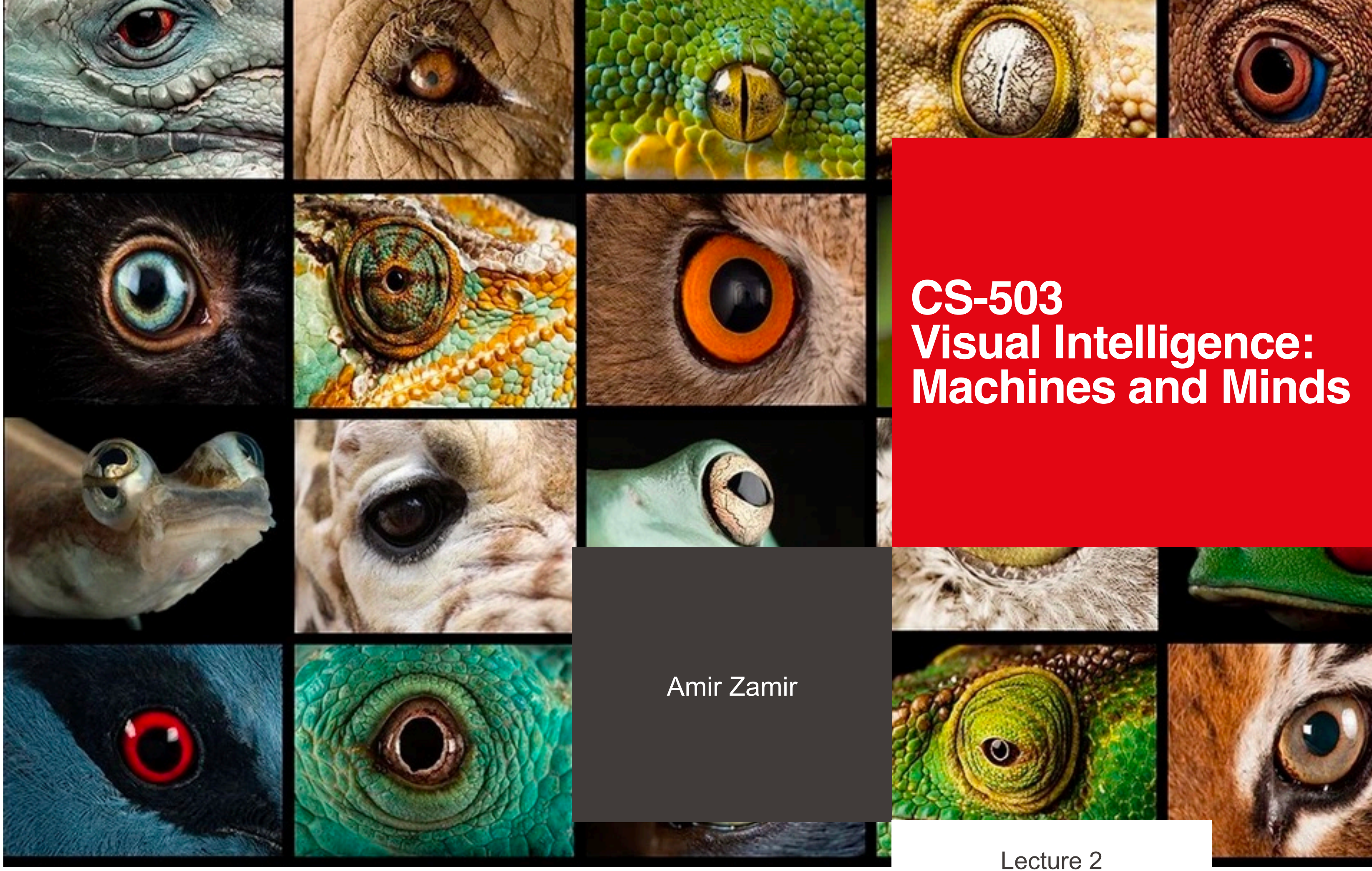


CS-503
Visual Intelligence: Machines and Minds

Amir Zamir

Lecture 2

École
polytechnique
fédérale
de Lausanne

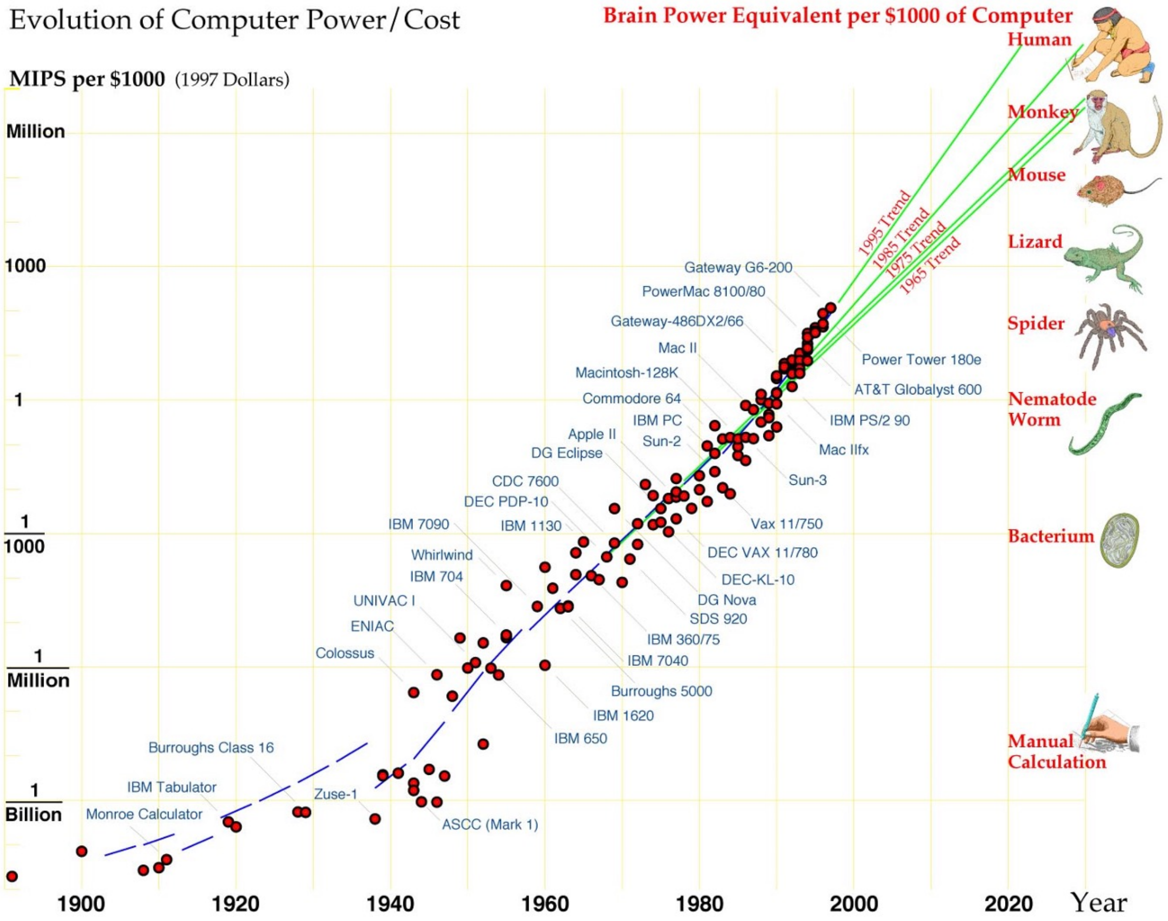
Recaps relevant to vision today

- **Data-driven/learning:**
  - e.g., Neural networks, AlexNet, etc.
- **Non-data-driven**
  - e.g., Image formation model, Image transformation, etc.
- “Method” recap, as opposed to historical credit assignment.

Fast Historical Recap

Historical review

Bigger picture



Moravec 1998

Historical review

- Emission Theory (ca. 400 BC)

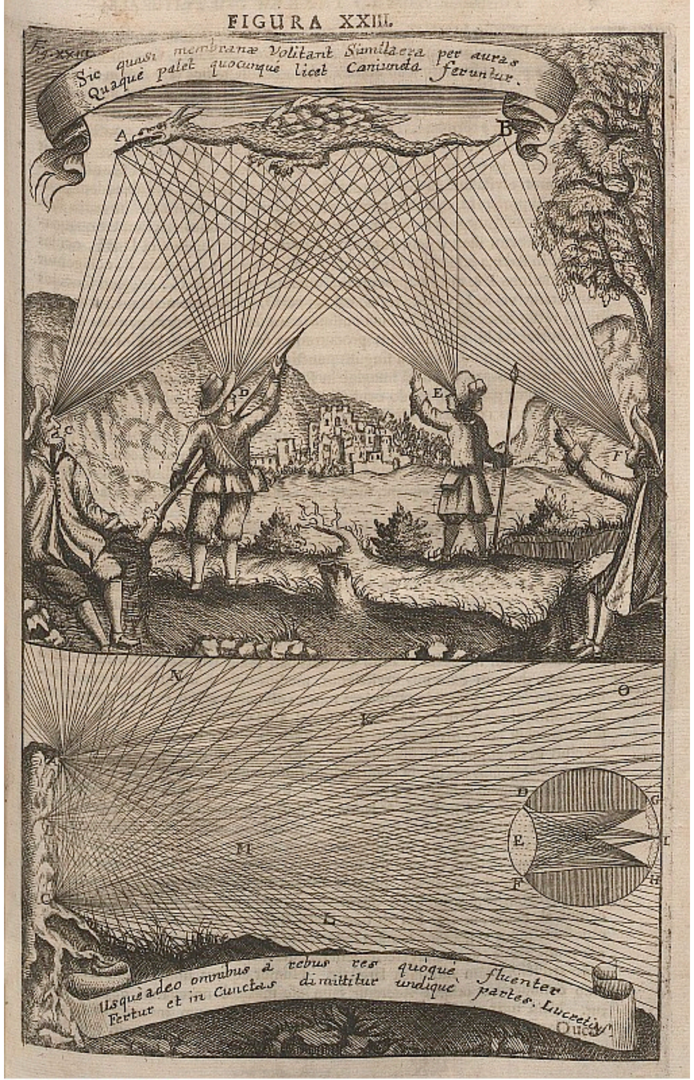
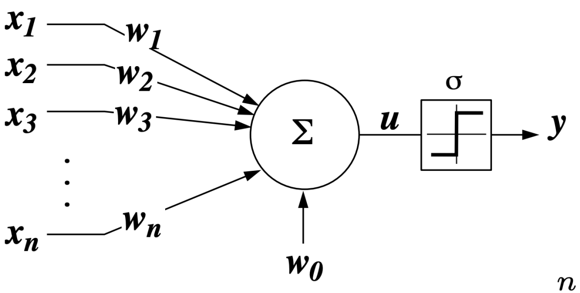


FIGURA XXIII.
Fig. XXIII.
Sic quasi membranæ Volitant Simulacra per auras
Quaquê patet quocunquê licet Canisunda feruntur.
A
B
C
D
E
F
N
O
K
M
L
D
E
F
G
I
H
Usquè adeo omnibus à rebus res quòquê fluenter
Fertur et in Cunctas dimittitur undiquè partes. Lucretius

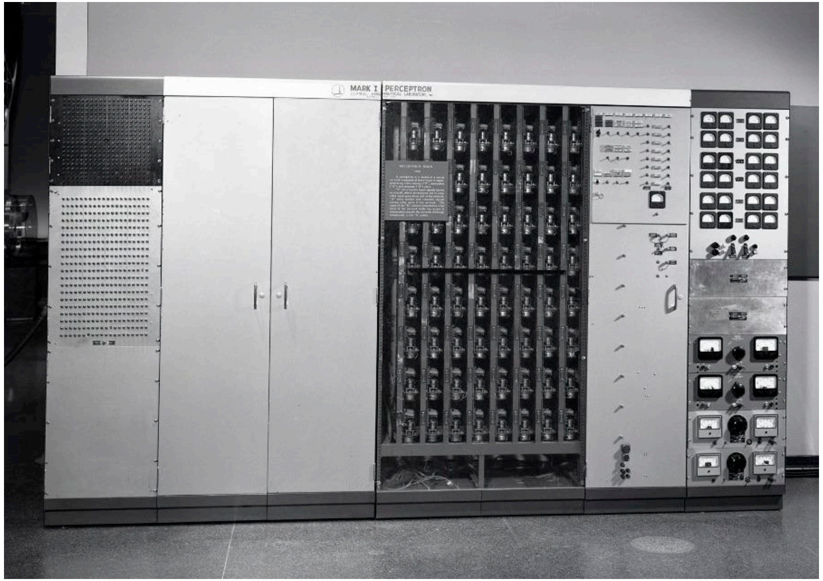
Historical review (neural networks)

- Perceptron model, Rosenblatt, 1958.



$$u = w_0 + \sum_{i=1}^{n} w_i\, x_i$$

$$y = \sigma(u)$$

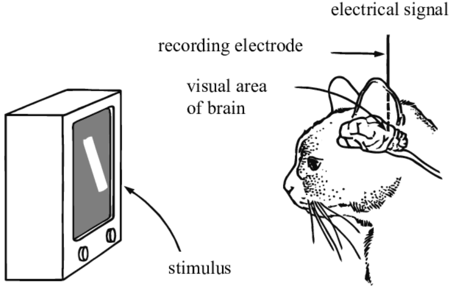


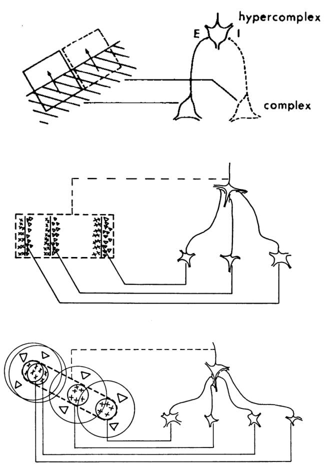
Historical review (neural networks)

- Perceptron model, Rosenblatt, 1958.

Historical review (neural networks)

- Hubel and Wiesel, ~1962.





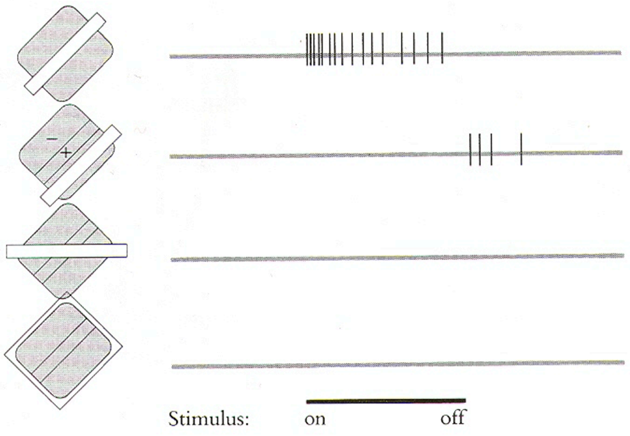
Hypercomplex

↑

Complex

↑

Simple



Historical review (neural networks)

- Neocognitron, Fukushima, 1980.
    - Modeled after Hubel and Wiesel.
    - Convolutional.
    - Multi-layer.
    - Hebbian Learning (no backpropagation).

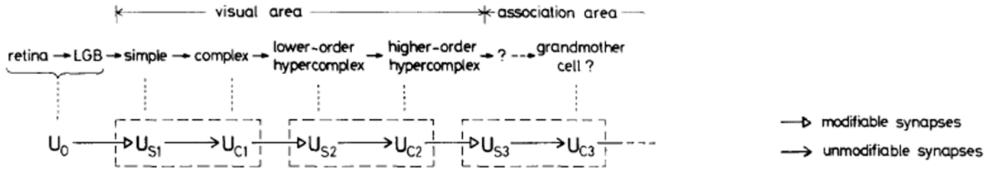


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

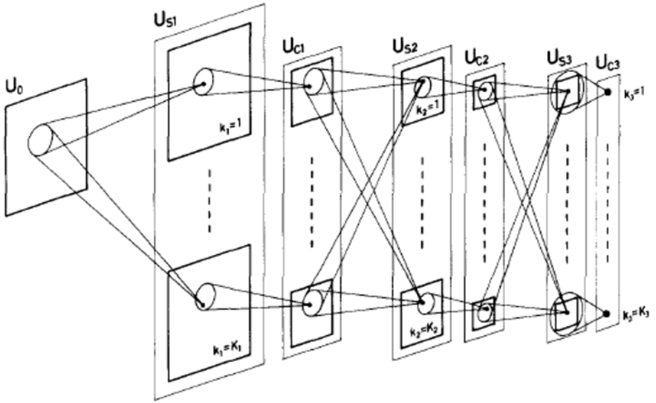
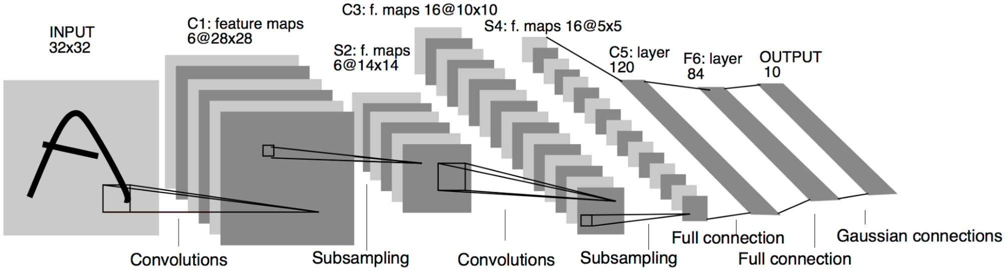


Fig. 2. Schematic diagram illustrating the interconnections between layers in the neocognitron

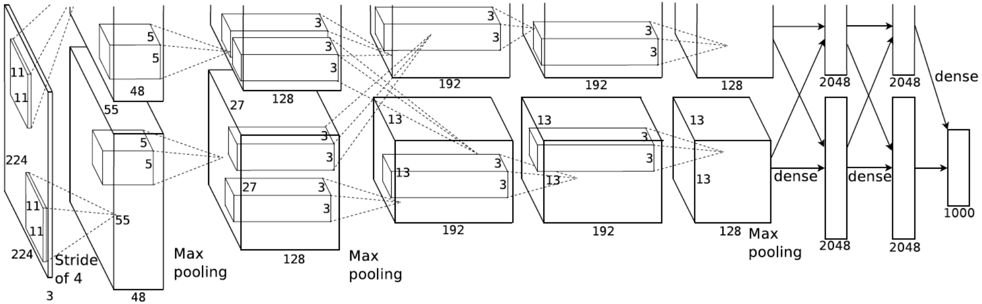
Historical review (neural networks)

- Backpropagation based learning. LeCun, Hinton, and co ~1989.
    - Backpropagation, as a method, already existed.
    - Most successful for handwritten digit recognition. Didn’t work for standard vision tasks.



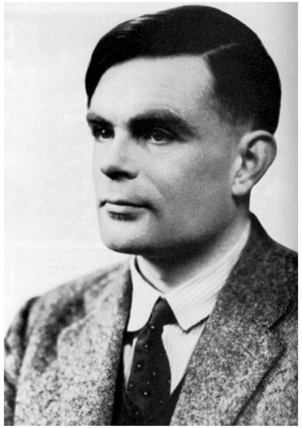
Historical review (neural networks)

- AlexNet, Krizhevsky, Sutskever & Hinton. 2012
    - Success at a standard vision task (ImageNet)
    - Deep Learning wave.
    - GPUs



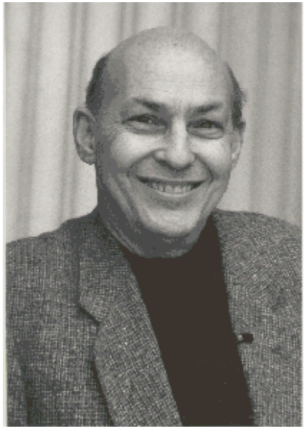
Historical review

- “AI”



Alan Turing

John von Neumann



Marvin Minsky

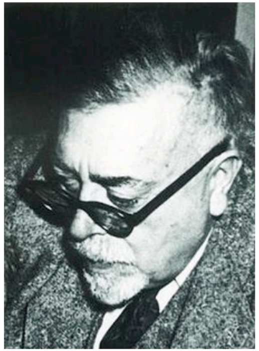
John McCarthy

Among the most challenging scientific questions of our time are the corresponding analytic and synthetic problems: How does the brain function? Can we design a machine which will simulate a brain?
-- *Automata Studies*, 1956

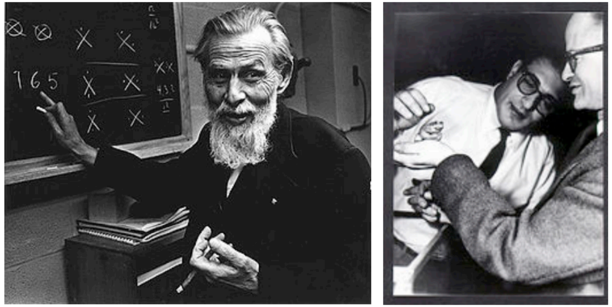
B. Olshausen

Historical review

- “Cybernetics”



Norbert Wiener



Warren McCulloch & Walter Pitts

Frank Rosenblatt

“The theory reported here clearly demonstrates the feasibility and fruitfulness of a quantitative statistical approach to the organization of cognitive systems. By the study of systems such as the perceptron, it is hoped that those fundamental laws of organization which are common to all information handling systems, machines and men included, may eventually be understood.” -- Frank Rosenblatt

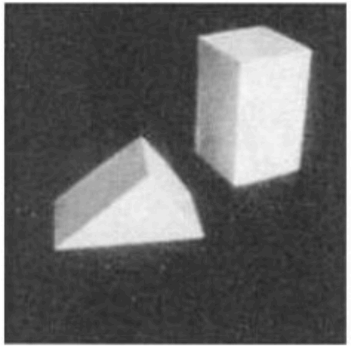
The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. In, *Psychological Review*, Vol. 65, No. 6, pp. 386-408, November, 1958.

B. Olshausen

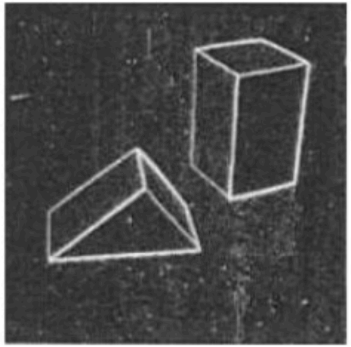
Historical review (computer vision)

- Larry Roberts Thesis 1963.
  - “Machine Perception of Three-Dimensional Solids”

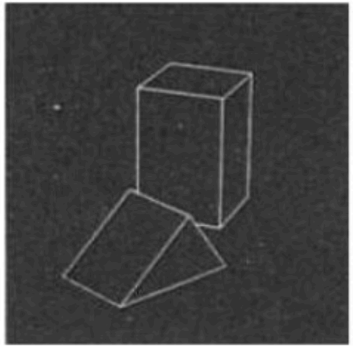
Larry Roberts
“Father of Computer Vision”



Input image



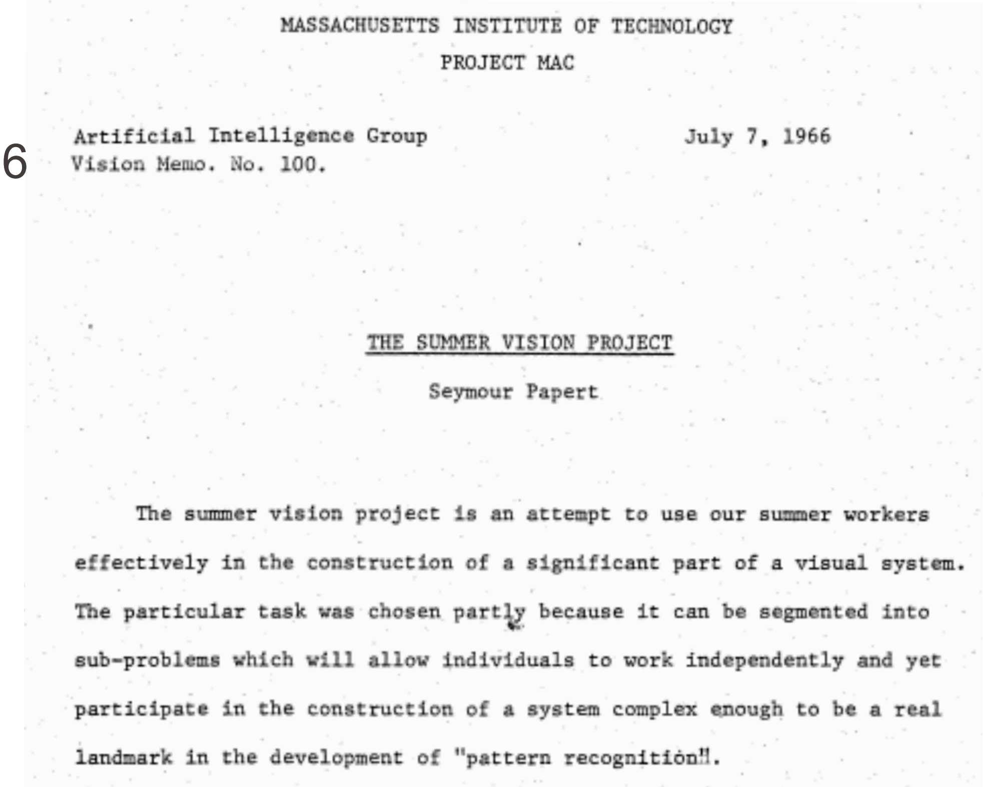
2x2 gradient operator



computed 3D model
rendered from new viewpoint

Historical review (computer vision)

- Larry Roberts Thesis 1963.
- The Summer Vision Project 1966



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".

Historical review (computer vision)

- **1960s**: Birth
  - Larry Roberts Thesis 1963.
  - The Summer Vision Project 1966
- **1970s**: Foundational work on image formation
- **1980s**: Applied mathematics: geometry, multi-scale analyses
- **1990s**: Geometric analysis. Vision+graphics. Resurfacing of statistical learning.
- **2000s**: Progress in visual recognition. Pre-deep learning. DPM. PASCAL.
- **2010s**: Deep Learning.

J. Malik

Historical review
AI

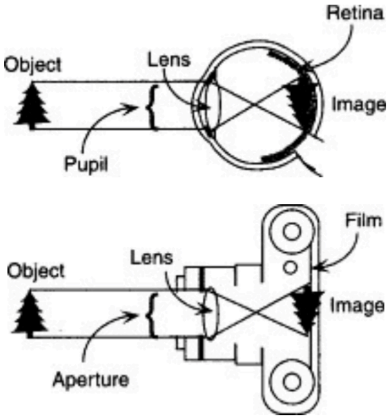
A good recap:

https://youtu.be/R3YFxF0n8n8

Fast Computer Vision Recap

Vision

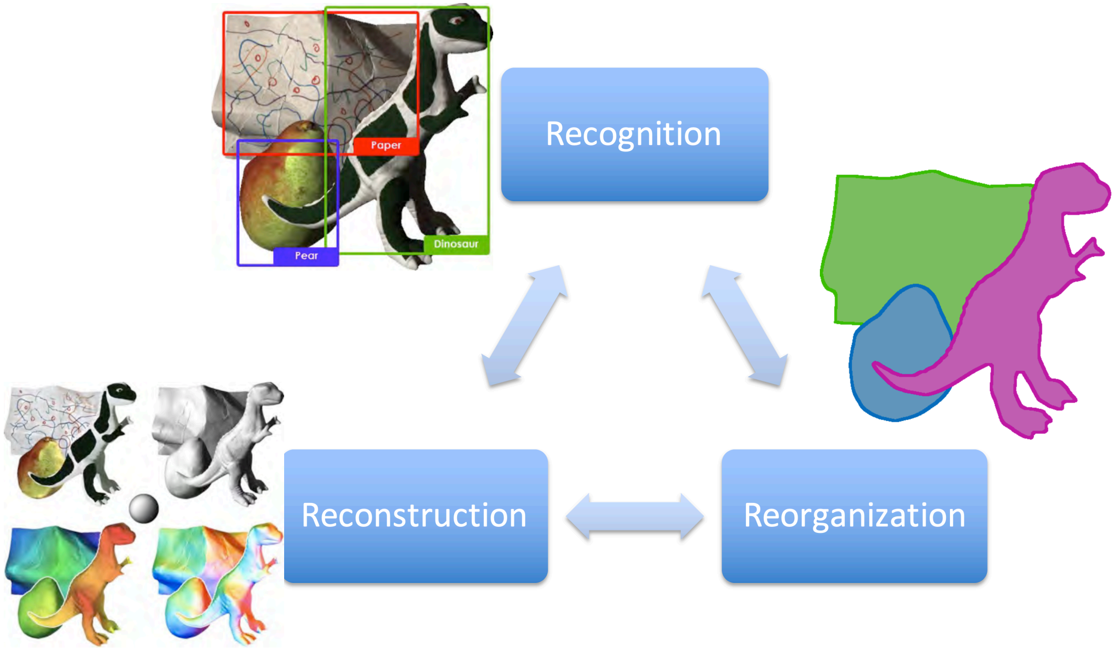
- defined as the process of acquiring knowledge about environmental objects and events by extracting information from the light they emit or reflect. [S. Palmer]
  - optical process: image formation
  - cognitive process: phenomena of visual perception
  - physiology of the visual nervous system: biological implementation



S. Palmer

One useful organization

- Three Rs of Computer Vision.
- Interaction between the Rs.



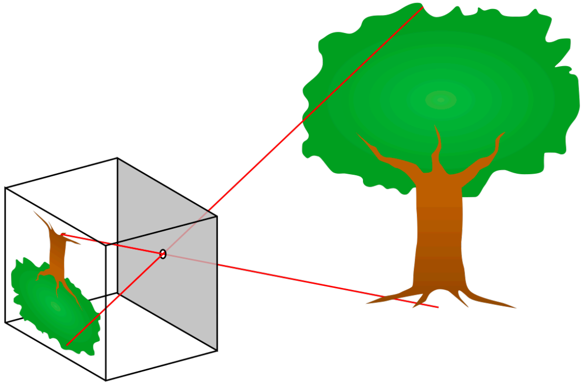
Recognition

Reconstruction

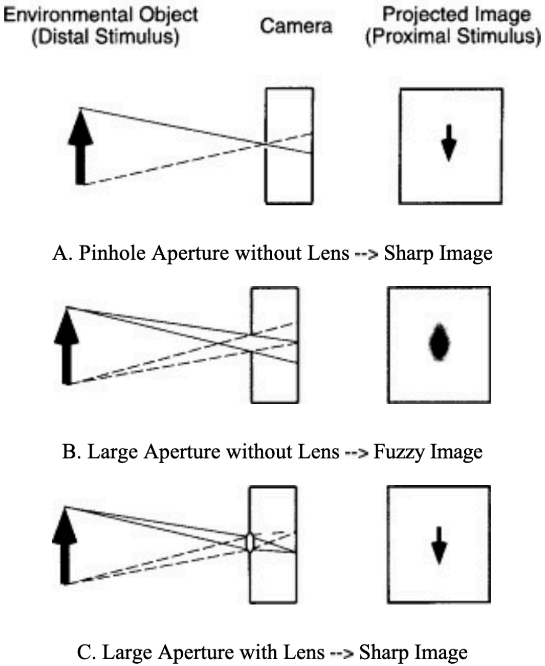
Reorganization

The three R's of computer vision: Recognition, reconstruction and reorganization. J. Malik et al. 2016.

Rudimentary Image Formation



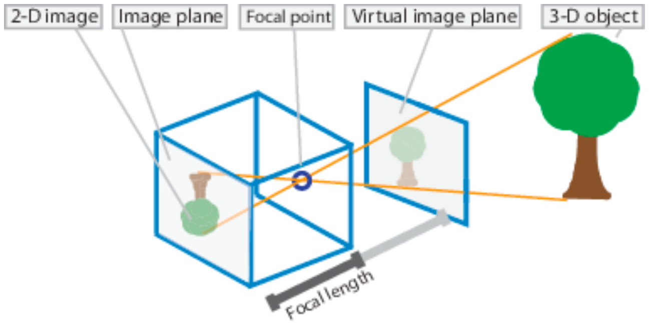
Rudimentary Image Formation



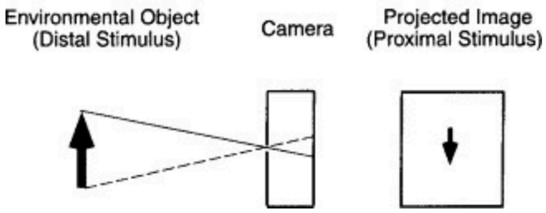
A. Pinhole Aperture without Lens --> Sharp Image

B. Large Aperture without Lens --> Fuzzy Image

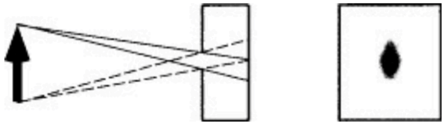
C. Large Aperture with Lens --> Sharp Image



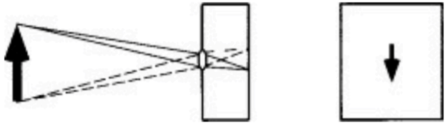
Rudimentary Image Formation



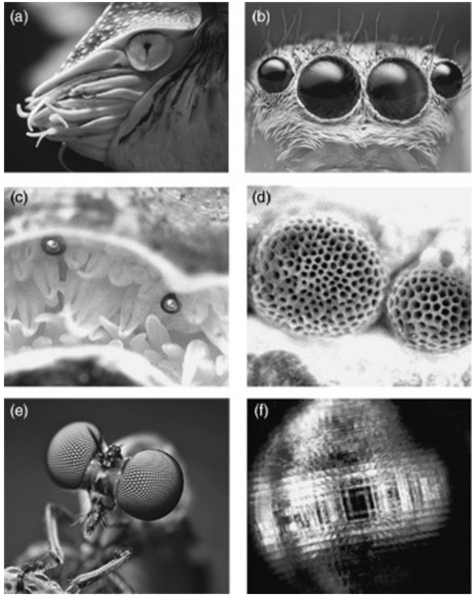
A. Pinhole Aperture without Lens --> Sharp Image



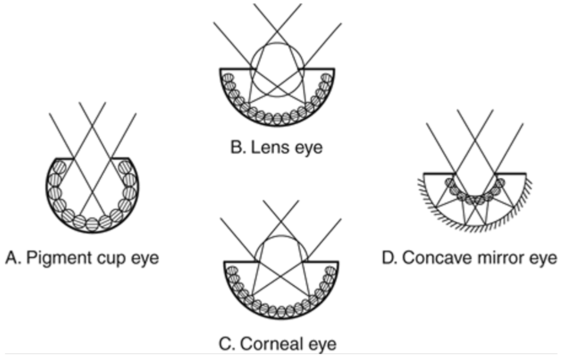
B. Large Aperture without Lens --> Fuzzy Image



C. Large Aperture with Lens --> Sharp Image



(a) The large pinhole eye of the cephalopod mollusc Nautilus. (b) Corneal eyes of a jumping spider Platycryptus. (c) Concave mirror eyes of the scallop Pecten. (d) Primitive compound eye of the ark clam Barbatia. (e) Compound eye of a male robberfly Holocephala. (f) Mirror compound eye of the shrimp Palaemonetes.



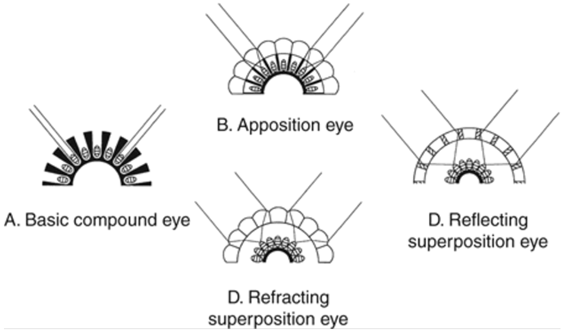

A. Basic compound eye
B. Apposition eye
D. Refracting superposition eye
D. Reflecting superposition eye

Image Transformation

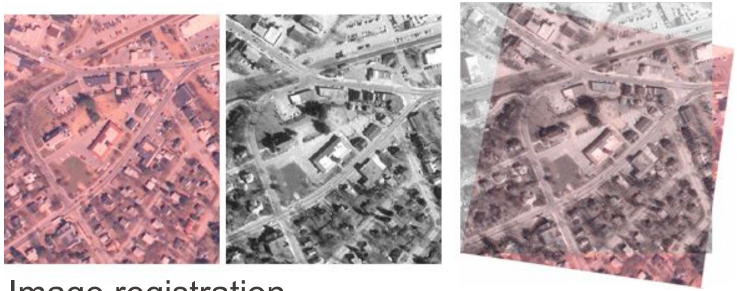


Image registration.

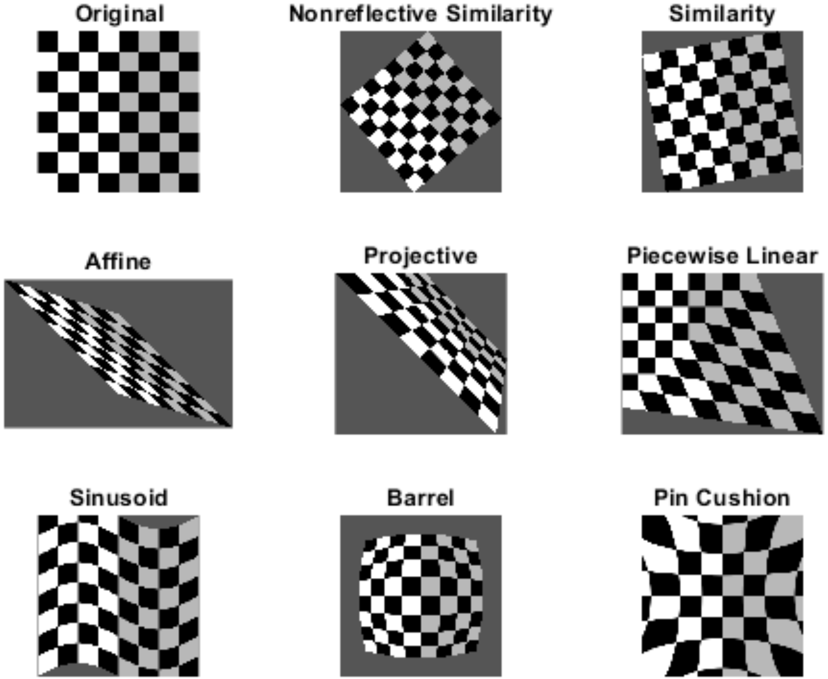
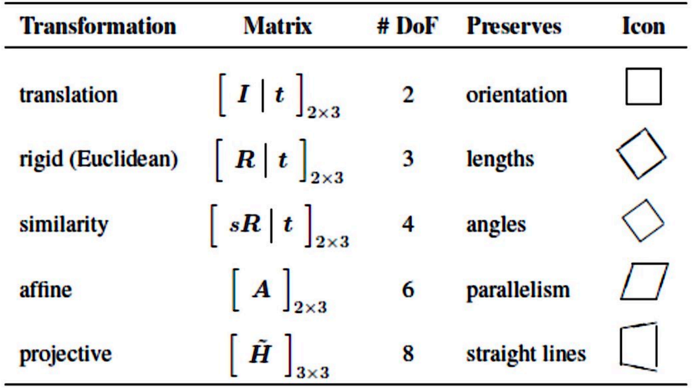
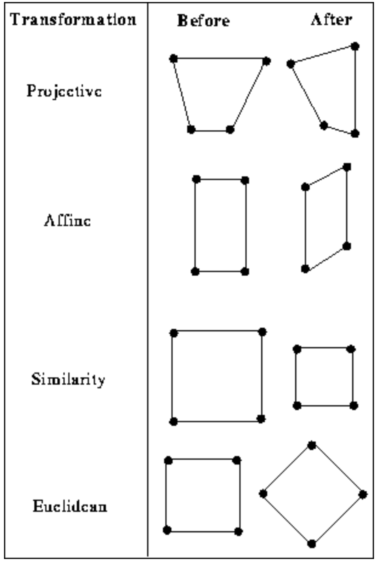


Image Transformation



Transformation	Matrix	# DoF	Preserves	Icon
translation	$\left[\, \boldsymbol{I} \mid \boldsymbol{t} \,\right]_{2\times 3}$	2	orientation	□
rigid (Euclidean)	$\left[\, \boldsymbol{R} \mid \boldsymbol{t} \,\right]_{2\times 3}$	3	lengths	◇
similarity	$\left[\, s\boldsymbol{R} \mid \boldsymbol{t} \,\right]_{2\times 3}$	4	angles	◇
affine	$\left[\, \boldsymbol{A} \,\right]_{2\times 3}$	6	parallelism	▱
projective	$\left[\, \tilde{\boldsymbol{H}} \,\right]_{3\times 3}$	8	straight lines	⏢



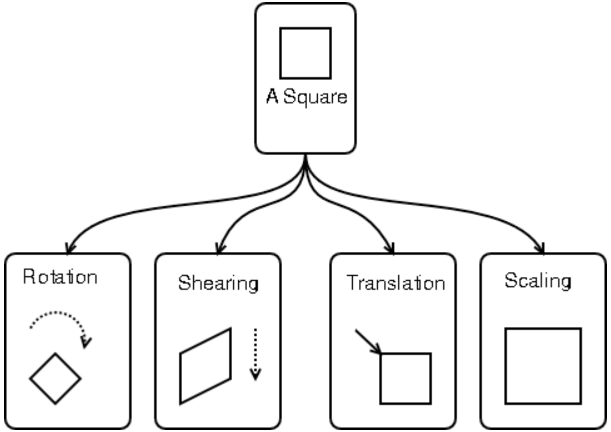
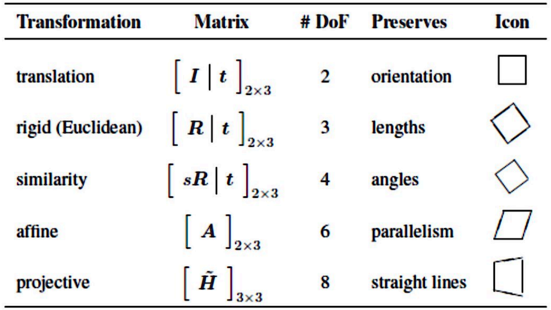
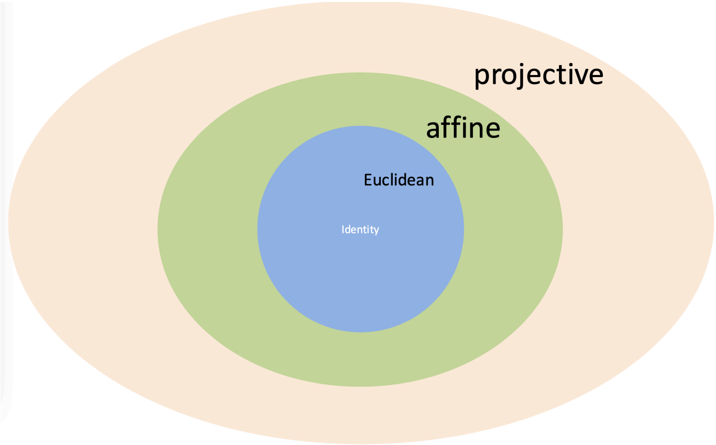
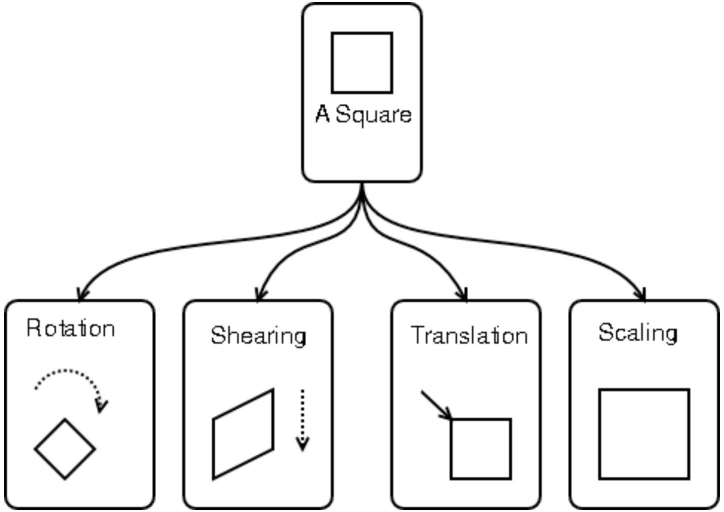


Image Transformation

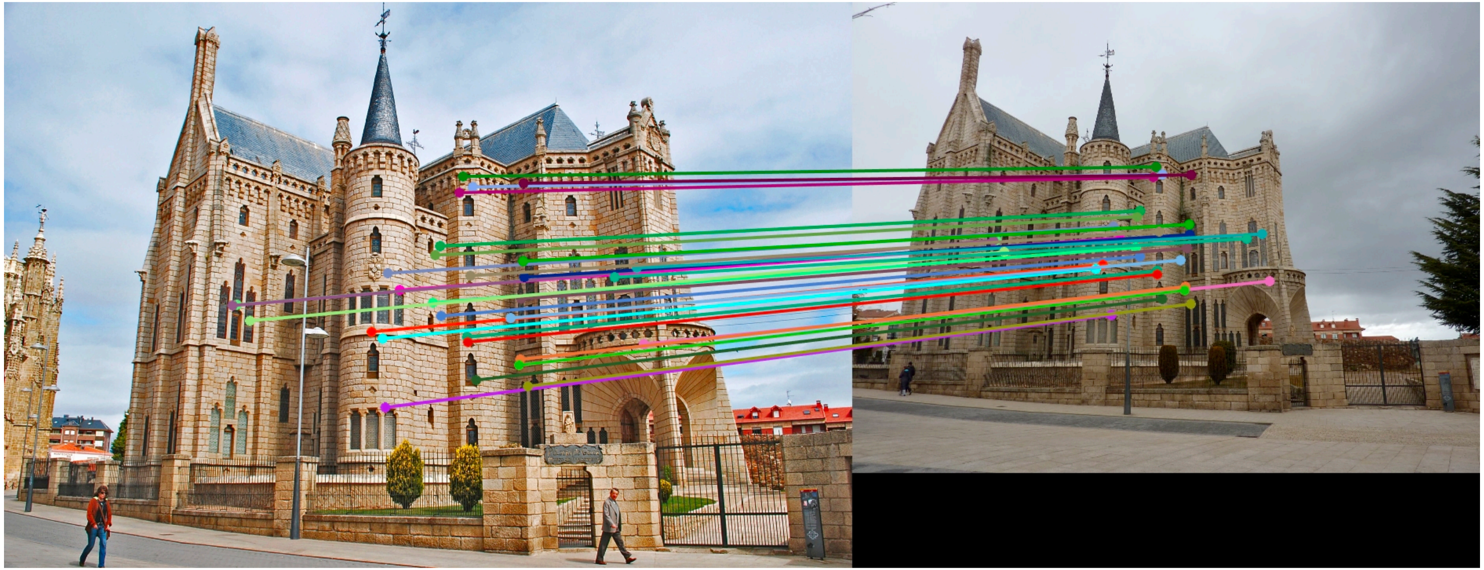


Transformation	Matrix	# DoF	Preserves	Icon
translation	$\left[\, \boldsymbol{I} \mid \boldsymbol{t} \,\right]_{2\times 3}$	2	orientation	□
rigid (Euclidean)	$\left[\, \boldsymbol{R} \mid \boldsymbol{t} \,\right]_{2\times 3}$	3	lengths	◇
similarity	$\left[\, s\boldsymbol{R} \mid \boldsymbol{t} \,\right]_{2\times 3}$	4	angles	◇
affine	$\left[\, \boldsymbol{A} \,\right]_{2\times 3}$	6	parallelism	▱
projective	$\left[\, \tilde{\boldsymbol{H}} \,\right]_{3\times 3}$	8	straight lines	⏢





Correspondences



Correspondences

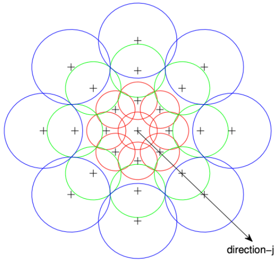
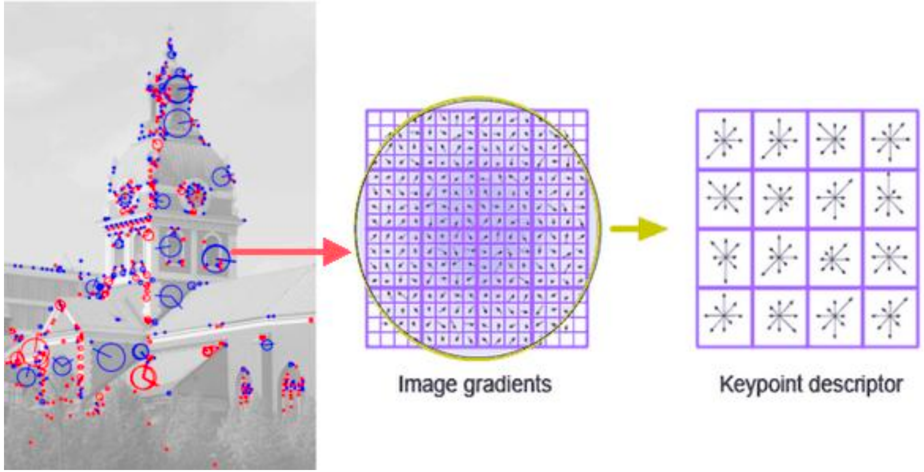


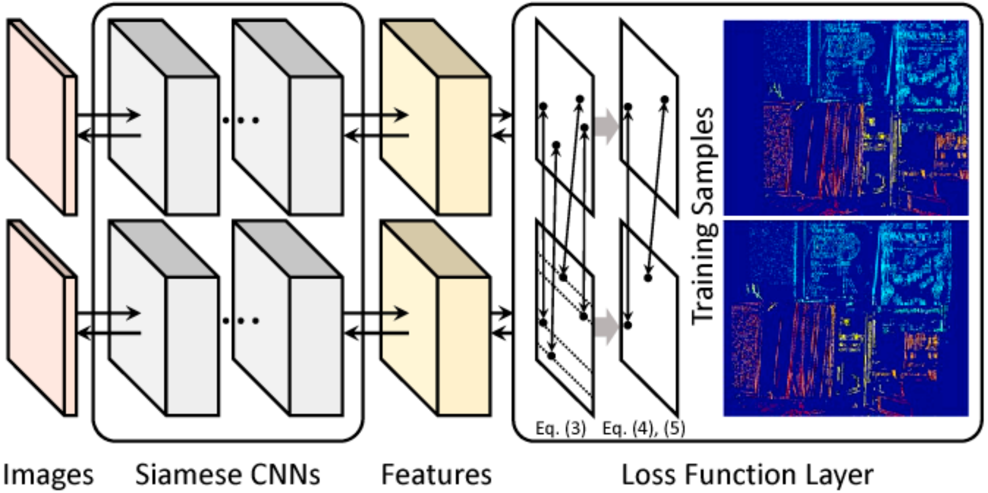
Figure 2. The DAISY descriptor. Each circle represents a region where the radius is proportional to the standard deviations of the Gaussian kernels and the '+' sign represents the locations where we sample the convolved orientation maps center being a pixel location where we compute the descriptor. By overlapping the regions we achieve smooth transitions between the regions and a degree of rotational robustness. The radii of the outer regions are increased to have an equal sampling of the rotational axis which is necessary for robustness against rotation.



SIFT, Lowe et al. 2003
DAISY, Tola et al. 2008

Correspondences

- Now gone deep.



e.g. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks (2015).

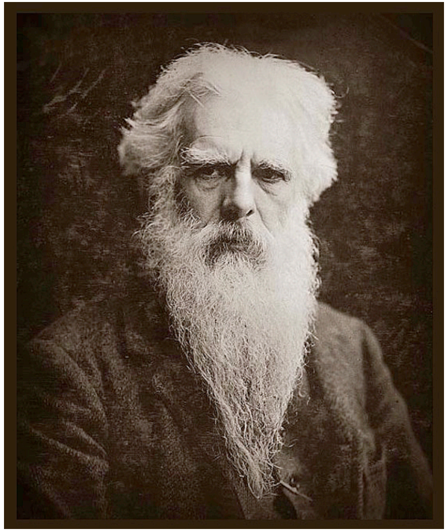
Dynamic Perspective

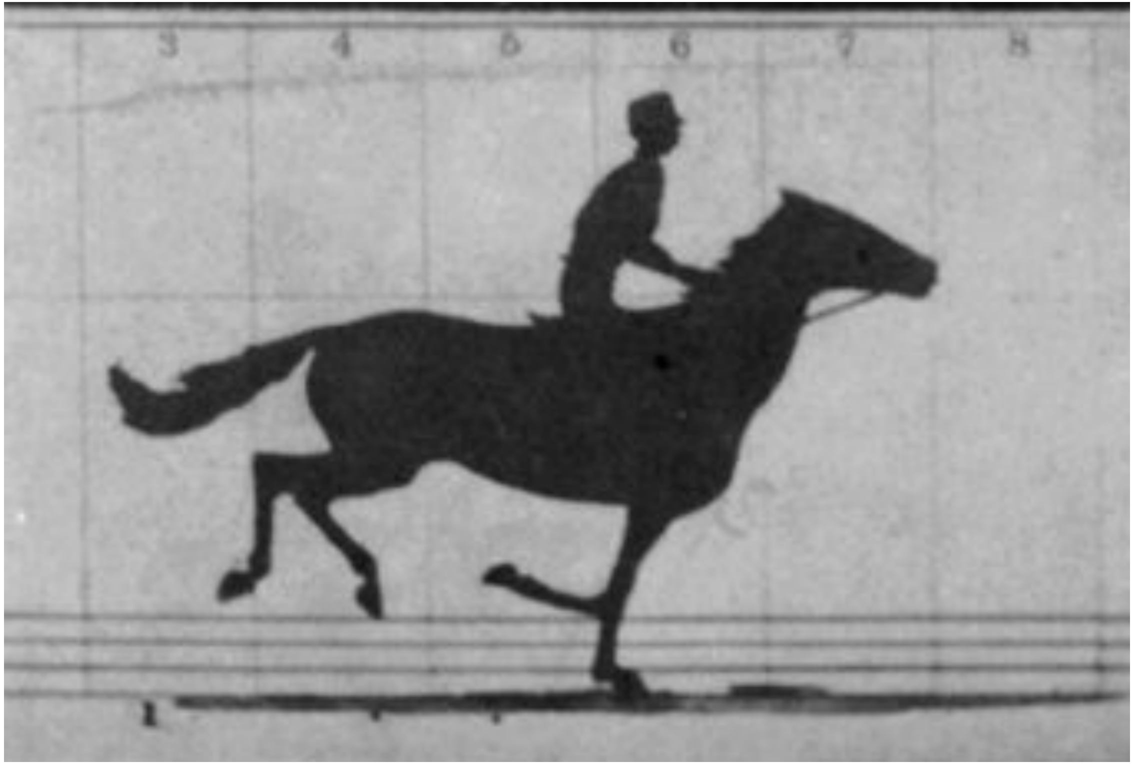
- Motion. Video.

Dynamic Perspective

- Motion. Video.

Eadweard Muybridge

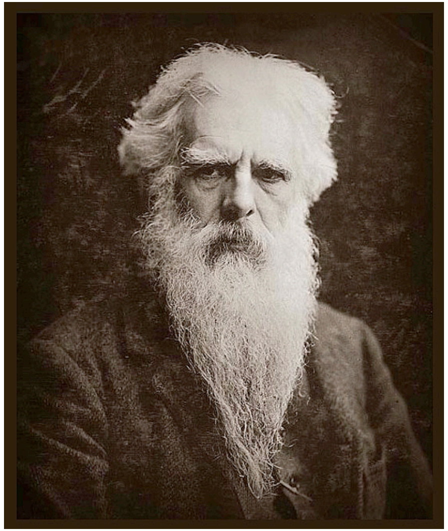


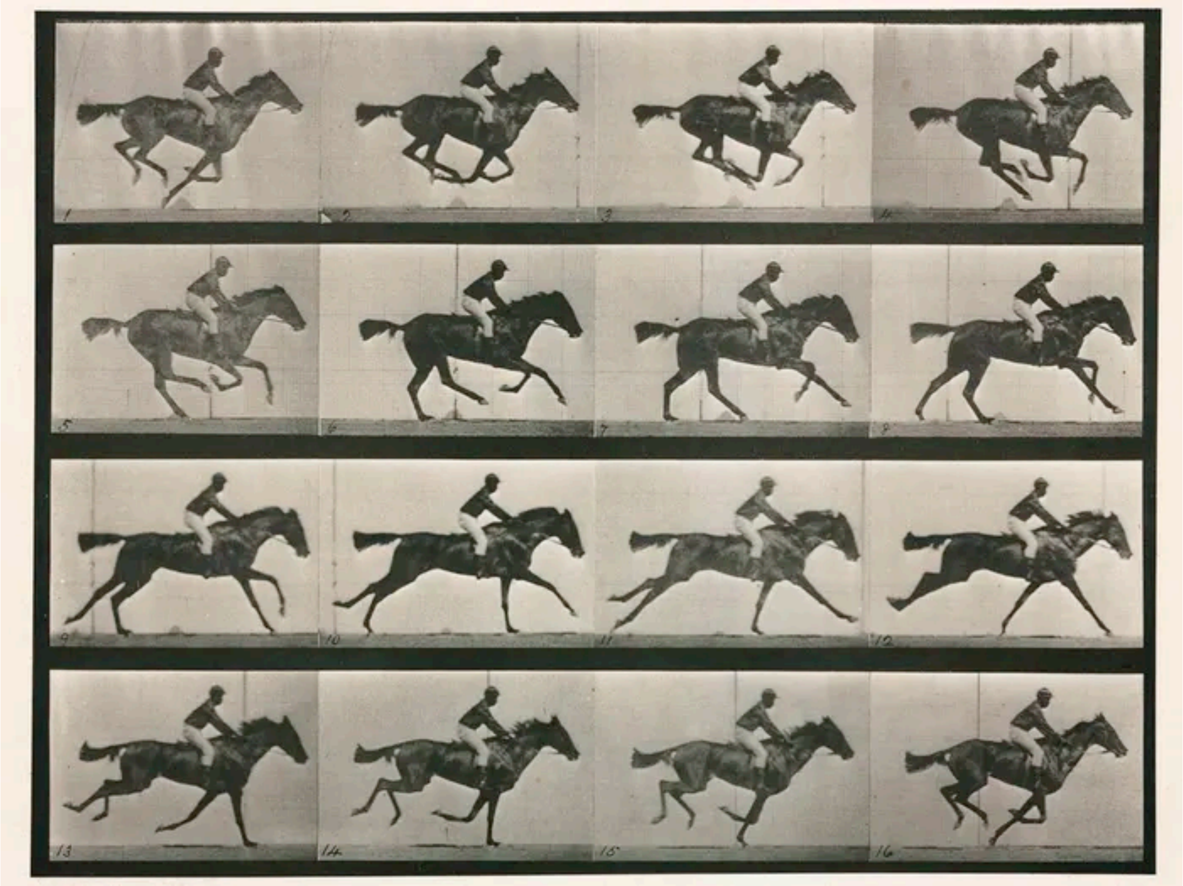


Dynamic Perspective

- Motion. Video.

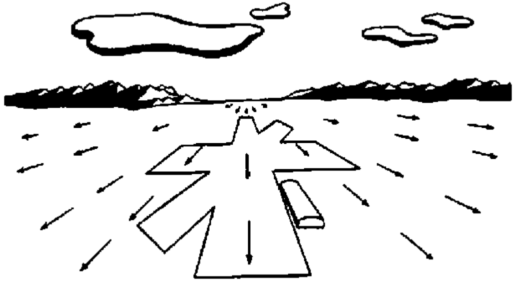
Eadweard Muybridge



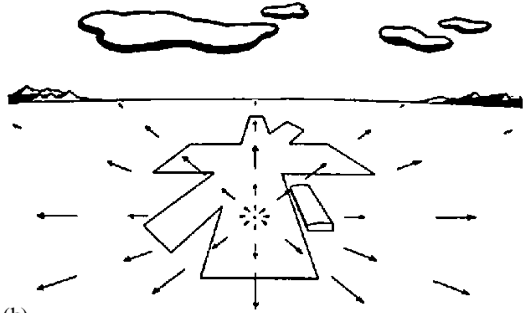


Dynamic Perspective

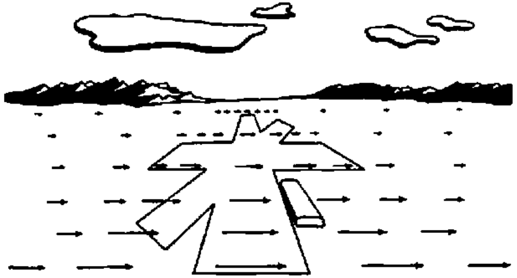
- Optical Flow.
- J J Gibson's examples:



(a)

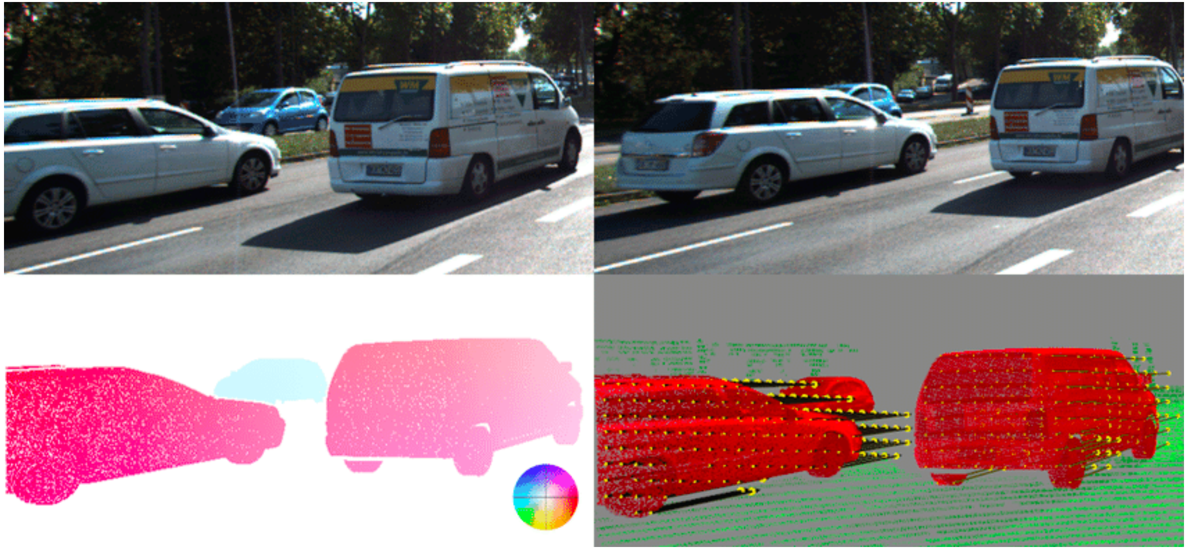


(b)



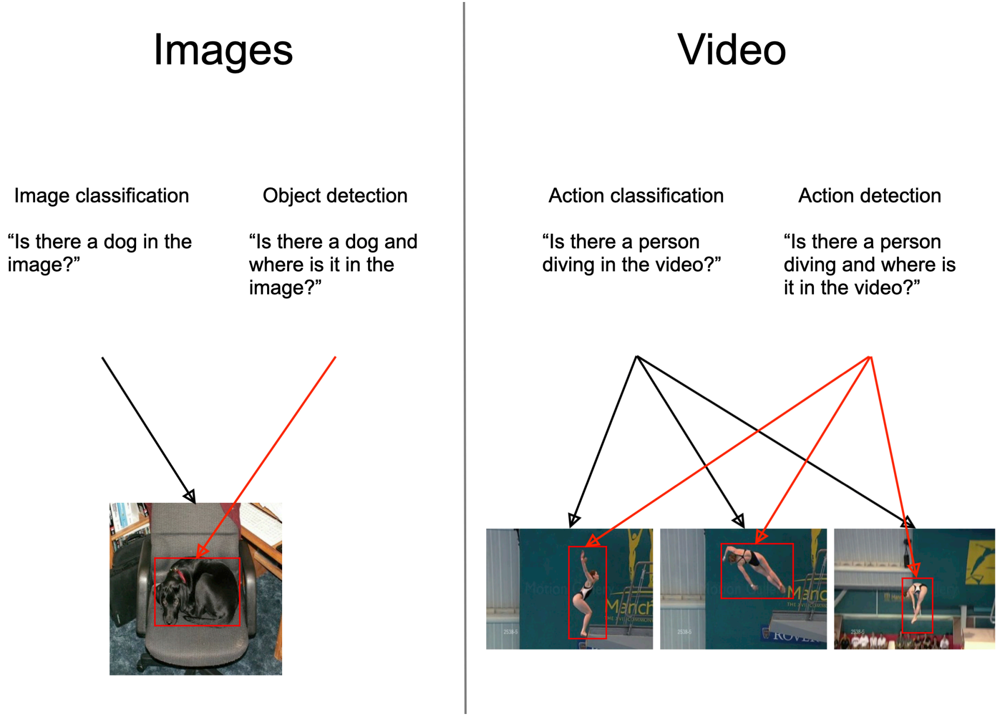
Dynamic Perspective

- Optical Flow



Dynamic Perspective

- Recognition in Videos
  - E.g. actions



Images

Image classification

“Is there a dog in the image?”

Object detection

“Is there a dog and where is it in the image?”

Video

Action classification

“Is there a person diving in the video?”

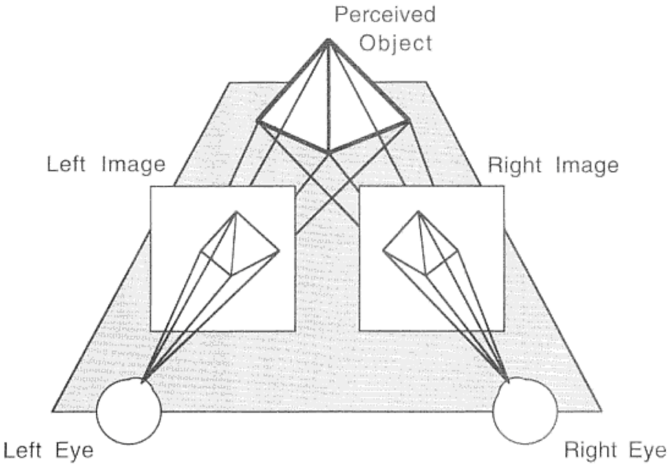
Action detection

“Is there a person diving and where is it in the video?”

Dynamic Perspective→3D

- Moving (or multi) Camera

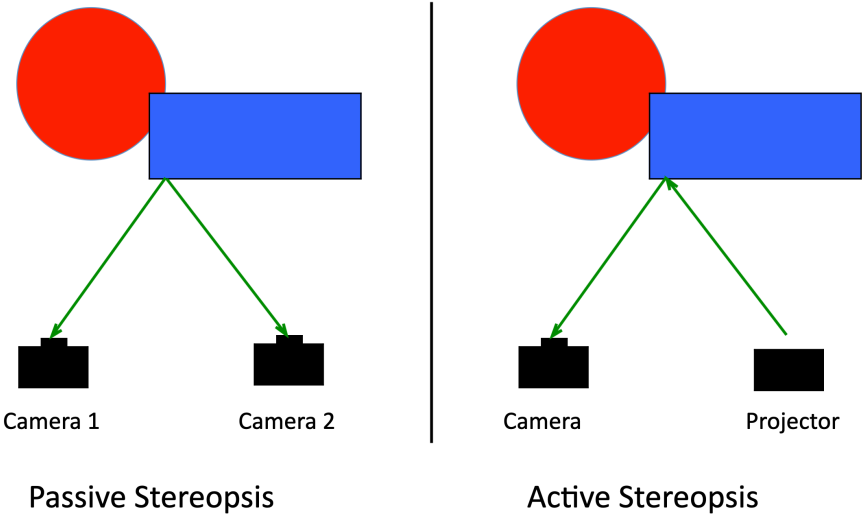
Binocular Stereopsis



3D

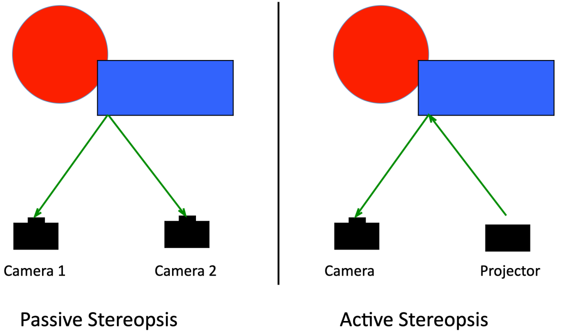
- Moving (or multi) Camera

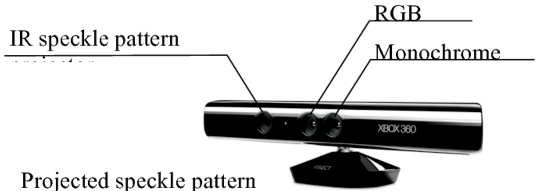
Depth from Triangulation



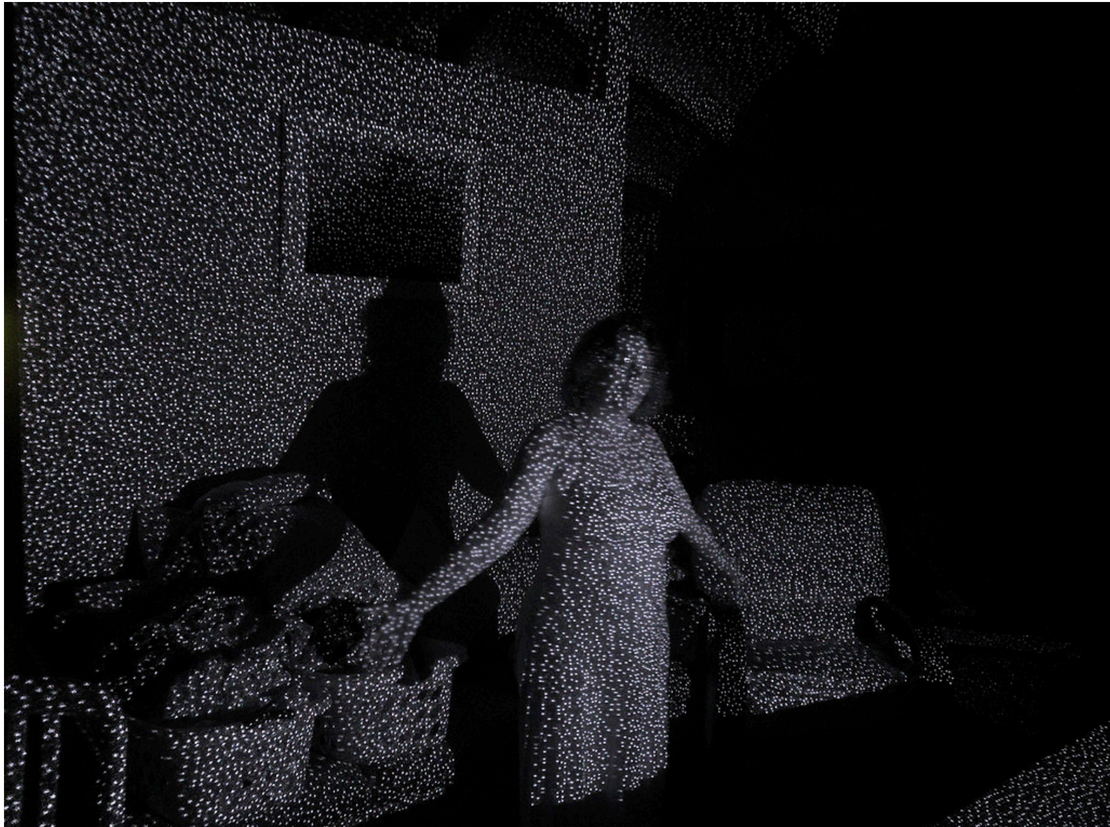
3D

Depth from Triangulation



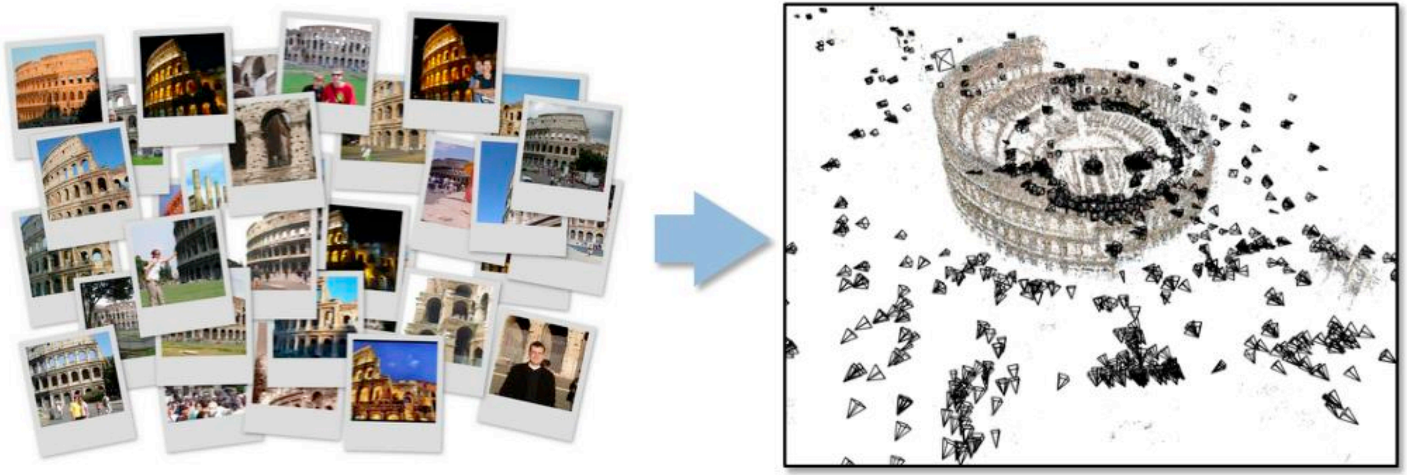


Projected speckle pattern



3D

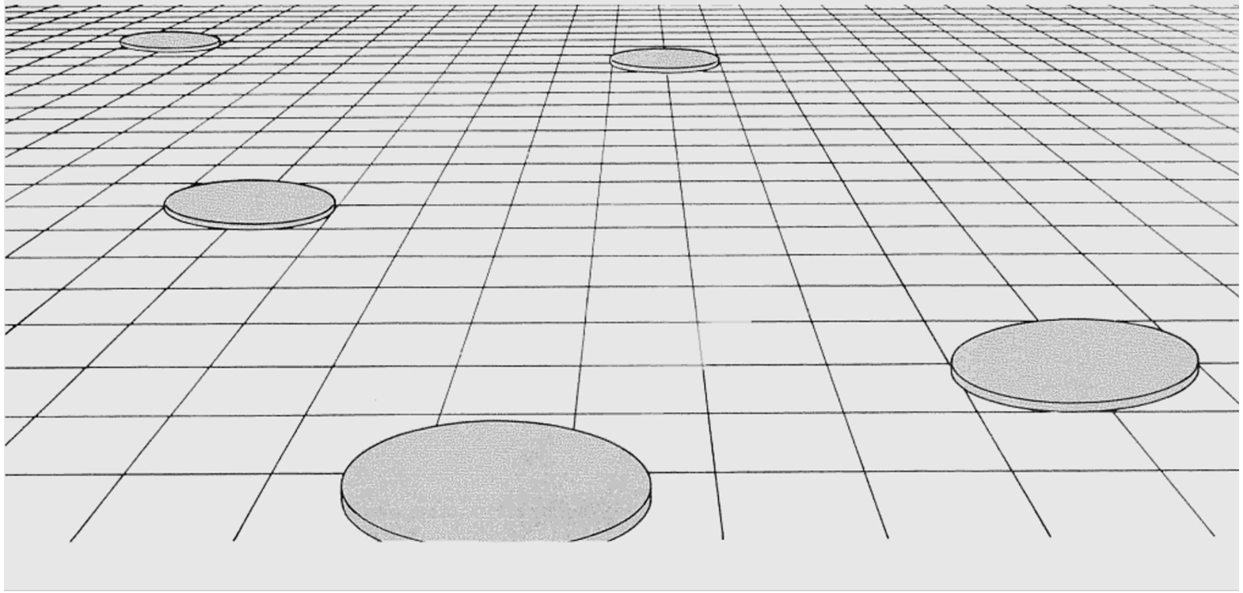
- Moving (or multi) Camera



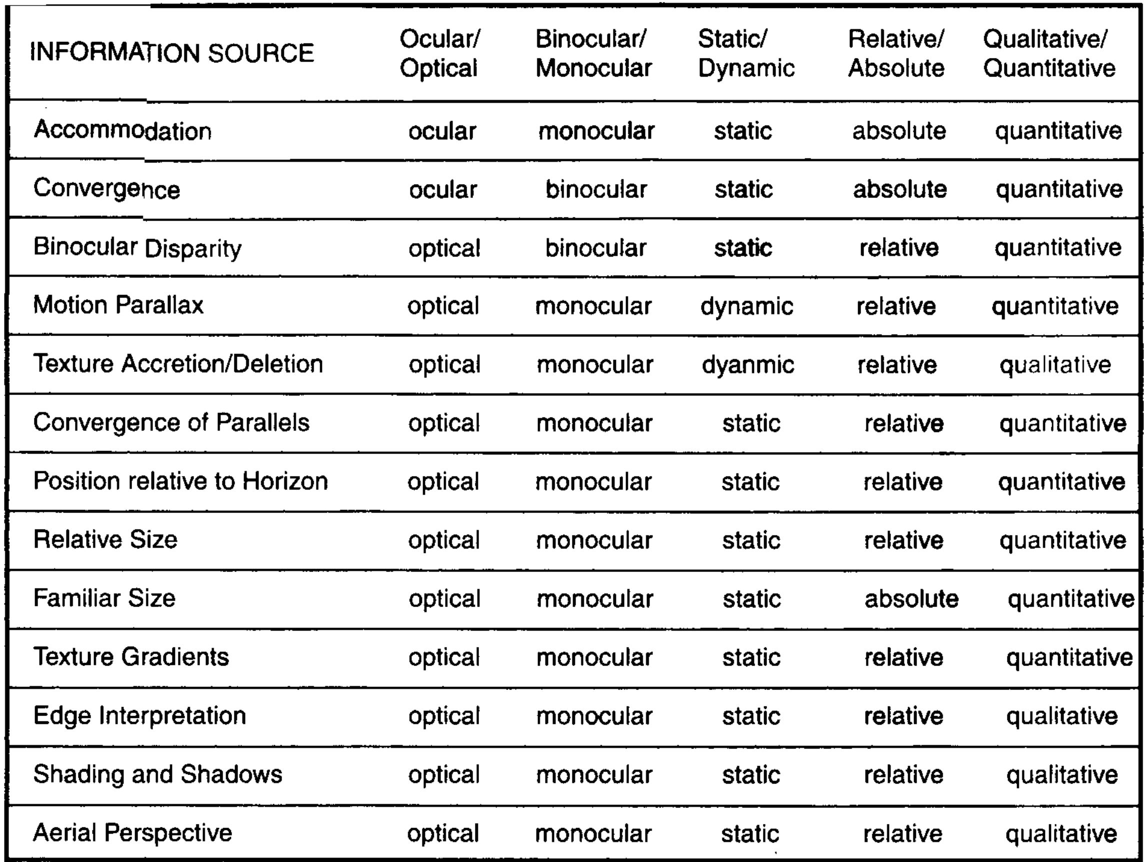
Building Rome in a Day. Snavely et al.

Other 3D estimation cues

- e.g. Pictorial cues



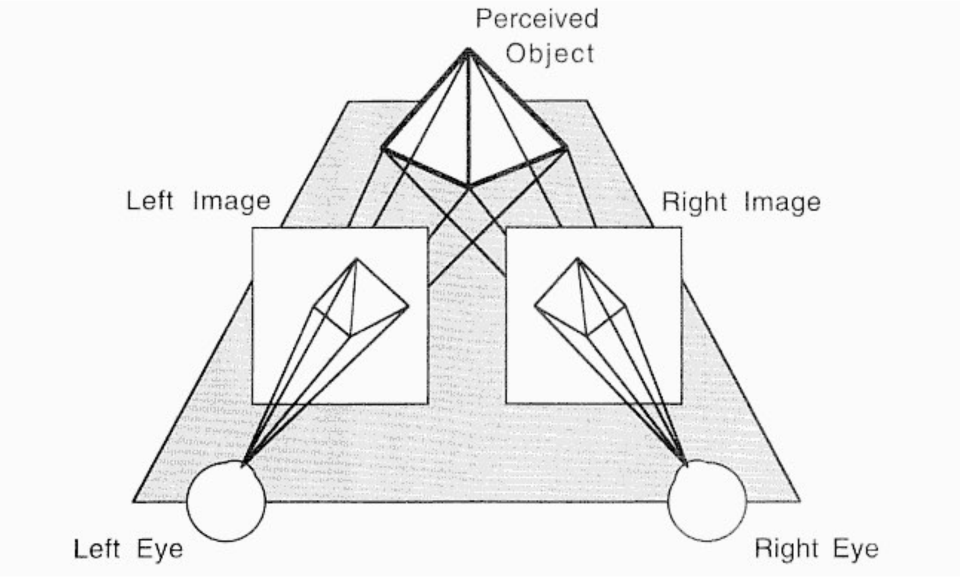
List of 3D cues
(Vision Science, Palmer)



INFORMATION SOURCE	Ocular/ Optical	Binocular/ Monocular	Static/ Dynamic	Relative/ Absolute	Qualitative/ Quantitative
Accommodation	ocular	monocular	static	absolute	quantitative
Convergence	ocular	binocular	static	absolute	quantitative
Binocular Disparity	optical	binocular	static	relative	quantitative
Motion Parallax	optical	monocular	dynamic	relative	quantitative
Texture Accretion/Deletion	optical	monocular	dyanmic	relative	qualitative
Convergence of Parallels	optical	monocular	static	relative	quantitative
Position relative to Horizon	optical	monocular	static	relative	quantitative
Relative Size	optical	monocular	static	relative	quantitative
Familiar Size	optical	monocular	static	absolute	quantitative
Texture Gradients	optical	monocular	static	relative	quantitative
Edge Interpretation	optical	monocular	static	relative	qualitative
Shading and Shadows	optical	monocular	static	relative	qualitative
Aerial Perspective	optical	monocular	static	relative	qualitative

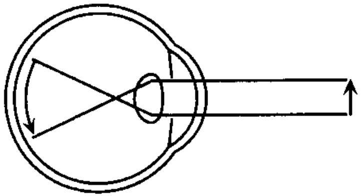
J. Malik

Binocular Stereopsis

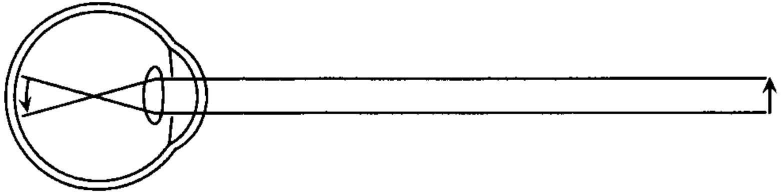


J. Malik

Accommodation



Thick Lens → Close

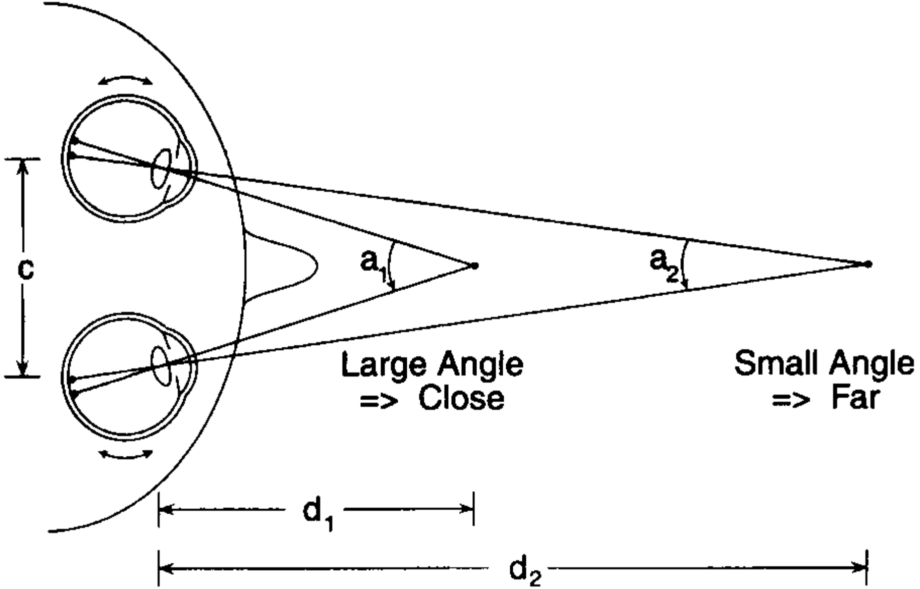


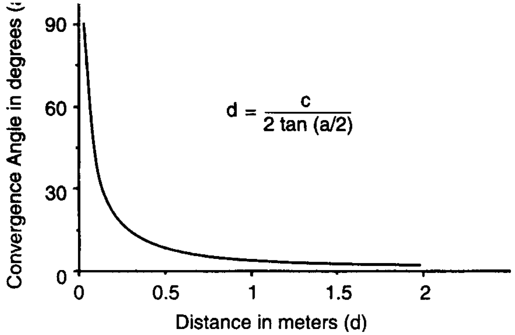
Thin Lens → Far

J. Malik

Convergence

Zamir

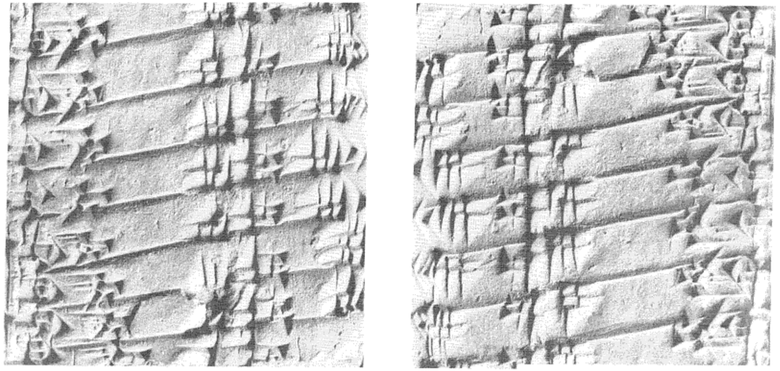




$$d = \frac{c}{2 \tan(a/2)}$$

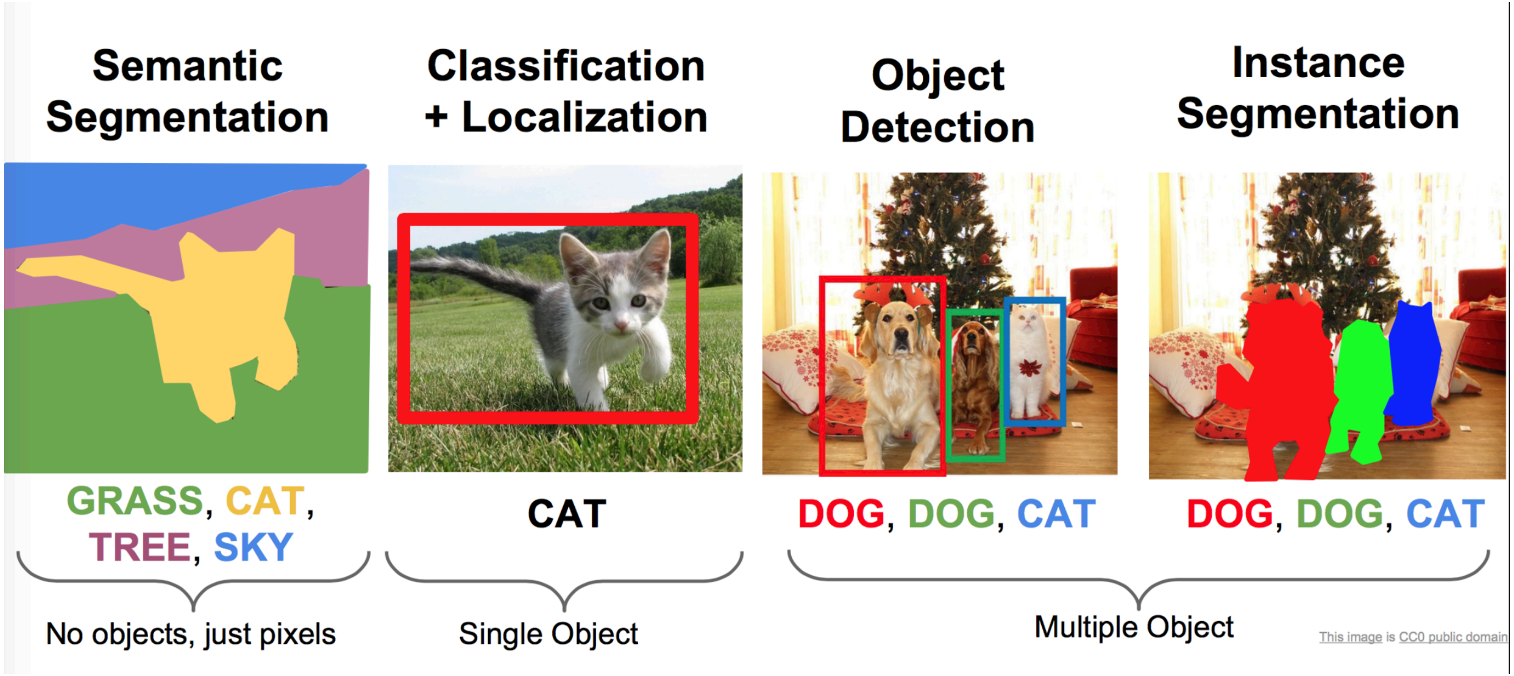
J. Malik

Shading



J. Malik

Recognition



Semantic Segmentation

GRASS, CAT, TREE, SKY

No objects, just pixels

Classification + Localization

CAT

Single Object

Object Detection

DOG, DOG, CAT

Instance Segmentation

DOG, DOG, CAT

Multiple Object

This image is CC0 public domain

Recognition

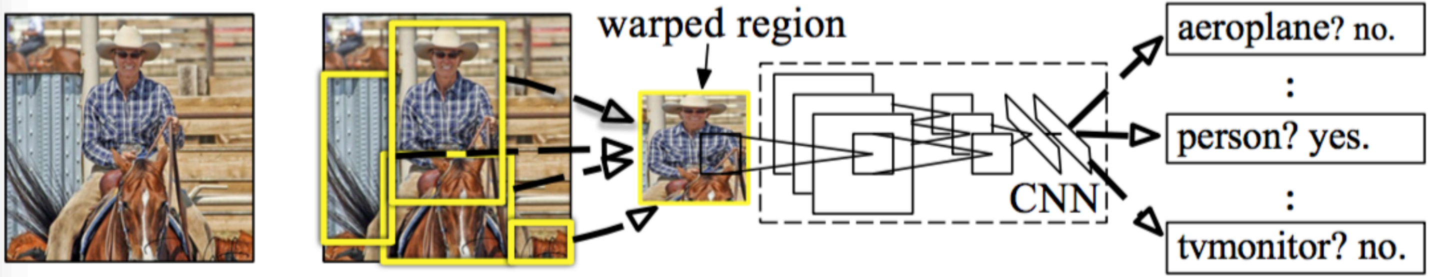


figure credit: R. Girshick et al.

input image

region proposals
~2,000

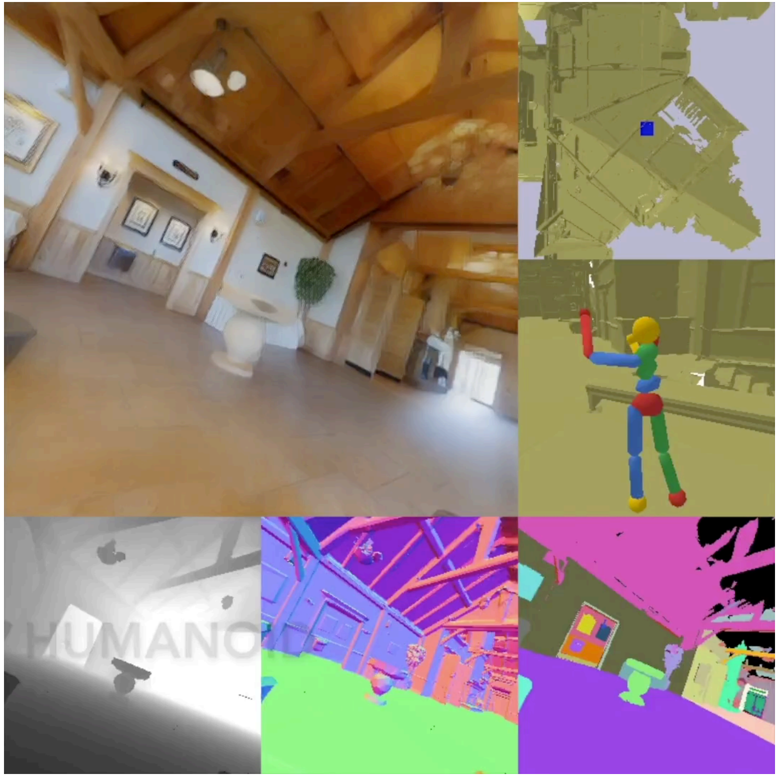
1 CNN for each region

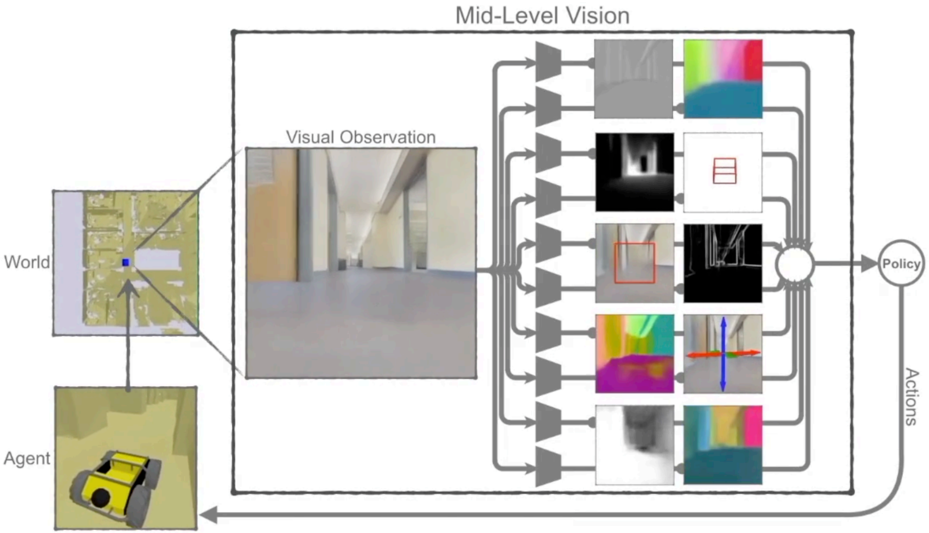
classify regions

R-CNN pipeline

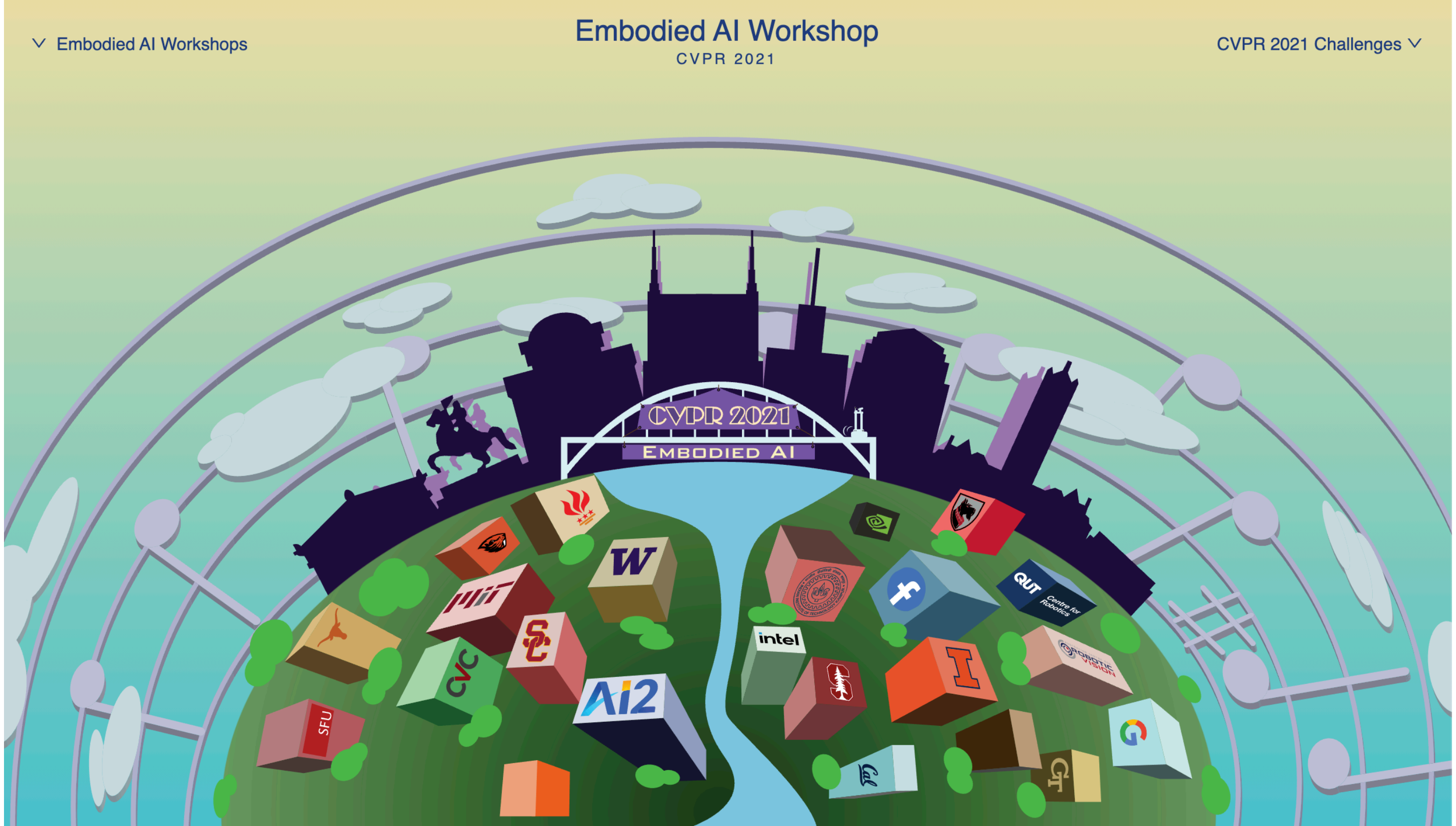
R. Girshick, J. Donahue, T. Darrell, & J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation". CVPR 2014

Embodied/Active Vision





Embodied/Active Vision



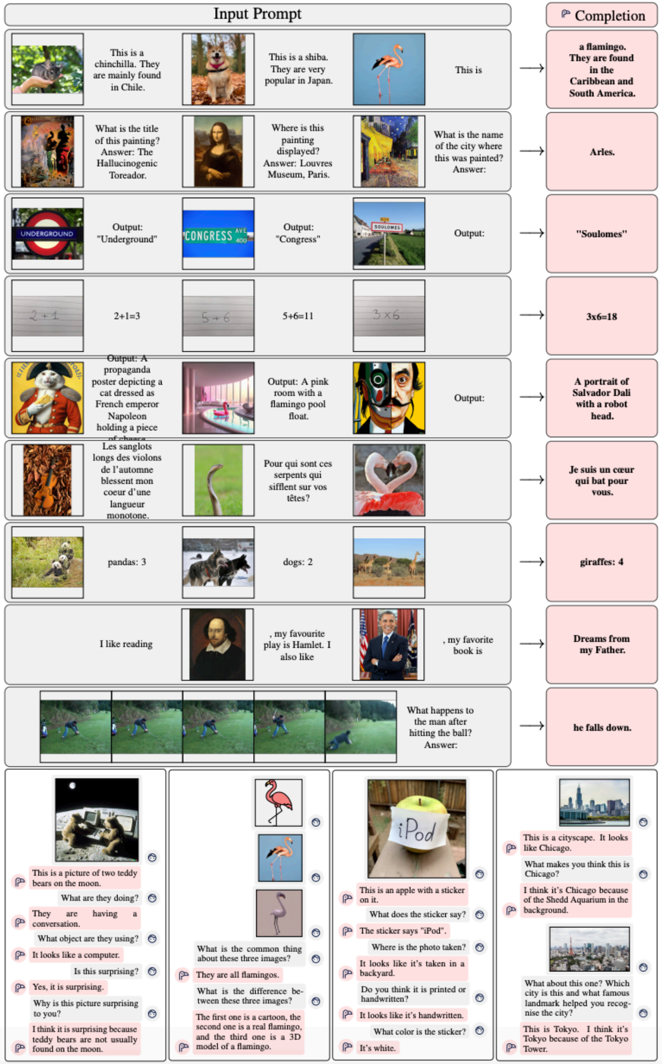
Embodied AI Workshops

Embodied AI Workshop

CVPR 2021

CVPR 2021 Challenges

"Foundation" models
Vision-Language Models
Multimodal Models



Input Prompt	Completion
This is a chinchilla. They are mainly found in Chile. This is a shiba. They are very popular in Japan. This is	a flamingo. They are found in the Caribbean and South America.
What is the title of this painting? Answer: The Hallucinogenic Toreador. Where is this painting displayed? Answer: Louvres Museum, Paris. What is the name of the city where this was painted? Answer:	Arles.
Output: "Underground" Output: "Congress" Output:	"Soulomes"
2+1=3 5+6=11	3x6=18
Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese. Output: A pink room with a flamingo pool float. Output:	A portrait of Salvador Dali with a robot head.
Les sanglots longs des violons de l'automne blessent mon coeur d'une langueur monotone. Pour qui sont ces serpents qui sifflent sur vos têtes?	Je suis un cœur qui bat pour vous.
pandas: 3 dogs: 2	giraffes: 4
I like reading , my favourite play is Hamlet. I also like , my favorite book is	Dreams from my Father.
What happens to the man after hitting the ball? Answer:	he falls down.

This is a picture of two teddy bears on the moon.

What are they doing?

They are having a conversation.

What object are they using?

It looks like a computer.

Is this surprising?

Yes, it is surprising.

Why is this picture surprising to you?

I think it is surprising because teddy bears are not usually found on the moon.

What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.

This is an apple with a sticker on it.

What does the sticker say?

The sticker says "iPod".

Where is the photo taken?

It looks like it's taken in a backyard.

Do you think it is printed or handwritten?

It looks like it's handwritten.

What color is the sticker?

It's white.

This is a cityscape. It looks like Chicago.

What makes you think this is Chicago?

I think it's Chicago because of the Shedd Aquarium in the background.

What about this one? Which city is this and what famous landmark helped you recognise the city?

This is Tokyo. I think it's Tokyo because of the Tokyo Tower.

Flamingo,, Alayrac et al, 2022.
LLaVA, Liu et al, 2023
4M, Mizrahi & Bachmann et al. 2023.

“Foundation” models
Vision-Language Models
Multimodal Models

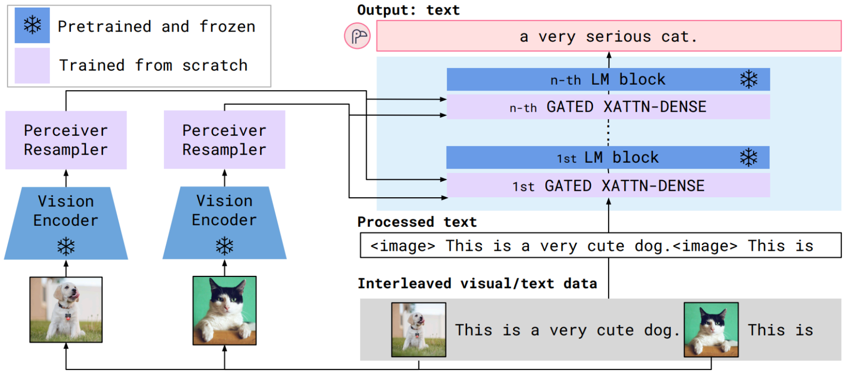
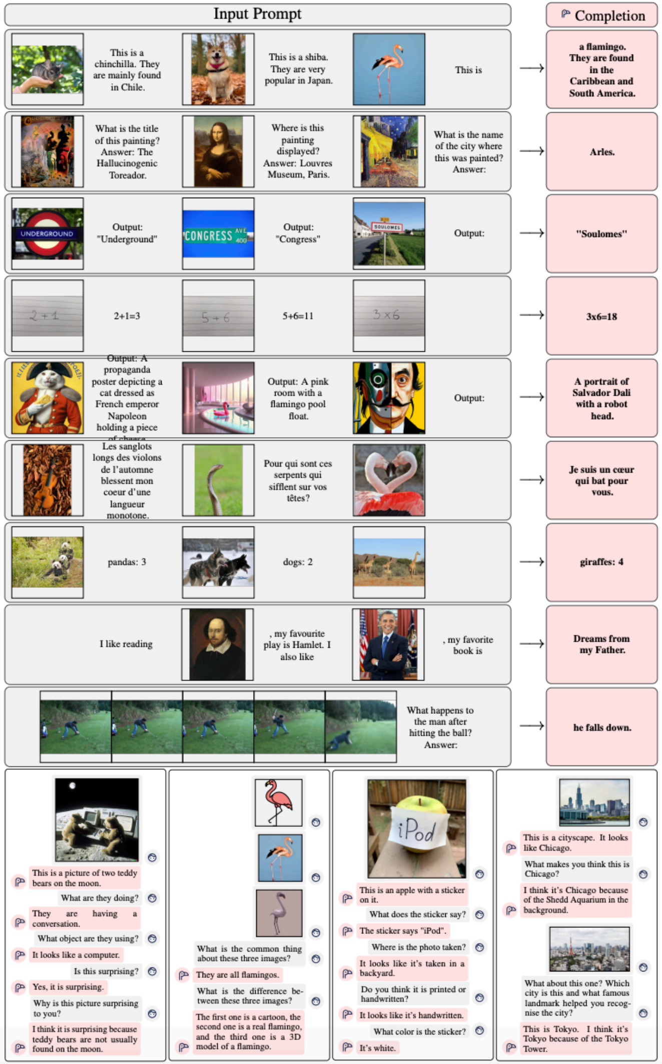


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

Flamingo,, Alayrac et al, 2022.
LLaVA, Liu et al, 2023
4M, Mizrahi & Bachmann et al. 2023.



Input Prompt			Completion
This is a chinchilla. They are mainly found in Chile.	This is a shiba. They are very popular in Japan.	This is	a flamingo. They are found in the Caribbean and South America.
What is the title of this painting? Answer: The Hallucinogenic Toreador.	Where is this painting displayed? Answer: Louvres Museum, Paris.	What is the name of the city where this was painted? Answer:	Arles.
Output: "Underground"	Output: "Congress"	Output:	"Soulomes"
2+1=3	5+6=11		3x6=18
Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.	Output: A pink room with a flamingo pool float.	Output:	A portrait of Salvador Dali with a robot head.
Les sanglots longs des violons de l'automne blessent mon coeur d'une langueur monotone.	Pour qui sont ces serpents qui sifflent sur vos têtes?		Je suis un cœur qui bat pour vous.
pandas: 3	dogs: 2		giraffes: 4
I like reading	, my favourite play is Hamlet. I also like	, my favorite book is	Dreams from my Father.
		What happens to the man after hitting the ball? Answer:	he falls down.

This is a picture of two teddy bears on the moon.
What are they doing?
They are having a conversation.
What object are they using?
It looks like a computer.
Is this surprising?
Yes, it is surprising.
Why is this picture surprising to you?
I think it is surprising because teddy bears are not usually found on the moon.

What is the common thing about these three images?
They are all flamingos.
What is the difference between these three images?
The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.

This is an apple with a sticker on it.
What does the sticker say?
The sticker says "iPod".
Where is the photo taken?
It looks like it’s taken in a backyard.
Do you think it is printed or handwritten?
It looks like it’s handwritten.
What color is the sticker?
It’s white.

This is a cityscape. It looks like Chicago.
What makes you think this is Chicago?
I think it’s Chicago because of the Shedd Aquarium in the background.
What about this one? Which city is this and what famous landmark helped you recognise the city?
This is Tokyo. I think it’s Tokyo because of the Tokyo Tower.

“Foundation” models
Vision-Language Models
Multimodal Models

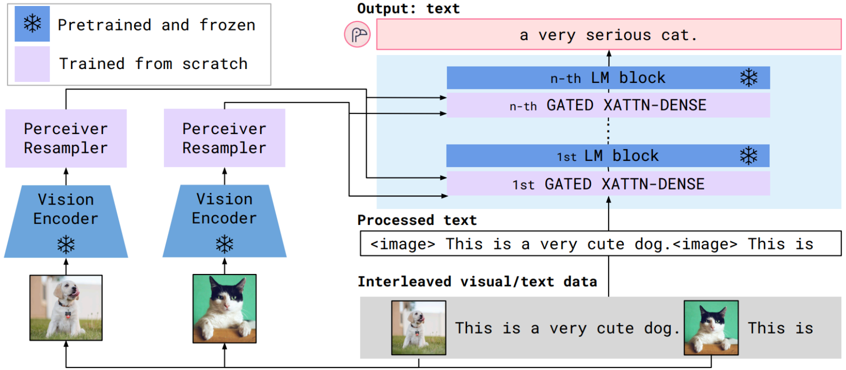
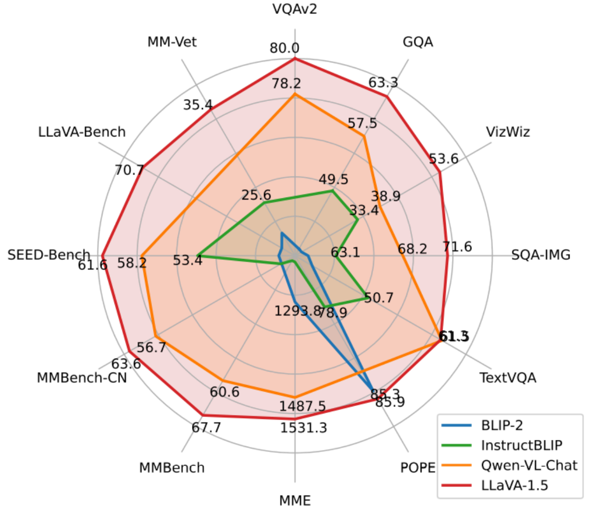
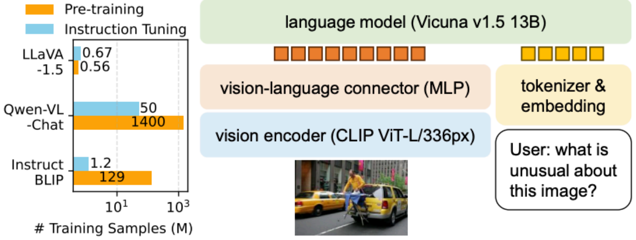


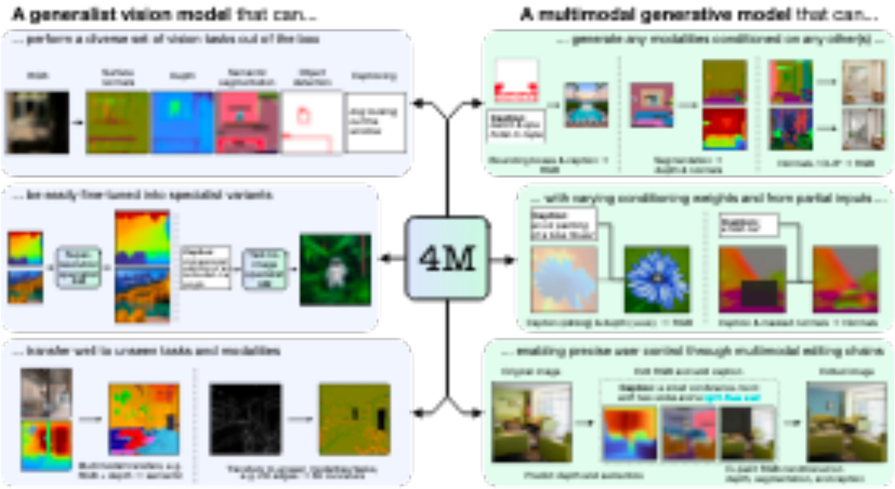
Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

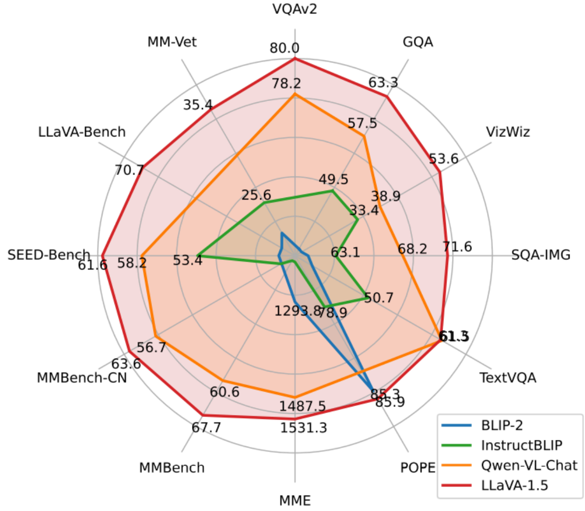


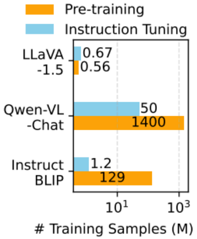


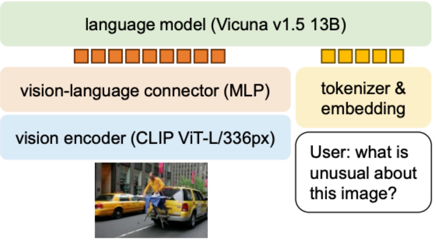
Flamingo,, Alayrac et al, 2022.
LLaVA, Liu et al, 2023
4M, Mizrahi & Bachmann et al. 2023.

“Foundation” models
Vision-Language Models
Multimodal Models


A generalist vision model that can...
... perform a diverse set of vision tasks out of the box
... be easily fine-tuned into specialist variants
... transfer well to unseen tasks and modalities
4M
A multimodal generative model that can...
... generate any modalities conditioned on any other(s) ...
... with varying conditioning weights and from partial inputs ...
... enabling precise user control through multimodal editing chains



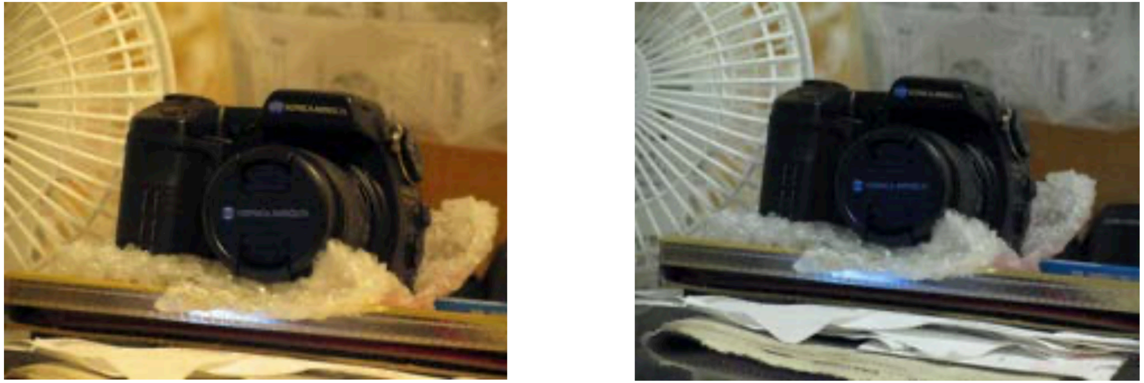



language model (Vicuna v1.5 13B)
vision-language connector (MLP)
tokenizer & embedding
vision encoder (CLIP ViT-L/336px)
User: what is unusual about this image?

Flamingo,, Alayrac et al, 2022.
LLaVA, Liu et al, 2023
4M, Mizrahi & Bachmann et al. 2023.

Image Processing Perspective

- e.g. White Balancing



- Manual
  - Choose color-neutral object in the photos and normalize
- Automatic (AWB)
  - Grey World: force average color of scene to grey
  - White World: force brightest object to white

Image Processing Perspective

- Point Processing

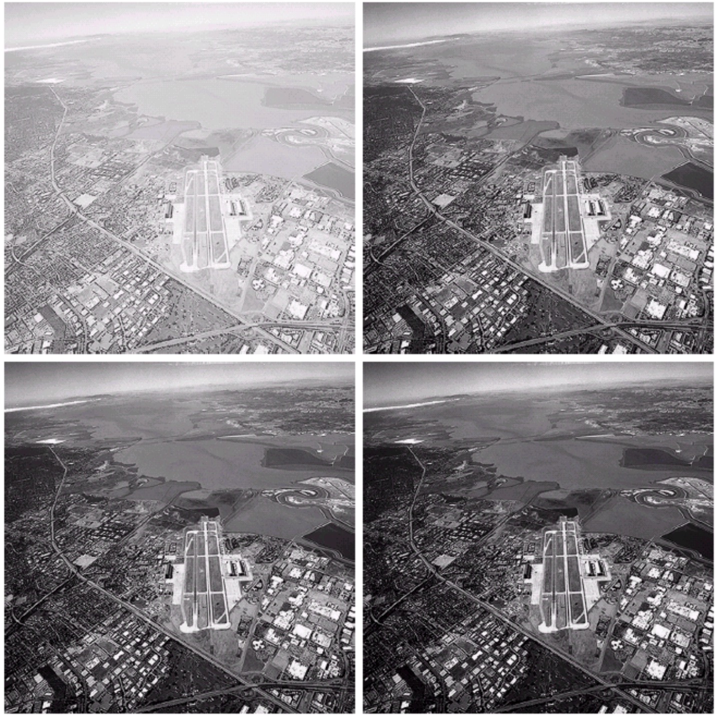
The simplest kind of range transformations are these independent of position x,y:

$$g = T(f)$$

This is called point processing.

e.g. Gain and Bias transform:

$$g(x,y) = a*f(x,y) + b$$

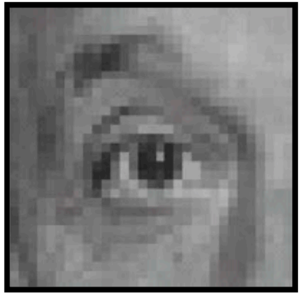


a b
c d

FIGURE 3.9 (a) Aerial image. (b)–(d) Results of applying the transformation in Eq. (3.2-3) with $c = 1$ and $\gamma = 3.0, 4.0$, and 5.0, respectively. (Original image for this example courtesy of NASA.)

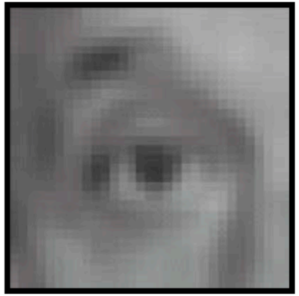
A. Efros.

Image Processing Perspective



Original

$$\frac{1}{9}\begin{array}{|c|c|c|}\hline 1 & 1 & 1 \\ \hline 1 & 1 & 1 \\ \hline 1 & 1 & 1 \\ \hline \end{array} =$$



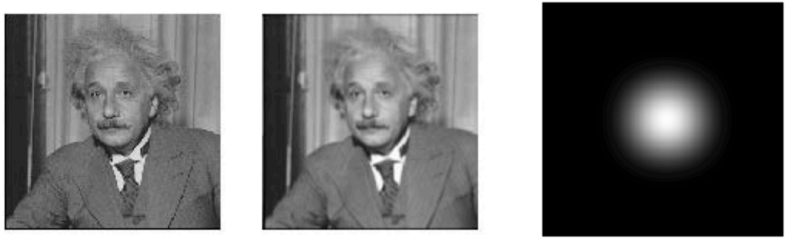
Blur (with a mean filter)

A. Efros.

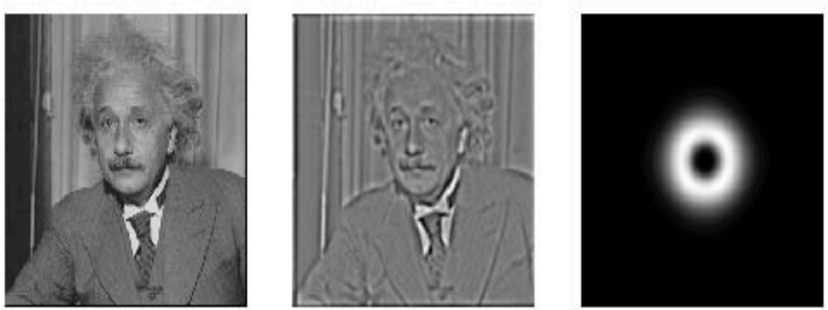
Image Processing Perspective

- e.g. Frequency Analyses

low-pass:



High-pass / band-pass:



Not in the recap (In SOTA lectures)

- Neural Network architecture
- Embodied vision simulators and active agent training
- Foundation models: language, multimodal, generative, etc. models.
- Generalization and Robustness

Some left out elephants

- Plenoptic function -> Radiance Fields -> Neural Radiance Fields

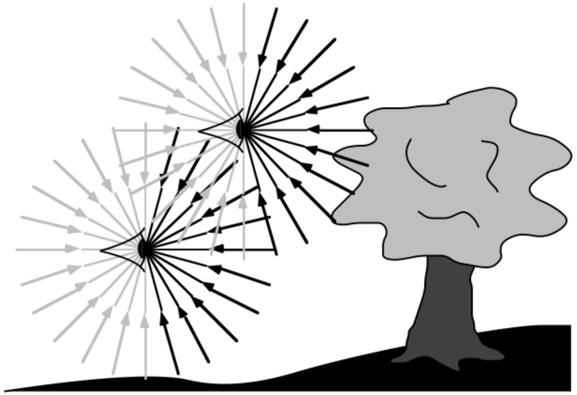
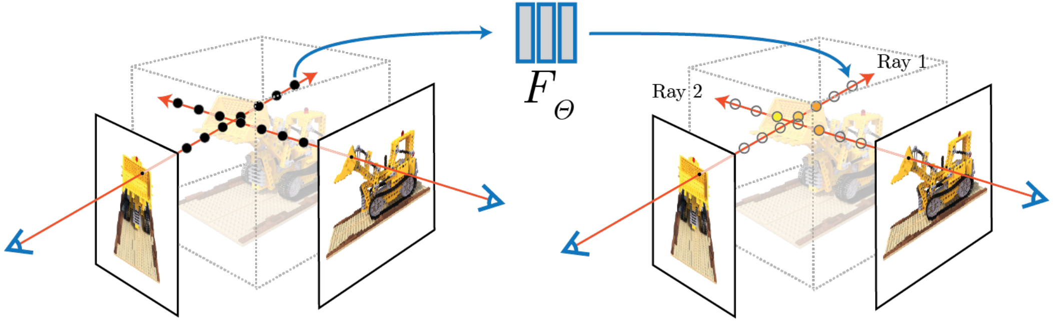


Fig.1.3
The plenoptic function describes the information available to an observer at any point in space and time. Shown here are two schematic eyes-which one should consider to have punctate pupils-gathering pencils of light rays. A real observer cannot see the light rays coming from behind, but the plenoptic function does include these rays.



Some left out elephants

- Plenoptic function -> Radiance Fields -> Neural Radiance Fields
- 3D Gaussain Splatting

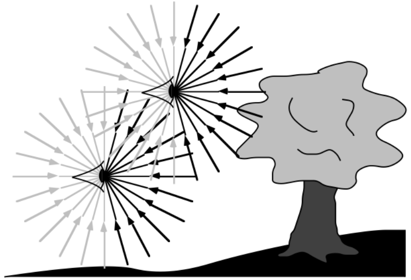
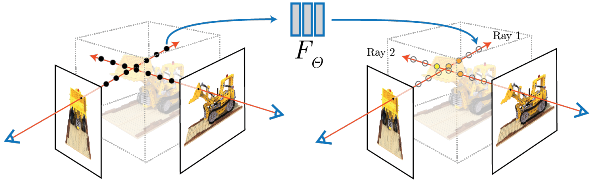
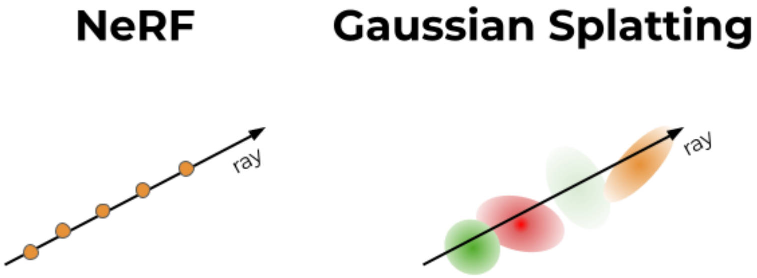
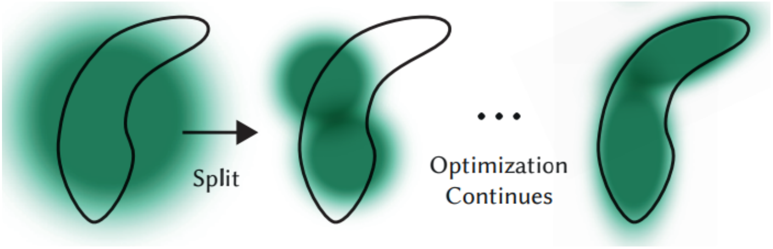


Fig.1.3
The plenoptic function describes the information available to an observer at any point in space and time. Shown here are two schematic eyes-which one should consider to have punctate pupils-gathering pencils of light rays. A real observer cannot see the light rays coming from behind, but the plenoptic function does include these rays.

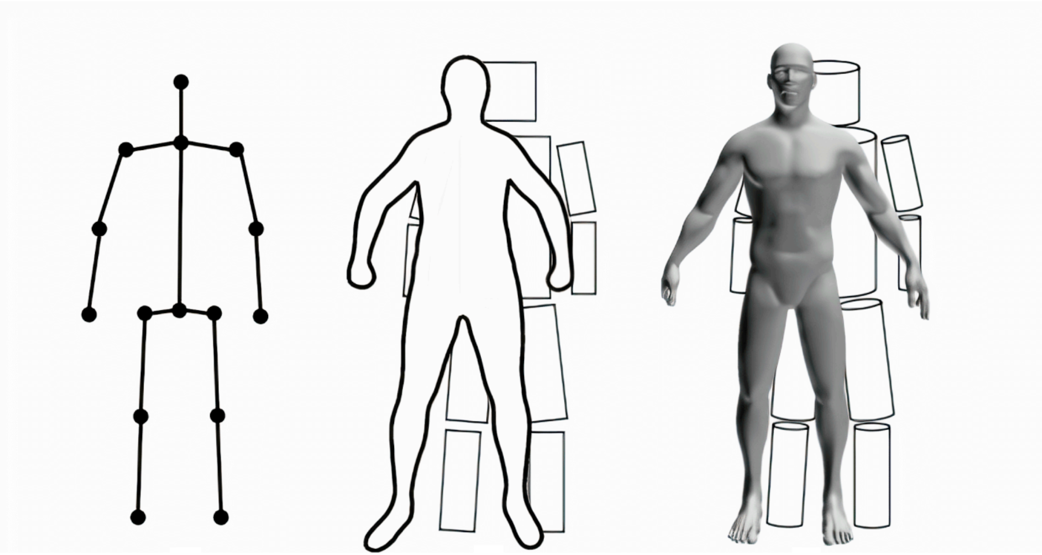






other (more specific) problems

HUMAN BODY MODELS

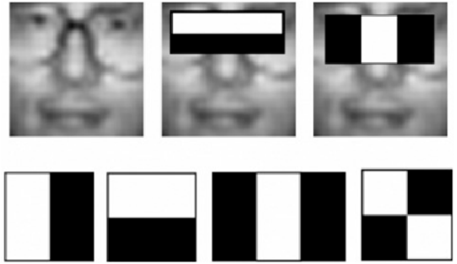


skeleton-based model

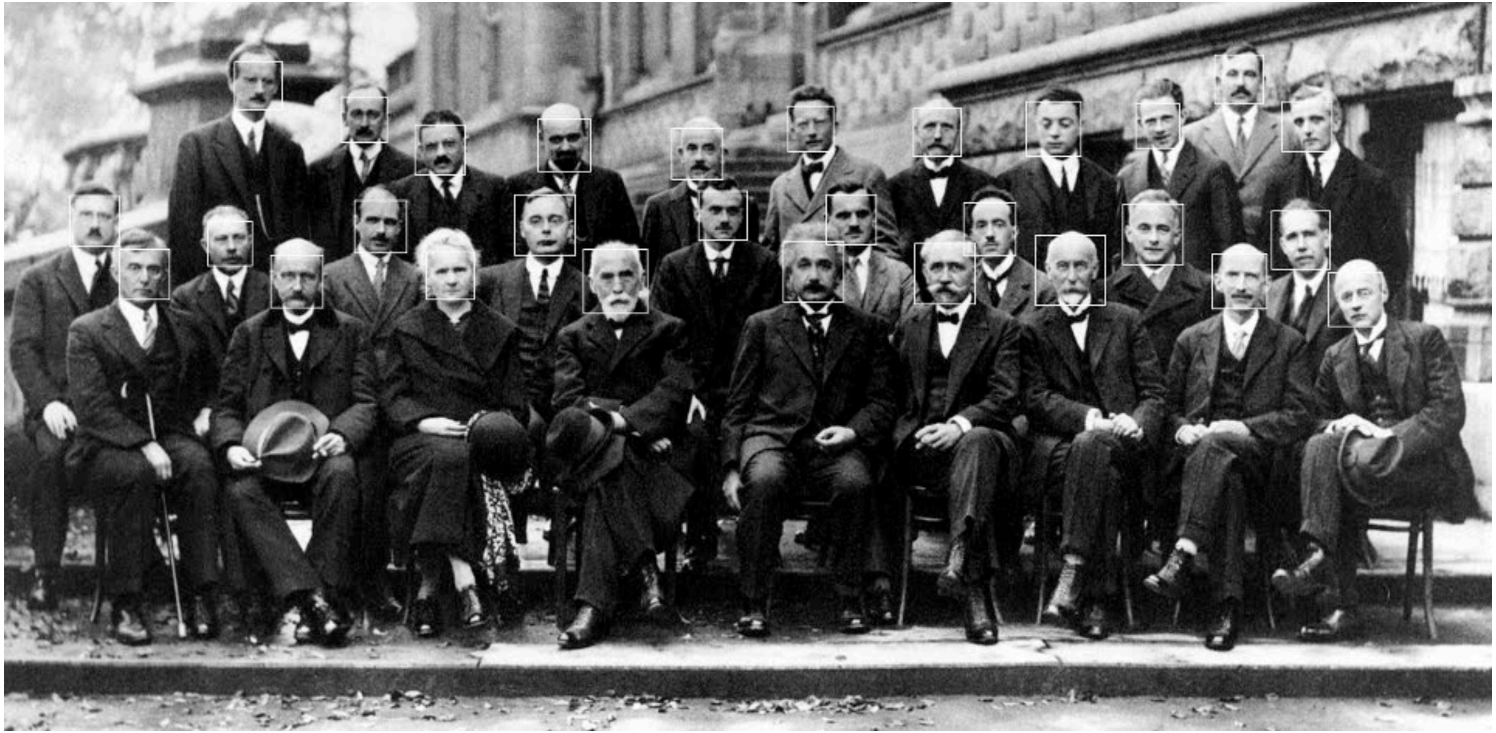
contour-based model

volume-based model

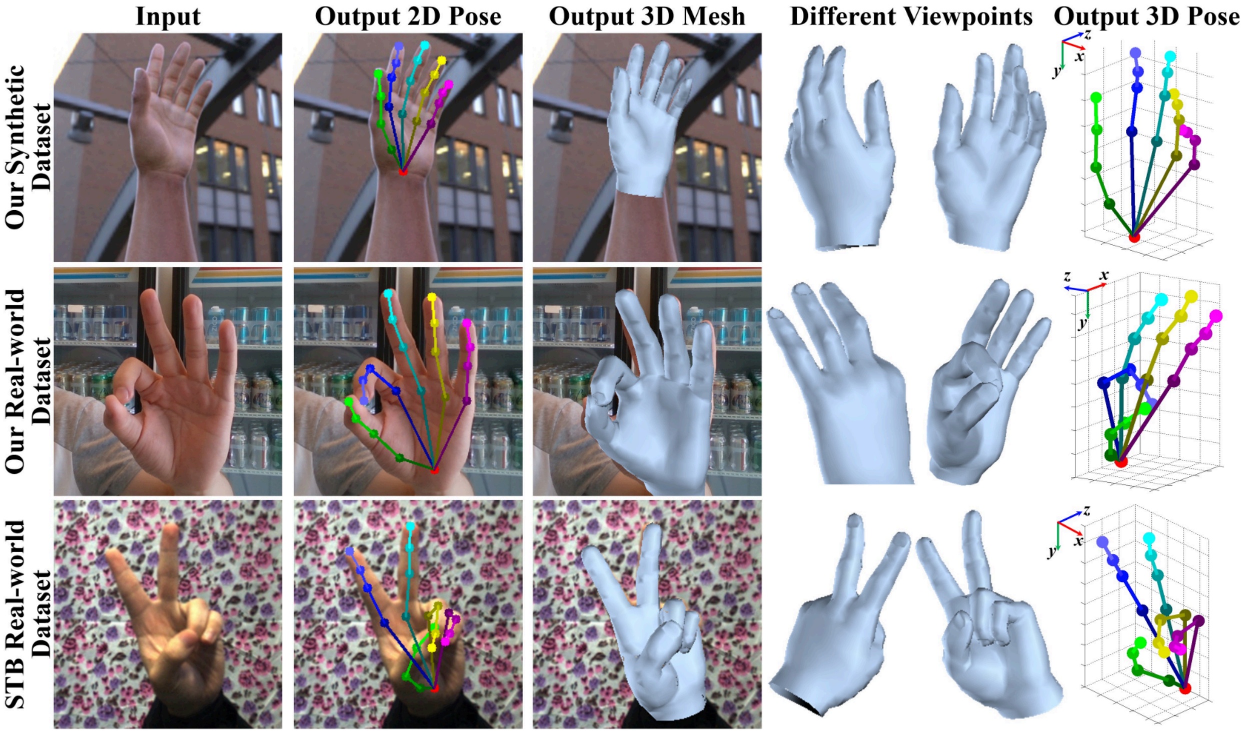
other (more specific) problems



Viola & Jones



other (more specific) problems



Ge et al, 2019

Questions?

https://vilab.epfl.ch/