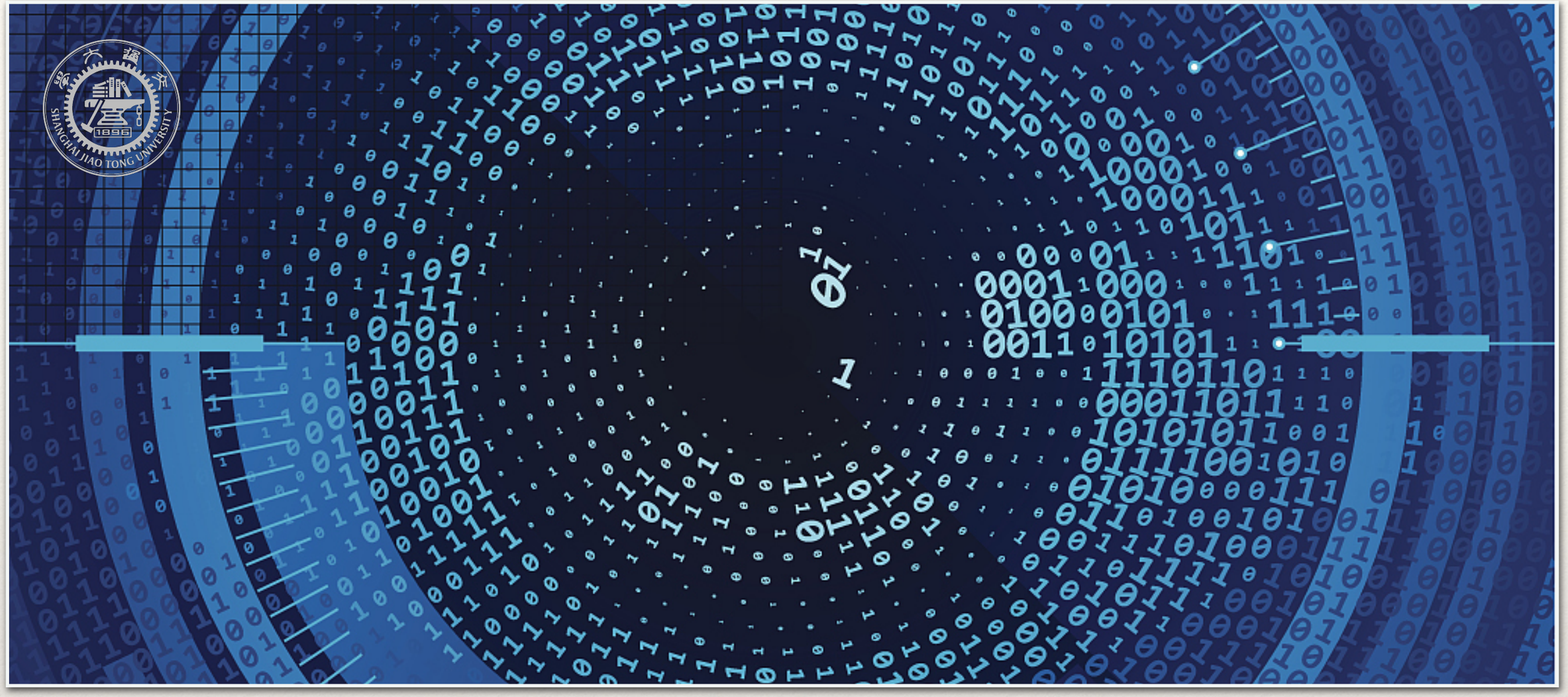


Shanghai Jiao Tong University

Computer Vision

Instructor: Xu Zhao

Class No.: C032703 F032528

Spring 2020

About Me

- Xu Zhao (赵旭), Professor at Department of Automation, SEIEE, SJTU
- Research direction: Computer Vision, Image Processing, Machine Learning, Pattern Recognition
- Email: zhaoxu@sjtu.edu.cn
- Office: 2-431 SEIEE Buildings

About this course

- **Motivation**
  - **Application**: promising and significant direction toward practical artificial intelligence.
  - **Research**: in CV field, many open problems need to be solved by inventing diverse methodologies from different research domains.
  - “Time and tide wait for no man.”

Lecture 1: Introduction

Contents

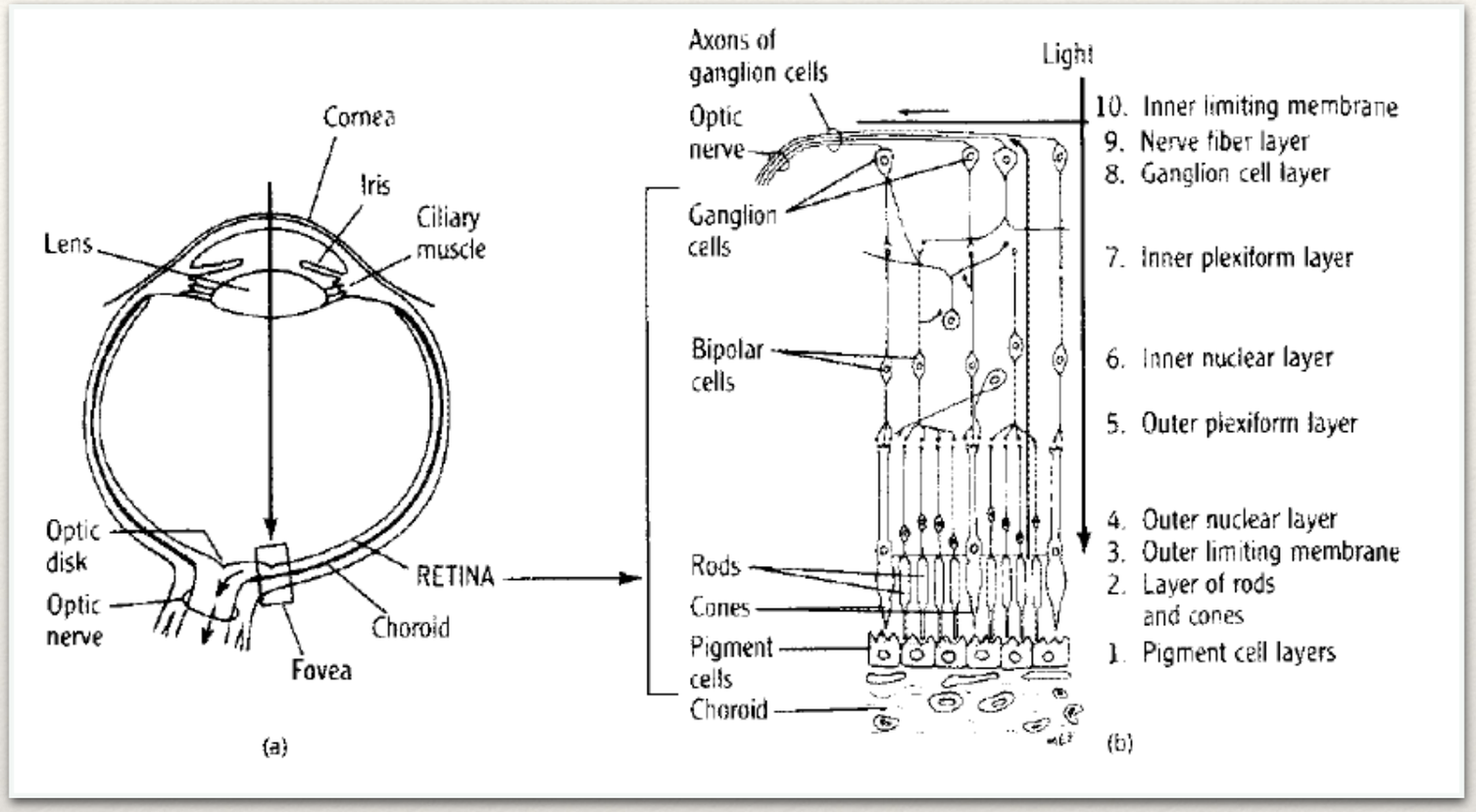
- **What is vision?**
  - Psychological perspective
  - David Marr's theory
- **What is computer vision?**
  - Motivation
  - History
  - Current state: applications and challenges
- **Course information**

What is vision?

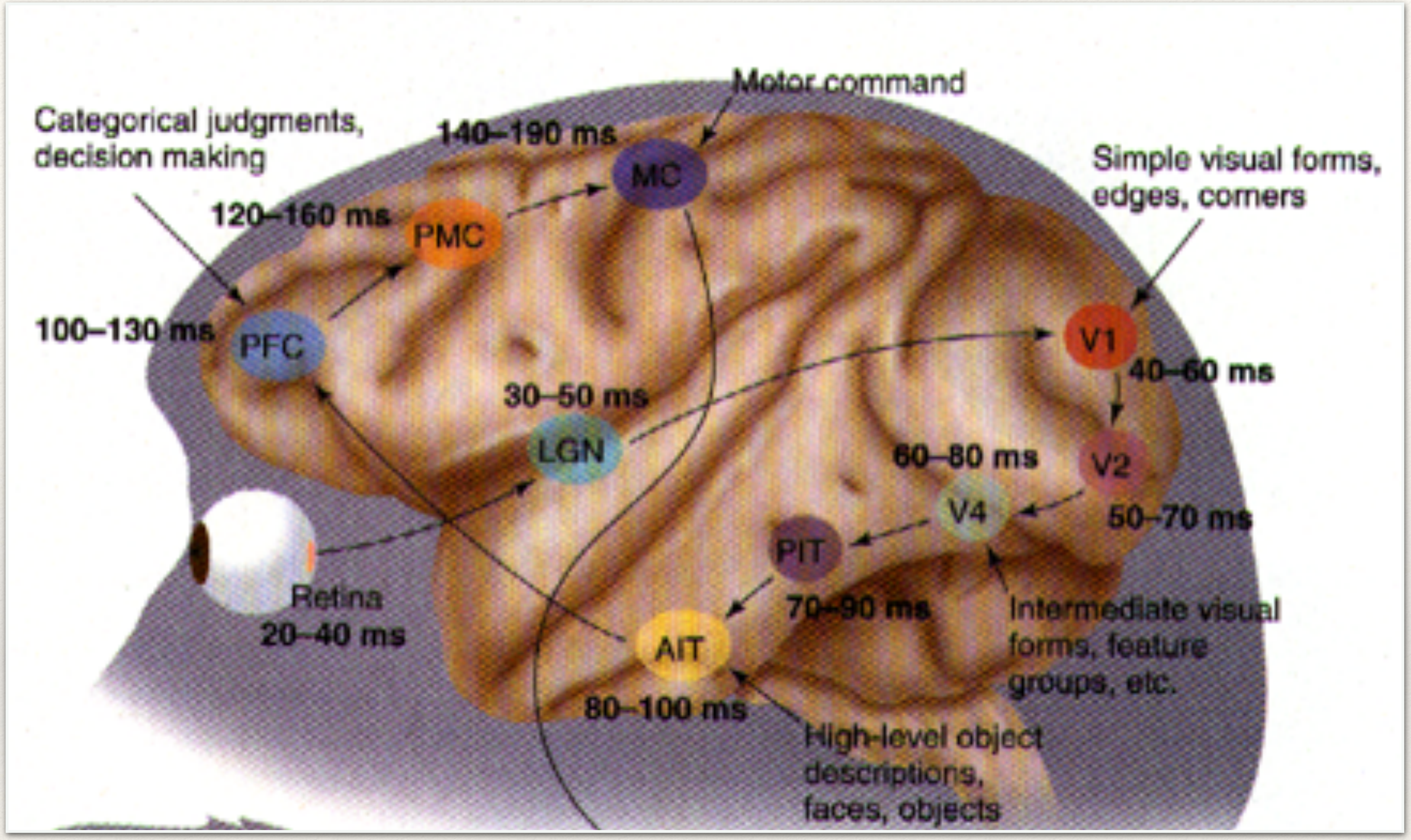
- What does it mean, to **see**?
  - **Philosophy perspective**: "To know what is where by looking". -plain man's answer. Vision is a process discovering from images what is present in the world and where it is, and furthermore, what action are taking place. [David Marr, Vision]
  - **Biological perspective**: The special sense by which the qualities of an object (such as color, luminosity, shape, and size) constituting its appearance are perceived through a process in which light rays entering the eye are transformed by the retina into electrical signals that are transmitted to the brain via the optic nerve. [Merriam-webster]

Human visual system

- The eye's spatial resolution is about $0.01°$ over a $150°$ field of view (not evenly spaced, there is a fovea and a peripheral region).
- Intensity resolution is about 11 bits/element, spectral range is 400–700nm.
- Temporal resolution is about 100 ms (10 Hz).
- Two eyes give a data rate of about 3 GBytes/s!



- Retina measures about 5×5 cm and contains 10^8 sampling elements (rods and cones).

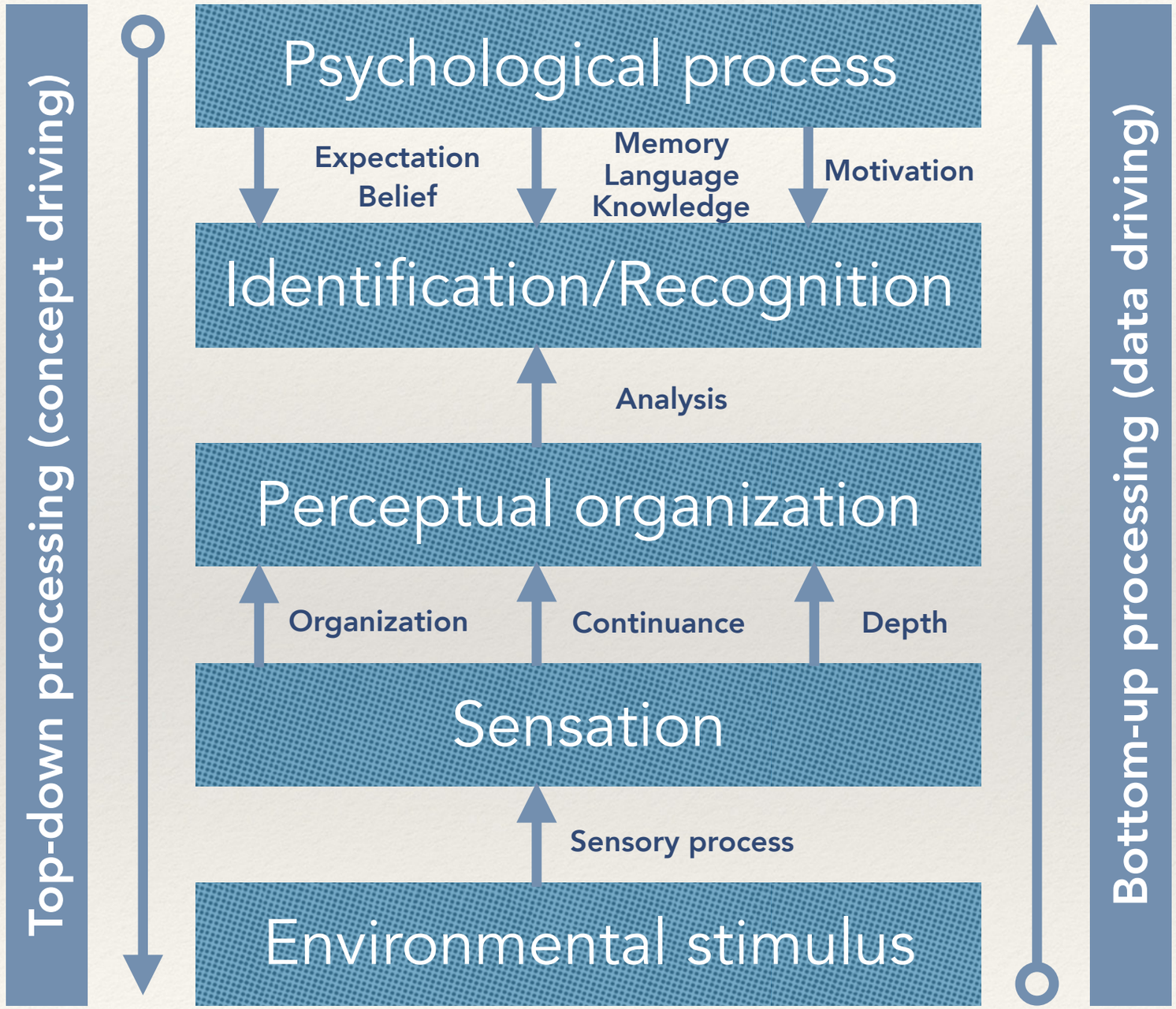


- Vision is the most powerful of our own senses.
- Around 1/3 of our brain is devoted to processing the signals from our eyes.
- The visual cortex has around $o(10^{11})$ neurons.

Visual perception

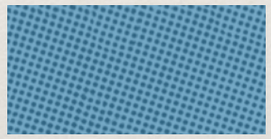
- **Sensation:** Provide basic fact in the view field and encoding the physical energy (optical signal) as neural signal which brain can recognize.
- **Perception**: A series of processes that organize and interpret sensory information from external objects and events.
  - Perceptual organization
  - Identification and recognition

Visual information processing

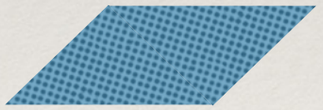


Example:

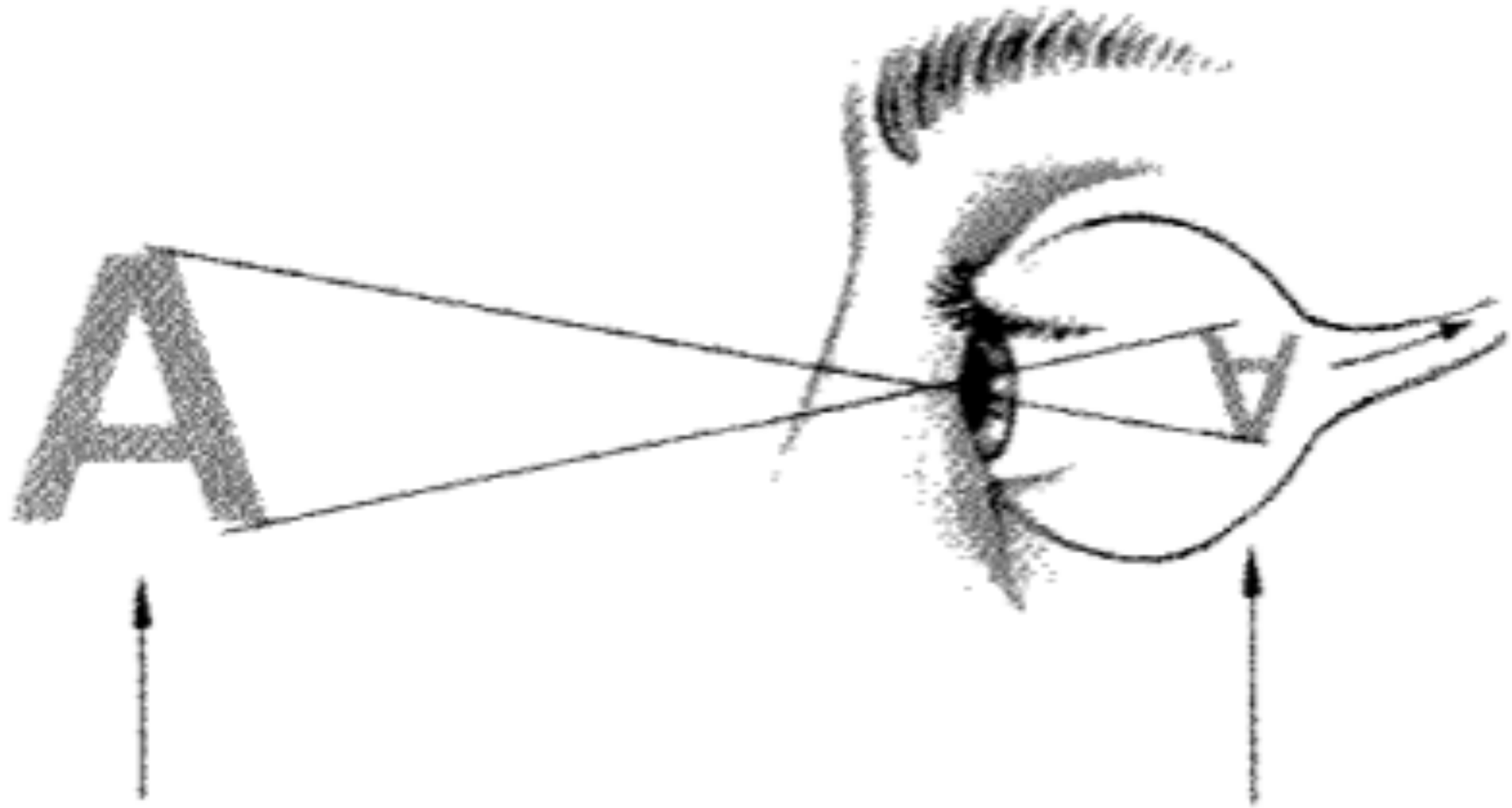
Recognized as kind of"Graph"



Rectangle



Retina image



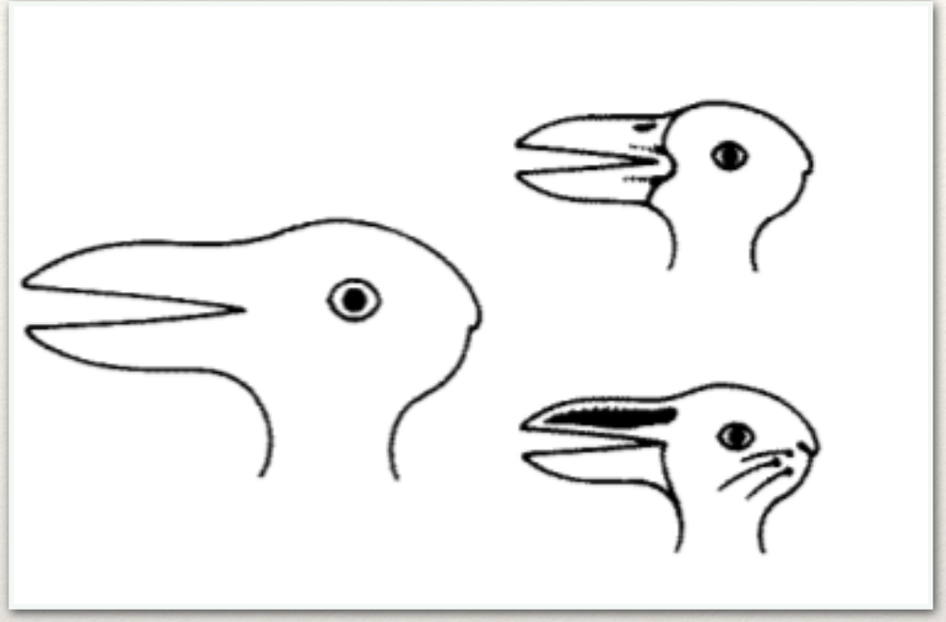
Distal stimulas

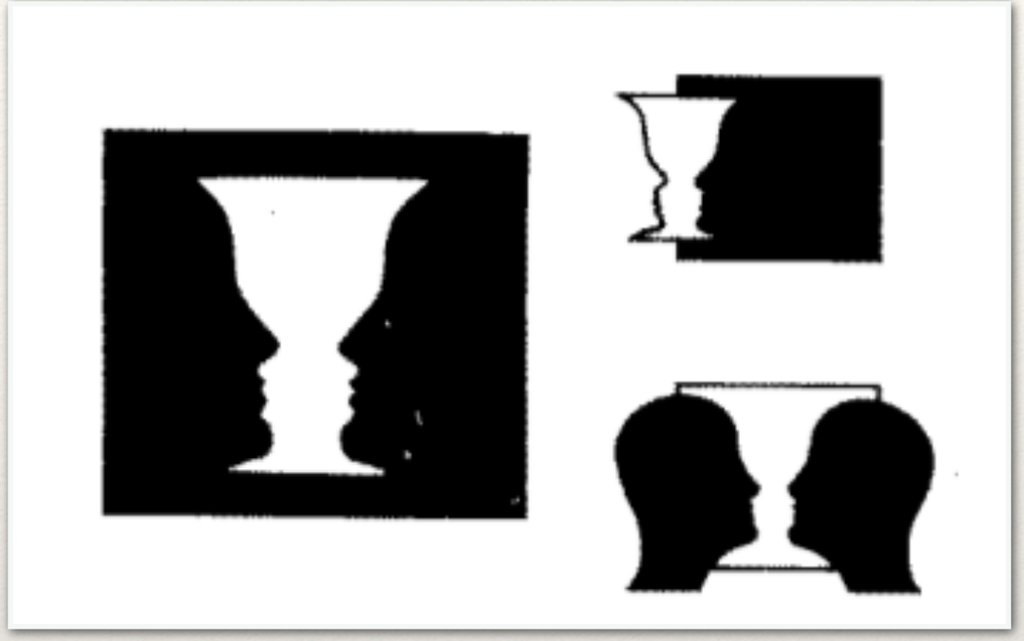
Proximal stimulas

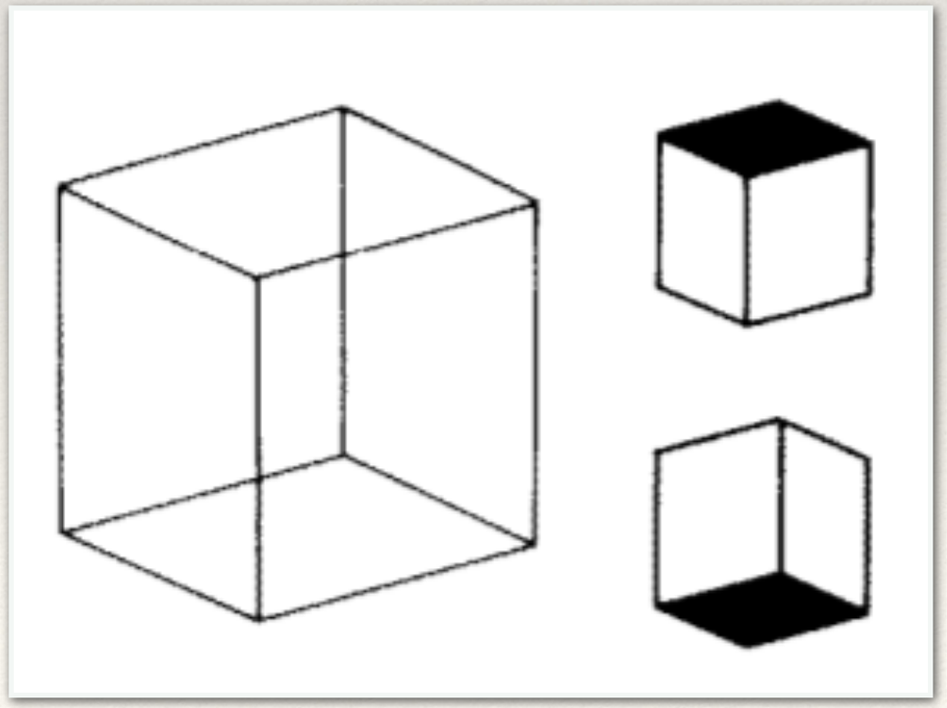
Perceptual organization

- Form the description to the objects
- Size, shape, motion, distance, orientation…
- Computing process integrating past experience and current input
- Integration: from parts to whole
- Fast and having no cognitive process

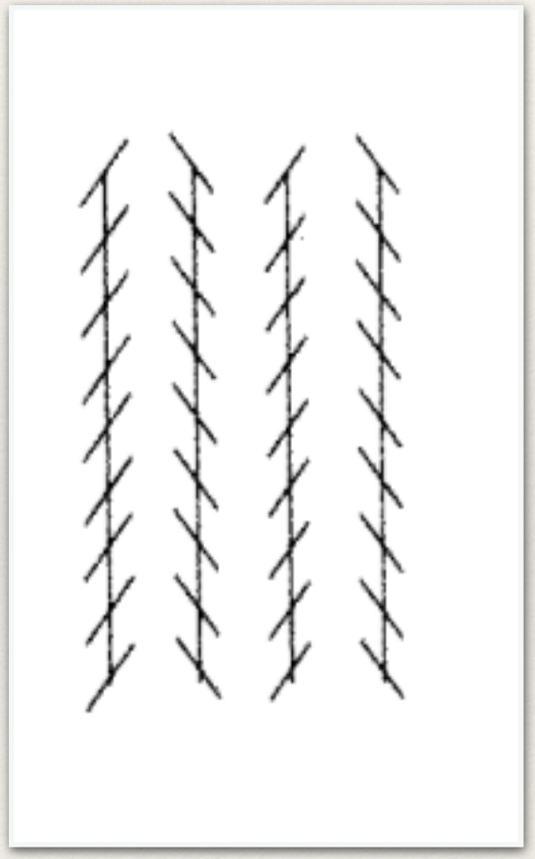
Challenges: ambiguity



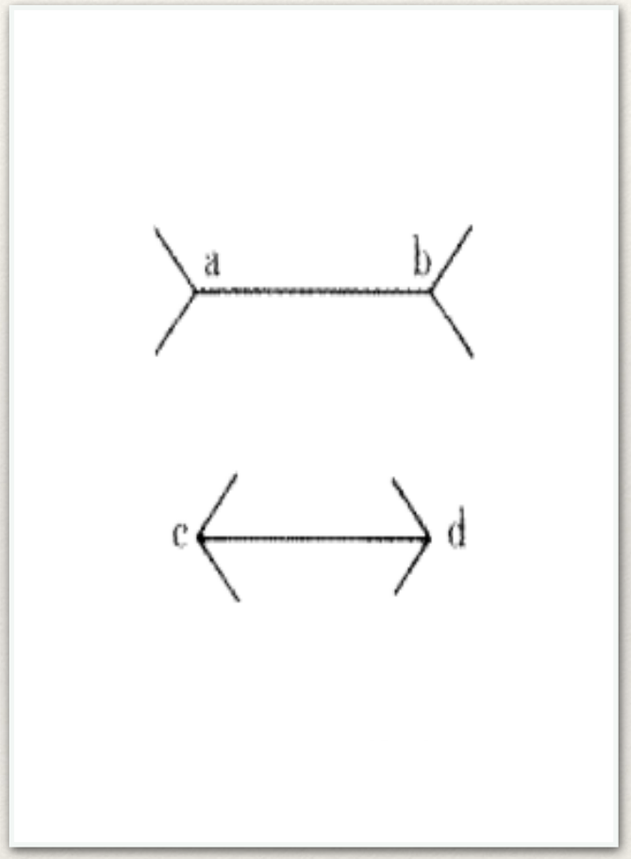




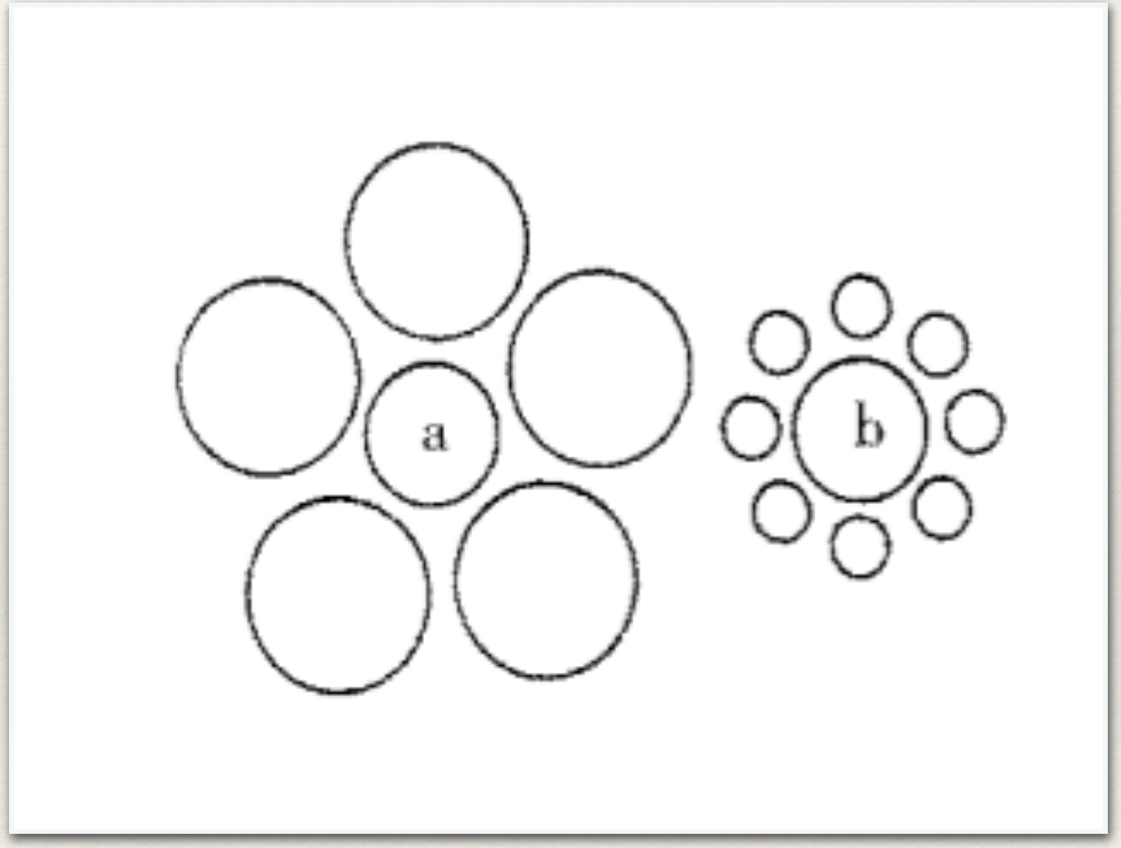
Challenges: illusion



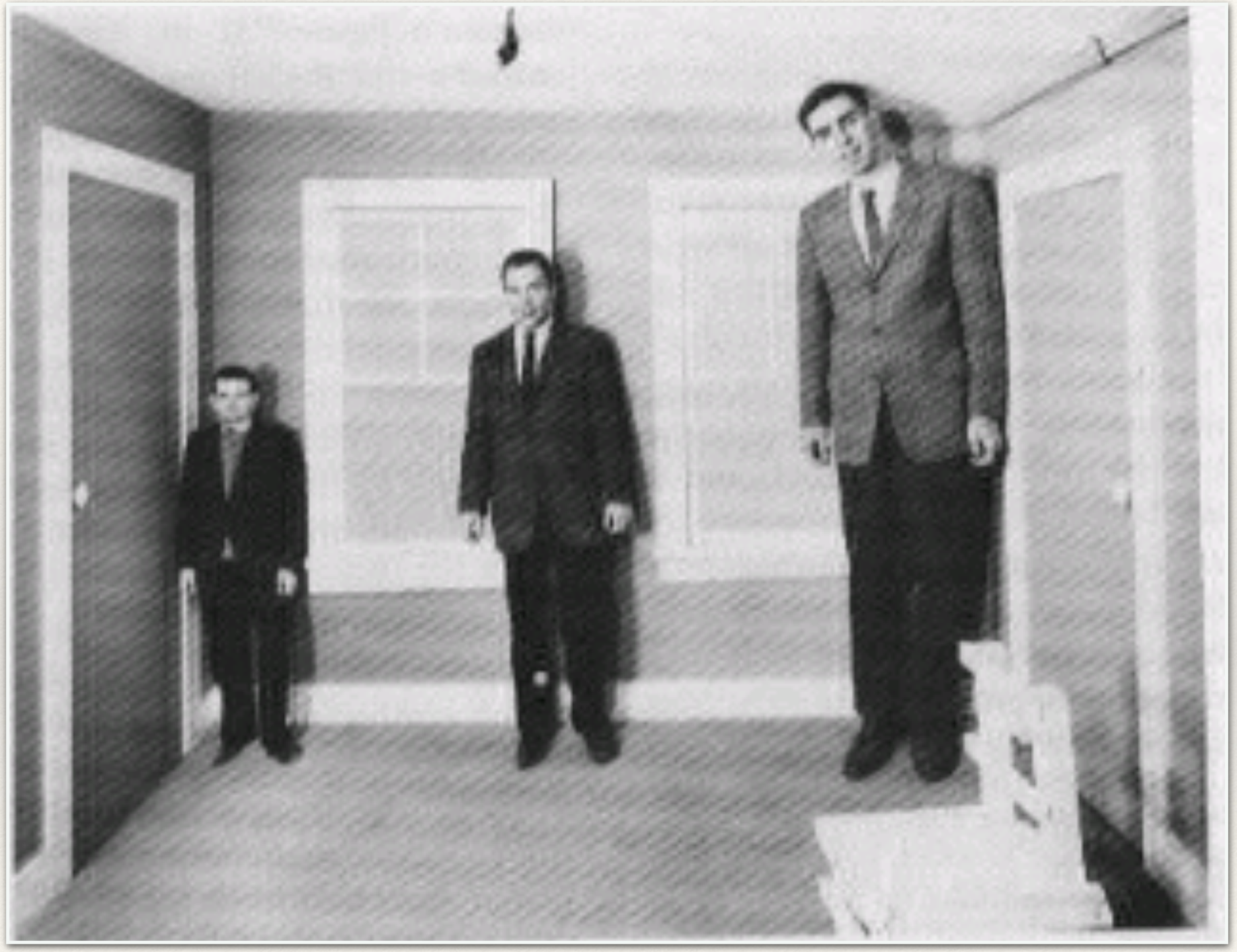
Zollner illusion



Muller Lyer illusion



Ebbinghaus Illusion

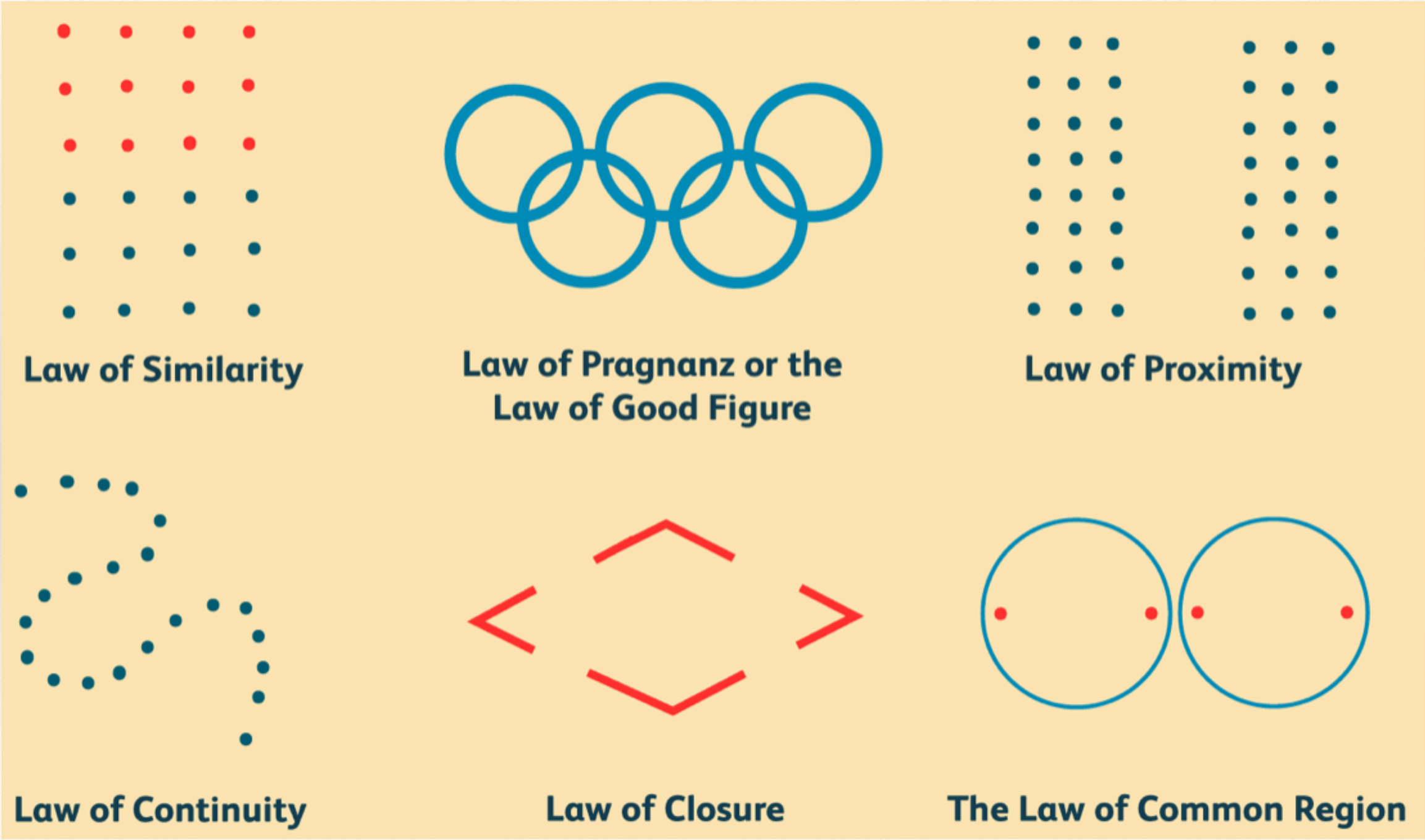


Ames room

Psychological approaches

- **Hermann von Helmholtz classical theory** (1866): unconscious inference, role of the experience
- **Gestalt psychology** (1920s): whole, structure, nature. The guiding principle behind the Gestalt movement was that the whole was greater than the sum of its parts.
- **Gibson theory of ecological optics** (1960s-1970s): focus on the attributes of the stimulus from the real world; perception is a kind of active exploring to stable environment.

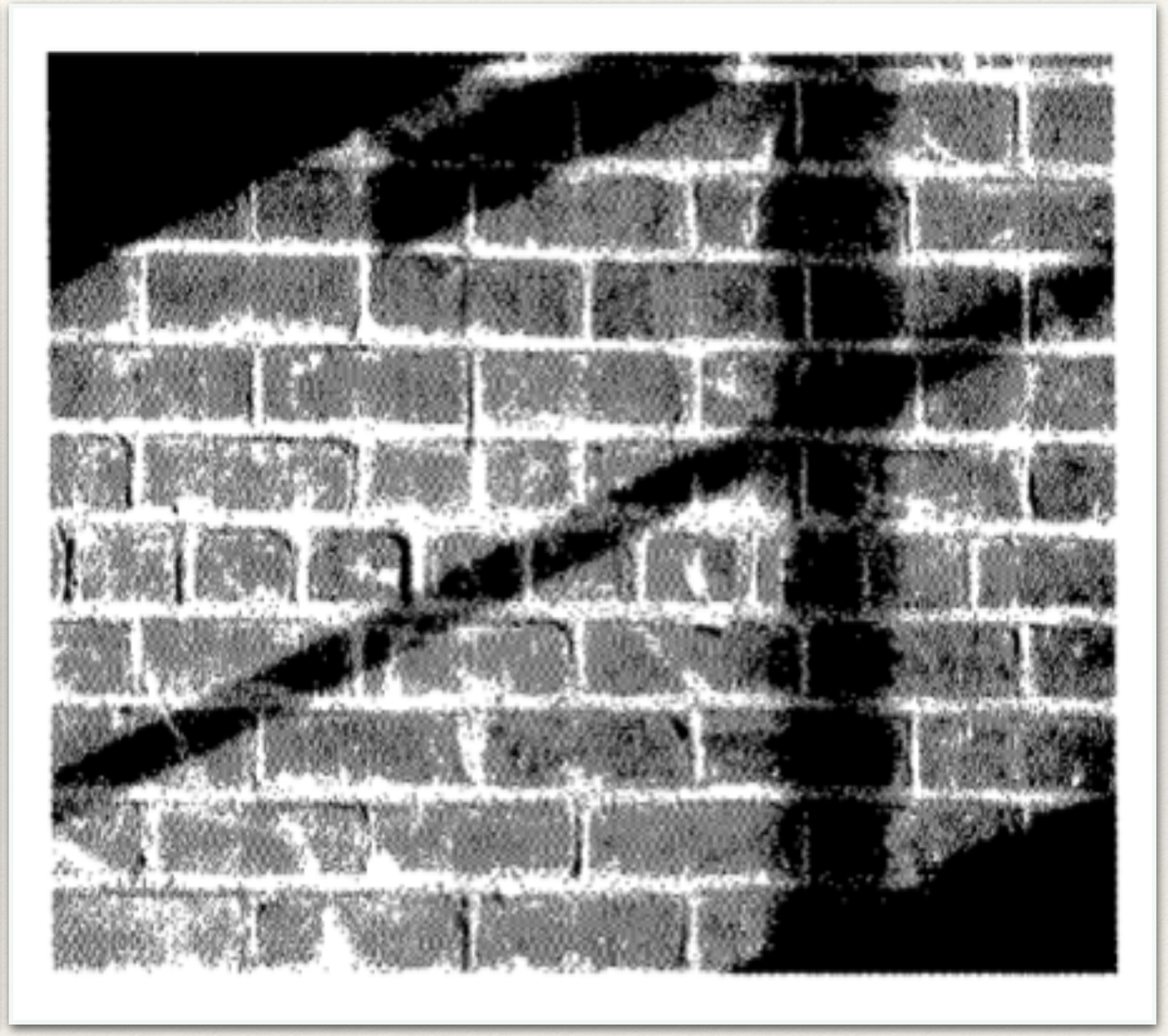
Examples of the Gestalt laws



Verywell / JR Bee

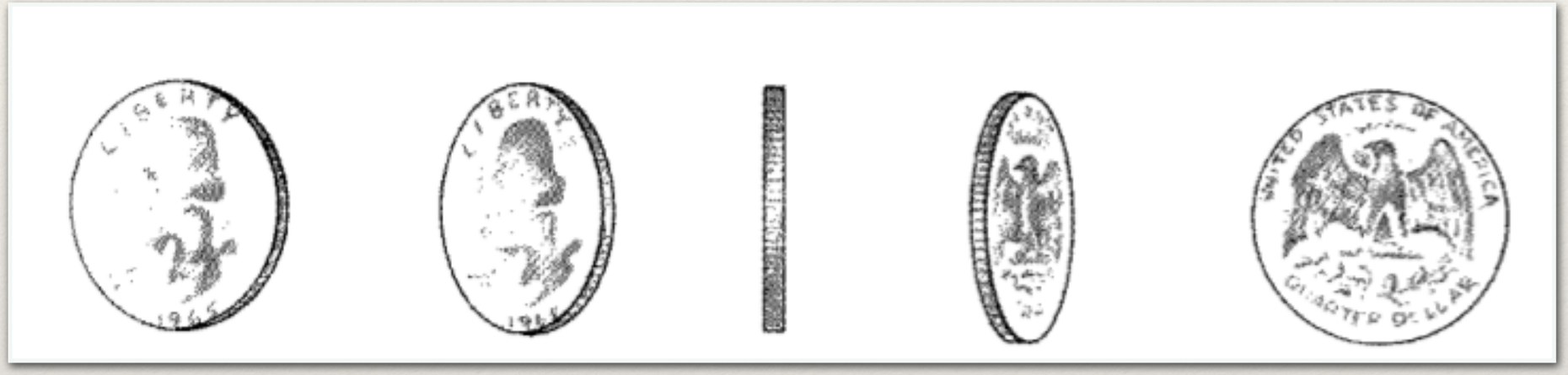
Perceptual constancy

Size



Lightness

Perceptual constancy

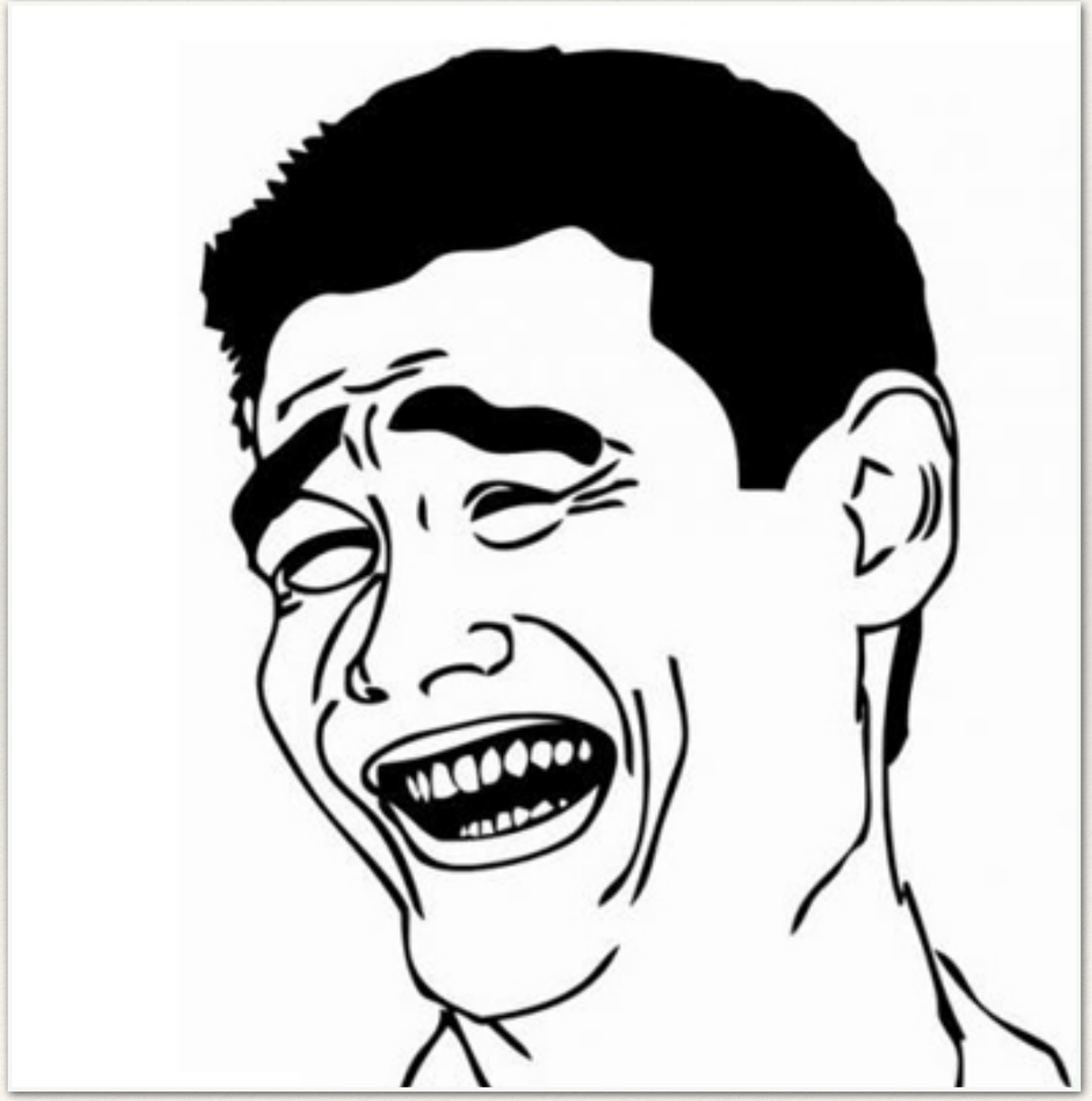


Shape

Orientation

Identification and recognition

- Top-down and bottom-up process

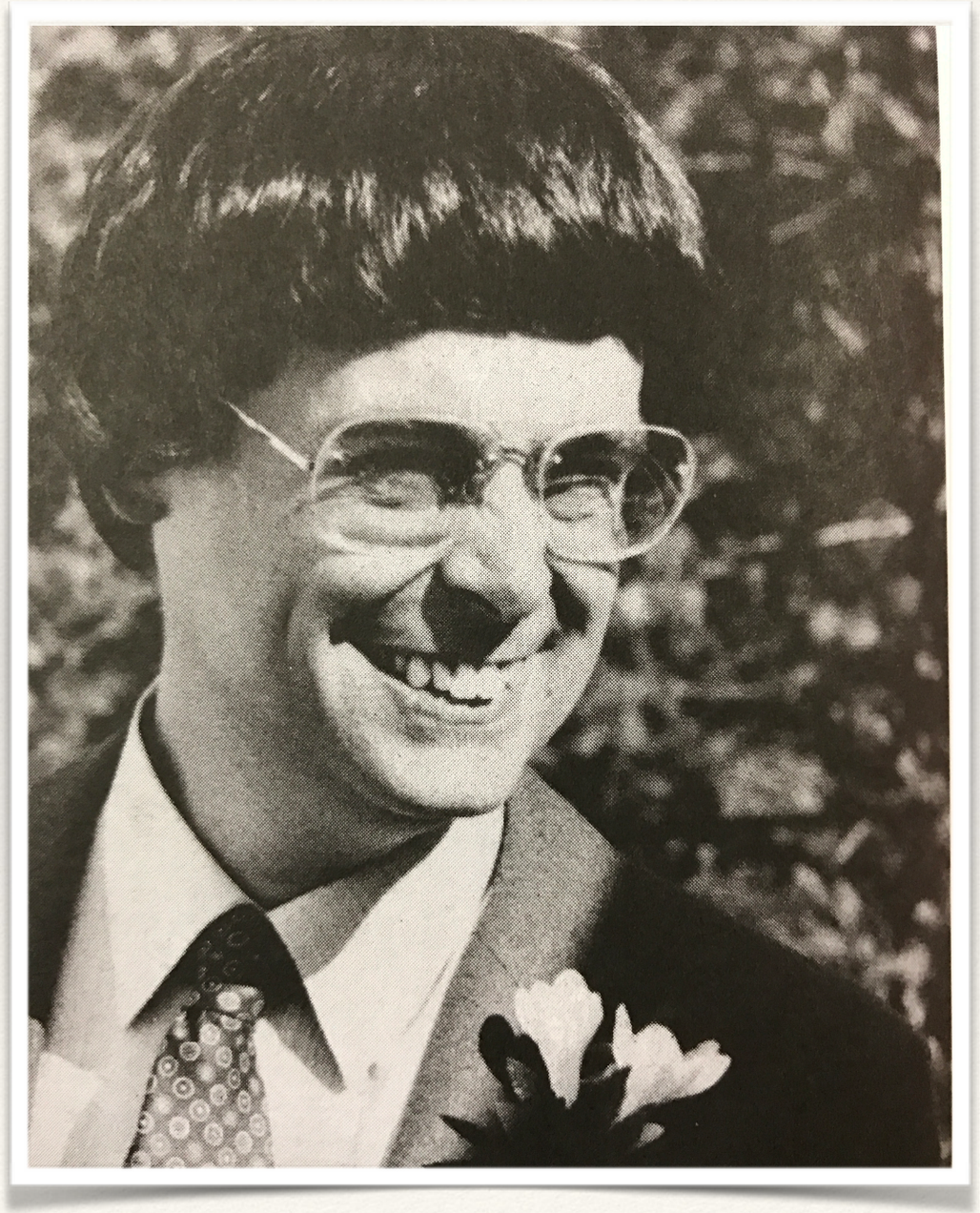


Identification and recognition

- Context plays critical role

David Marr's theory

- **David Marr (1945-1980): Pioneer scientist of computer vision**. David Courtenay Marr was a British neuroscientist and physiologist. Marr integrated results from psychology, artificial intelligence, and neurophysiology into new models of visual processing. His work was very influential in computational neuroscience and led to a resurgence of interest in the discipline. - WiKi



David Marr's theory

- Vision is a complex information processing task.
- A **representation** is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this.
- "And I shall call the result of using a representation to describe a given entity a **description** of the entity in that representation" - Marr and Nishihara, 1978

David Marr's theory

- The three levels at which any machine carrying out an information-processing task must be understood [Vision, David Marr]
  - **Computational theory**: 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?
  - **Representation and algorithm**: 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?
  - **Hardware implementation**: How can the representation and algorithm be realized physically?

Representational framework

Name	Purpose	Primitives
Images	Represent intensity	Intensity value at each point in the image
Primal sketch	Makes explicit important information about the 2D image, primarily the intensity changes there and their geometrical distribution and organization.	Zero-crossings Blobs, Terminations and discontinuities, Edge segments, Virtual lines, Groups, Curvilinear organization Boundaries
2.5 D sketch	Makes explicit the orientation and rough depth of the visible surfaces and contours of discontinuities in these quantities in a viewer centered coordinate frame	Local surface orientation (the "needles" primitives); Distance from viewer; Discontinuities in depth; Discontinuities in surface orientation
3-D model representation	Describes shapes and their spatial organization in an object-centered coordinate frame, using a modular hierarchical representation that includes volumetric primitives as well as surface primitives.	3-D models arranged hierarchically, each one based on a spatial configuration of a few sticks or axes, to which volumetric or surface shape primitives are attached

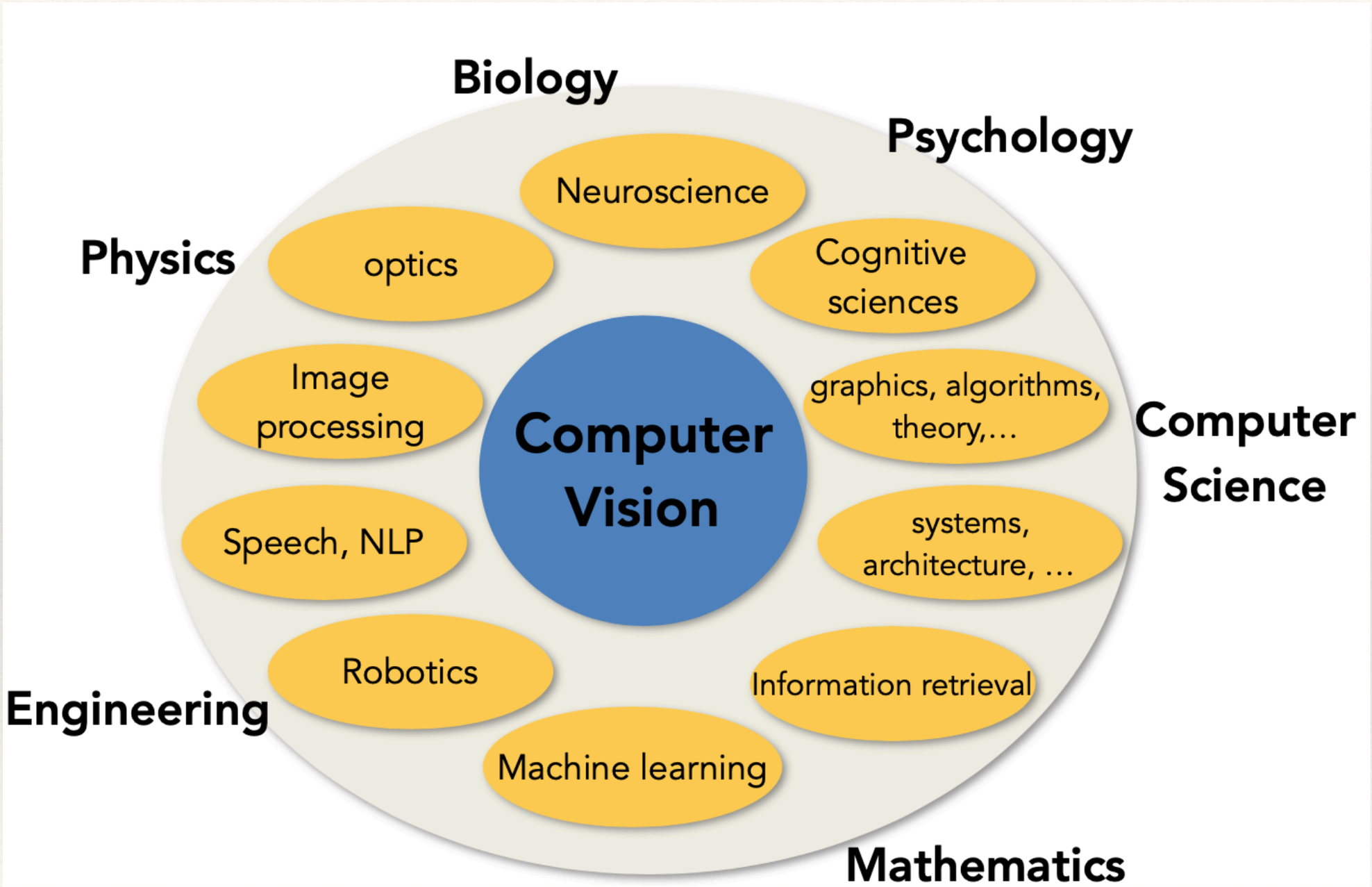
What is computer vision?

- **Vision** is about discovering from images what is present in the scene and where it is.
- In **Computer Vision** a camera (or several cameras) is linked to a computer. The computer interprets images of a real scene to obtain information useful for tasks such as navigation, manipulation and recognition.
- **Computer vision** aims to recover useful information about a (3D) scene from its 2D projections (images).

Motivation

- Replicate human vision to allow a machine to see
  - Central to that problem of Artificial Intelligence
  - Many industrial applications
- Gain insight into how we see
  - Vision is explored extensively by neuroscientists to gain an understanding of how the brain operates

What is it related to?



Fei-FeiLi & JustinJohnson & SerenaYeung

History

- **1970s: Early thriving**
  - 1966: Marvin Minsky, at MIT asked his undergraduate student Gerald Jay Sussman to "spend the summer linking a camera to a computer and getting the computer to describe what it saw"
  - 1970s: ***Scene understanding*** by finding edges and then inferring the 3D structure
  - 1970s: Three-dimensional modeling of non-polyhedral objects, ***generalized cylinder***
  - *1973:* ***pictorial structures*** for object recognition, Fischler and Elschlager
  - *1980s:* qualitative approach to understanding intensities and shading variations. ***Shape from shading.***
  - 1970s-1980s: ***Stereo*** corresponding algorithms.
  - 1970-1980: simultaneously recovering ***3D structure*** and camera motion
  - 1970s-1980s: Intensity based ***optical flow***
  - 1970s: David Marr: ***three level visual information processing system***

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966

Vision Memo. No. 100.

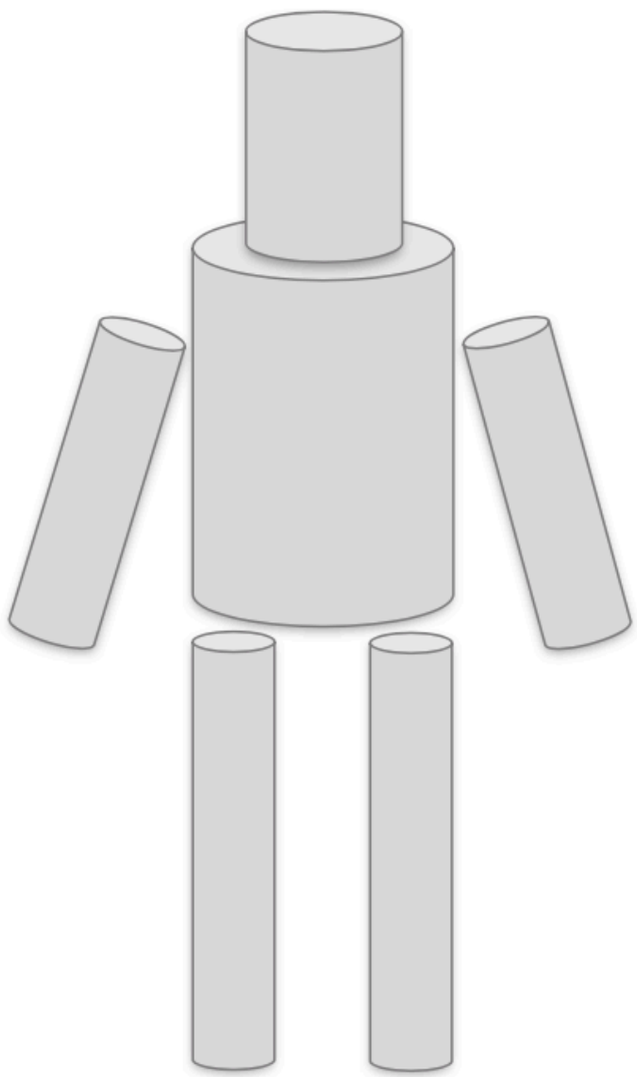
THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

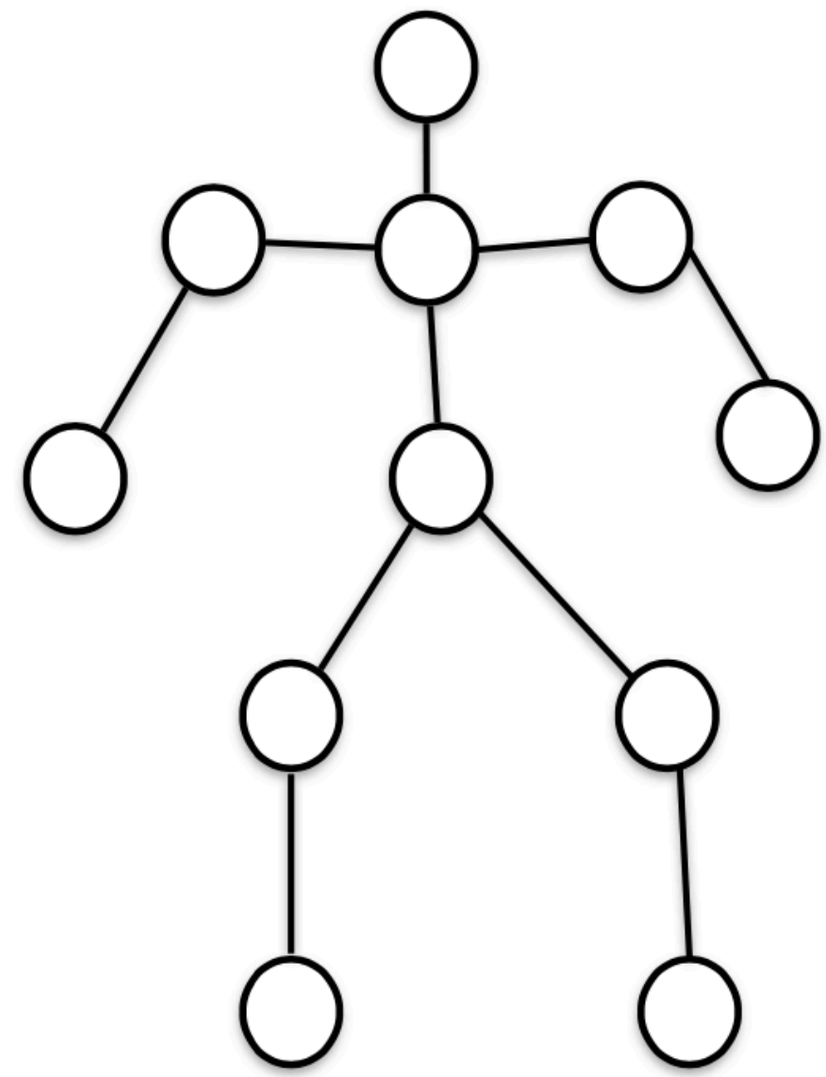
- Generalized Cylinder

Brooks & Binford, 1979



- Pictorial Structure

Fischler and Elschlager, 1973



History

- **1980s: Utilization of sophisticated mathematical techniques**
  - ***Motion and structure*** from feature correspondences
  - ***Image pyramid*** and wavelets for multiple resolution image processing
  - ***Shape-from-X*** (stereo, texture, shading, focus)
  - ***Canny*** for edge detection (1986)
  - ***Active contours*** (snake, 1988)
  - ***MRF*** applied into CV (1984)
  - ***Variational optimization*** problems and regularization

Motion and Structure from Feature Correspondences: A Review

THOMAS S. HUANG, FELLOW, IEEE, AND ARUN N. NETRAVALI, FELLOW, IEEE

We present a review of algorithms and their performance for determining three-dimensional (3D) motion and structure of rigid objects when their corresponding features are known at different times or are viewed by different cameras. Three categories of problems are considered, depending upon whether the features are two- (2D) or three-dimensional (3D) and the type of correspondence: a) 3D to 3D (i.e., locations of corresponding features in 3D space are known at two different times), b) 2D to 3D (i.e., locations of features in 3D space and their projection on the camera plane are known), and c) 2D to 2D (i.e., projections of features on the camera plane are known at two different times). Features considered include points, straight lines, curved lines, and corners. Emphasis is on problem formulation, efficient algorithms for solution, existence and uniqueness of solutions, and sensitivity of solutions to noise in the observed data. Algorithms described have been used in a variety of applications. Some of these are: a) positioning and navigating 3D objects in a 3D world, b) camera calibration, i.e., determining location and orientation of a camera by observing 3D features whose location is known, c) estimating motion and structure of moving objects relative to a camera. We mention some of the mathematical techniques borrowed from algebraic geometry, projective geometry, and homotopy theory that are required to solve these problems, list unsolved problems, and give some directions for future research.

such a way that they require solution of simultaneous transcendental, polynomial, or linear equations in multiple variables which represent the structure of the object and its 3D motion as characterized by rotation and translation. It is this commonality that we wish to exploit by formulating the problems in seemingly disparate applications in a common format and then discussing mathematical tools for solving them.

The list of applications in sensing, modeling, and interpretation of motion and structure from corresponding features observed at different times is rather long. It depends upon the type of features observed. We categorize the problems discussed in this paper and their applications as follows:

i) 3D-to-3D feature correspondences.
   applications:

   a) Motion estimation using stereo or other range finding devices
   b) positioning a known 3D object using stereo or other range-finding devices.

History

- **1990s: Existing techniques continued to be explored, learning based appeared**
  - Projective invariants for recognition, ***structure from motion***
  - ***Factorization*** techniques
  - ***Bundle adjustment***
  - ***Physics-based*** vision, physical models of radiance transport and color image formation
  - Dense stereo correspondence algorithms, ***multi-view stereo algorithms***
  - Tracking: ***particle filter***
  - Image segmentation: mean-shift, ***normalized cut***
  - ***Statistical learning*** based
  - Interaction with ***computer graphics***

Normalized Cut (Shi & Malik, 1997)

Image is CC BY 3.0

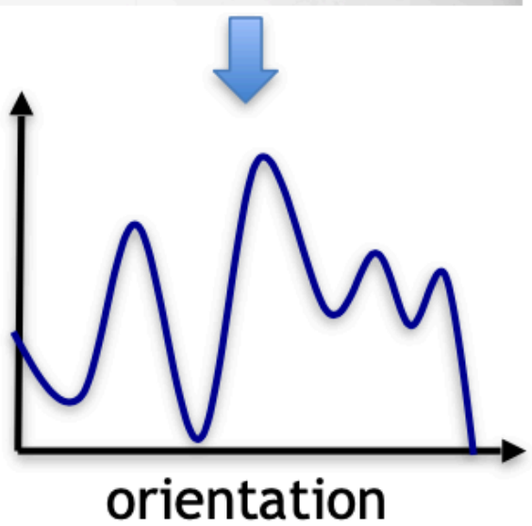
Image is public domain

Image is CC-BY SA 3.0

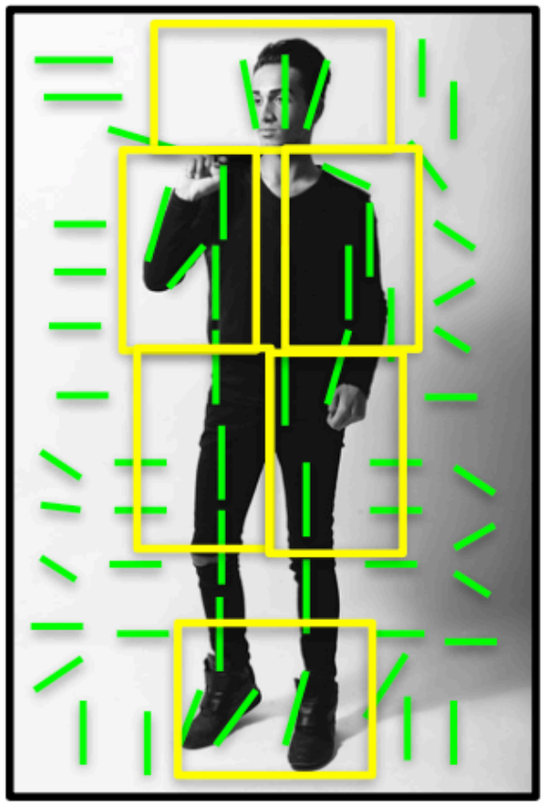
History

- **2000s: Sophisticated machine learning techniques**
  - Projective feature-based techniques (combined with learning) for ***object recognition***
  - ***Pictorial structure***: Felzenszwalb and Huttenlocher
  - ***Interest point*** features
  - Complex ***global optimization*** problems
  - ***Probabilistic graph models***
    - Loopy belief propagation (LBP)
    - ***Graph cut***
  - ***Dimensionality reduction***

Image is CC0 1.0 public domain



Histogram of Gradients (HoG)
Dalal & Triggs, 2005



Deformable Part Model
Felzenswalb, McAllester, Ramanan, 2009

History

- **2010s: Full prosperity of deep learning**
  - ***Big data driven learning*** for CV task
  - ***Deep learning*** and ***CNN***
  - ***GPU***

IMAGENET Large Scale Visual Recognition Challenge

Steel drum

The Image Classification Challenge:

1,000 object classes

1,431,167 images

Output:
Scale
T-shirt
<u>Steel drum</u>
Drumstick
Mud turtle

✔

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

Fei-FeiLi & JustinJohnson & SerenaYeung

IMAGENET Large Scale Visual Recognition Challenge

Year 2010

NEC-UIUC

Dense descriptor grid: HOG, LBP

Coding: local coordinate, super-vector

Pooling, SPM

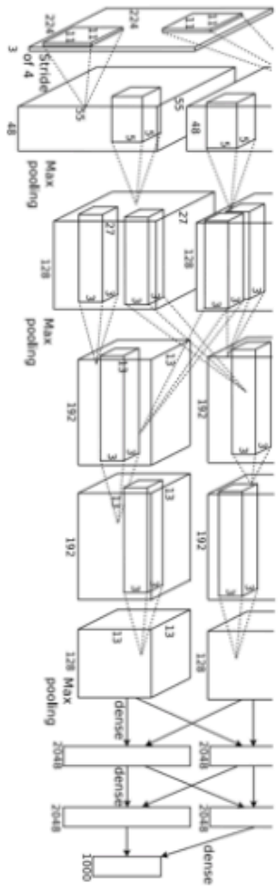
Linear SVM

[Lin CVPR 2011]

Lion image by Swissfrog is licensed under CC BY 3.0

Year 2012

SuperVision

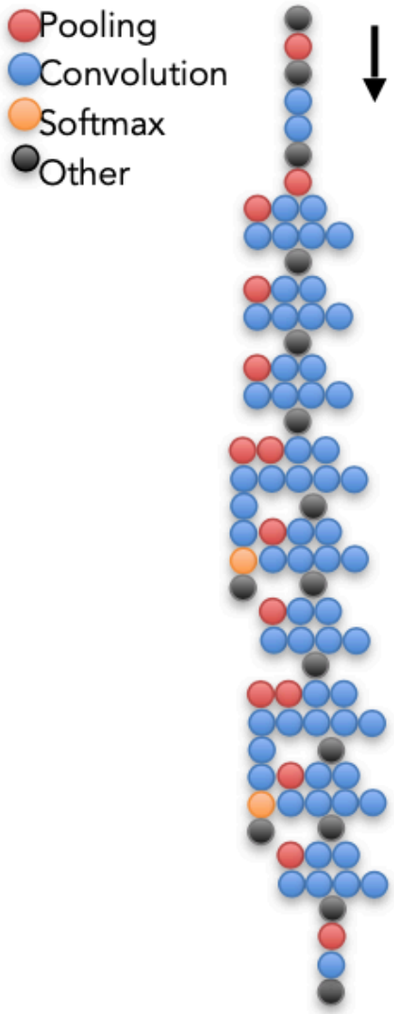


[Krizhevsky NIPS 2012]

Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

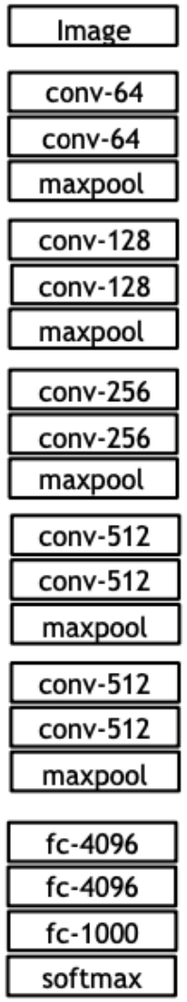
Year 2014

GoogLeNet



[Szegedy arxiv 2014]

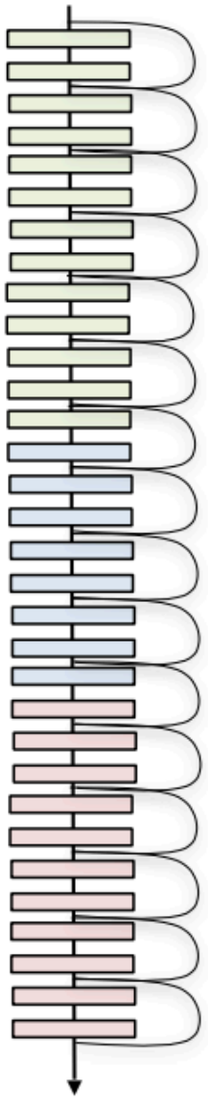
VGG



[Simonyan arxiv 2014]

Year 2015

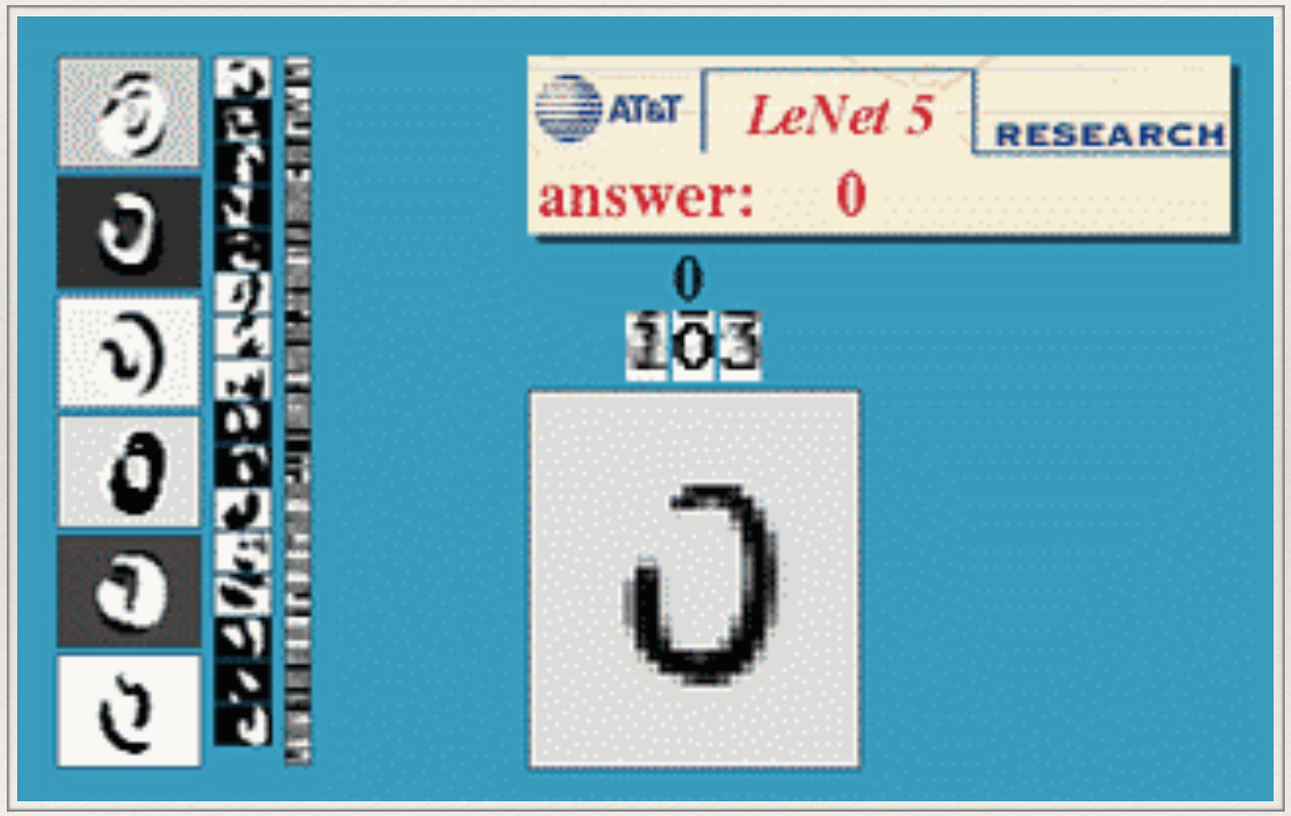
MSRA

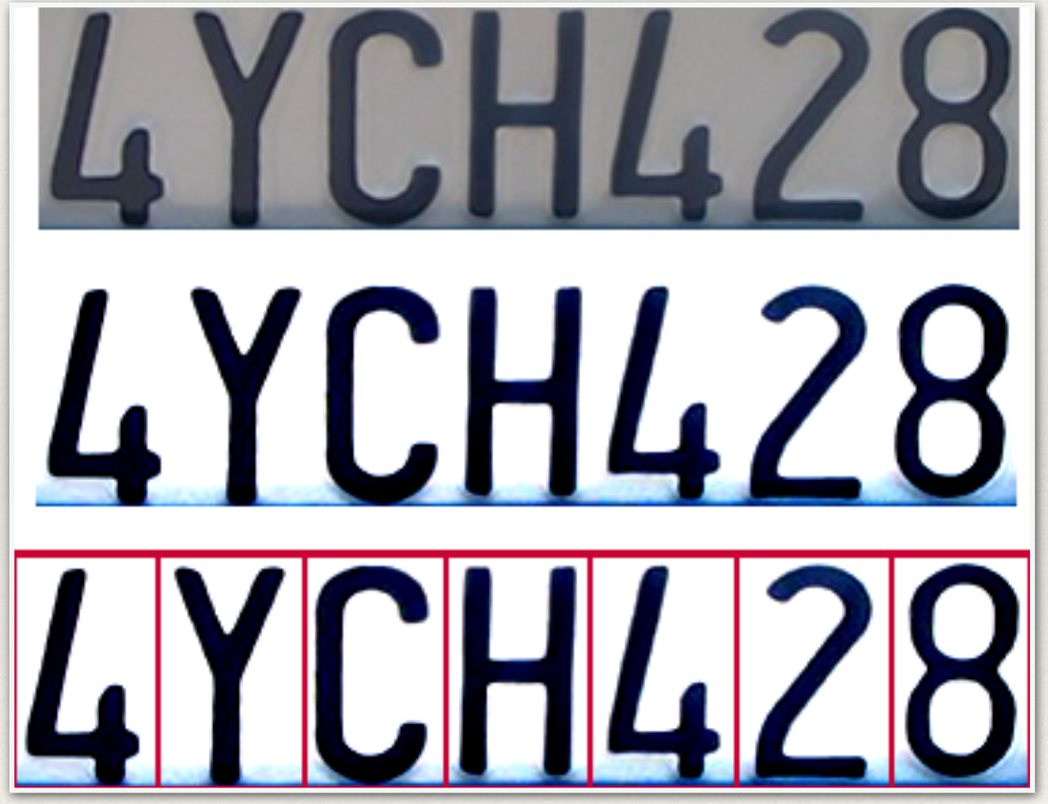


[He ICCV 2015]

Applications: OCR

- Technology to convert images of text into text

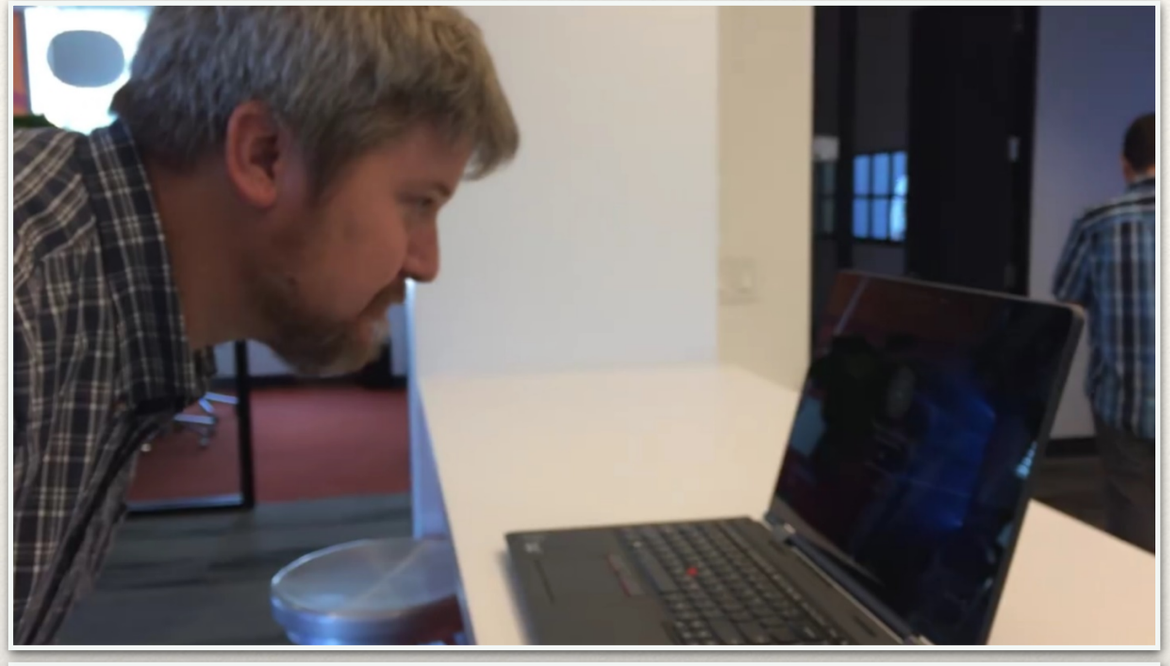


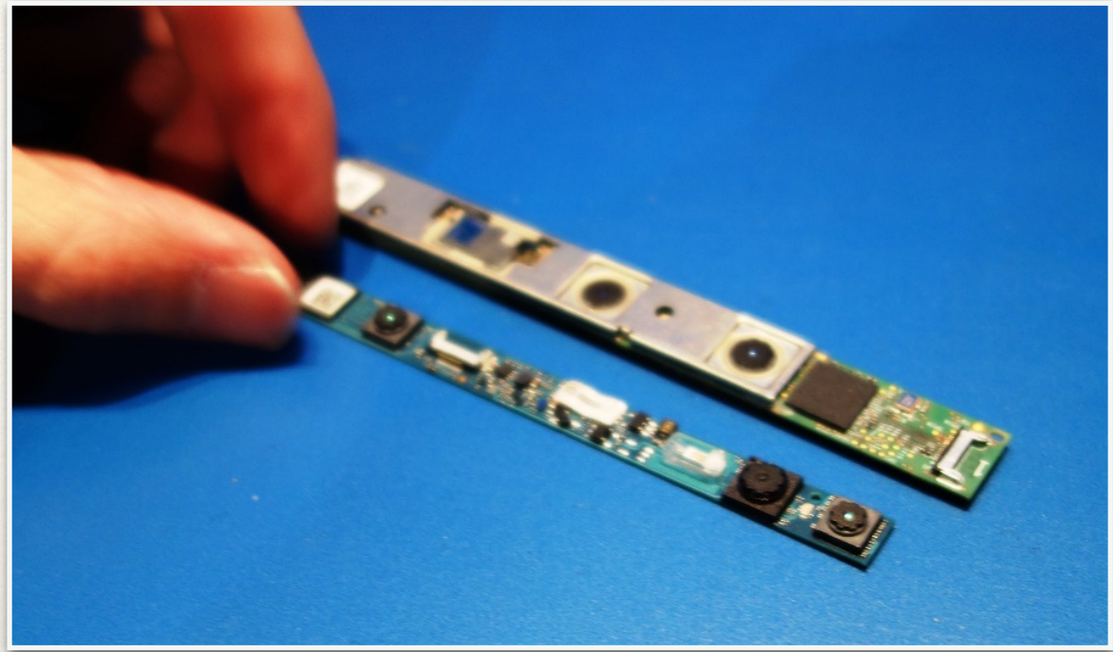


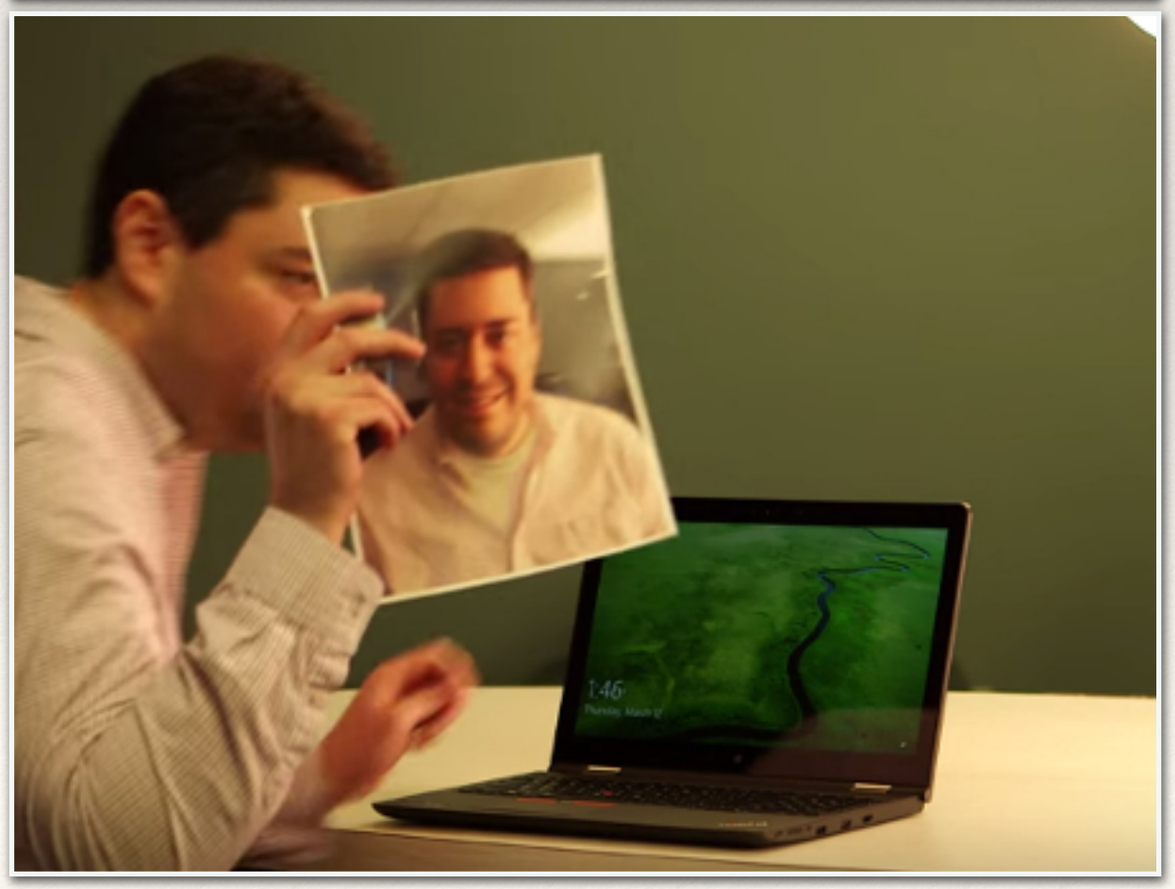
Applications: Vision-based biometrics

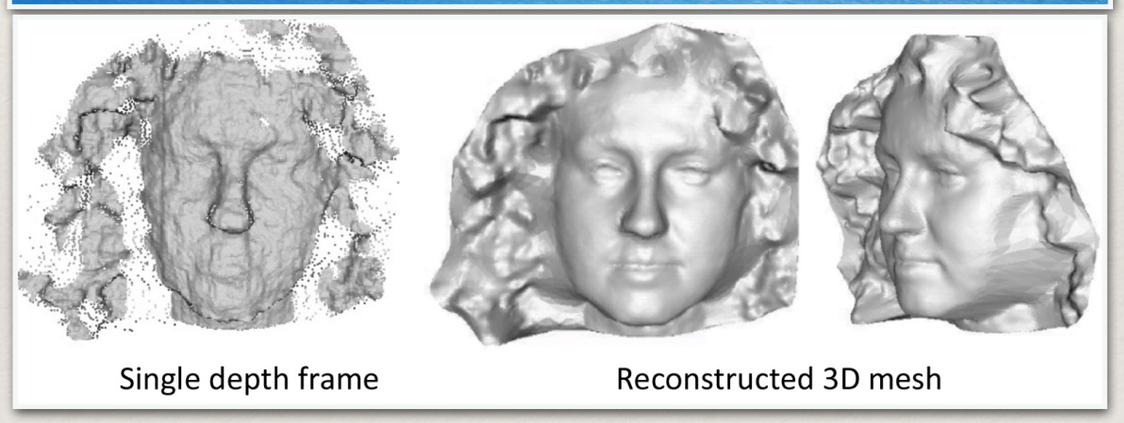
- Face, Fingerprint, Iris, Palm, Gait…

Facial login without a password…



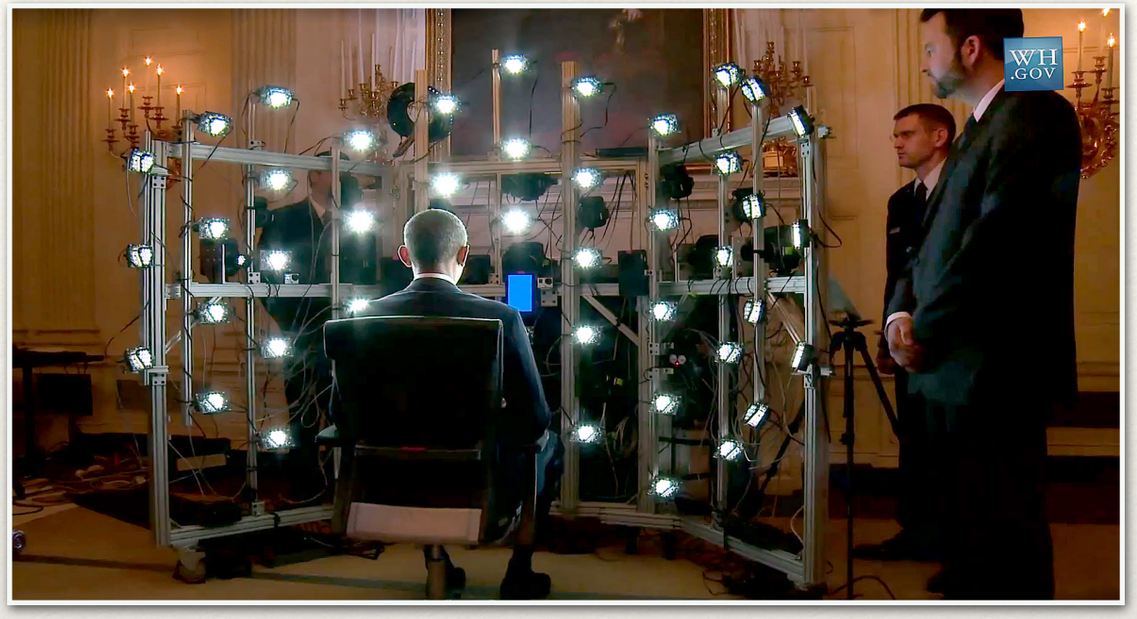


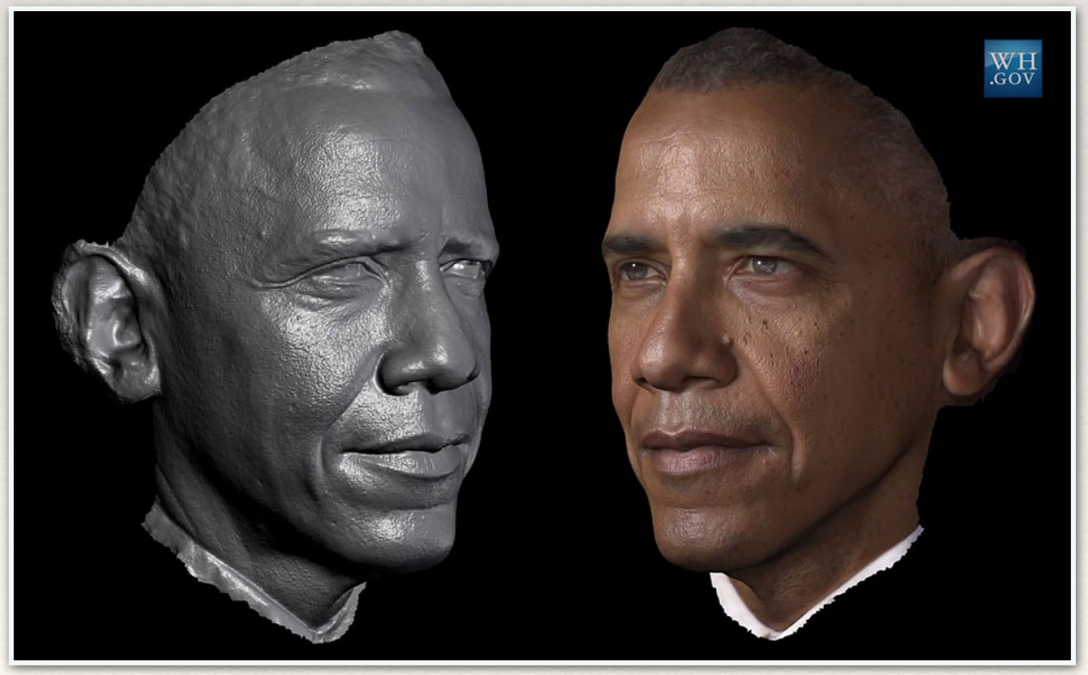


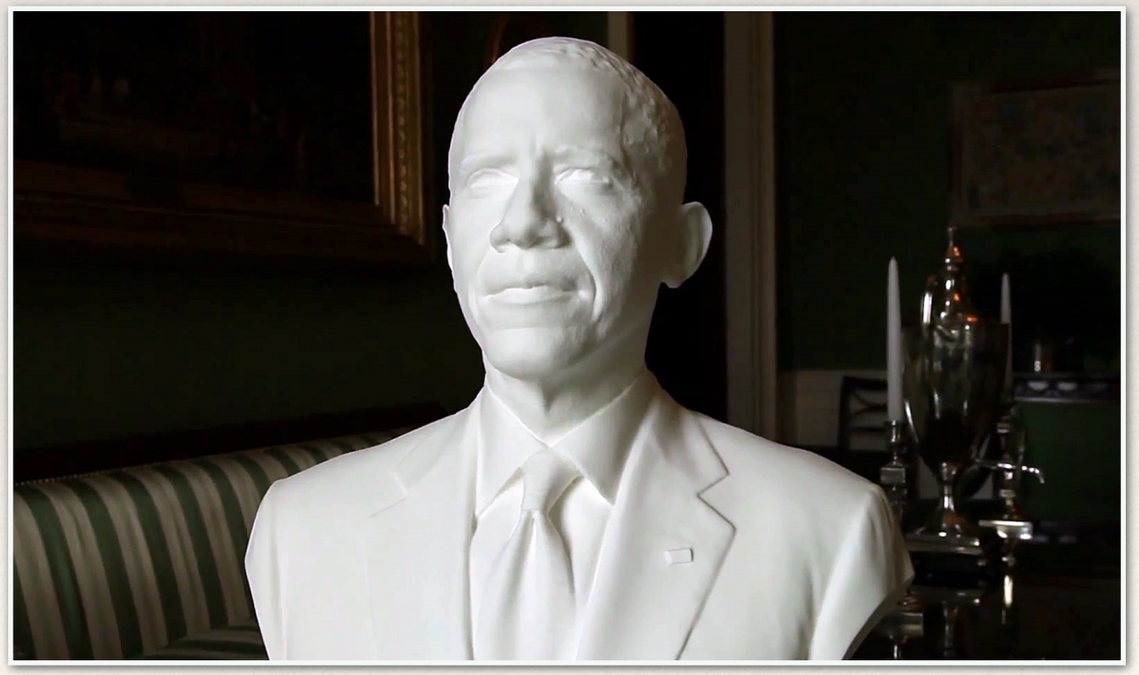


Liang et al. 2014

Applications: Human shape capture



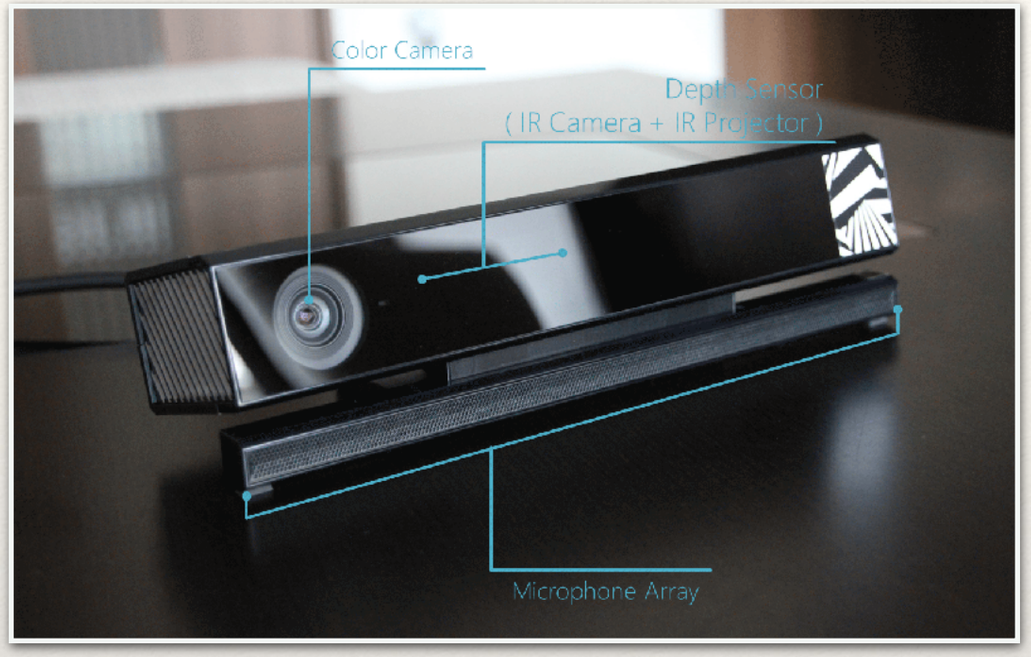




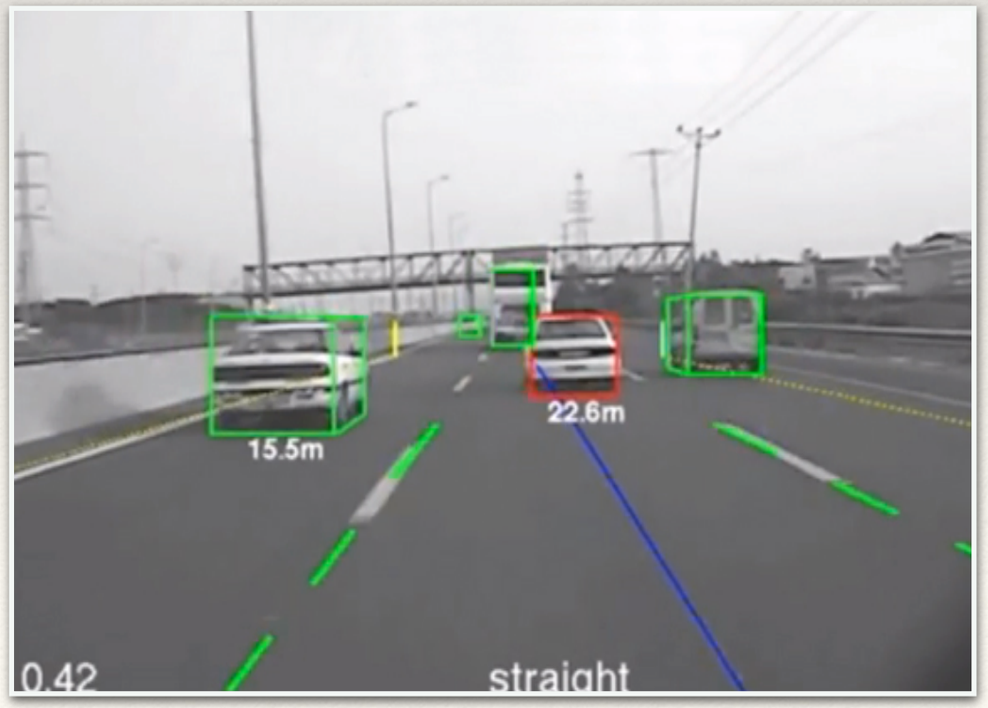
http://gl.ict.usc.edu/Research/presidentialportrait/

Applications: Motion capture

Applications: Human computer interaction



Applications: Autonomous driving

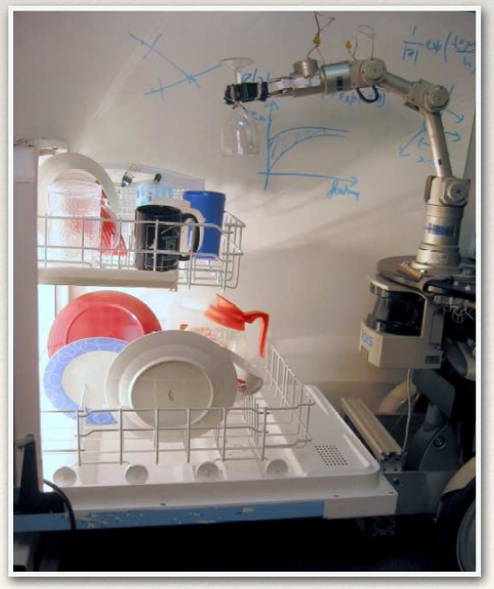


Applications: Industrial robots

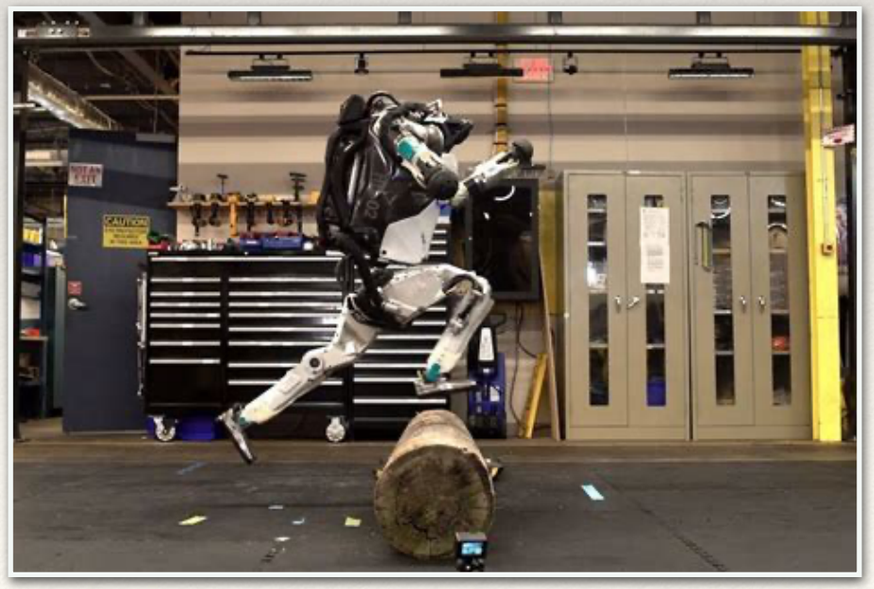
Vision-guided robots position nut runners on wheels

Applications: Mobile robot

Logistic robots



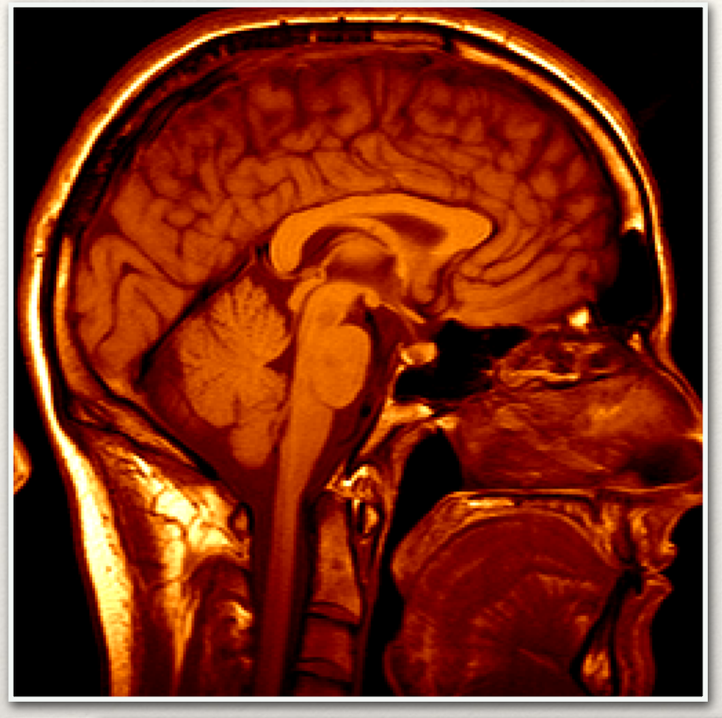
Saxena et al. 2008
STAIR at Stanford



Robotics from Boston dynamics

Applications: Medical

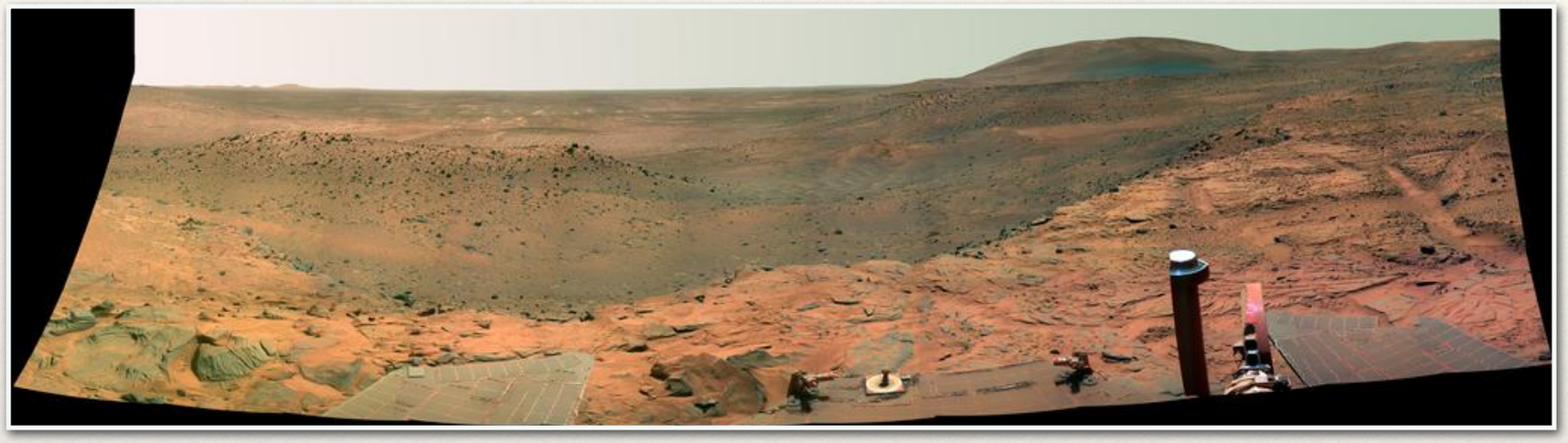
Da Vinci surgical robotics

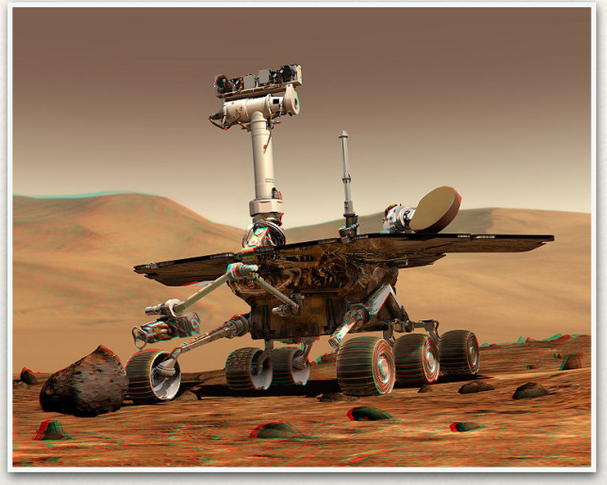


3D imaging
MRI, CT

Image guided surgery
Grimson et al., MIT

Applications: Vision in space





NASA'S Mars Exploration Rover Spirit captured this westward view from atop a low plateau where Spirit spent the closing months of 2007.

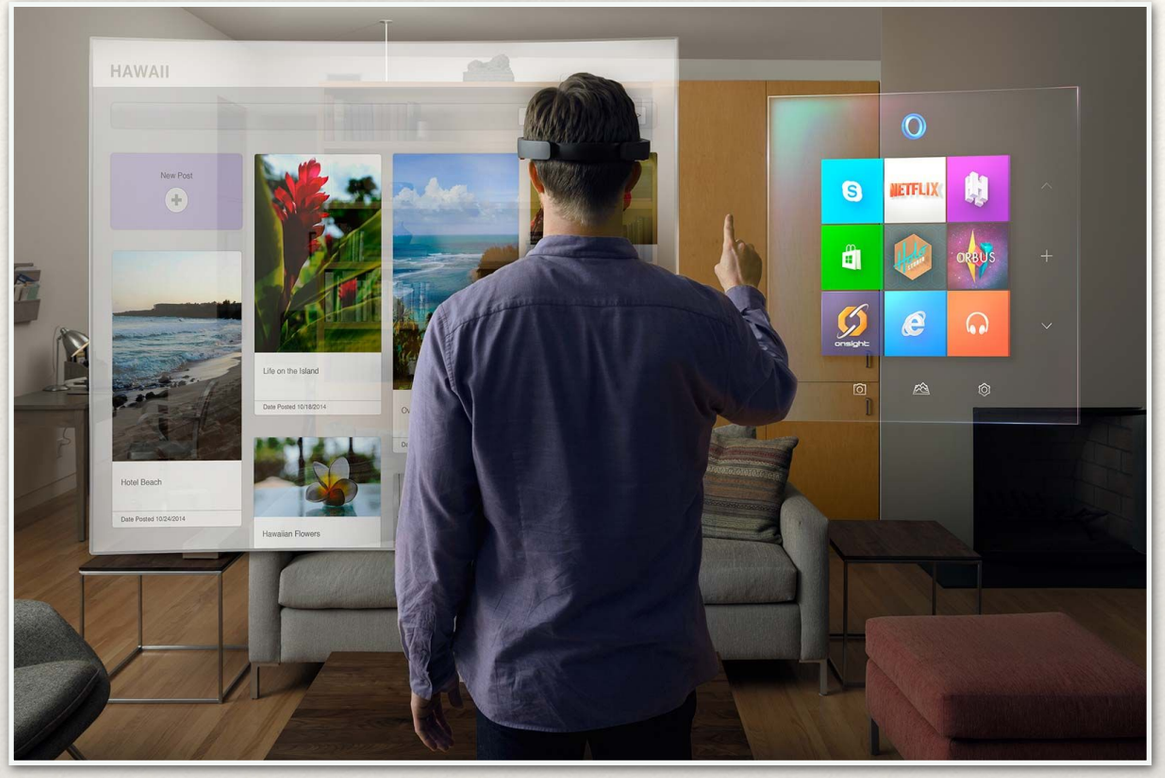
Vision systems used for several tasks

- Panorama stitching
- 3D terrain modeling
- Obstacle detection, position tracking
- For more, read "Computer Vision on Mars" by Matthies et al.

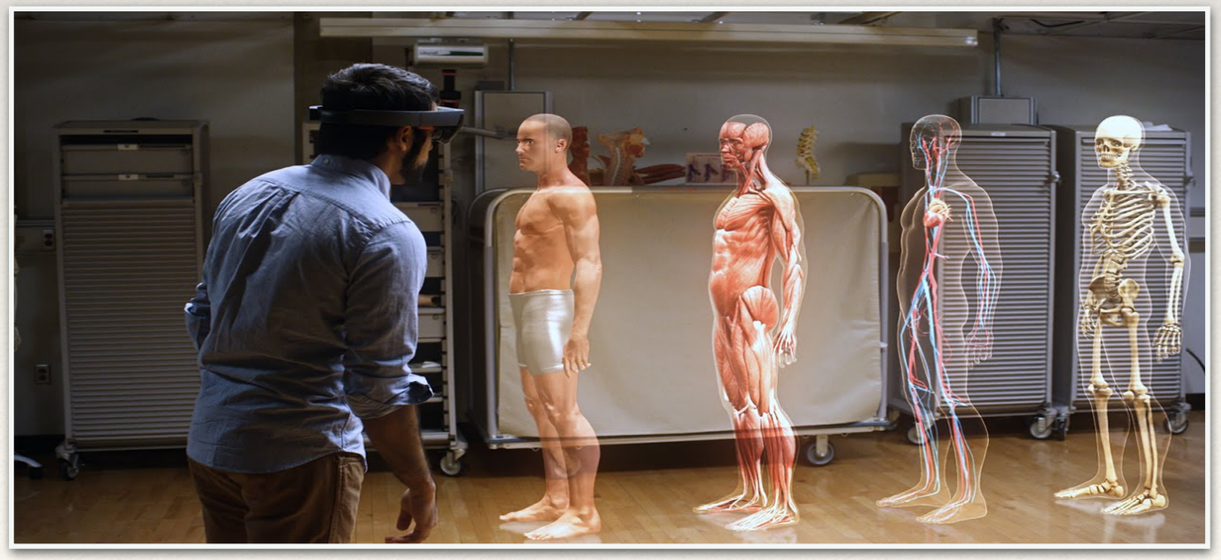
Applications: Sports

Free viewpoint video. Canon 2017

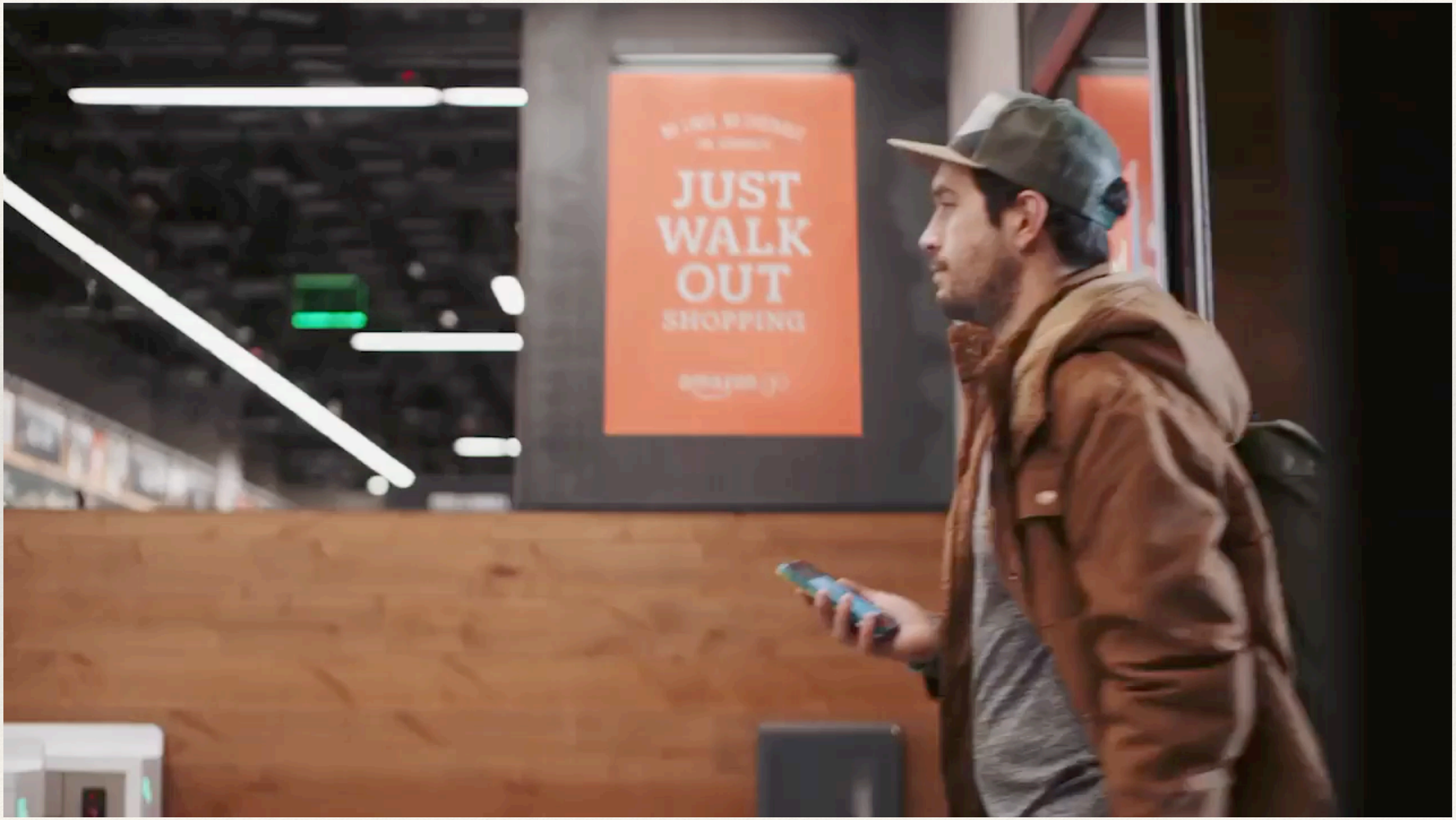
Applications: Augmented Reality and Virtual Reality



MS HoloLens, Oculus, Magic Leap,
ARCore / ARKit , Google glass …

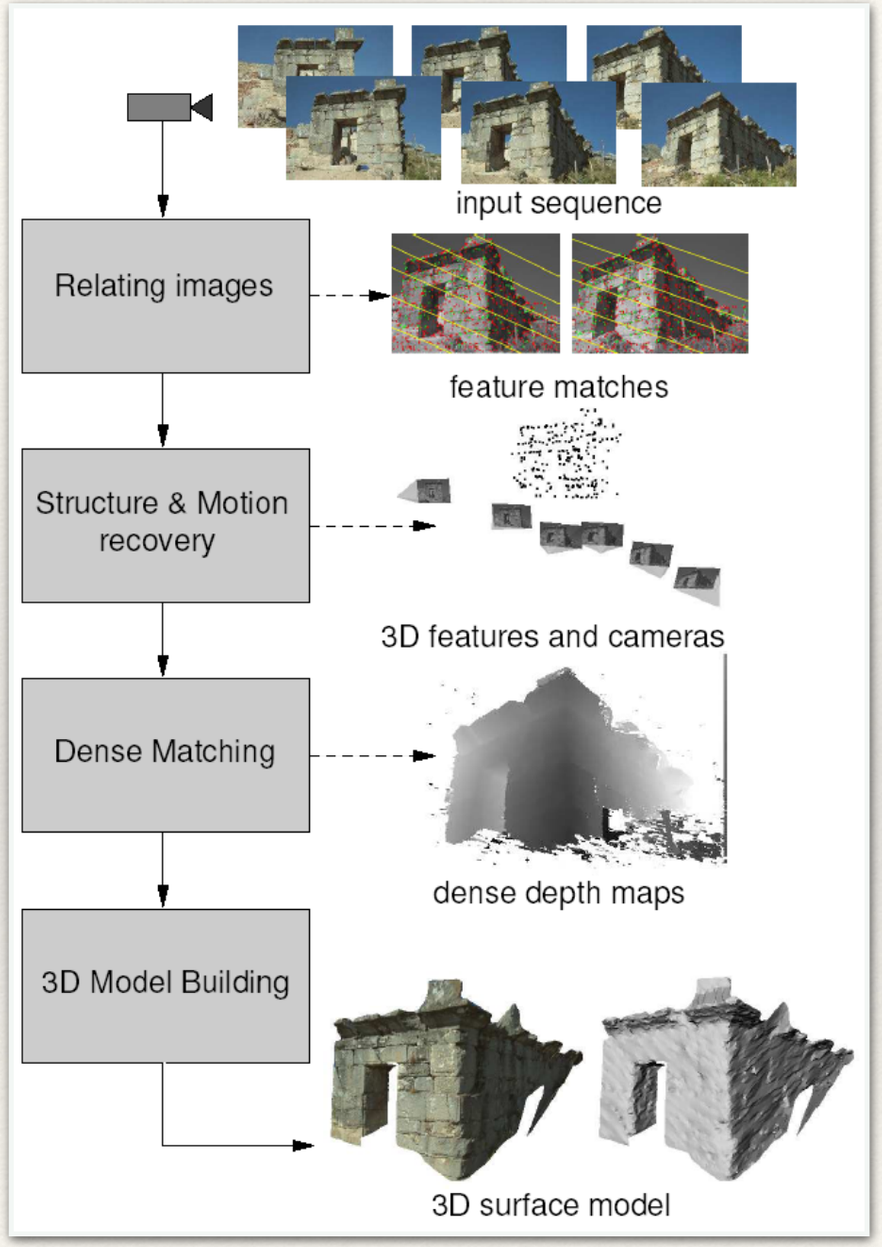


Applications: Automatic retail



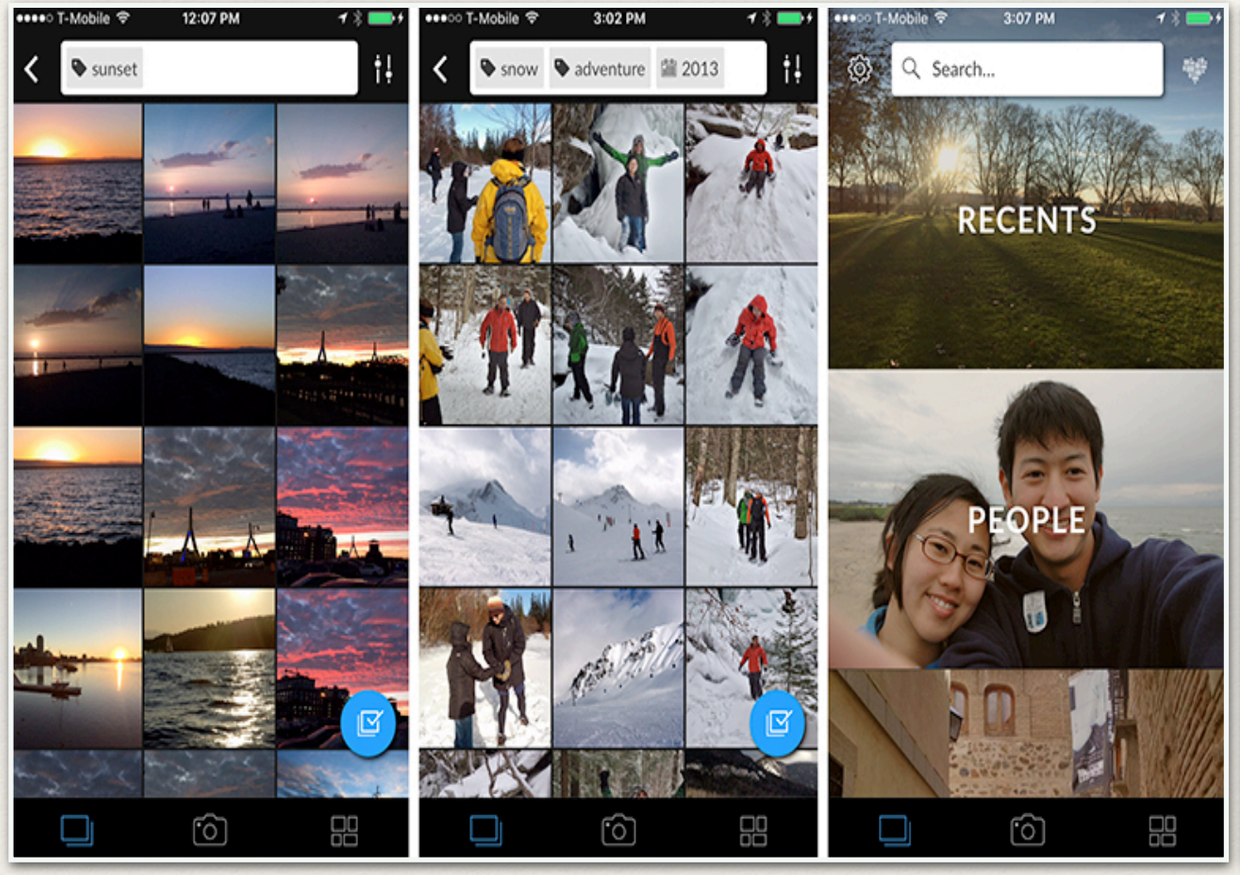
Applications: 3D reconstruction

Building Rome in a Day: Agarwal et al. 2009



Pollefeys et al.

Applications: Advanced photo search



It's a challenging task

- Computer vision is an **inverse problem**: insufficient information to recover some unknowns.
- A lot of efforts need to be made to **disambiguate** between potential solutions.
- Real visual world is more than **complex**, and one can view it through countless viewpoints…
- **Image formation** is a complex function takes input as many variables to transform 3D as 2D.

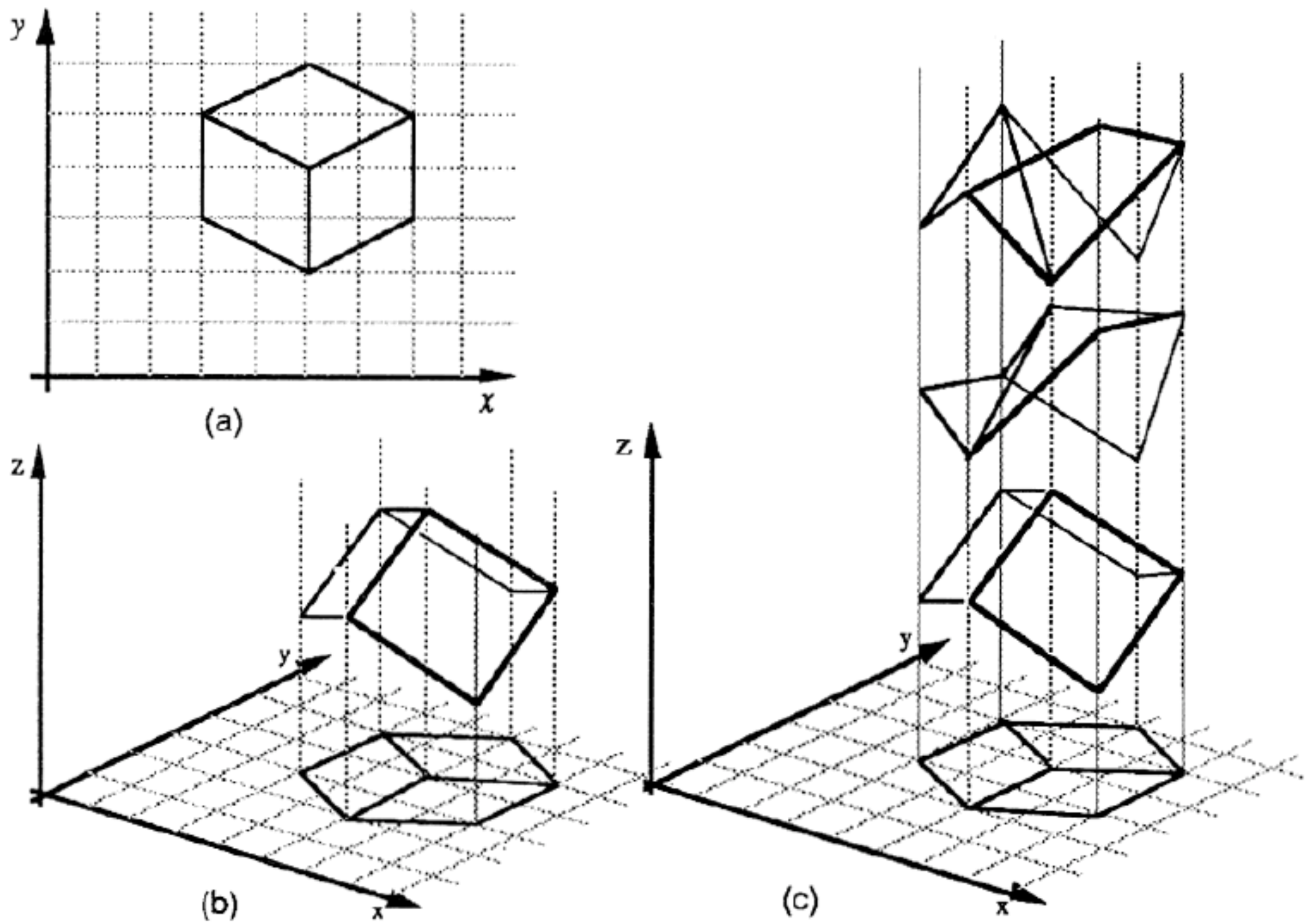
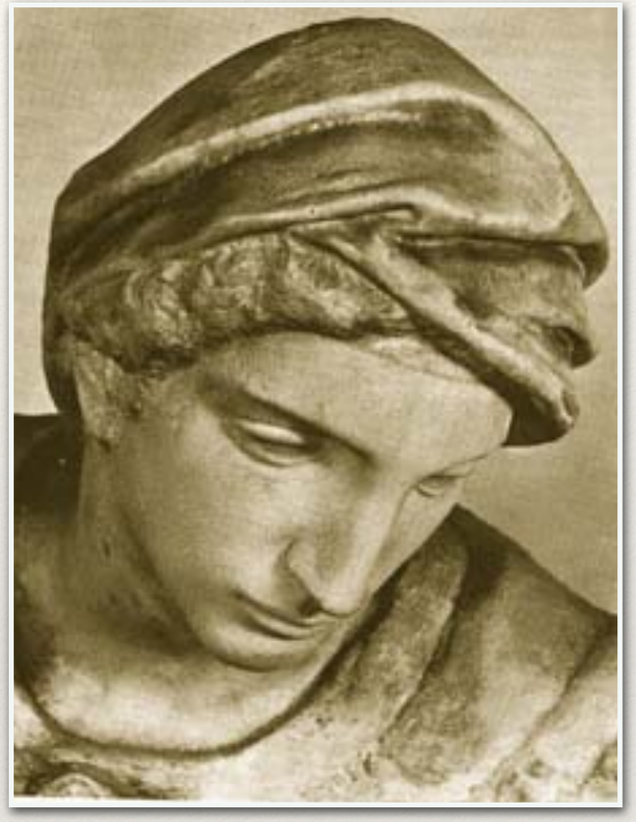


Figure 1. (a) A line drawing provides information only about the x, y coordinates of points lying along the object contours. (b) The human visual system is usually able to reconstruct an object in three dimensions given only a single 2D projection (c) Any planar line-drawing is geometrically consistent with infinitely many 3D structures.

From [Sinha and Adelson 1993]

Challenges: View point variation



Michelangelo 1475-1564

Challenges: Occlusion

Magritte, 1957

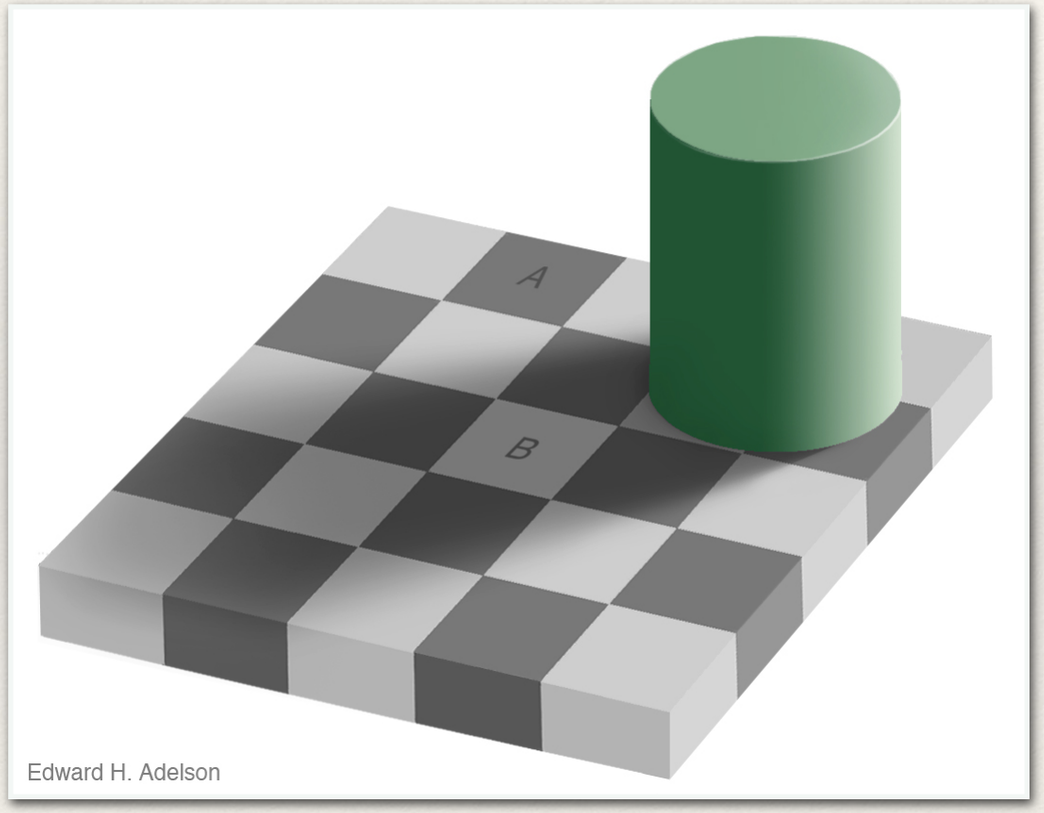
Challenges: Background clutter

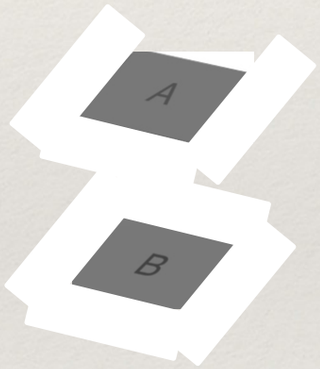
Klimt, 1913

Challenges: Object intra-class variation

Slide by Fei-Fei, Fergus & Torralba

Challenges: Measurement vs. perception





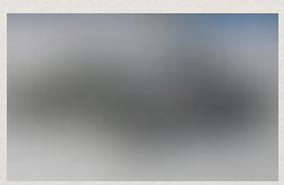
Challenges: Illumination

Challenges: Scale

Challenges: Deformation

Mr. Bean

Challenges: Local appearance ambiguity



Slide by Fei-Fei, Fergus & Torralba

Course information

- All in **Canvas** https://oc.sjtu.edu.cn/login/canvas
  - Sildes
  - Video
  - Assignments
  - Discussion
  - Announcement
- Keeping in touch: Wechat group
- Office hours:
  - Time: 16:00-17:00, Thursday
  - Location: 2-431, SEIEE Buildings

Grading policy

- 3 Problem sets: $15\% \times 3 = 45\,\%$
- Attendance: $10\,\%$
- Final project: $45\,\%$
  - Team work: **4 people** each
  - Proposal: introduce your idea with 1-2 pages
  - Final report: CVPR paper format
  - Final presentation: 10 mins each project

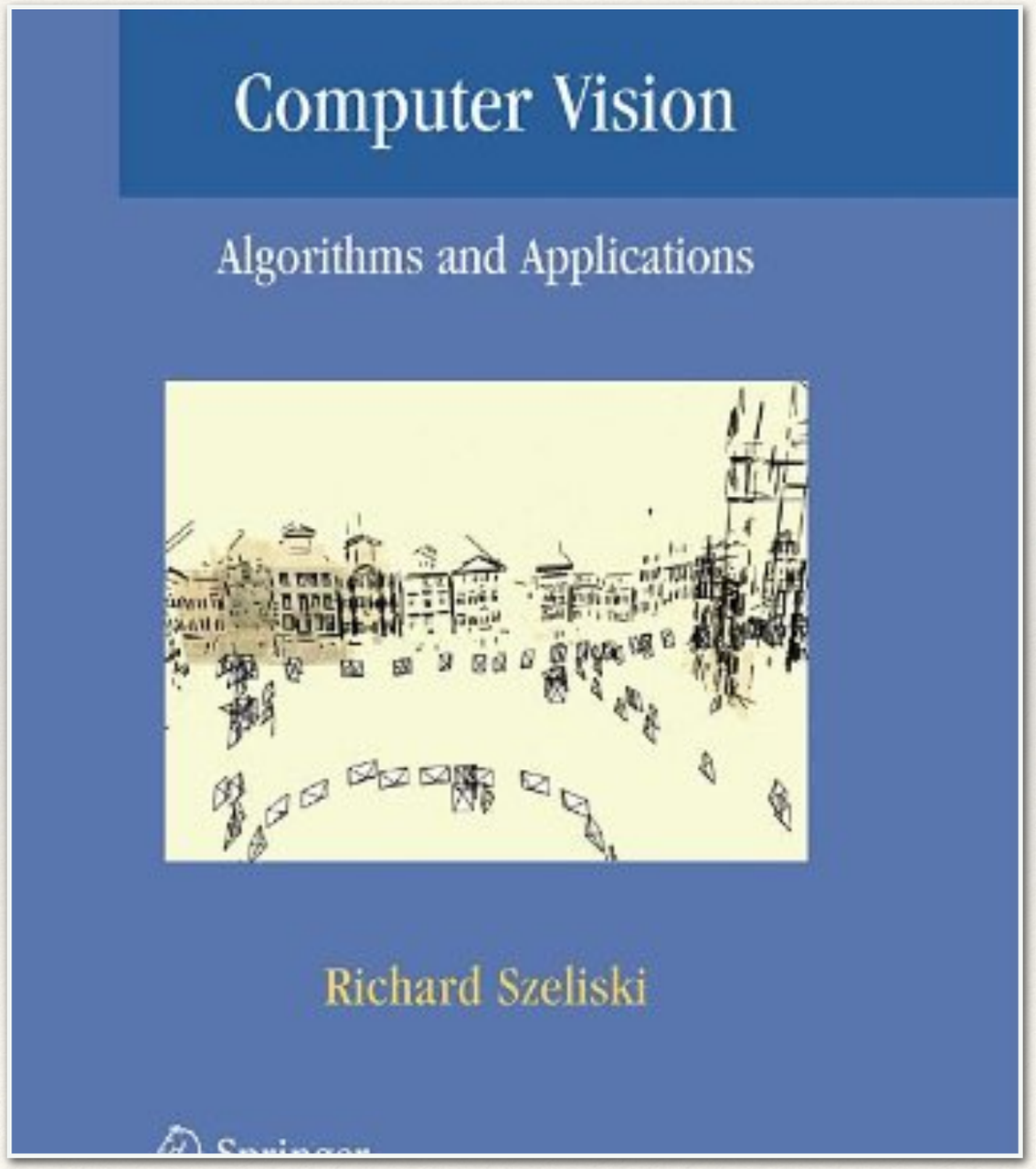
Syllabus

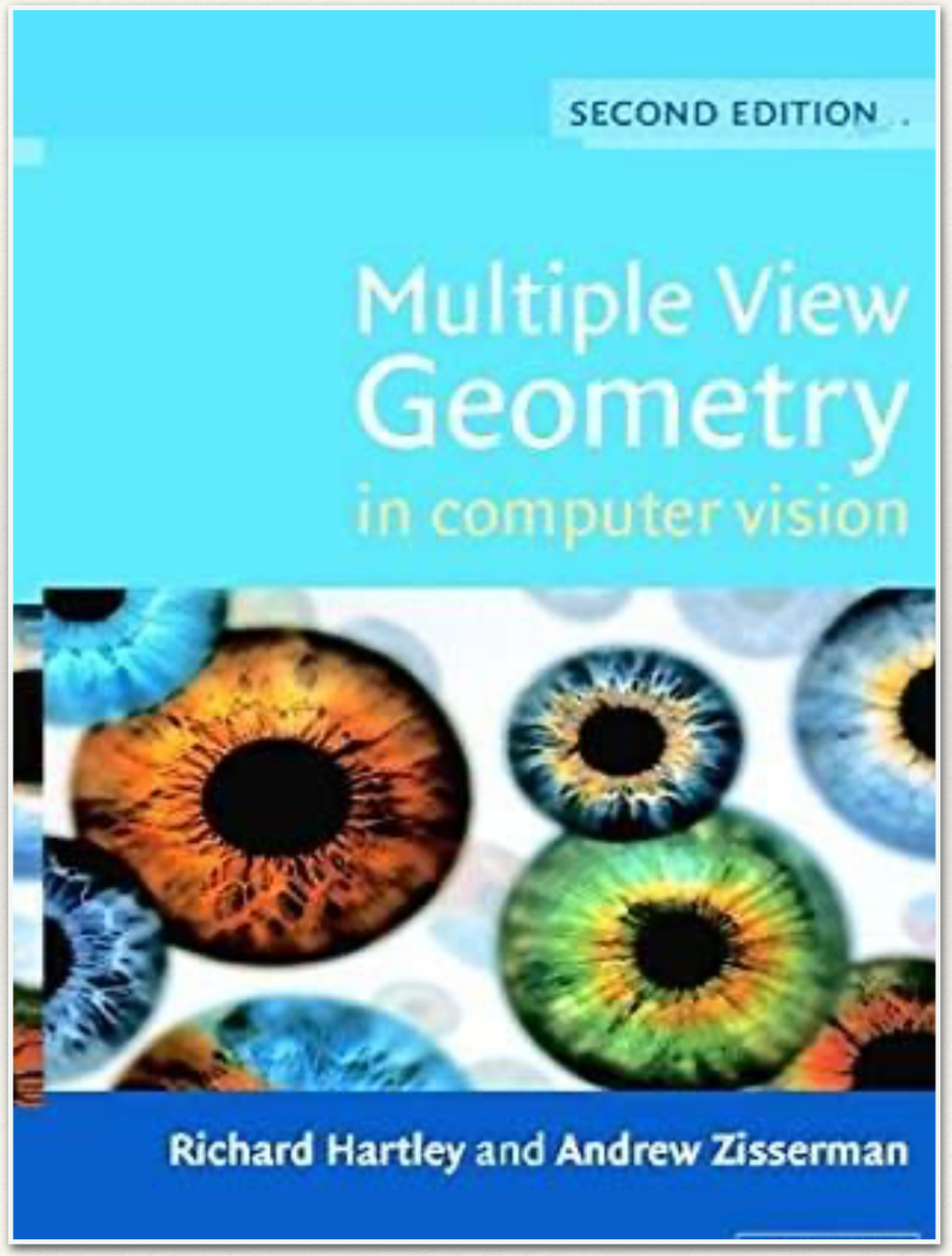
Event type	Contents	Hours	Week No.
Unit 0	**Introduction**	**3**	**1**
Lecture 1	Introduction	3	1
Unit I	**Image formation**	**3**	**2**
Lecture 2	Geometric fundamentals, camera and calibration	3	2
Unit II	**Visual features: low-level features**	**9**	**3-5**
Lecture 3	Local image features	3	3
Lecture 4	Motion features: optical flow and beyond	3	4
R&P	TBD	3	5

R&P: Reading and Practicing

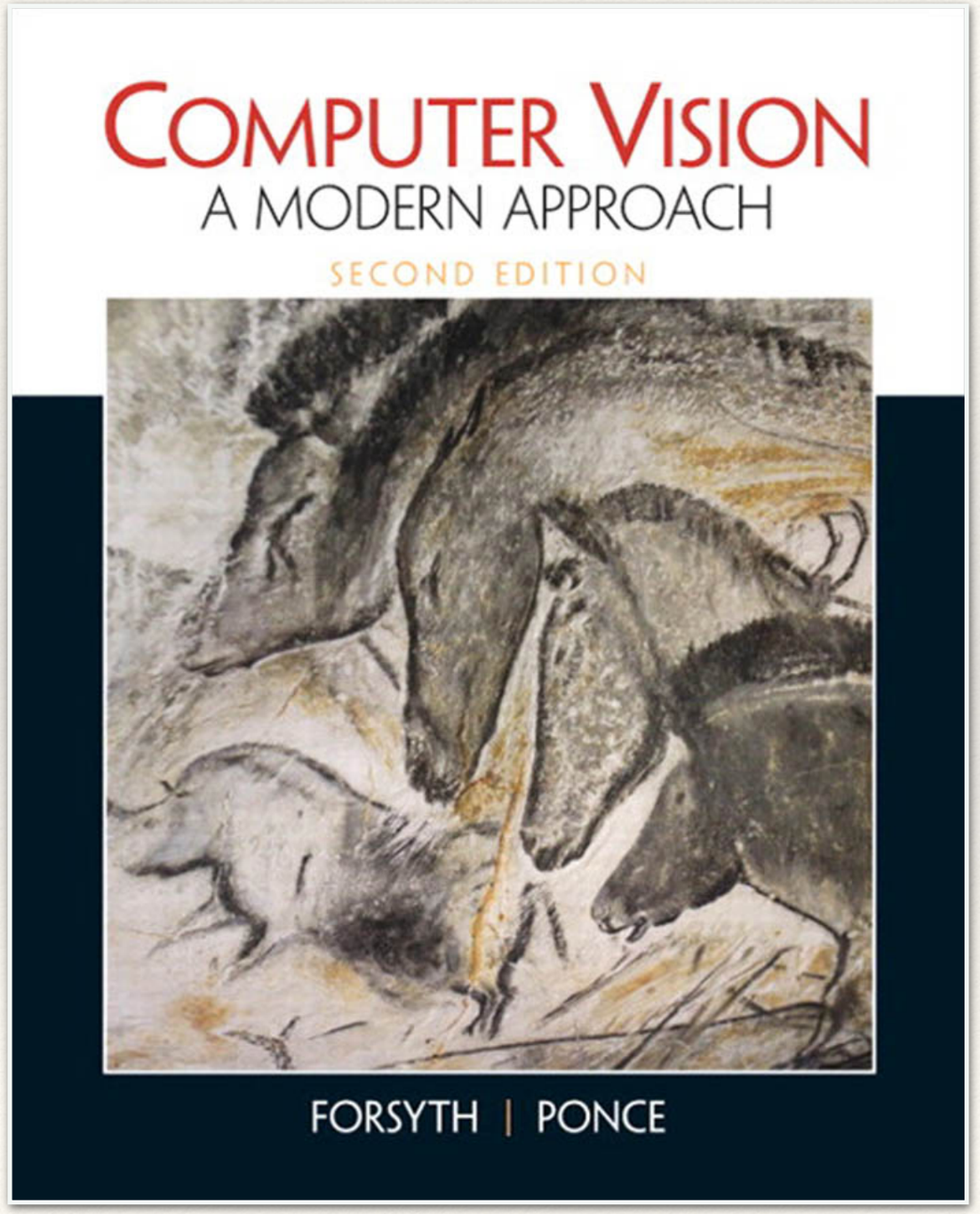
Event type	Contents	Hours	Week No.
Unit III	**Multiple views**	**9**	**6-8**
Lecture 5	Stereopsis	3	6
Lecture 6	Structure from motion	3	7
Lecture 7	Registration	3	8
Unit IV	**Grouping and fitting: mid-level visions**	**9**	**9-11**
Lecture 8	Segmentation	3	9
Lecture 9	Fitting	3	10
R&P	TBD	3	11
Unit V	**Recognition: high-level vision**	**9**	**12-14**
Lecture 10	Deep learning and CNN	3	12
Lecture 11	Object detection	3	13
Lecture 12	Object/scene/activity categorization	3	14
Final project	**Project presentation and evaluation**	**6**	**15-16**

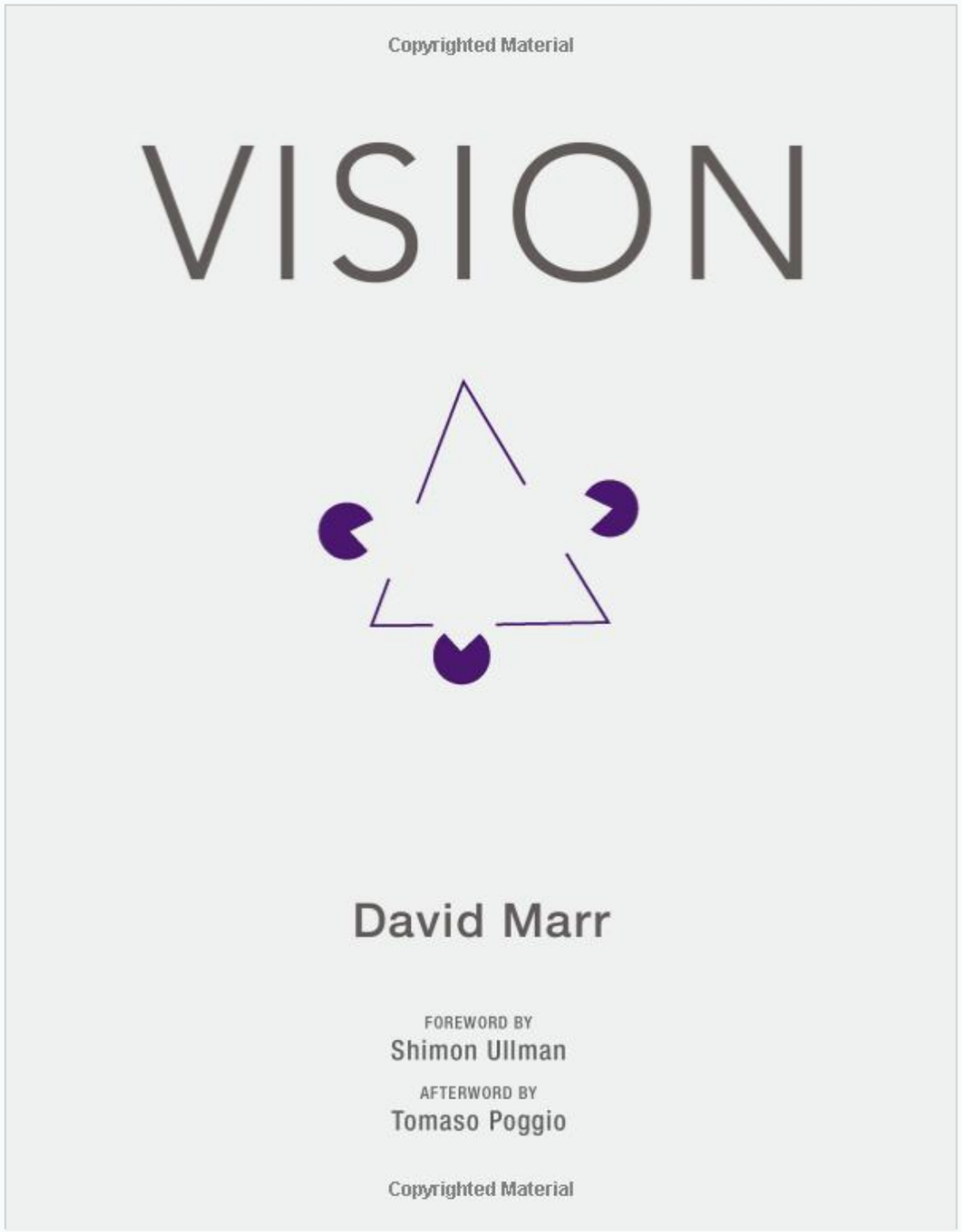
Reference


Computer Vision
Algorithms and Applications
Richard Szeliski

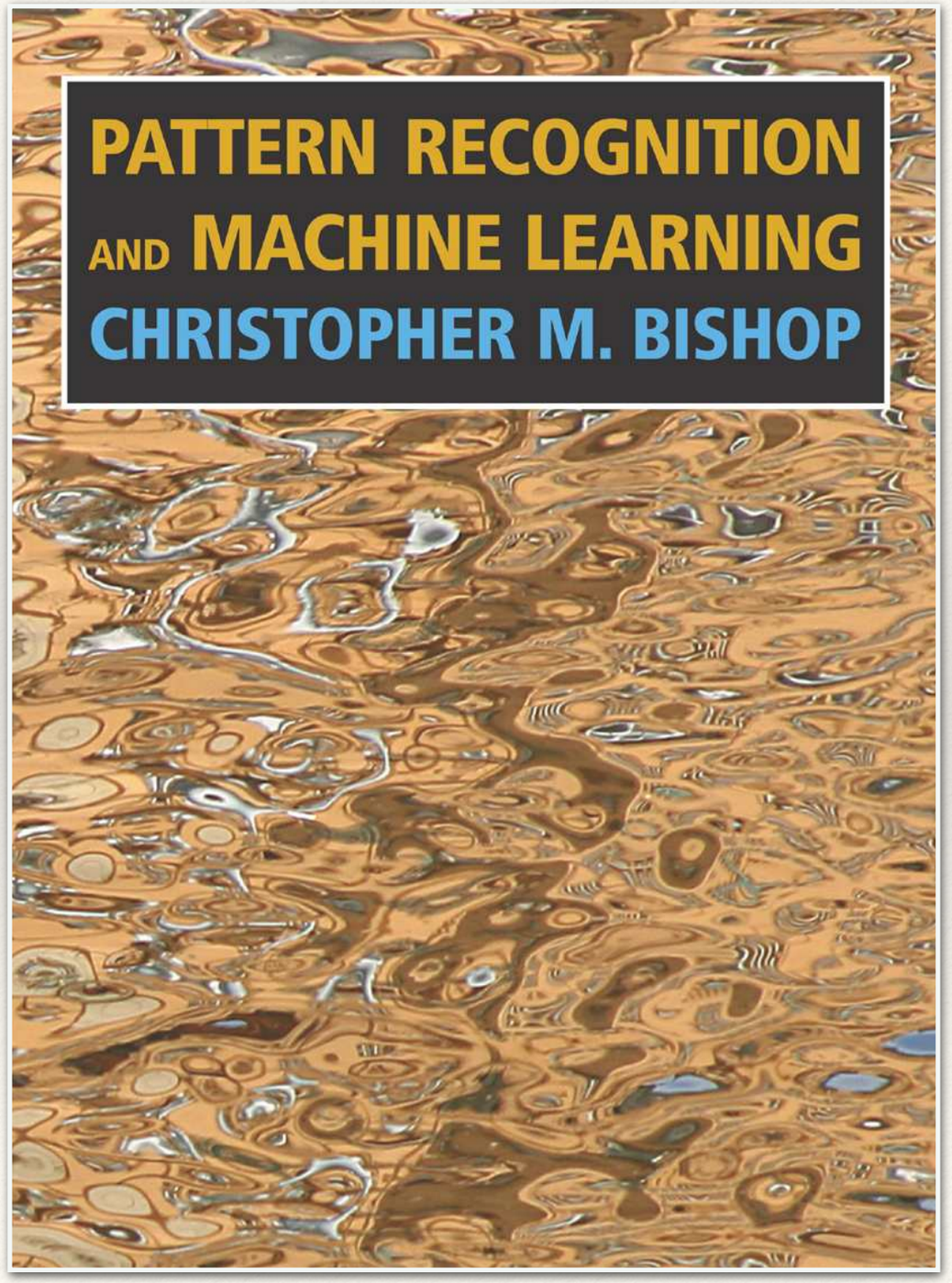

SECOND EDITION
Multiple View Geometry
in computer vision
Richard Hartley and Andrew Zisserman

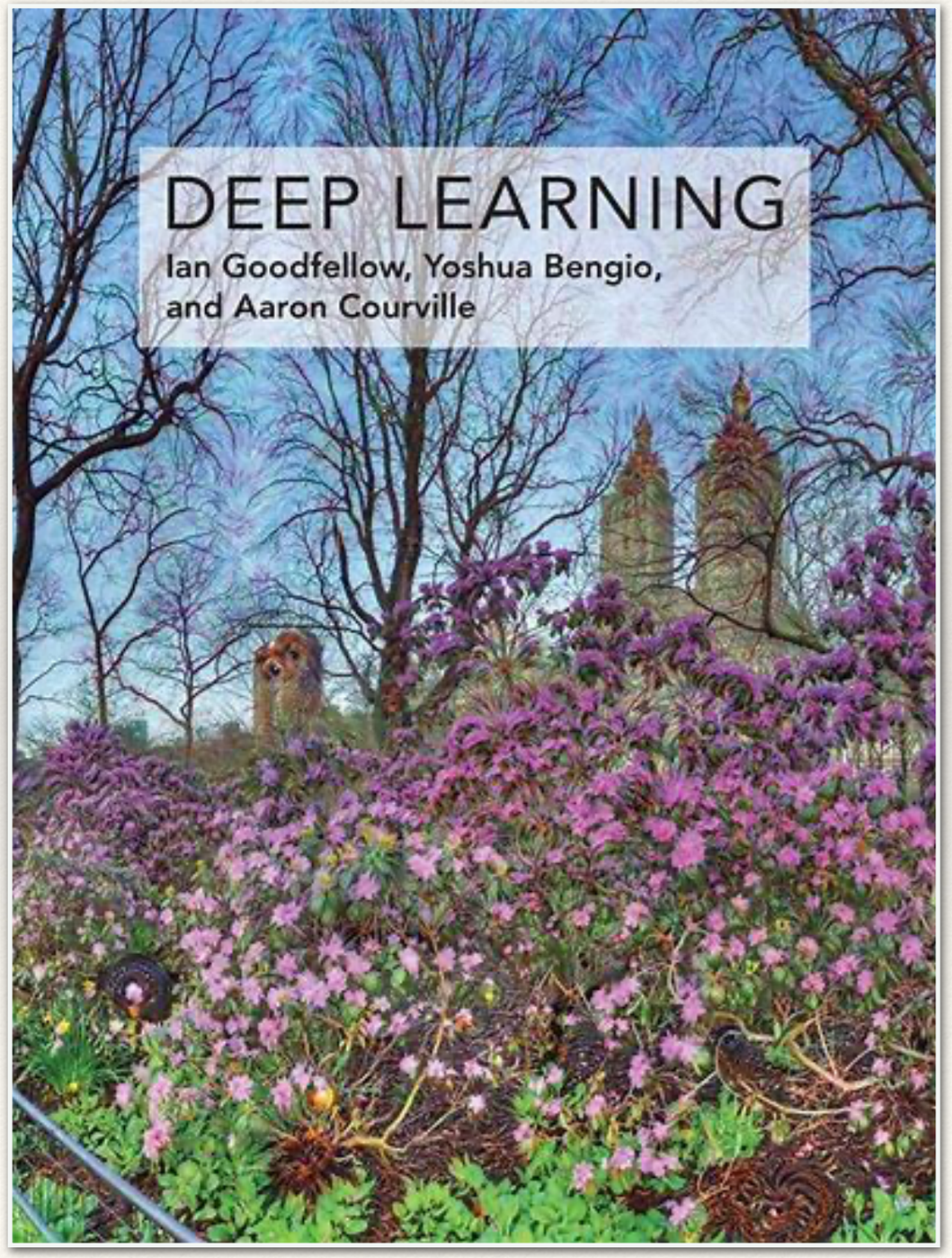
Reference


COMPUTER VISION
A MODERN APPROACH
SECOND EDITION
FORSYTH | PONCE


Copyrighted Material
VISION
David Marr
FOREWORD BY
Shimon Ullman
AFTERWORD BY
Tomaso Poggio
Copyrighted Material

Reference


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio,
and Aaron Courville

Other resources

- Conference: CVPR, ICCV, ECCV, ACCV, AAAI,…
- Journal: IEEE TPAMI, IJCV, IEEE TIP,…
- arXiv