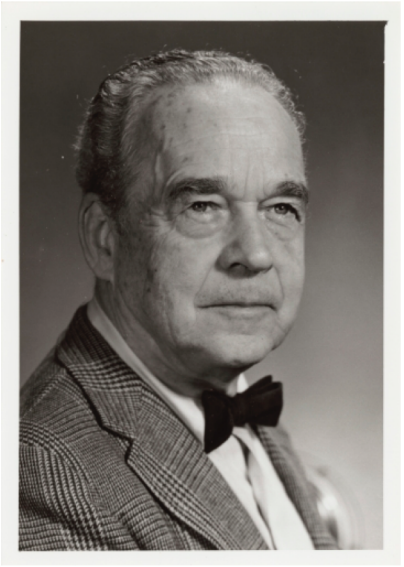
Theories of perception

Giorgione, The Three Philosophers, c. 1505

David Marr (1945-1980)



James Jerome Gibson (1904-1979)

Jan Johan Koenderink (b. 1943)

David Marr (1945-1980)

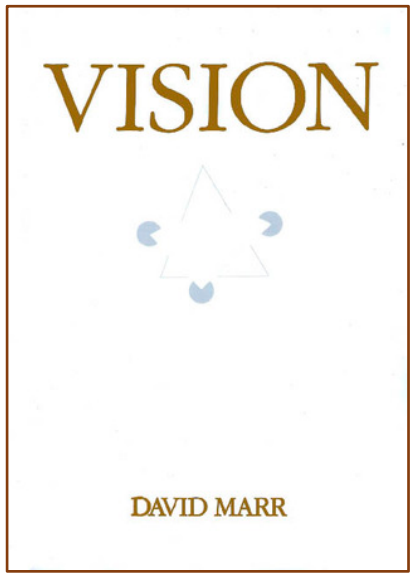
- Ph.D. in theoretical neuroscience, Cambridge, 1969
  - Models of the cerebellum (1969), neocortex (1970), hippocampus (1971)
- Joined MIT AI Lab in 1973, became professor of psychology in 1977
  - Stereo algorithms (with Tommaso Poggio), 1976-79
  - 3D object representation (with Keith Nishihara), 1978
  - Edge detection (with Ellen Hildreth), 1980

Bio

Marr's Vision

- Posthumous book: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information (1982)

In December 1977, certain events occurred that forced me to write this book a few years earlier than I had planned. Although the book has important gaps, which I hope will soon be filled, a new framework for studying vision is already clear and supported by enough solid results to be worth setting down as a coherent whole.



Full text

Marr's motivation (ch. 1)

- Vision is *hard*

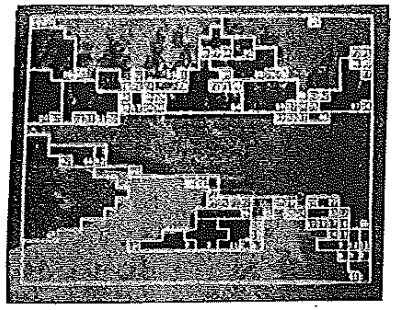
> The first great revelation was that the problems are difficult. Of course, these days this fact is a commonplace. But in the 1960s almost no one realized that machine vision was difficult. The field had to go through the same experience as the machine translation field did in its fiascoes of the 1950s before it was at last realized that here were some problems that had to be taken seriously. The reason for this misperception is that we humans are ourselves so good at vision. The notion of a feature detector was well established by Barlow and by Hubel and Wiesel, and the idea that extracting edges and lines from images might be at all difficult simply did not occur to those who had not tried to do it. It turned out to be an elusive problem: Edges that are of critical importance from a three-dimensional point of view often cannot be found at all by looking at the intensity changes in an image. Any kind of textured image gives a multitude of noisy edge segments; variations in reflectance and illumination cause no end of trouble; and even if an edge has a clear existence at one point, it is as likely as not to fade out quite soon, appearing only in patches along its length in the image. The common and almost despairing feeling of the early investigators like B.K.P. Horn and T.O. Binford was that practically anything could happen in an image and furthermore that practically everything did.

Marr's motivation (ch. 1)

- Vision is *hard*
- We may not be able to figure out the right solution right away, but at least we should start by establishing a sound methodology
  - Marr explicitly considered and rejected low-level neurophysiology, empirical "hacking", and blocks world simplification

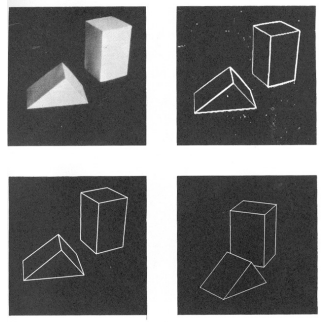
Hubel & Wiesel (1959)

(source)



(B-2) Output of the non-semantic weakest boundary melted first region grower.

Yakimovsky & Feldman (1973)



Roberts (1963)

Towards an *information processing* theory of vision

Computational theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Figure 1–4. The three levels at which any machine carrying out an information-processing task must be understood.

Computational theory description of vision

- What should be the goal of vision?
  - “Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information.”
- What should be the input?
  - “In the case of human vision, the initial representation is in no doubt – it consists of **arrays of image intensity values** as detected by the photoreceptors in the retina.”
- What should be the output?
  - ”[The purpose of vision is] **building a description of the shapes and positions of things** from images… It also tells about the illumination and about the reflectances of the surfaces that make the shapes – their brightnesses and colors and visual textures – and about their motion. But these things seemed secondary; they could be hung off a theory in which **the main job of vision was to derive a representation of shape**.”

Proposed algorithmic pipeline

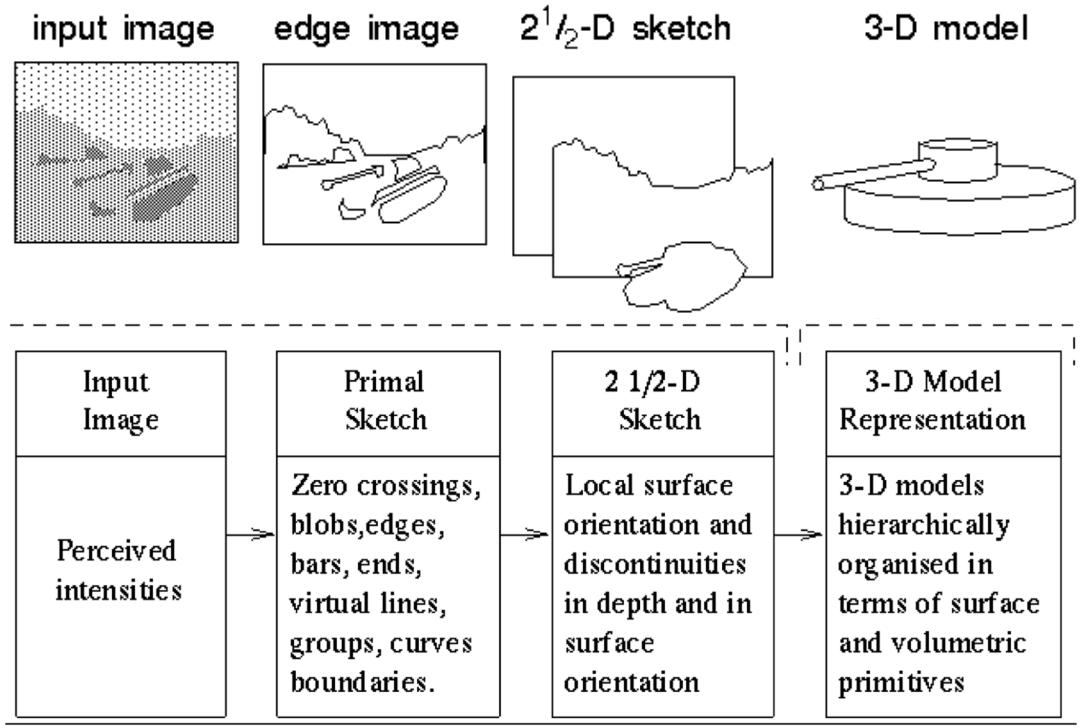


Image source

So, what's the big deal?

- Marr's book was a major milestone
  - Critical summary of key developments in study of human and computer vision to date
  - Unprecedented attempt at a unified account of the entire visual system
- Computational framework was very appealing to computer vision researchers from a "software engineering" perspective
  - Abstraction, modularity, feedforward pipeline
- Theories meshed well with the dominant computer vision paradigms
  - Vision as "inverse graphics" or "inverse optics"
  - Emphasis on recovery of general-purpose 3D representations composed of simple geometric primitives
  - Convenient division of vision problems into "low-level", "mid-level", and "high-level"

Special issue dedicated to Marr: Perception 41(9), 2012

What about the bad stuff?

- None of the particulars of Marr's approach have panned out either on the human or the computer vision side

Figure 3–1. The interpretation of some images involves more complex factors as well as more straightforward visual skills. This image devised by R. C. James may be one example. Such images are not considered here.

What about the bad stuff?

- None of the particulars of Marr's approach have panned out either on the human or the computer vision side
- Principles of modularity and feedforward processing don't hold for human vision
  - P. Churchland, V.S. Ramachandran, and T. Sejnowski, A critique of pure vision, 1994
- Humans do not recover veridical, task-independent 3D representations
  - W. Warren, Does This Computational Theory Solve the Right Problem? Marr, Gibson, and the Goal of Vision, Perception 41(9), 2012
- Marr dismissed statistical approaches, did not even consider learning
- Even the goals, inputs, and outputs of a vision system are very much open to debate (as discussed next)

James Jerome Gibson (1904-1979)

- Ph.D. in psychology, Princeton, 1928
- Taught at Smith college, served in the Aviation Psychology Program during WWII, then taught at Cornell
- Books:
  - *The Perception of the Visual World* (1950)
  - *The Sense Considered as Perceptual Systems* (1966)
  - *The Ecological Approach to Visual Perception* (1979)

Bio

“Ecological optics” doctrine

(Based on *The Ecological Approach to Visual Perception,* 1979)

- Perception must be studied in the context of an organism’s environment and biological function
  - “To perceive is to be aware of the **surfaces of the environment** and **oneself** in it… The full awareness of surfaces includes their **layout**, their **substances**, their **events**, and their **affordances**.”
- Perception is **embodied** and **active**, its key goal is **control of behavior**
- To understand perception, one must use **ecological physics** (optics, geometry, etc.), i.e., concepts for understanding the environments of animals and people that are relevant for behavior
  - E.g., absolute space and time are ecologically meaningless

“Ecological optics” doctrine (cont.)

- Percetion starts not with the “retinal image”, but with the **ambient optic array**
  - Gibson regards “retinal image” as a harmful fiction. This image (if it even exists) changes constantly, whereas our awareness of the visual world is stable and unchanging
  - “To be an array means to have an arrangement, and to be ambient at a point means to surround a position in the environment that could be occupied by an observer.”
  - The ambient optic array is **structured** into nested components corresponding to distinct parts of the environment (roughly speaking, “objects”)

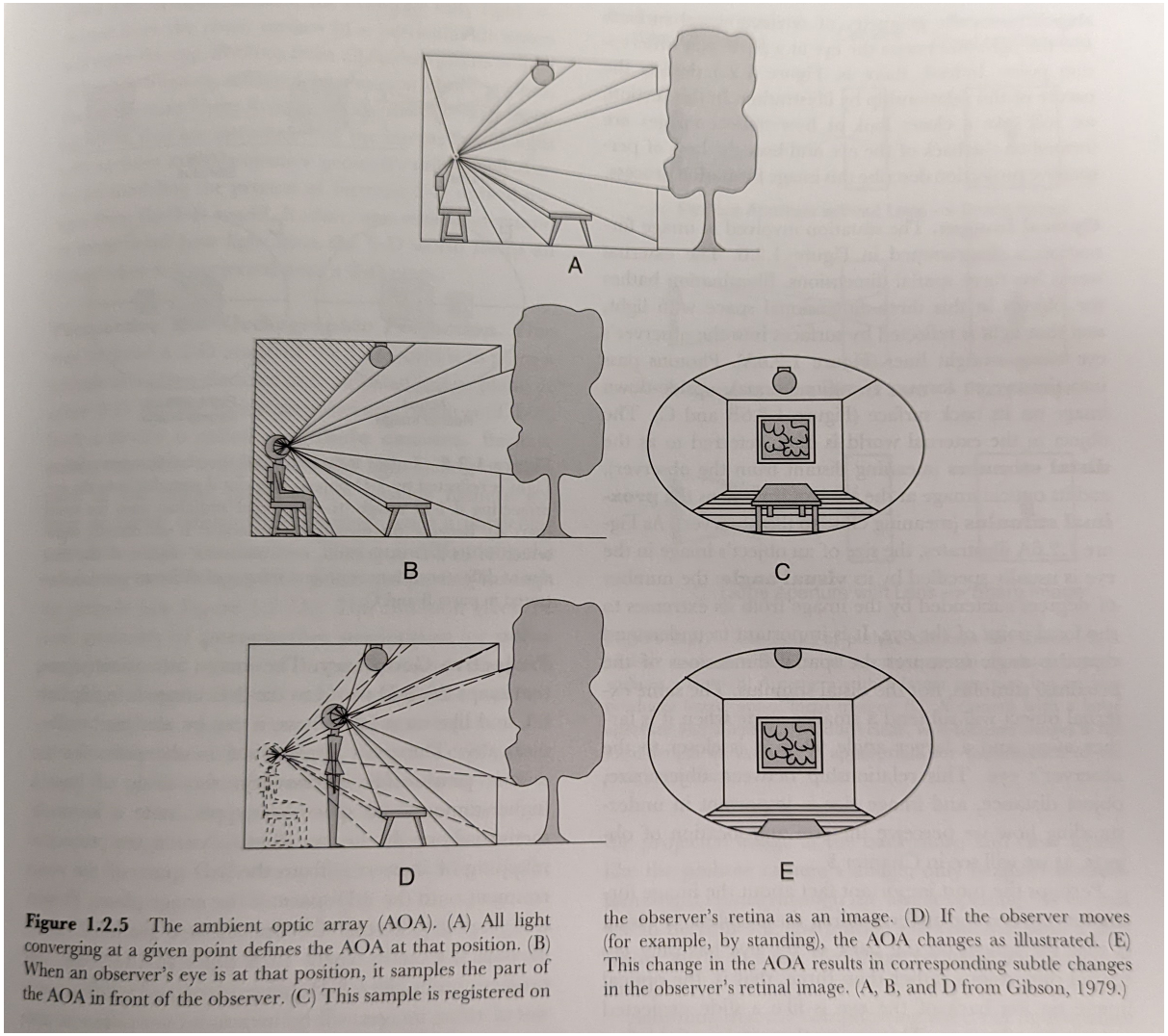


Figure 1.2.5 The ambient optic array (AOA). (A) All light converging at a given point defines the AOA at that position. (B) When an observer's eye is at that position, it samples the part of the AOA in front of the observer. (C) This sample is registered on the observer's retina as an image. (D) If the observer moves (for example, by standing), the AOA changes as illustrated. (E) This change in the AOA results in corresponding subtle changes in the observer's retinal image. (A, B, and D from Gibson, 1979.)

Source: S. Palmer, Vision Science

“Ecological optics” doctrine (cont.)

- Percetion starts not with the “retinal image”, but with the **ambient optic array**
- Perception happens by **direct information pickup**, or “the concurrent registering of both persistence and change in the flow of structured stimulation”
  - “Direct” means not mediated by information processing or internal representations: “The perceptual system simply extracts the **invariants** from the flowing array; it *resonates* to the invariant structure or is *attuned* to it”
- For an active observer, perception is **mostly unambiguous**
  - Gibson views the case of “monocular arrested vision” as “unnatural” and dismisses illusions that arise from it

Affordances

- The goal of perception is providing the agent with information relevant for control of behavior, encapsulated in **affordances**
  - "The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill. The verb to afford is found in the dictionary, but the noun affordance is not. I have made it up."

"throwable"

"sittable-upon"

"drinkable-from"

Source: S. Palmer via A. Efros

Affordances

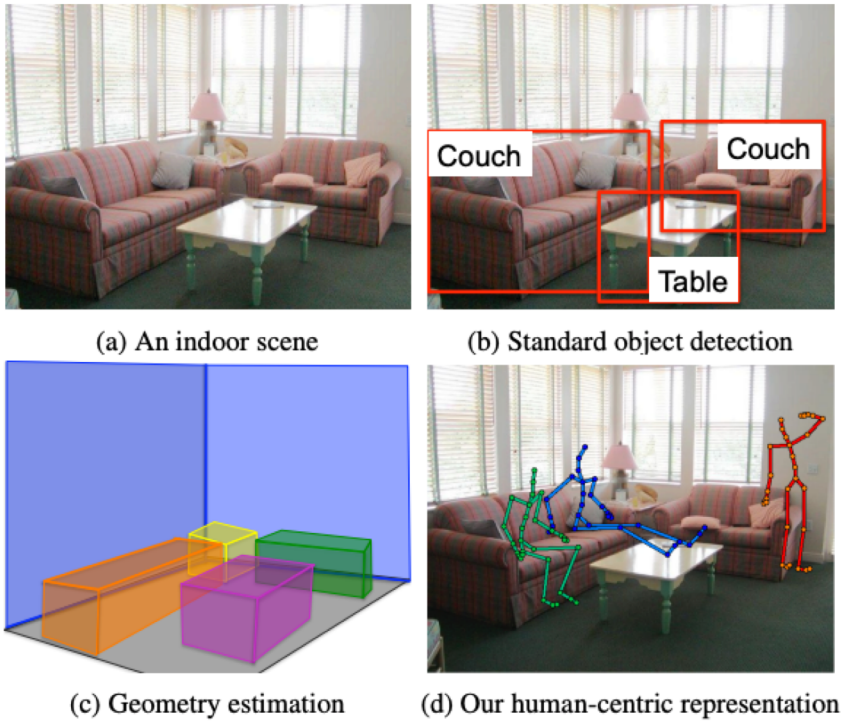
- The goal of perception is providing the agent with information relevant for control of behavior, encapsulated in **affordances**
  - "The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill. The verb to afford is found in the dictionary, but the noun affordance is not. I have made it up."
- Affordances are **intrinsic invariants** that can be had by any feature of the environment (place, object, surface, substance, or event)
  - "A fire affords warmth on a cold night; it also affords being burnt. An approaching object affords either contact without collision or contact with collision; a tossed apple is one thing, but a missile is another. For one of our early ancestors, an approaching rabbit afforded eating whereas an approaching tiger afforded being eaten."

Perception of affordances vs. categorization

- “The perceiving of an affordance is not a process of perceiving a value-free physical object to which meaning is somehow added in a way that no one has been able to agree upon; it is a process of perceiving **a value-rich ecological object**.”
- “The theory of affordances rescues us from the philosophical muddle of assuming fixed classes of objects, each defined by its common features and then given a name… **You do not have to classify and label things in order to perceive what they afford.**”

Are affordances all we need?

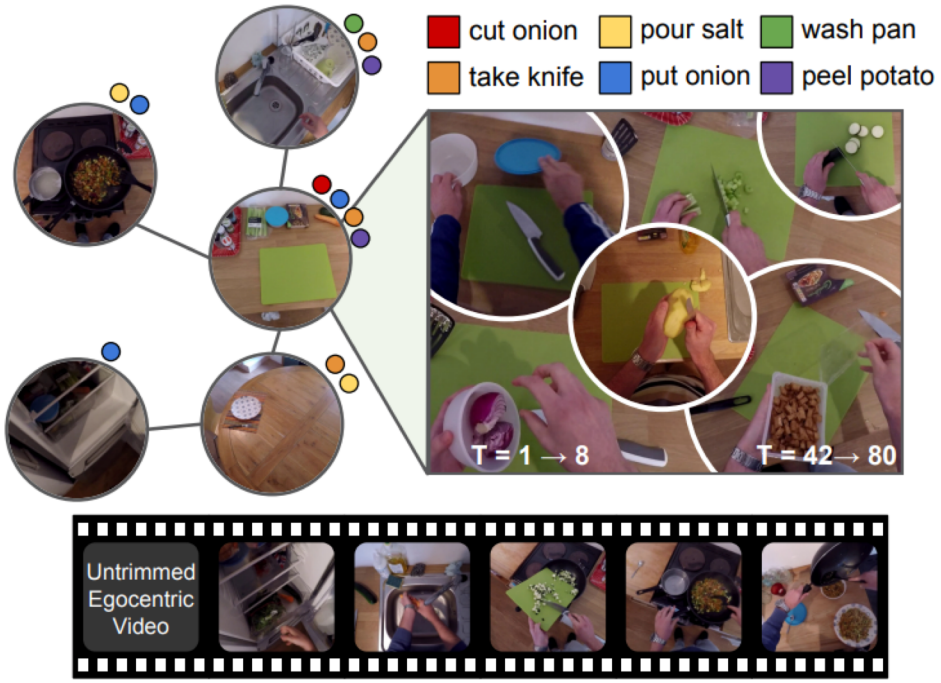
Affordances in computer vision



(a) An indoor scene (b) Standard object detection

(c) Geometry estimation (d) Our human-centric representation

Gupta et al. (2011)



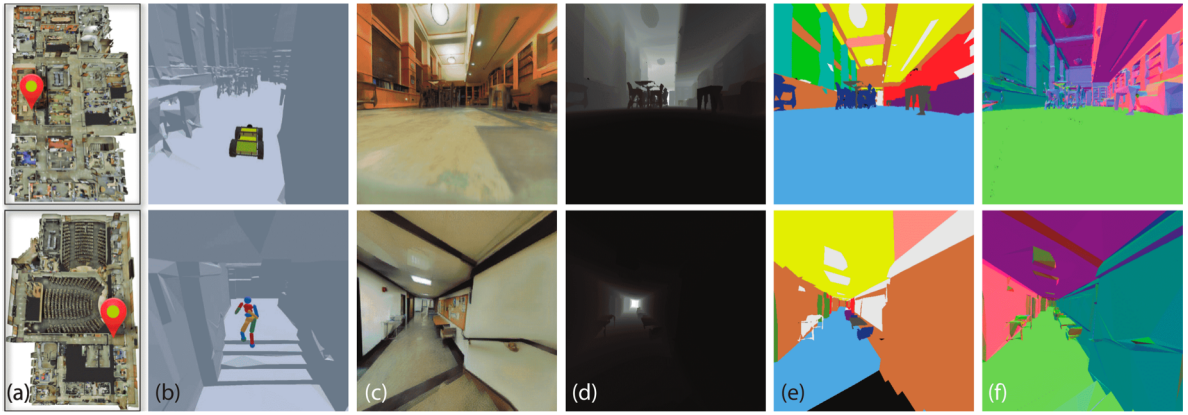
Nagarajan et al. (2019)

Gibson: What's the big deal?

- Emphasized the role of an active, embodied observer, and aspects of environment relevant for behavior
  - Warned of limitations of "snapshot vision"
  - Attached primary importance to motion and control
  - Pointed out that recovery of persistence and change, or world and observer, are two sides of the same coin
- Direct perception is not a completely crazy idea: In many cases, cues relevant for action can indeed be perceived without going through a full general-purpose visual pipeline
- Concept of affordances proved very influential

Gibson's legacy

- Same with his ideas about active perception and control
  - S. Soatto, Actionable information in vision, 2010
  - F. Xia et al., Gibson Env: Real-World Perception for Embodied Agents, CVPR 2018



"We must perceive in order to move, but we must also move in order to perceive" – J. J. Gibson

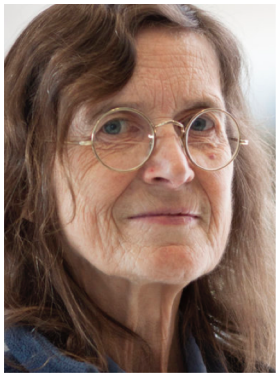
What about the bad stuff?

- Gibson completely rejected questions of representation and information processing
- According to Marr, Gibson was “misled by the apparent simplicity of the act of seeing” and seriously underestimated the complexity of extracting invariants
- Viewed perception as primarily a function of the environment and downplayed the role of the observer
- Paid only cursory lip service to learning

Jan Johan Koenderink (b. 1943)

Bio

- Ph.D. 1972, Utrecht University
- Professor of physics and astronomy at Utrecht University, 1978-2008
- Contributions (many with Andrea van Doorn)
  - Motion and optical flow (1975, 1976)
  - Stereopsis (1976)
  - Aspect graphs (1976, 1979)
  - Scale space theory (1984)
  - Properties of smooth 3D shapes and 2D contours (1982, 1984, 1992)
  - Affine structure from motion (1991)
  - Local grayvalue invariants (1987, 1994)
  - 3D shape perception (1992, 1993, 1996)
  - Surface reflectance (1983, 1998, 1999)
  - Perception and art (2015)



Andrea van Doorn

2D contour and 3D shape

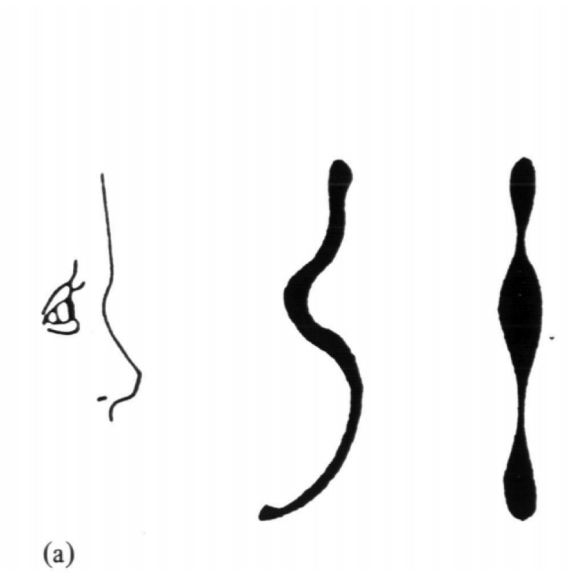


Figure 3. (a) A figure taken from Marr (1982). The suggestion is that convexities and concavities in the projection of the snake have to do with relative *distances* rather than with local shapes.

J. Koenderink. What does the occluding contour tell us about solid shape? Perception 13 (321-330), 1984

2D contour and 3D shape

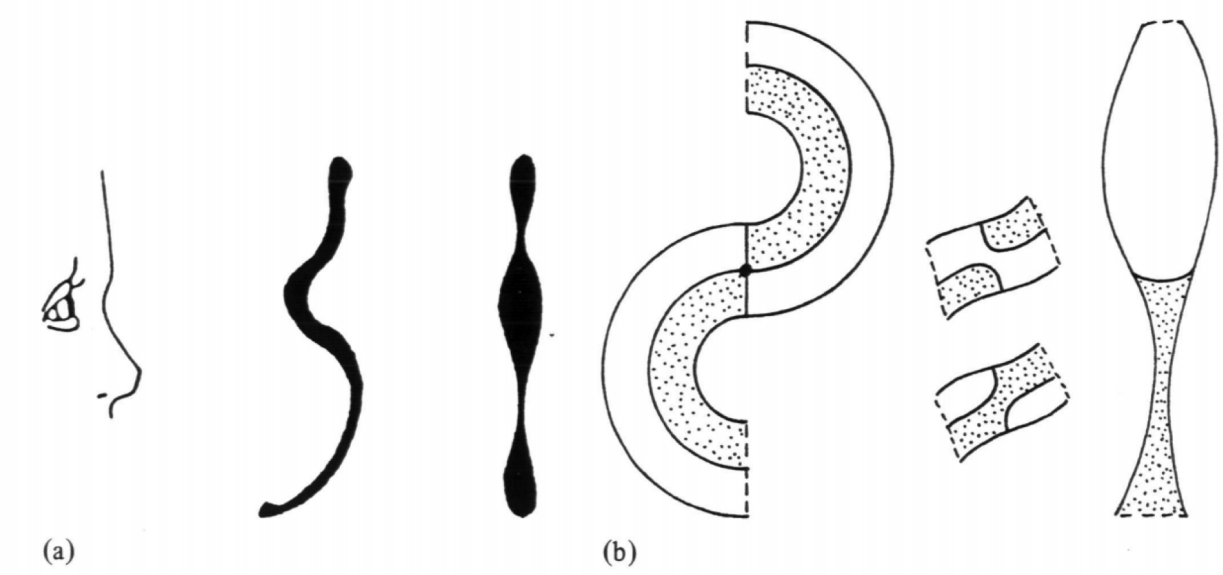


Figure 3. (a) A figure taken from Marr (1982). The suggestion is that convexities and concavities in the projection of the snake have to do with relative *distances* rather than with local shapes. (b) A torus cut into two and pasted together again. The shaded regions are anticlastic, the other regions synclastic. The small insets show the generic case after a small deformation. In projection (right), this 'snake' has convexities where the body is locally egg-shaped, concavities where the body is locally saddle-shaped, inflexions at flexional curves of the body.

J. Koenderink. What does the occluding contour tell us about solid shape? Perception 13 (321-330), 1984

2D contour and 3D shape

Figure 4. Details from Dürer's "Samson killing the lion". (Bartsch #2; the print dates from 1498.)

J. Koenderink. What does the occluding contour tell us about solid shape? Perception 13 (321-330), 1984

Locally orderless images

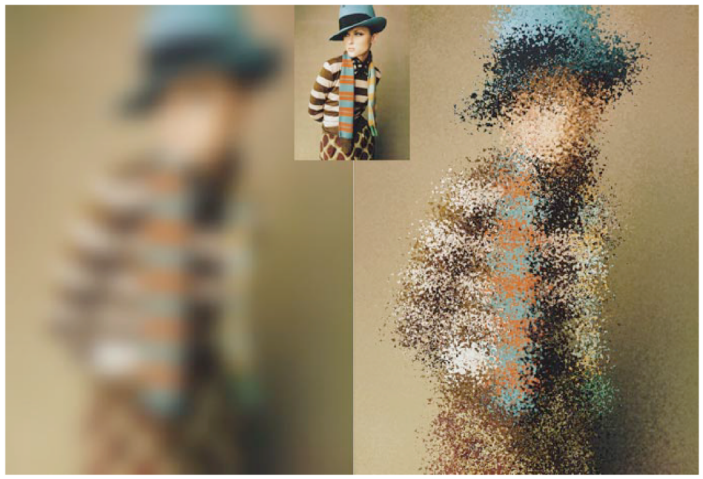


Fig. 11. A fashion image (small inset) blurred (left) and disordered (right) by the same amount. Notice that the spatial resolutions of the blurred and the disordered image are indeed similar, but that the disordered image has retained pixel values that are lost in the blurred image where they were averaged out. Though "unsharp", the disordered rendering has thus retained a vestige of image structure at the original scale. Aesthetically (as well as information technically) the disordered rendering is much to be preferred over the blurred image, in fact it is not unlike impressionist renderings.

J. Koenderink and A. van Doorn, The structure of locally orderless images, IJCV 31 (159-168), 1999

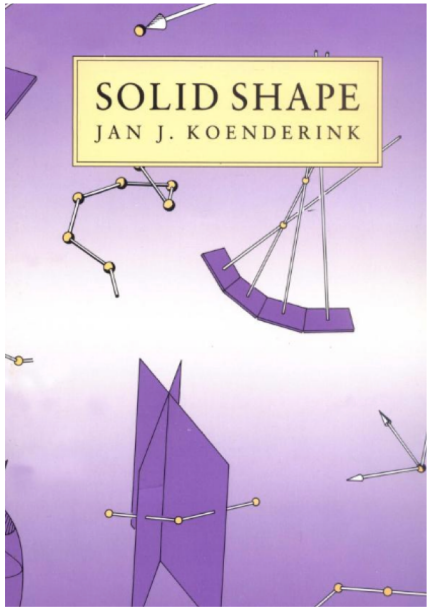
Locally orderless images

Instances of locally orderless *perceptions* are quite frequently encountered in various contexts. Cases of *amblyopia* ("lazy eye") have been described (Hess 1982) in which the observer is able to distinguish fine black and white stripes from uniform gray but cannot distinguish vertical from horizontal stripes or read text at a similar level of resolution. This condition has been termed "tarachopia" or scrambled vision. It seems likely that the peripheral visual field of *normal* observers has a similar locally orderless structure (Metzger 1975), and so has the central visual field for finest details (Helmholtz 1866). That such cases are *typical* of normal perception, rather than the exception, has been forcefully argued on phenomenological grounds by Ruskin (1857 and 1873), for instance (Ruskin 1873):

> *"Go to the top of Highgate Hill on a clear summer morning at five o'clock, and look at Westminster Abbey. You will receive an impression of a building enriched with multitudinous vertical lines. Try to distinguish one of these lines all the way down from the next to it: You cannot. Try to count them: You cannot. Try to make out the beginning or end of any of them: You cannot. Look at it generally, and it is all symmetry and arrangement. Look at it in its parts, and it is all inextricable confusion."*

J. Koenderink and A. van Doorn, The structure of locally orderless images, IJCV 31 (159-168), 1999

Books



1991

2010

See also: E-Books

Koenderink: What's the big deal?

- Style of work is that of a mathematical theorist of perception, starting with some visual phenomenon and creating an elegant mathematical formalization to describe it
- Some of these formalizations have been quite influential and even useful when translated into more accessible terms and suitably operationalized
- Research is guided by a strong sense of taste and aesthetics

A grand theory of perception?

J. Koenderink, Sentience, 2019

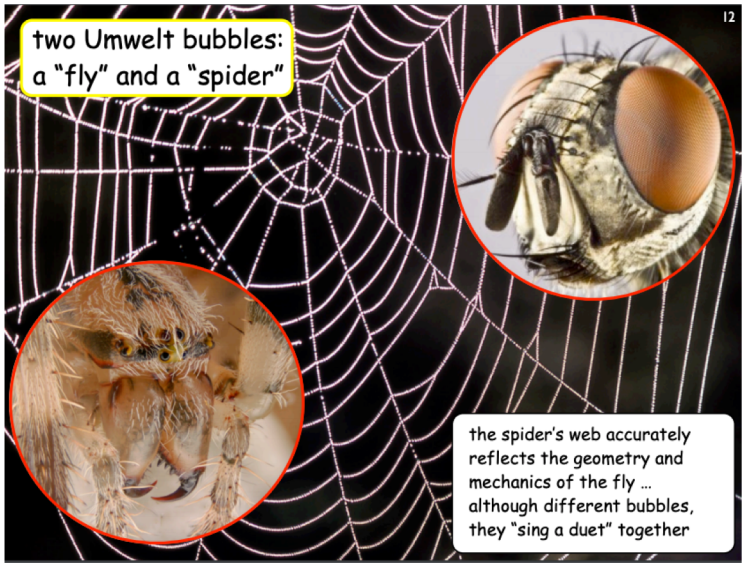
A grand theory of perception?

- Heavily influenced by Jakob von Uexküll
  - German biologist, 1864-1944

J. Koenderink, Sentience, 2019

Sensory-action worlds

- Each organism has its own *umwelt* or "surrounding world"
  - This is the organism's sensory and action world. It is determined by biology "bounds the universe from the perspective of the animal"



J. Koenderink, Sentience, 2019

Source: Koenderink's slides

Sensory-action worlds

- Each organism has its own *umwelt* or “surrounding world”
  - This is the organism’s sensory and action world. It is determined by biology “bounds the universe from the perspective of the animal”

“Indra's net” illustrates Leibniz’s monadology: the “monads have no windows”, but each reflects all others in pre-established harmony…

von Uexküll’s “Umwelts” also mesh - harmony is not pre-established by a Creator, but von Uexküll does perceive overall structure, not a chaos

J. Koenderink, Sentience, 2019

Source: Koenderink's slides

The AI viewpoint

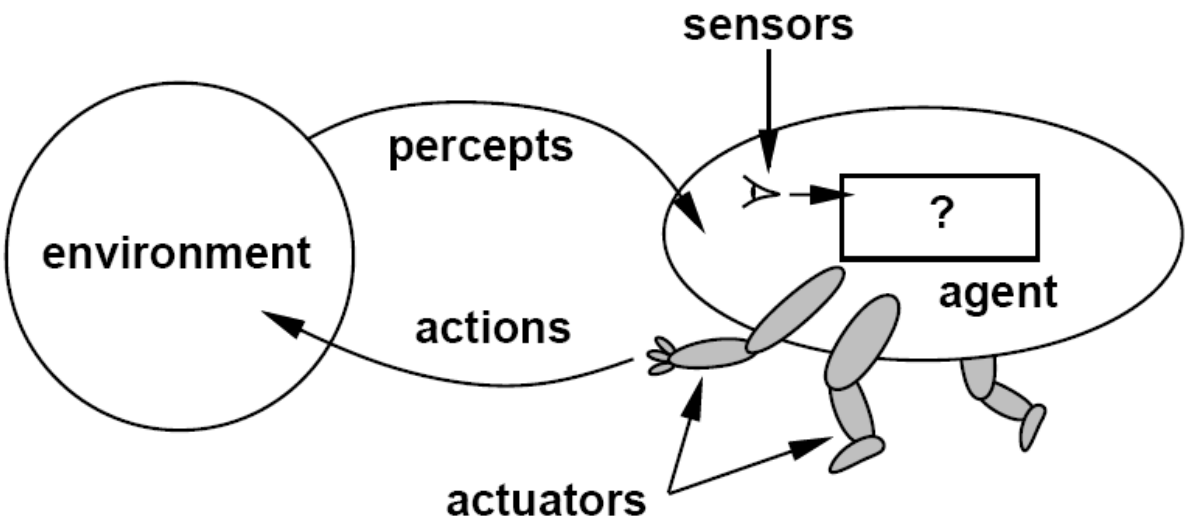
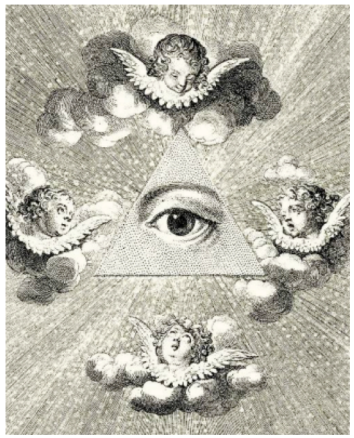


Figure from Russell & Norvig

Sensory-action worlds

- Each organism has its own *umwelt* or "surrounding world"
  - This is the organism's sensory and action world. It is determined by biology "bounds the universe from the perspective of the animal"
  - Absolute time and space don't exist from the organism's point of view
  - Gibson had a similar idea, but he still implicitly assumed a "God's eye" view that Koenderink rules out



J. Koenderink, Sentience, 2019

Perception-action cycles

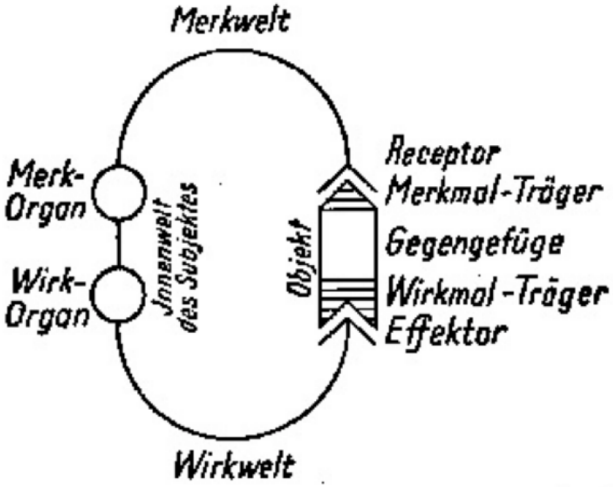
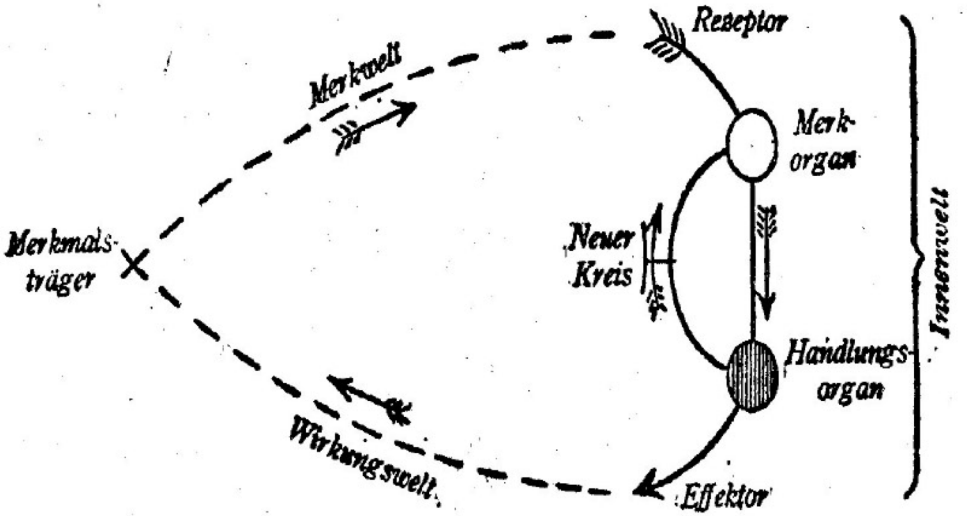


Abb. 3. Funktionskreis



Figur 4.

Figures from von Uexküll's *Theoretische Biologie*, 1920

The AI viewpoint

Reflex agent

- Consider how the world IS
- Choose action based only on current percept
- Do not consider the future consequences of actions

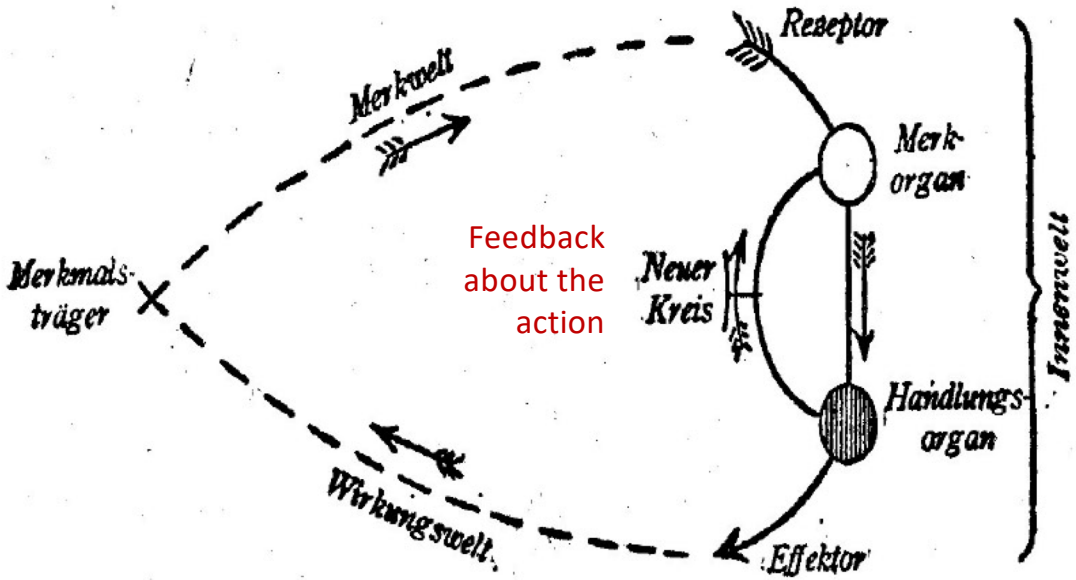
Predictive agent

- Consider how the world WOULD BE
- Decisions based on (hypothesized) consequences of actions
- Must have a model of how the world evolves in response to actions

Sources: D. Klein, P. Abbeel, my AI slides

Sensorimotor feedback loop

Input from the world



The world itself no longer matters!

Feedback about the action

"Inner world" of the agent or interface between the agent and the world

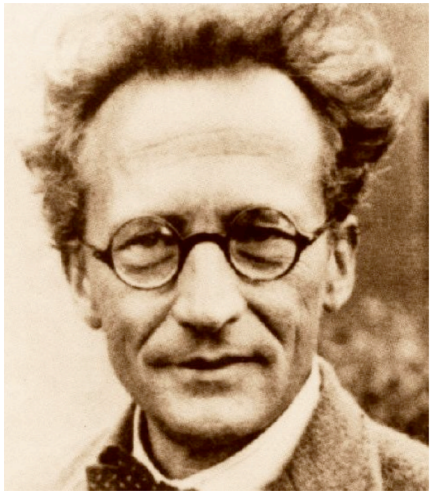
Action to affect the world

J. Koenderink, Sentience, 2019

Figure from von Uexküll's Theoretische Biologie, 1920

The awareness “hypothesis”

- The “new loop” is the source of the organism’s **sentience** or **awareness**
  - In particular, **discrepancies** between the predictions of the feedback mechanism and the observed state of the world generate “sparks of awareness” (a view held by Erwin Schrödinger)



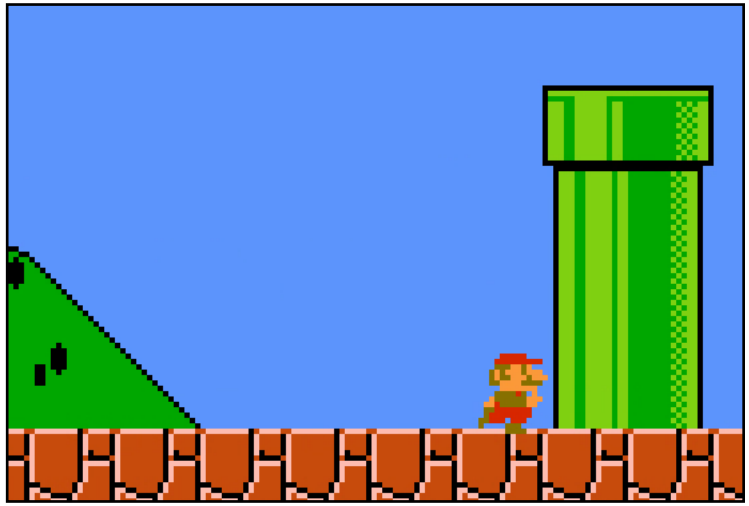
Erwin Schrödinger's in "Mind and Matter" proposes a "psychophysical linking hypothesis" that connects the functional tones to meanings and qualities:

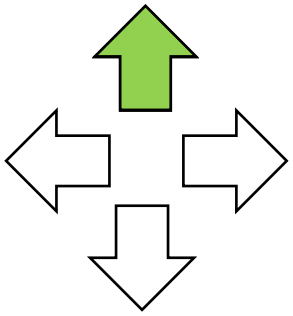
if an expectation is falsified in perception, you "meet nature" - it is a moment of learning: "it discharges a spark of awareness"

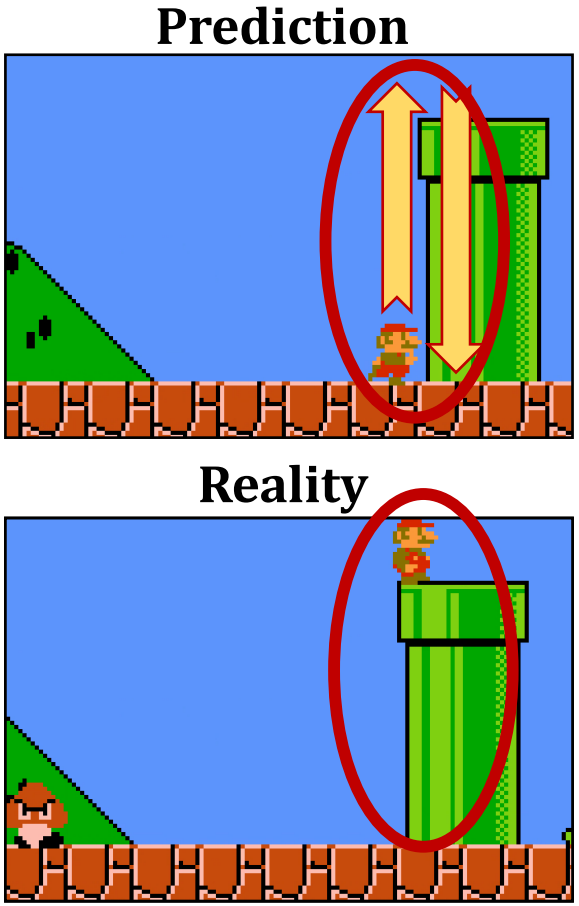
J. Koenderink, Sentience, 2019

Source: Koenderink's slides

Connection: Curiosity-based exploration







Source: D. Pathak et al. (via A. Efros)

D. Pathak et al. Curiosity-driven Exploration by Self-supervised Prediction. ICML 2017

Interface theory of perception

- The “new loop” creates a complete **interface** between the organism and the world. The organism does not experience the world in any other way except through this interface
  - However, the world is still perceived as being “out there” and it can still kill us

D. Hoffman, The interface theory of perception, *Object Categorization: Computer and Human Vision Perspectives*, 2009
See also https://www.quantamagazine.org/the-evolutionary-argument-against-reality-20160421/

Interface theory of perception

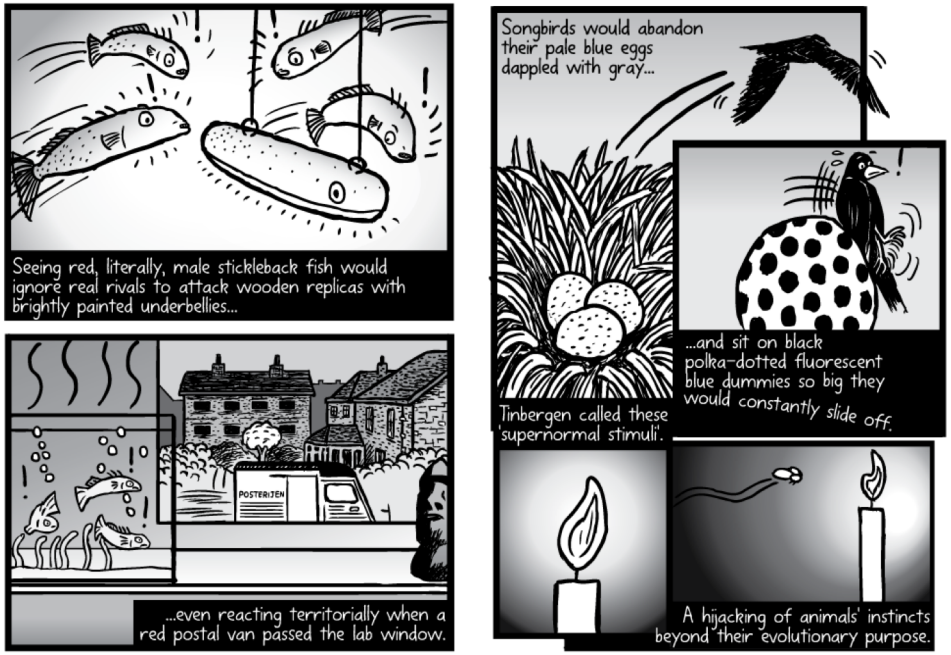
- Conventional view
  - **Principle of Faithful Depiction:** A primary goal of perception is to recover, or estimate, objective properties of the physical world. A primary goal of perceptual categorization is to recover, or estimate, the objective statistical structure of the physical world.
  - Palmer: "Evolutionarily speaking, visual perception is useful only if it is reasonably accurate. Indeed, vision is useful precisely because it is so accurate. By and large, **what you see is what you get**."
- Interface theory
  - "The error in this argument is fundamental: Natural selection optimizes fitness, not veridicality."
  - **Bayes' Circle:** We can only see the world through our posteriors. When we measure priors and likelihoods in the world, our measurements are necessarily filtered through our posteriors. Using our measurements of priors and likelihoods to justify our posteriors thus leads to a vicious circle.

D. Hoffman, The interface theory of perception, *Object Categorization: Computer and Human Vision Perspectives*, 2009

Source: A. Efros

Non-veridicality of perception

- Perception evolved not to produce “accurate” representations of the world, but to further organisms’ fitness
  - It is easy to “hack” many organisms with *supernormal stimuli*

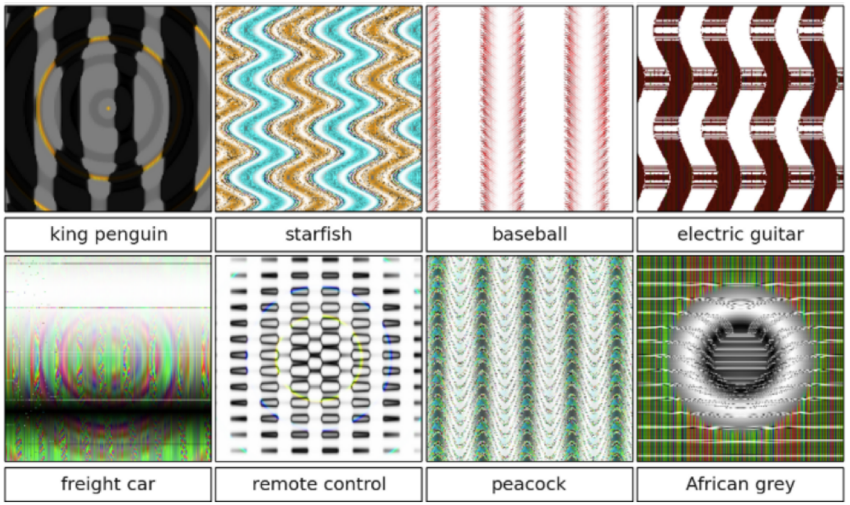


Source

(Wikipedia)

Non-veridicality of perception

- Perception evolved not to produce “accurate” representations of the world, but to further organisms’ fitness
  - It is easy to “hack” many organisms with *supernormal stimuli*



Supernormal stimuli for neural networks?

A. Nguyen, J. Yosinski, J. Clune, Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, CVPR 2015

Interface theory of perception

- **Reconstruction Thesis**: Perception *reconstructs* certain properties and categories of the objective world.

- **Construction Thesis:** Perception *constructs* the properties and categories of an organism's perceptual world.

D. Hoffman, The interface theory of perception, *Object Categorization: Computer and Human Vision Perspectives*, 2009

Source: A. Efros

The process of perception

- Perception is a fundamentally **active**, **creative** process that generates theories about the world based on sensory input and retains the theory that best fits the input

Looking is an action, as is pretty clear in this picture of Toshiro Mifune. The notion that vision is a passive act in which the world spoon–feeds you with information is nonsense. Optical meaning is actively hunted for.

The famous — although fictive — detective Sherlock Holmes plays a major role in my account of the theory of psychogenesis.

J. Koenderink, Sentience, 2019

Source: Koenderink's slides

The process of perception

- Perception is a fundamentally **active**, **creative** process that generates theories about the world based on sensory input and retains the theory that best fits the input
  - Contrary to Marr, perception is most definitely not “inverse optics”
  - Contrary to Gibson, perception is not a function primarily of the environment. It can frequently be ambiguous and is heavily driven by the organism’s goals, desires, and internal state

J. Koenderink, Sentience, 2019

Perception as controlled hallucination

Video by Antonio Torralba (starring Rob Fergus)

But actually…

Video by Antonio Torralba (starring Rob Fergus)

Implications

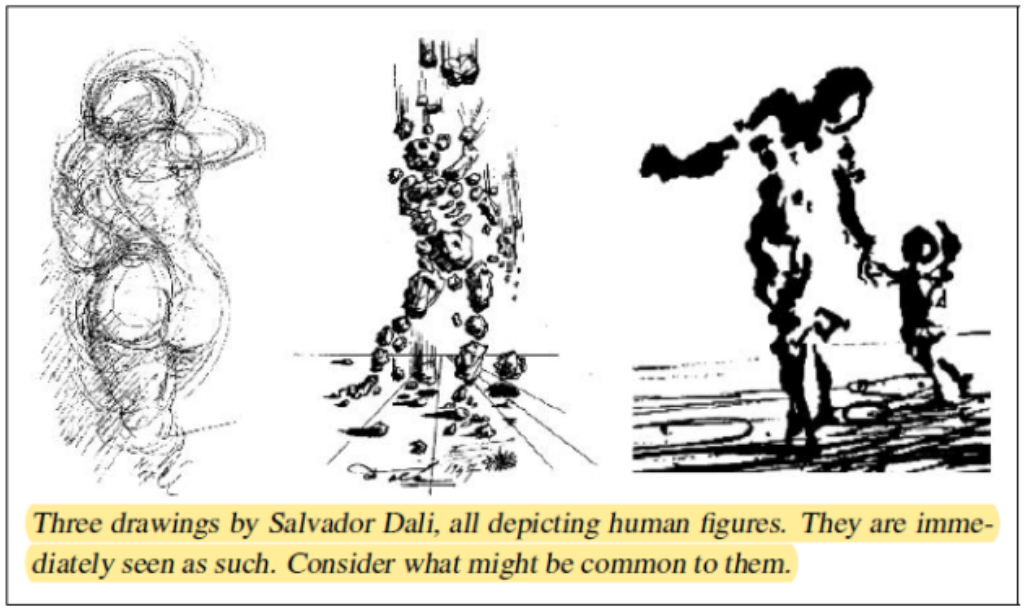
- "Perceptual organization" cannot be primarily a bottom-up process as Marr saw it

Figure 3–1. The interpretation of some images involves more complex factors as well as more straightforward visual skills. This image devised by R. C. James may be one example. Such images are not considered here.

Figure from Marr

Implications

- “Perceptual organization” cannot be primarily a bottom-up process as Marr saw it
  - Koenderink: “Edges are imposed, not detected”



Three drawings by Salvador Dali, all depicting human figures. They are immediately seen as such. Consider what might be common to them.

What about recognition?

- Koenderink agrees with Gibson that “object categories” don’t make sense

"Now! ... *That* should clear up a few things around here!"

What about recognition?

- Koenderink agrees with Gibson that “object categories” don’t make sense

Abb. 52 Förster und Eiche.

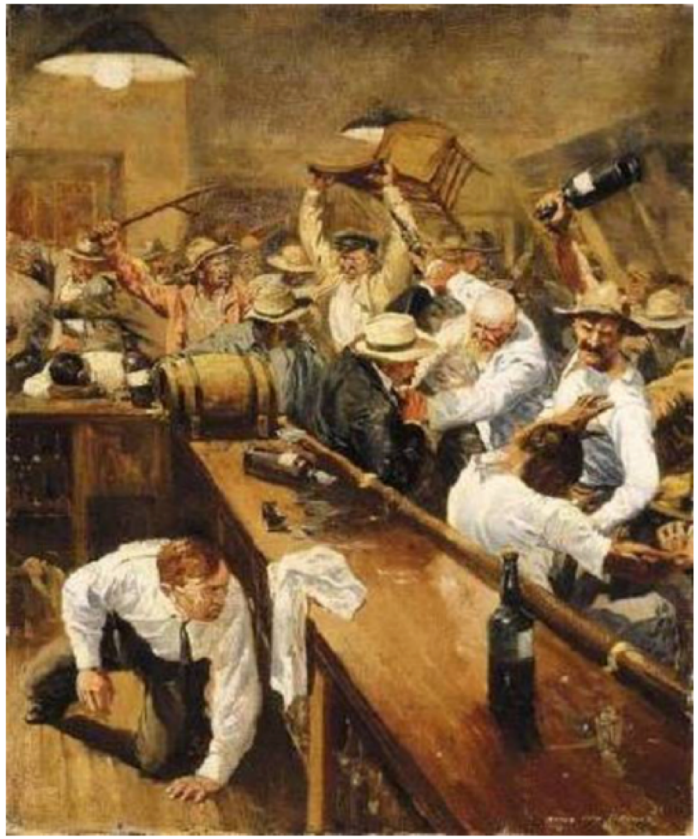
Abb. 53 Mädchen und Eiche.

A tree seen by different sentient beings. The little girl is scared by the hallucinated face, the forester is mainly occupied with the diameter of the trunk, which is an indicator of the bulk of wood offered by the tree. Other examples discussed by Jakob von Uexküll are a fox, an ant and various others. The tree is a different object to all these sentient beings.

J. Koenderink, Sentience, 2019

What about recognition?

- At best, categories are “bundles of cues that play a role in actions”
  - Similar to Gibson’s affordances but to Koenderink, there is no such thing as an intrinsic affordance



In a bar fight tables and bottles gain a new meaning. Tables are not just “situponable,” or “furniture.” Bottles do not say “drink me.”

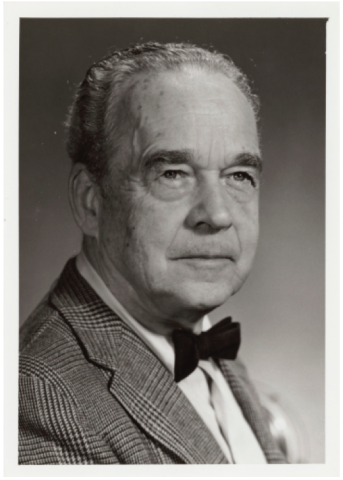
Gibson’s notion that the affordance of an object is a property of the object, like its weight, or size, evidently fails to reach the heart of the matter. Anything can mean anything, depending upon mood and situational awareness.

J. Koenderink, Sentience, 2019

Summary

- Marr, Gibson, and Koenderink all asked about the nature of vision and came up with different answers

"Vision is a computational process that transforms the retinal image into an objective representation of 3D shape."



"There is no computation. There is no retinal image. There are no representations. There is no 3D shape. There is only direct pickup of ecologically relevant variants and invariants. Vision is in the world, not the observer."

"There is no objective world, only the observer's *umwelt*. Thus, vision cannot be in the world but is a creative act of the observer."

Take-aways?

- The time may be right to take on the “crazier” ideas of Gibson and Koenderink
  - We need to study embodied vision
  - We need to build models with feedback
  - We need to focus on “ecologically meaningful” tasks (and object classification is most likely not it)
  - We need to integrate discriminative and generative models