Perception, Learning and Control

Jitendra Malik

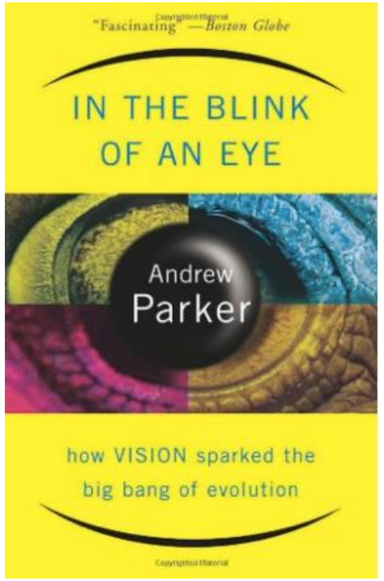
UC Berkeley

Phylogeny of Intelligence

Cambrian Explosion
540 million years ago

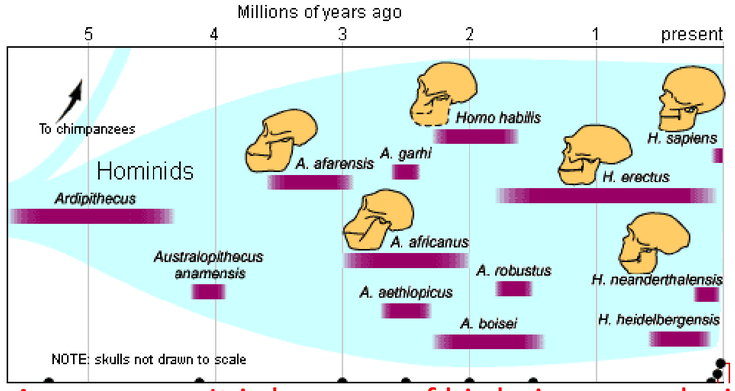
Variety of life forms,
almost all phyla emerge

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



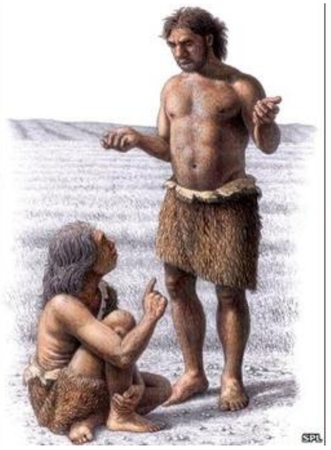
Bipedalism

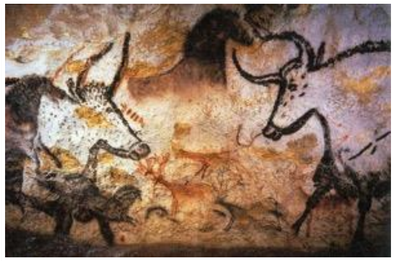
Opposable thumb

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

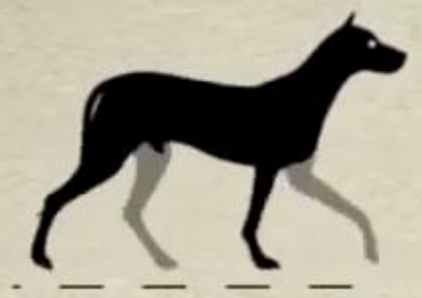
Symbolic behavior

The evolutionary progression

- Vision and Locomotion
- Manipulation
- Language

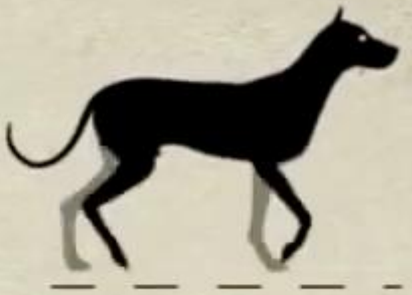
Successes in AI seem to follow the same order!

Animal Gaits



walk

amble



pace

trot

canter

gallup / run

-Stephen Cunnane

Computational Gaits

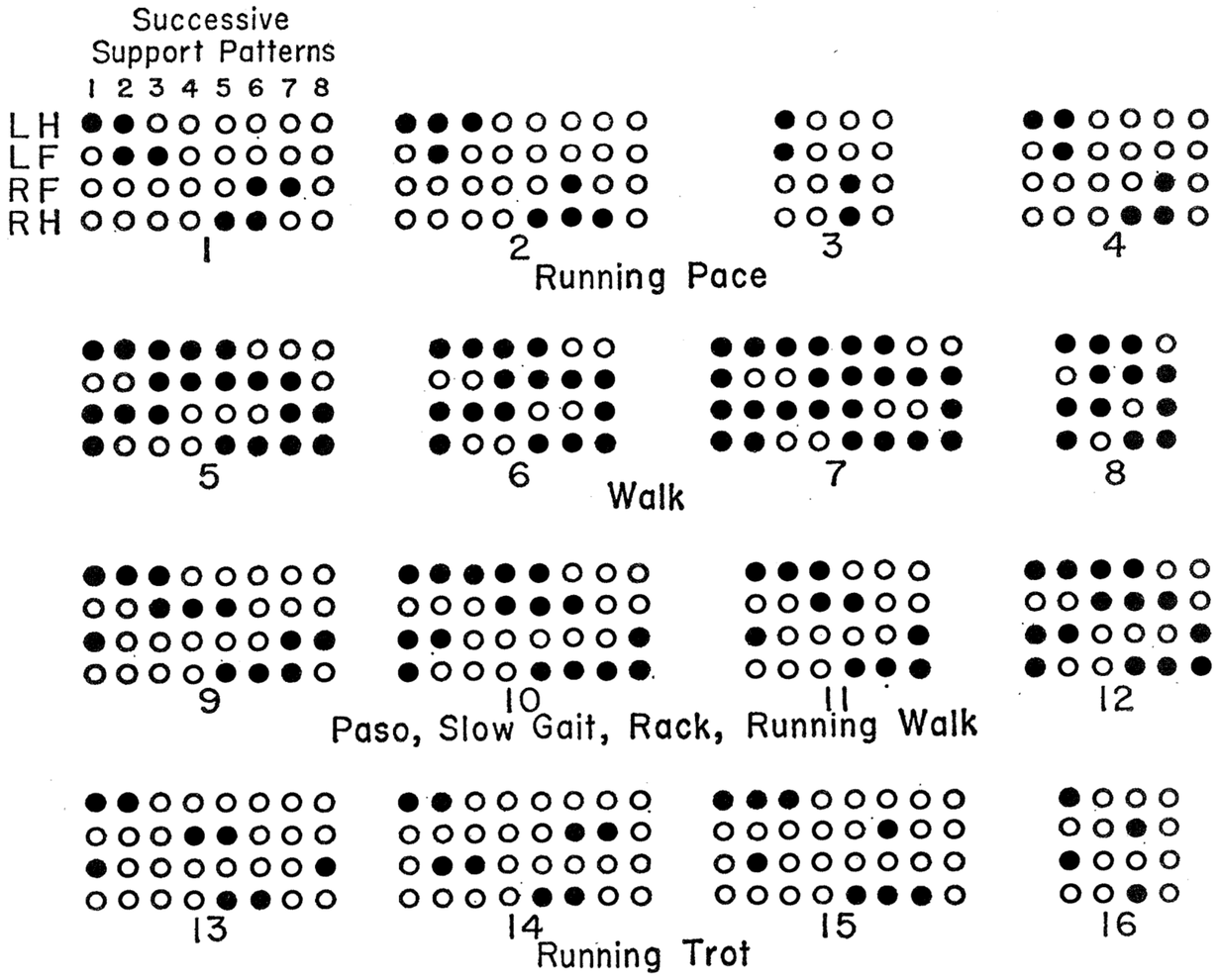


Fig. 3. Sixteen of the many support sequences that might be used by horses doing the gaits indicated. The initials L, R, F, and H stand for left, right, fore, and hind feet. Black circles indicate feet supporting weight; open circles, unweighted feet. Within each diagram, a vertical row of four circles shows a particular pattern of support. Thus, in the fifth support pattern of support sequence No. 1, only the RH foot is on the ground. Each sequence starts with the footfall of the LH foot. Sequences 1, 5, 9, 10, 13, 14, and 15 are relatively common for horses.

Symmetrical Gaits of Horses

Gaits can be expressed numerically and analyzed graphically to reveal their nature and relationships.

Milton Hildebrand

Hildebrand (1965)

Computational Gaits — Central Pattern Generators

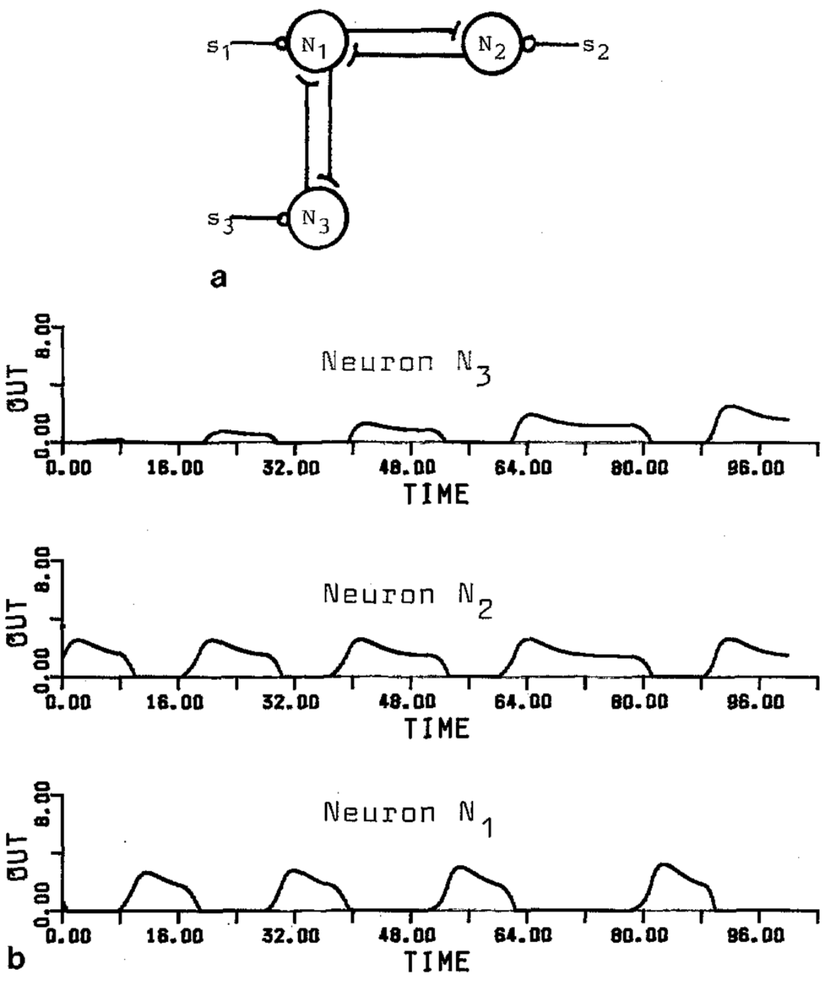


Fig. 3a and b. Network II. **a** Structure. **b** Outputs; $a_{ij}=1.5$, $s_1=s_2=5$, $s_3(t)=0.05t$; the other parameters are the same as in Fig. 2b

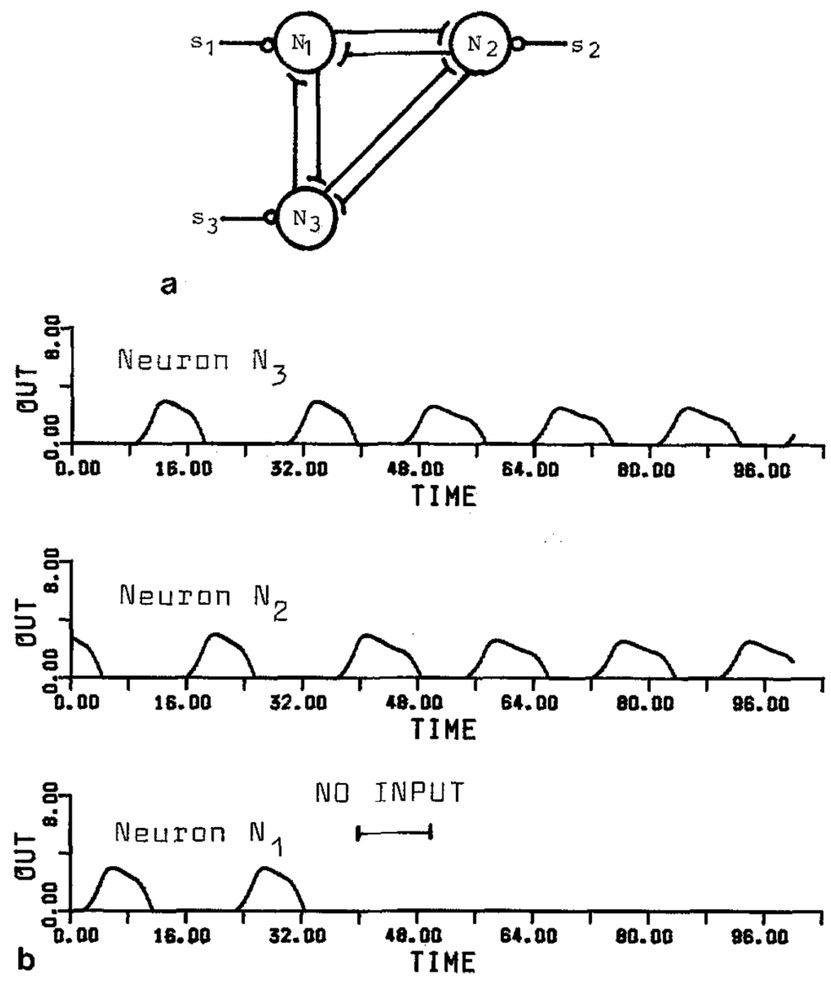


Fig. 4a and b. Network III. **a** Structure. **b** Output; $a_{ij}=1.5$ for every i and j; $s_1(t)=0$ for $40<t<50$, otherwise $s_1(t)=5$; $s_1=s_2=5$ at every t; the other parameters are the same as in Fig. 2b

Matsuoka, Kiyotoshi. "Mechanisms of frequency and pattern control in the neural rhythm generators." *Biologic al cybernetics* (1987)

Computational Gaits —Raibert Controller

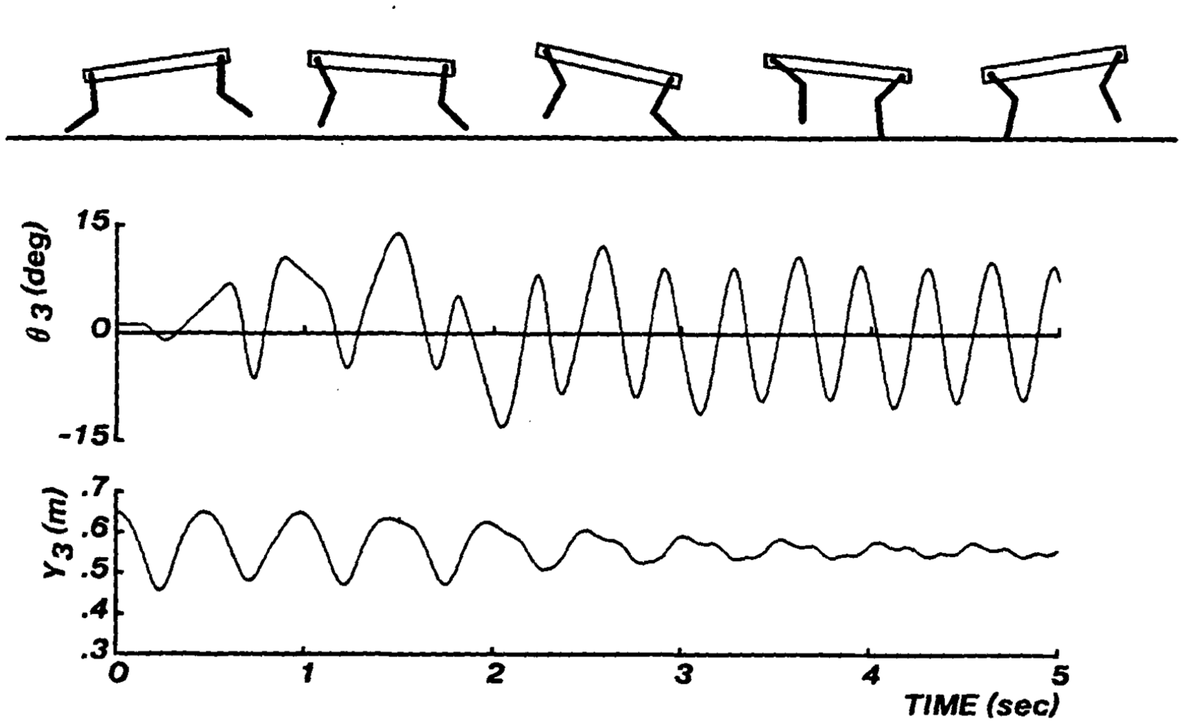


Figure 1-2: Planar two-legged model running with a bounding gait. Each leg of the model is controlled independently to regulate hopping height and forward velocity. The body rocks back and forth in a passively stabilized oscillation, with very little up and down motion of the center of gravity. When running is initiated there is a random pattern of rocking, but it soon stabilizes. TOP: The cartoon shows behavior when running at about 4 m/sec. MIDDLE: Attitude of body. BOTTOM: Altitude of body.

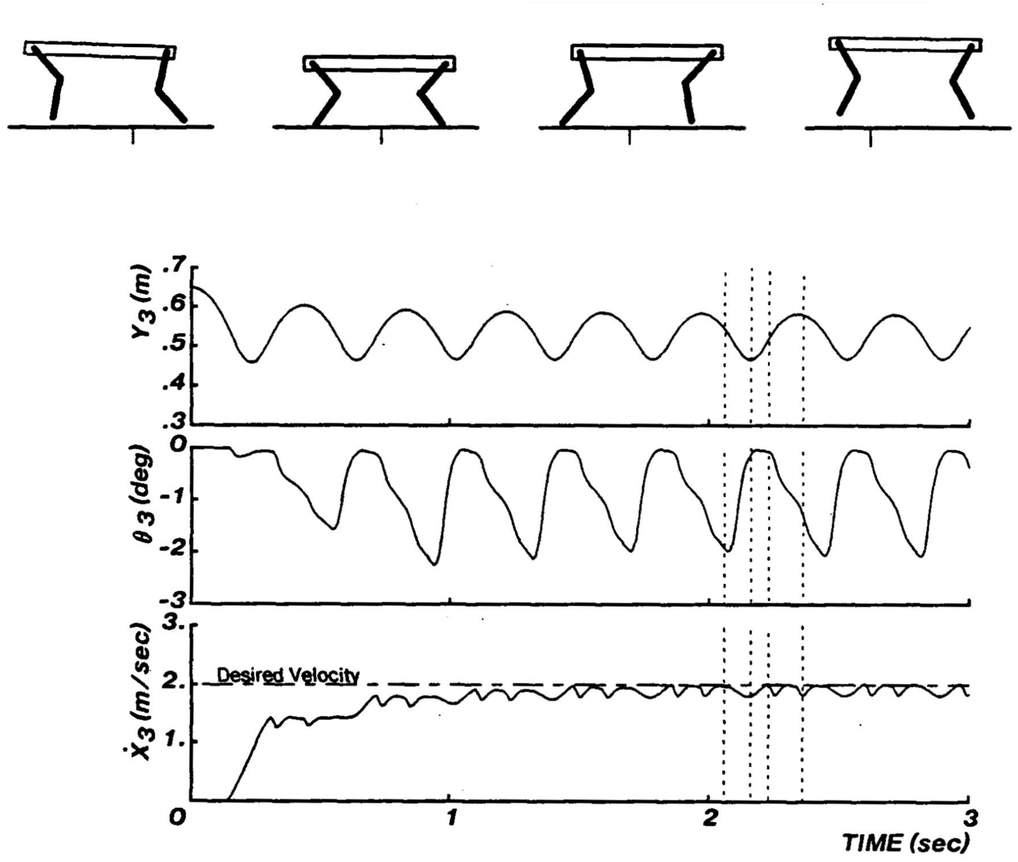


Figure 3-5: The dog trotting. The legs used a constant thrust for vertical control with differential thrust for attitude control. During flight, each foot swung forward to the estimated center of its hip-print. During stance, hip torques corrected the forward velocity. The errors in θ_3 occurred when the feet left the ground and the legs were swept forward. This caused the body to pitch nose downward.

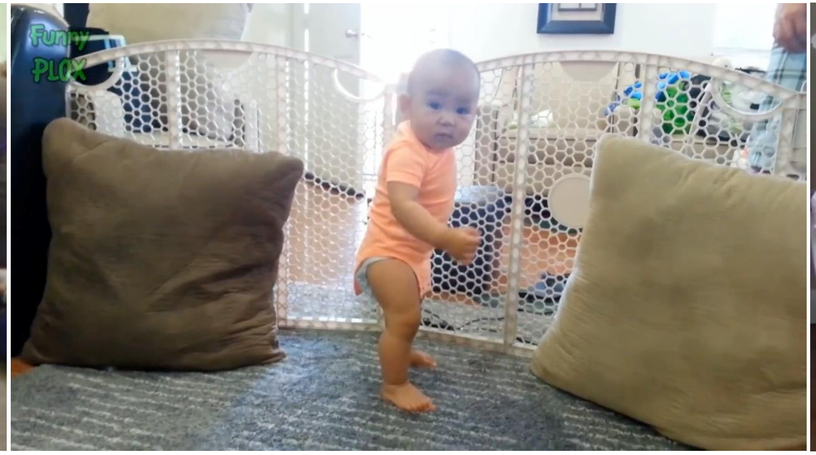
Raibert, Marc H., et al. *Dynamically Stable Legged Locomotion*. MIT ARTIFICIAL INTELLIGENCE LAB, 1983.


Spot - Boston Dynamics

Alternating Gait

Crawl Gait

Gaits in Real Life



Gaits in Real Life

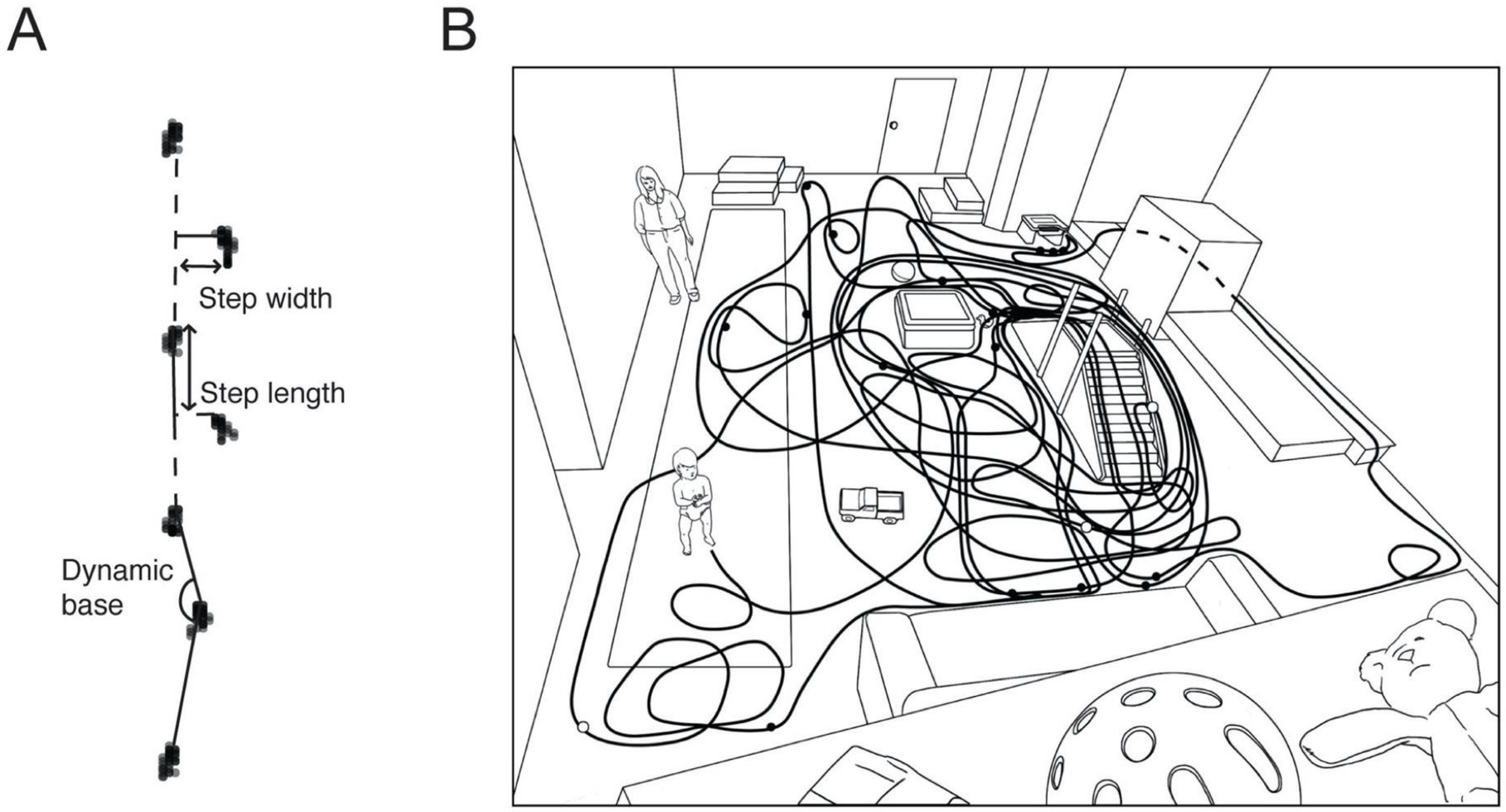


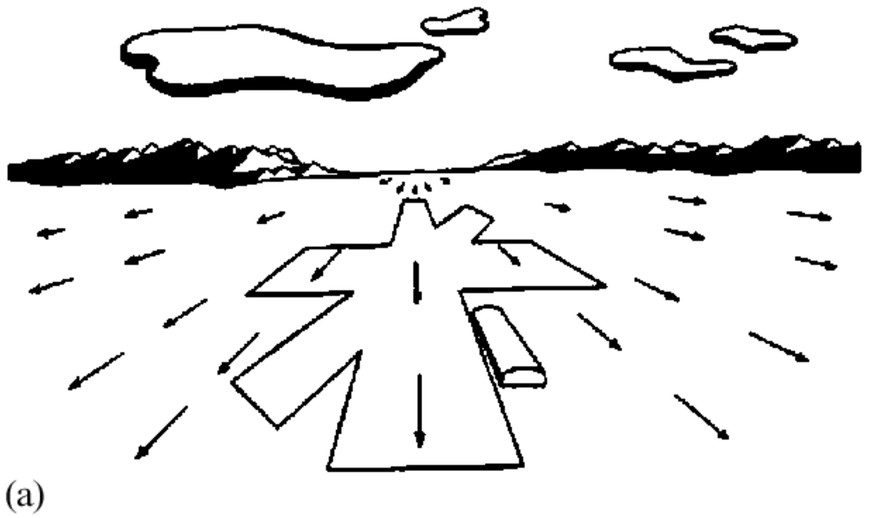
Figure 2.
(A) Sample footprint path showing the calculation of standard gait measures obtained on the pressure-sensitive gait carpet (step length, step width, and dynamic base angle). (B) Schematic drawing of the laboratory playroom. The squiggly line represents a typical path in

Cole, Whitney G., Scott R. Robinson, and Karen E. Adolph. "Bouts of steps: The organization of infant exploration." *Developmental psychobiology* (2016).

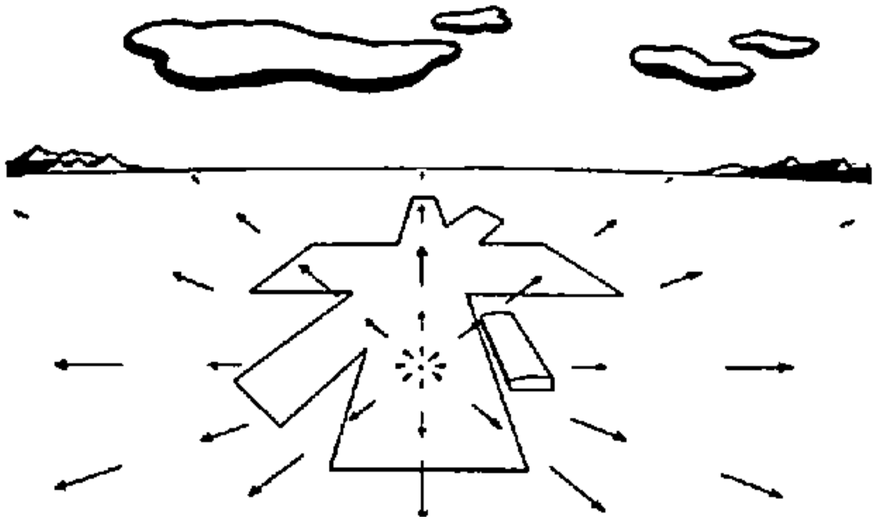
Problem 1

Gaits are not a complete model of all legged movement

Coupled Perception and Action



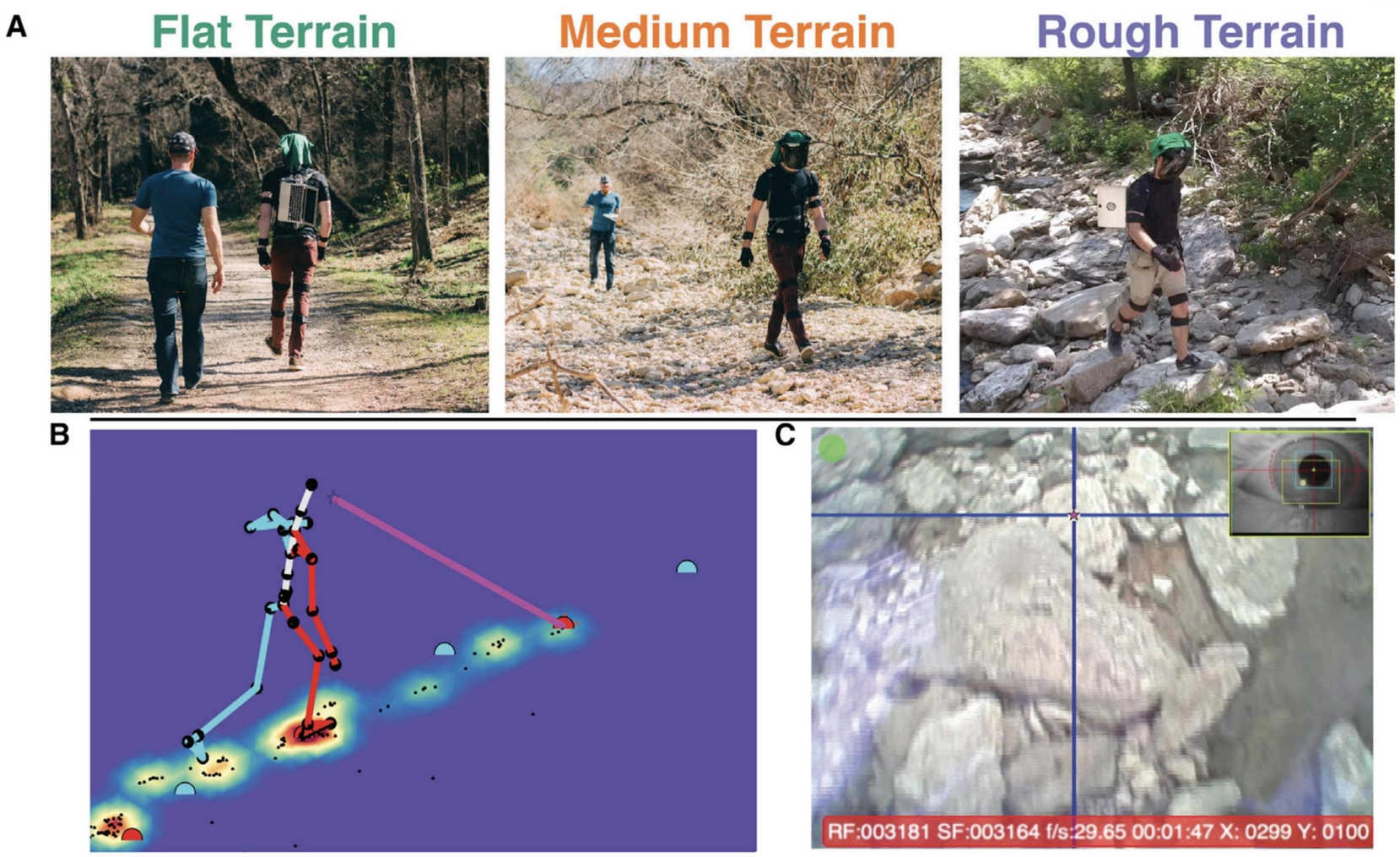
(a)



“We see in order to move, and we move in order to see”

— J.J. Gibson

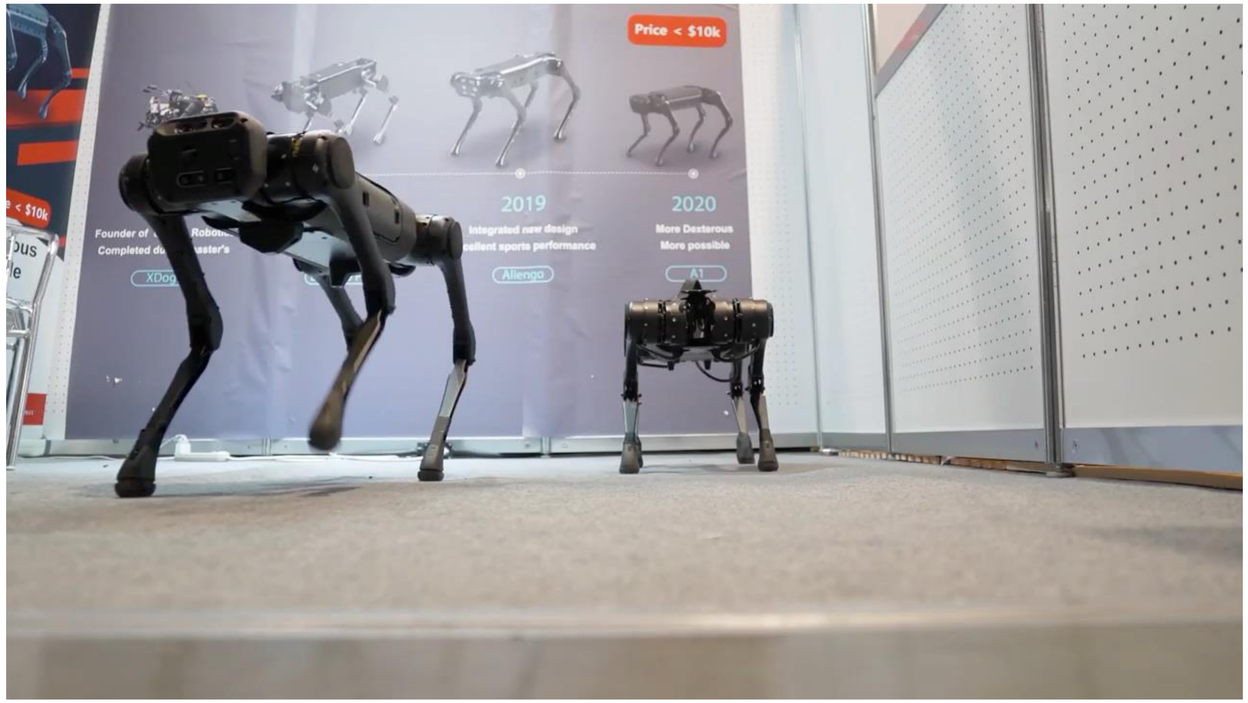
Coupled Perception and Action



Matthis, Jonathan Samir, Jacob L. Yates, and Mary M. Hayhoe. "Gaze and the control of f

Weakly Coupled Vision

Spot — Boston Dynamics



Laikago — Unitree Robotics

Problem 2

Vision is missing, or only weakly coupled

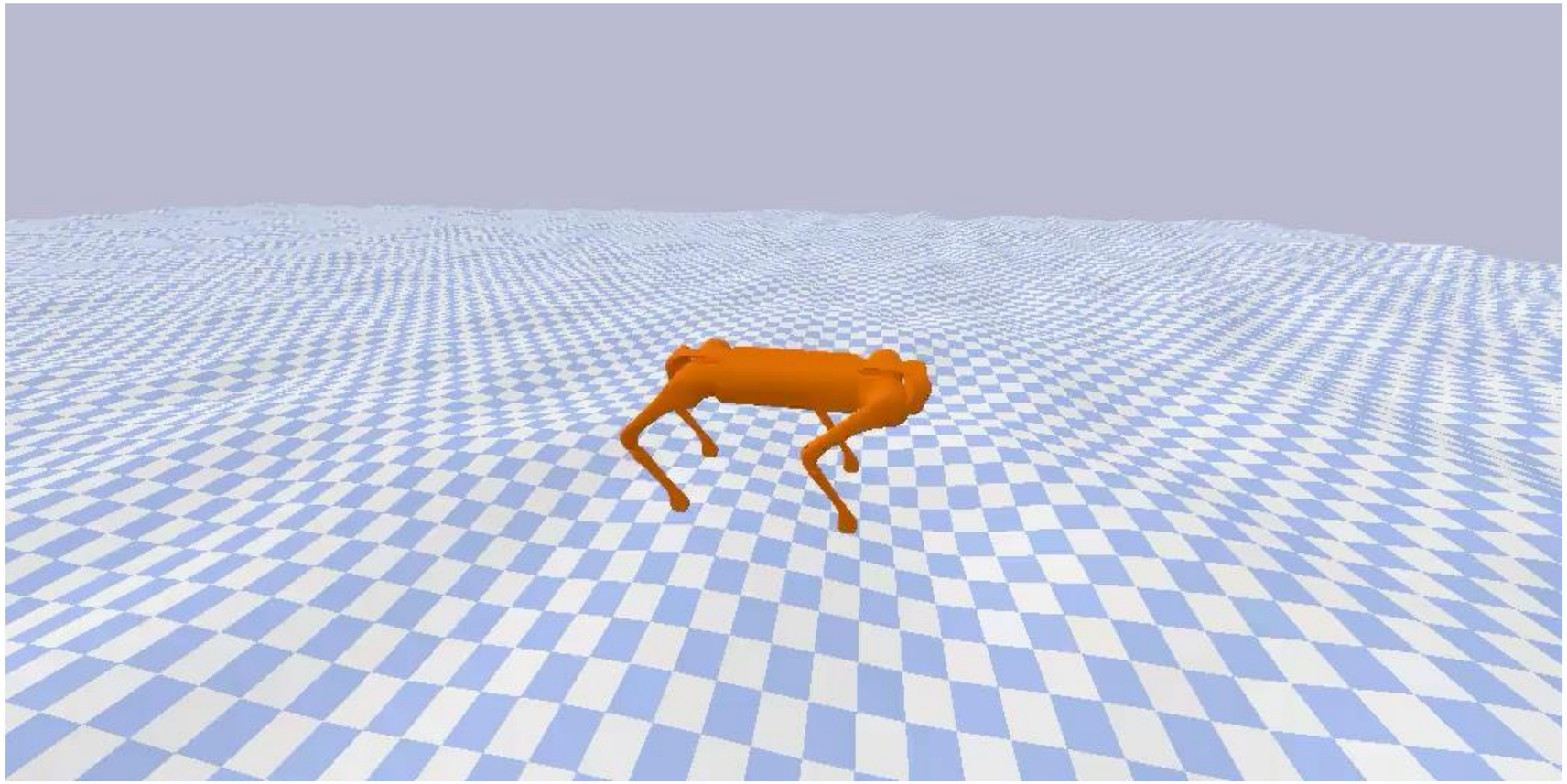
Our Axioms

1. Solve general legged locomotion (no explicit gait models).
2. Use vision from ground up.
3. Use learning as much as possible.

Visual Control of Legged Locomotion

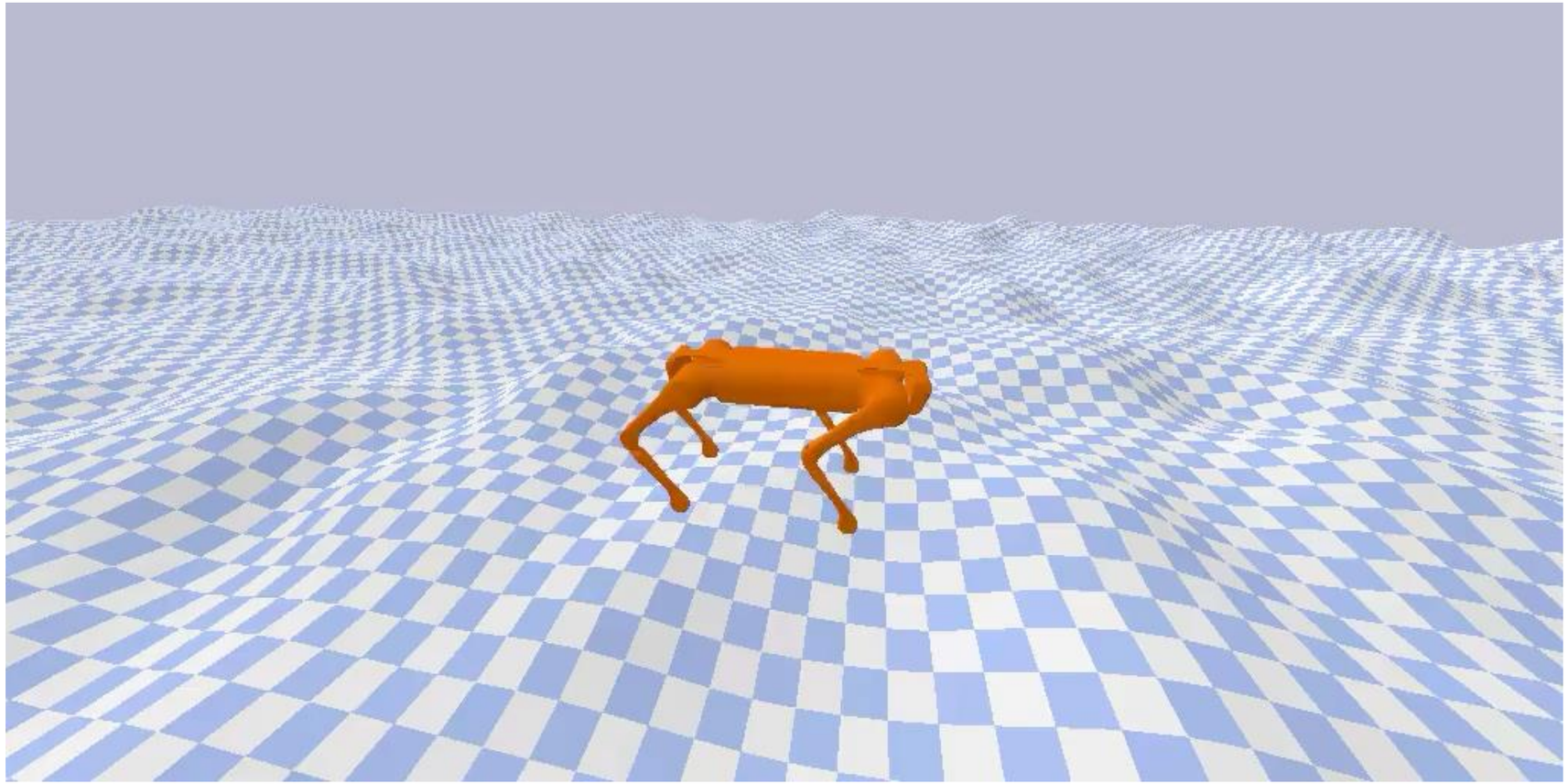
Ashish Kumar
Deepak Pathak
Stuart Anderson
Jitendra Malik

Uneven Terrain with Depth



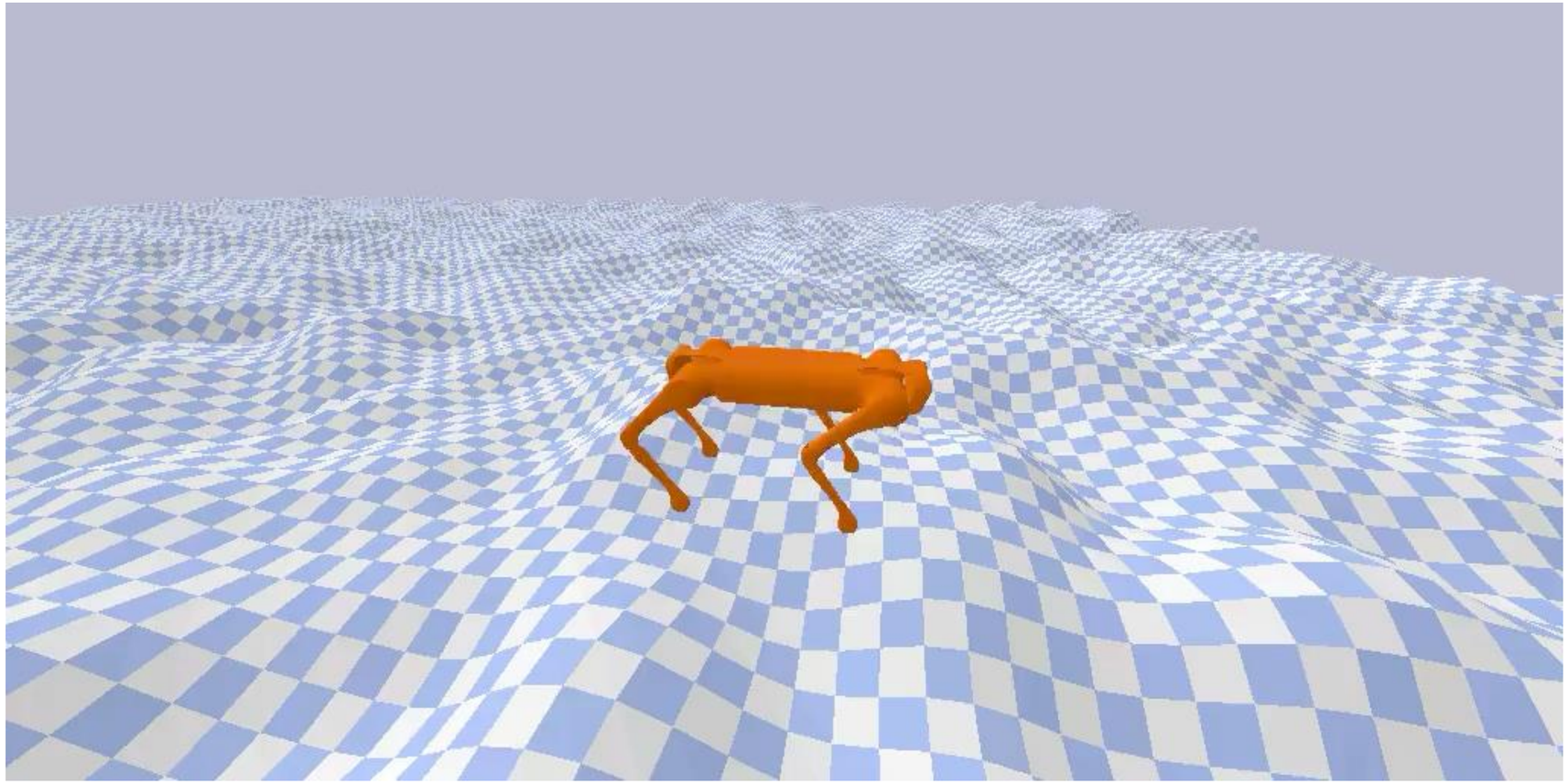
- Trained with PPO, small CNN (~4ms on a GPU)
- Rewards contain energy penalties inspired from Biomechanics

Uneven Terrain with Egocentric Depth



- Testing on Medium Difficulty

Uneven Terrain with Egocentric Depth



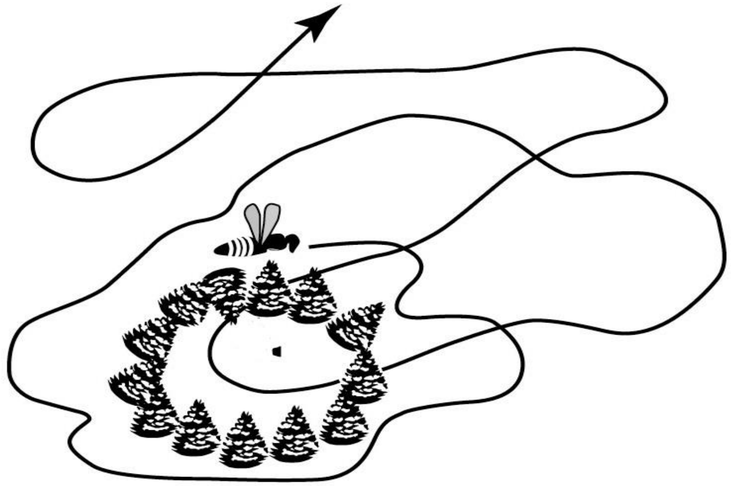
- Testing on High Difficulty

Latency in Vision

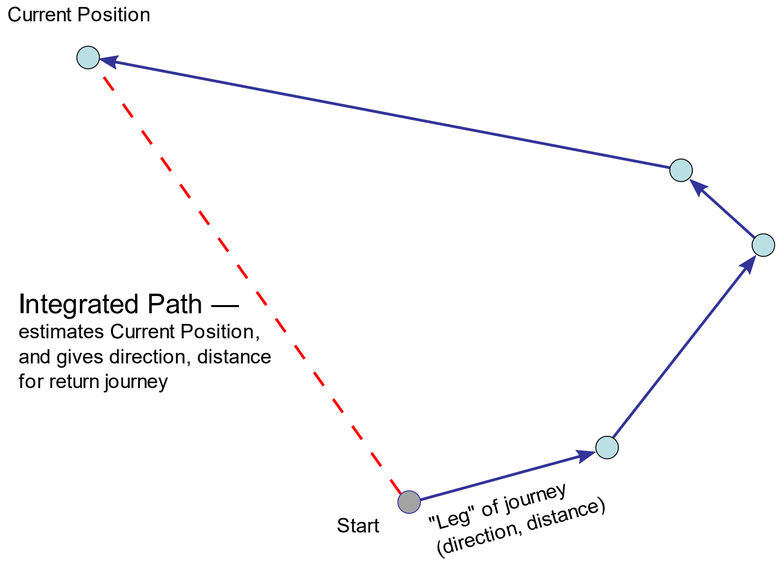
1. High Level footstep (low freq) planner and low level stepping function (high freq) — Train HL followed by LL / train LL followed by HL

2. Train with updating the vision features with less frequency

Animal Navigation: Landmarks and Maps



Niko Tinbergen (1951)



VOL. 55, NO. 4 JULY, 1948

THE PSYCHOLOGICAL REVIEW

COGNITIVE MAPS IN RATS AND MEN [1]

BY EDWARD C. TOLMAN

THE

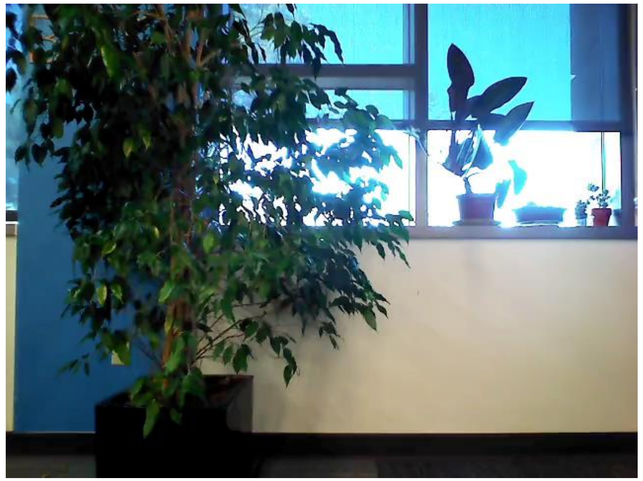
HIPPOCAMPUS

AS A COGNITIVE MAP

JOHN O'KEEFE
AND
LYNN NADEL

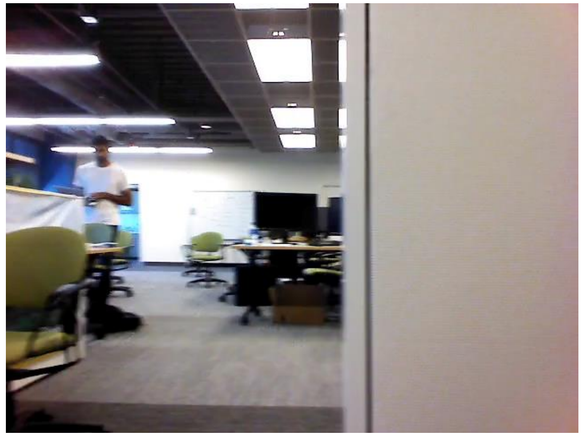
Learn Skills that Enable a Robot to Move Around in Novel Environments

Robot w/camera

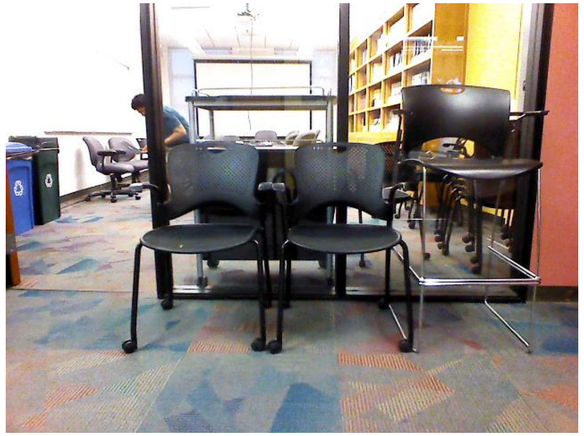


In novel environment

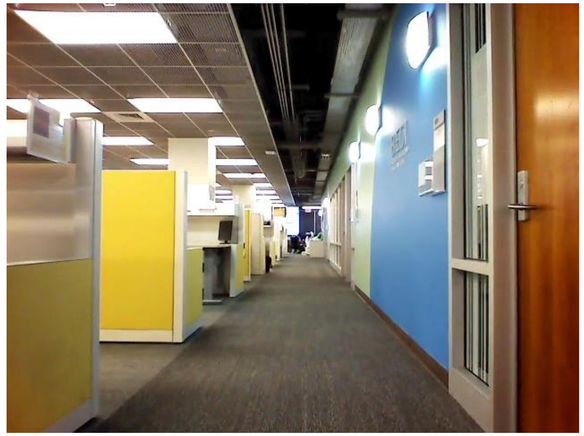
Go around obstacles



Come out of cubicles



Go through door



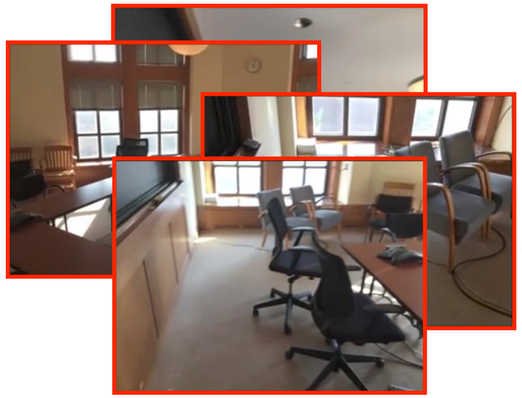
Go down hallway

The classical robotics solution is SLAM (Simultaneous Localization and Mapping)

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

Classical Mapping and Planning

3D Reconstruction

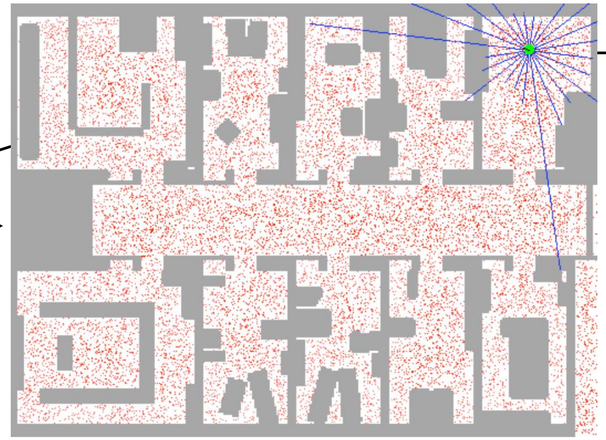


Observed Images

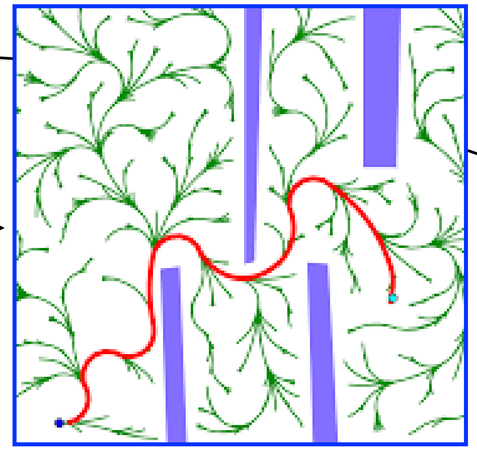
Mapping

3D Reconstruction

Policy Execution



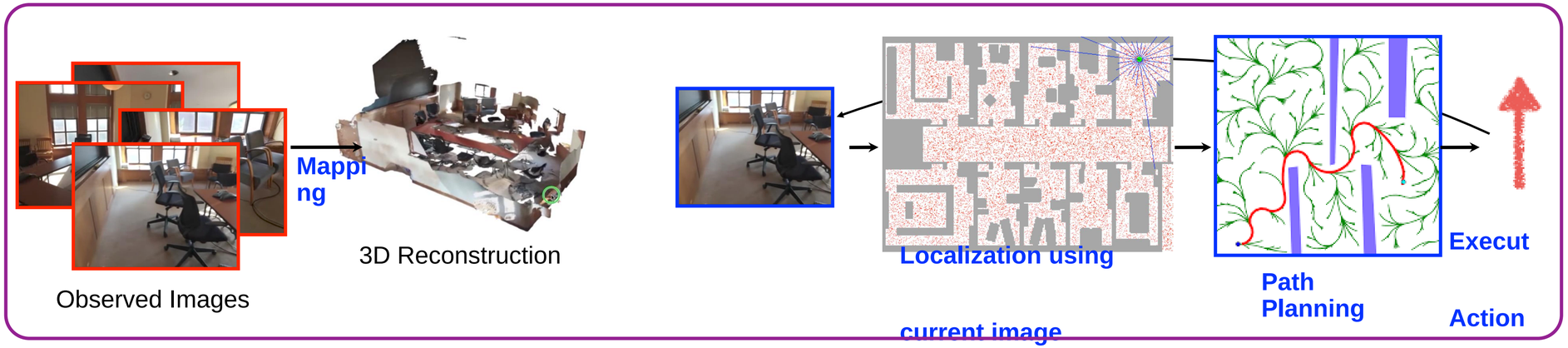
Localization using current image



Path Planning

Execute Action

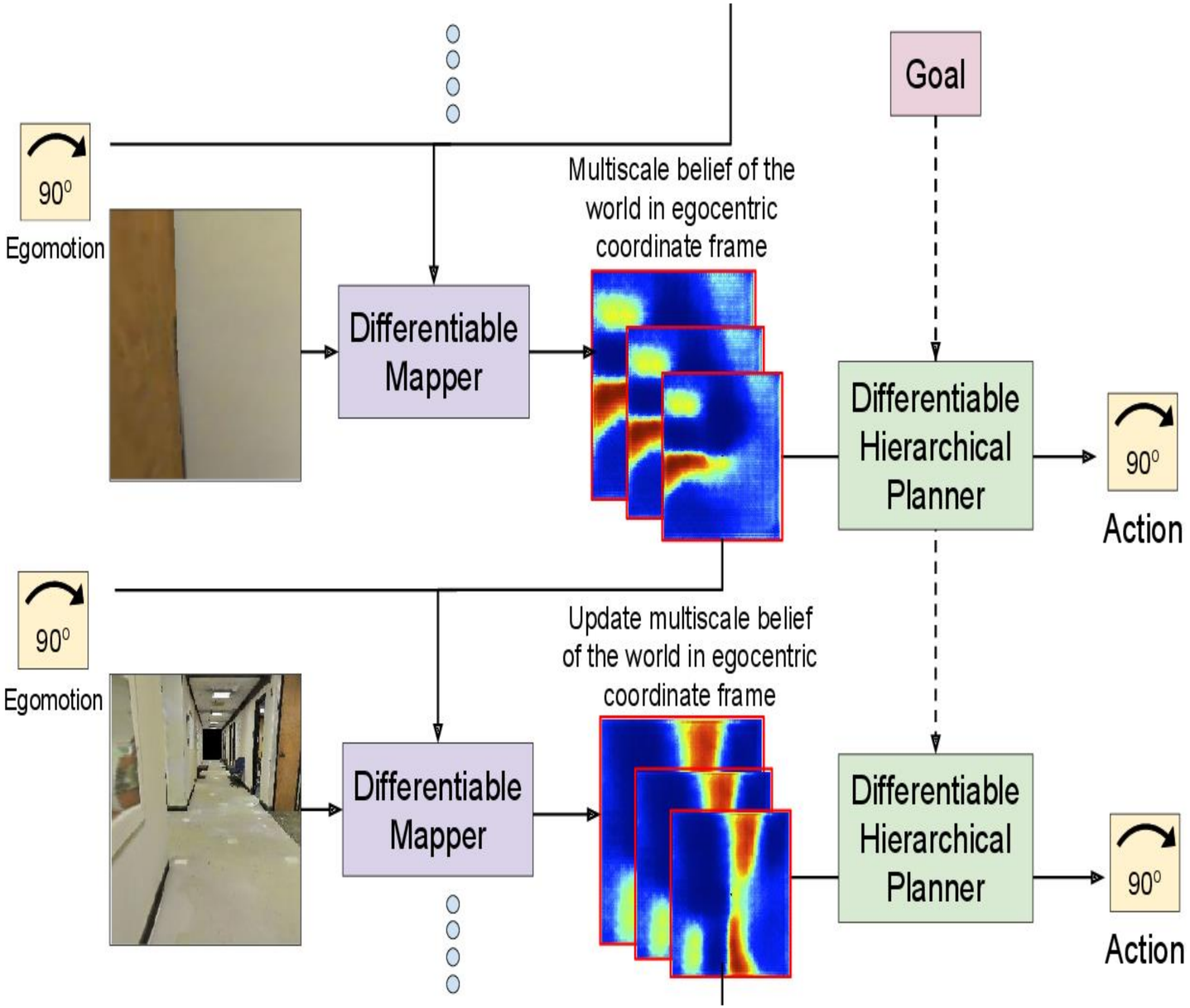
Classical Mapping and Planning



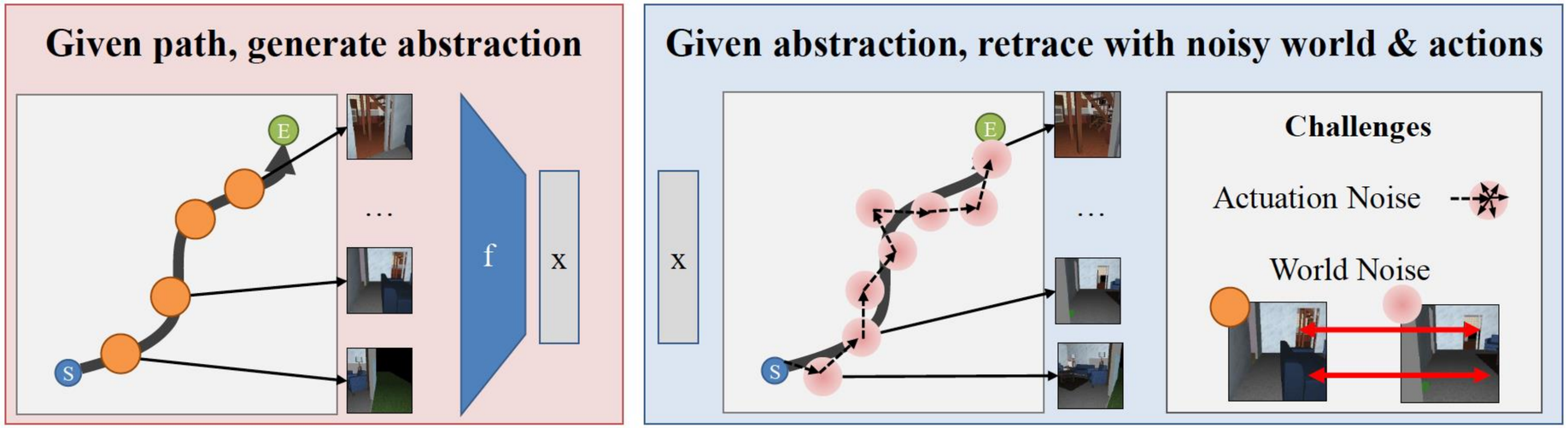
- **Unnecessary:**
  - Precise reconstruction of everything is not necessary.
  - Precise localization may also not be necessary.

- **Insufficient:**
  - Only geometry, no semantics.
  - Nothing is known till it is explicitly observed, failure to exploit experience with similar past environments.
  - Not robust to changes in the environment.
  - No way to encode semantic primitives (go down the hallway, go through the doorway).

Cognitive Mapping & Planning – Gupta et al (2017)



Visual Memory for Robust Path Following
Kumar, Gupta, Fouhey, Levine & Malik (2018)



Perception and Interaction

Language

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

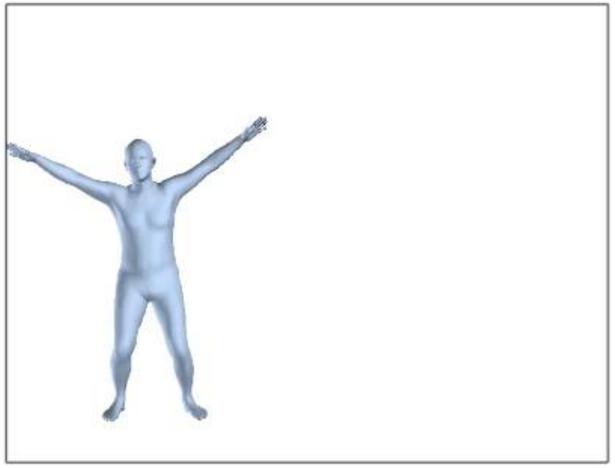
Model-free Reinforcement Learning has very high sample eomplexity

OpenAI et al, Learning Dexterous In-Hand Manipulation, arXiv 2018
OpenAI et al, Solving Rubik’s Cube with a Robot Hand, arXiv 2019

Learning by Imitating Others

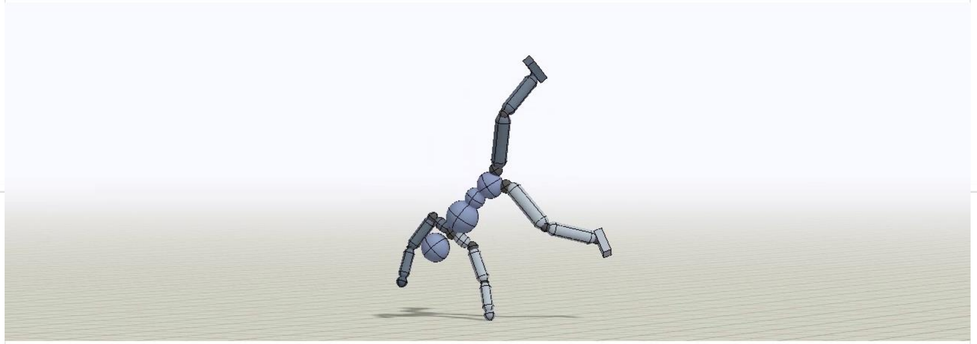
Reinforcement Learning of Skills from Videos : Peng, Kanazawa, Malik, Abbeel and Levine

Video: Cartwheel B



Reference Motion

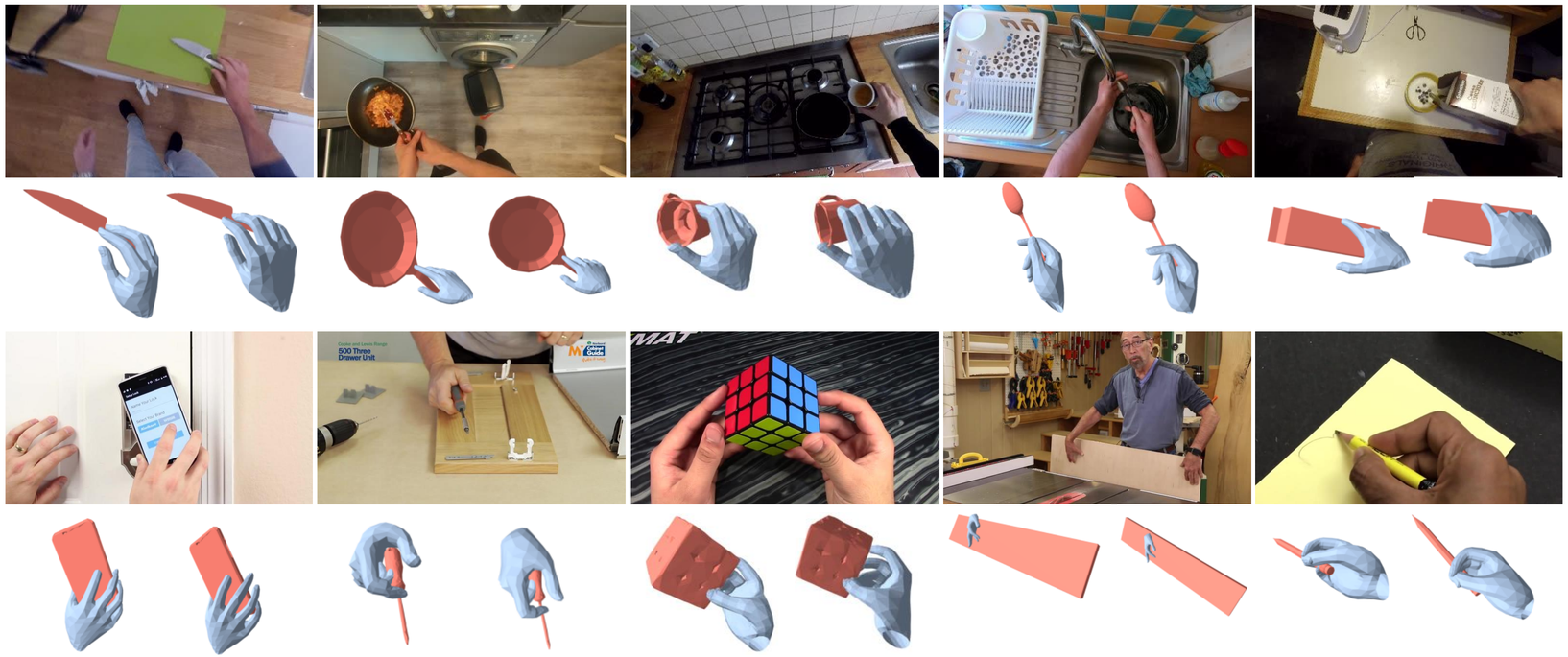
Policy



The policies learn to recover from significant perturbations.

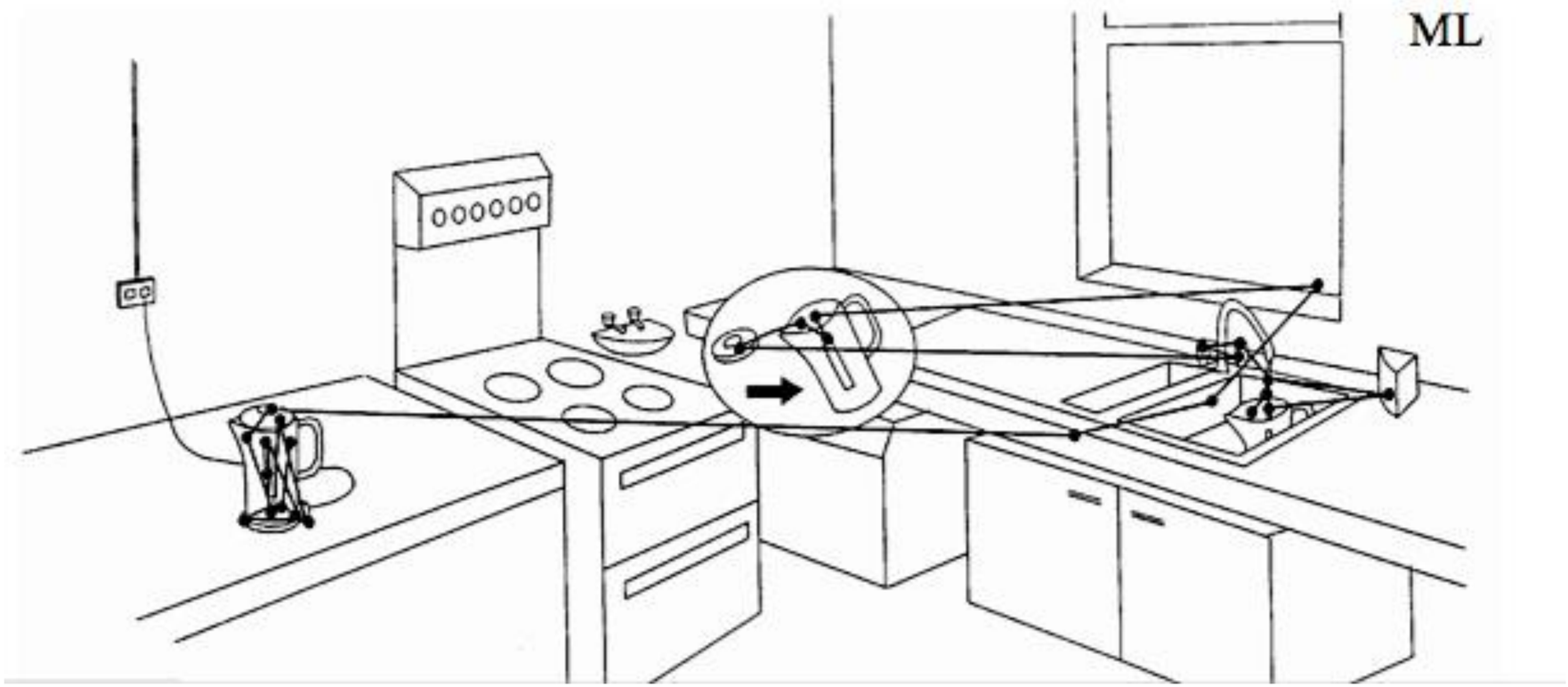
Human-Object Interactions in the Wild

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, Jitendra Malik



Eye Movements while making a cup of tea

Land, Mennie and Rusted (1999)



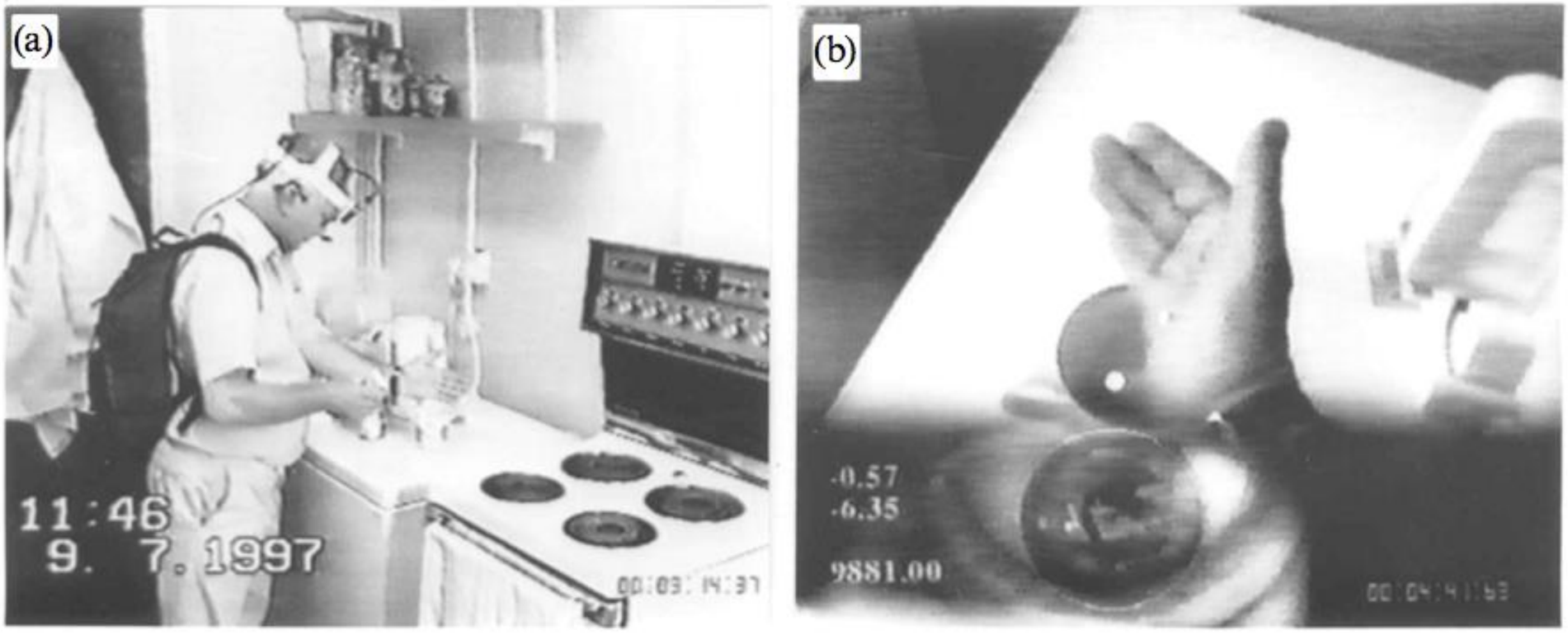
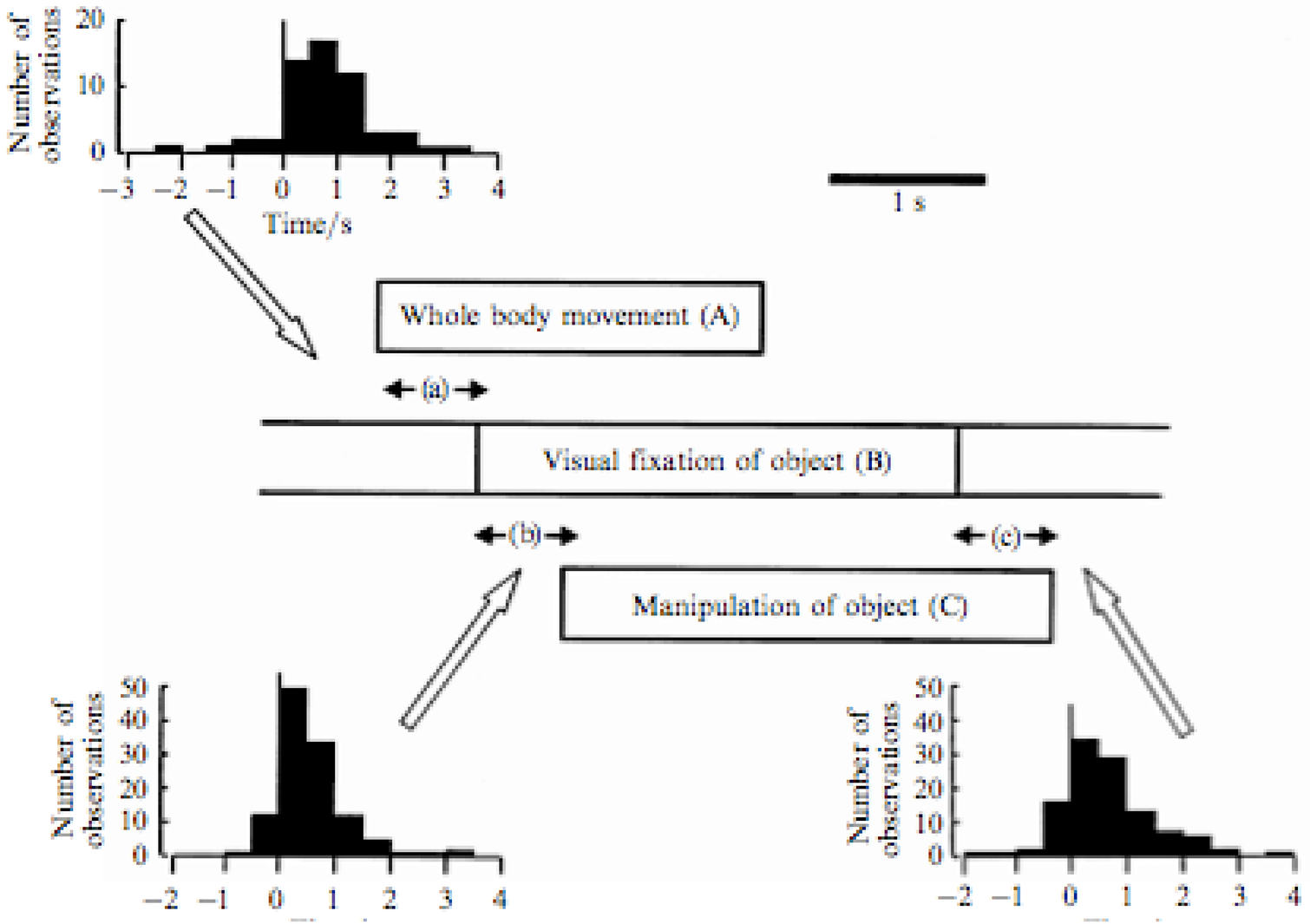


Figure 1. Prints from (a) the activity video, and (b) eye-movement video of the same instant, when the sweetener is dropped into the mug (3.14 on figure 3). The head-mounted camera and


Number of observations
20
10
0
−3 −2 −1 0 1 2 3 4
Time/s
1 s
Whole body movement (A)
(a)
Visual fixation of object (B)
(b)
(c)
Manipulation of object (C)
Number of observations
50
40
30
20
10
0
−2 −1 0 1 2 3 4
Number of observations
50
40
30
20
10
0
−2 −1 0 1 2 3 4

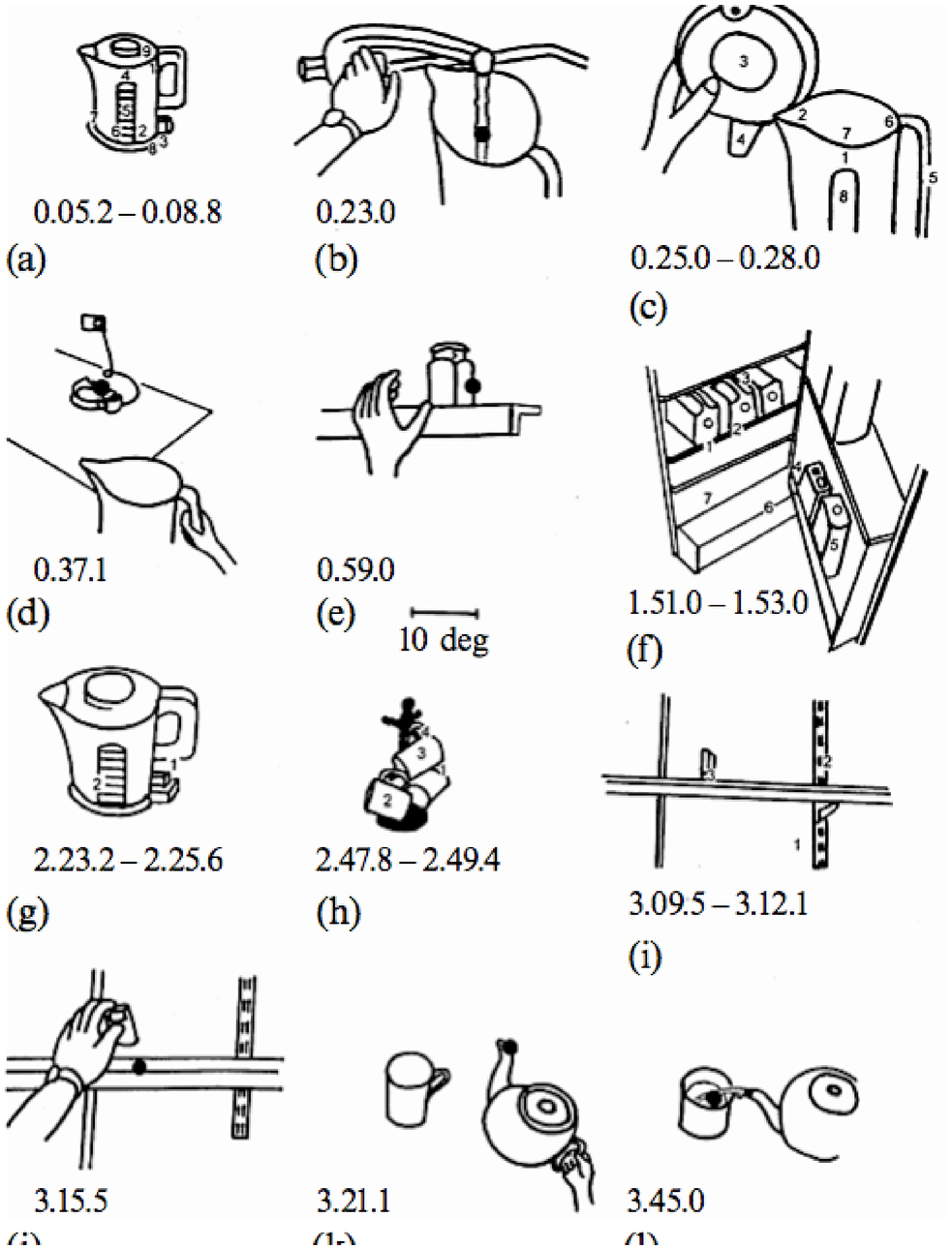
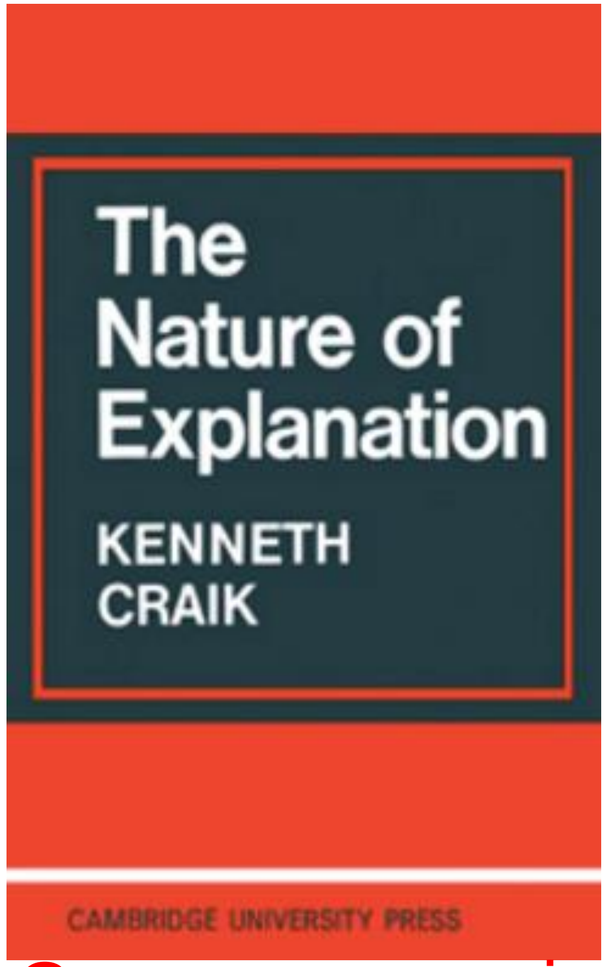


Figure 8. Examples of fixation patterns drawn from the eye-movement videotape. Sequences of successive fixation positions are indicated by numbers on the figures, and single fixations by single black dots. Numbers beneath each figure refer to timings in figure 3. 10 deg scale in centre applies to all figures. (a) Initial examination of kettle. (b) Tap control via water stream. (c) Fitting lid to kettle (drawing made at fixation 4). (d) Moving kettle to base: base is fixated. (e) Hand being directed to the tea-caddy. (f) Search around the inside of fridge 2. The tea-making milk is located at fixation 5. (g) Fixations checking the switch and gauge of the kettle when waiting for it to boil. (h) Selecting a mug. Hand goes to fixation 4. (i) Relocating sweetener prior to use requires 3 fixations. Sweetener last seen 68 s earlier. (j) Replacing sweetener 5 s after (i). Location on shelf is fixated first. (k) Swirling teapot: checking spout. (l) Pouring tea: receiving vessel fixated.

AI systems need to build “mental models”



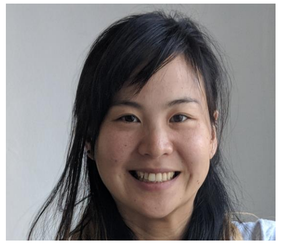
If the organism carries a `small-scale model’ of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present 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)

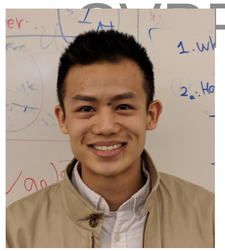
Commonsense is not just facts, it is a collection of models

Learning 3D Human Dynamics from Video

Angjoo Kanazawa*, Jason Zhang*, Panna Felsen*, Jitendra Malik

CVPR 2019



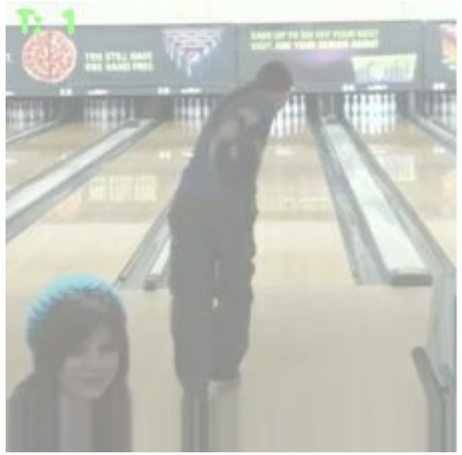


* Equal contribution

Auto-regressive prediction of 3D motion from video

Input

Video



Ground Truth

Video

Predicted Future

Different

Viewpoint

Predicting 3D Human Dynamics from Video, Zhang, Felsen, Kanazawa, Malik(ICCV 2019)

Input
Video