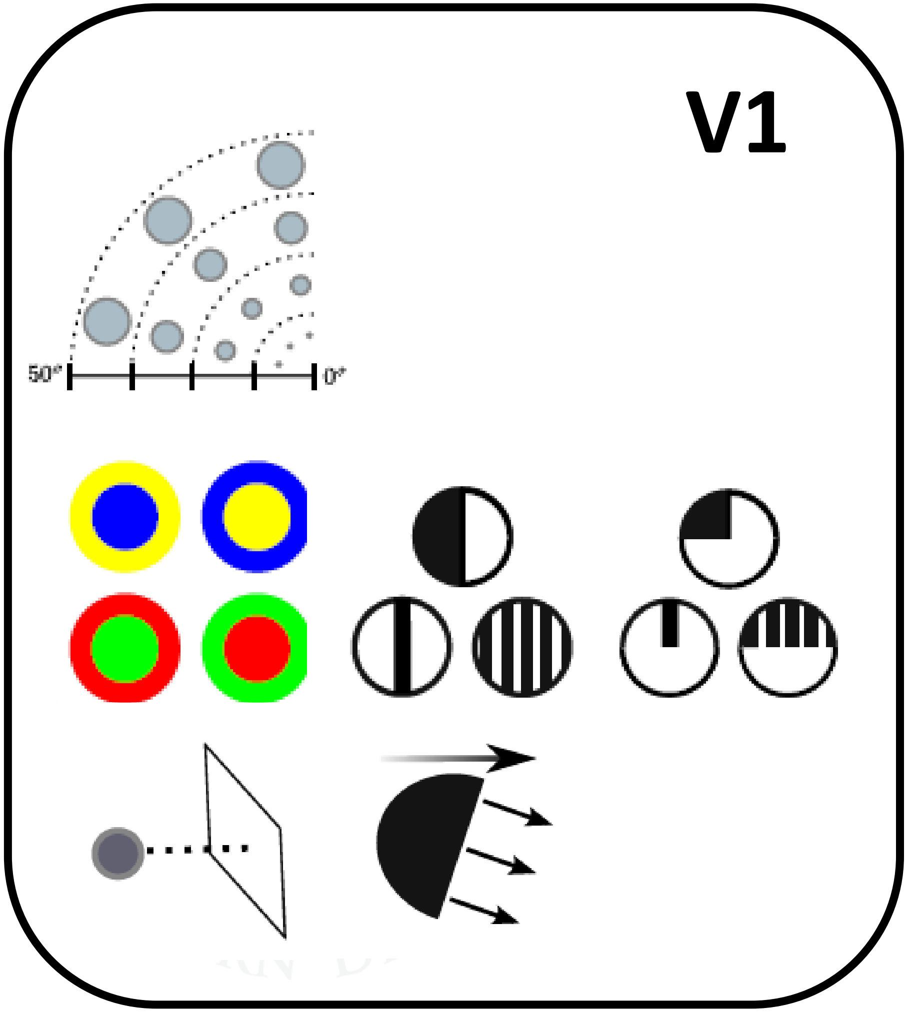


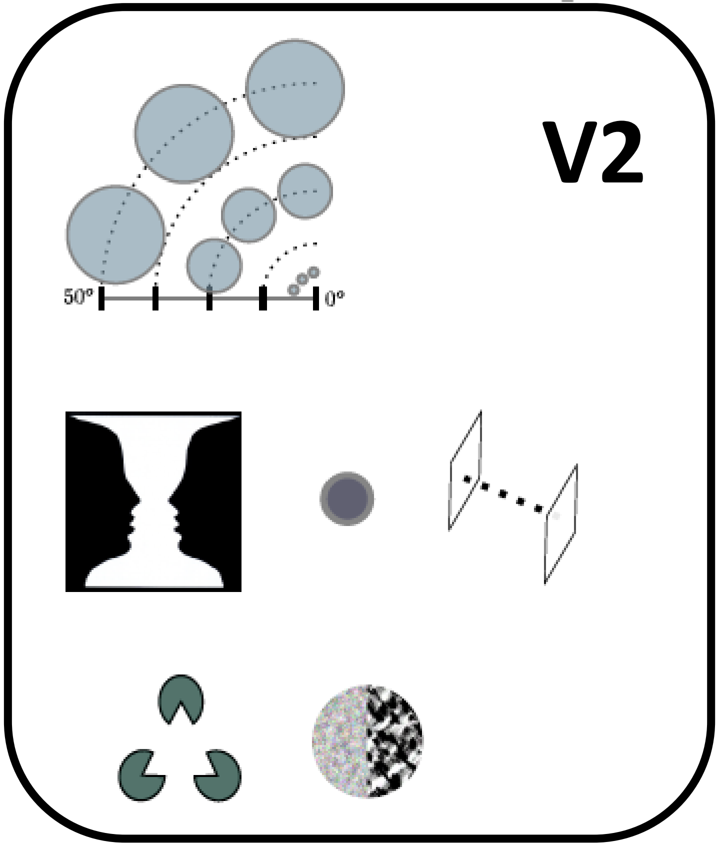
Retinotopic Organization

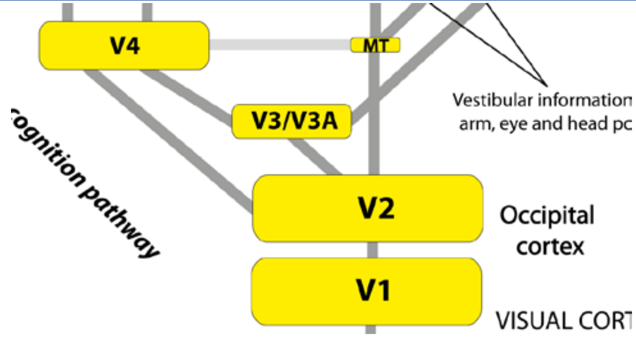


Occipital Cortex: V1 and V2



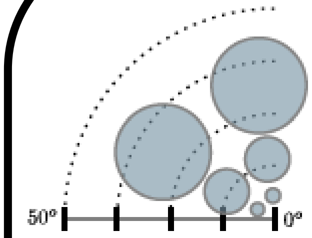






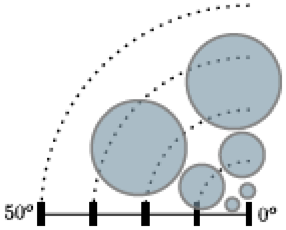
V4 and MT

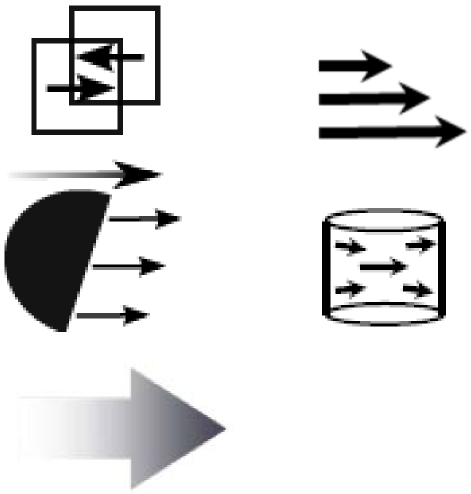
V4



Concept of Hue as Object Property
Linguistic Concept of 'red' or 'blue'

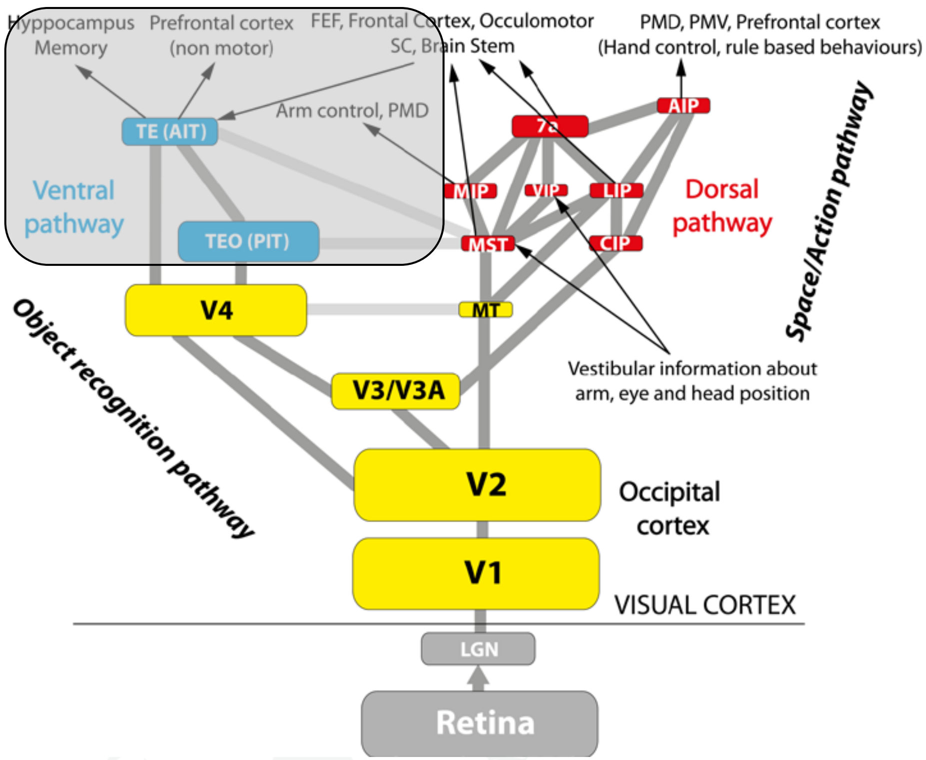
MT

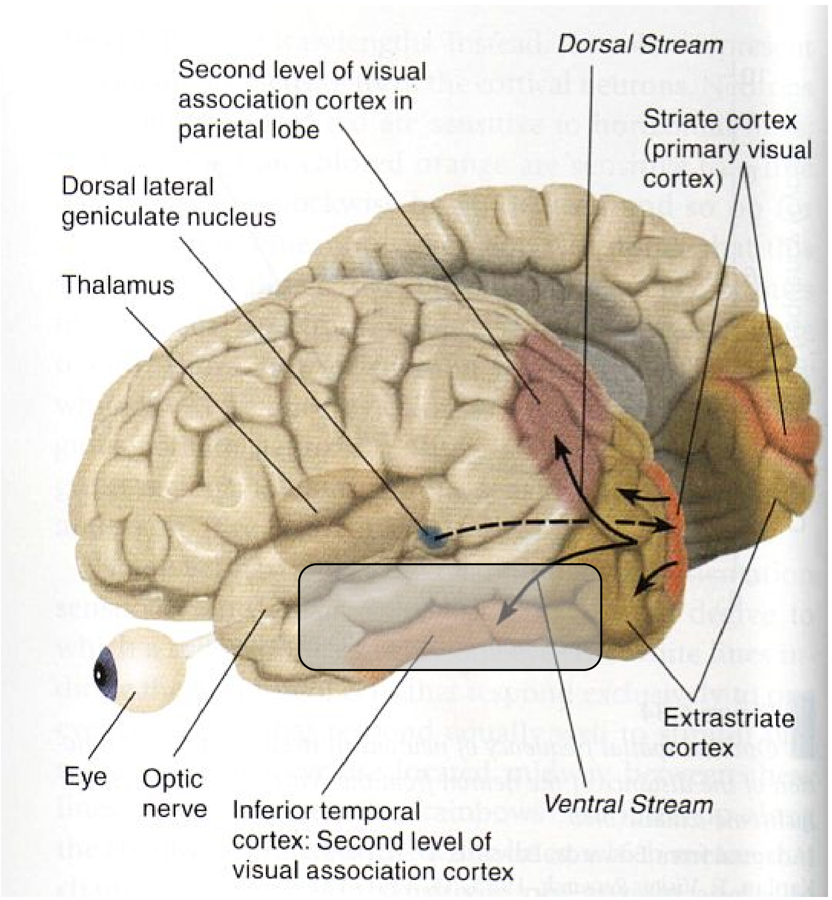




2D Motion 3D Motion

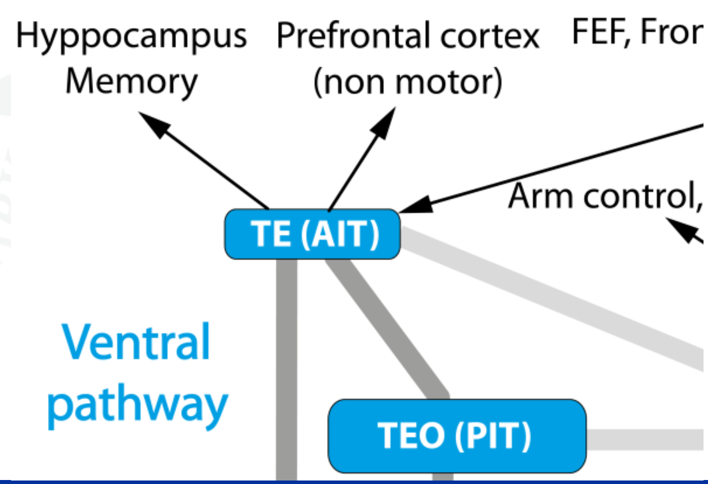
Ventral Pathway



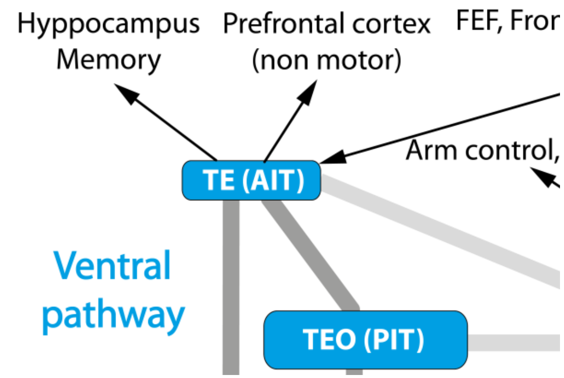


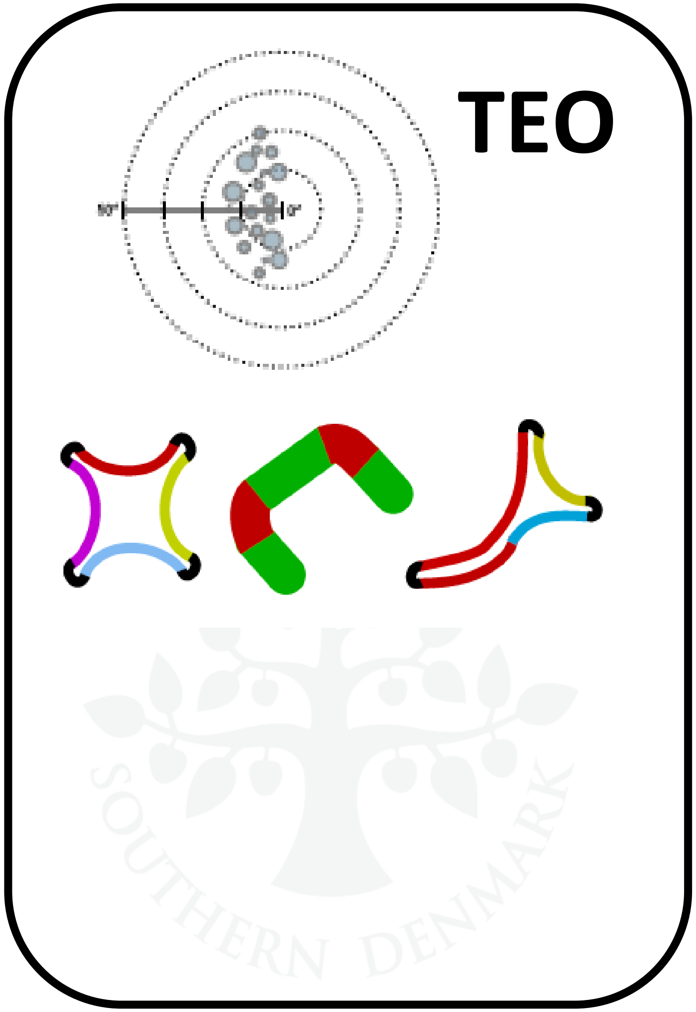
Ventral Pathway

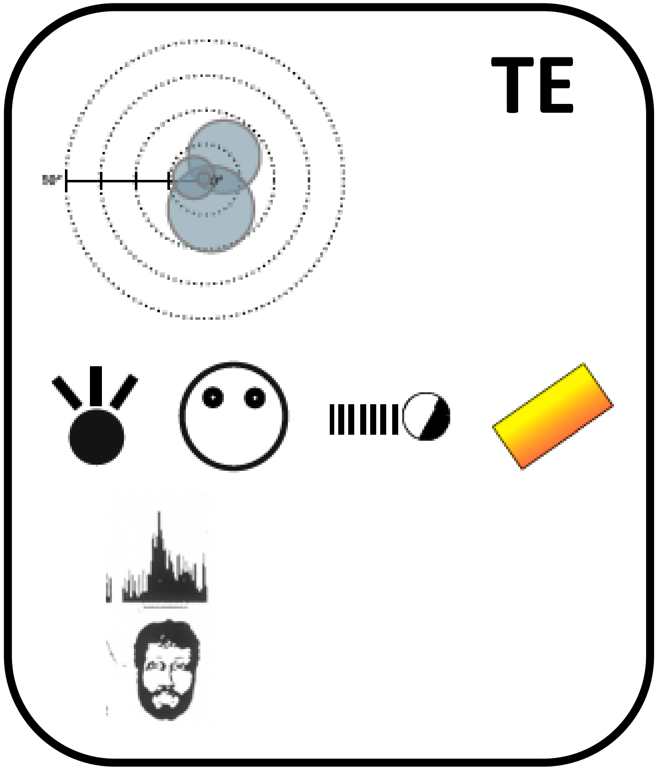
- **More than 70% of the visual cortex**
  - Occipital Cortex 3340mm^2
  - Ventral Pathway 770mm^2
  - Dorsal Pathway 585mm^2
- **Processing**
  - Object Recognition and Categorization
  - Many suggestions for how to divide into areas

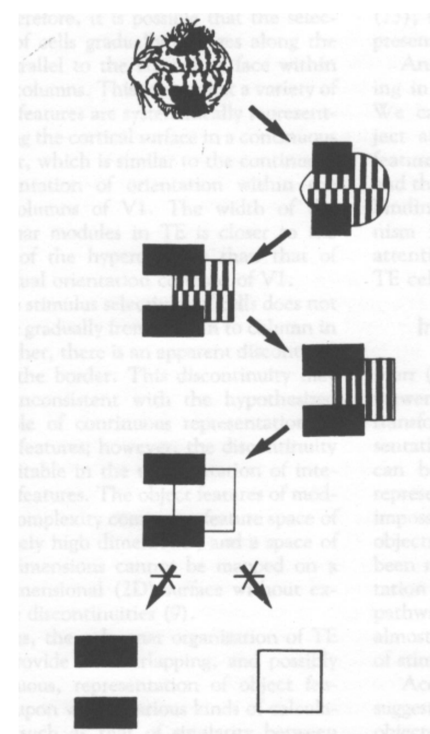


Ventral Pathway: TEO and TE



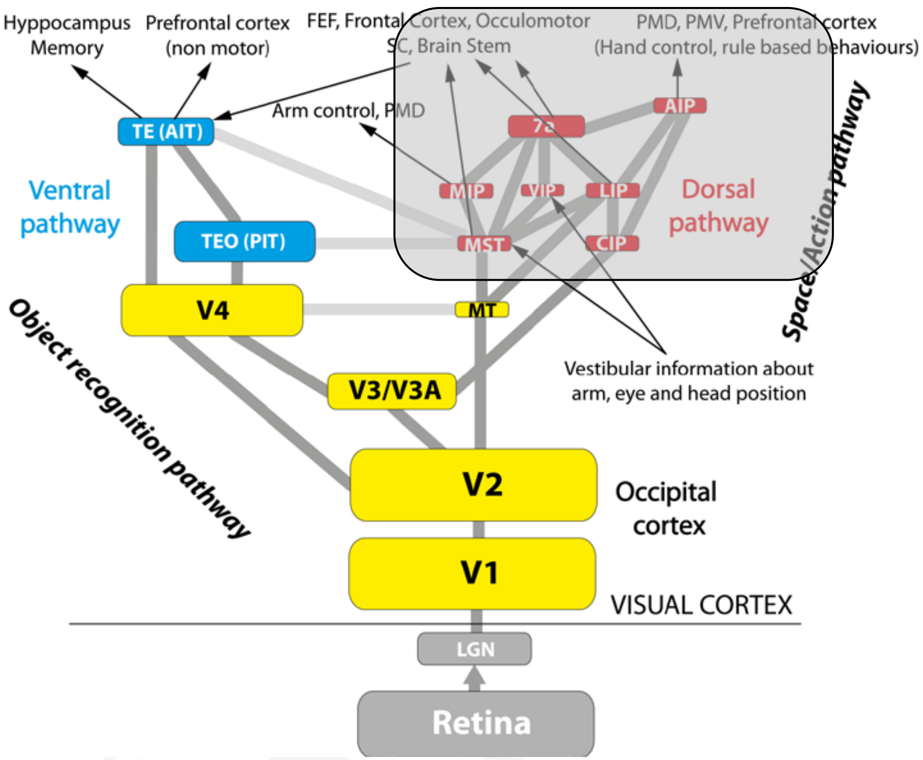


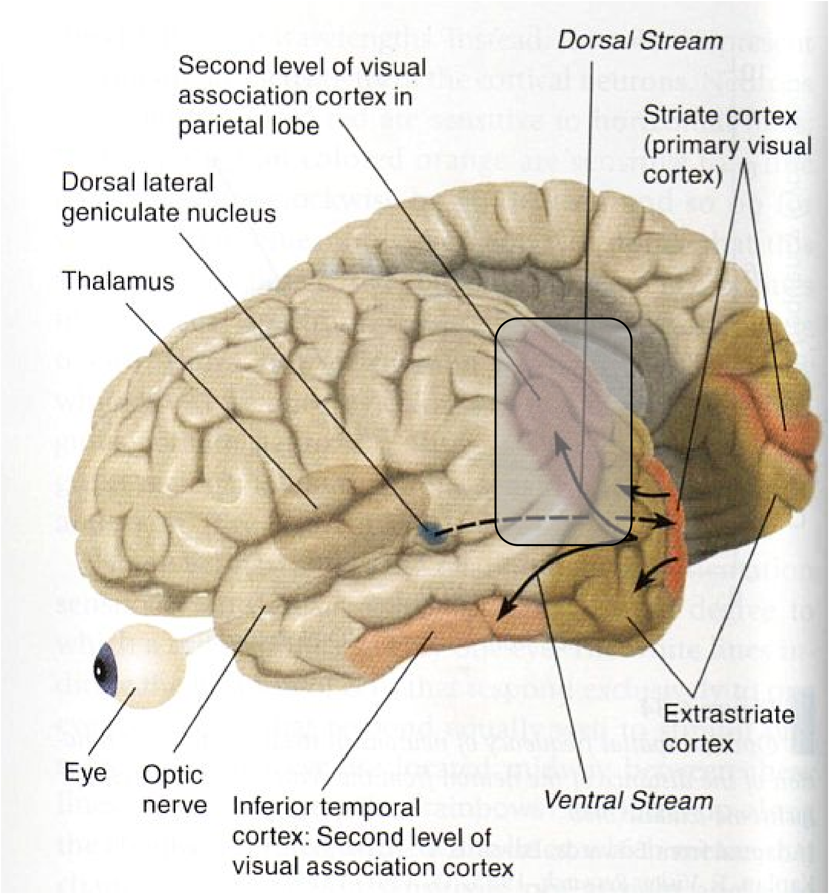




Tanaka

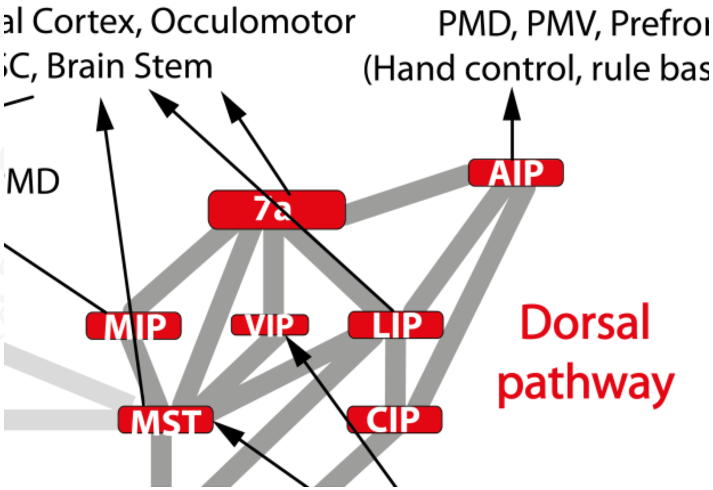
Dorsal Pathway



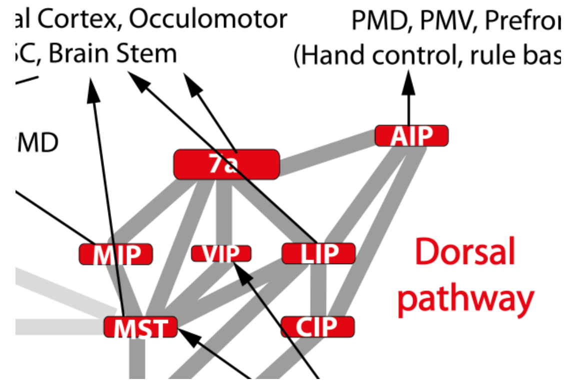


Dorsal Pathway

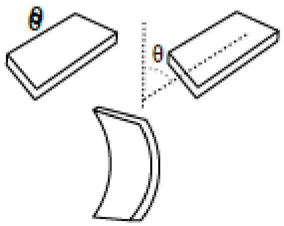
- **More than 70% of the visual cortex**
  - Occipital Cortex 3340mm^2
  - Ventral Pathway 770mm^2
  - Dorsal Pathway 585mm^2
- **Processing**
  - Much less known than Ventral Pathway
  - Many more distinguished areas
  - Coding visual information related to action and position in space



Dorsal Pathway

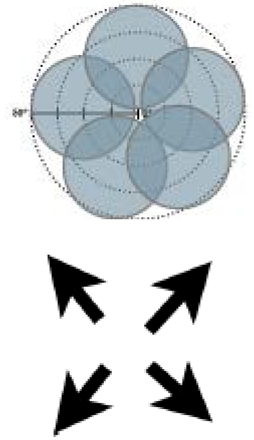


CIP



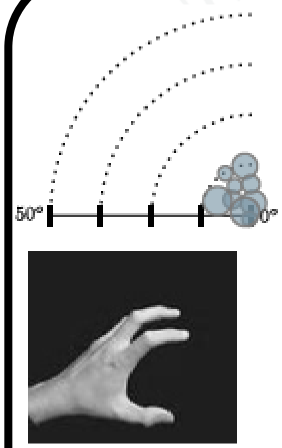
Cue invariant 3D shape

MST



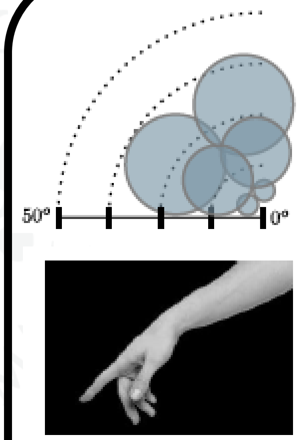
Ego-motion

AIP



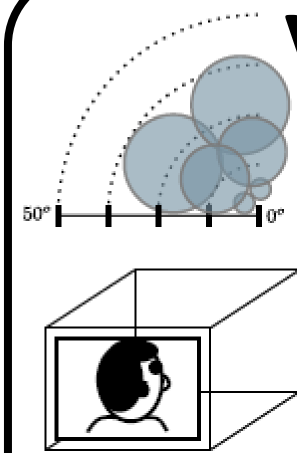
Hand shape and affordances

MIP



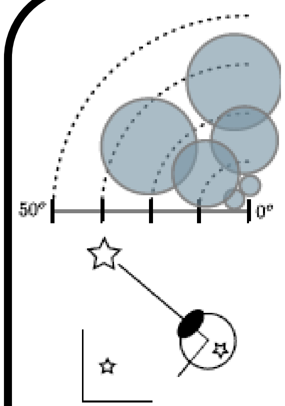
Reaching

VIP



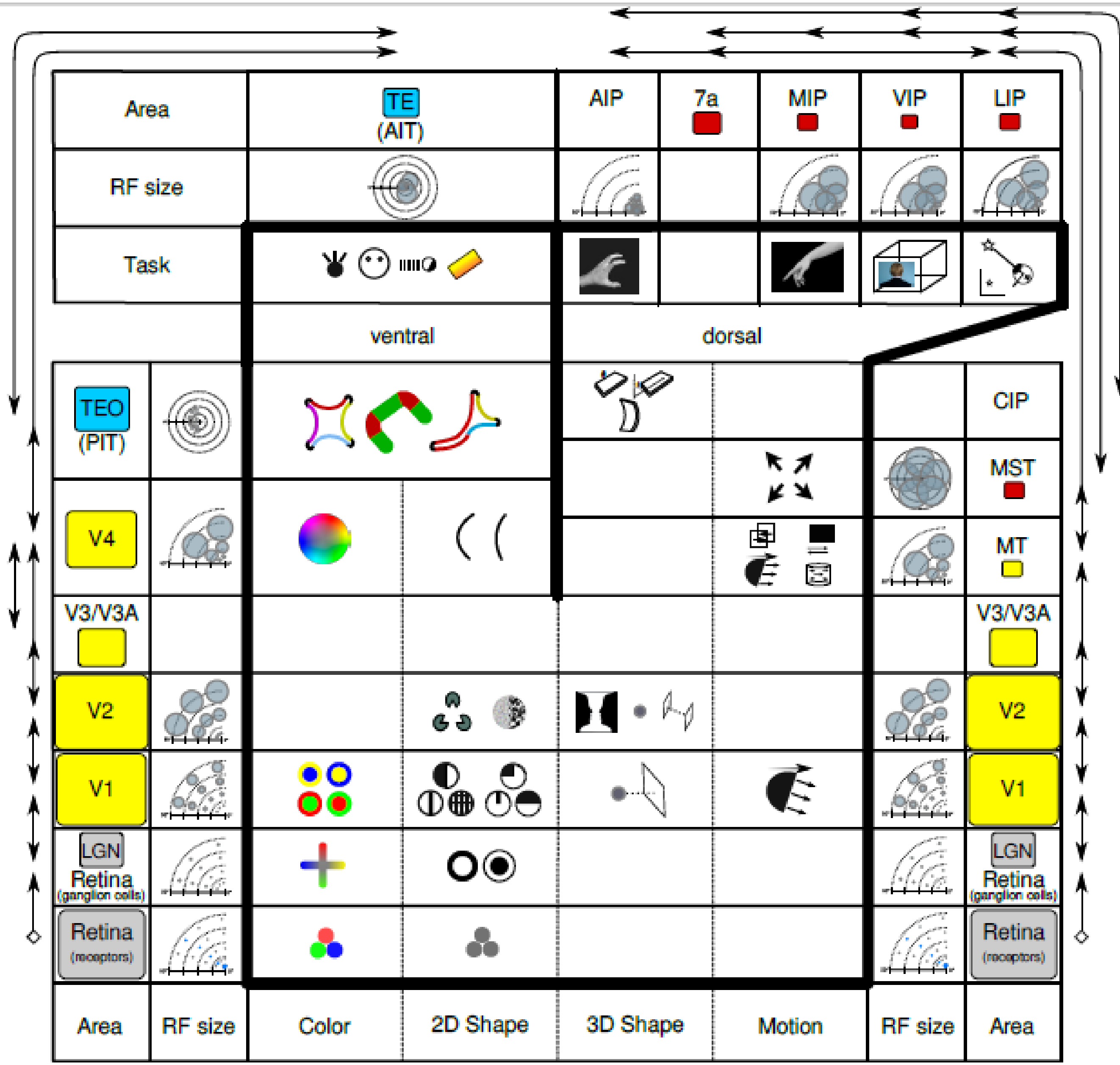
Ego-space

LIP



Saccadic related retinotopic repr.

Vertical View

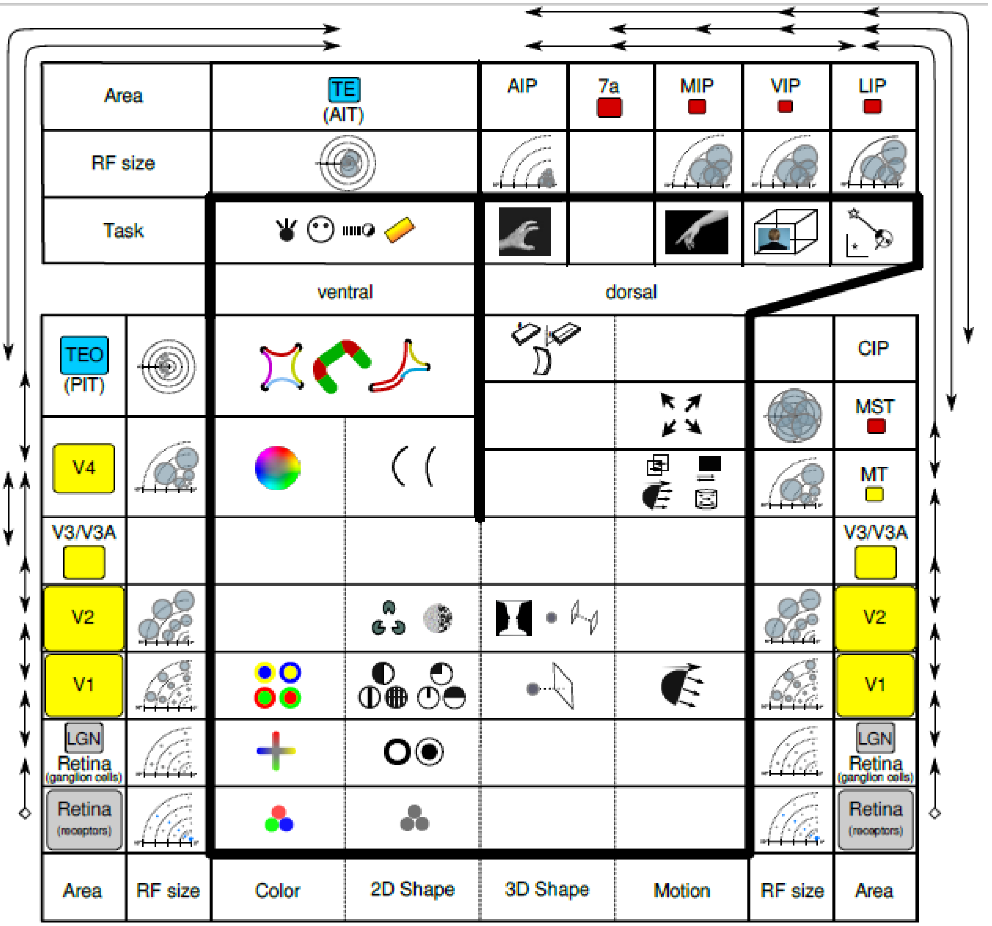


What do we know about primate's vision which is relevant for engineers?

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

Richness of representation

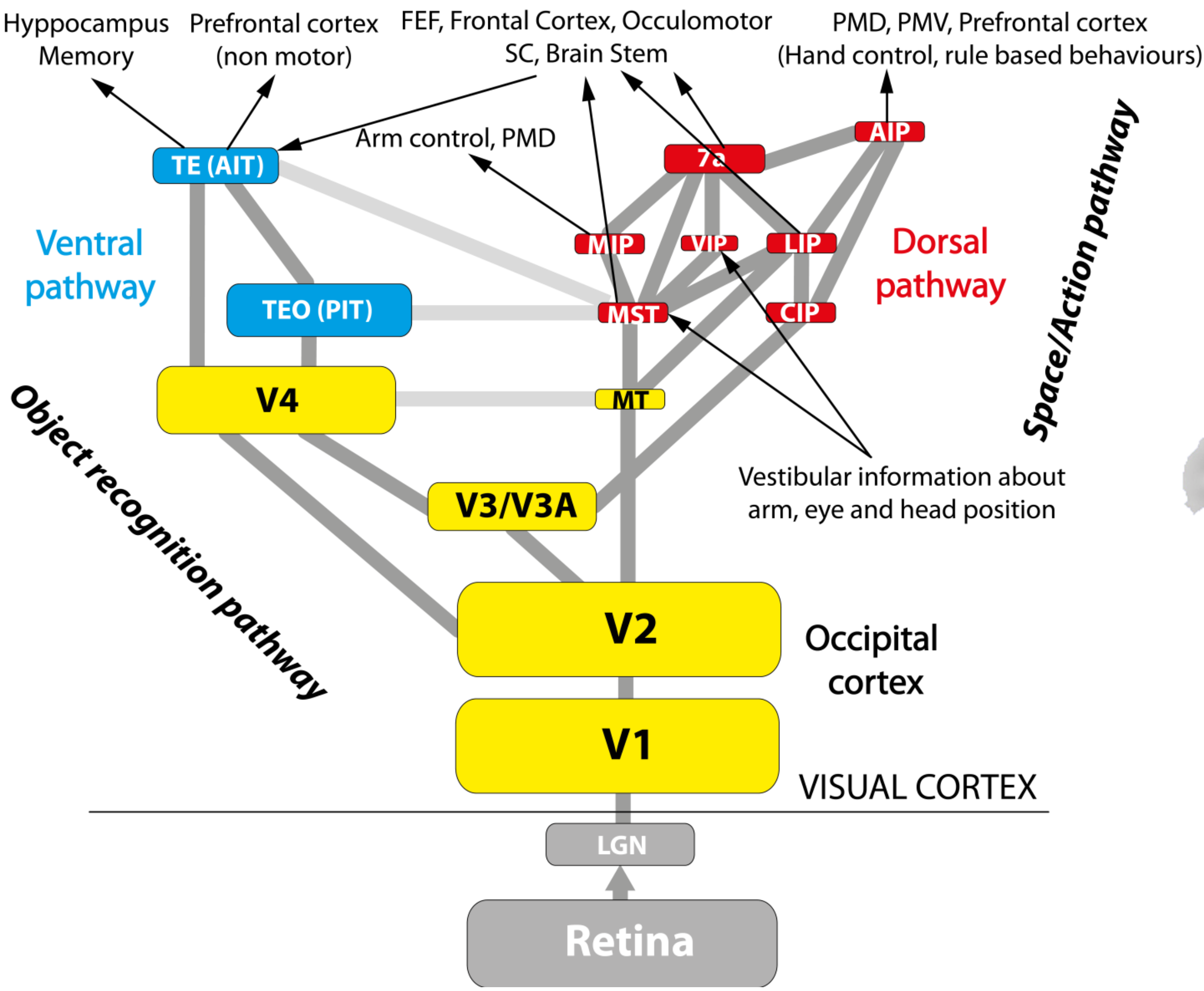
- **The occipital cortex provides a huge variety of visual aspects at different levels of granularity and different levels of abstractions**
  - Zoo of features
  - Challenge: Designing/learning this hierarchy is difficult but maybe required
- **What is important for learning a certain task or category is unclear**
  - Challenge: Learning algorithms that are able to deal with such a huge and at the same time highly structured input space

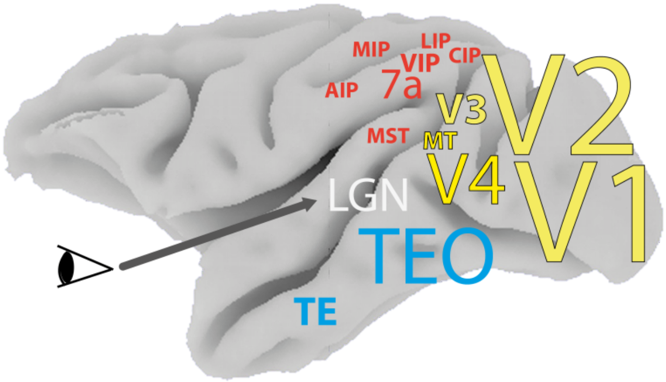


What do we know about primate's vision which is relevant for engineers and linguists?

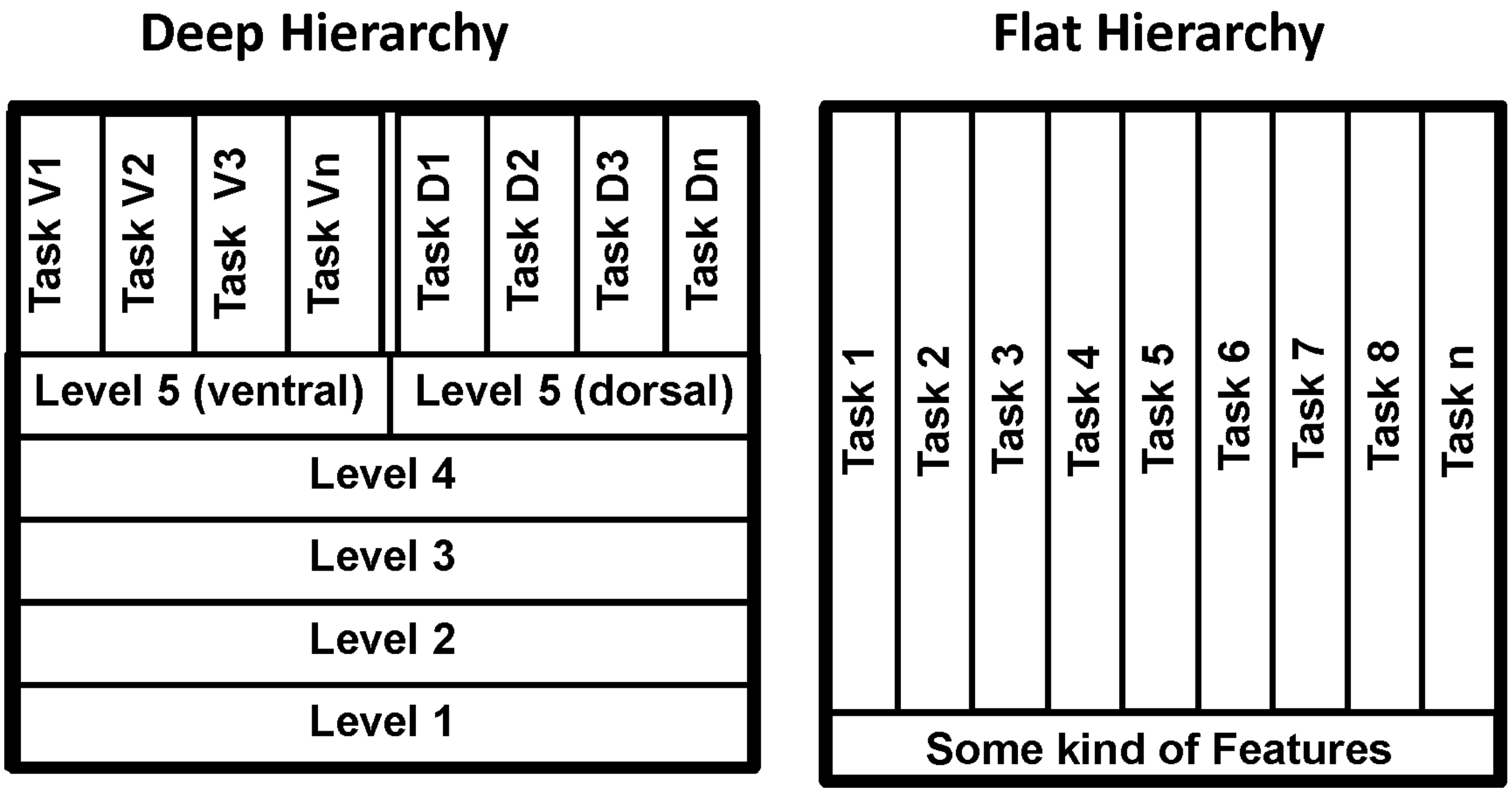
- Richness of representation
- **Deep Hierarchy versus flat Architectures**
- Separation of information

Deep Hierarchary





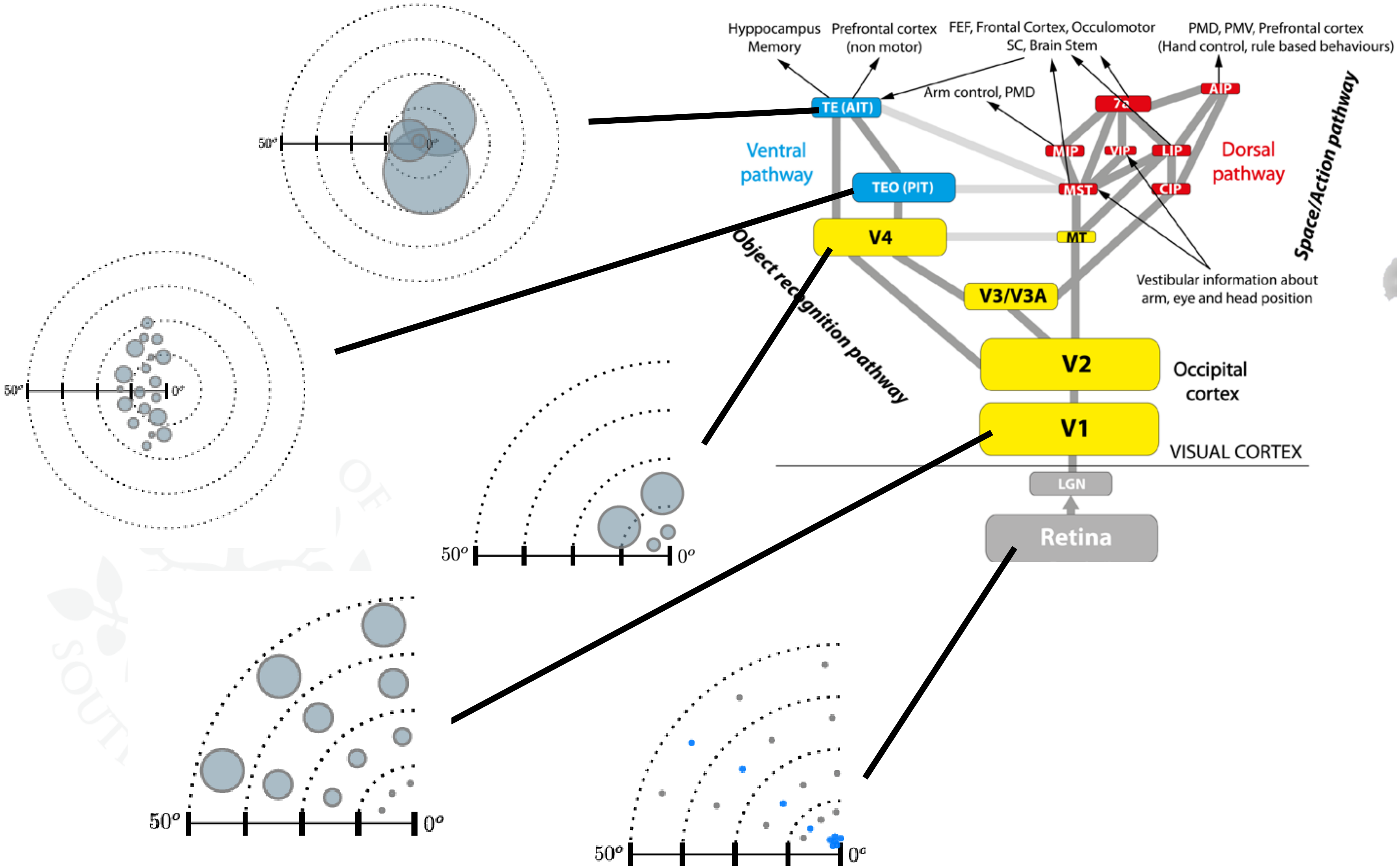
Flat versus deep Hierarchies



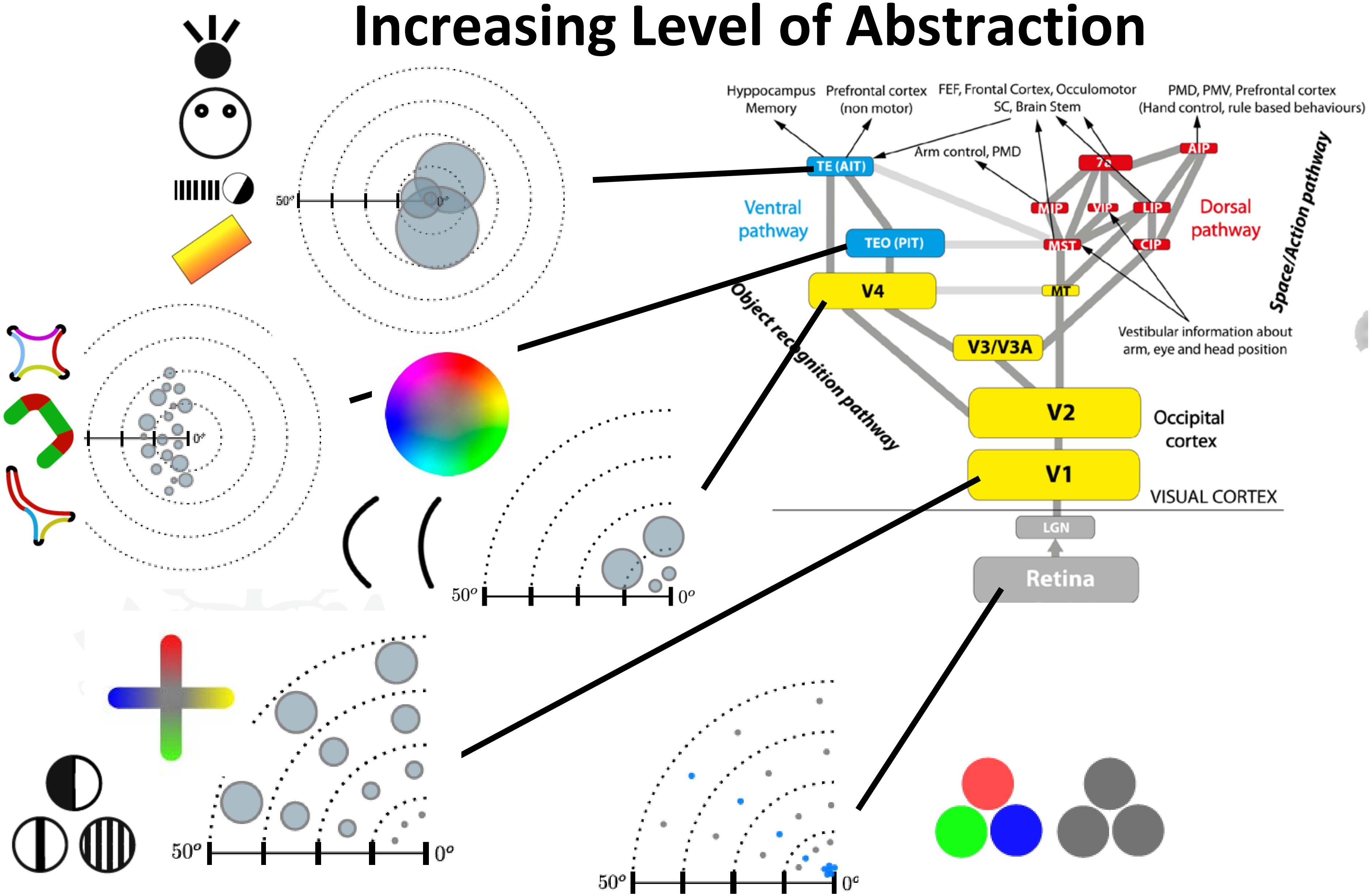
Example of a flat hierarchy

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

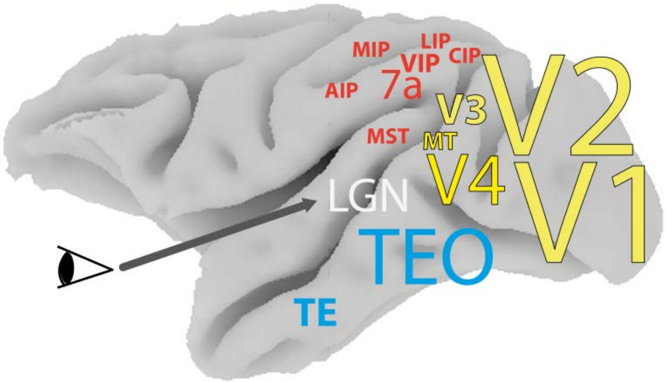


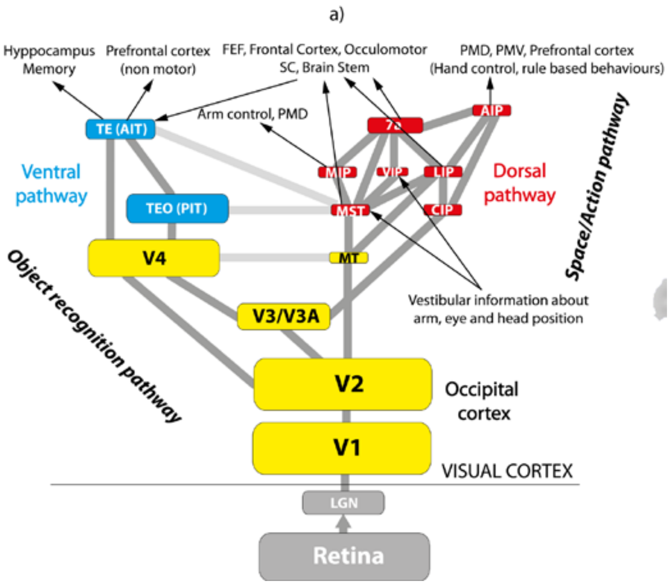
Increasing Level of Abstraction



Flat versus deep hierarchies

- **Flat Hiererachies are inefficient**
  - **No sharing of computational recources**
  - **Transfer of experience across tasks is facilitated within the same representations**

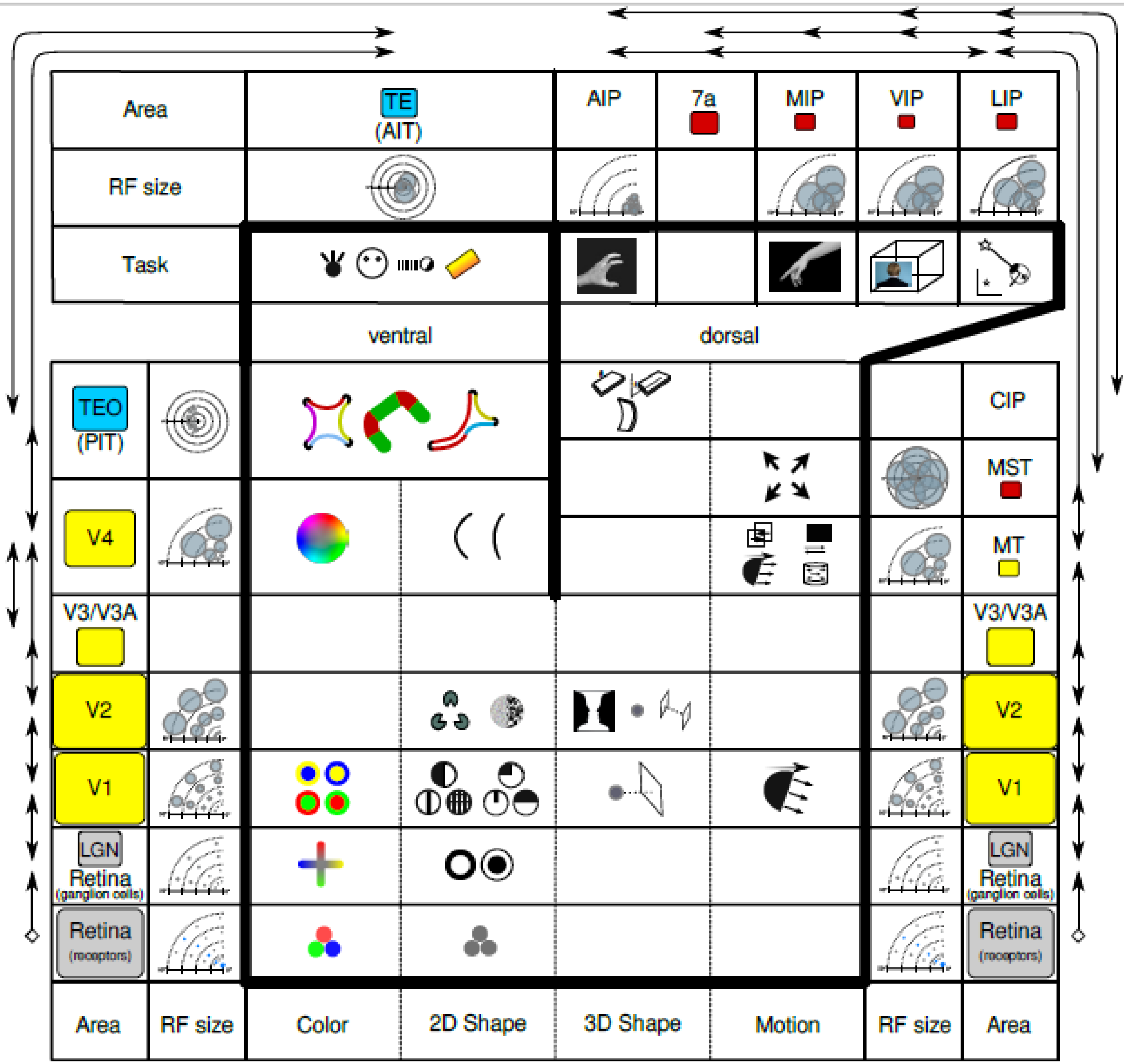




UNIVERSITY OF SOUTHERN DENMARK

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
V3/V3A
V2
V2
V1
V1
LGN
Retina
(ganglion cells)
LGN
Retina
(ganglion cells)
Retina
(receptors)
Retina
(receptors)
Area
RF size
Color
2D Shape
3D Shape
Motion
RF size
Area

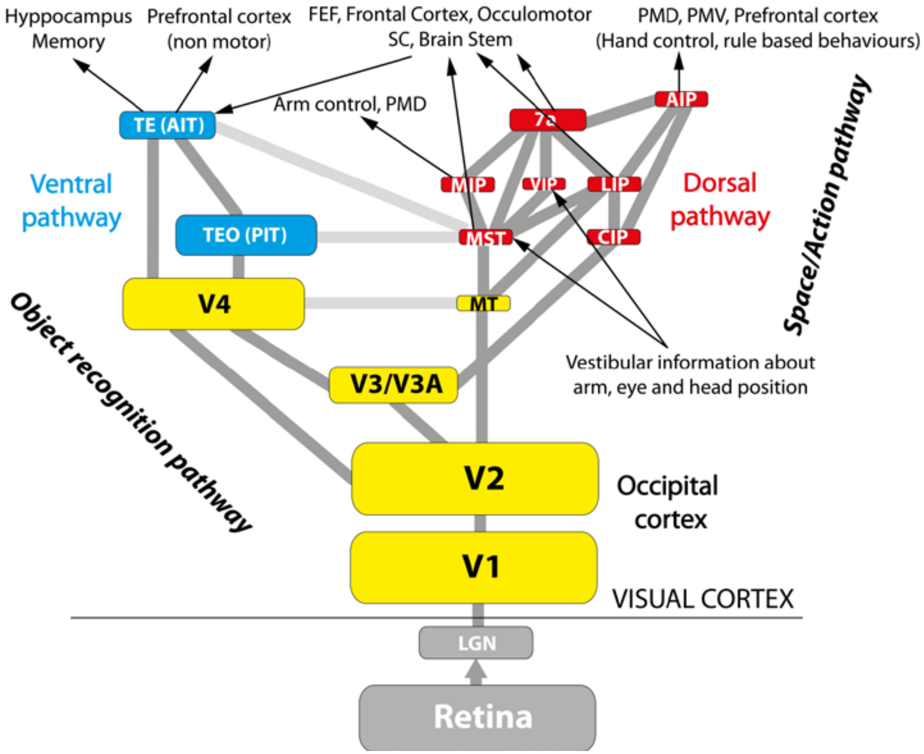
Separation of Information

- 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)

UNIVERSITY OF SOUTHERN DENMARK

Overview

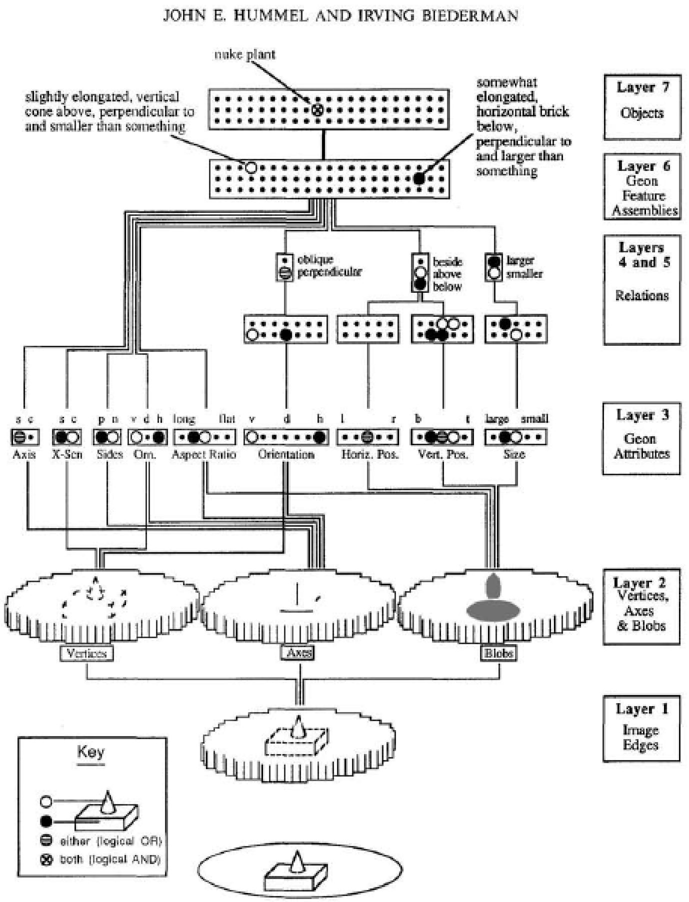
- Background Information
- The primate's vision system: A deep Hierarchy
- **SotA and Problems of research on deep hierarchical systems**
- Reflections



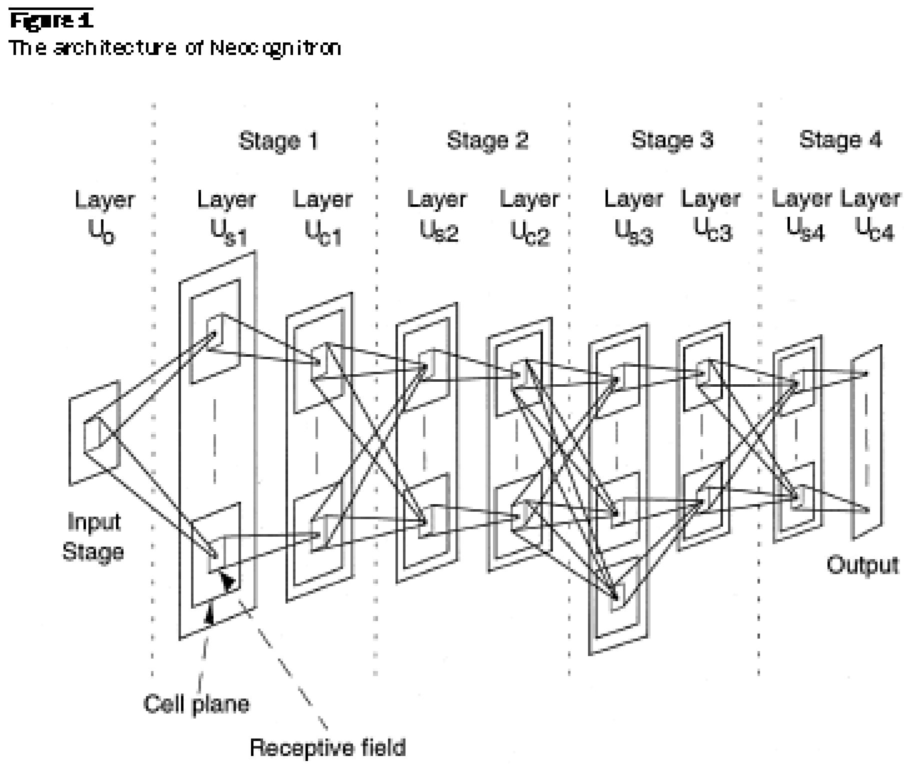
Research on Deep Hierarchies (non-exhaustive)

- **Meta reasoning**
  - Tsotsos, Geman et al. , Mel and Fiser,
- **Learning of Hierarchical Vision Systems**
  - Amit, Hawkins, Leonardis, Piater, Ullman, DiCarlo and Cox, Ommer and Buhmann , Serre and Poggio, Bengio, Wiskott, Hinton
- **Design of Hierarchical Vision Systems**
  - Biederman and Hummel, Fukushima, Pugeault and Kruger

Biederman and Fukushima

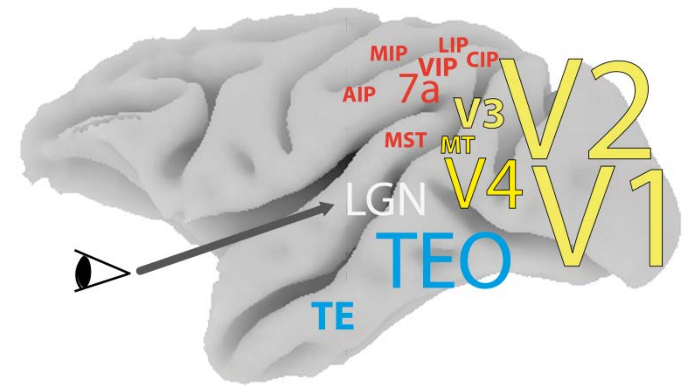


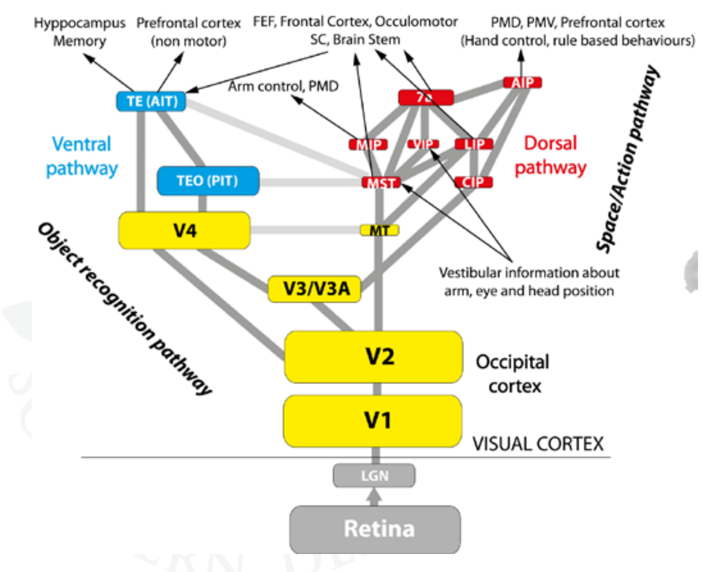
John E. Hummel and Irving Biederman (1992). Dynamic Binding in a Neural Network for Shape Recognition

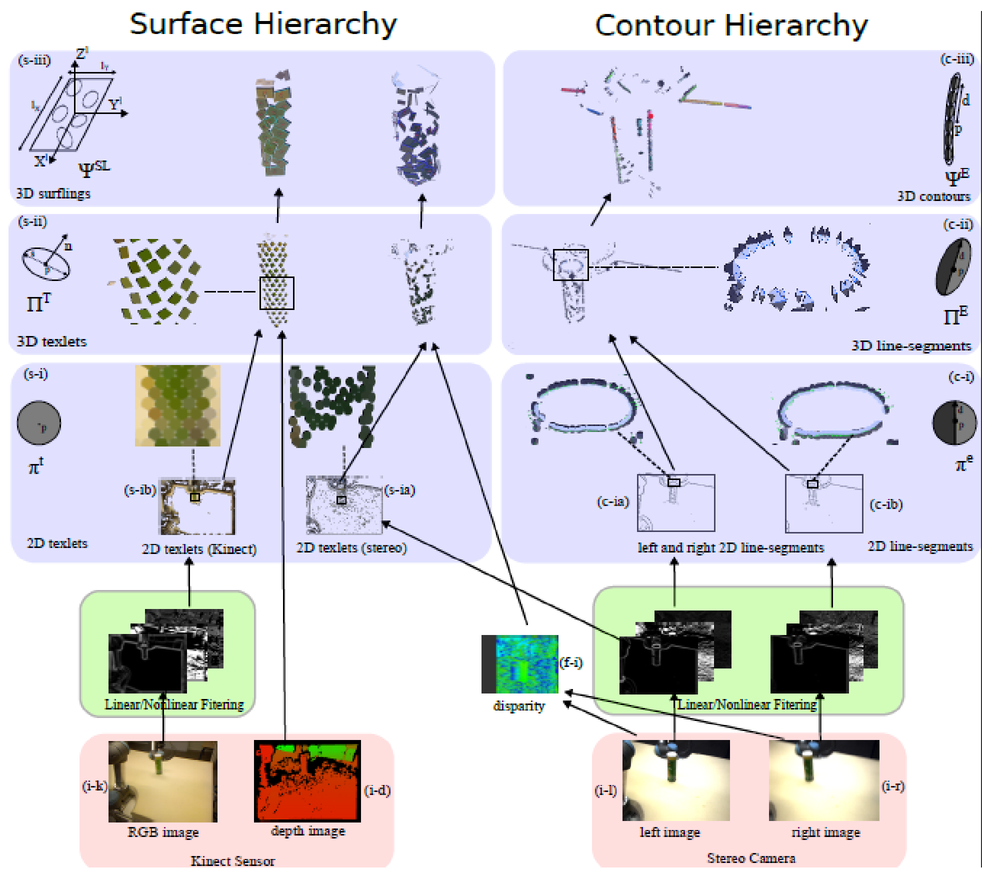


Kunihiko Fukushima 1987

Early Cognitive Vision System

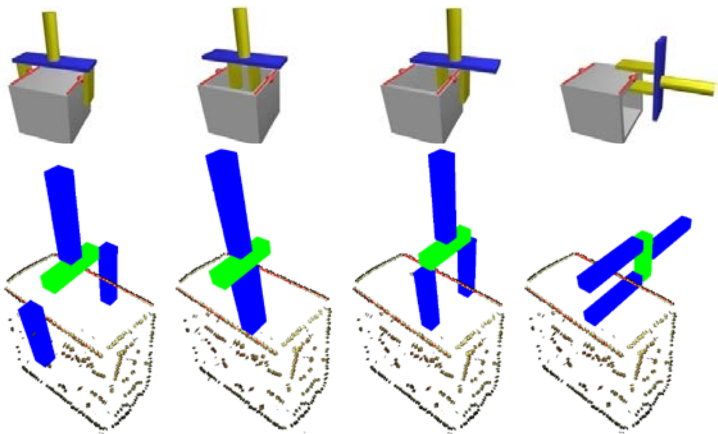




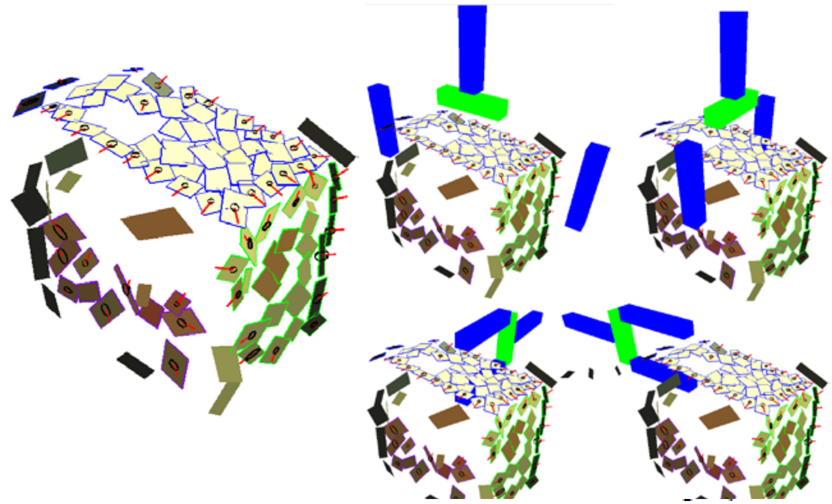


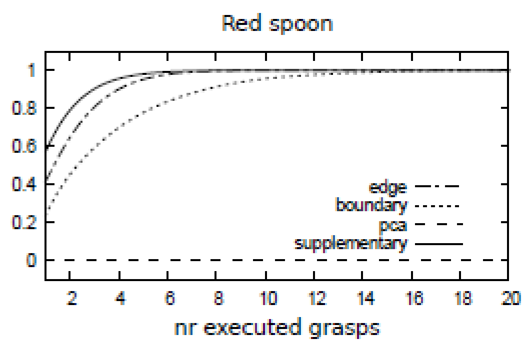
Edge and Surface based Grasp Affordances

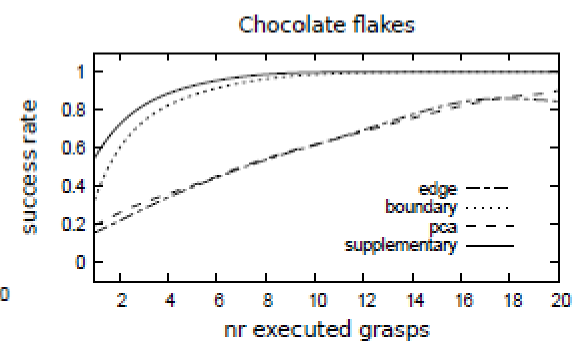
Edge based



Surface based

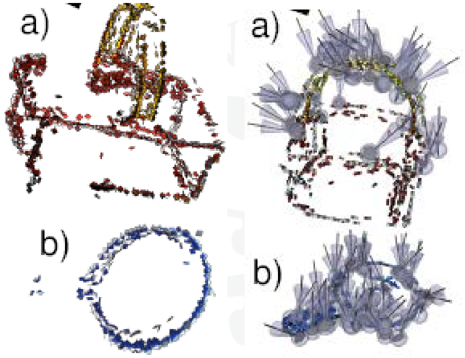


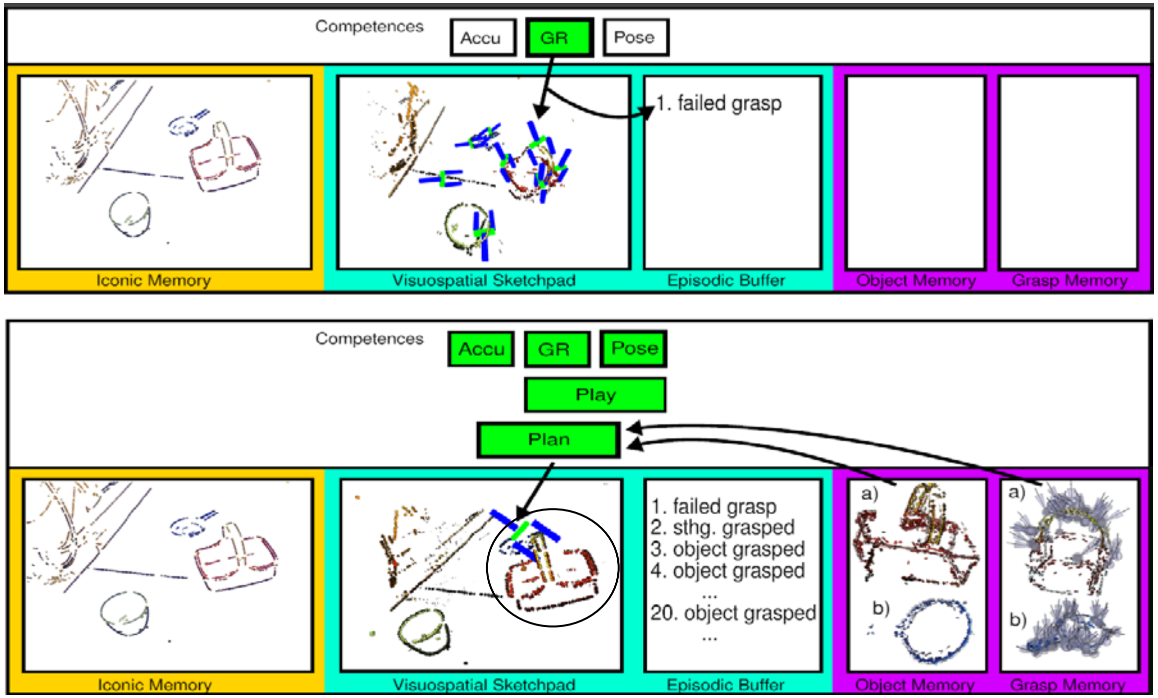




M. Popović, G. Kootstra, J. A. Jørgensen, D. Kragic and N. Krüger. Grasping Unknown Objects using an Early Cognitive Vision System for General Scene Understanding. IROS 2011 (nominated as one of the finalists for an IROS Awards)
G. Kootstra, M. Popovic, J. A. Jorgensen, K. Kuklinski, K. Miatliuk, D.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

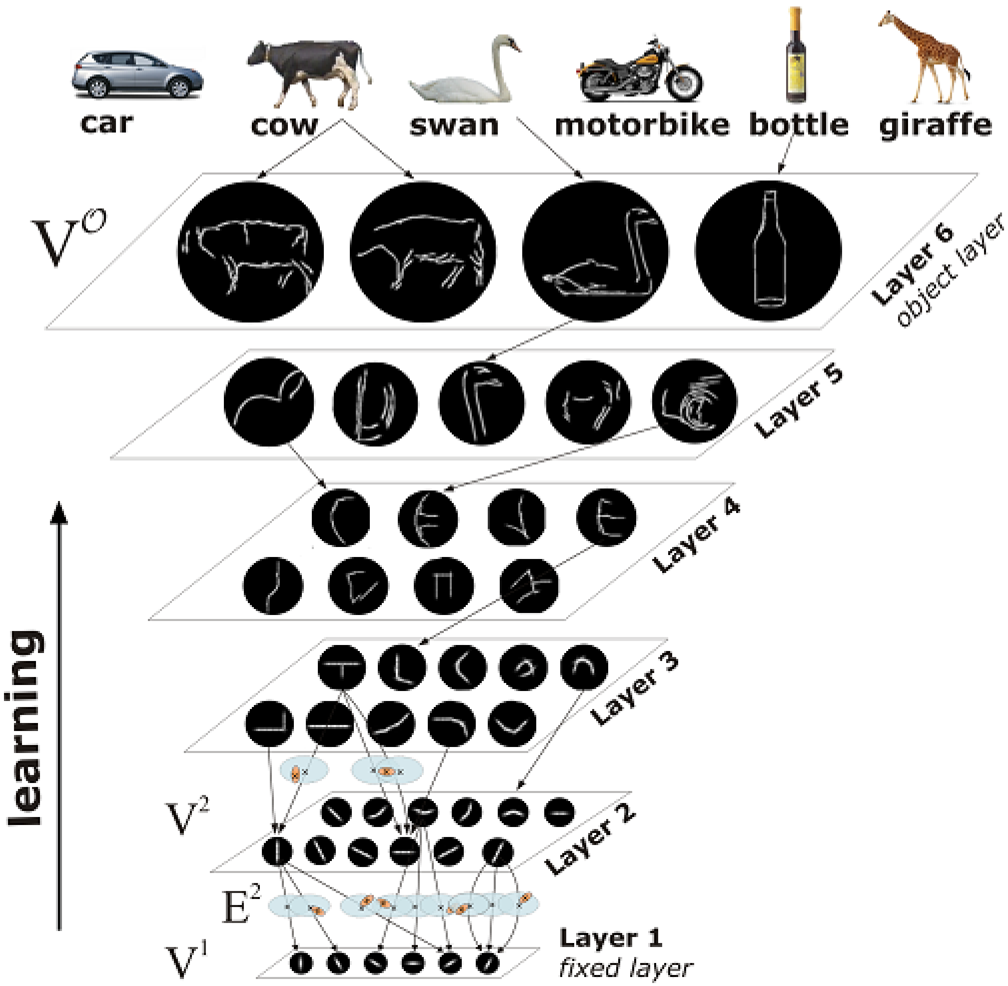
Bootstrapping Robots: Grounding objects and grasping affordances

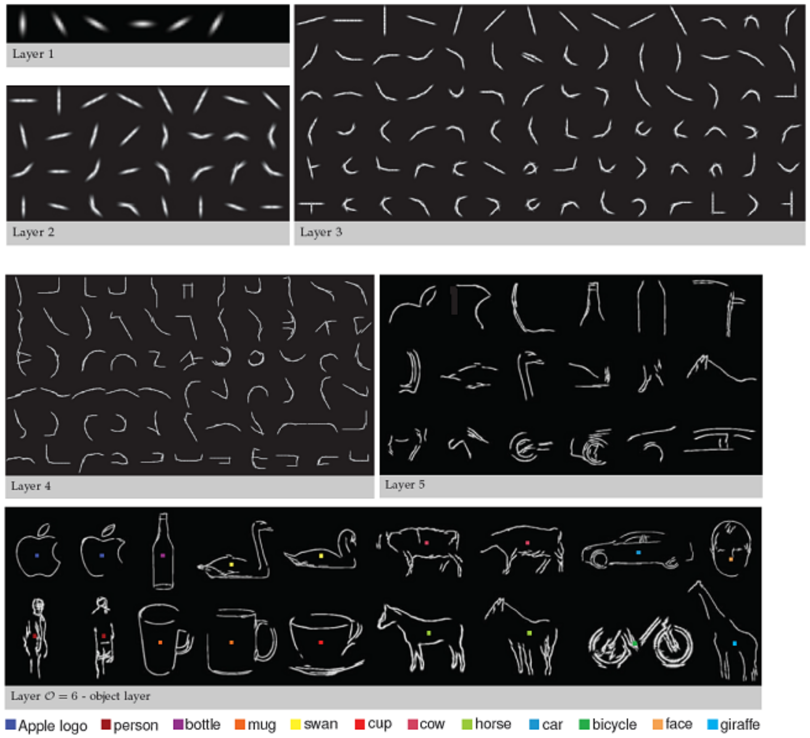




F. Guerin, D. Kraft and N. Krüger. A Survey of the Ontogeny of Tool Use: From Sensorimotor Experience to Planning. IEEE Transactions on Autonomous Mental Development, 5(1), pp. 18–45, 2013.
D. Kraft, R. Detry, N. Pugeault, E. Başeski, F. Guerin, J. Piater and N. Krüger. Development of Object and Grasping Knowledge by Robot Exploration.Autonomous Mental Development, IEEE Transactions on, vol.2, no.4, pp.368-383, Dec. 2010.

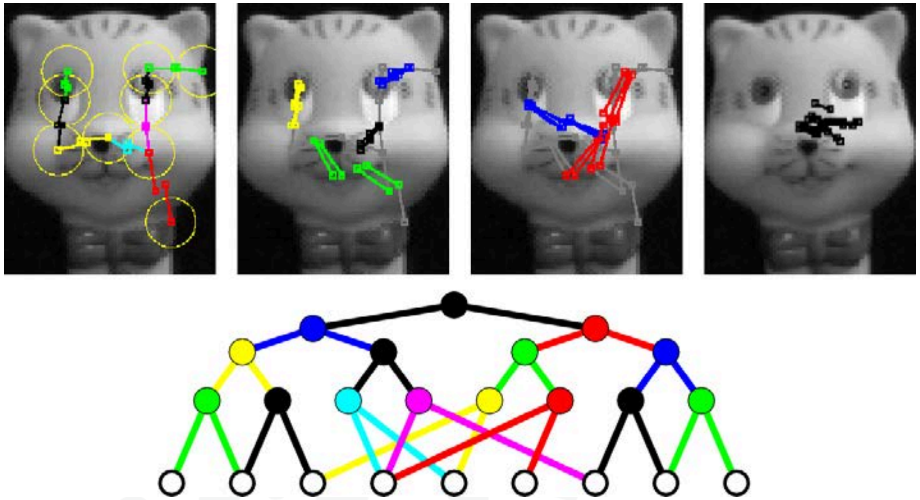
Learning Hierarchies: Work from Ales Leonardis

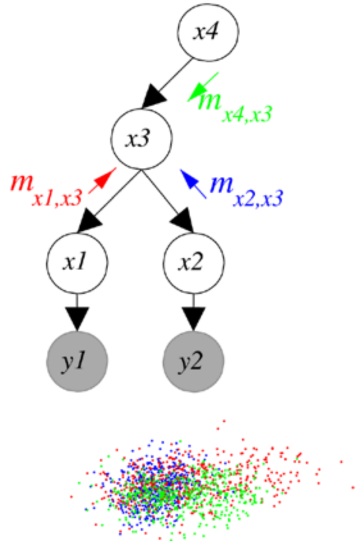


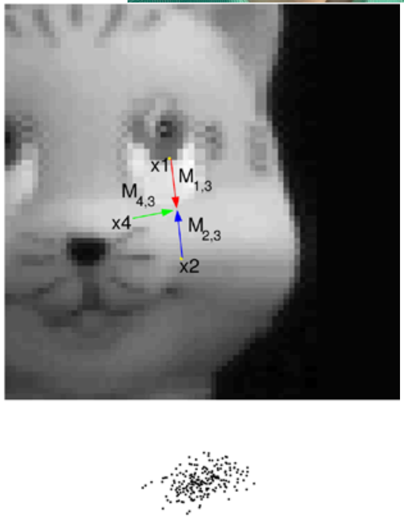


Learning Hierarchies: Work from Justus Piater

Layered Graphical Model







- Each vertex represents a (composite or primitive) feature.
- Each edge is annotated with a spatial relation (scale-normalized distance and relative orientation).

Revival of deep neural net working

- Deep Nets seem to recently beat other algorithms on important benchmarks
- Christian Szegedy et al. (2014). Intriguing properties of neural networks. ICLR 2014.
  (quotes from article of Mike James)
  - A single neuron's feature is no more interpretable as a meaningful feature than a random set of neurons.
  - Every deep neural network has "blind spots" in the sense that there are inputs that are very close to correctly classified examples that are misclassified.

Car Not A Car!

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?**
  - It is worthwhile to build/learn deep hierarchical systems
  - Number of levels
  - Receptive field size
  - What features to extract at what stage in the hierarchy