Introduction to Computer Vision

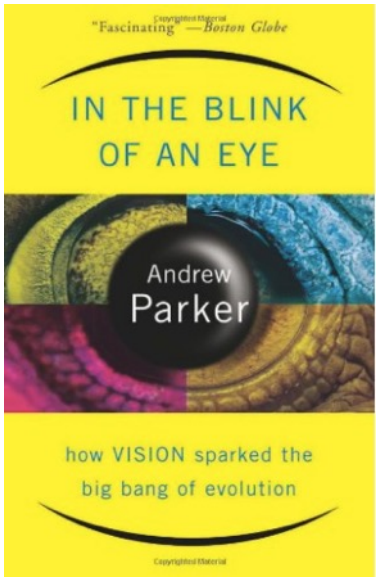
Jitendra Malik
UC Berkeley

Phylogeny of Intelligence

Cambrian Explosion
540 million years ago

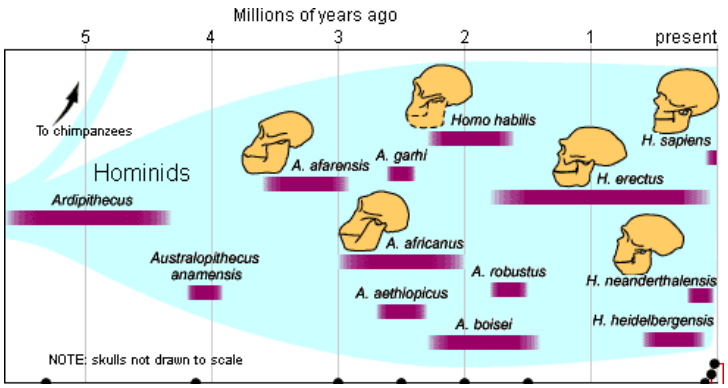
Variety of life forms,
almost all phyla emerge

Animals that could
see and move



Gibson: we see in order to move and we move in order to see

Hominid evolution, last 5 million years



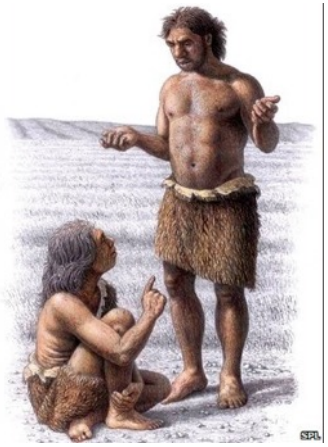
Bipedalism

Opposable thumb

Tool use

Anaxogaras: It is because of his being armed with hands that man is the most intelligent animal

Modern humans, last 50 K years



Language

Abstract thinking

Symbolic behavior

The evolutionary progression

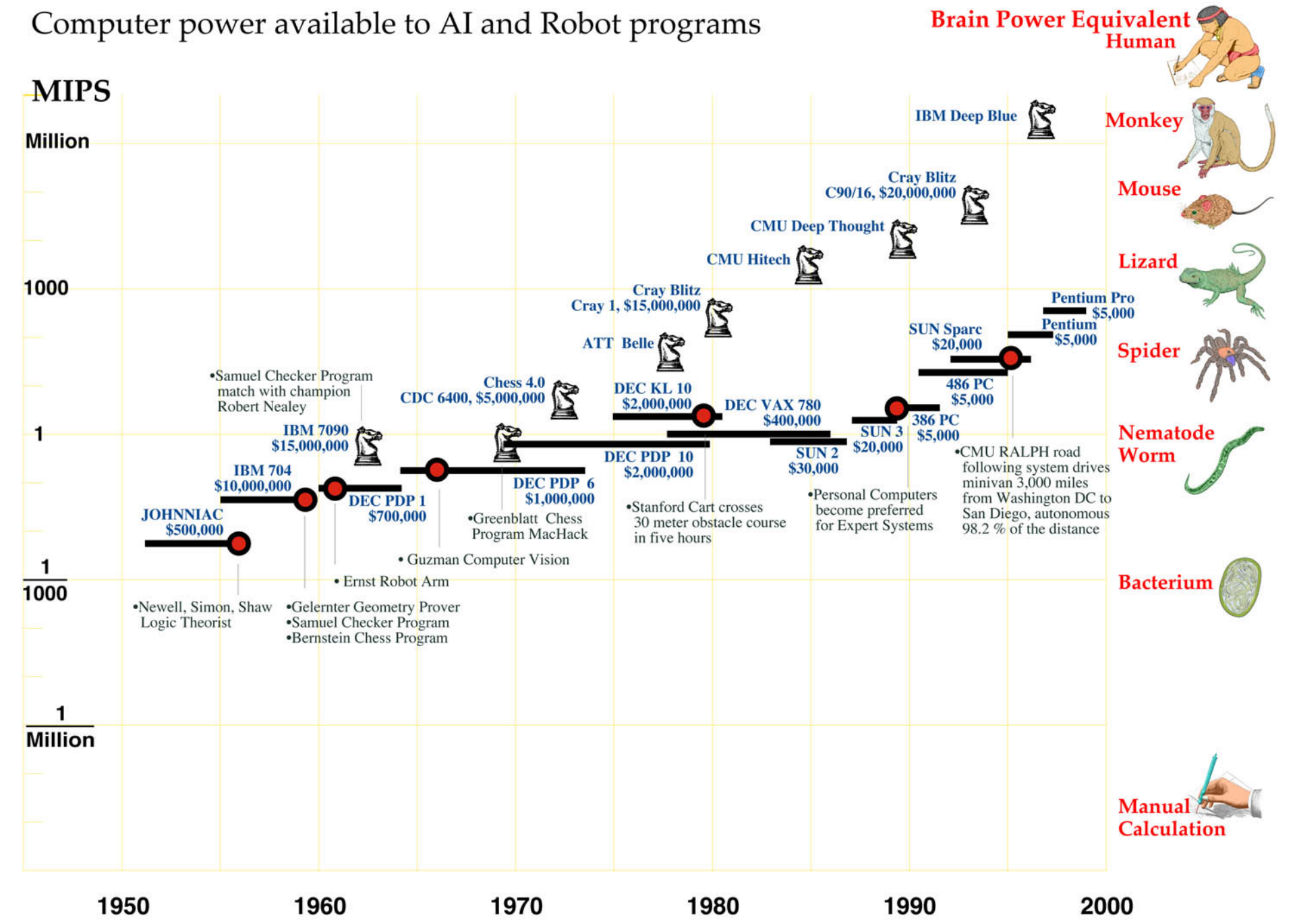
- Vision and Locomotion
- Manipulation
- Language

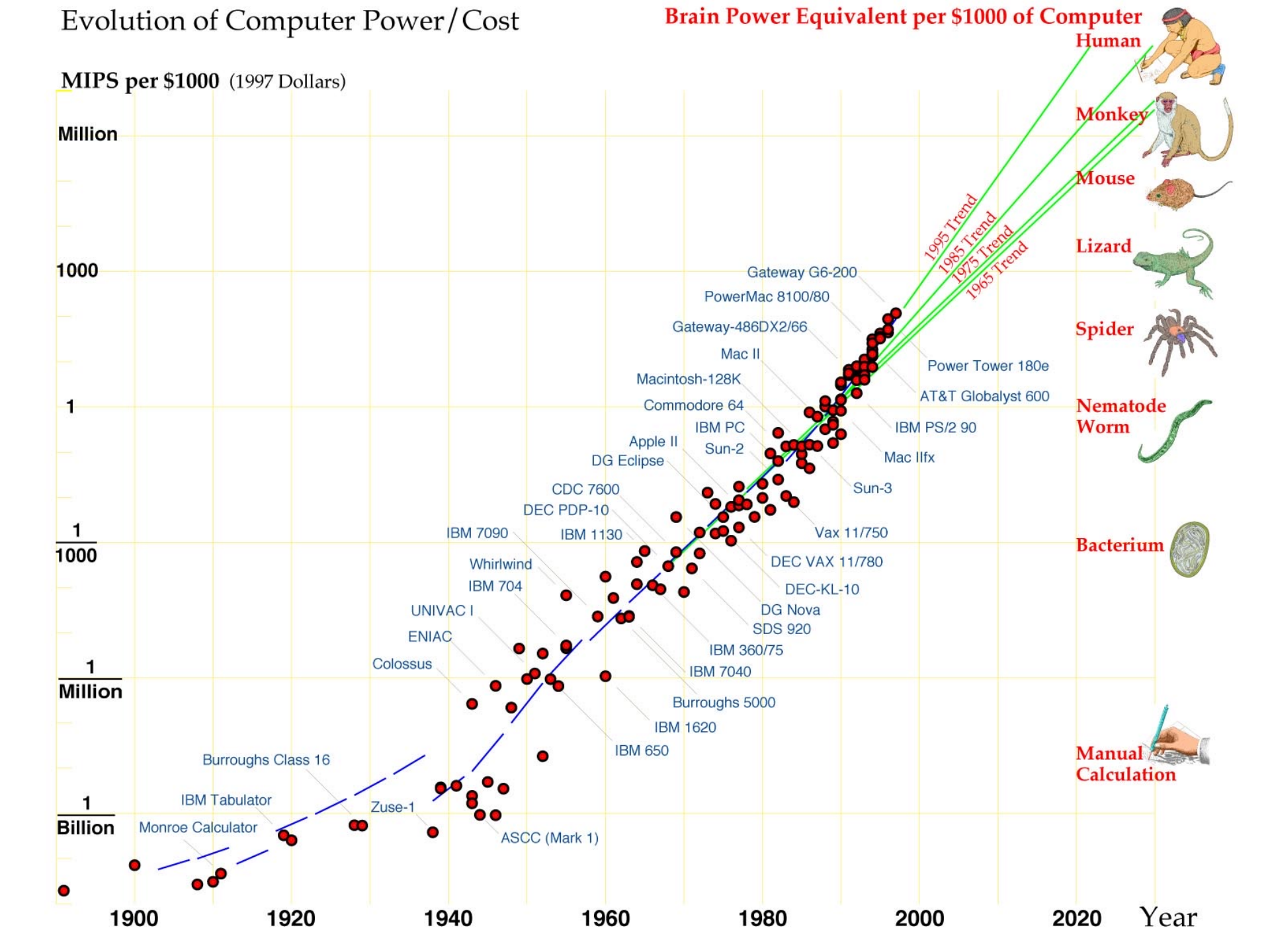
Moravec's argument(1998)

ROBOT: Mere Machine To Transcendent Mind

- 1 neuron = 1000 instructions/sec
- 1 synapse = 1 byte of information
- Human brain then processes 10^14 IPS and has 10^14 bytes of storage
- In 2000, we have 10^9 IPS and 10^9 bytes on a desktop machine
- Assuming Moore's law we obtain human level computing power in 2025, or with a cluster of 100 nodes in 2015.

Computer power available to AI and Robot programs

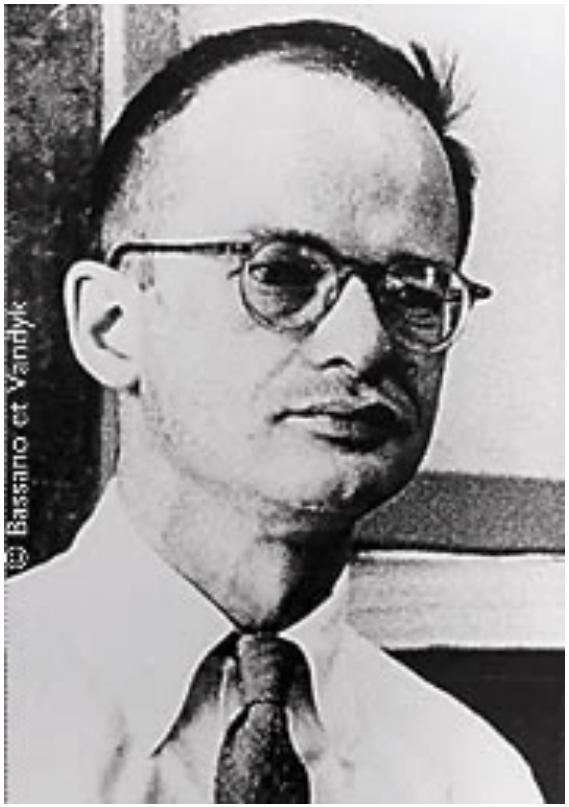



Evolution of Computer Power/Cost
MIPS per $1000 (1997 Dollars)
Brain Power Equivalent per $1000 of Computer
Human
Monkey
Mouse
Lizard
Spider
Nematode Worm
Bacterium
Manual Calculation
1995 Trend
1985 Trend
1975 Trend
1965 Trend
Million
1000
1
1/1000
1/Million
1/Billion
Gateway G6-200
PowerMac 8100/80
Gateway-486DX2/66
Mac II
Macintosh-128K
Commodore 64
IBM PC
Apple II
Sun-2
DG Eclipse
CDC 7600
DEC PDP-10
IBM 7090
IBM 1130
Whirlwind
IBM 704
UNIVAC I
ENIAC
Colossus
IBM 650
IBM 1620
Burroughs 5000
IBM 7040
IBM 360/75
SDS 920
DG Nova
DEC-KL-10
DEC VAX 11/780
Vax 11/750
Sun-3
Mac IIfx
IBM PS/2 90
AT&T Globalyst 600
Power Tower 180e
Burroughs Class 16
IBM Tabulator
Monroe Calculator
Zuse-1
ASCC (Mark 1)
1900
1920
1940
1960
1980
2000
2020
Year

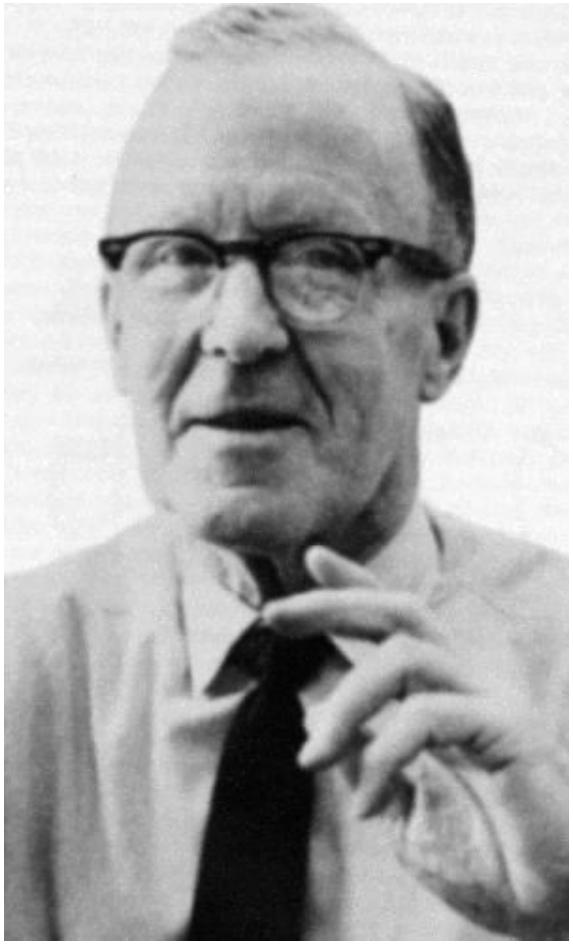
Some early history…

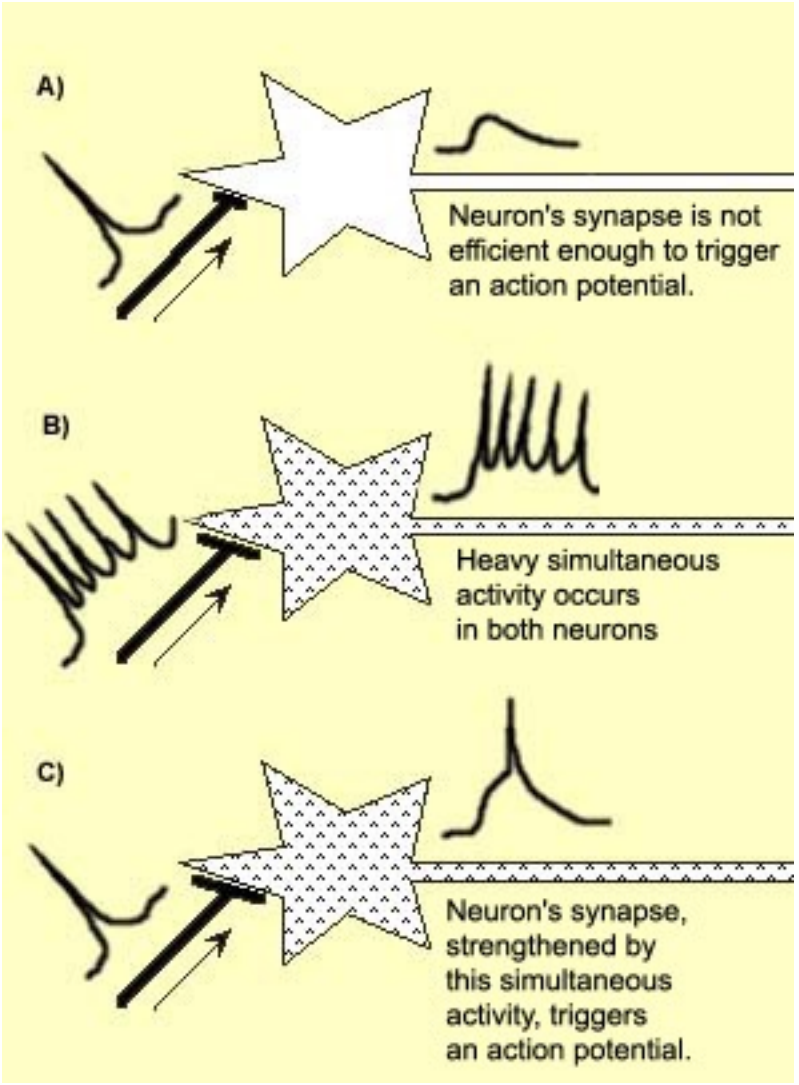
McCulloch & Pitts (1943)

A logical calculus of the ideas immanent in nervous activity



D. Hebb and Synaptic Learning





Turing's suggestion

Perception and Interaction

Language

456 A. M. TURING :

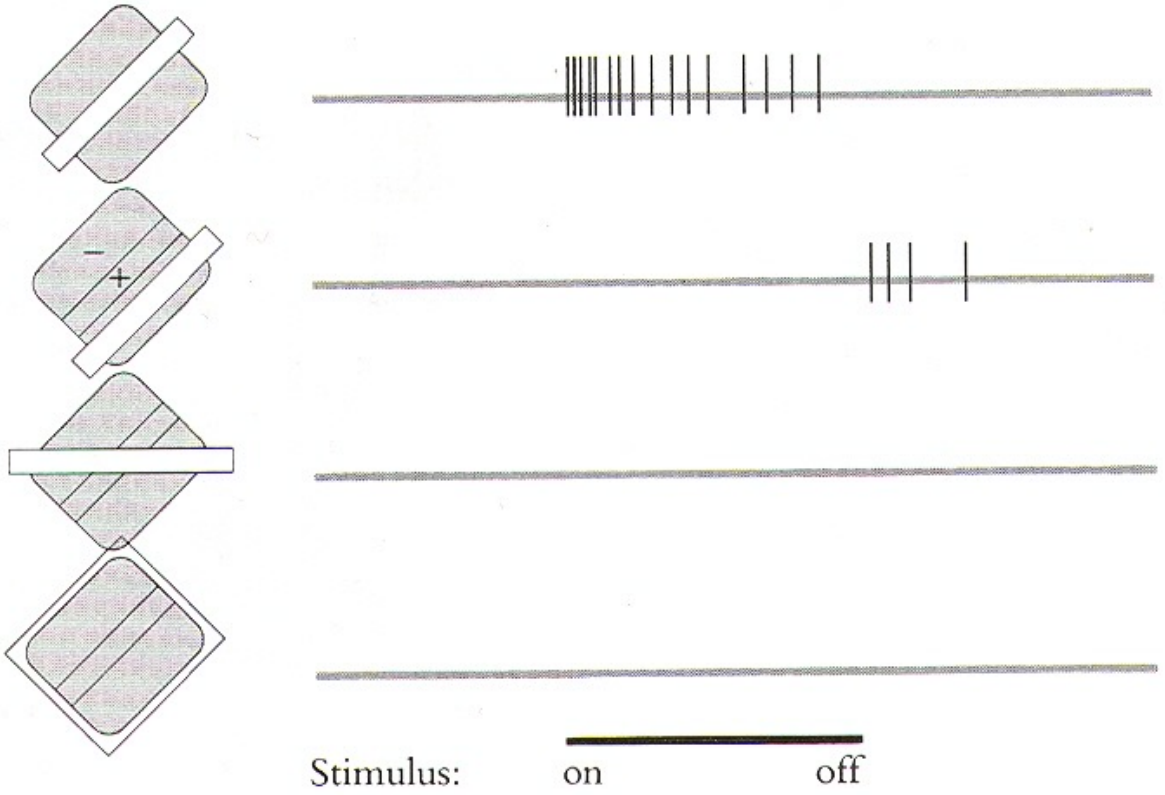
Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's ? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

Turing (1950)
Computing Machinery
And Intelligence

Paradigms for mechanizing intelligence ~1960

- Classic AI (McCarthy, Minsky, Newell, Simon)
  - Games, theorem-proving, reasoning
  - Search, represent and reason in first-order logic
- Pattern Recognition (Rosenblatt, Widrow)
  - Classification, Associative memory
  - Learning (Perceptrons …)
- Estimation and Control (Bellman, Kalman)
  - Decide action in uncertain, time-varying environment
  - Markov Decision Processes, adaptive control …

Hubel and Wiesel (1962) discovered orientation sensitive neurons in V1



Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima

NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan

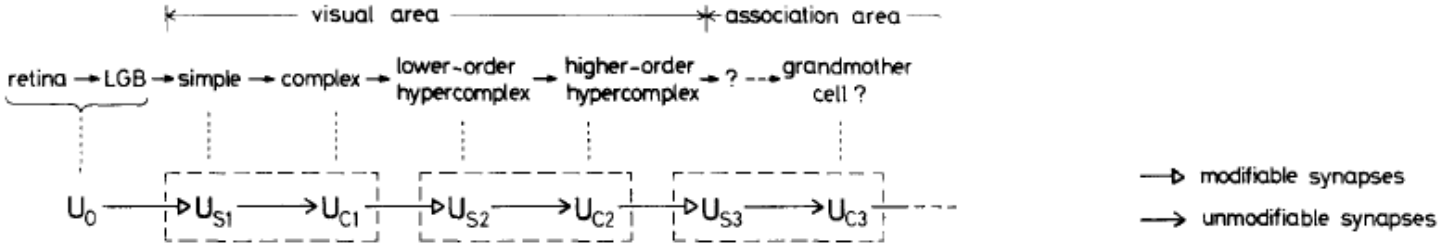


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

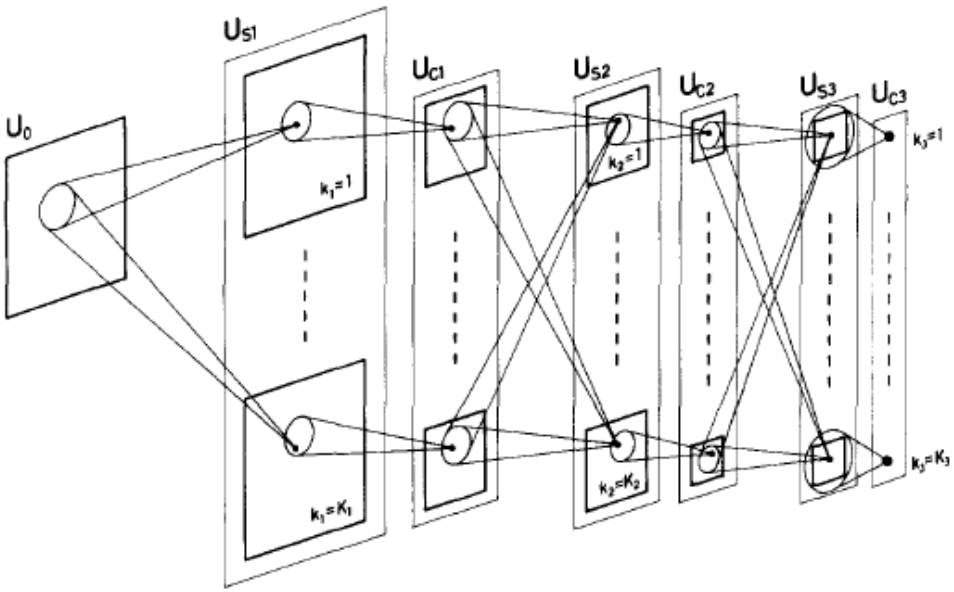


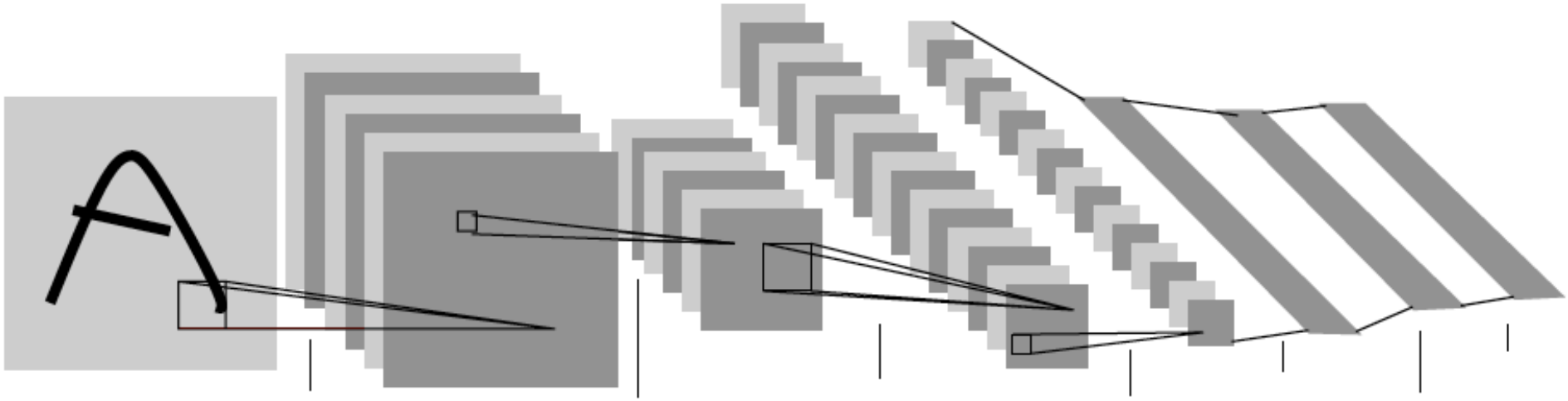
Fig. 2. Schematic diagram illustrating the interconnections between layers in the neocognitron

Biol. Cybernetics 36, 193–202 (1980)

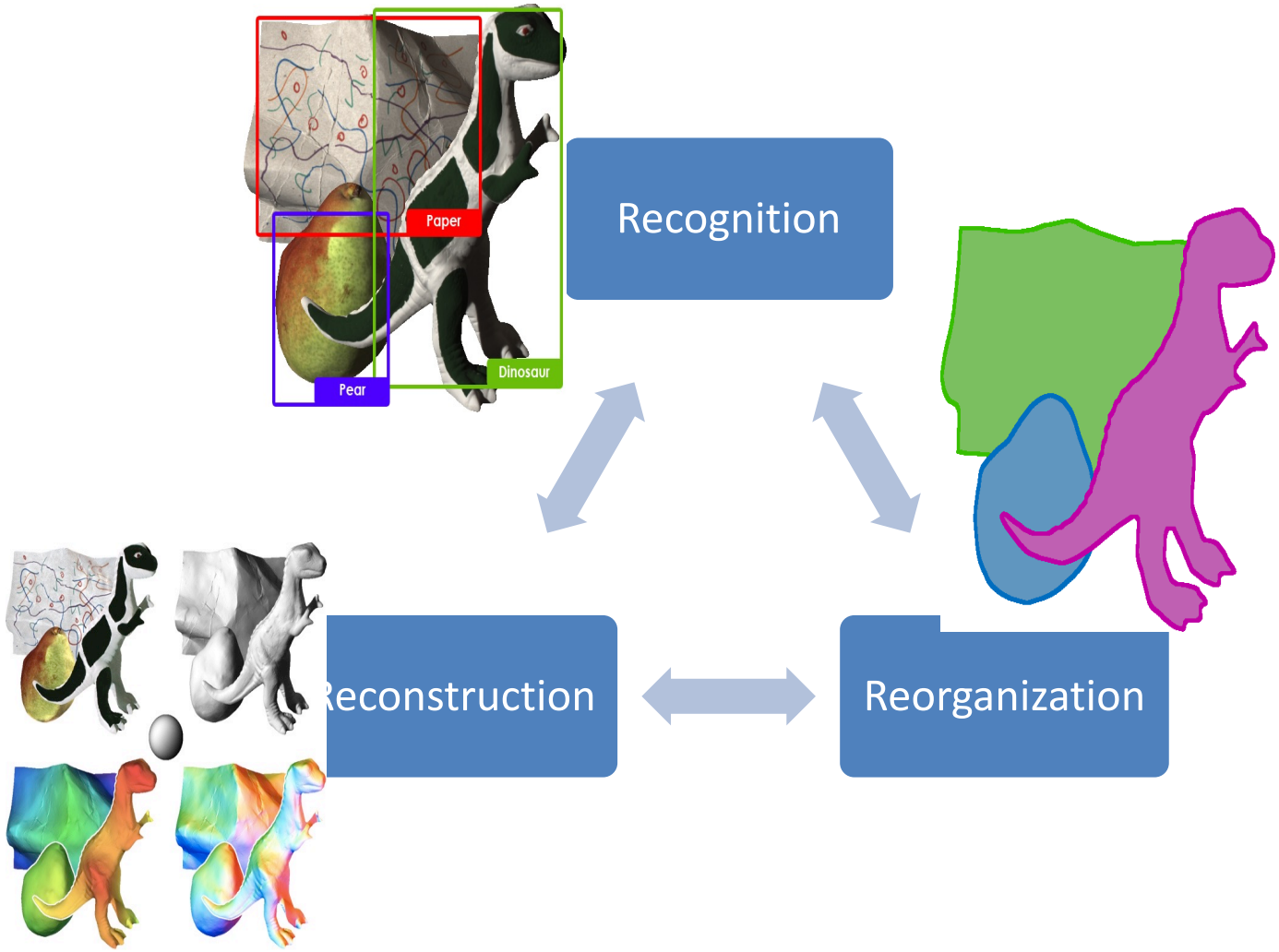
Convolutional Neural Networks (LeCun et al)

Used backpropagation to train the weights in this architecture

- First demonstrated by LeCun et al for handwritten digit recognition(1989)
- Applied in sliding window paradigm for tasks such as face detection in the 1990s.
- However was not competitive on standard computer vision object detection benchmarks in the 2000s.
- Thanks to availability of faster computing (GPUs) and large amounts of labeled data (Imagenet) we have seen an amazing renaissance led by Krizhevsky, Sutskever & Hinton (2012)



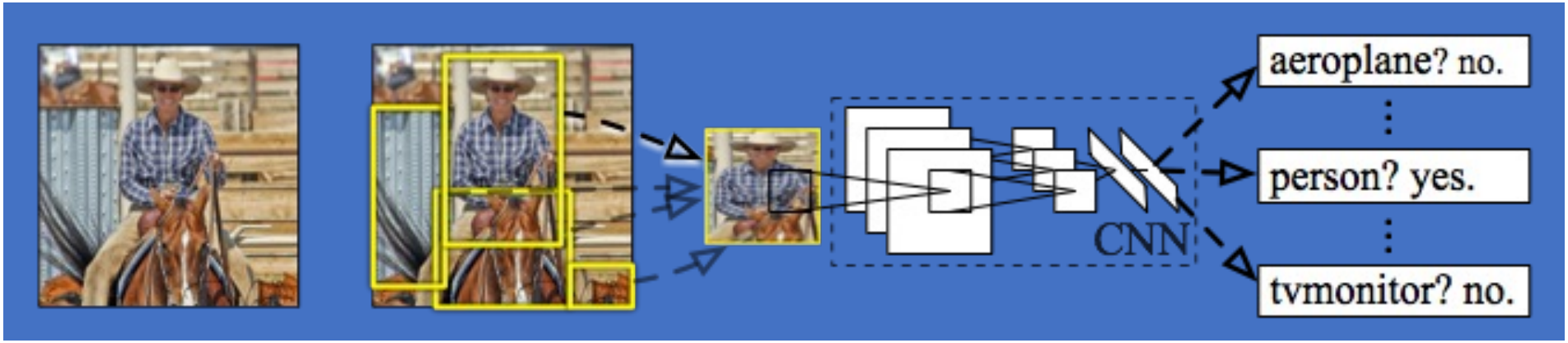
The 3R's of Vision: Recognition, Reconstruction & Reorganization



Talk at POCV Workshop, CVPR 2012

R-CNN: Regions with CNN features

Girshick, Donahue, Darrell & Malik (CVPR 2014)



Input image

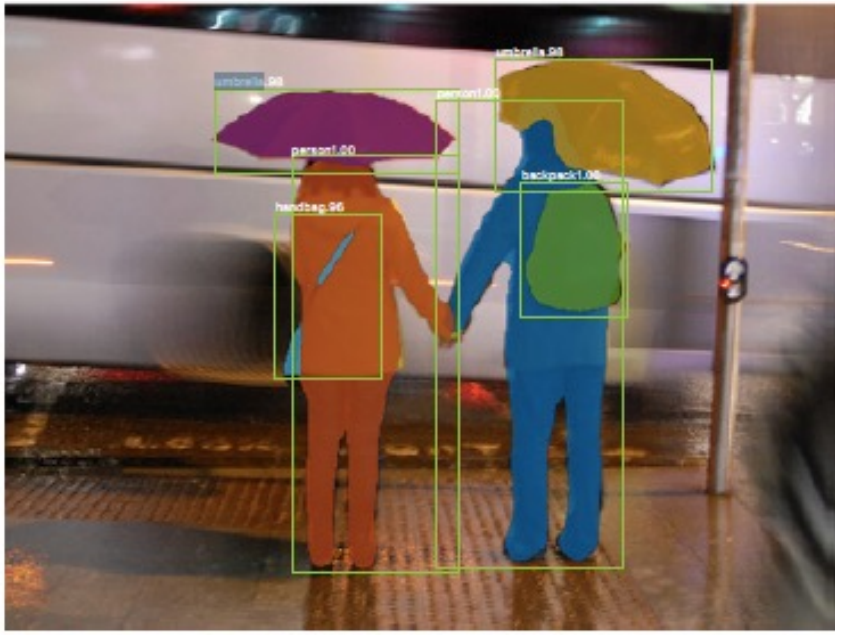
Extract region proposals (~2k / image)

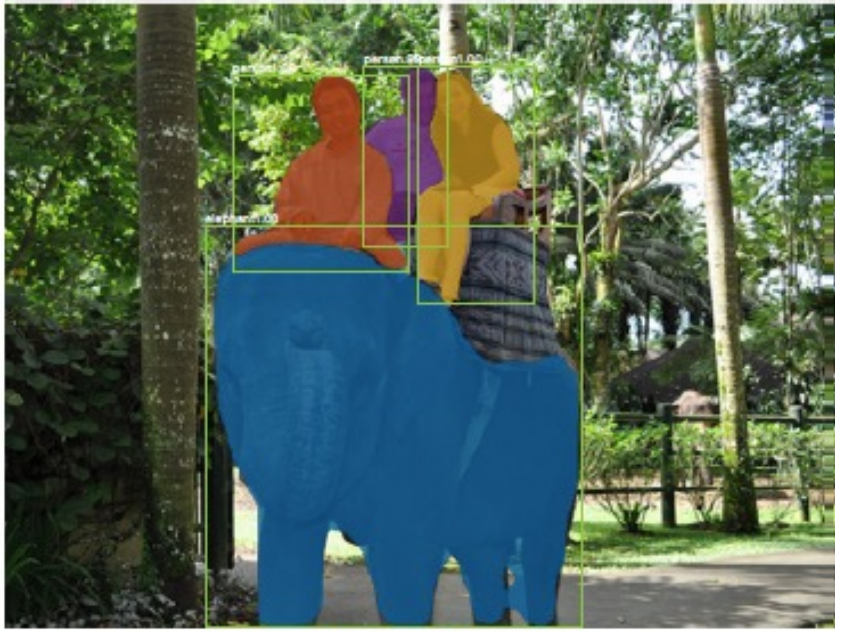
Compute CNN features

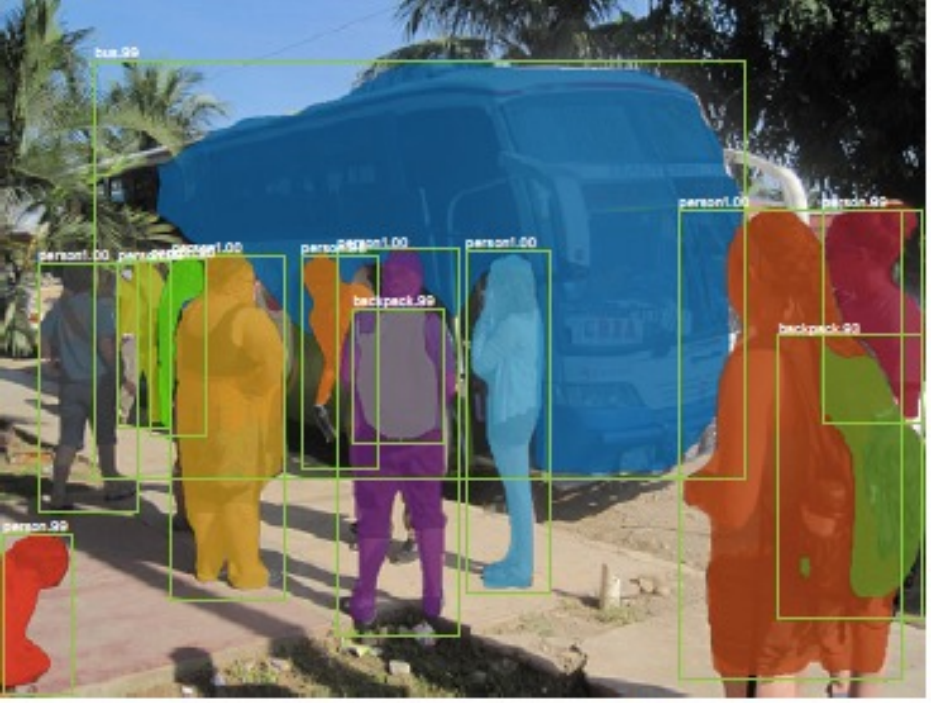
Classify regions (linear SVM)

Object detection, one of the most canonical problems of computer vision, yielded to deep learning

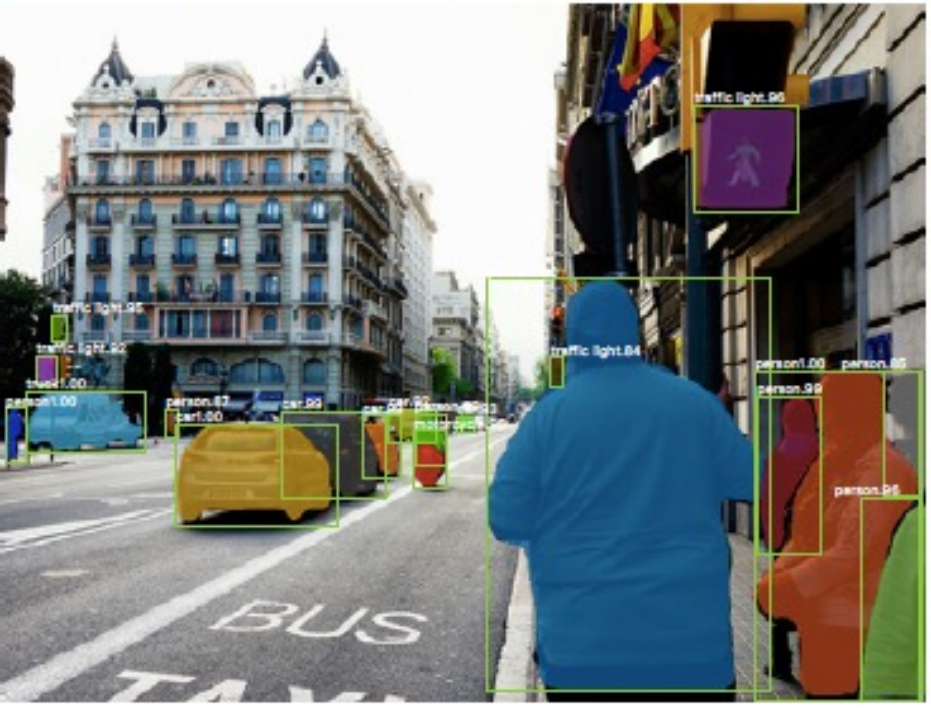
Mask R-CNN : He, Gkioxari, Dollar & Girshick (2017)

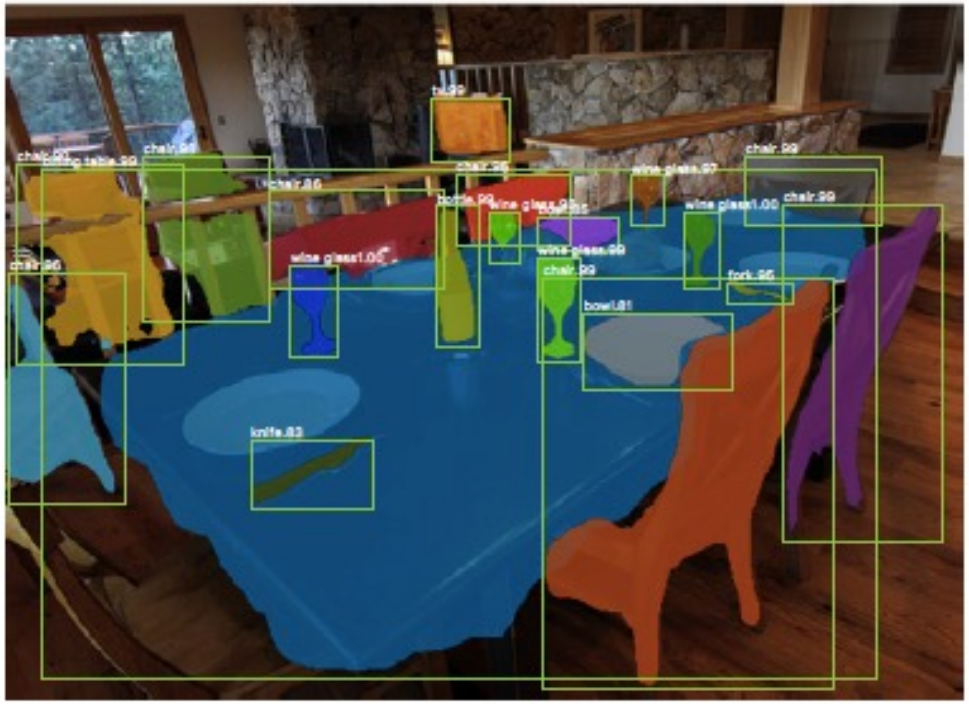





bus.99
person1.00
person.99
backpack.99
backpack.93

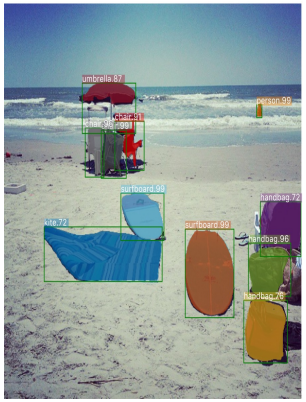

person1.00
person.99
sheep.99
sheep.96
sheep.93
sheep.91
sheep1.00


traffic light.96
traffic light.94
traffic light.84
person1.00
person.99
person.95
BUS

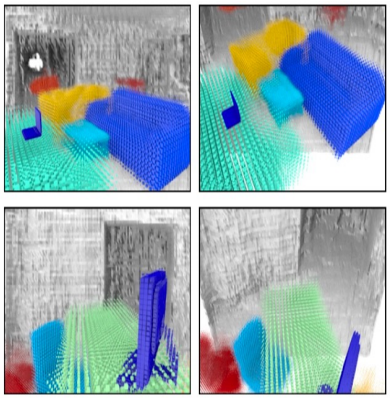

chair.99
chair.96
wine glass.97
wine glass1.00
bowl.81
fork.96
knife.83

The Future: Seeing in 3D

Box ~1970 - 2015

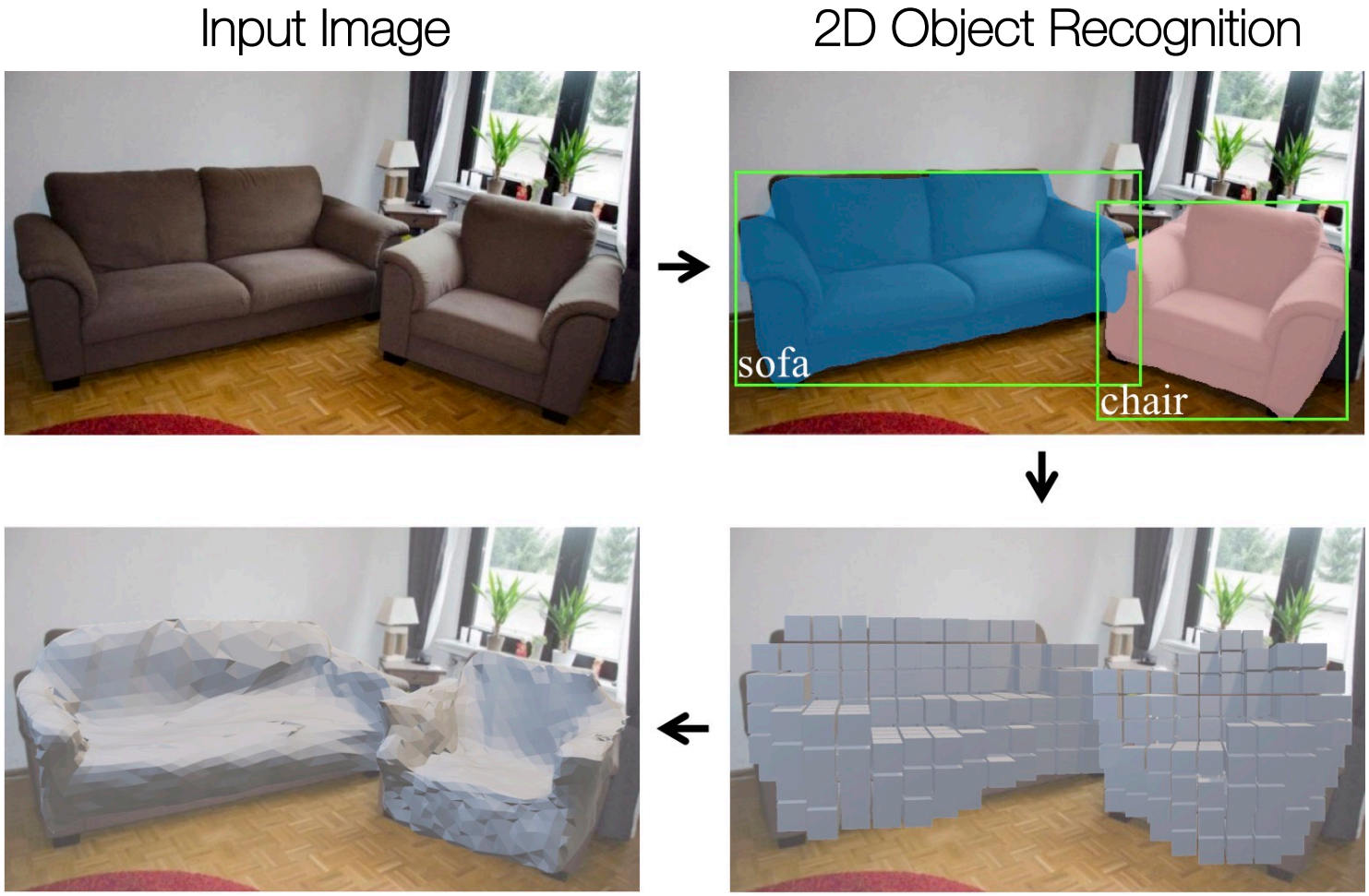


Mask 2015 - ???



3D shape: early days

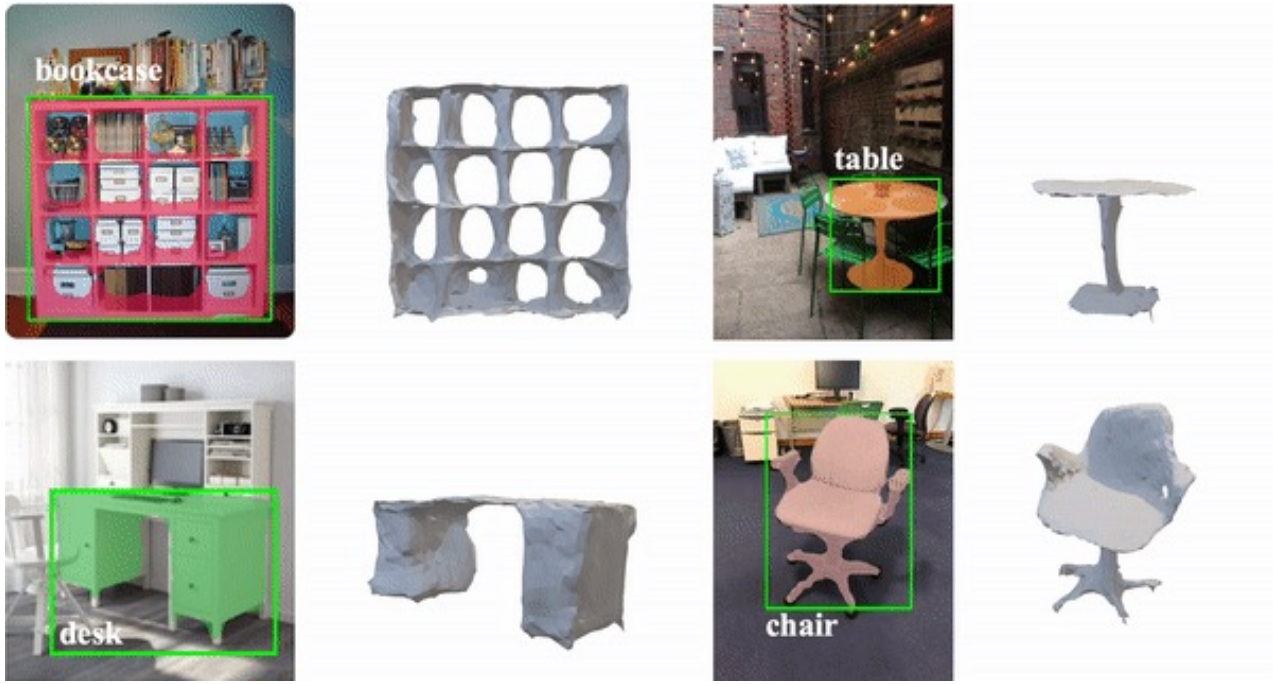
- Recognize everything *and understand its shape*
- Detailed understanding of semantics and geometry



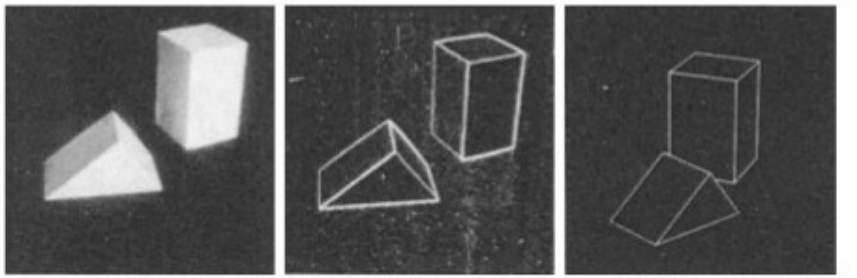
Mesh R-CNN

Gkioxari, Malik and Johnson (2019)

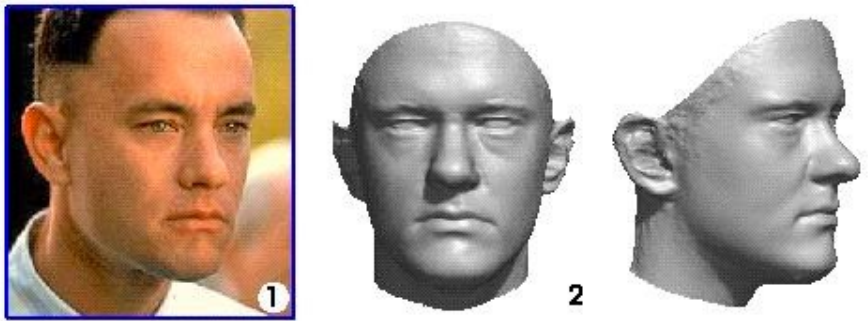
Results on Pix3D



Classical work in 3D Vision



Blocks World



Morphable Models

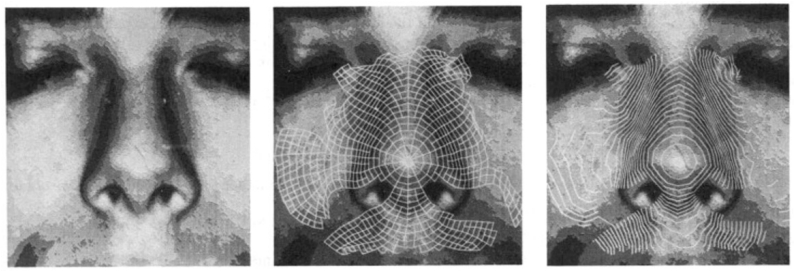
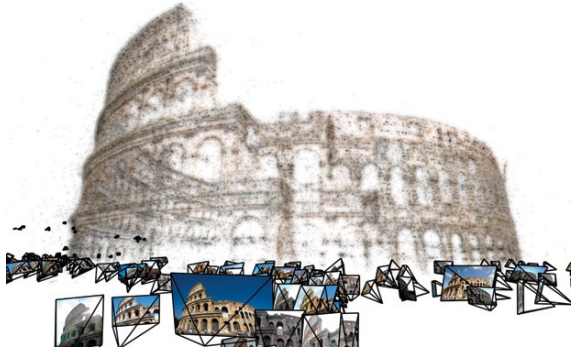


Figure 11-7. The shape-from-shading method is applied here to the recovery of the shape of a nose. The first picture shows the (crudely quantized) gray-level image available to the program. The second picture shows the base characteristics superimposed, while the third shows a contour map computed from the elevations found along the characteristic curves.

Shape from Shading



Structure from Motion

SLAM (Simultaneous Localization And Mapping)

Video Credits: Mur-Artal et al., Palmieri et al.

Human 3D Mesh Recovery from a Single Image

Kanazawa, Black, Jacobs and Malik

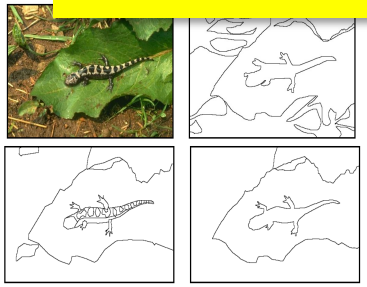
http://people.eecs.berkeley.edu/~jathushan/PHALP/

Challenges facing AI today

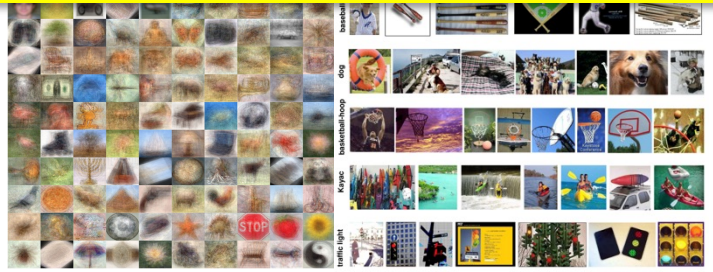
- Static dataset AI vs embodied AI
- Perception disconnected from action
- The semantic gap

We made a lot of progress with supervised learning..

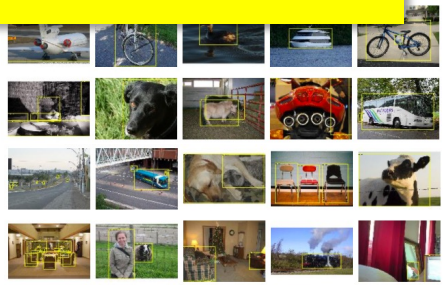
A "disembodied" well-curated moment in time



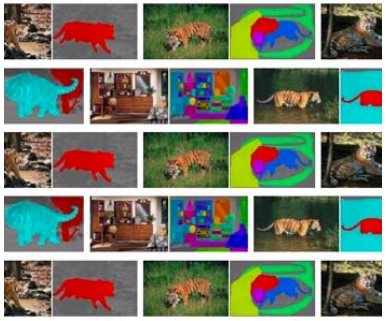
Berkeley Segments (2001)



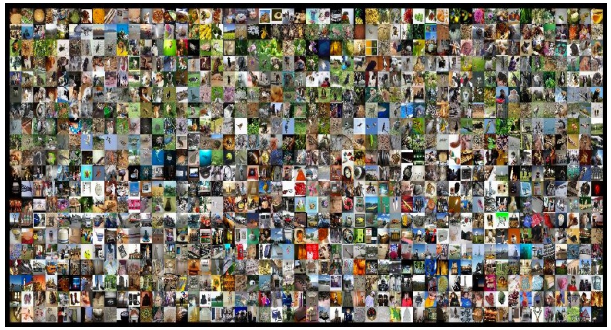
Caltech 101 (2004), Caltech 256 (2006)



PASCAL (2007-12)

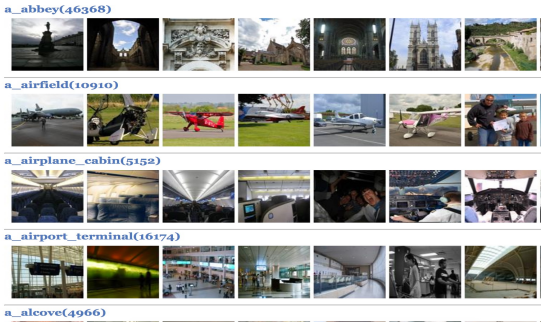


LabelMe (2007)



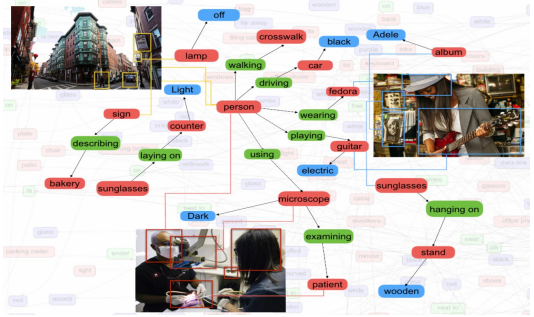
ImageNet (2009)

SUN (2010)



Places (2014)

MS COCO (2014)



Visual Genome (2016)

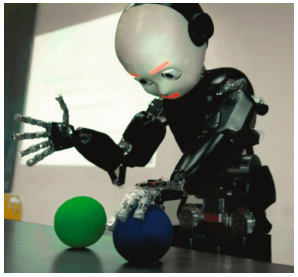
First-person perception and learning

Status quo:

Learning and inference with "disembodied" images/videos.

On the horizon:

Visual learning in the context of agent goals, interaction, and multi-sensory observations.



Around the World in 3,000 Hours of Egocentric Video

CVPR 22

Ego4D team

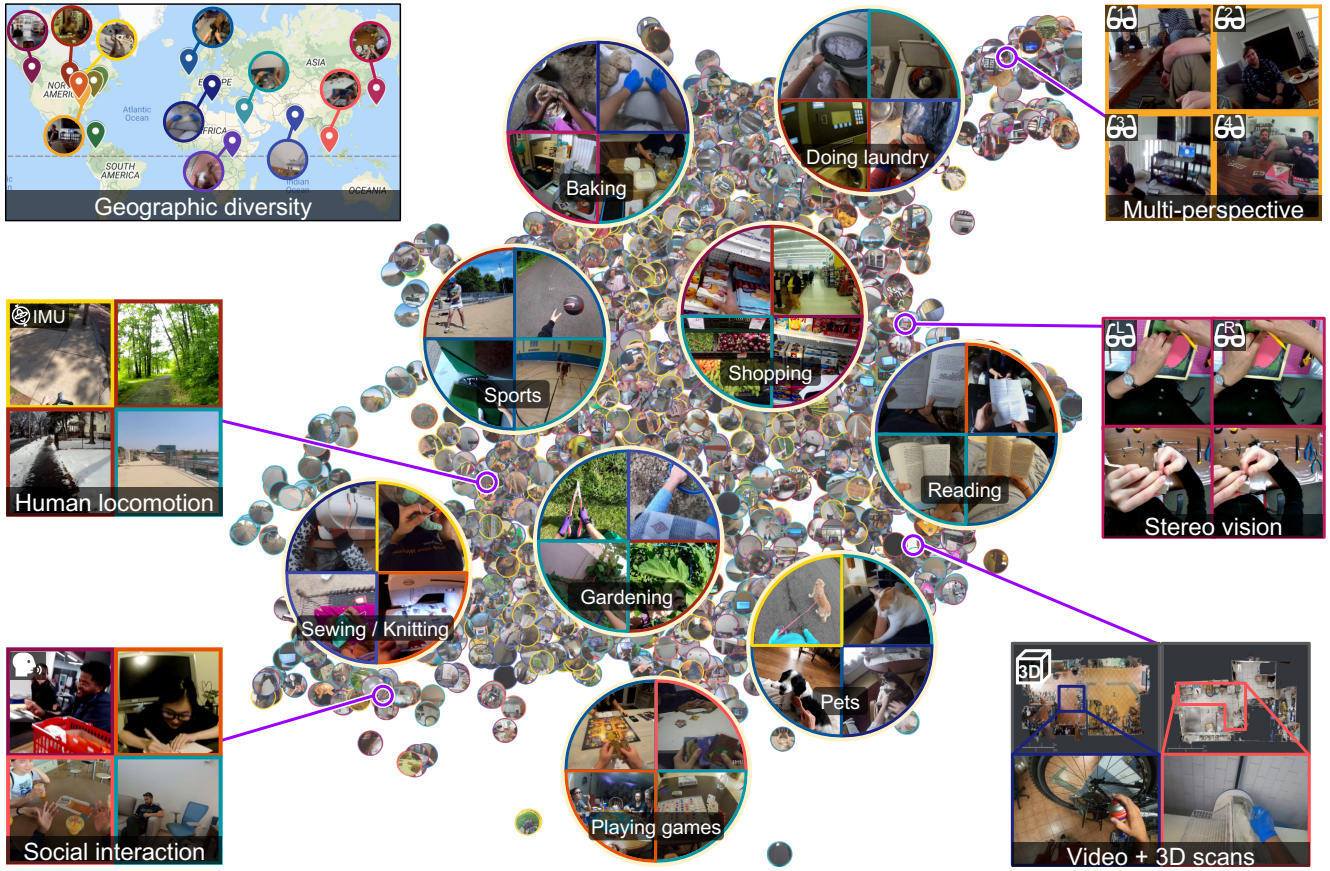
Ego4D: Around the World in 3,000 Hours of Egocentric Video

Kristen Grauman[1,2], Andrew Westbury[1], Eugene Byrne*[1], Zachary Chavis*[3], Antonino Furnari*[4], Rohit Girdhar*[1], Jackson Hamburger*[1], Hao Jiang*[5], Miao Liu*[6], Xingyu Liu*[7], Miguel Martin*[1], Tushar Nagarajan*[1,2], Ilija Radosavovic*[8], Santhosh Kumar Ramakrishnan*[1,2], Fiona Ryan*[6], Jayant Sharma*[3], Michael Wray*[9], Mengmeng Xu*[10], Eric Zhongcong Xu*[11], Chen Zhao*[10], Siddhant Bansal[17], Dhruv Batra[1], Vincent Cartillier[1,6], Sean Crane[7], Tien Do[3], Morrie Doulaty[13], Akshay Erapalli[13], Christoph Feichtenhofer[1], Adriano Fragomeni[9], Qichen Fu[7], Christian Fuegen[13], Abrham Gebreselasie[12], Cristina González[14], James Hillis[5], Xuhua Huang[7], Yifei Huang[15], Wenqi Jia[6], Weslie Khoo[16], Jachym Kolar[13], Satwik Kottur[13], Anurag Kumar[5], Federico Landini[13], Chao Li[5], Yanghao Li[1], Zhenqiang Li[15], Karttikeya Mangalam[1,8], Raghava Modhugu[17], Jonathan Munro[9], Tullie Murrell[1], Takumi Nishiyasu[15], Will Price[9], Paola Ruiz Puentes[14], Merey Ramazanova[10], Leda Sari[5], Kiran Somasundaram[5], Audrey Southerland[6], Yusuke Sugano[15], Ruijie Tao[11], Minh Vo[5], Yuchen Wang[16], Xindi Wu[7], Takuma Yagi[15], Yunyi Zhu[11], Pablo Arbeláez†[14], David Crandall†[16], Dima Damen†[9], Giovanni Maria Farinella†[4], Bernard Ghanem†[10], Vamsi Krishna Ithapu†[5], C. V. Jawahar†[17], Hanbyul Joo†[1], Kris Kitani†[7], Haizhou Li†[11], Richard Newcombe†[5], Aude Oliva†[18], Hyun Soo Park†[3], James M. Rehg†[6], Yoichi Sato†[15], Jianbo Shi†[19], Mike Zheng Shou†[11], Antonio Torralba†[18], Lorenzo Torresani†[1,20], Mingfei Yan†[5], Jitendra Malik[1,8]

[1]Facebook AI Research (FAIR), [2]University of Texas at Austin, [3]University of Minnesota, [4]University of Catania, [5]Facebook Reality Labs, [6]Georgia Tech, [7]Carnegie Mellon University, [8]UC Berkeley, [9]University of Bristol, [10]King Abdullah University of Science and Technology, [11]National University of Singapore, [12]Carnegie Mellon University Africa, [13]Facebook, [14]Universidad de los Andes, [15]University of Tokyo, [16]Indiana University, [17]International Institute of Information Technology, Hyderabad, [18]MIT, [19]University of Pennsylvania, [20]Dartmouth

Ego4D data: everyday activity around the world

- 3,025+ hours of video
- 855+ camera wearers
- Geographic diversity
- Occupational diversity
- Unscripted daily life activit


Geographic diversity
Baking
Doing laundry
Multi-perspective
IMU
Human locomotion
Sports
Shopping
Reading
Stereo vision
Gardening
Sewing / Knitting
Pets
Social interaction
Playing games
3D
Video + 3D scans

Challenges facing AI today

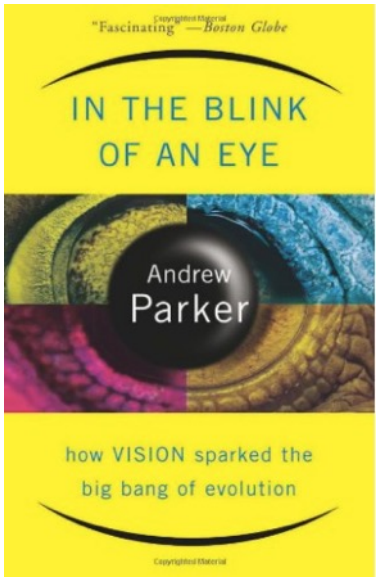
- Static dataset AI vs embodied AI
- Perception disconnected from action
- The semantic gap

Phylogeny of Intelligence

Cambrian Explosion
540 million years ago

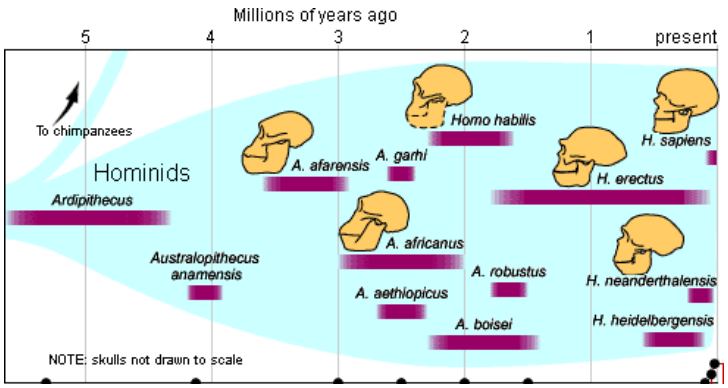
Variety of life forms,
almost all phyla emerge

Animals that could
see and move



Gibson: we see in order to move and we move in order to see

Hominid evolution, last 5 million years



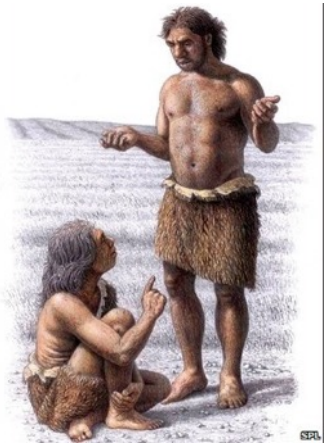
Bipedalism

Opposable thumb

Tool use

Anaxogaras: It is because of his being armed with hands that man is the most intelligent animal

Modern humans, last 50 K years

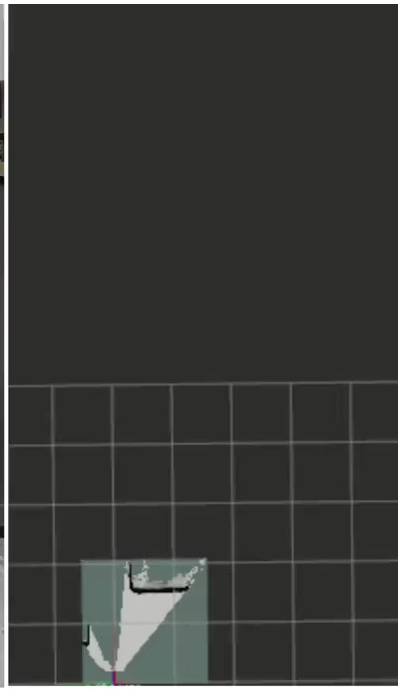


Language

Abstract thinking

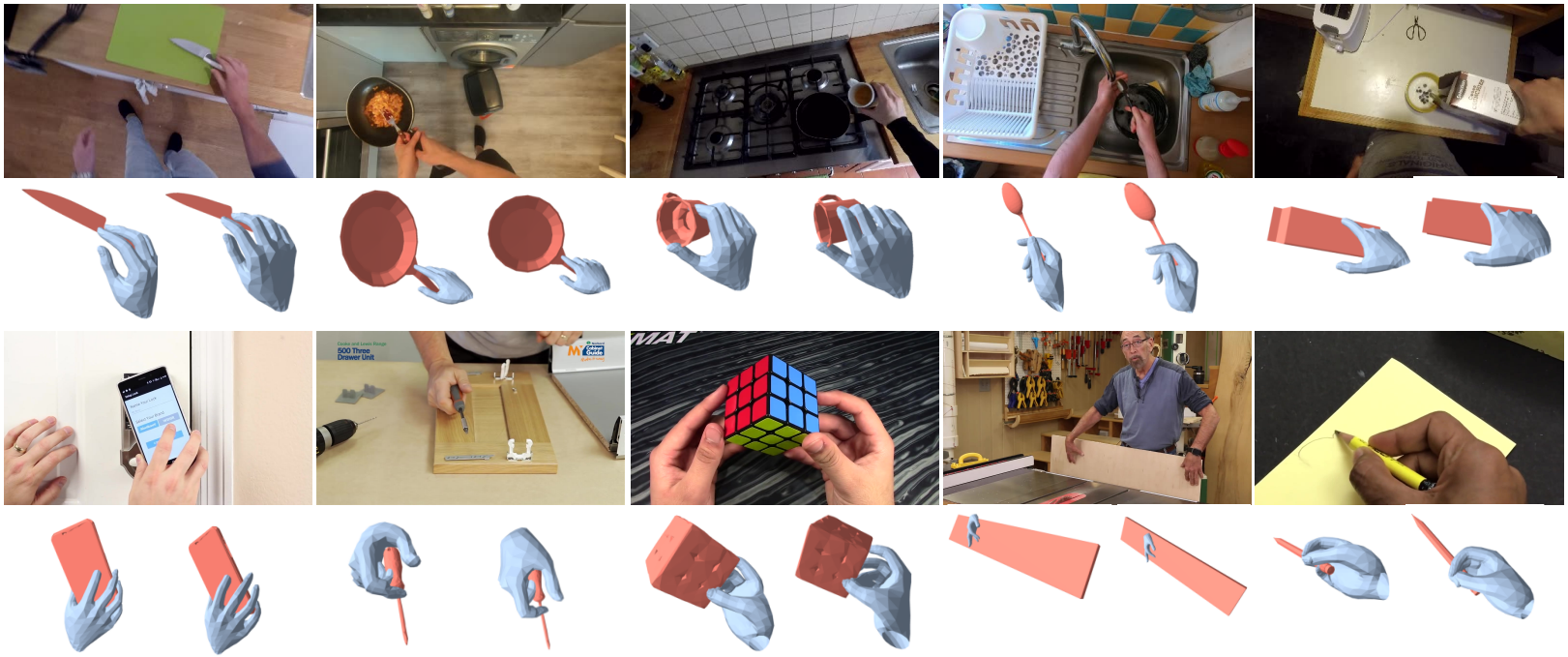
Symbolic behavior

Navigation in Cluttered Indoor Setting



Recognizing Human-Object Interactions in the Wild

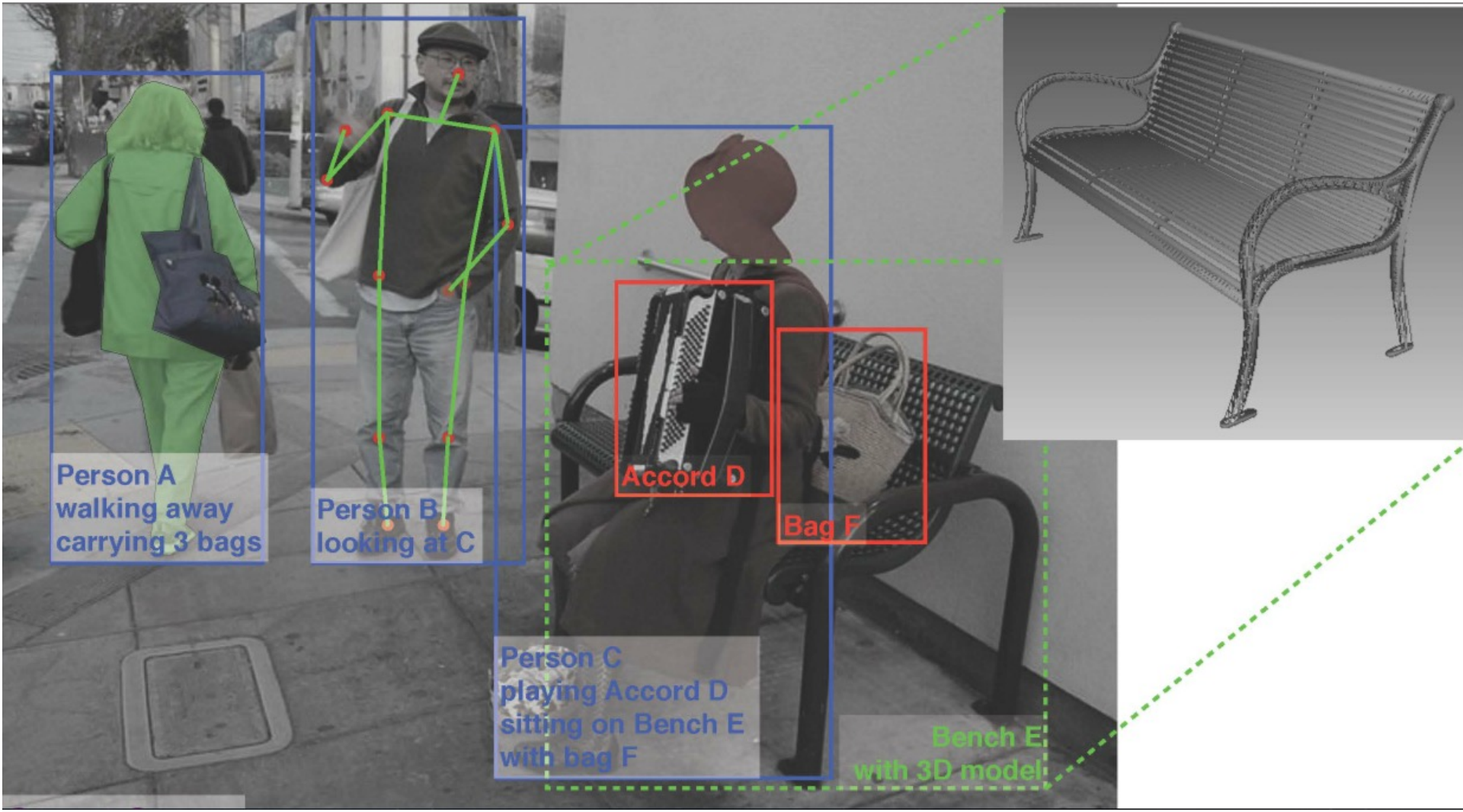
Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, Jitendra Malik



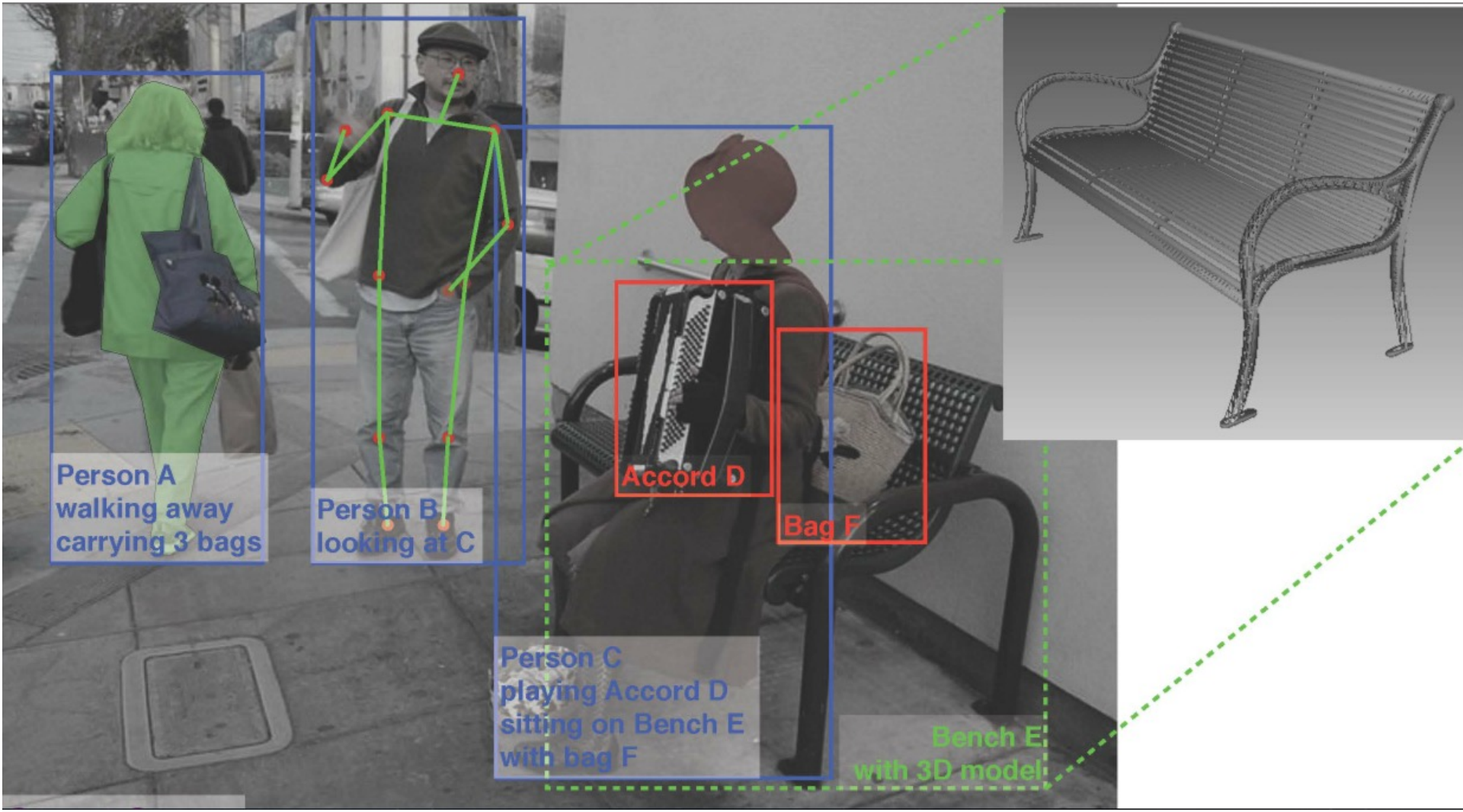
Challenges facing AI today

- Static dataset AI vs embodied AI
- Perception disconnected from action
- The semantic gap

What we might be able to infer…



What we would like to infer…



Will person B put some money into Person C's tip bag?

The semantic gap exists in present day NLP as well

ELSEVIER

Physica D: Nonlinear Phenomena
Volume 42, Issues 1–3, June 1990, Pages 335-346

The symbol grounding problem

Stevan Harnad

Show more

https://doi.org/10.1016/0167-2789(90)90087-6

Get rights and content

Abstract

There has been much discussion recently about the scope and limits of purely symbolic models of the mind and about the proper role of connectionism in cognitive modeling. This paper describes the "symbol grounding problem": How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols? The problem is analogous to trying to learn Chinese from a Chinese/Chinese dictionary alone. A candidate solution is sketched: Symbolic representations must

AI systems need to build “mental models”

If the organism carries a `small-scale model’ of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it (Craik, 1943,Ch. 5, p.61)

Commonsense is not just facts, it is a collection of models

Where should we go next?

- Turing’s Baby

Ontogeny of Intelligence

Perception and Interaction

Language

456 A. M. TURING :

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's ? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

Turing (1950)
Computing Machinery
And Intelligence

The Development of Embodied Cognition: Six Lessons from Babies

Linda Smith & Michael Gasser

Abstract. The embodiment hypothesis is the idea that intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity. In this paper we offer six lessons for *developing* embodied intelligent agents suggested by research in developmental psychology. We argue that starting as a baby grounded in a physical, social and linguistic world is crucial to the development of the flexible and inventive intelligence that characterizes humankind.

The Six Lessons

- Be multi-modal
- Be incremental
- Be physical
- Explore
- Be social
- Use language

- I think this provides the right structure for viewing the stages of inbuilt, supervised by observation, supervised by interaction, supervised by culture

We can only see a short distance ahead, but we can see plenty there that needs to be done.
-Alan Turing