Reconnaissance d'objets et vision artificielle

https://imagine.enpc.fr/~varolg/teaching/recvis23/

Gül Varol (gul.varol@enpc.fr)

and

Jean Ponce, Armand Joulin, Josef Sivic, Ivan Laptev, Cordelia Schmid, and Mathieu Aubry

Mardis 16h00-19h0, salle Dussane

Planches disponibles après les cours

Nous cherchons
toujours
des stagiaires
à la fin du semestre

Initiation à la vision artificielle

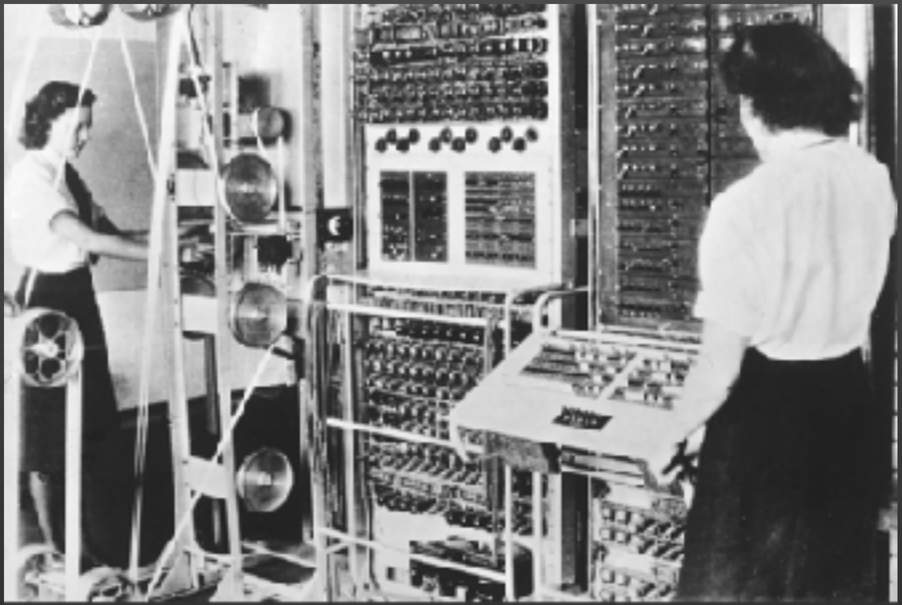
Jean Ponce
(jean.ponce@ens.fr)

Mardis, salle E. Noether, ENS, 9h-12h

Il y a d'autres choses que la reconnaissance visuelle dans la vie

Outline

- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry
- Image processing



What?

Description:

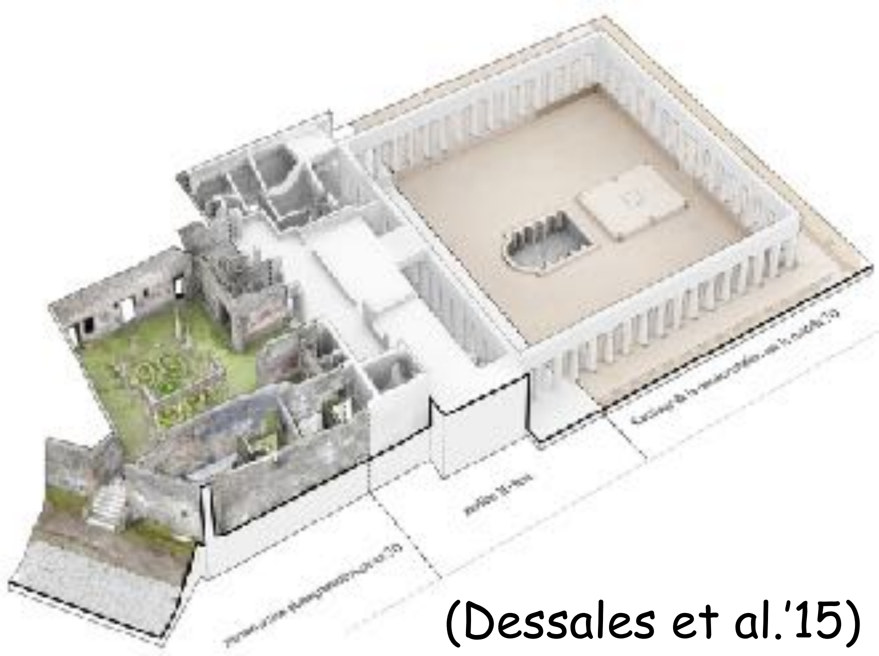
- Street scene
- Bar
- Chairs
- People drinking coffee
- Ashtray, etc

Why?

NAO (Aldebaran Robotics)

(Mairal, Bach, Ponce, PAMI'12)

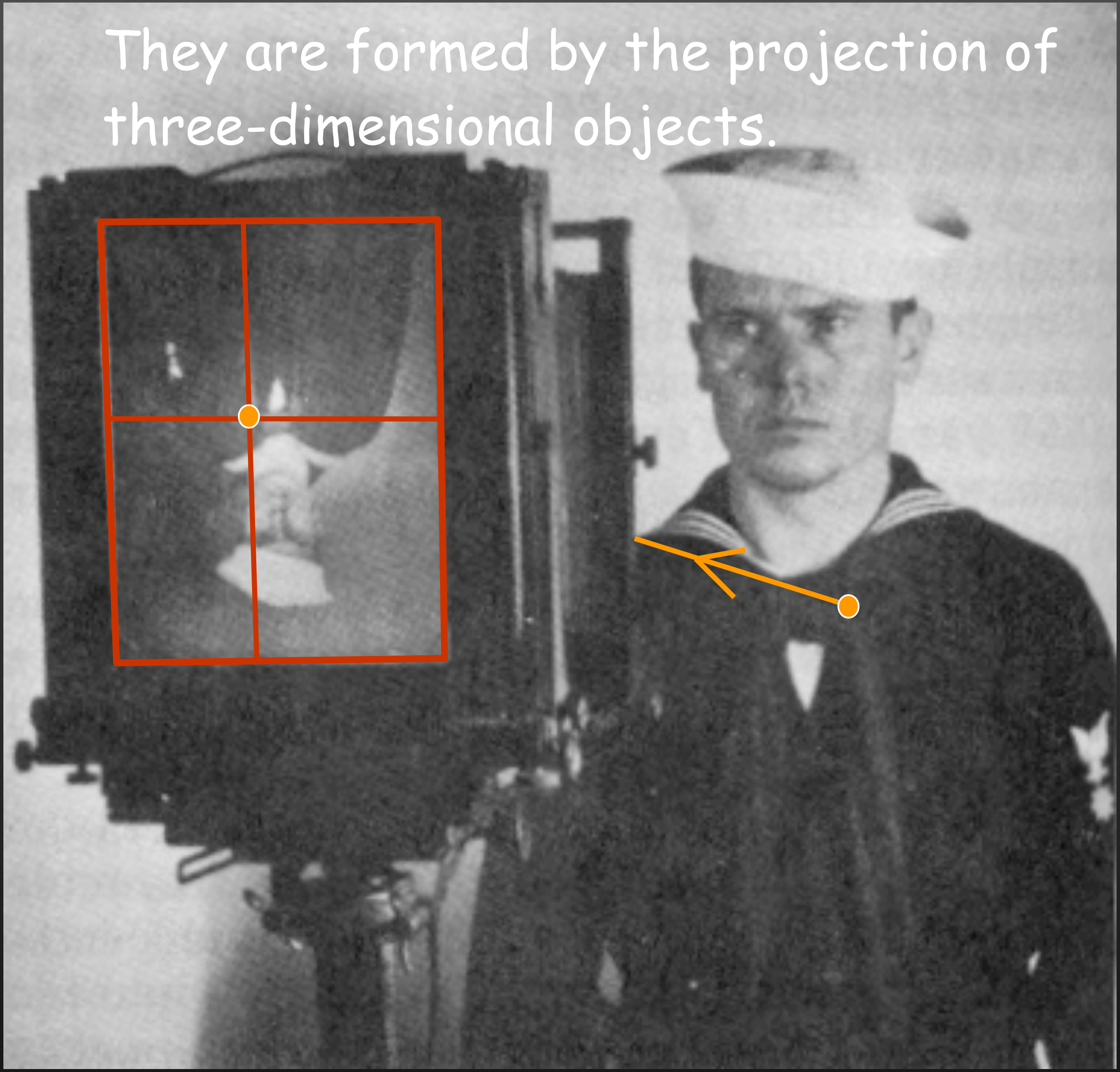
Why?



(Dessales et al.'15)

CMU's Chimp

Facebook's Moments


They are formed by the projection of three-dimensional objects.

Images are brightness/color patterns drawn in a plane.

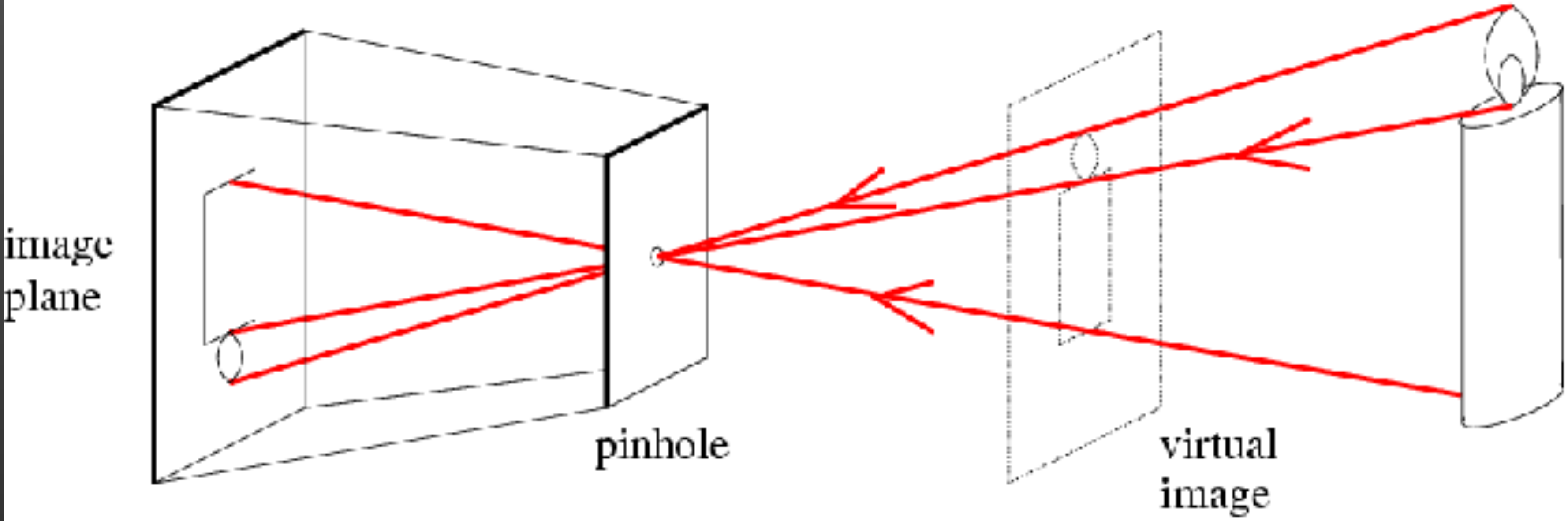
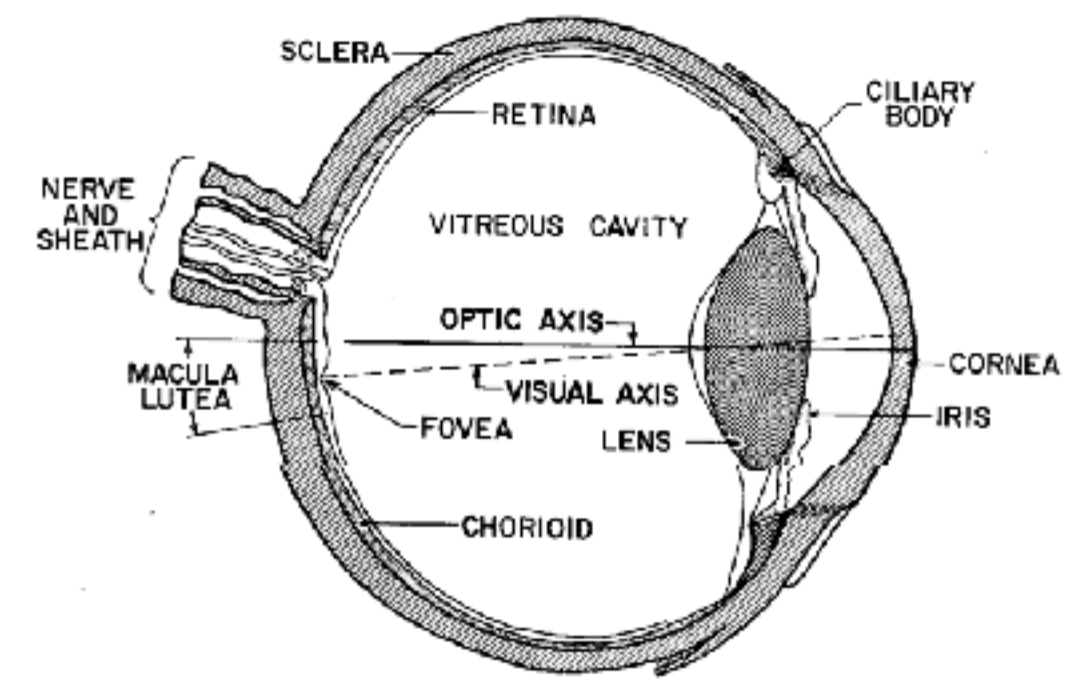
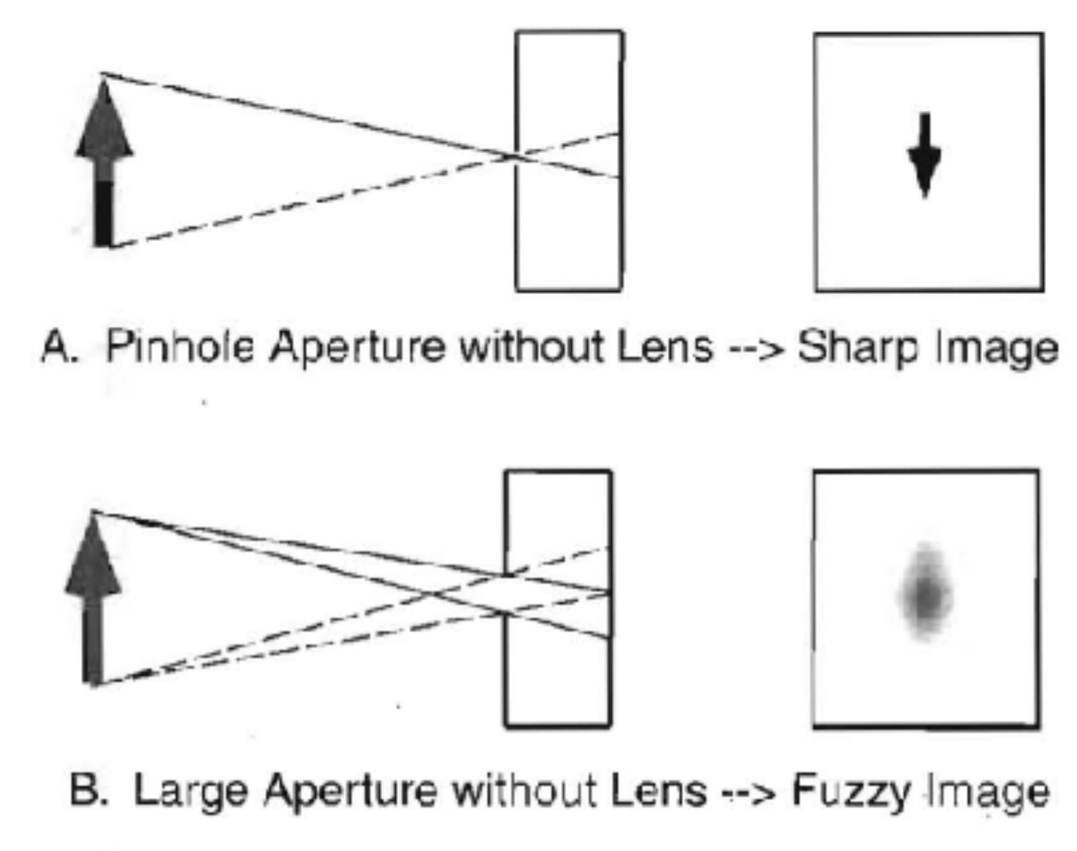


image plane
pinhole
virtual image


SCLERA
RETINA
CILIARY BODY
NERVE AND SHEATH
VITREOUS CAVITY
OPTIC AXIS
CORNEA
MACULA LUTEA
VISUAL AXIS
FOVEA
LENS
IRIS
CHORIOID

Pinhole camera: trade-off between sharpness and light transmission


A. Pinhole Aperture without Lens --> Sharp Image
B. Large Aperture without Lens --> Fuzzy Image

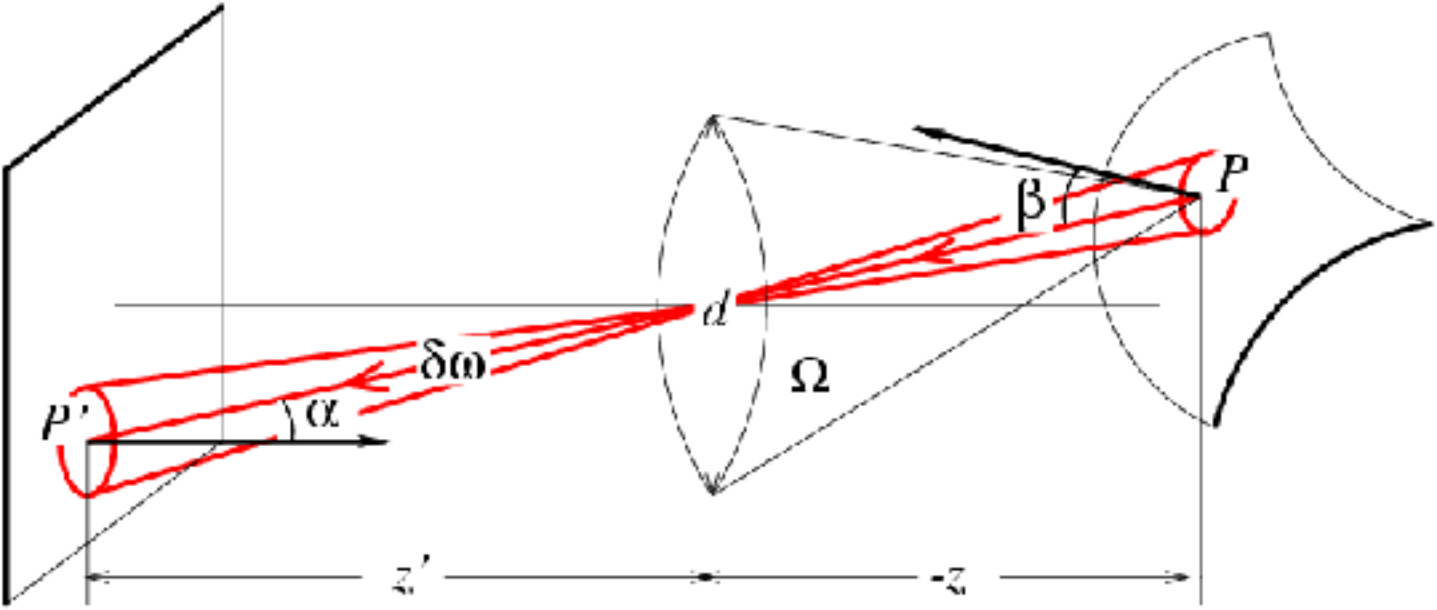
Camera Obscura in Edinburgh

Source: S. Lazebnik

Advantages of lens systems

Lenses

- can focus sharply on close and distant objects
- transmit more light than a pinhole camera

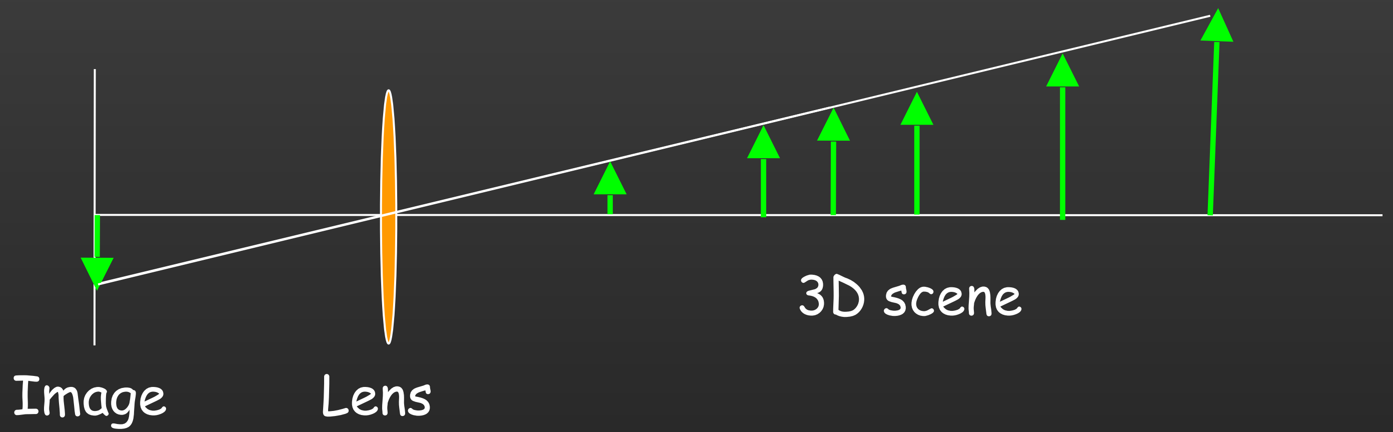


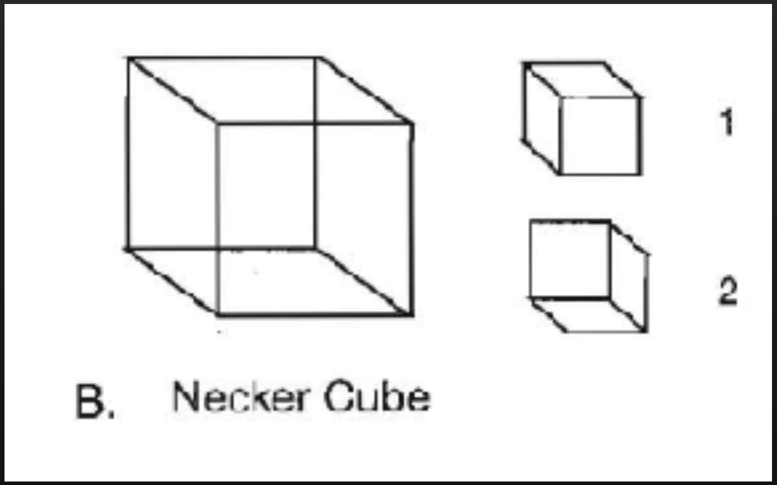
$$\boxed{E = (\Pi/4)\left[\,(d/z')^2 \cos^4\alpha\,\right] L}$$

Fundamental problem

3D world is “flattened” to 2D images

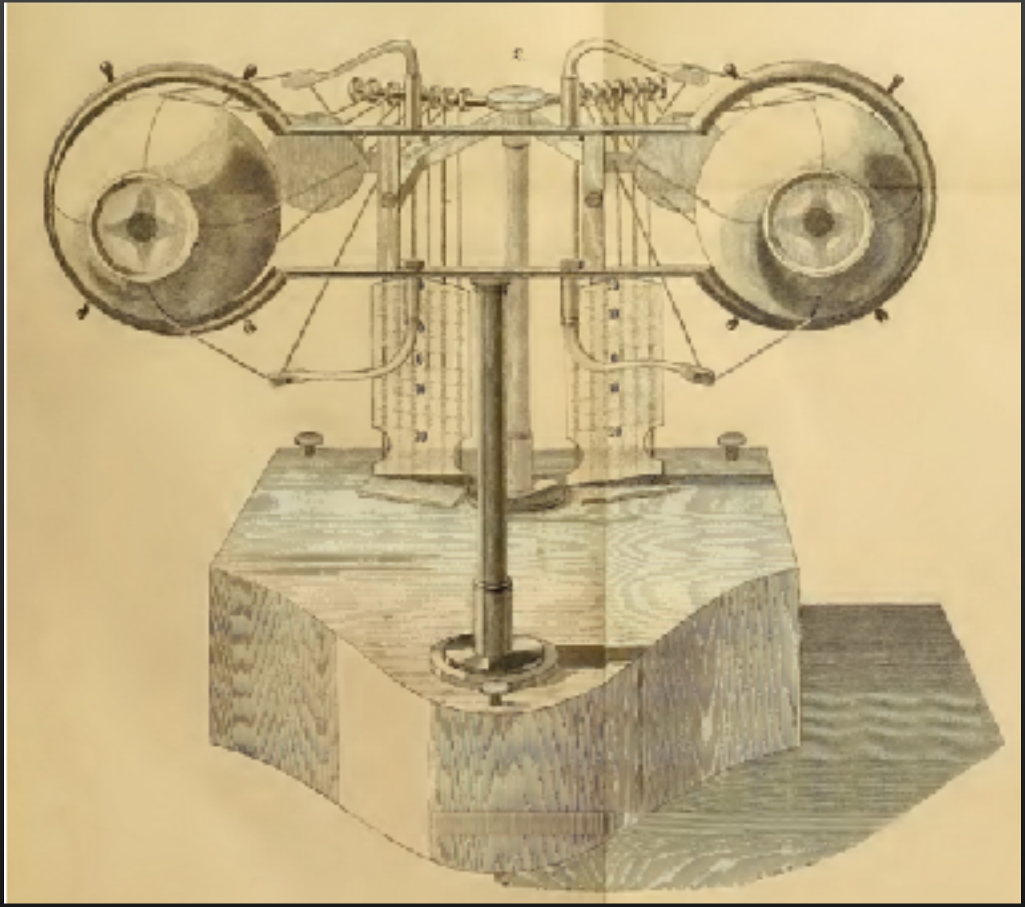
➡ Loss of information



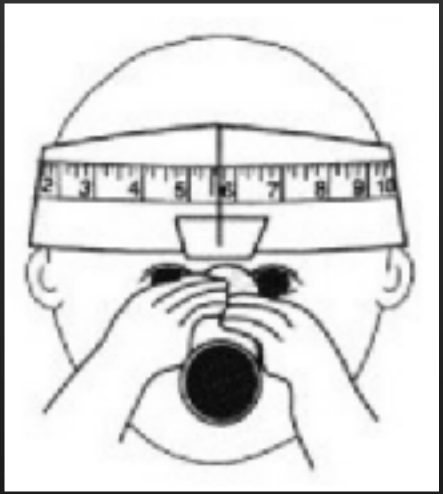


Source: S. Lazebnik

A (naïve) detour through human perception: Seeing with two eyes

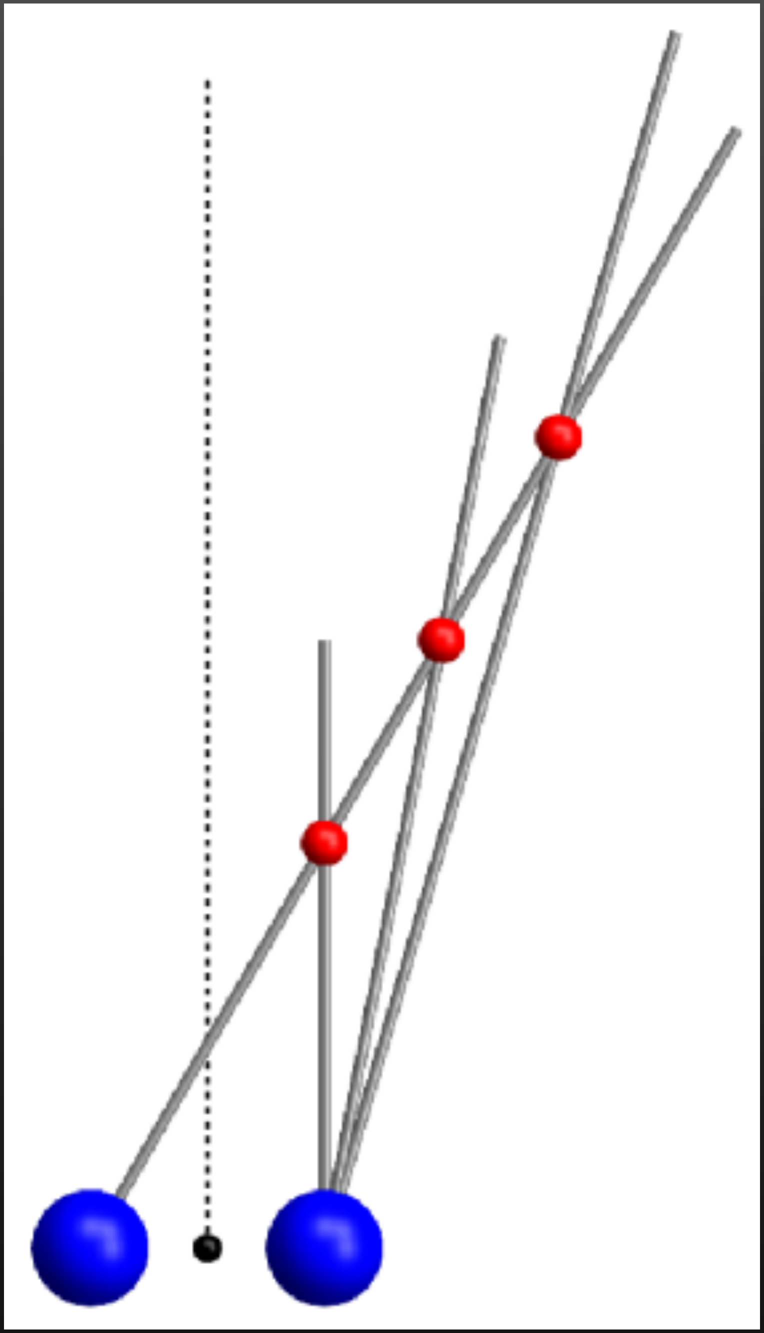


Christian Georg Theodor Ruete



Dominant eye vs Cyclopean vision

Source: J.J. Koenderink



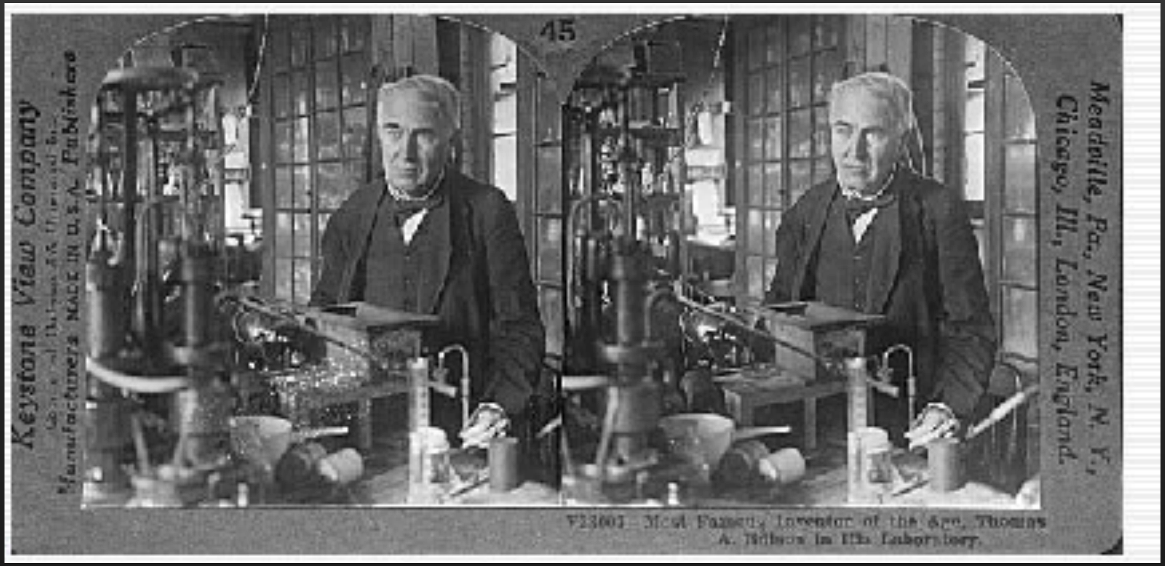
Seeing in depth:

The vergence angle reveals absolute range

Source: J.J. Koenderink

Binocular stereo

Two images can be fused to give a sense of depth!


Keystone View Company
Manufacturers MADE IN U.S.A. Publishers
45
Meadville, Pa., New York, N. Y.,
Chicago, Ill., London, England.
V23001 Most Famous Inventor of the Age, Thomas
A. Edison in His Laboratory.

Stereograms: Invented by Charles Wheatstone, 1838

Triangulation

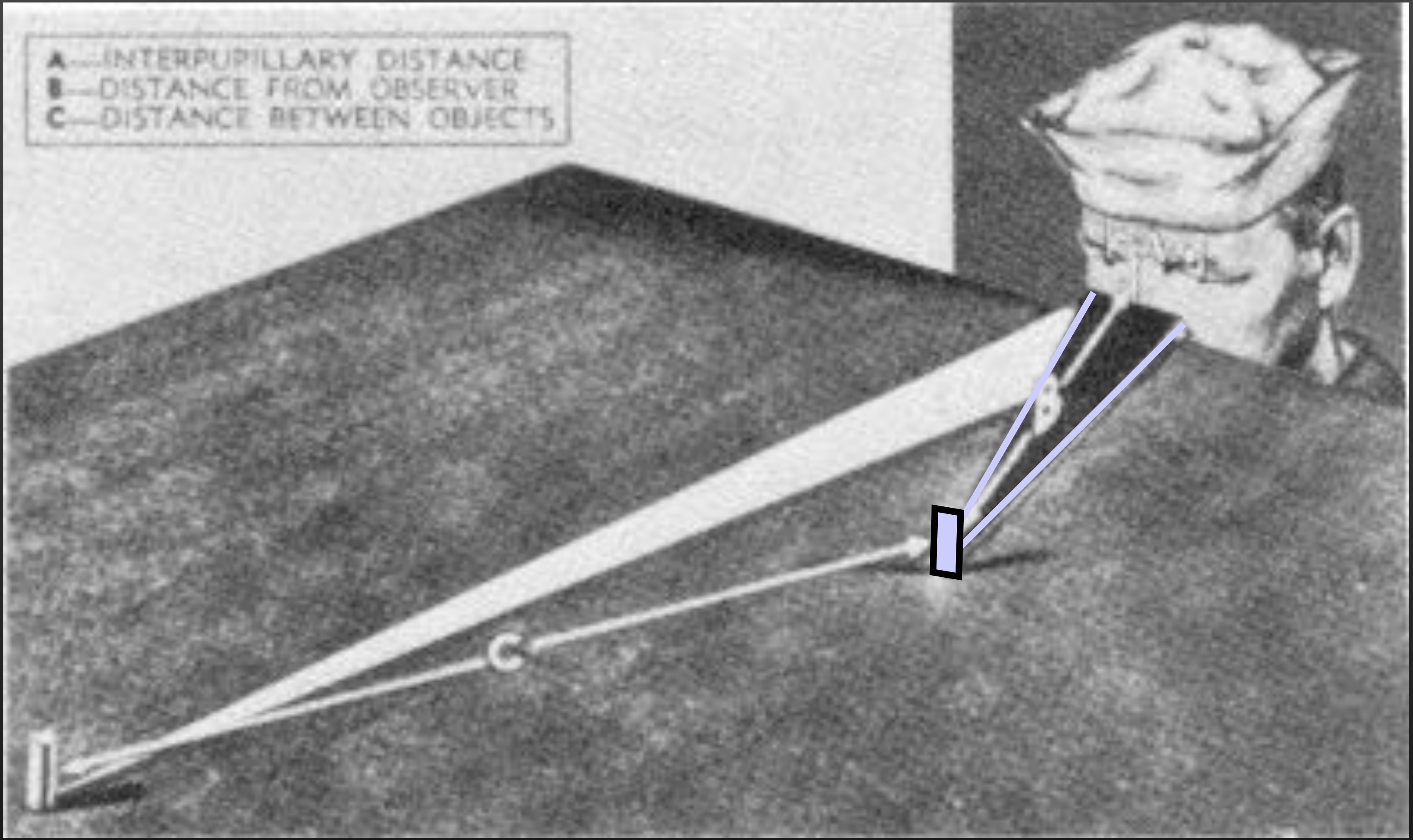


Figure from “US Navy Manual of Basic Optics and Optical Instruments”, Bureau of Naval Personnel. Reprinted by Dover Publications, Inc., 1969.

Triangulation

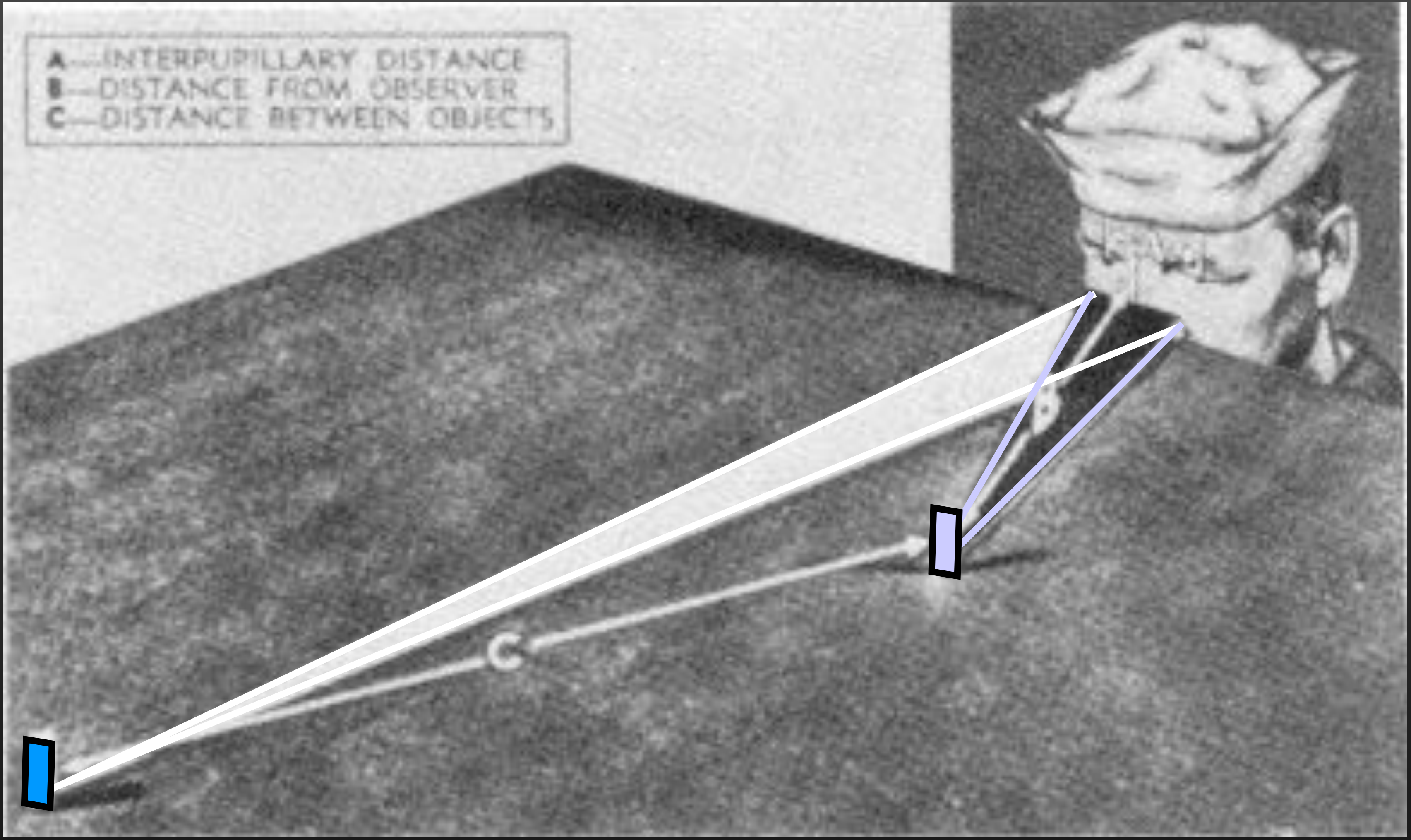


Figure from "US Navy Manual of Basic Optics and Optical Instruments", Bureau of Naval Personnel. Reprinted by Dover Publications, Inc., 1969.

Why movies look "flat" on TV

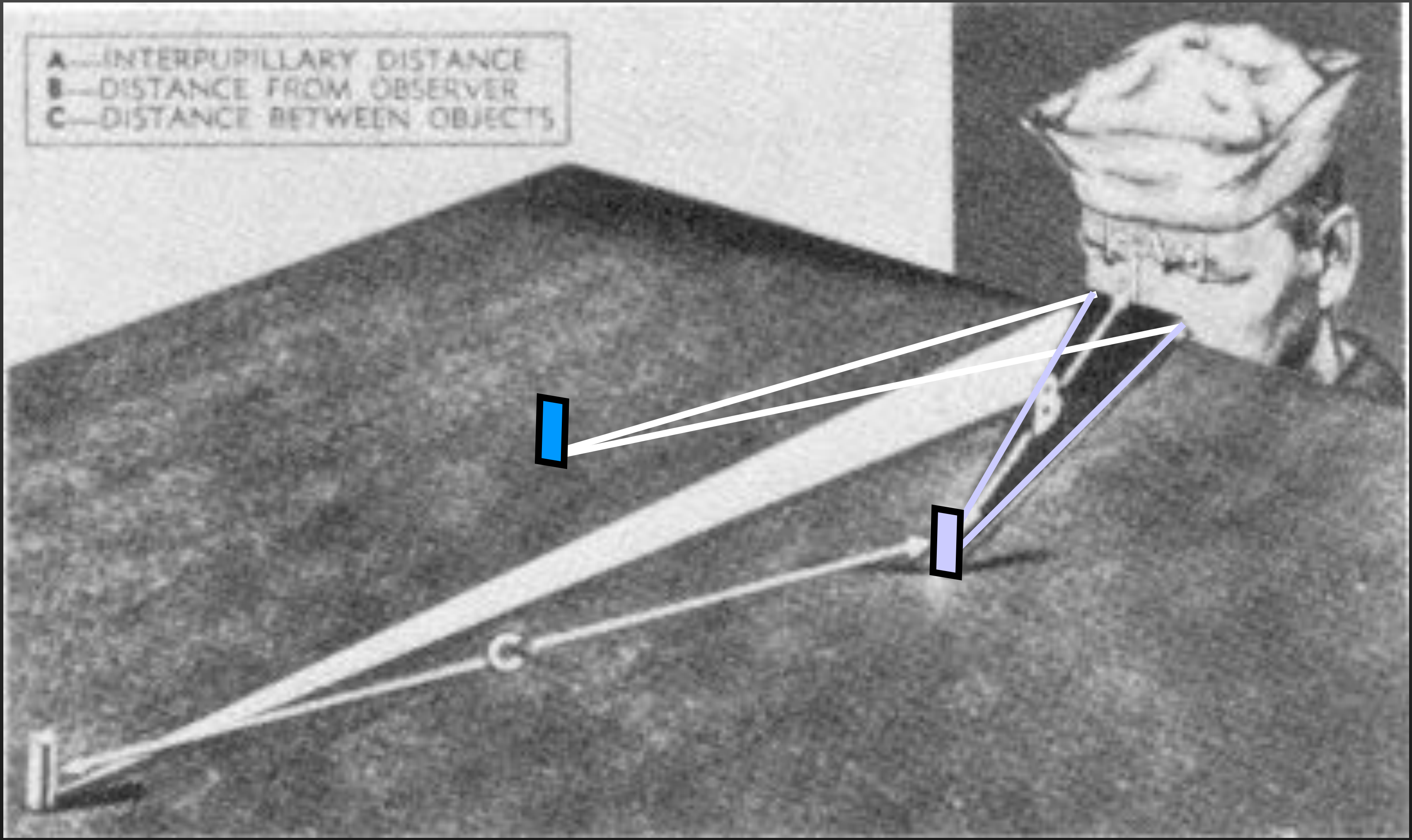


Figure from "US Navy Manual of Basic Optics and Optical Instruments", Bureau of Naval Personnel. Reprinted by Dover Publications, Inc., 1969.

Why movies look "flat" on TV

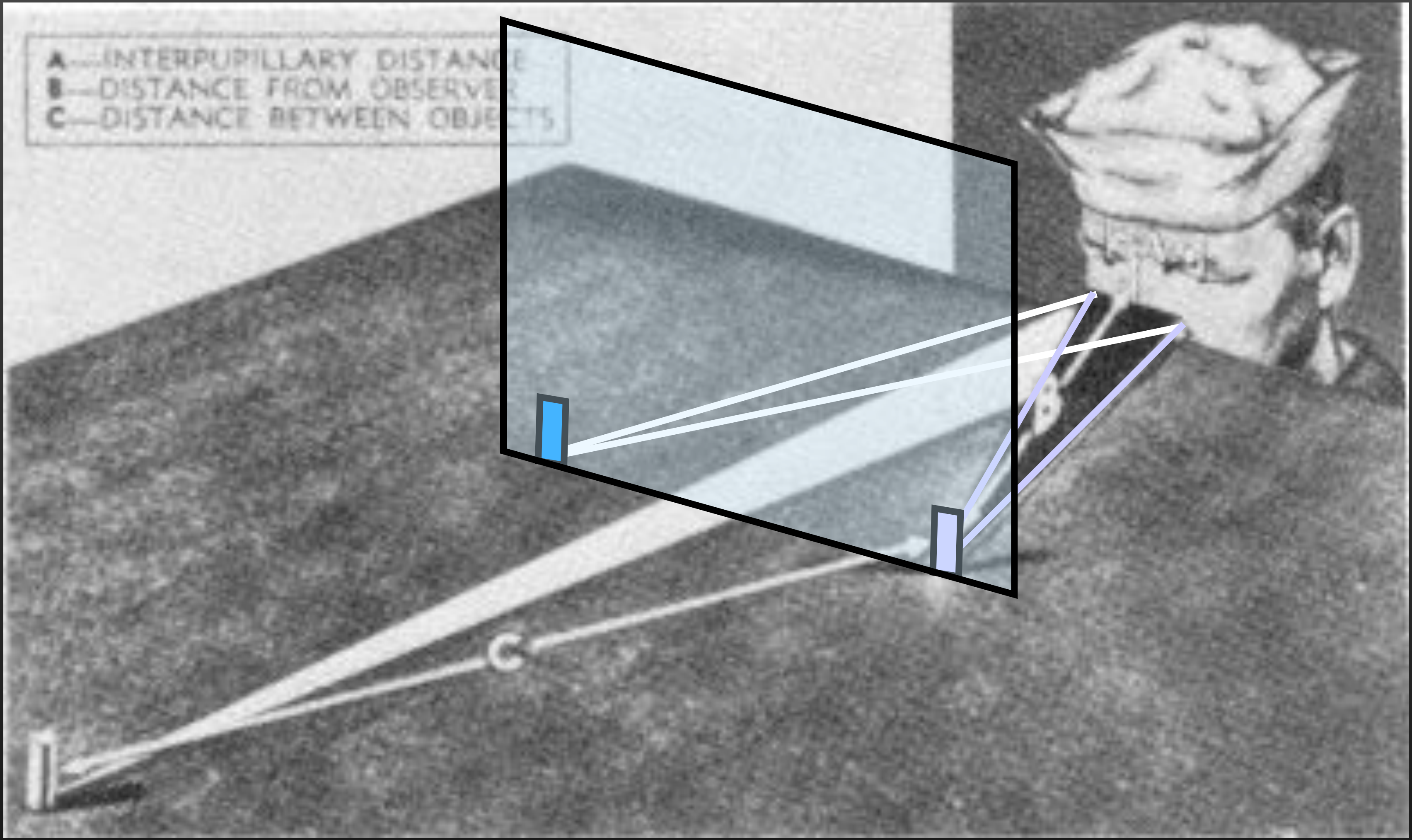
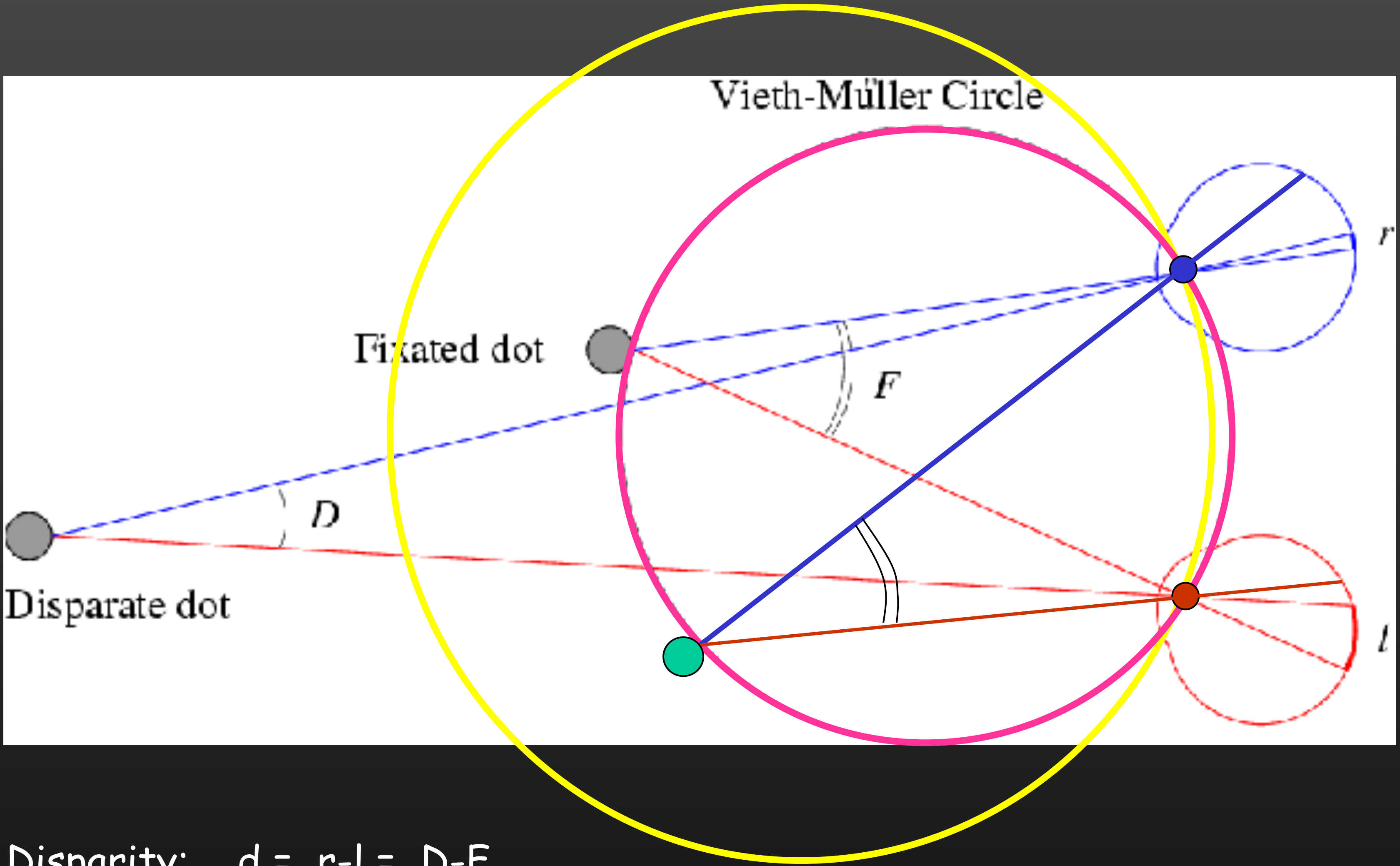


Figure from "US Navy Manual of Basic Optics and Optical Instruments", Bureau of Naval Personnel. Reprinted by Dover Publications, Inc., 1969.

Triangulation for human eyes



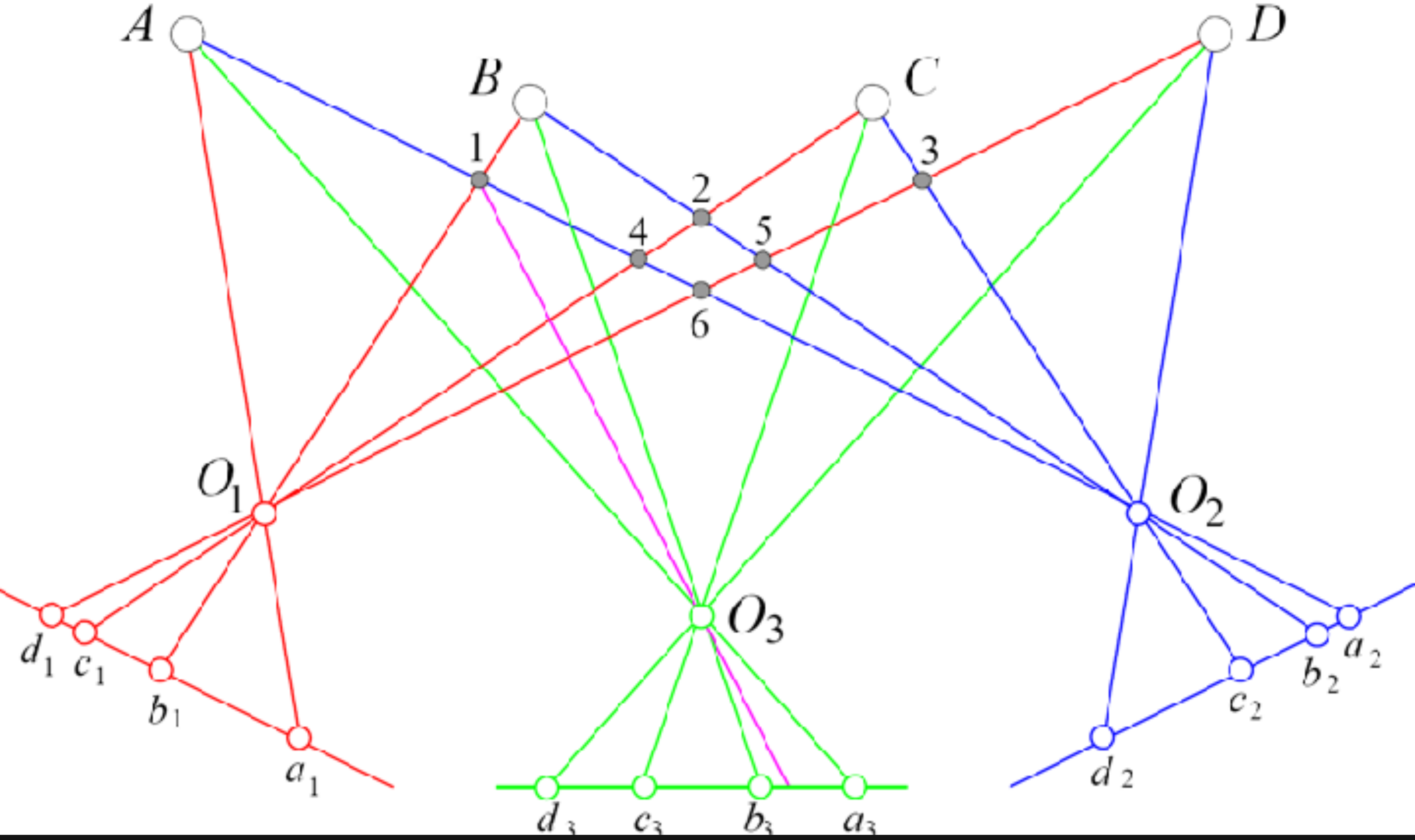
Disparity: d = r-l = D-F.

d<0

In 3D, the horopter.

Binocular fusion: A problem of correspondence

A third eye might come in handy

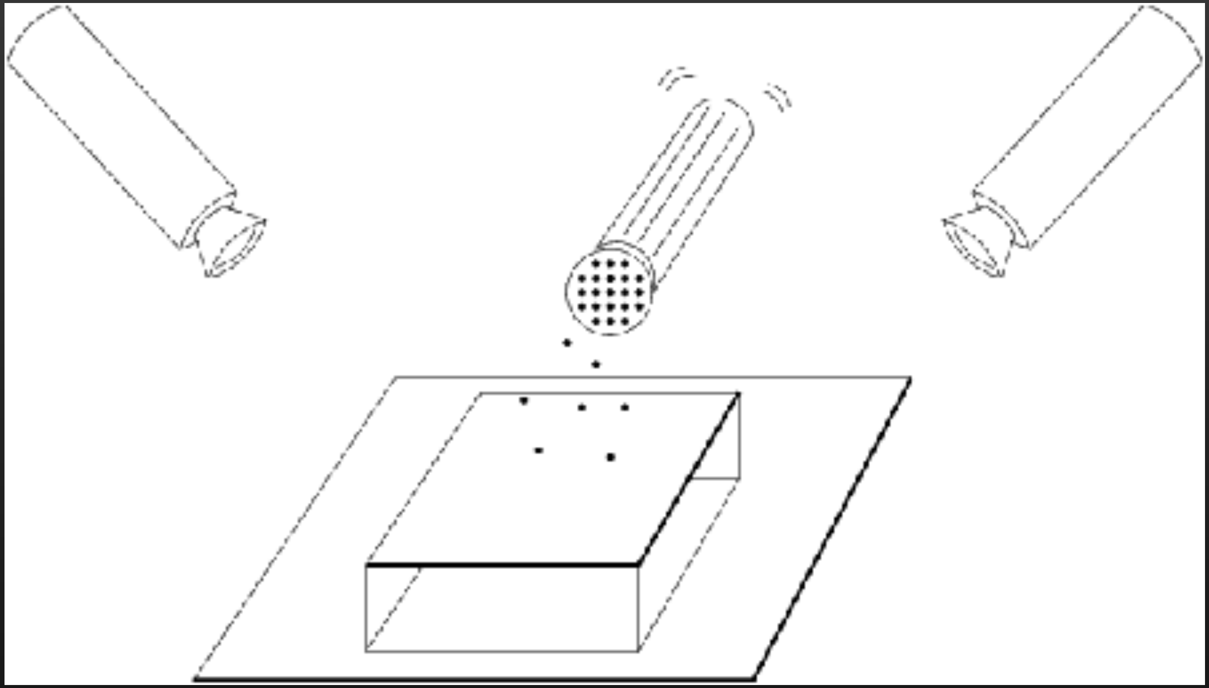


What is the mechanism behind human binocular fusion?

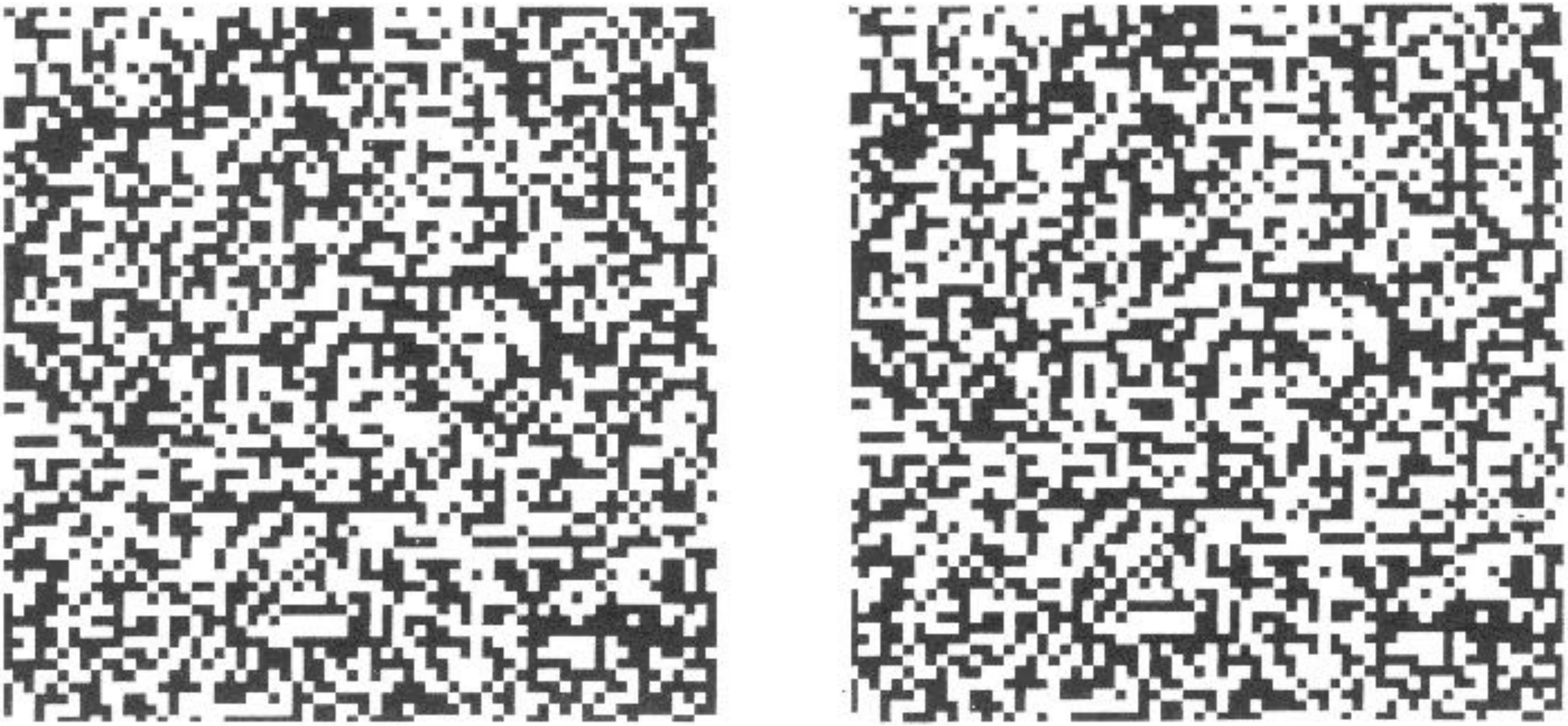
How are the correspondences established?

Julesz (1971): Is the mechanism for binocular fusion
a monocular process or a binocular one??

- There is anecdotal evidence for the latter (camouflage).

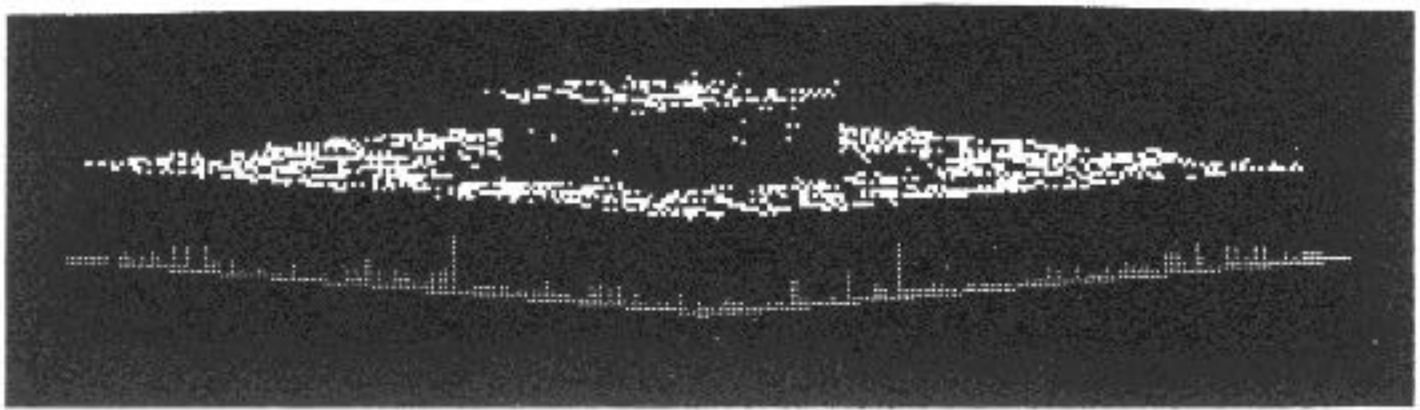


- Random dot stereograms provide an objective answer

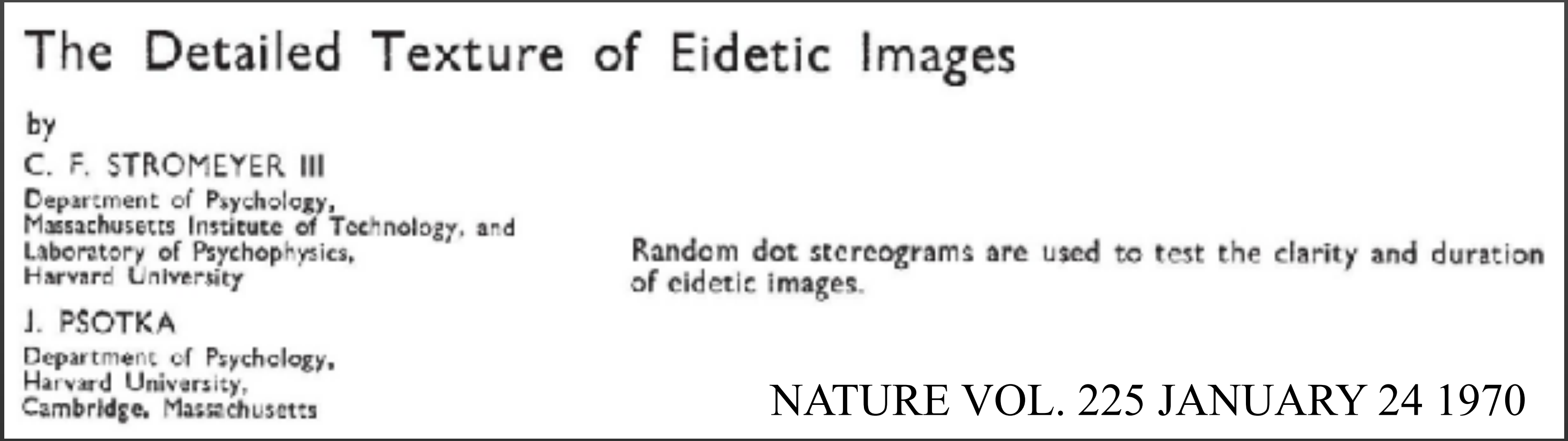


Left

Right



The curious case of Elizabeth Stromeyer



The Detailed Texture of Eidetic Images

by

C. F. STROMEYER III

Department of Psychology,
Massachusetts Institute of Technology, and
Laboratory of Psychophysics,
Harvard University

J. PSOTKA

Department of Psychology,
Harvard University,
Cambridge, Massachusetts

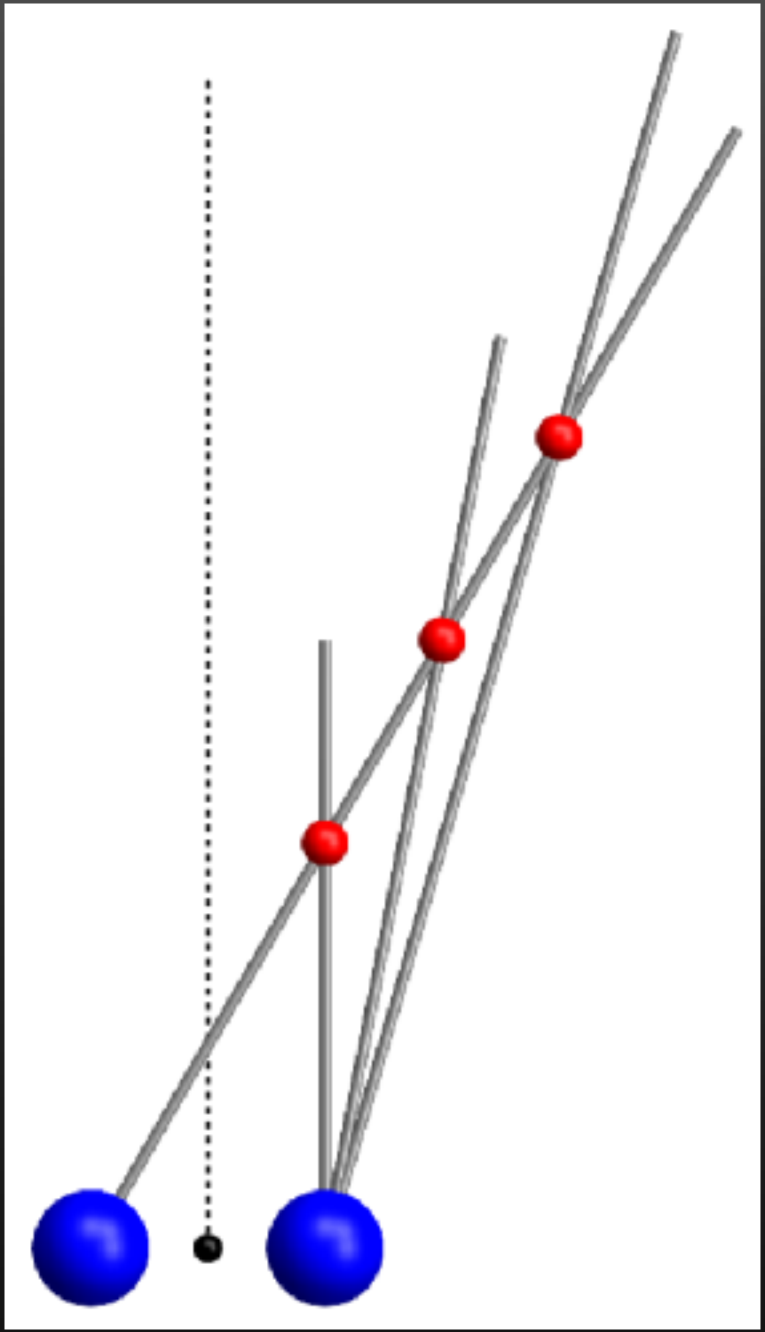
Random dot stereograms are used to test the clarity and duration of eidetic images.

NATURE VOL. 225 JANUARY 24 1970

Excerpts:

We have found, quite by accident, an observer who can accurately report the figure seen in depth when the interval between the observations is as great as 24 h. The observer never guessed or hesitated in making reports, but immediately reported the figures and claimed the task was "ridiculously easy".

Recently we have successfully carried out double-blind random-dot stereogram experiments with our observer; neither the experimenter nor the observer knew what the figures were. Patterns with ten-thousand elements were used with intervals as long as 3 days; and million dot patterns with intervals as long as 4 h.



Back to depth perception:

The vergence angle reveals absolute range

But (Helmholtz 1860’s):

- There is evidence showing that vergence angles cannot be measured precisely.
- People get fooled by bas-relief sculptures.
- Relative depth can be judged accurately.

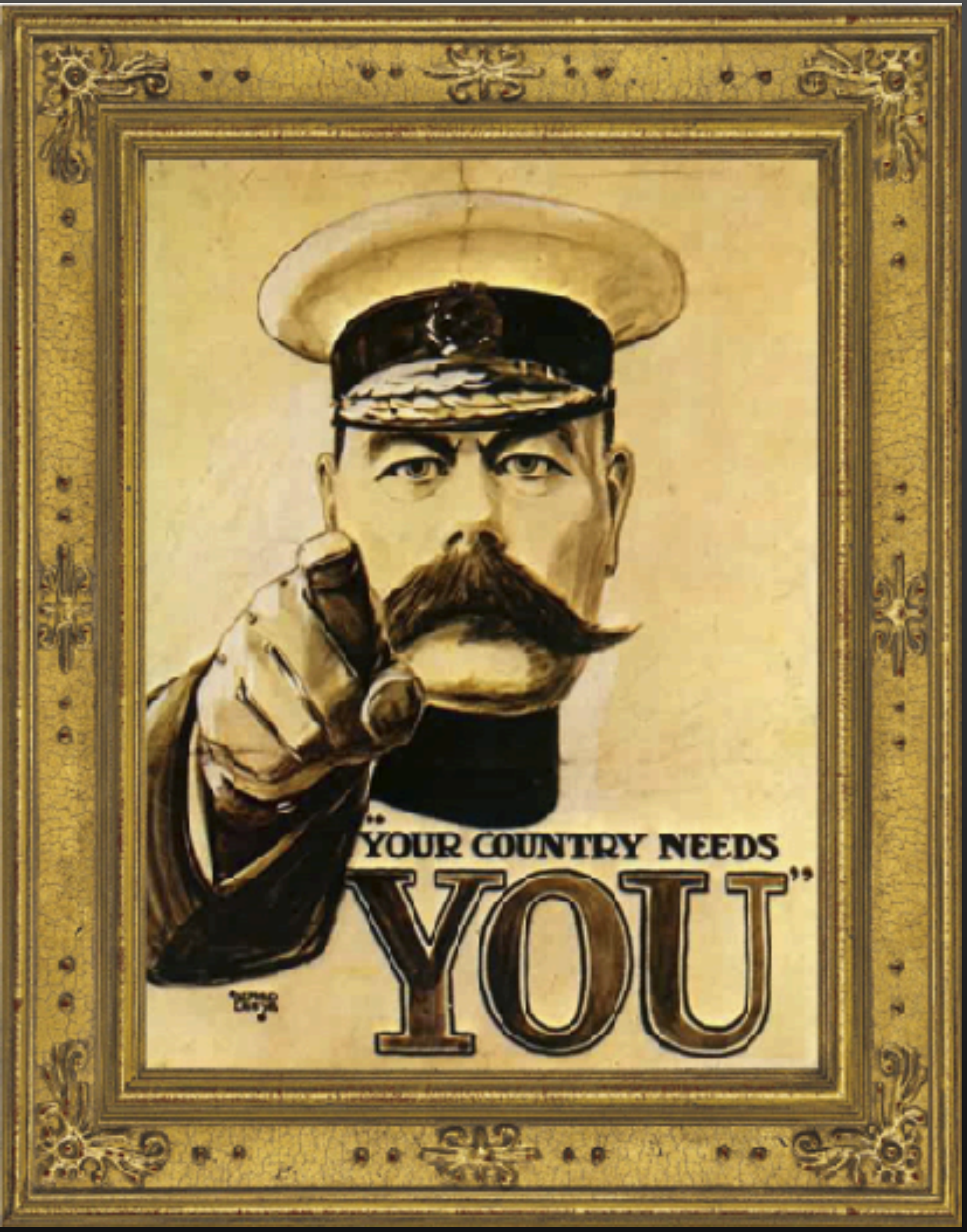
Steropsis is spatial (3D) vision.
It is not limited to binocuar steropsis.
Of course the lamb “sees depth” too!

Source: J.J. Koenderink

"There is little doubt that we share awareness with (at least) the other vertebrates. They should be your friends, even if they eat you when hungry, and even if you eat them."

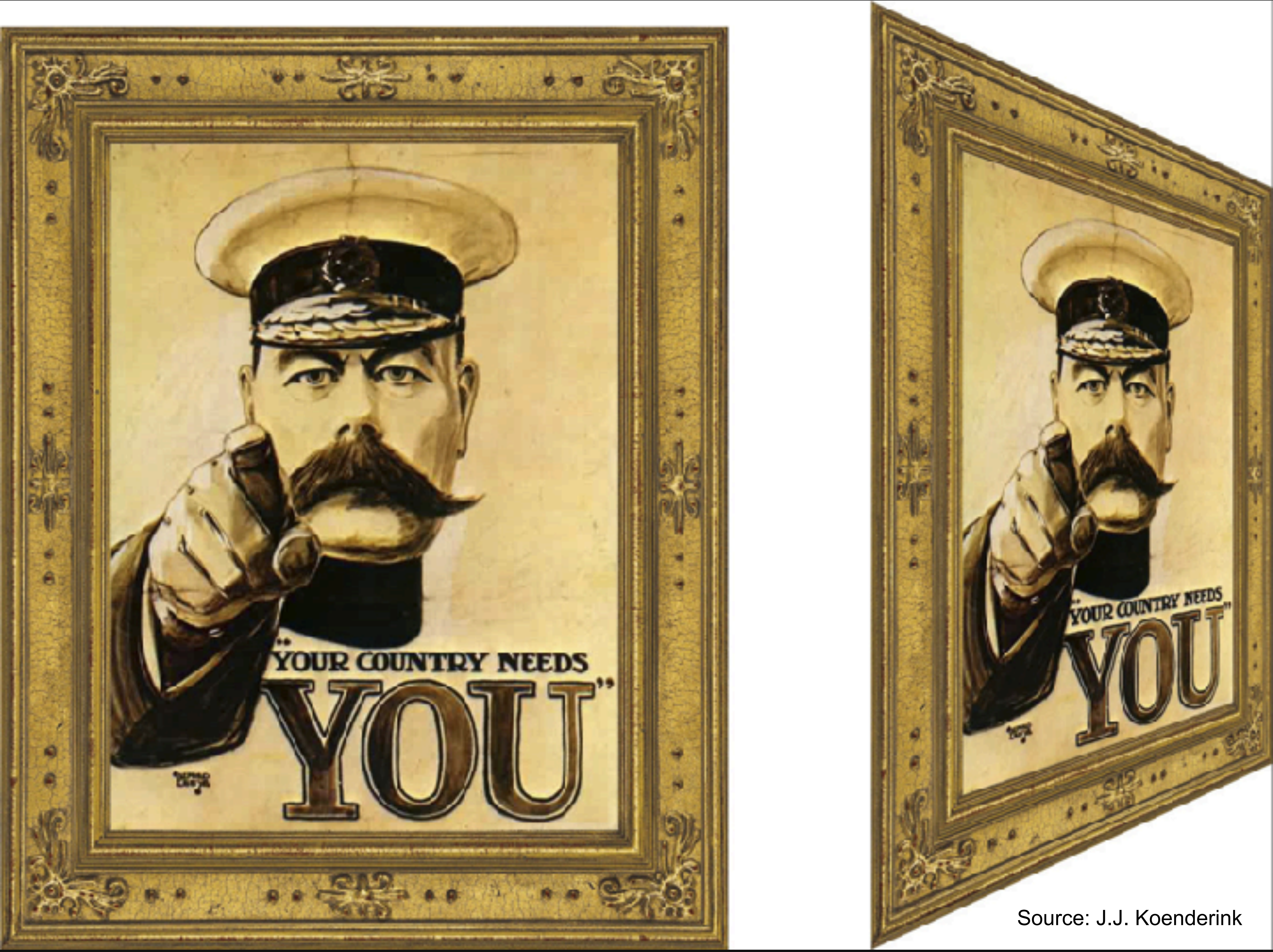
And close one eye, for Pete's sake. Does the world suddenly look flat to you?

Source: J.J. Koenderink

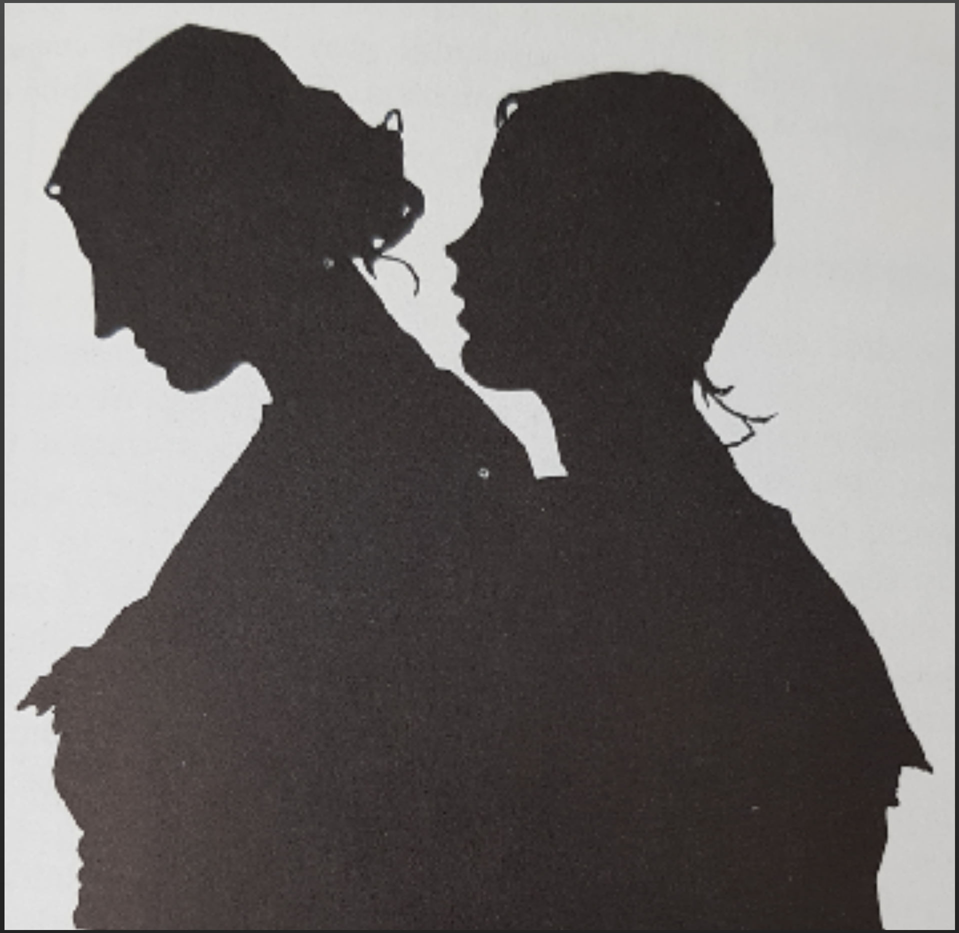


Let us look into the picture

What happens if I turn the frame 30 degrees in depth?



Source: J.J. Koenderink

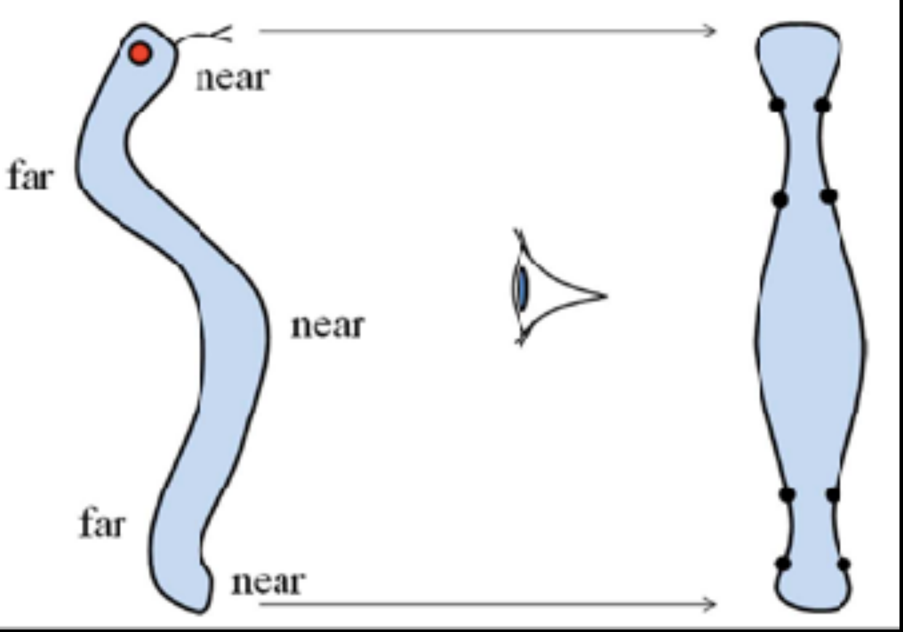


Source: B.P.K. Horn

«This particular picture may lead us to believe that mere silhouettes can convey a great deal of information about three-dimensional objects. The artist's carefully chosen viewpoint and our familiarity with the subject matter conspire to give us this impression. Silhouettes of unfamiliar objects, taken from randomly chosen viewpoints, are typically quite difficult to interpret. » (Horn, Robot vision, 1986)

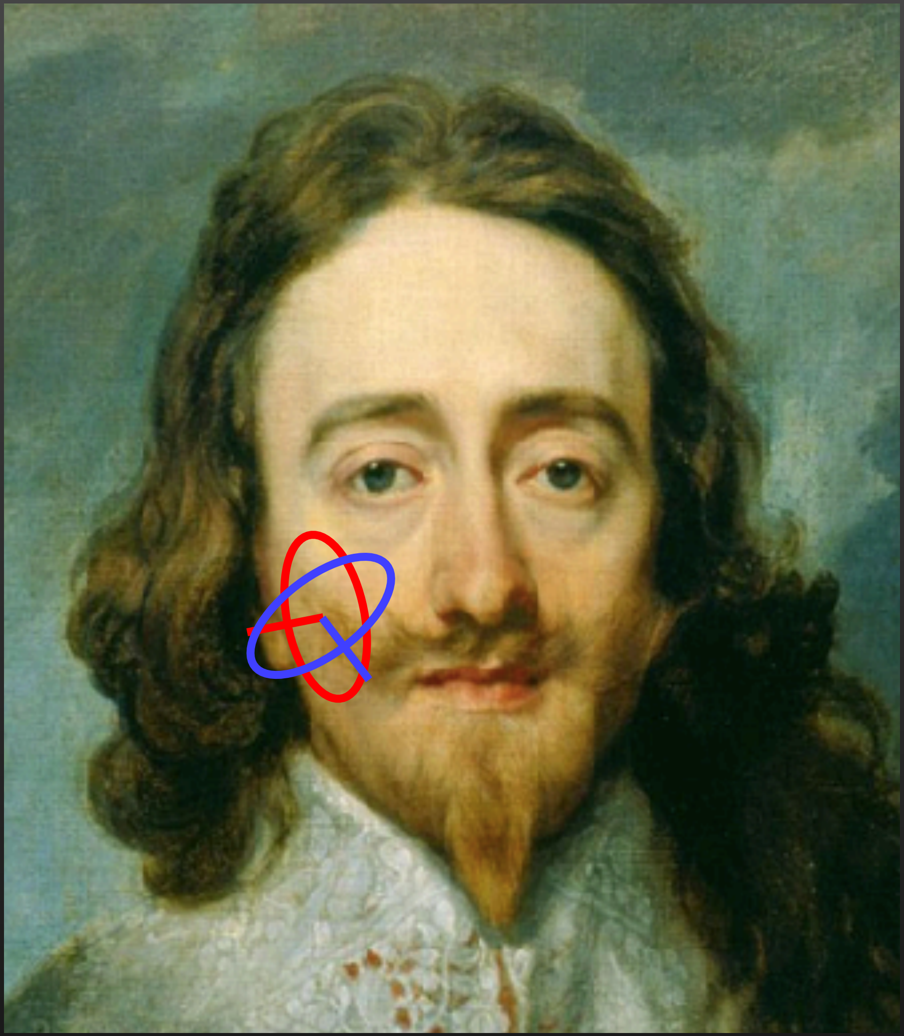
What does the occluding contour tell us about shape?

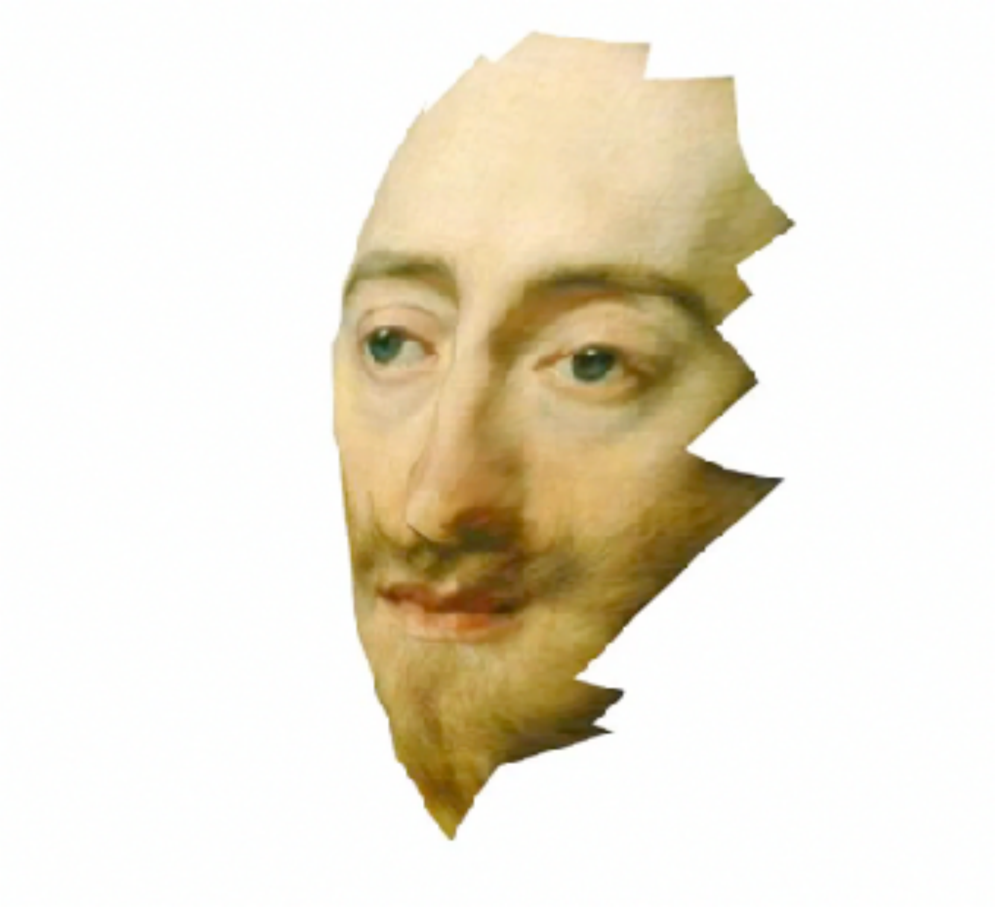
Nothing (Marr'77) ?



Where do the concave points project?

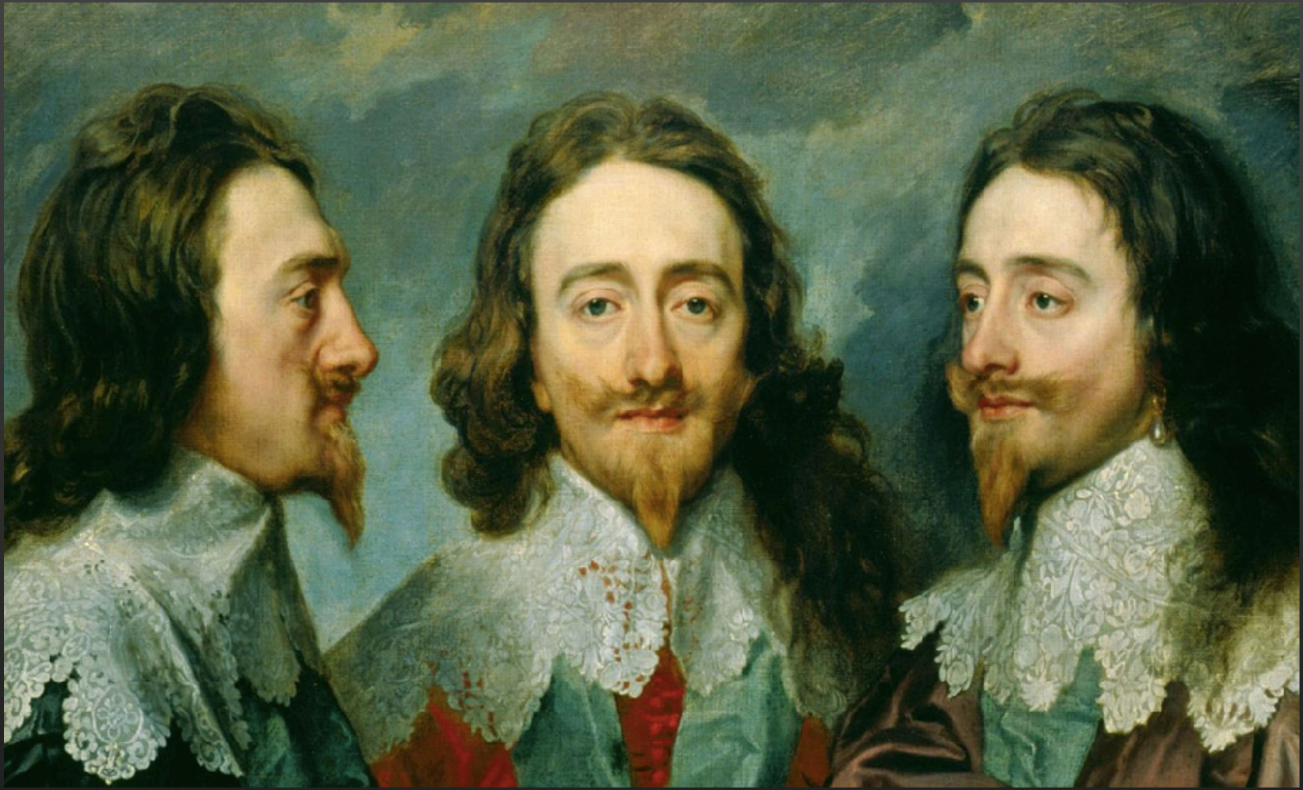
Probing shape perception from pictures with gauge figures

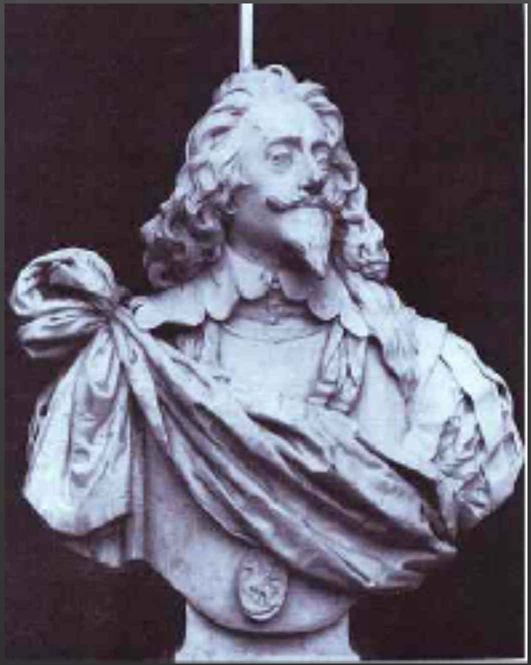




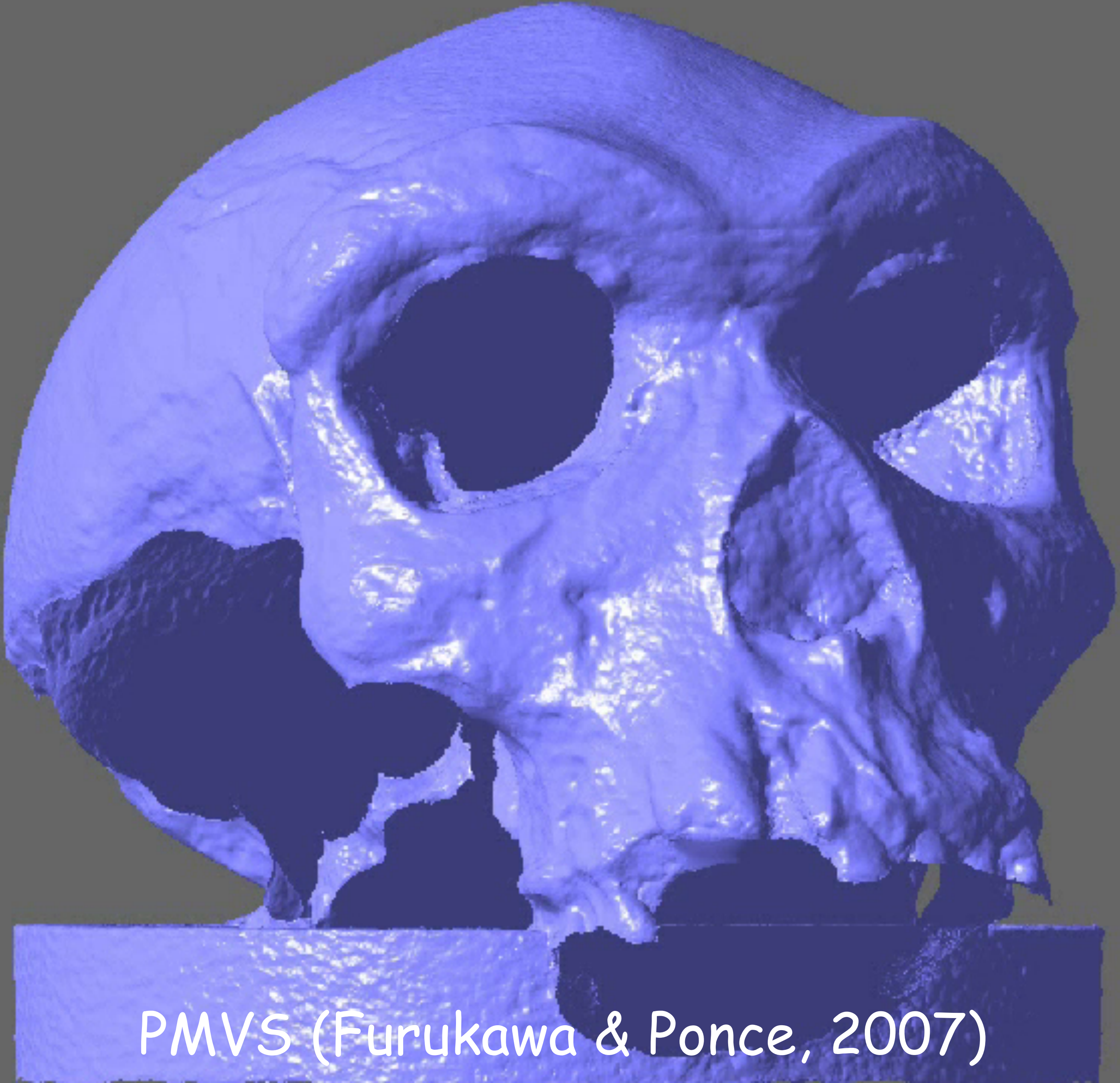
Van Dyck's portrait of Charles I (detail)

Source: J.J. Koenderink



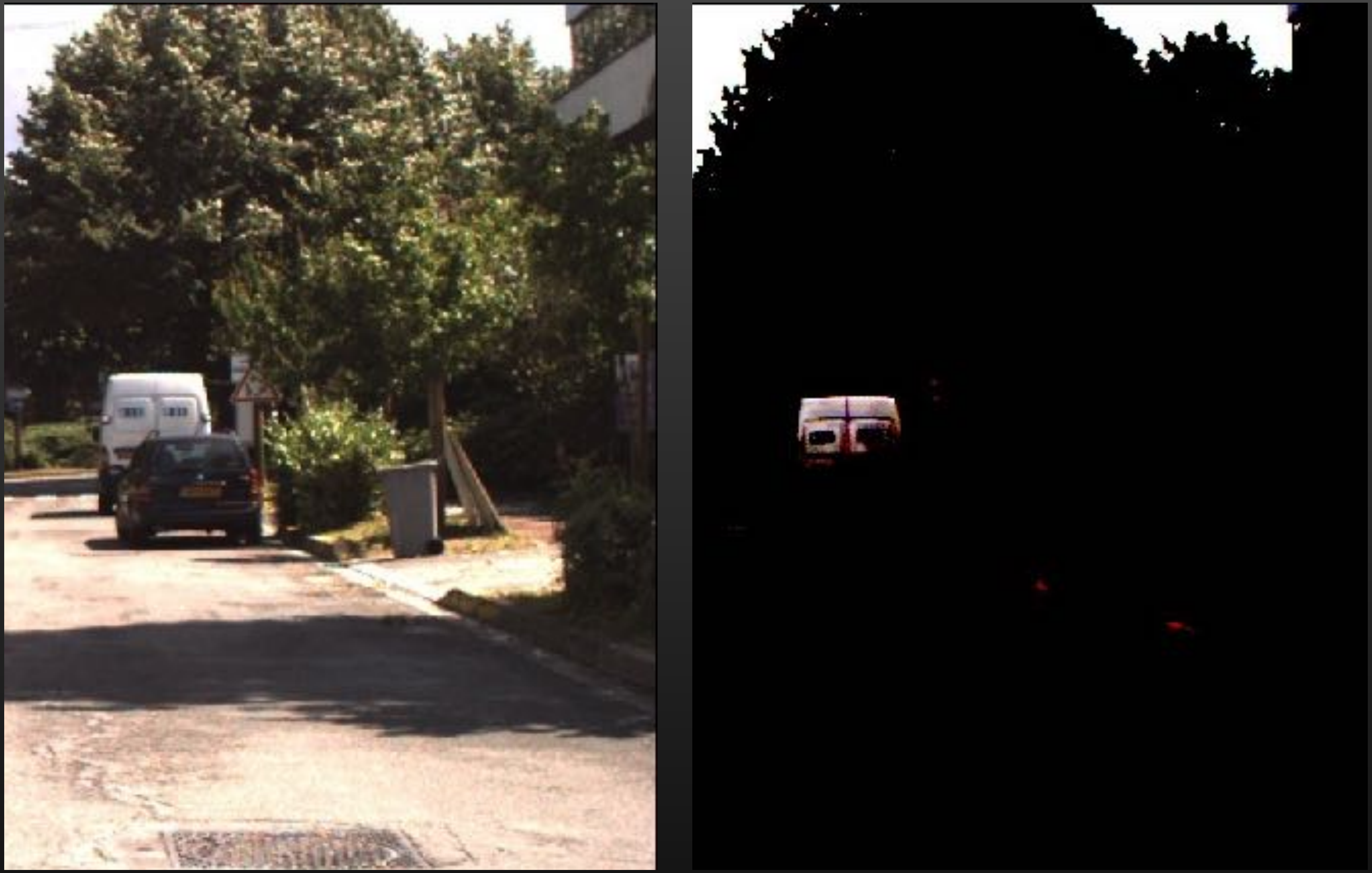


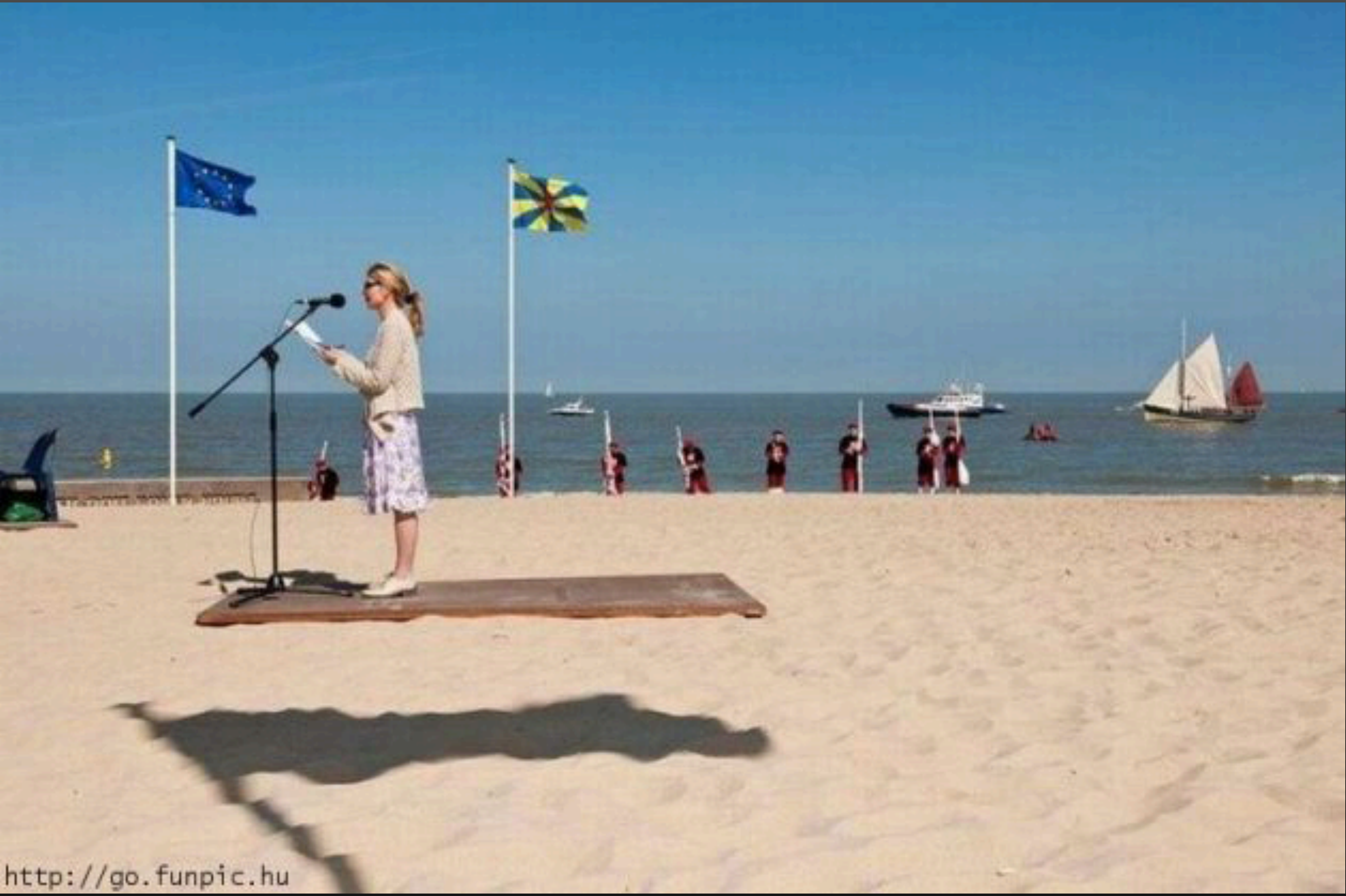
Van Dyck’s triple portrait of Charles I with a copy of Bernini’s bust and an engraving by von Voerst of the bust



PMVS (Furukawa & Ponce, 2007)

What is happening with the shadows?



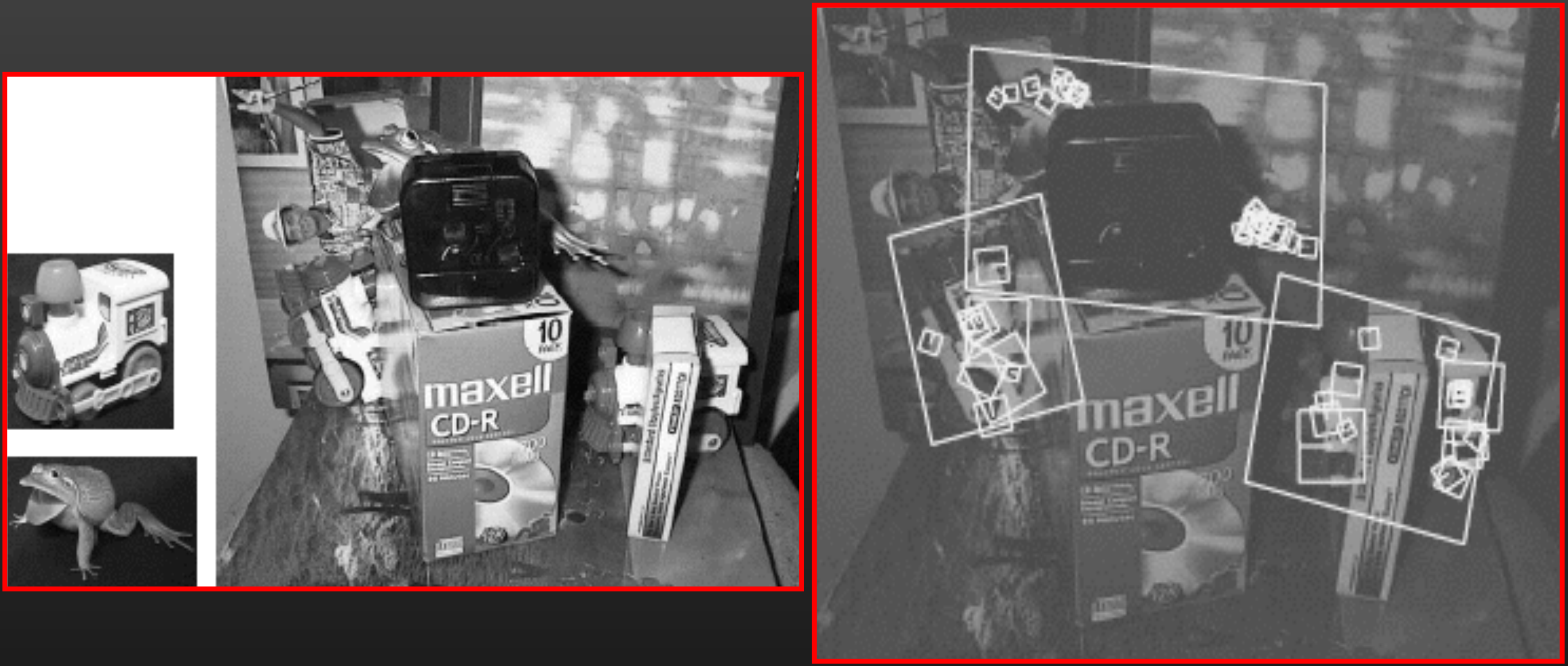


Source: J.J. Koenderink

Outline

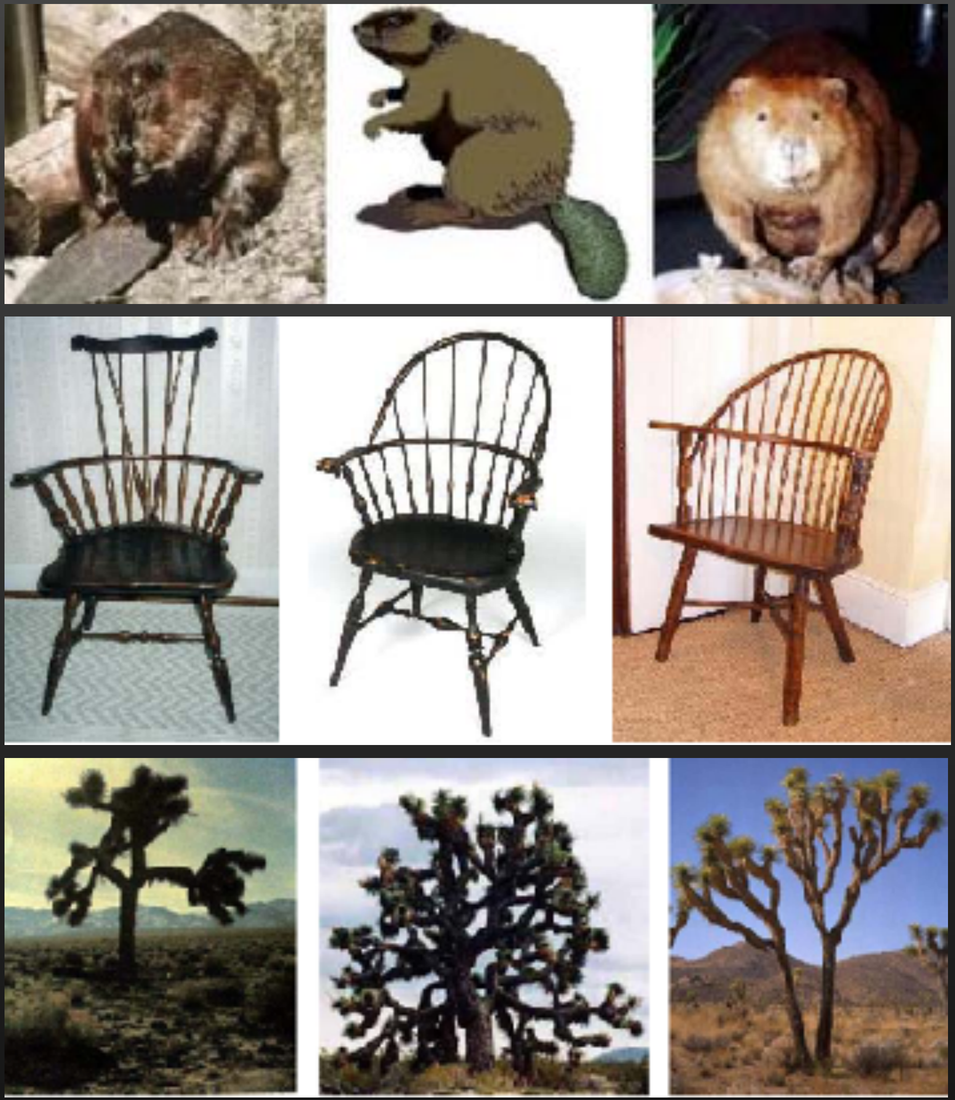
- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry
- Image processing

Specific object detection



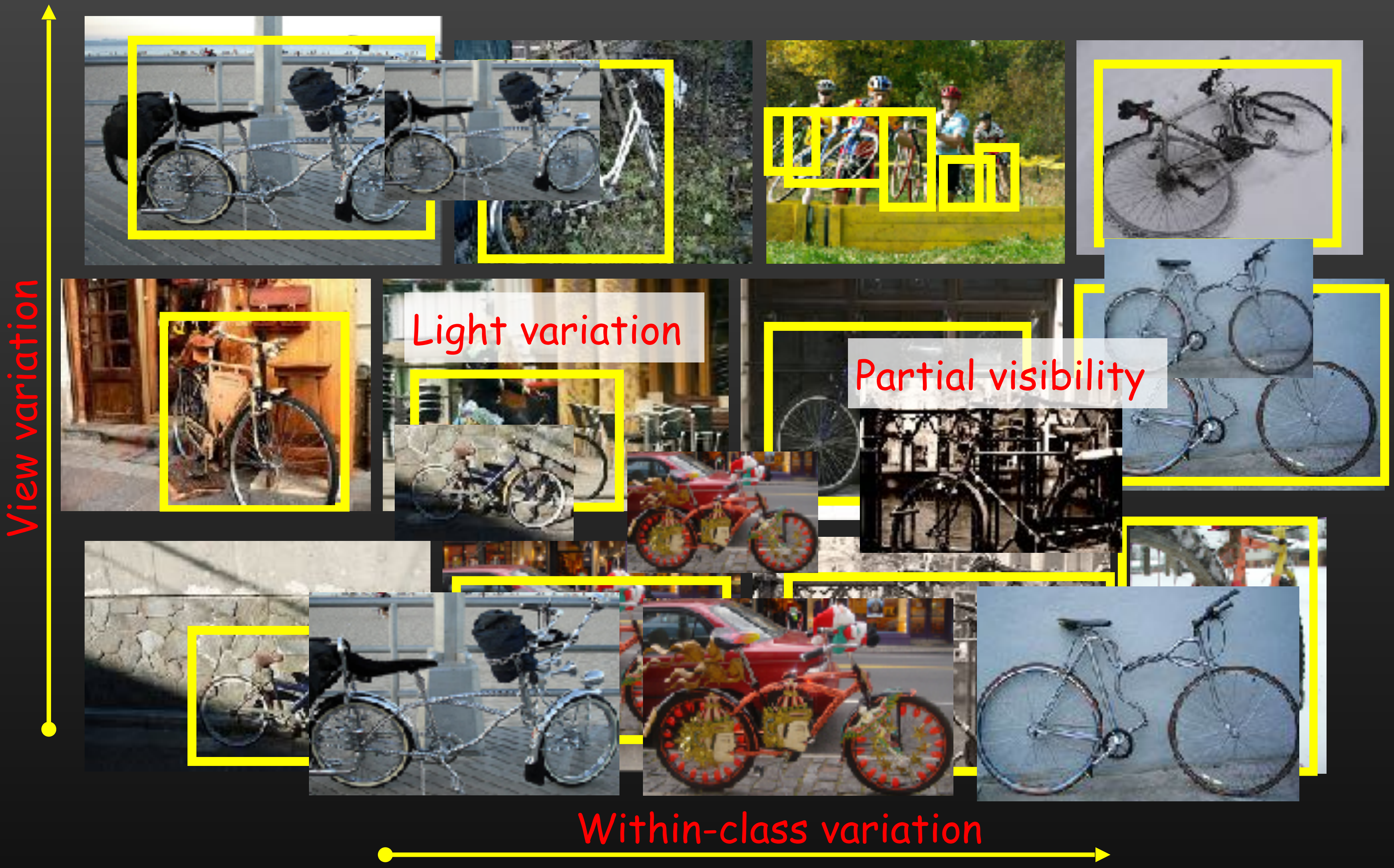
(Lowe, 2004)

Image classification



Caltech 101 : http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Object category detection



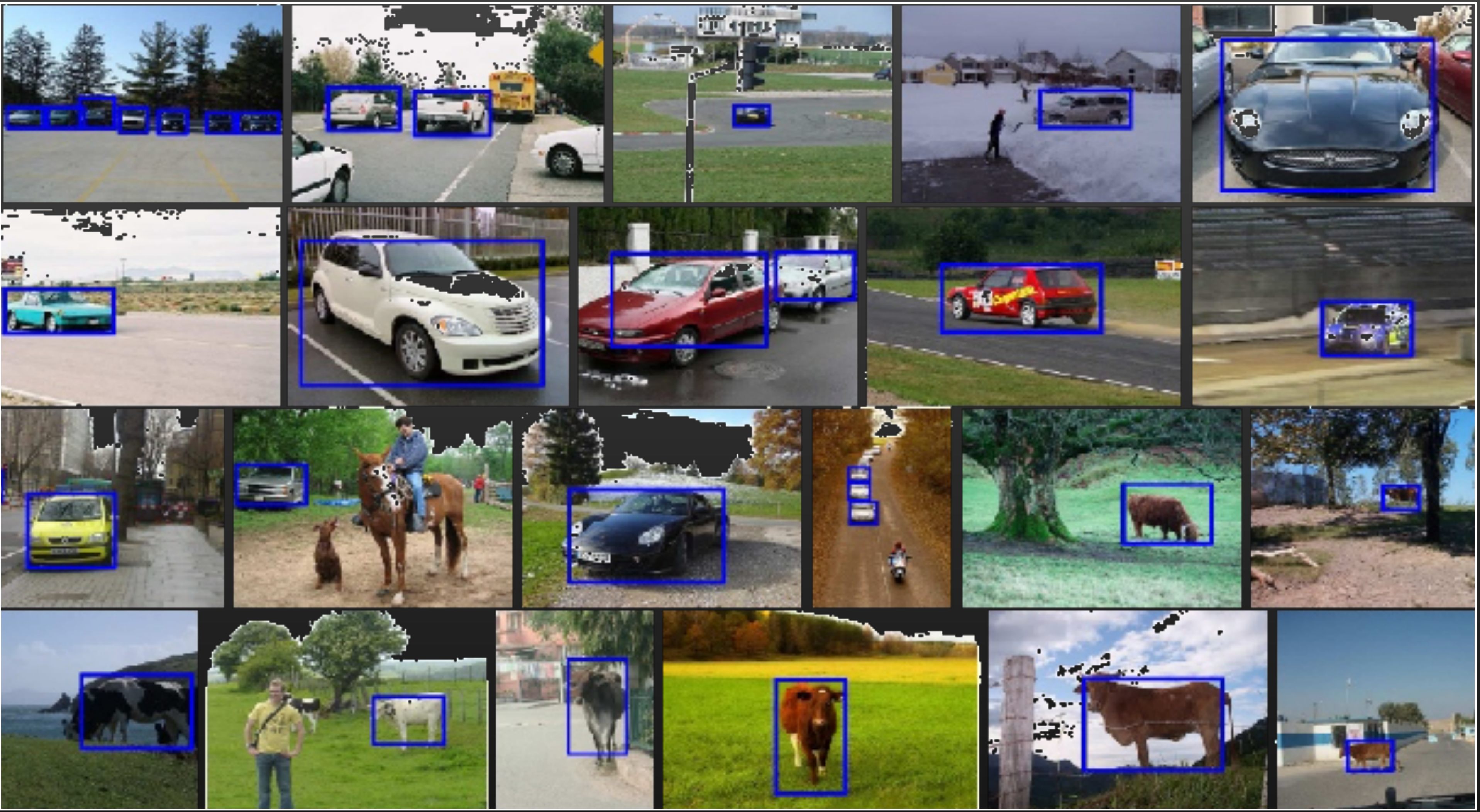
View variation

Light variation

Partial visibility

Within-class variation

Example: part-based models



Qualitative experiments on Pascal VOC'07 (Kushal, Schmid, Ponce, 2008)

Scene understanding

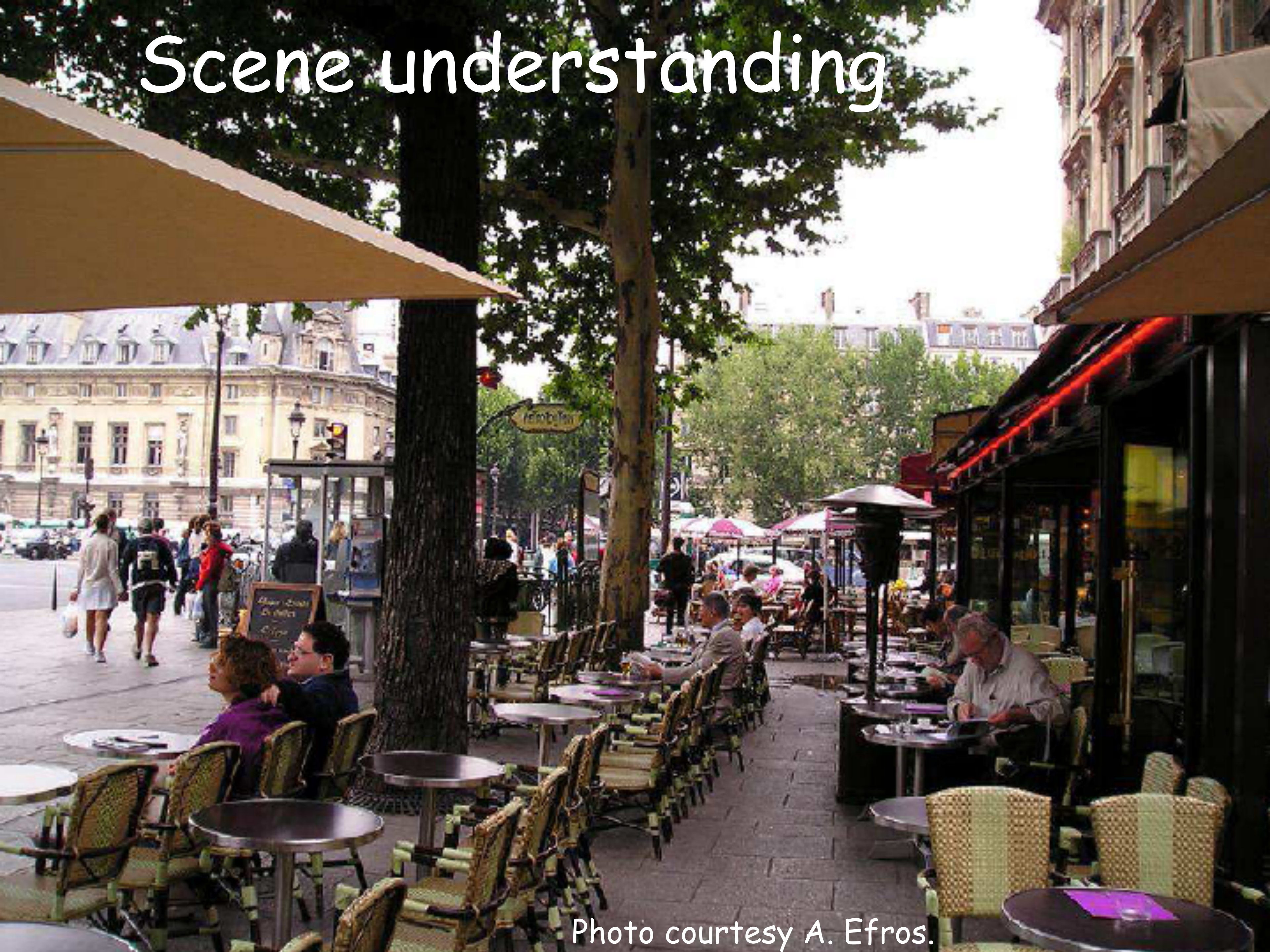


Photo courtesy A. Efros.

Computer vision books

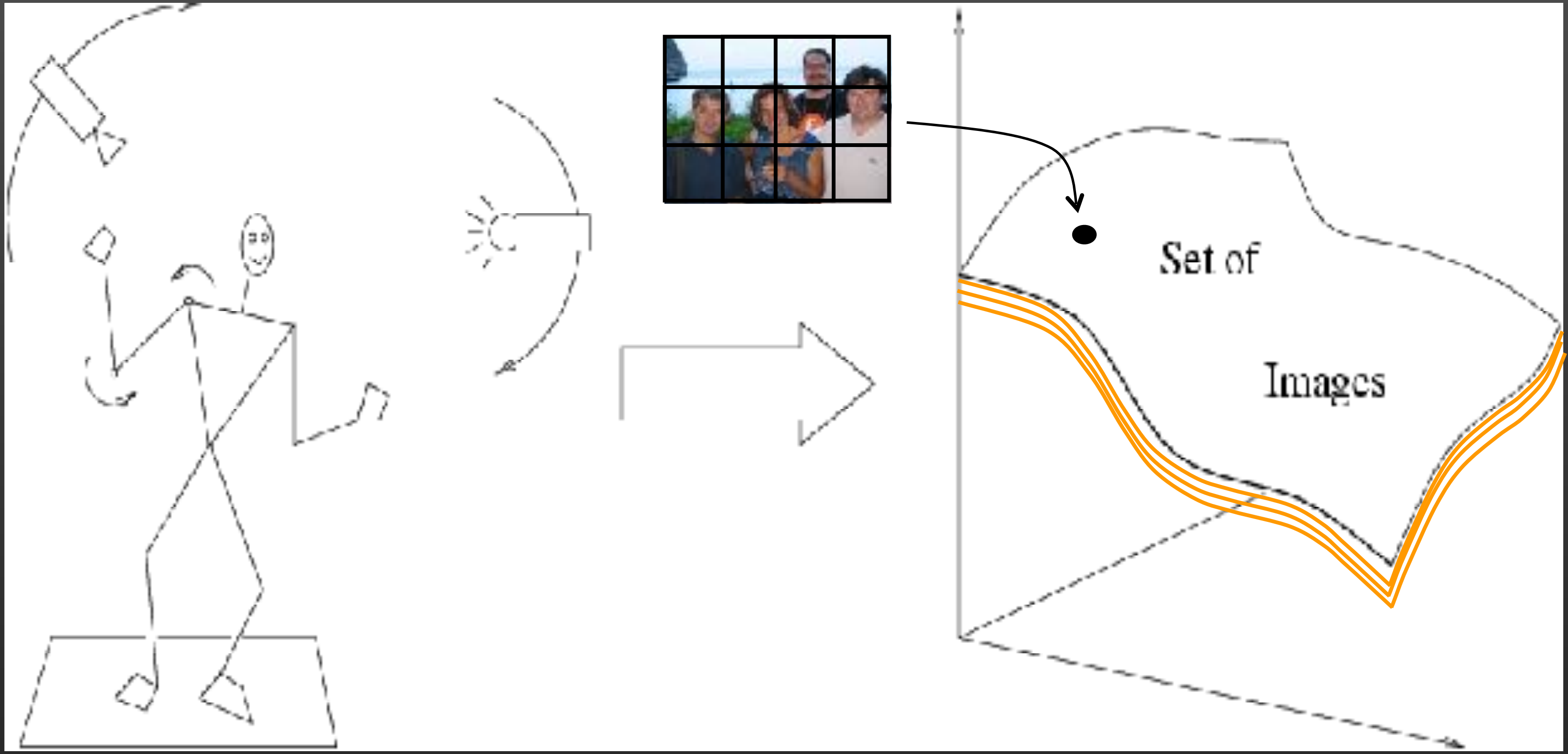
- D.A. Forsyth and J. Ponce, "Computer Vision: A Modern Approach", Prentice-Hall/Pearson, 2nd edition, 2011.

- J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, "Toward category-level object recognition", Springer LNCS, 2007.
- R. Szeliski, "Computer Vision: Algorithms and Applications", Springer, 2010.
- O. Faugeras, Q.T. Luong, and T. Papadopoulo, "Geometry of Multiple Images," MIT Press, 2001.
- R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision", Cambridge University Press, 2004.

- J.J. Koenderink, "Solid Shape", MIT Press, 1990, and http://

Outline

- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry
- Image processing

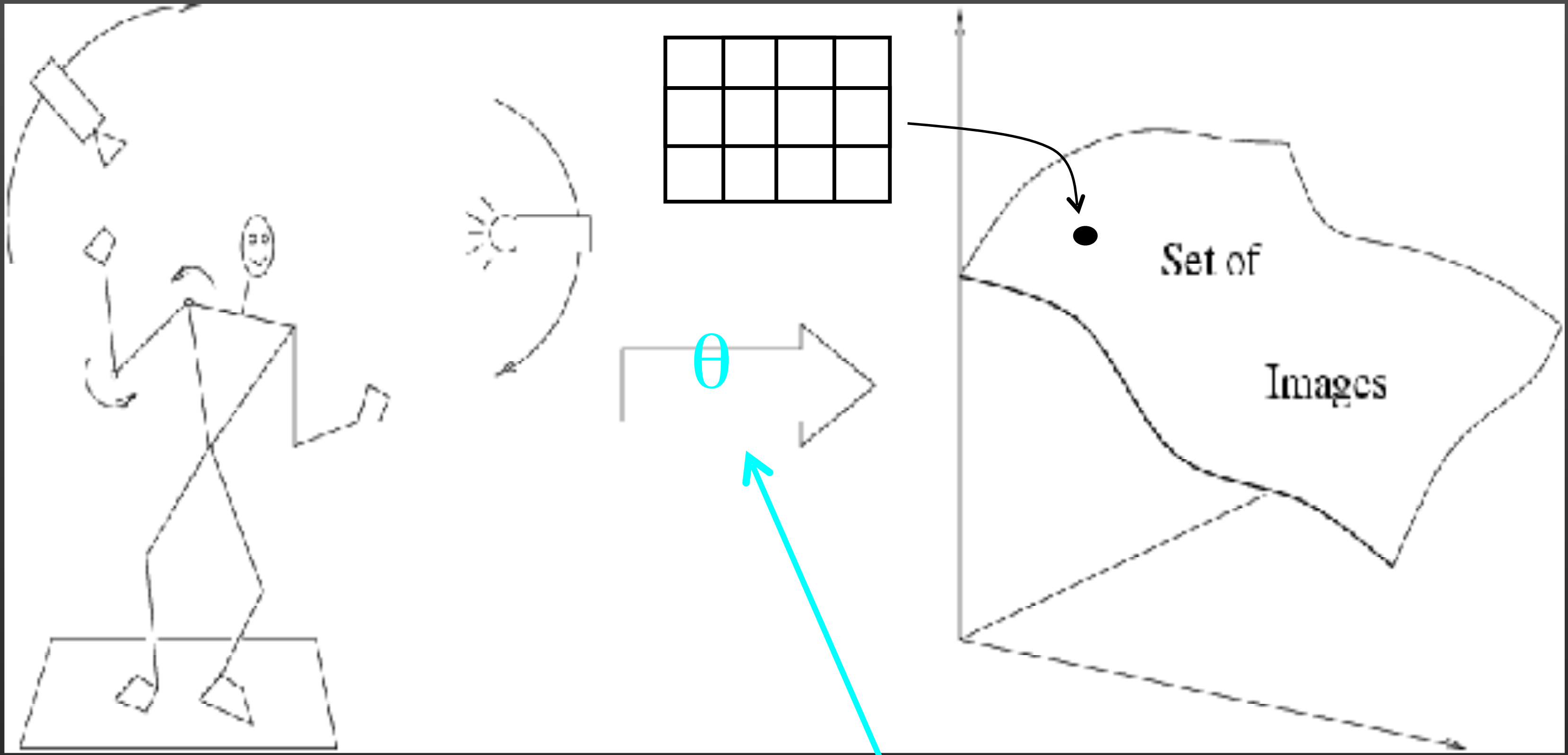


Variability: Camera position

Illumination

Internal parameters

Within-class variations

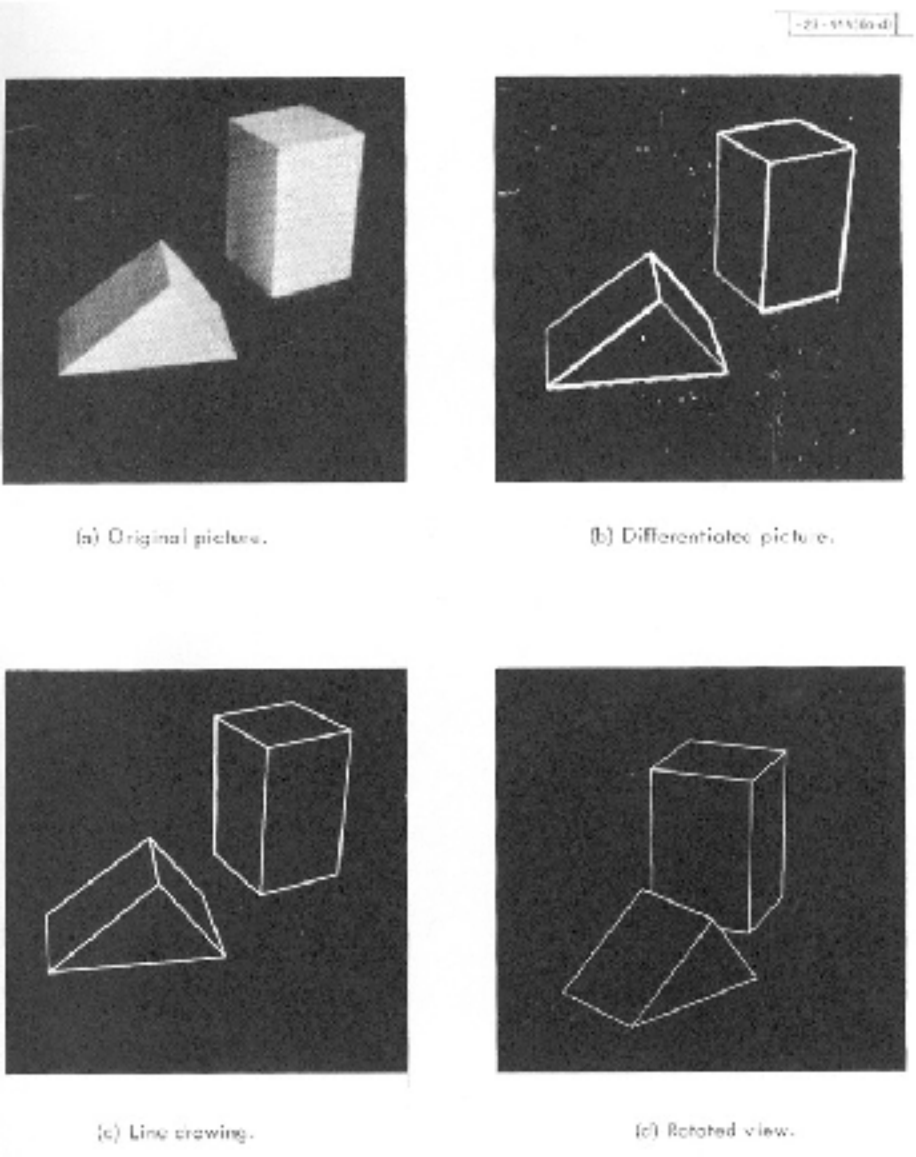


Variability:

- Camera position
- Illumination
- Internal parameters

Roberts (1963); Lowe (1987); Faugeras & Hebert (1986); Grimson & Lozano-Perez (1986); Huttenlocher & Ullman (1987)

Origins of computer vision



(a) Original picture.

(b) Differentiated picture.

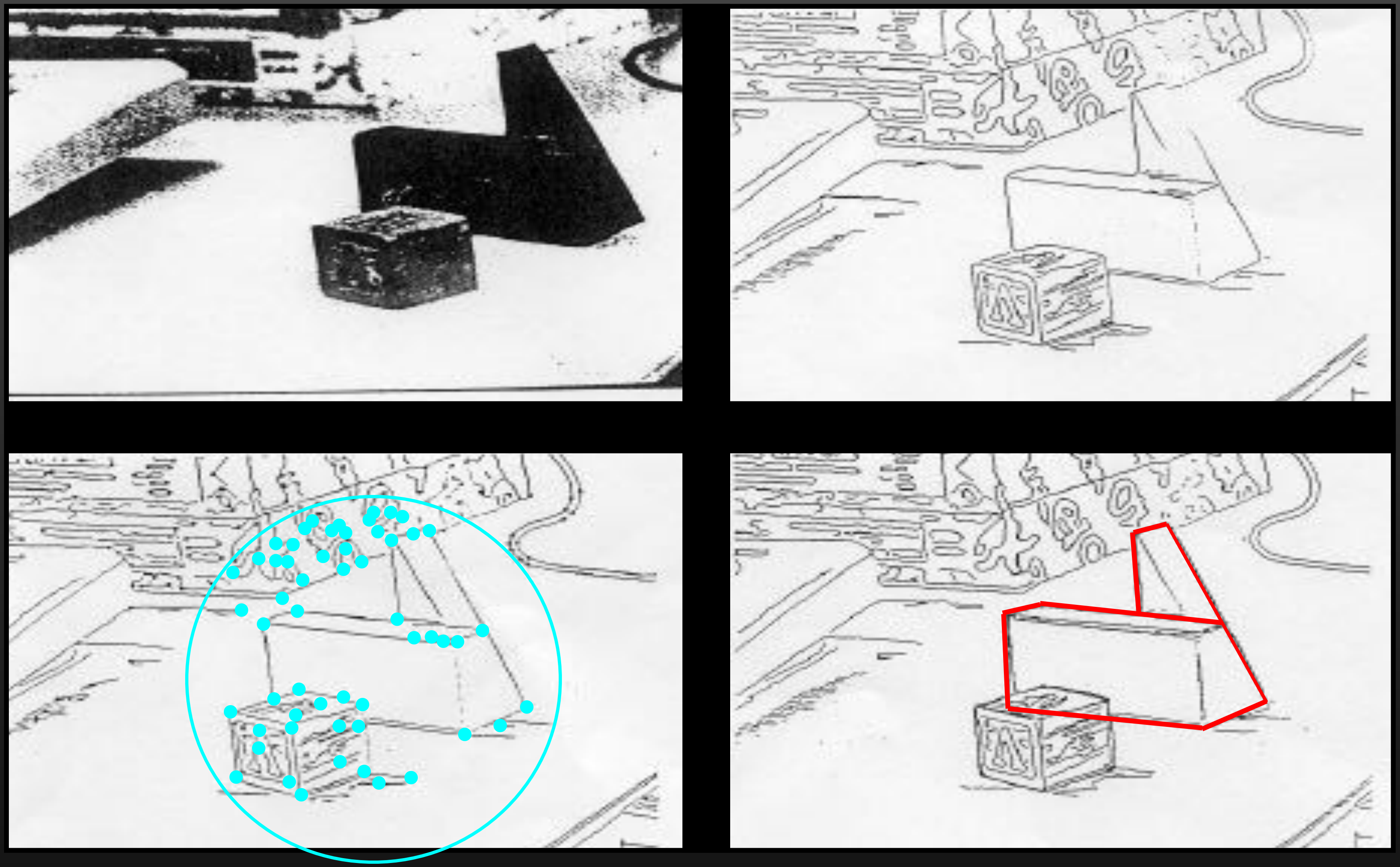
(c) Line drawing.

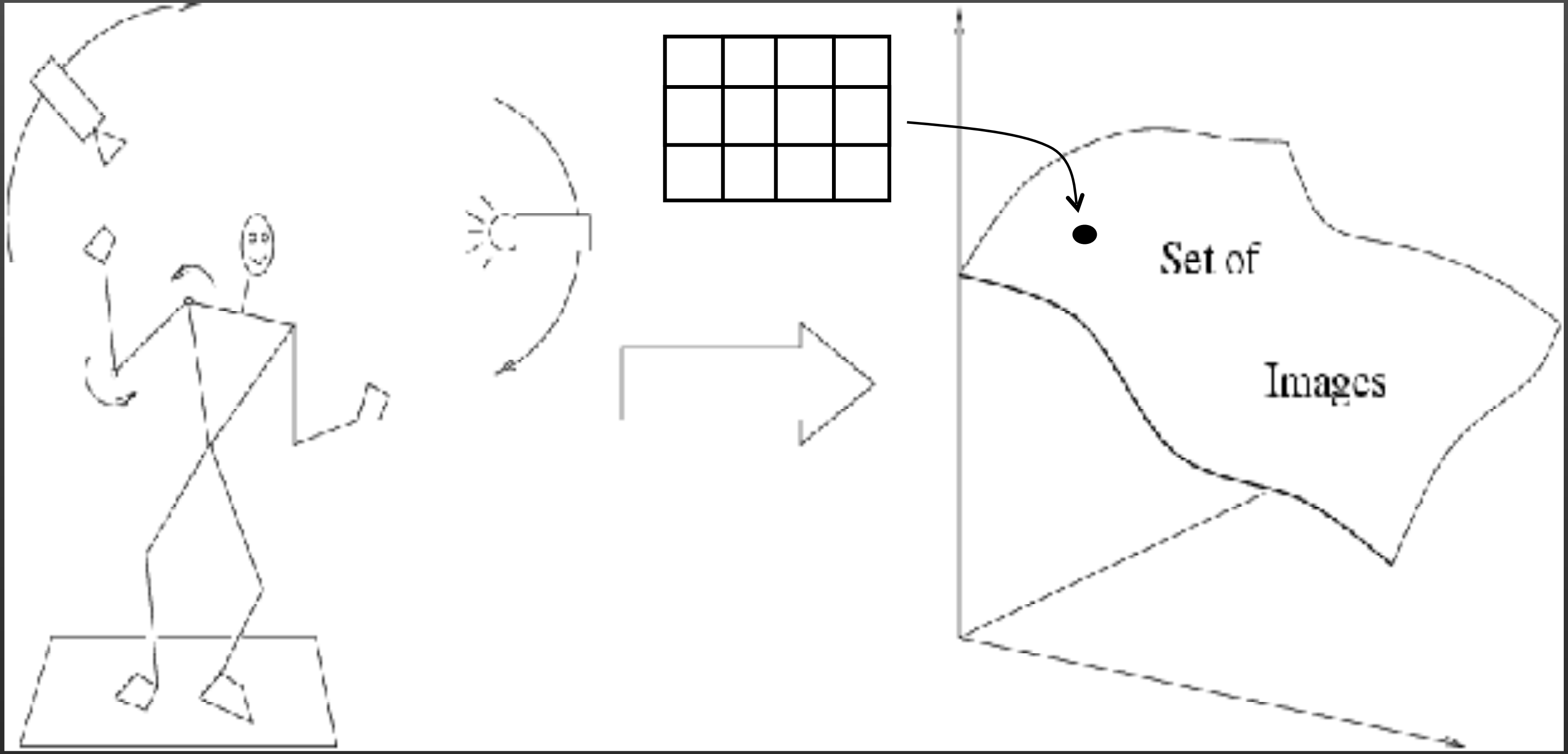
(d) Rotated view.

L. G. Roberts, *Machine Perception of Three Dimensional Solids,* Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

photo credit: Joe Mundy

Huttenlocher & Ullman (1987)





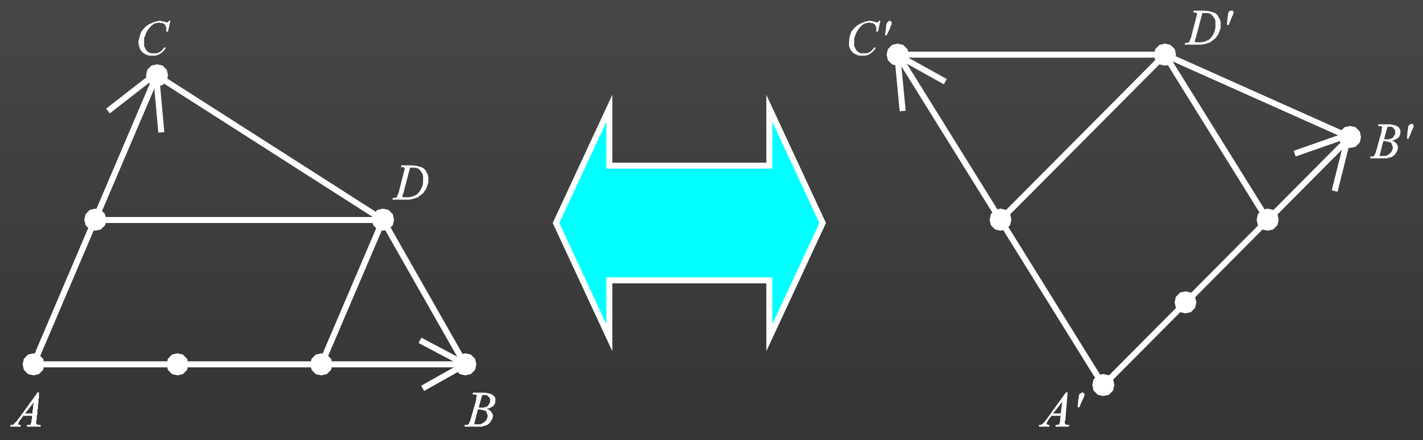
~~Variability~~

Invariance to:

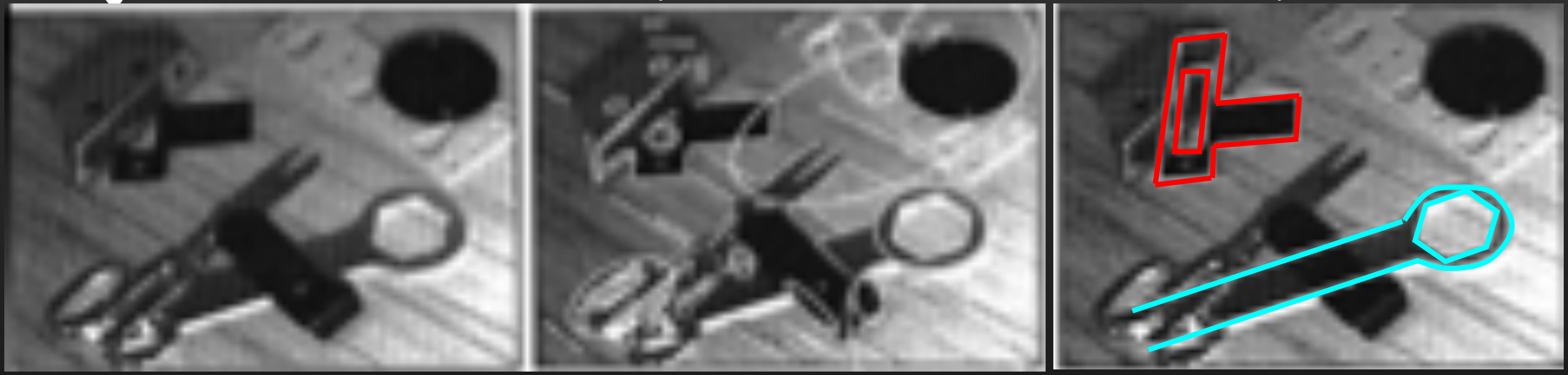
- Camera position
- Illumination
- Internal parameters

Duda & Hart (1972); Weiss (1987); Mundy et al. (1992-94); Rothwell et al. (1992); Burns et al. (1993)

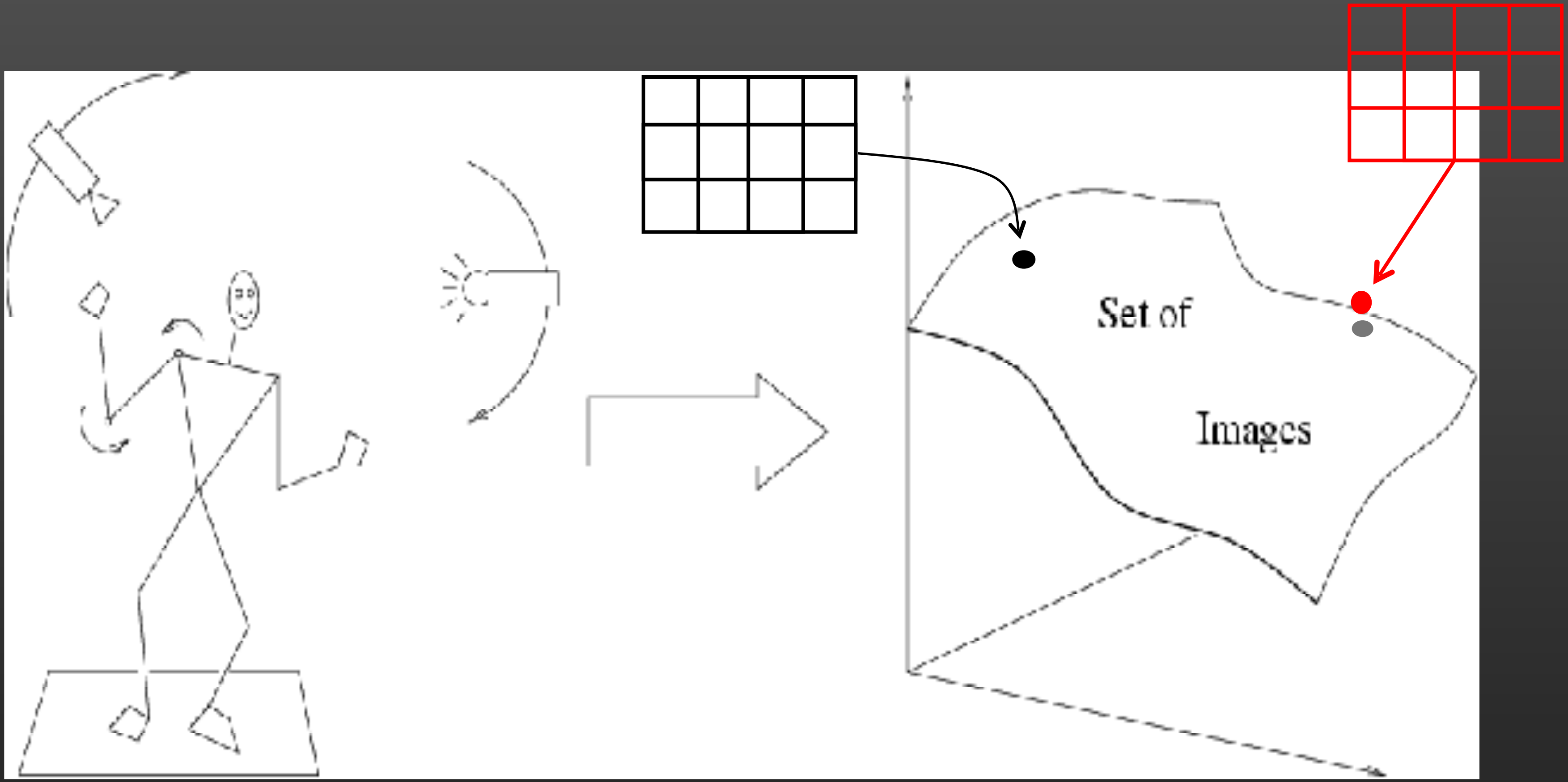
Example: affine invariants of coplanar points



Projective invariants (Rothwell et al., 1992):



BUT: True 3D objects do not admit monocular viewpoint invariants (Burns et al., 1993) !!

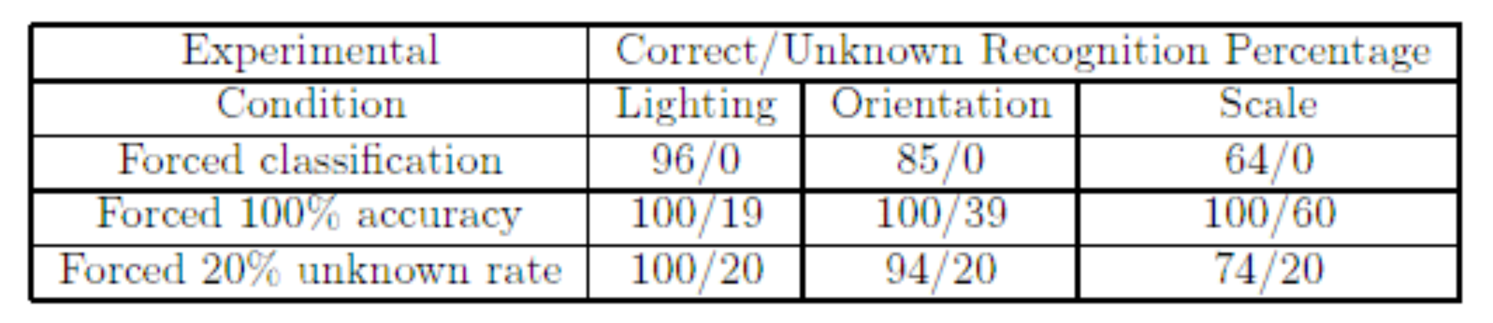


Empirical models of image variability:

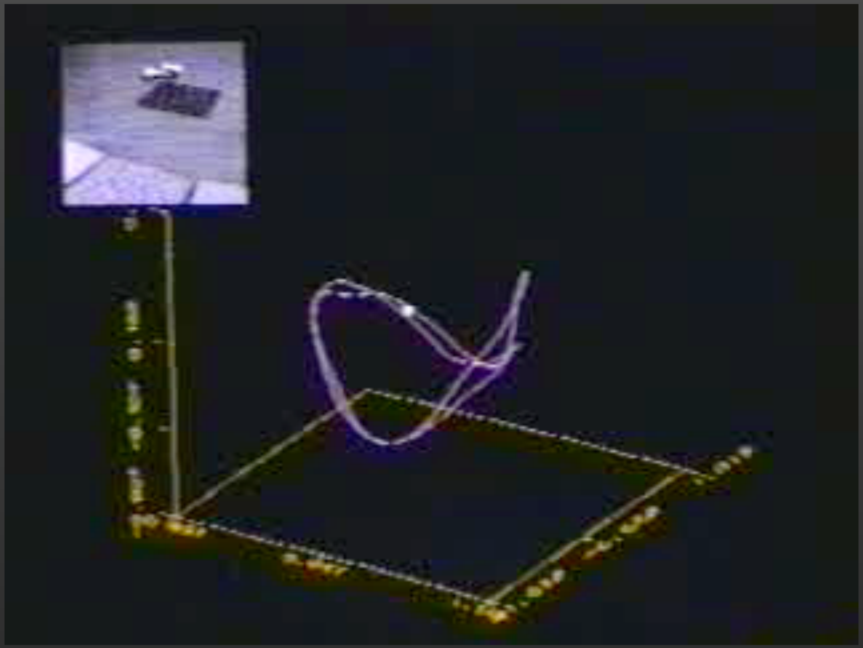
Appearance-based techniques

Turk & Pentland (1991); Murase & Nayar (1995); etc.

Eigenfaces (Turk & Pentland, 1991)



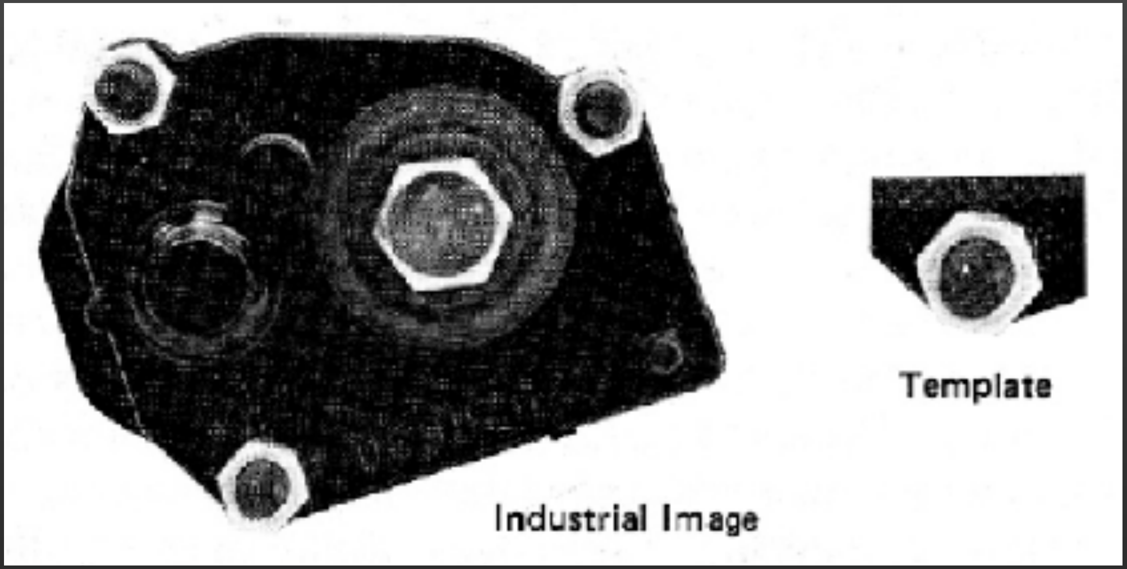
Experimental	Correct/Unknown Recognition Percentage		
Condition	Lighting	Orientation	Scale
Forced classification	96/0	85/0	64/0
Forced 100% accuracy	100/19	100/39	100/60
Forced 20% unknown rate	100/20	94/20	74/20



Appearance manifolds

(Murase & Nayar, 1995)

Correlation-based template matching (60s)



Ballard & Brown (1980, Fig. 3.3). Courtesy Bob Fisher and Ballard & Brown on-line.

- Automated target recognition
- Industrial inspection
- Optical character recognition
- Stereo matching
- Pattern recognition

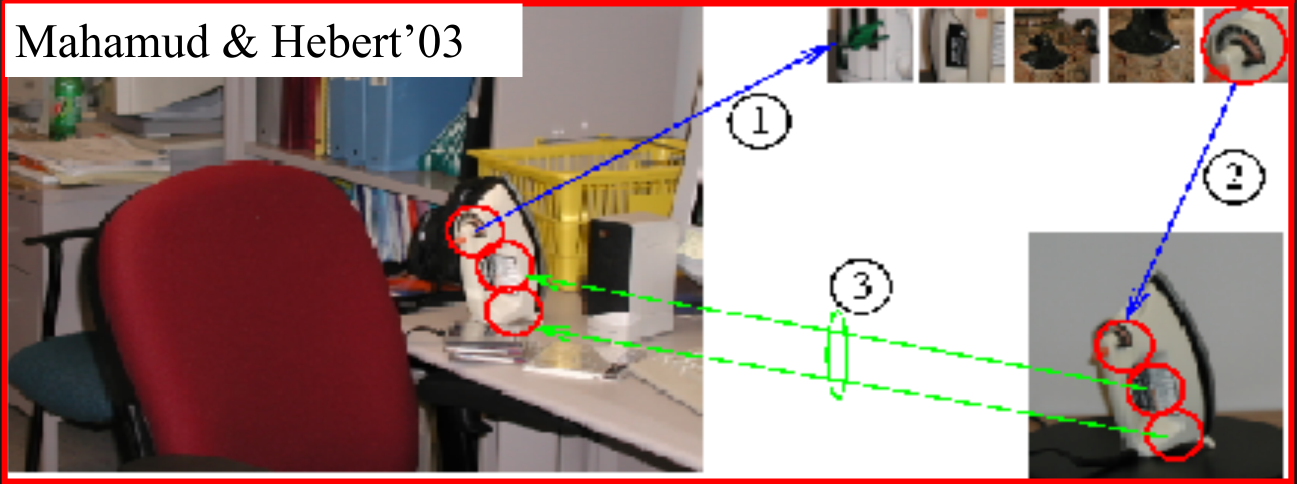
In the late 1990s, a new approach emerges:
Combining local appearance, spatial constraints, invariants, and classification techniques from machine learning.

Query

Retrieved (10° off)

Schmid & Mohr’97

Lowe’02


Mahamud & Hebert’03

Late 1990s: Local appearance models

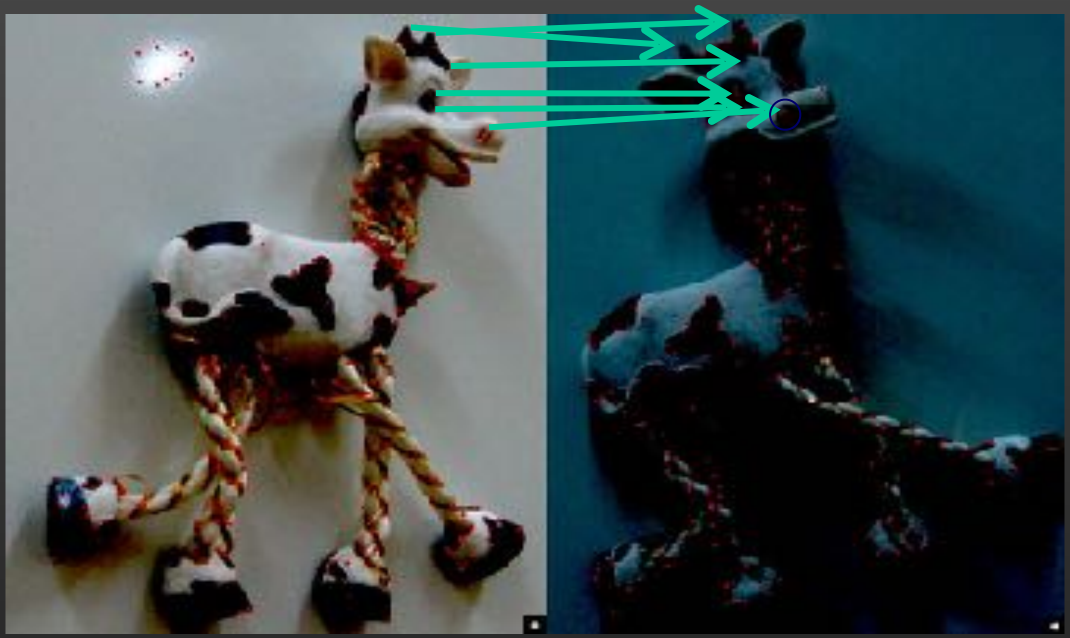
(Image courtesy of C. Schmid)

Late 1990s: Local appearance models

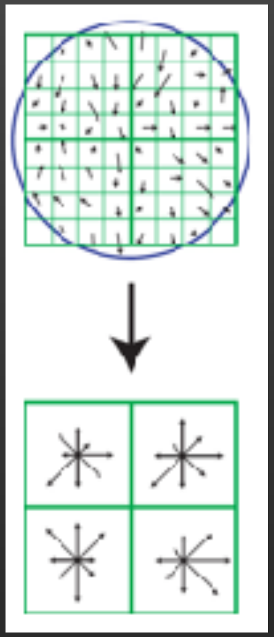
(Image courtesy of C. Schmid)

- Find features (interest points).

Late 1990s: Local appearance models



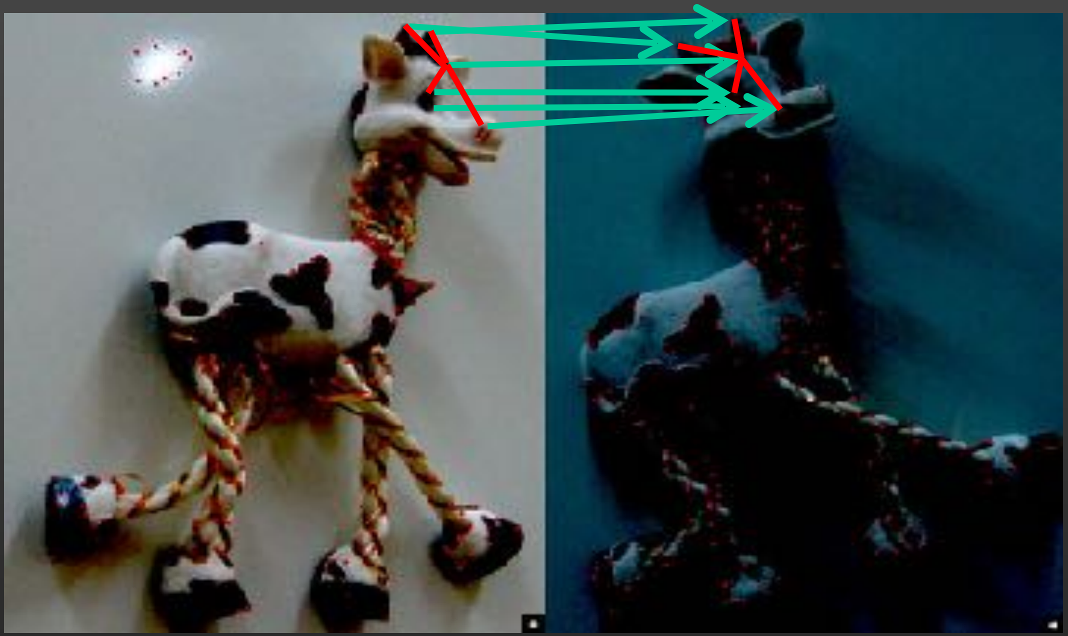
(Image courtesy of C. Schmid)



(Lowe 2004)

- Find features (interest points).
- Match them using local invariant descriptors (jets, SIFT).

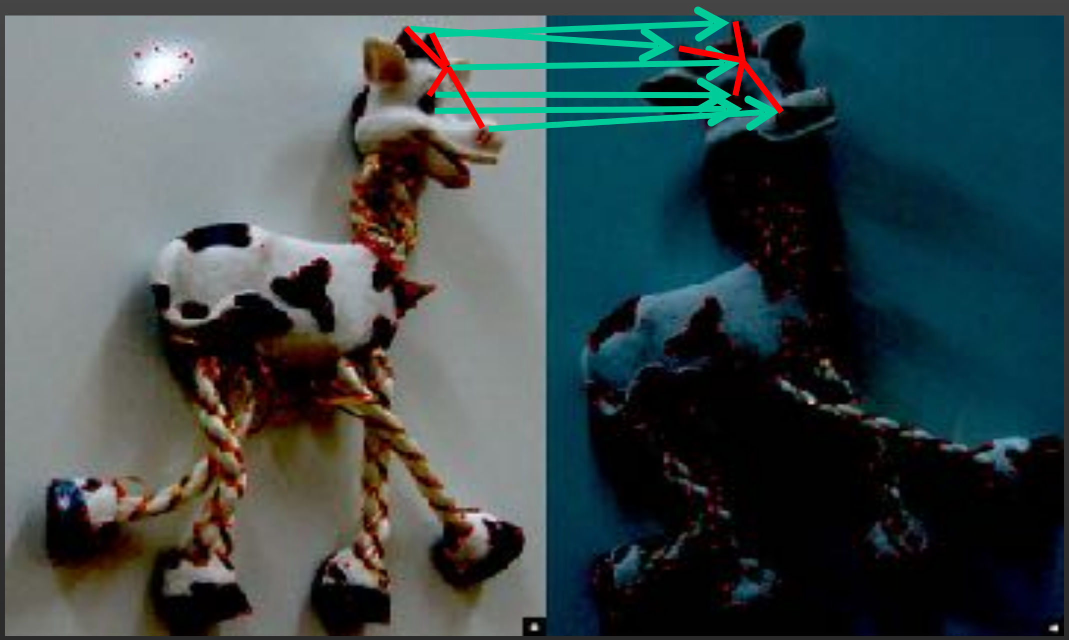
Late 1990s: Local appearance models



(Image courtesy of C. Schmid)

- Find features (interest points).
- Match them using local invariant descriptors (jets, SIFT).
- Optional: Filter out outliers using geometric consistency.

Late 1990s: Local appearance models



(Image courtesy of C. Schmid)

- Find features (interest points).
- Match them using local invariant descriptors (jets, SIFT).
- Optional: Filter out outliers using geometric consistency.
- Vote.

See, for example, Schmid & Mohr (1996); Lowe (1999);Tuytelaars & Van Gool, (2002); Rothganger et al. (2003); Ferrari et al., (2004).

Bags of words: Visual "Google"

(Sivic & Zisserman, ICCV'03)

"Visual word" clusters

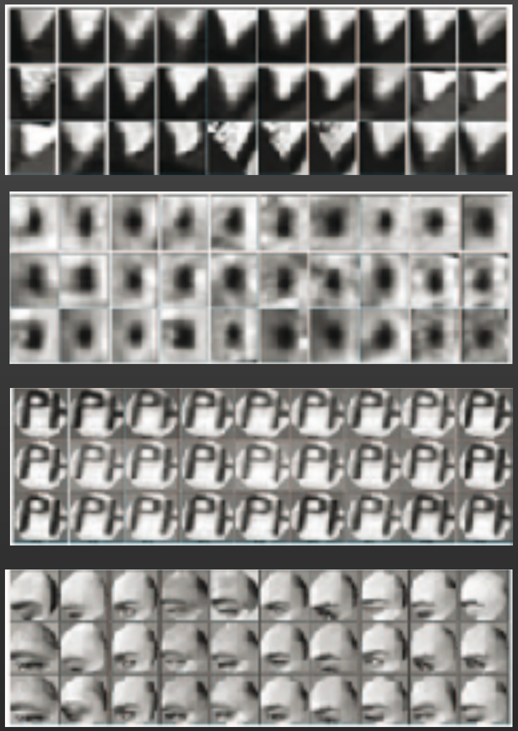
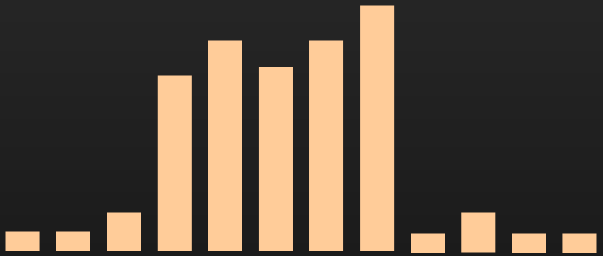


Image retrieval in videos

Select a region and click on Submit to search for an object:

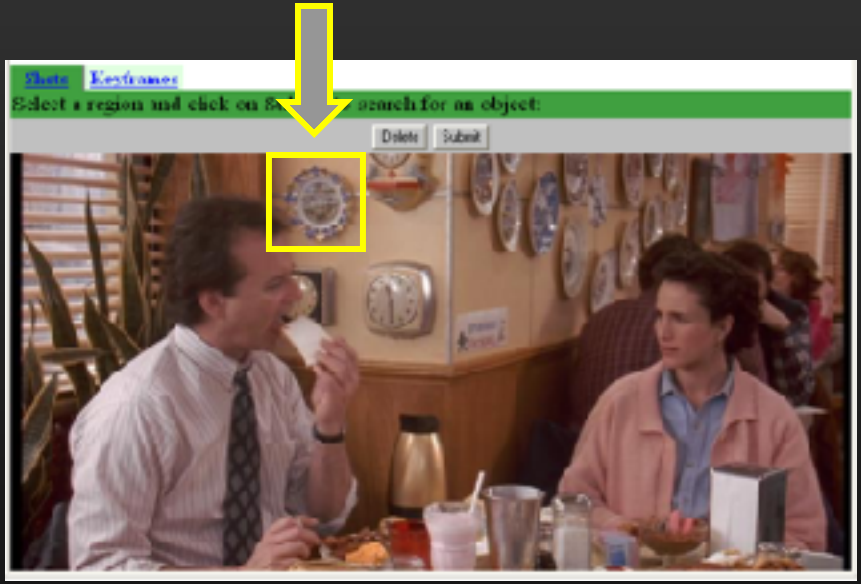


Vector quantization into histogram (the "bag of words")

Bags of words: Visual "Google"

(Sivic & Zisserman, ICCV'03)

Select a region



Select a region and click on S search for an object:

Retrieved shots

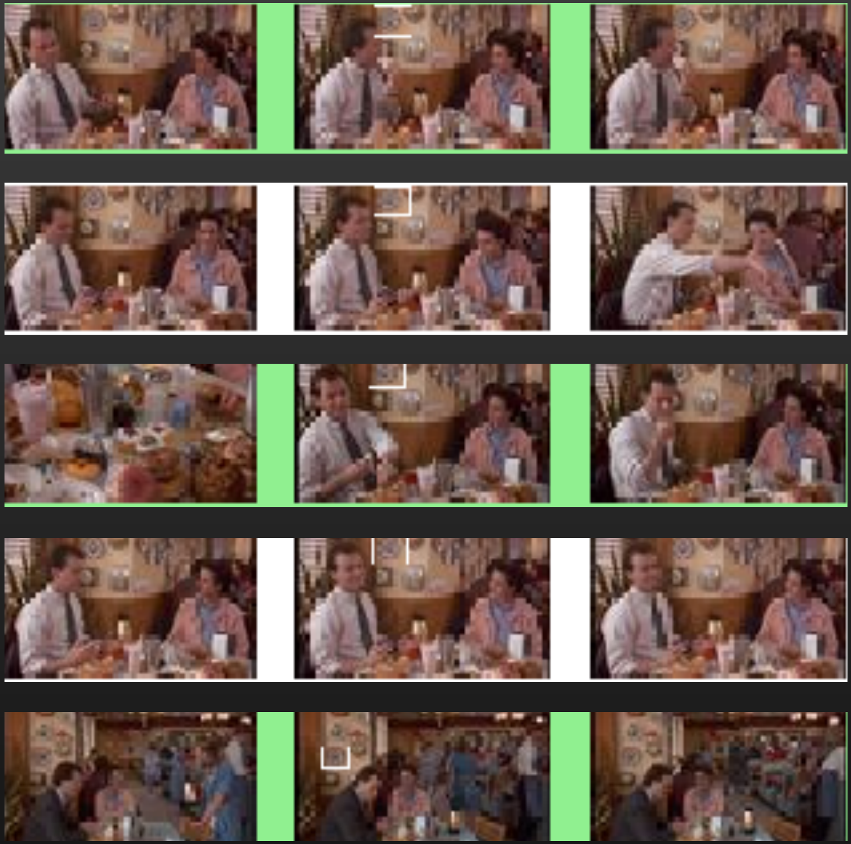
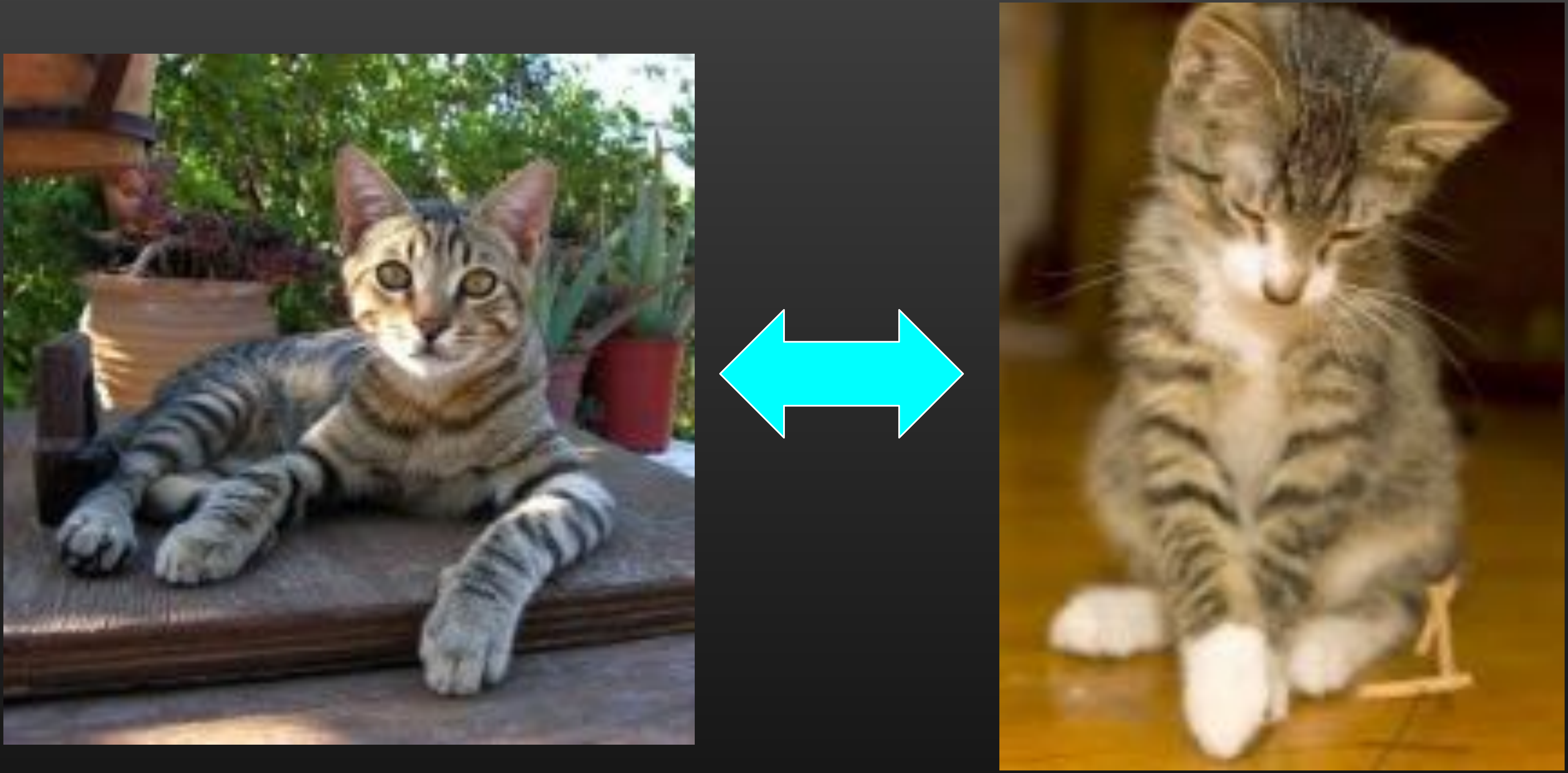
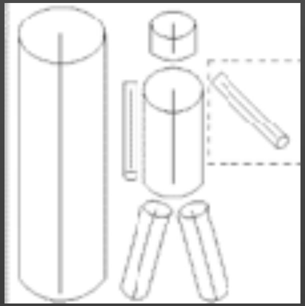


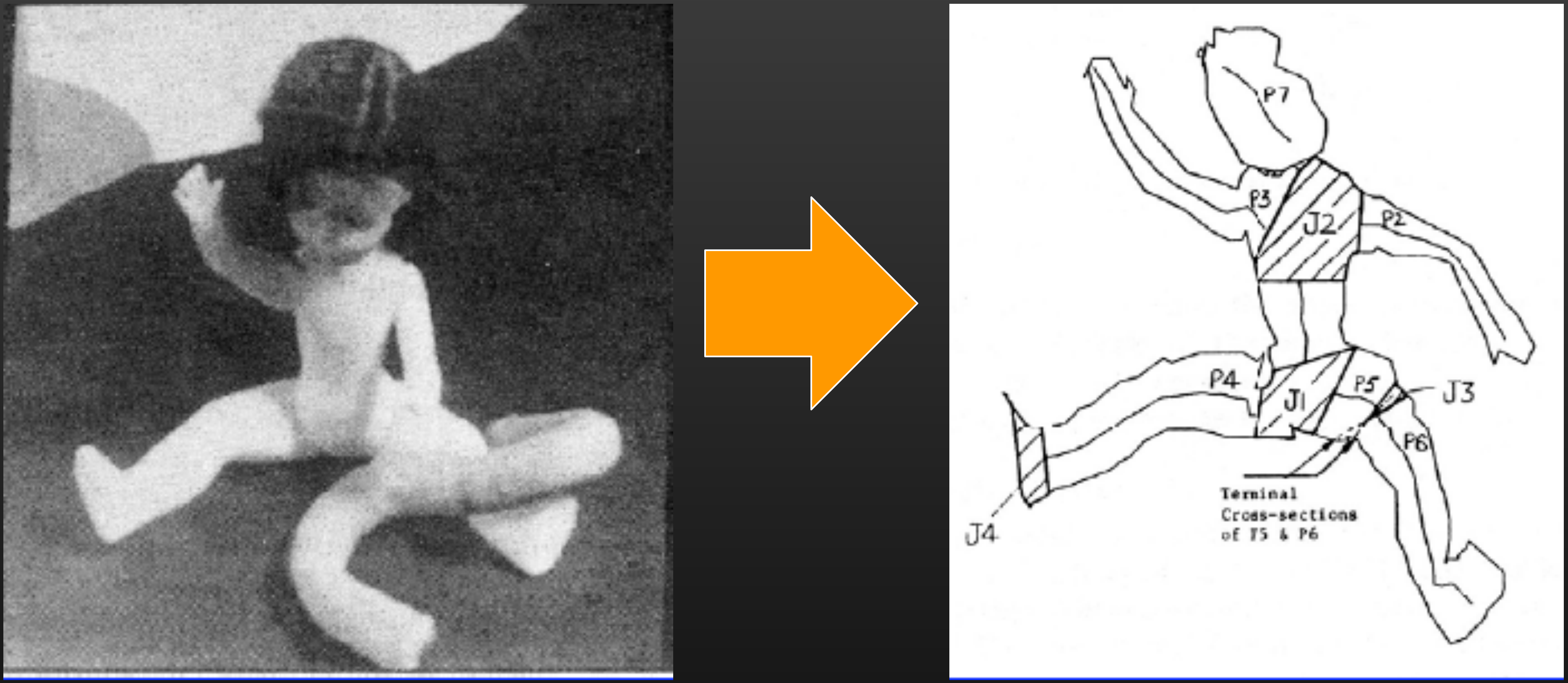
Image categorization is harder





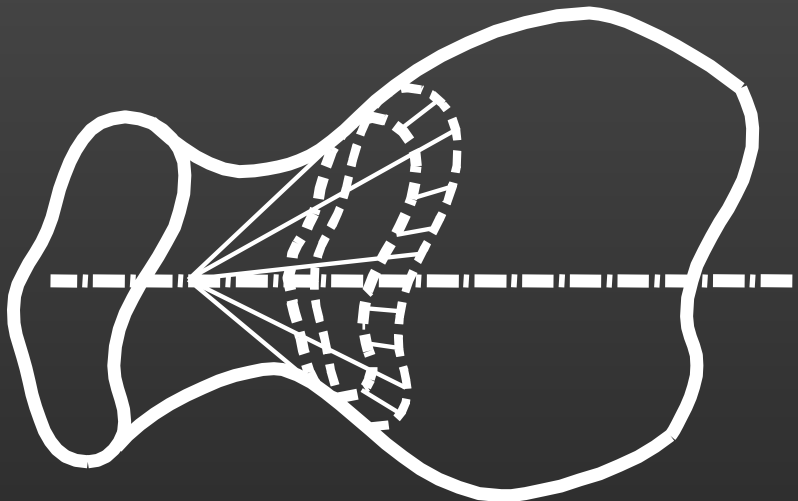
Structural part-based models

(Binford, 1971; Marr & Nishihara, 1978)

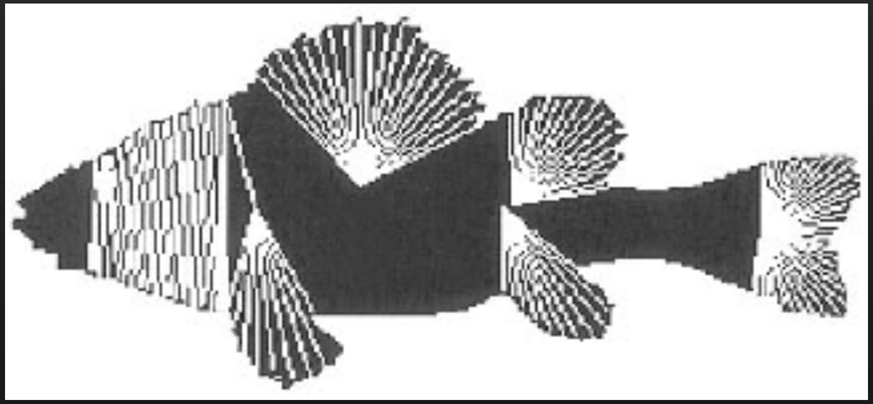


(Nevatia & Binford, 1972)

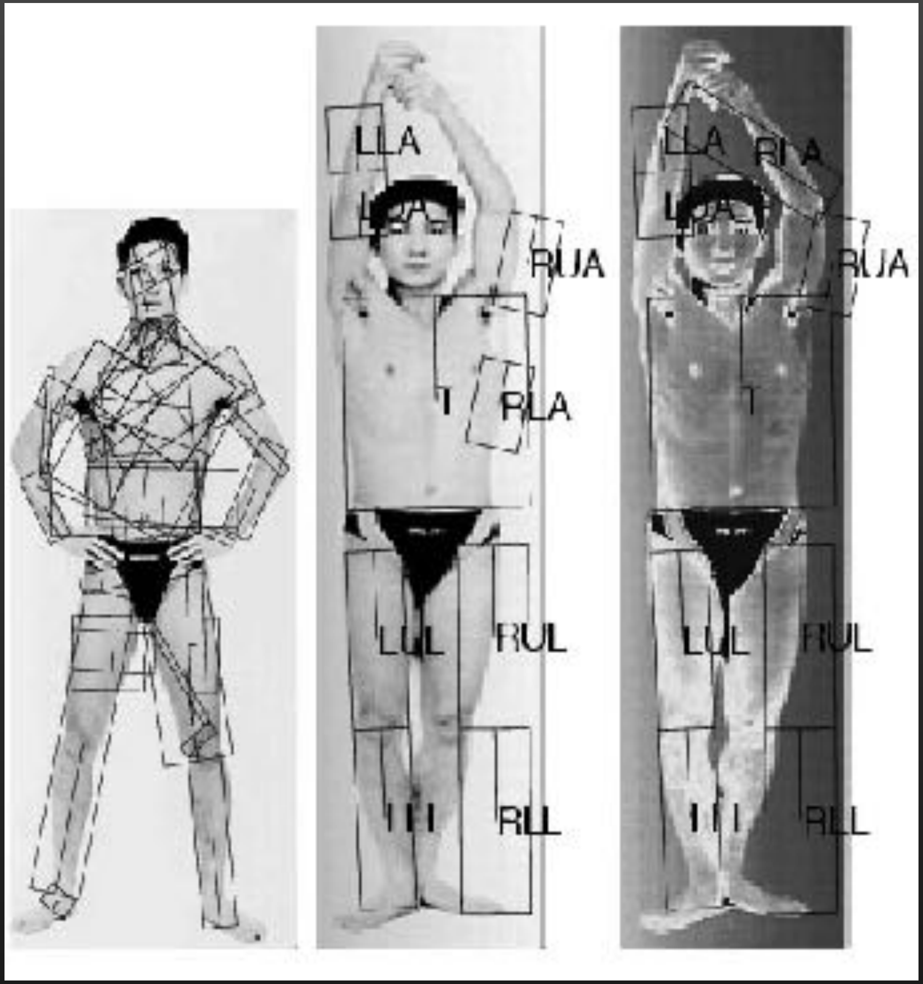
Helas, this is hard to operationalize



Ponce et al. (1989)

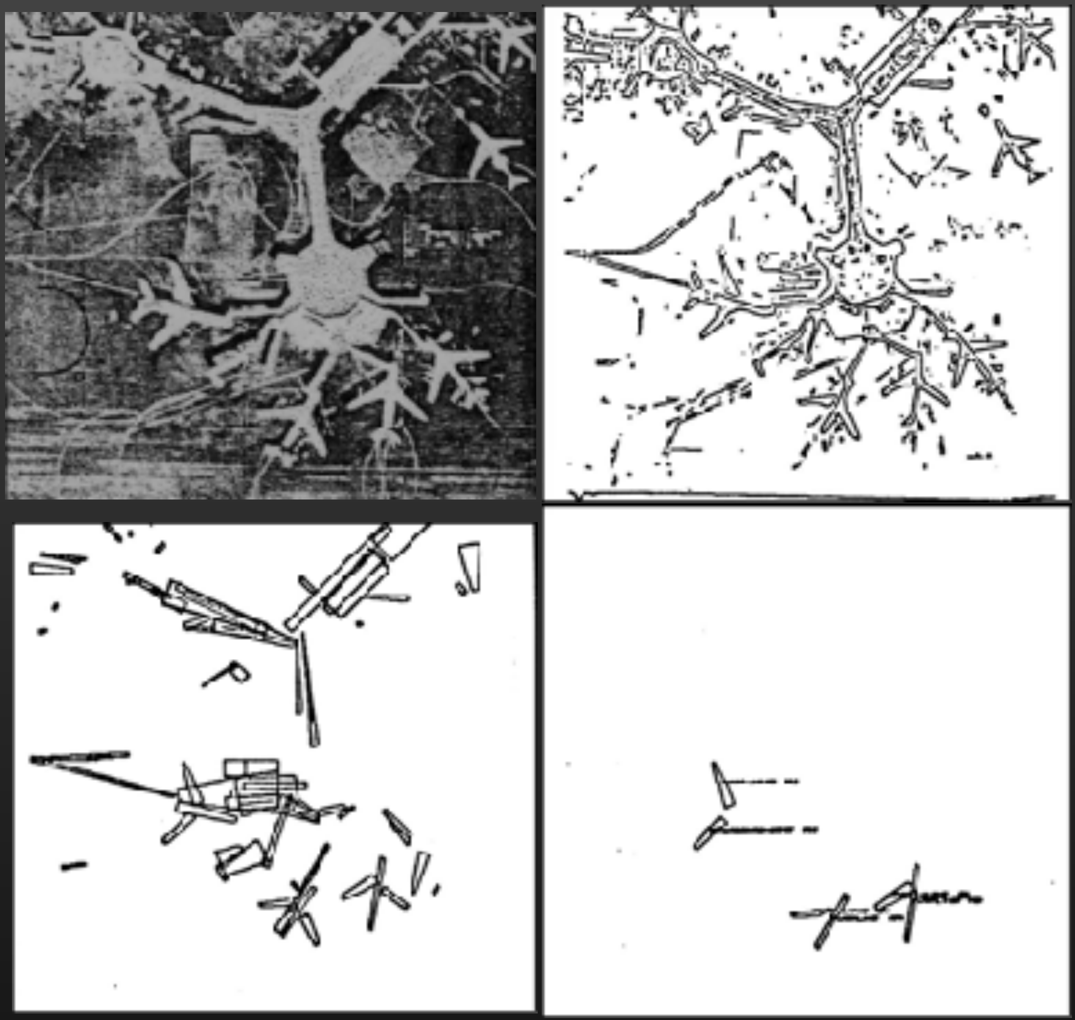


Zhu and Yuille (1996)



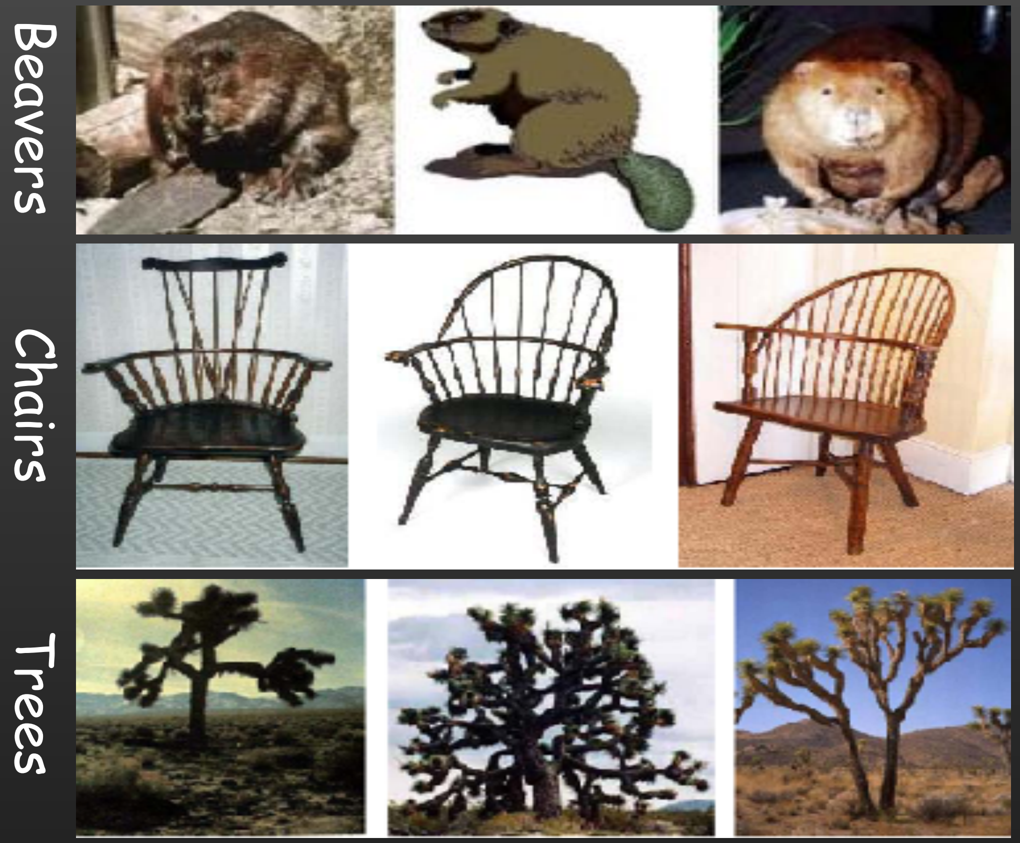
Ioffe and Forsyth (2000)

Ultimate GCs: ACRONYM



(Brooks & Binford, 1981)

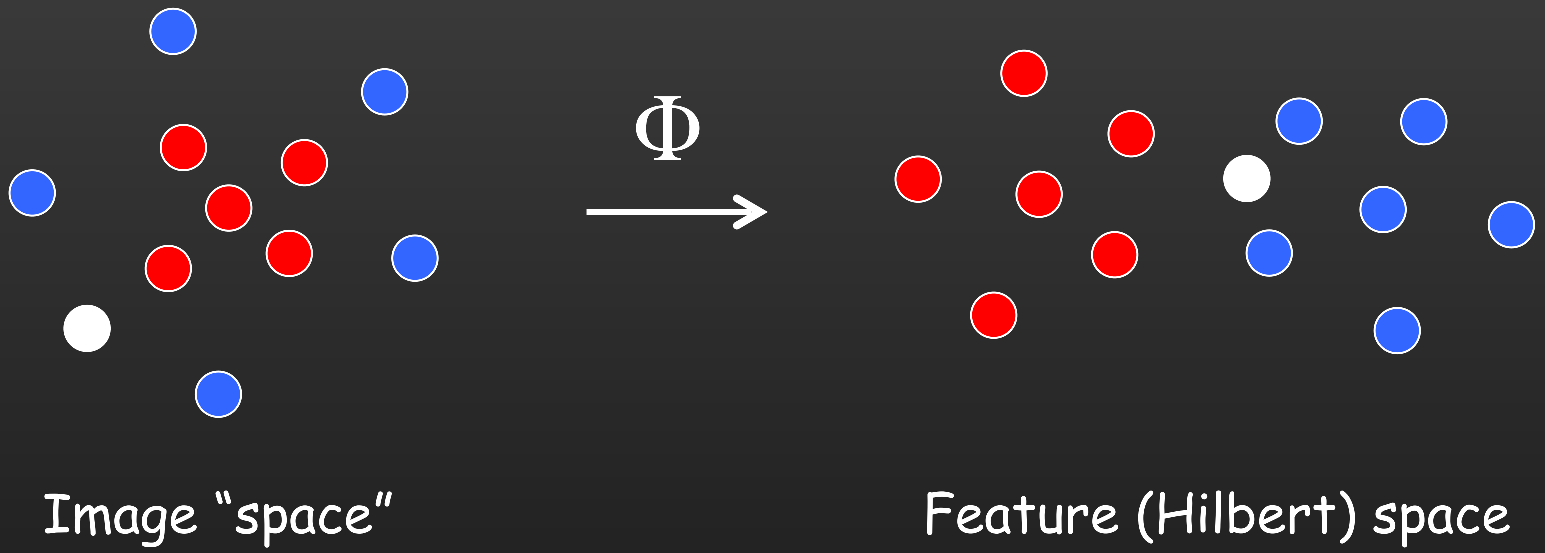
Categorization as supervised classification



Labelled training examples

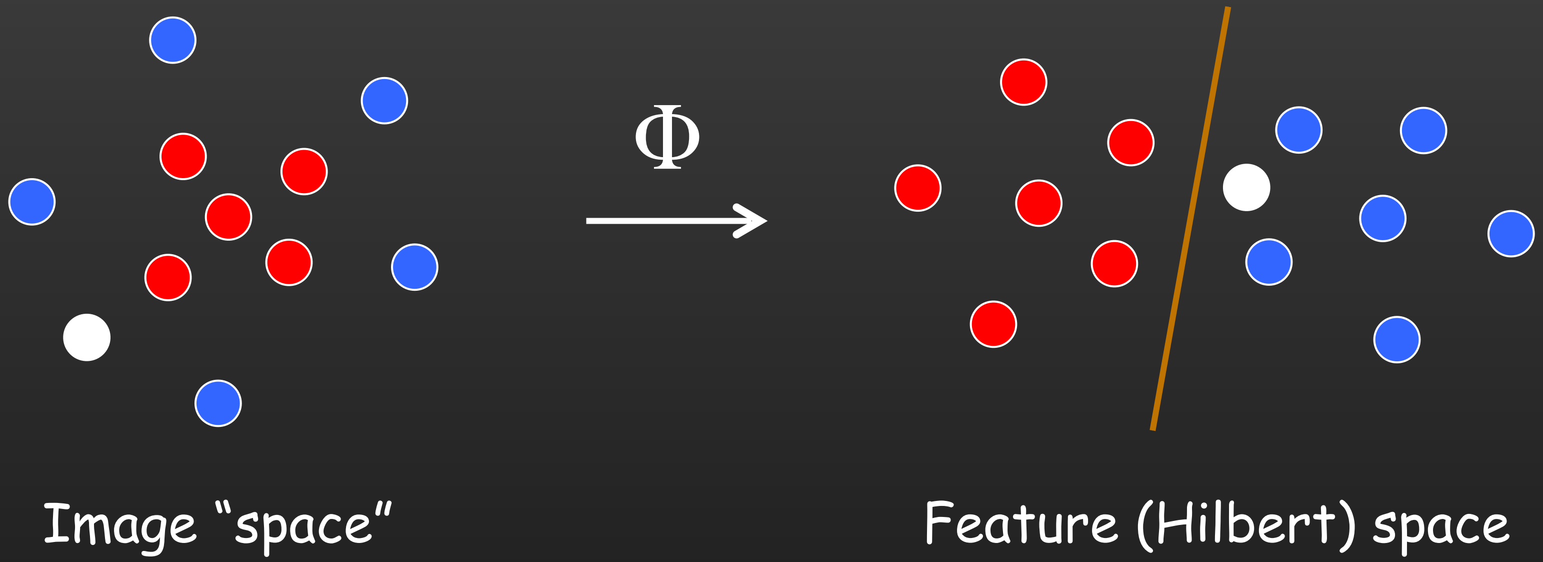
Test image

Image categorization as supervised classification



$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{n} \ell(\underbrace{z_n}_{\text{Label}}, \underbrace{f}_{\text{Prediction function}}(\phi(\overset{\text{Training datum}}{x_n}))) + \Omega(f)$$

Image categorization as supervised classification



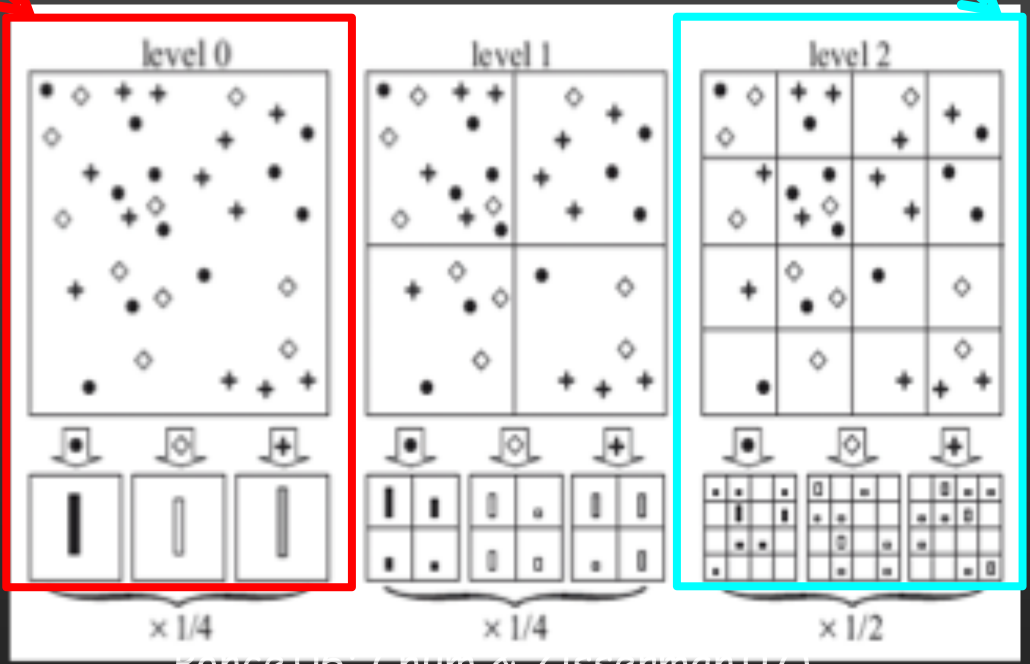
$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{n} \ell(z_n, \underset{\nwarrow\ \text{affine}}{f}(\phi(x_n))) + \Omega(f)$$

Spatial pyramids

BoW (Csurka et al.'04)

HOG (Dalal & Triggs'05)

(Koenderink & van Doorn'99)



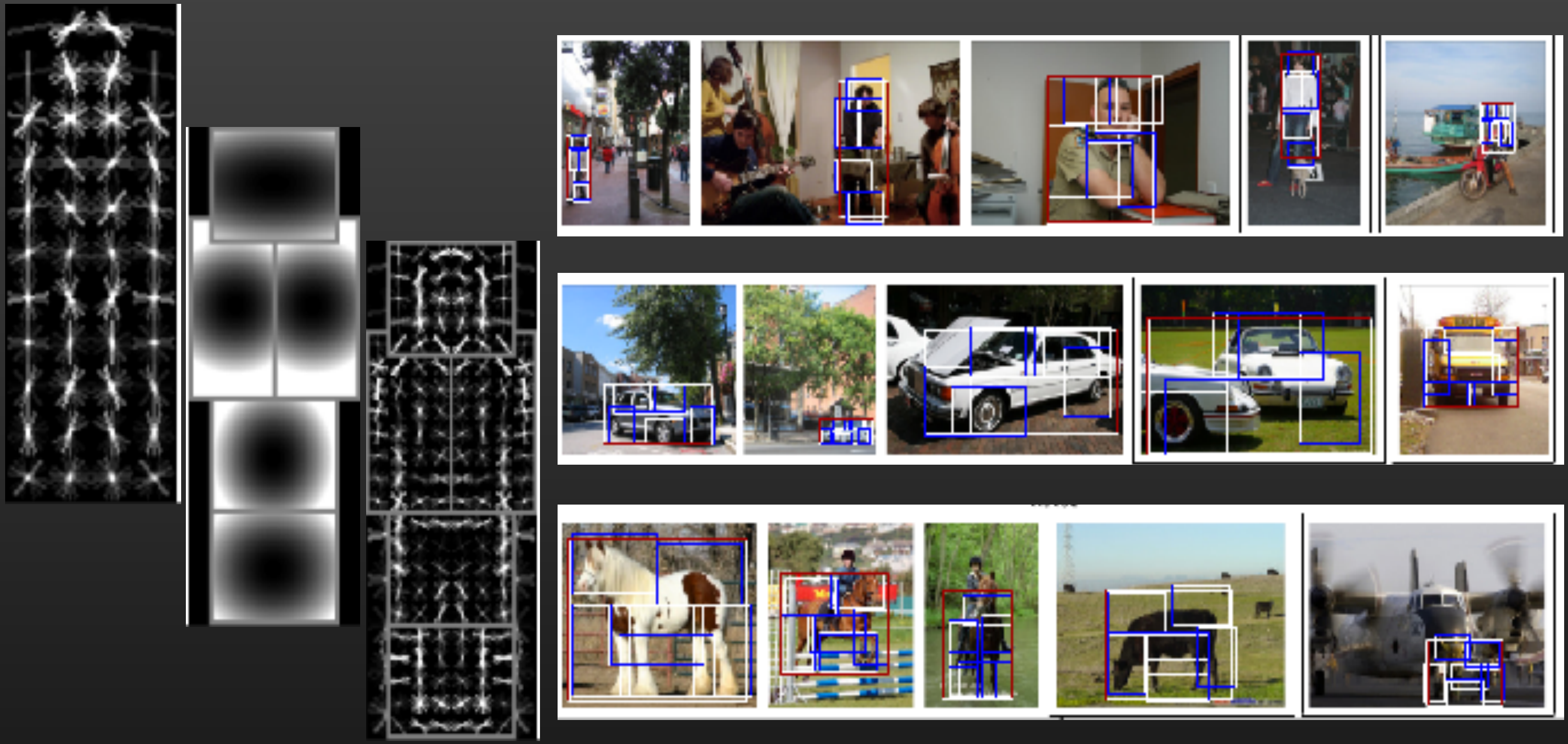
Ponce'06; Chum & Zisserman'07)

(Lazebnik, Schmid, Ponce'06)

- Bags of words=orderless models=histograms of visual words
- Spatial pyramids=locally orderless models
- Classifier: support vector machine=a linear classifier

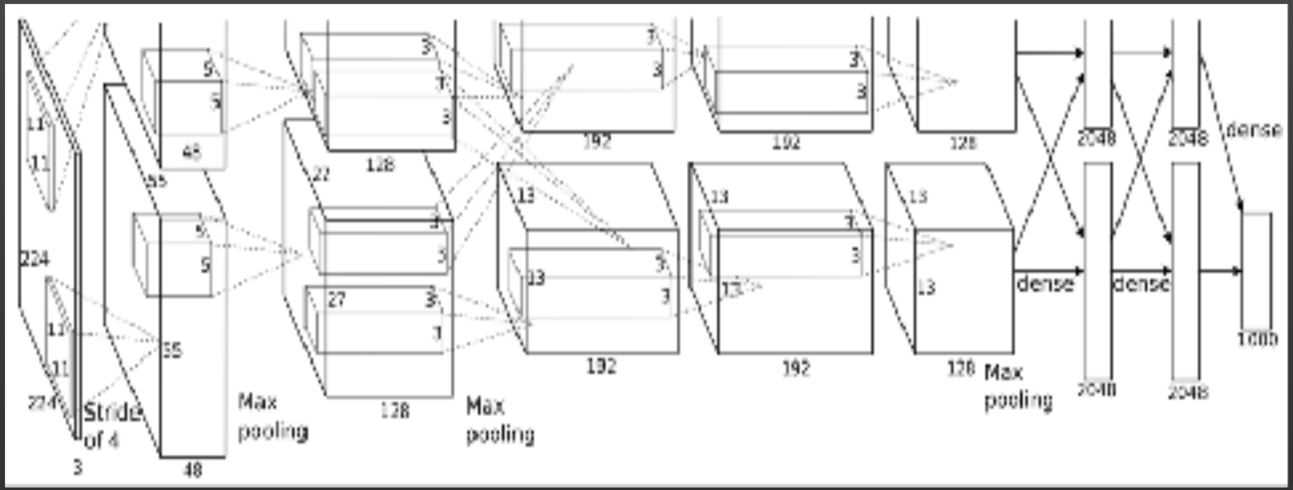
(Swain & Ballard'91, Grauman & Darrell'05, Zhang et al.'06, Felzenszwalb'08)

Discriminatively trained part-based models

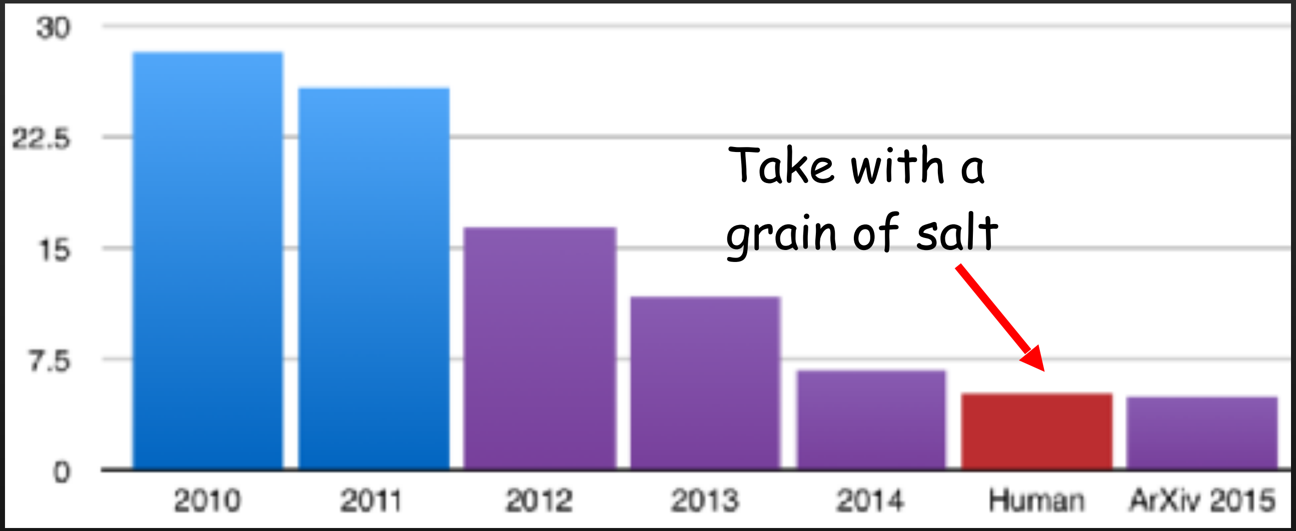


(Felzenszwalb, Girshick, McAllester, Ramanan'08)

The “revolution” of deep learning in 2012



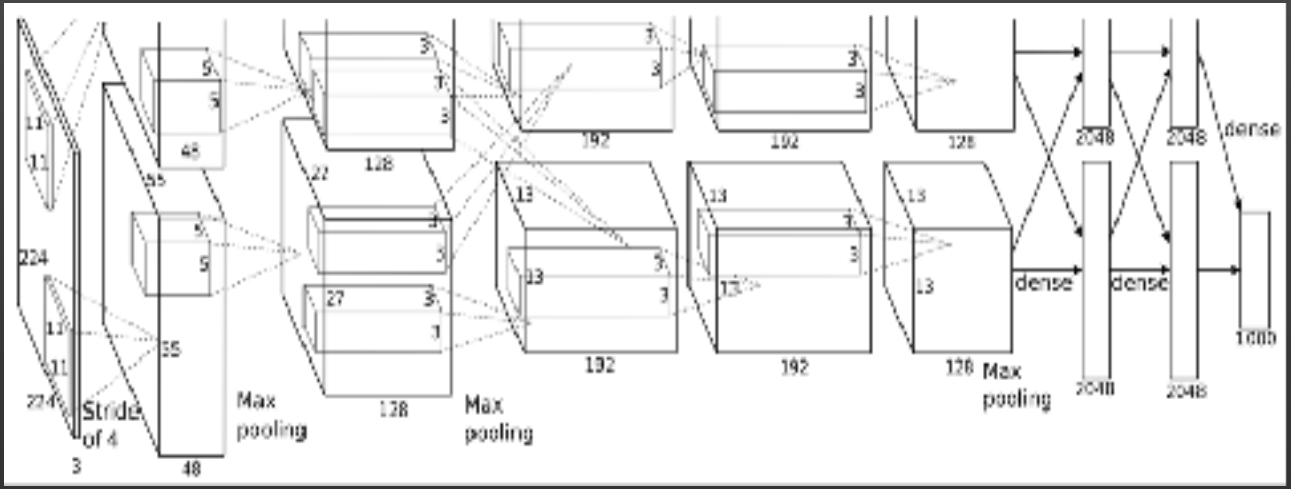
(Krizhevsky, Sutskever, Hinton, 2012)



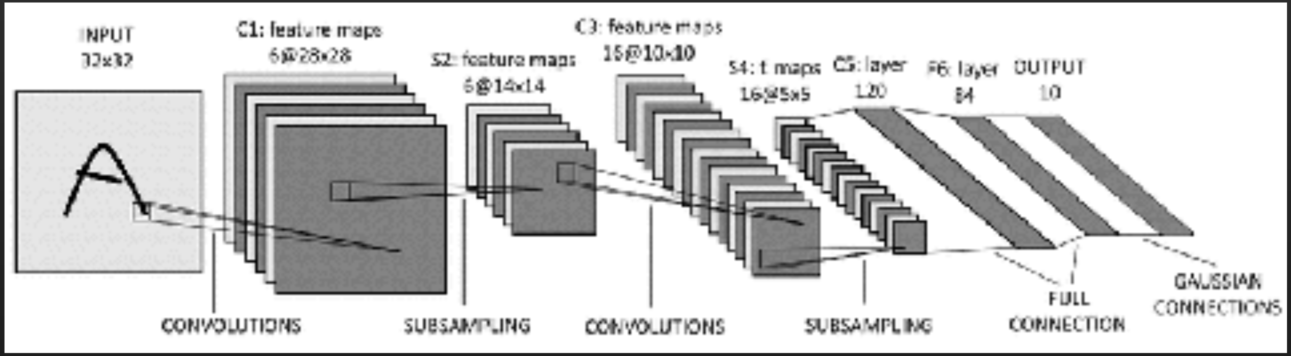
Take with a grain of salt

(And ResNets, GANs, RNNs, LSTMs, etc. [Schmidhuber’14, LeCun et al.’15])

The “revolution” of deep learning in 2012



(Krizhevsky, Sutskever, Hinton, 2012)



answer: 31

Convolutional nets early 90s (LeCun et al.’98)
(And one should not forget Pomerleau 1980s.)

(And ResNets, GANs, RNNs, LSTMs, etc. [Schmidhuber’14, LeCun et al.’15])

Image categorization as supervised classification

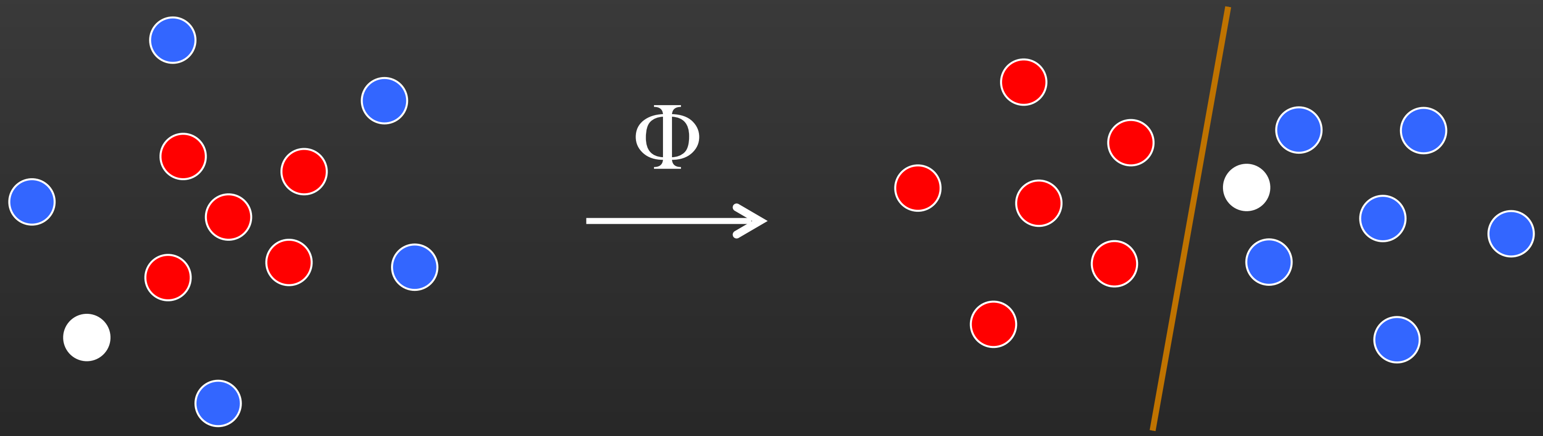
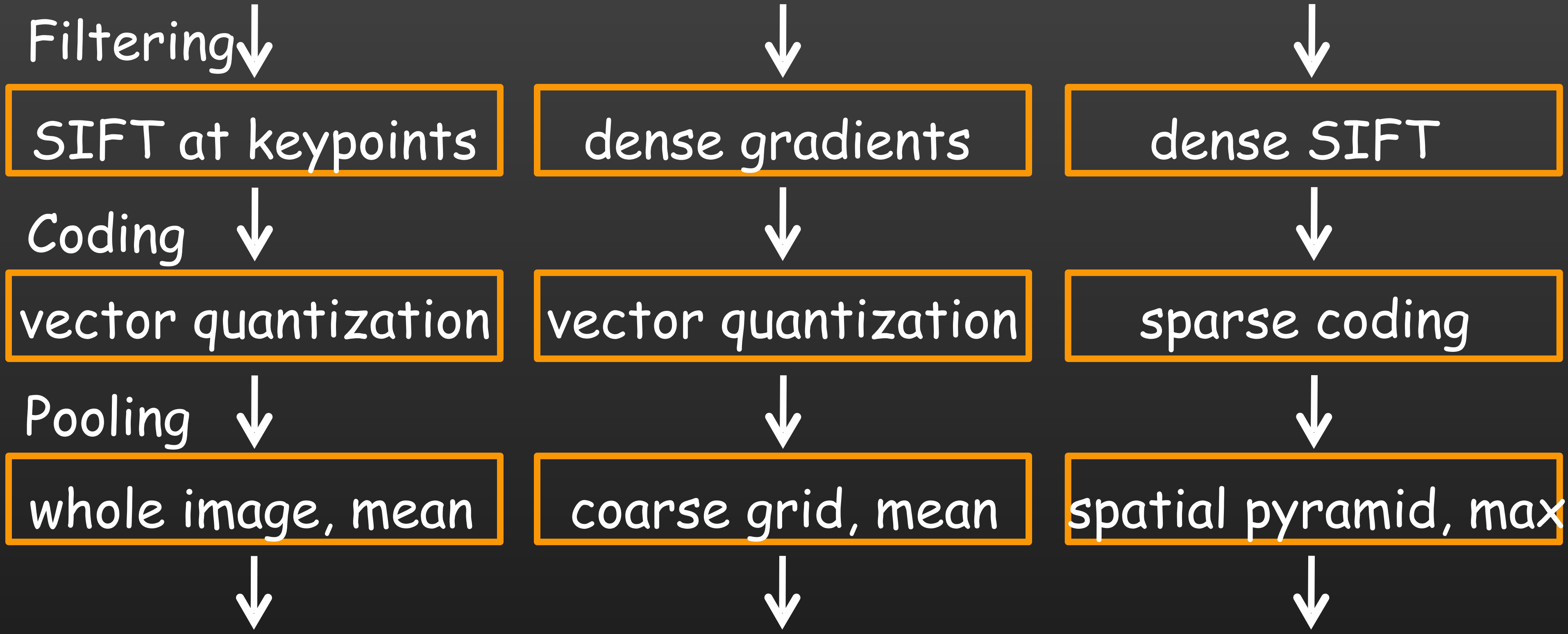


Image "space"

Feature (Hilbert) space

$$\min_{\substack{f\in\mathcal{F}\\ \theta}} \frac{1}{N}\sum_{n} \ell(z_n, f(\underset{\theta}{\phi}(x_n))) + \Omega(f)$$

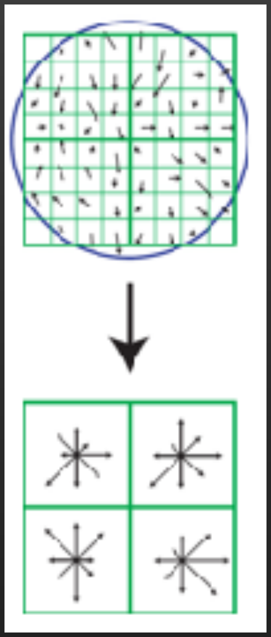
A common architecture for image classification



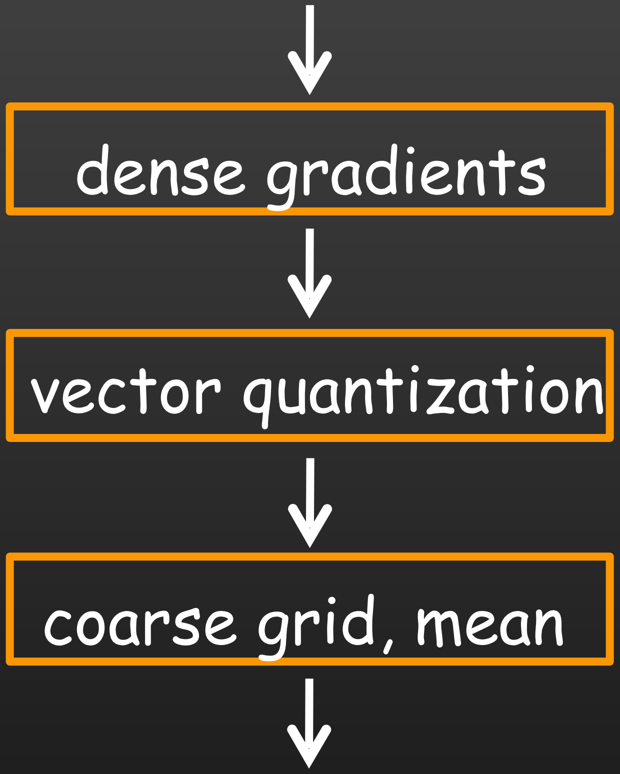
Filtering	SIFT at keypoints	dense gradients	dense SIFT
Coding	vector quantization	vector quantization	sparse coding
Pooling	whole image, mean	coarse grid, mean	spatial pyramid, max

(Lowe'04, Csurka et al.'04, Dalal & Triggs'05)
(Yang et al.'09-10, Boureau et al.'10, Mallat'11)

A common architecture for image classification



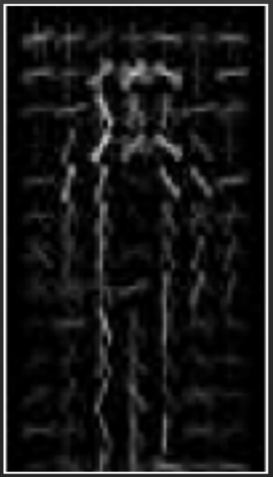
SIFT



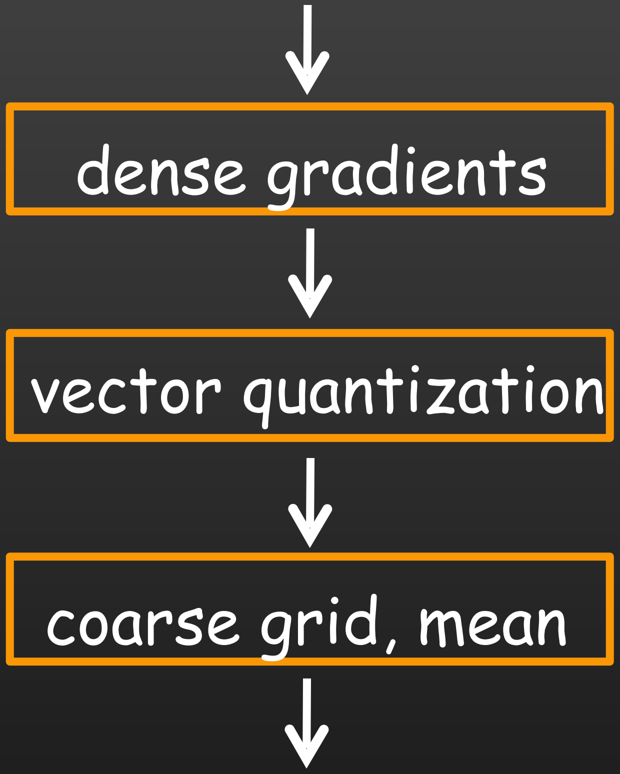
dense gradients

vector quantization

coarse grid, mean



HOG



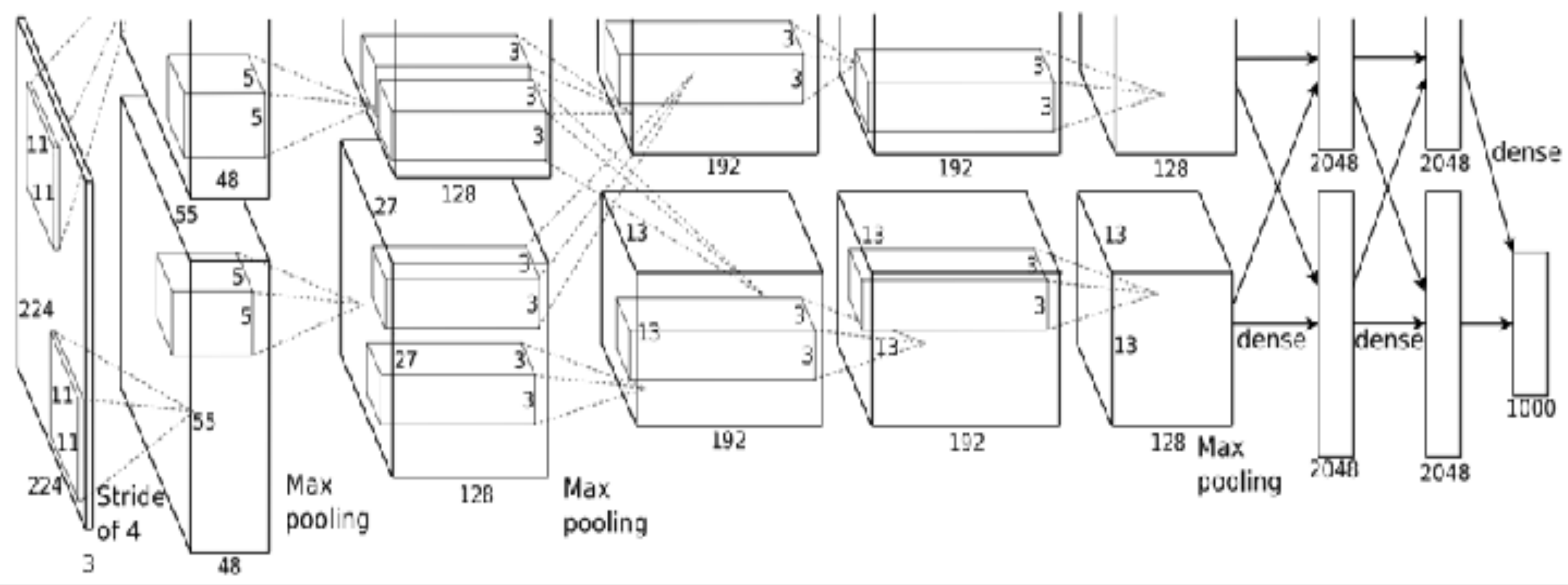
dense gradients

vector quantization

coarse grid, mean

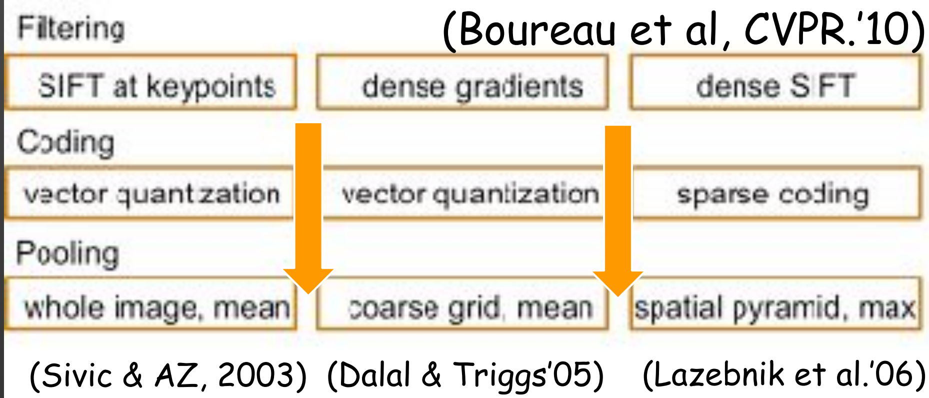
(Lowe'04, Csurka et al.'04, Dalal & Triggs'05)
(Yang et al.'09-10, Boureau et al.'10, Mallat'11)

A common architecture for image classification



(Deep learning: Krizhevsky, Sutskever, Hinton, 2012)

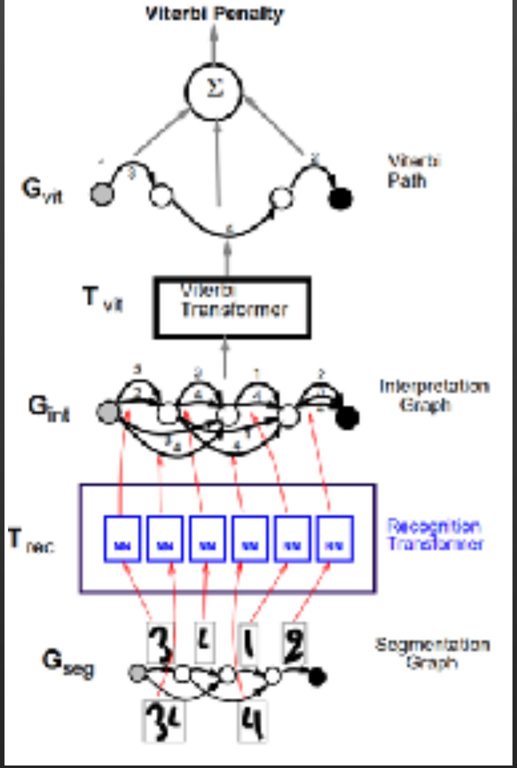
Beyond pattern recognition



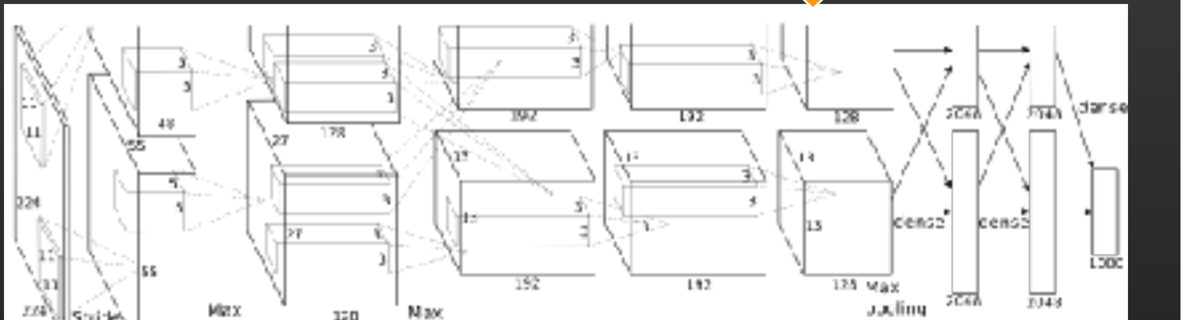
(Boureau et al, CVPR.'10)

	(Sivic & AZ, 2003)	(Dalal & Triggs'05)	(Lazebnik et al.'06)
Filtering	SIFT at keypoints	dense gradients	dense SIFT
Coding	vector quantization	vector quantization	sparse coding
Pooling	whole image, mean	coarse grid, mean	spatial pyramid, max

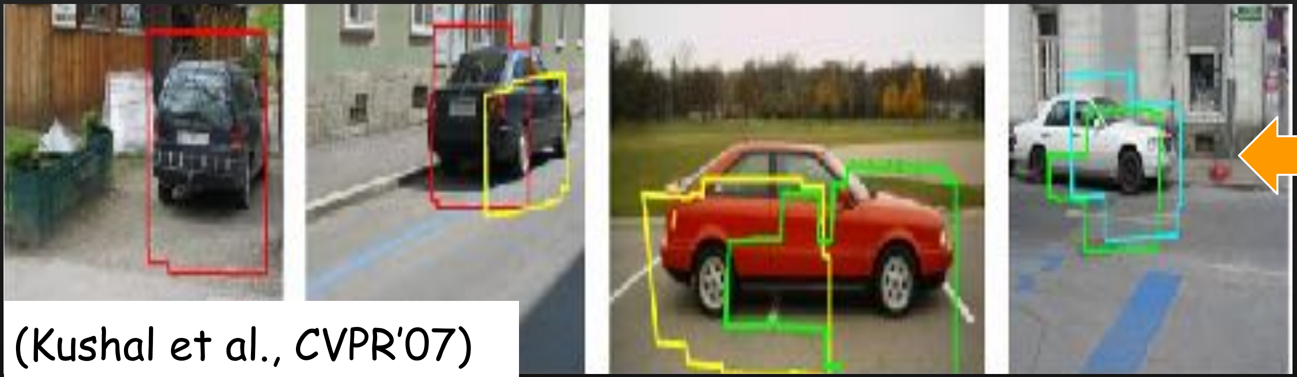
Deep learning (LeCun et al.'98)



Graph transformer networks



CNNs (Krizhevsky et al.'12)



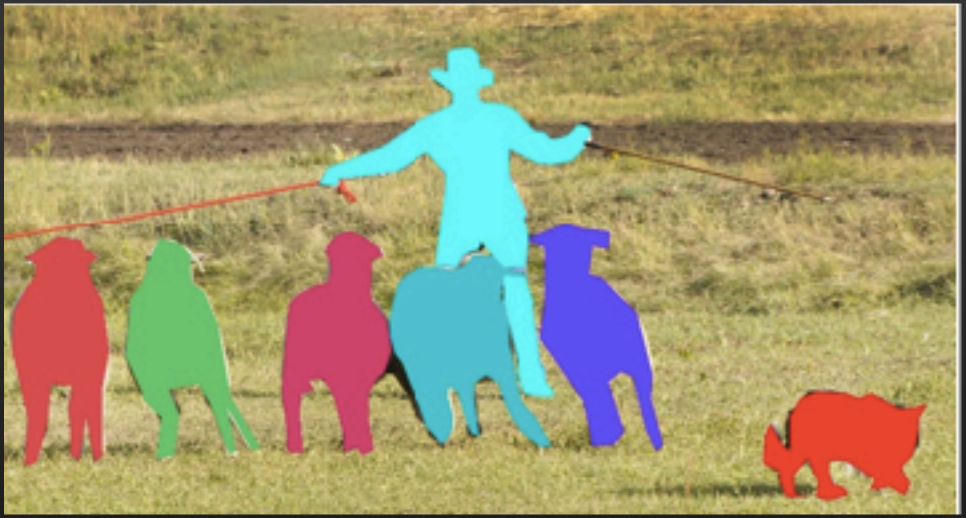
(Kushal et al., CVPR'07)

Didn't work so well but the problem is important!

Supervision: Where do the labels come from?

- A trend toward manually annotating the whole wide world with crowd sourcing

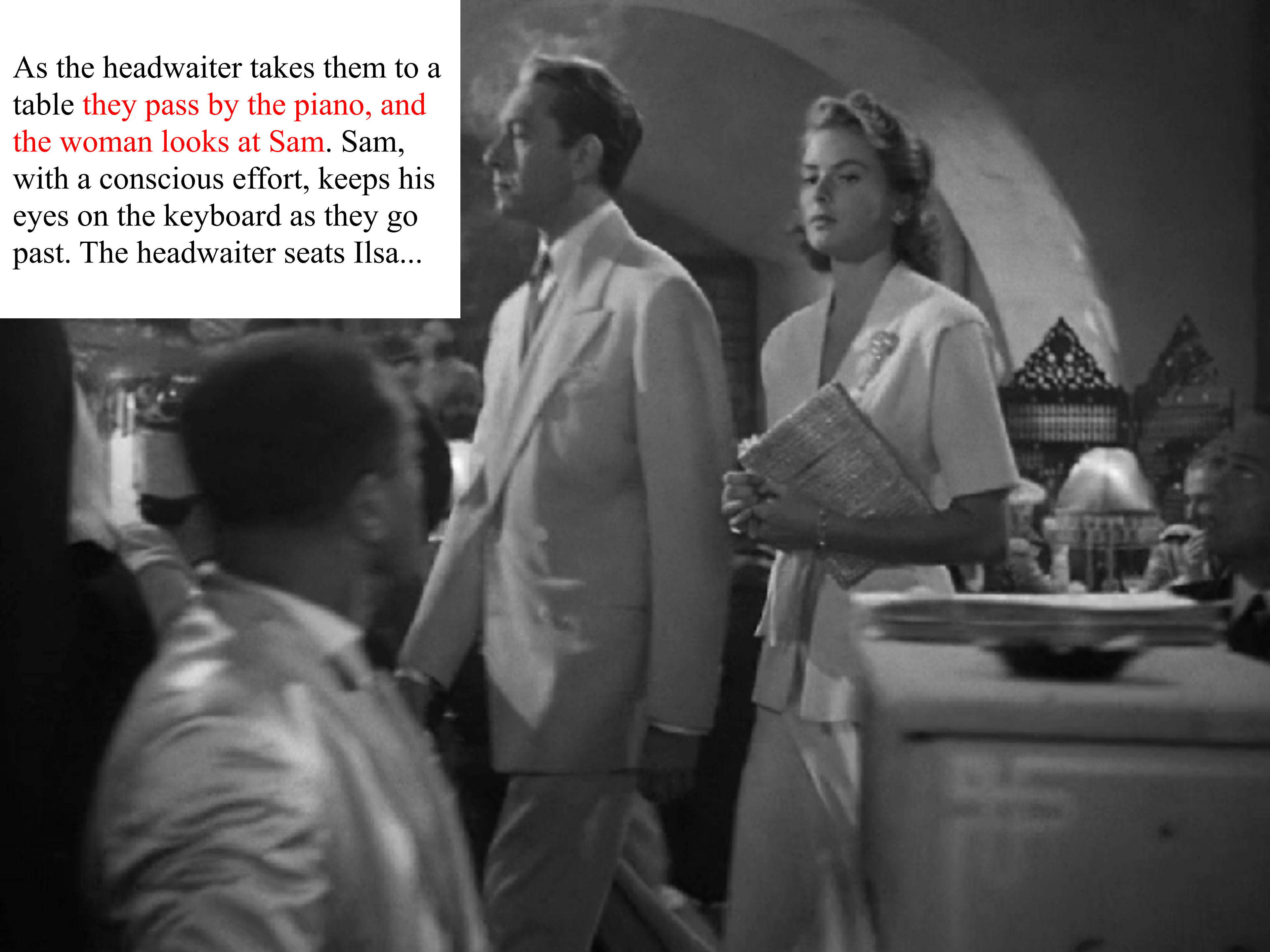
- Example: MS COCO (Lin et al., 2015) :328K images of 91 object categories



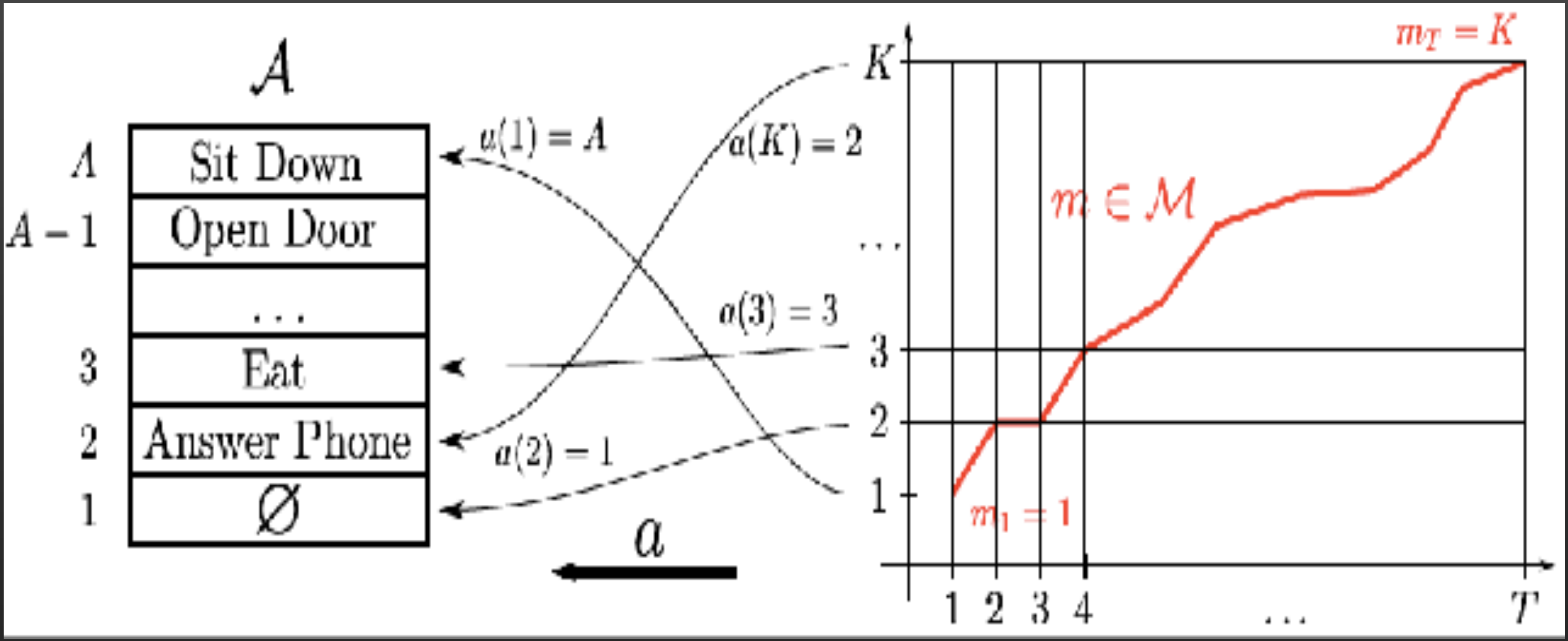
Scaling up: Little or no supervision

(Russell et al., 2008; Deng et al., 2009; Everingham et al., 2010; Xiao et al., 2010)

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



Action labeling under ordering constraints (Bojanowski et al., ECCV'14, CVPR'15)



Dictionary Script metadata a Alignment m

$$\min_{f \in \mathcal{F}} \left[\sum_{n=1}^{N} \min_{m \in \mathcal{M}} \frac{1}{T} \sum_{t=1}^{T} \ell\left(a_n(m_t), f(x_n(t))\right) \right] + \lambda \Omega(f)$$

Temporal action localization

Clip number 0101

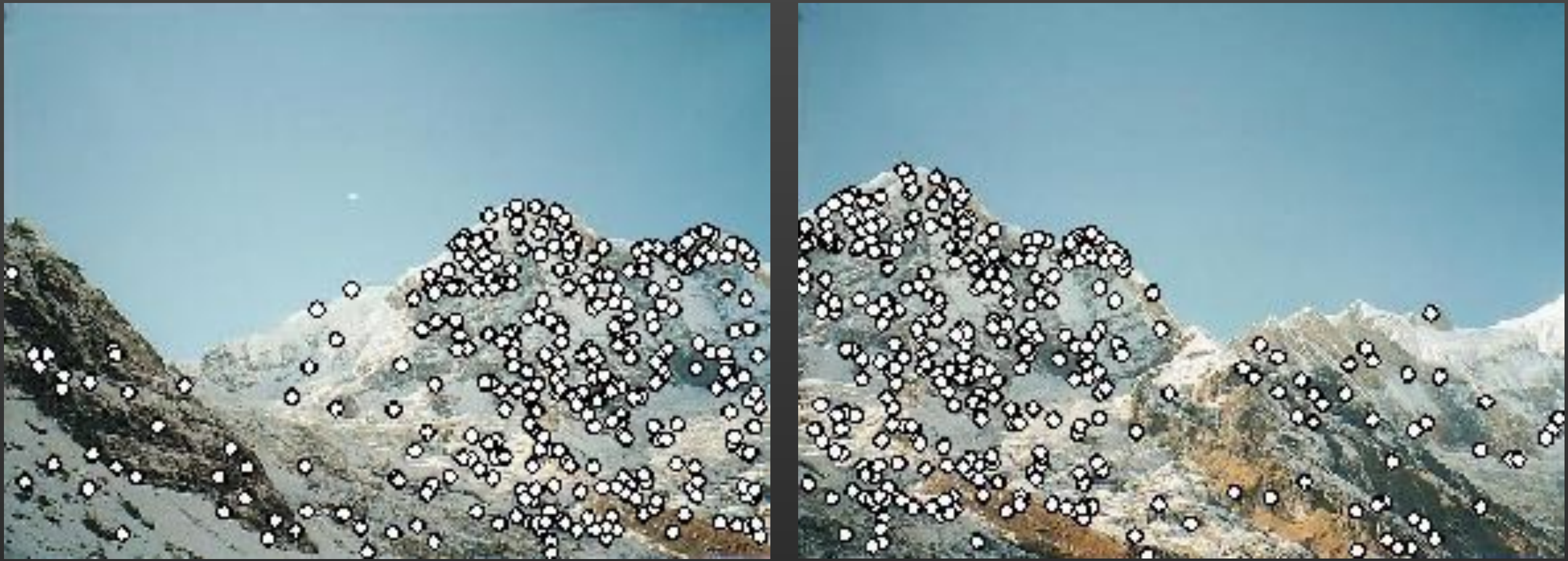
(Bojanowski et al., CVPR'15)

Outline

- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry
- Image processing

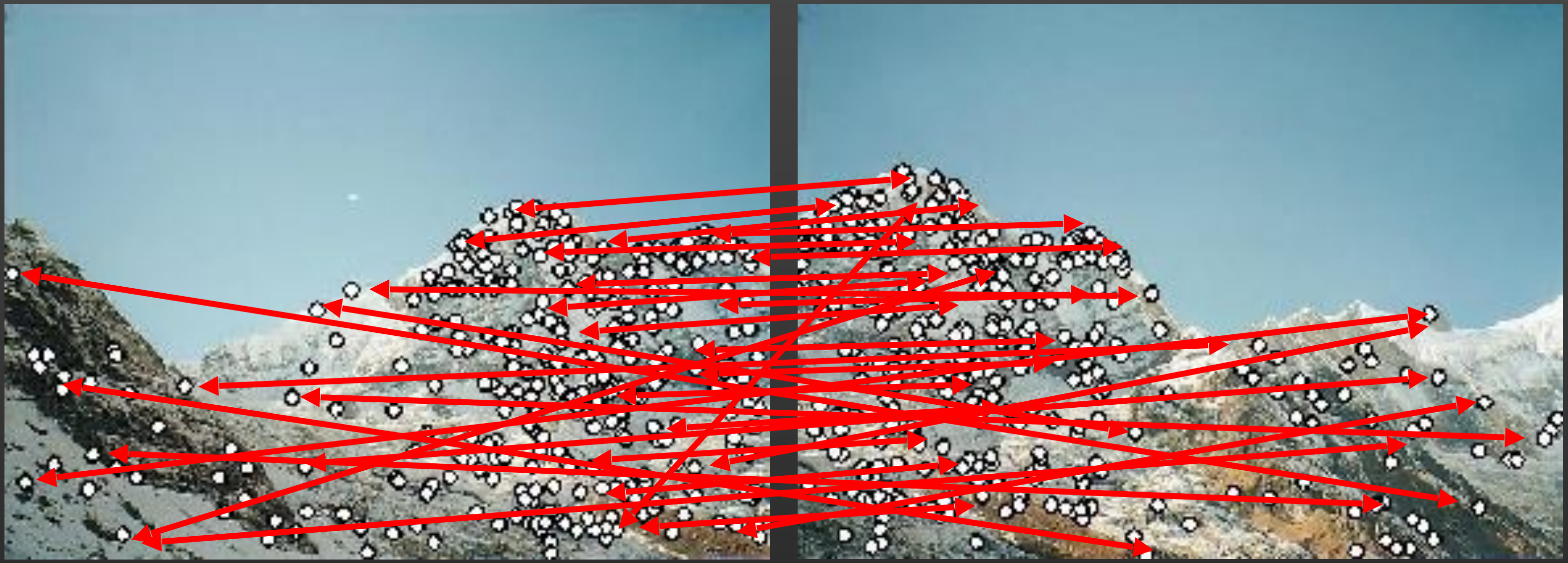
Feature-based alignment outline

Feature-based alignment outline



Extract features

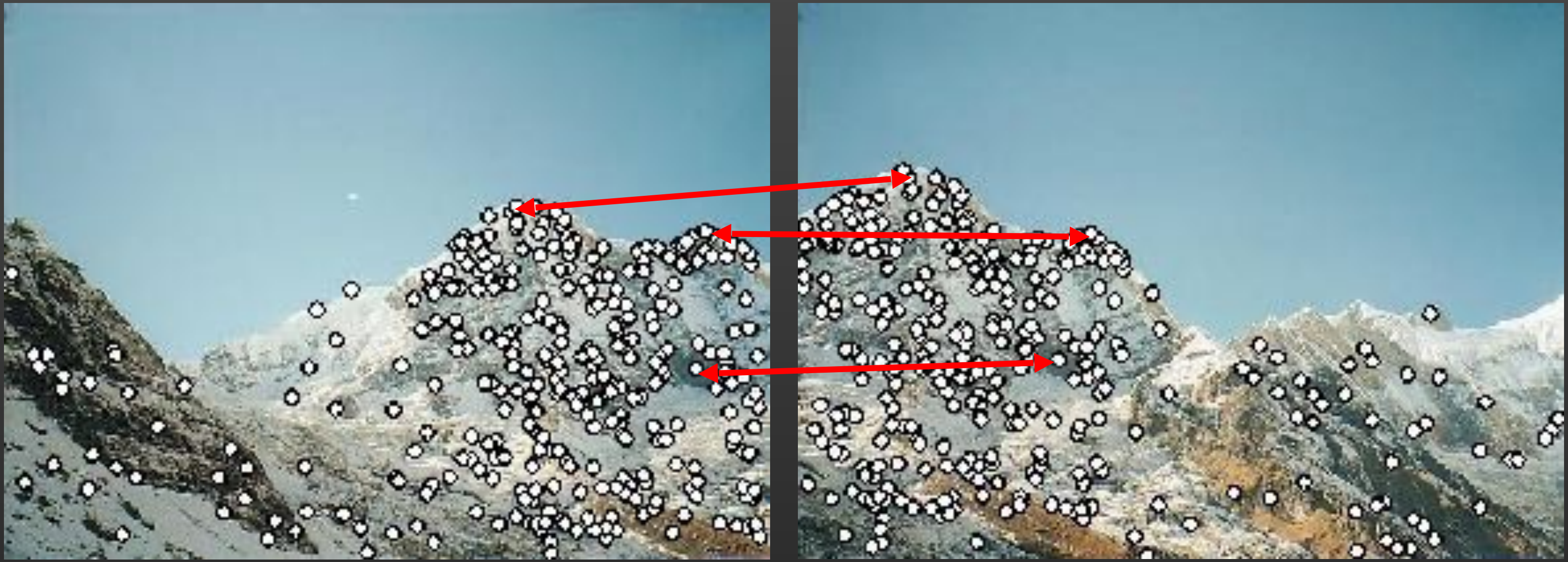
Feature-based alignment outline



Extract features

Compute *putative matches*

Feature-based alignment outline



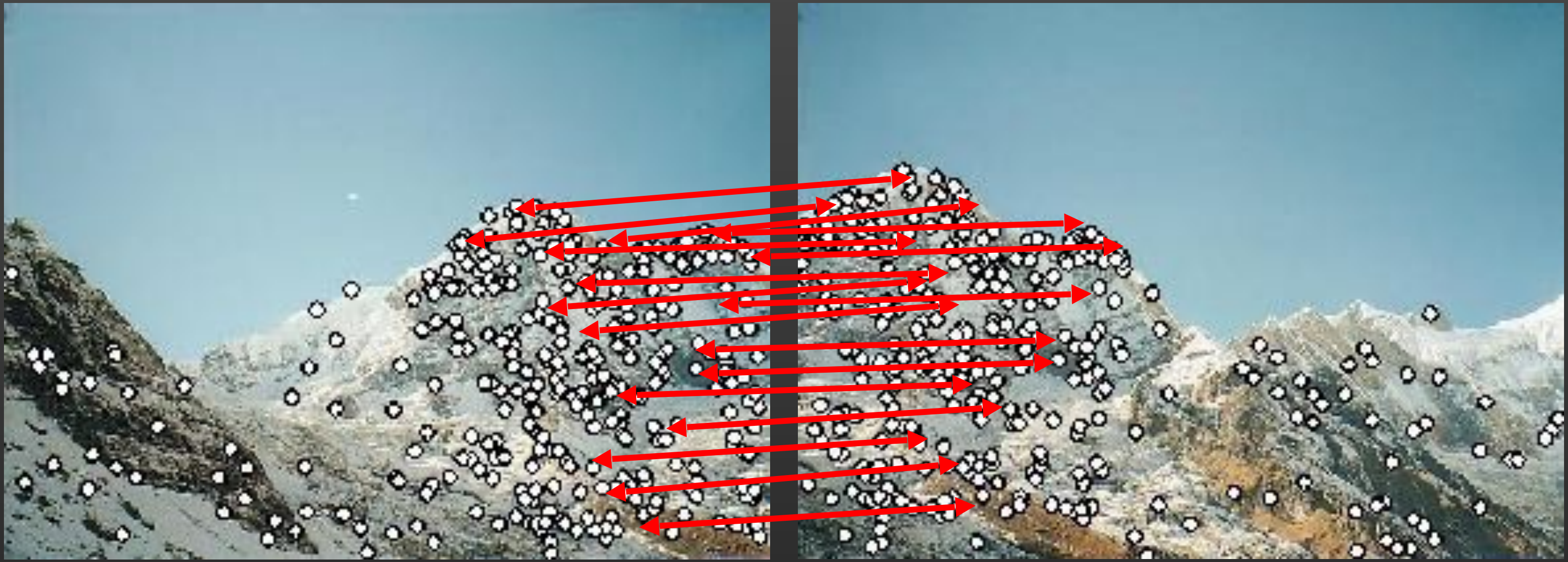
Extract features

Compute *putative matches*

Loop:

- *Hypothesize* transformation *T* (small group of putative matches that are related by *T*)

Feature-based alignment outline



Extract features

Compute *putative matches*

Loop:

- *Hypothesize* transformation *T* (small group of putative matches that are related by *T*)
- *Verify* transformation (search for other matches consistent with *T*)

Feature-based alignment outline

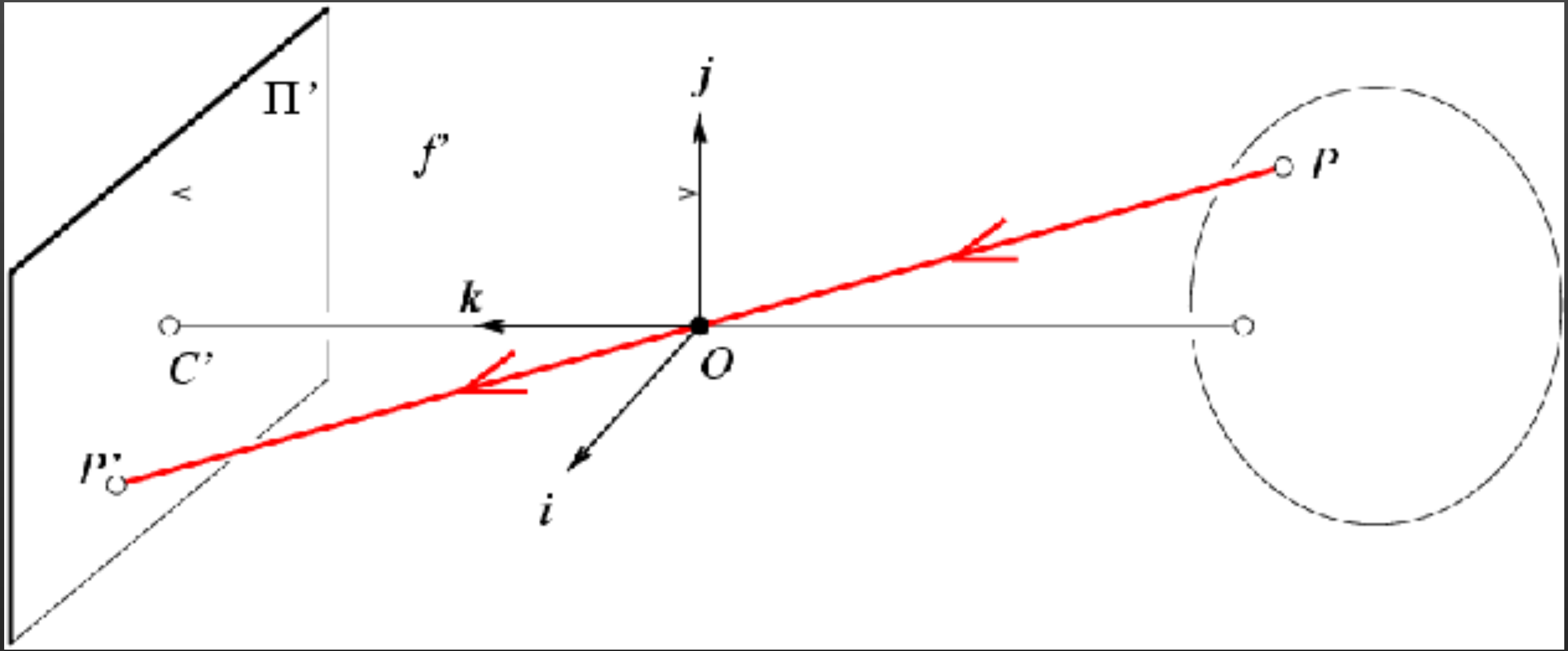
Extract features

Compute *putative matches*

Loop:

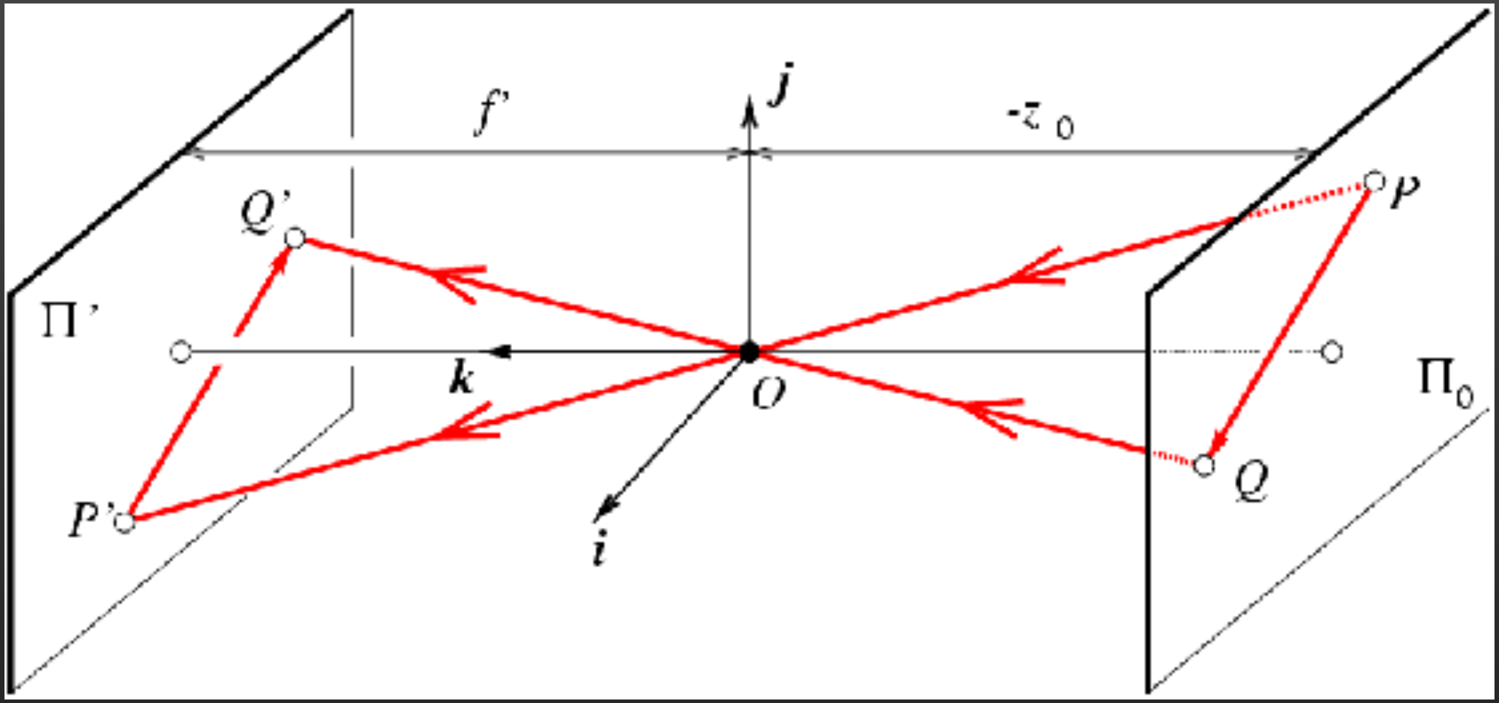
- *Hypothesize* transformation *T* (small group of putative matches that are related by *T*)
- *Verify* transformation (search for other matches consistent with *T*)

Pinhole perspective equation



NOTE: z is always negative..

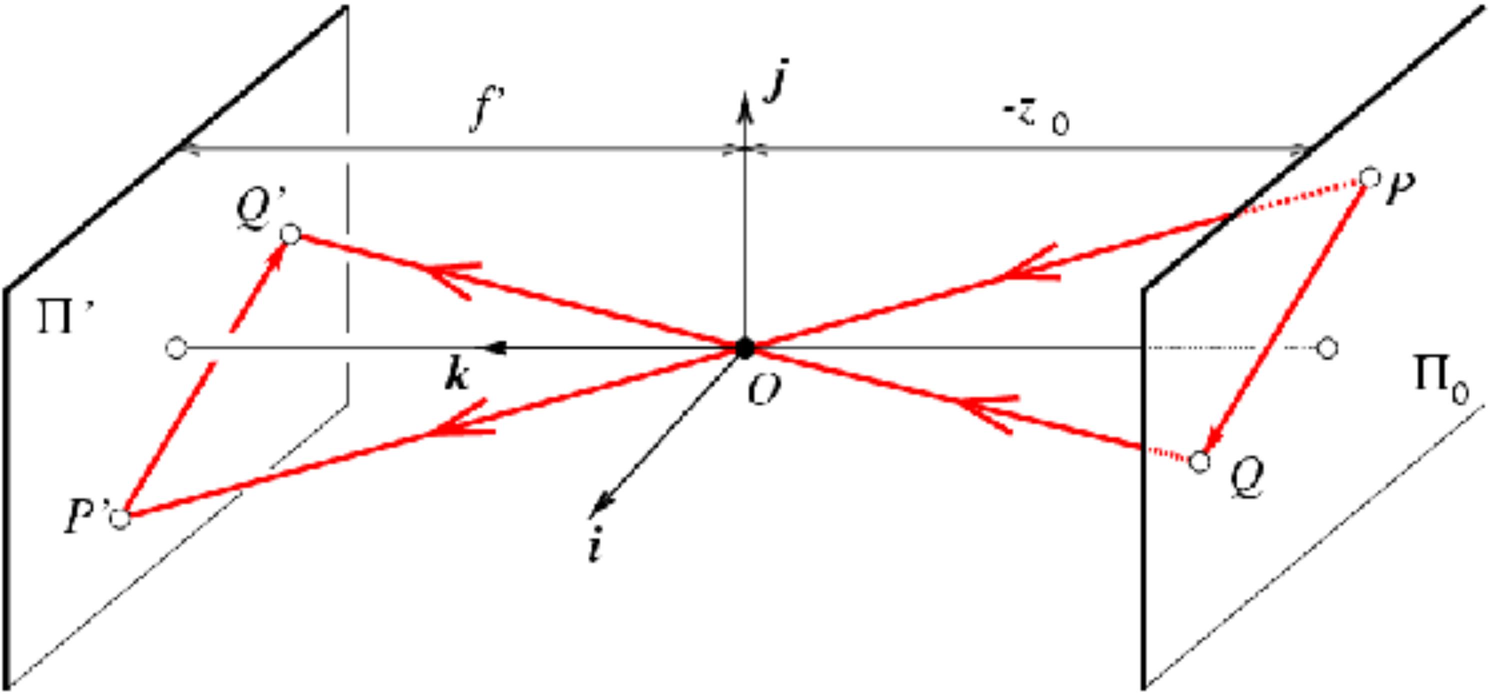
Affine models: Weak perspective projection

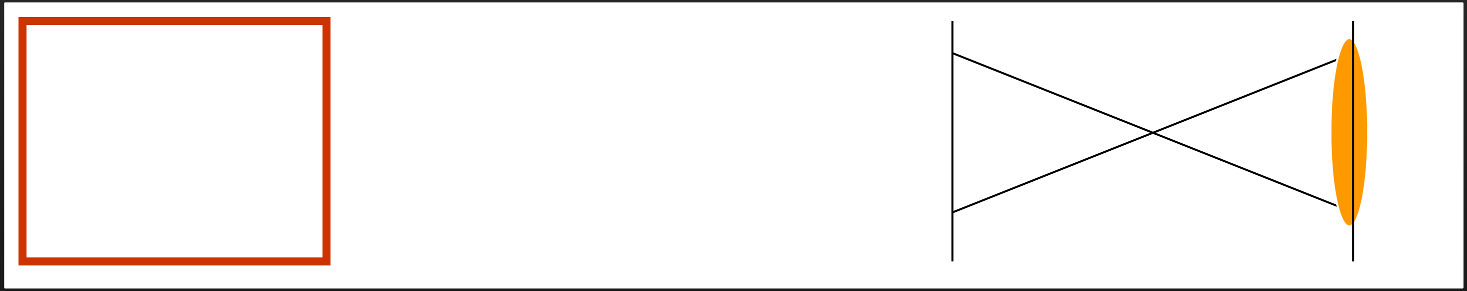


is the magnification.

When the scene relief is small compared its distance from the camera, m can be taken constant: weak perspective projection.

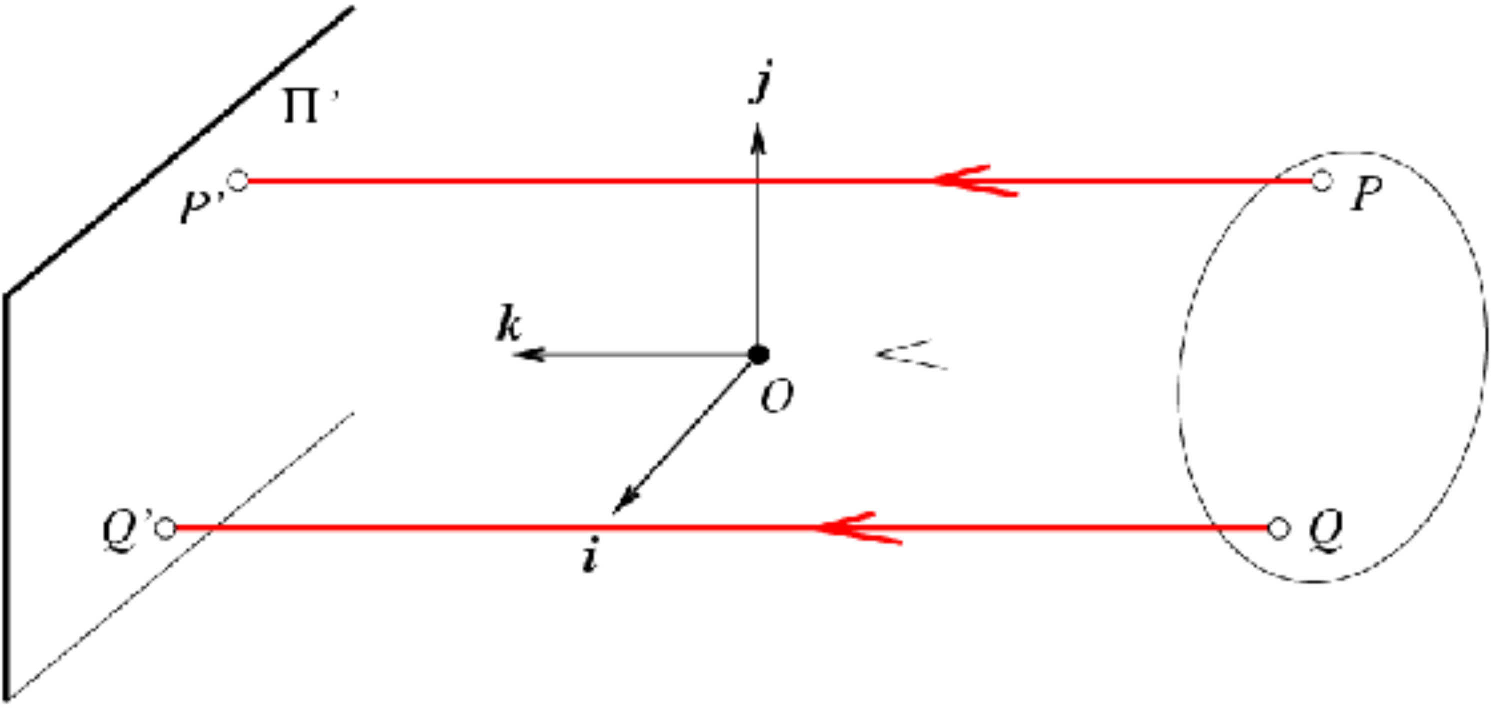
Affine models: Weak perspective projection





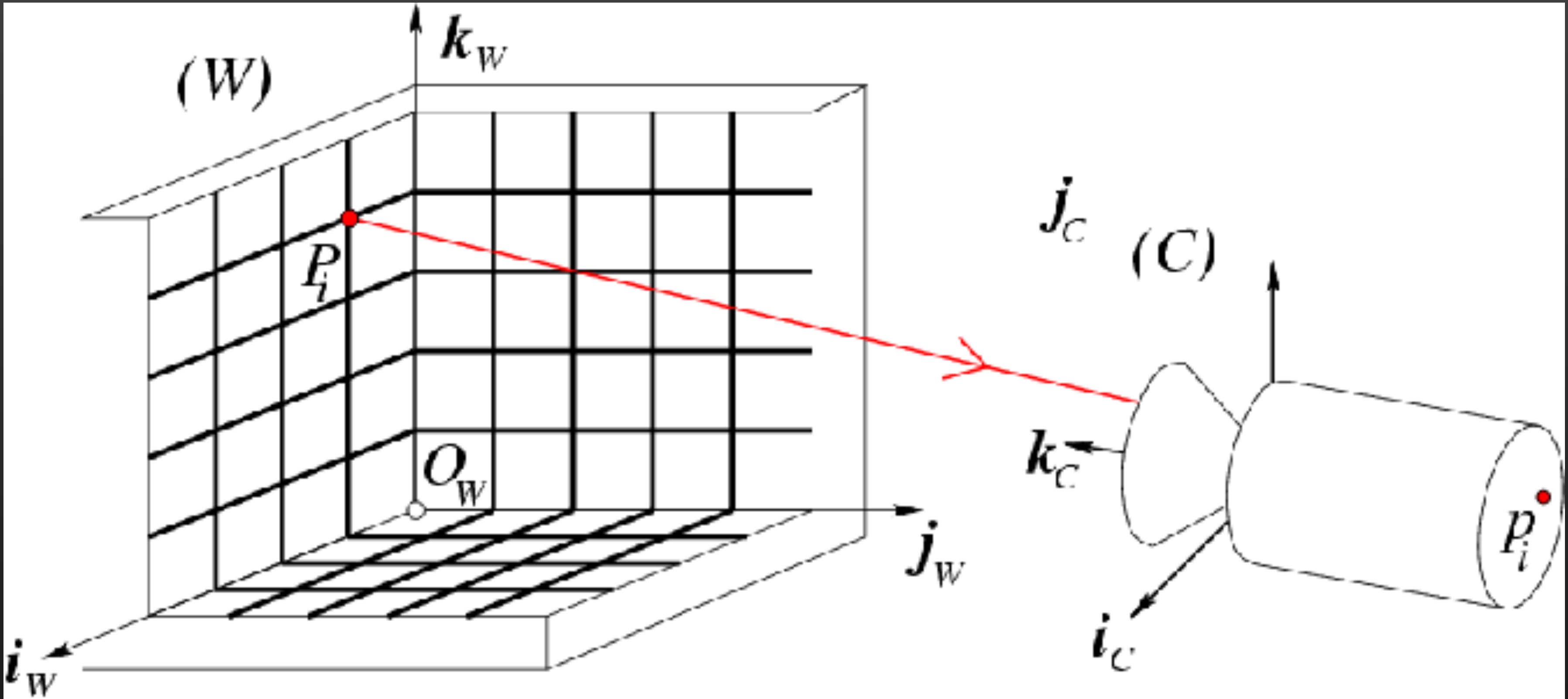
When the scene relief is small compared its distance from the camera, m can be taken constant: weak perspective projection.

Affine models: Orthographic projection



When the camera is at a (roughly constant) distance from the scene, take m=1.

Analytical camera geometry



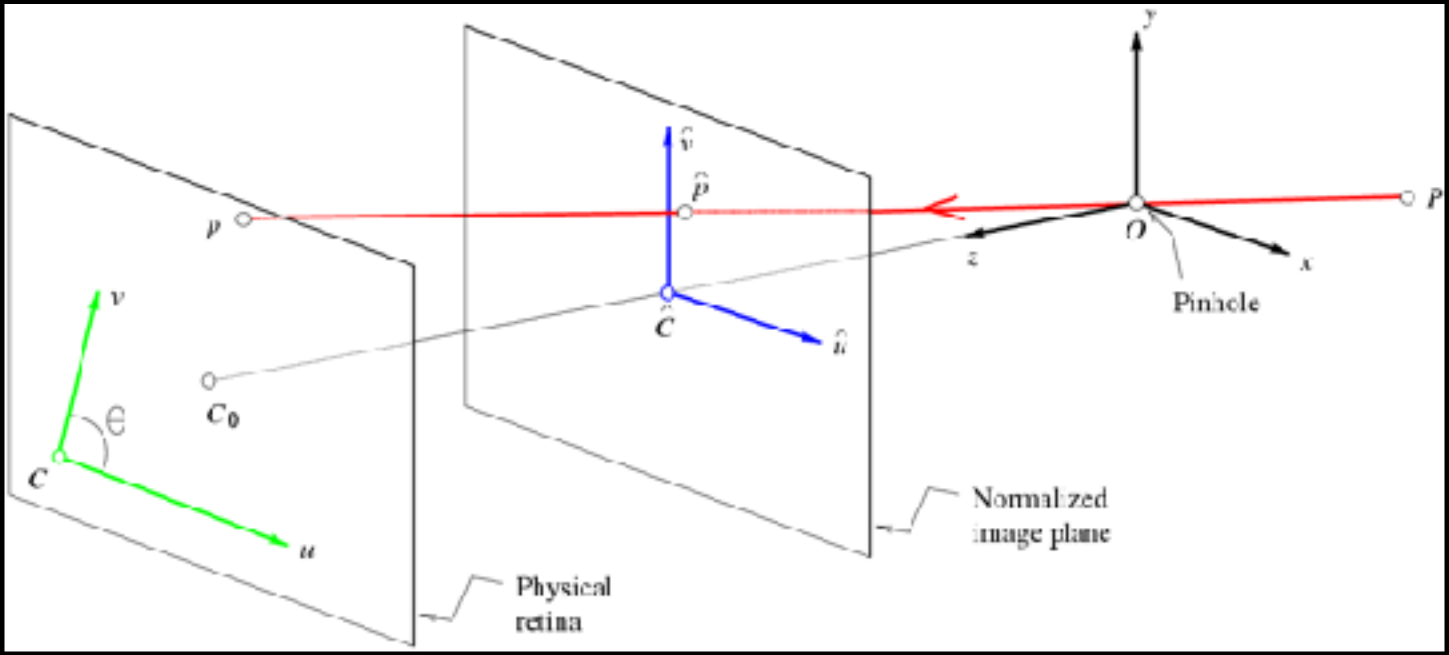
The intrinsic parameters of a camera

Units:

k,l : pixel/m

f : m

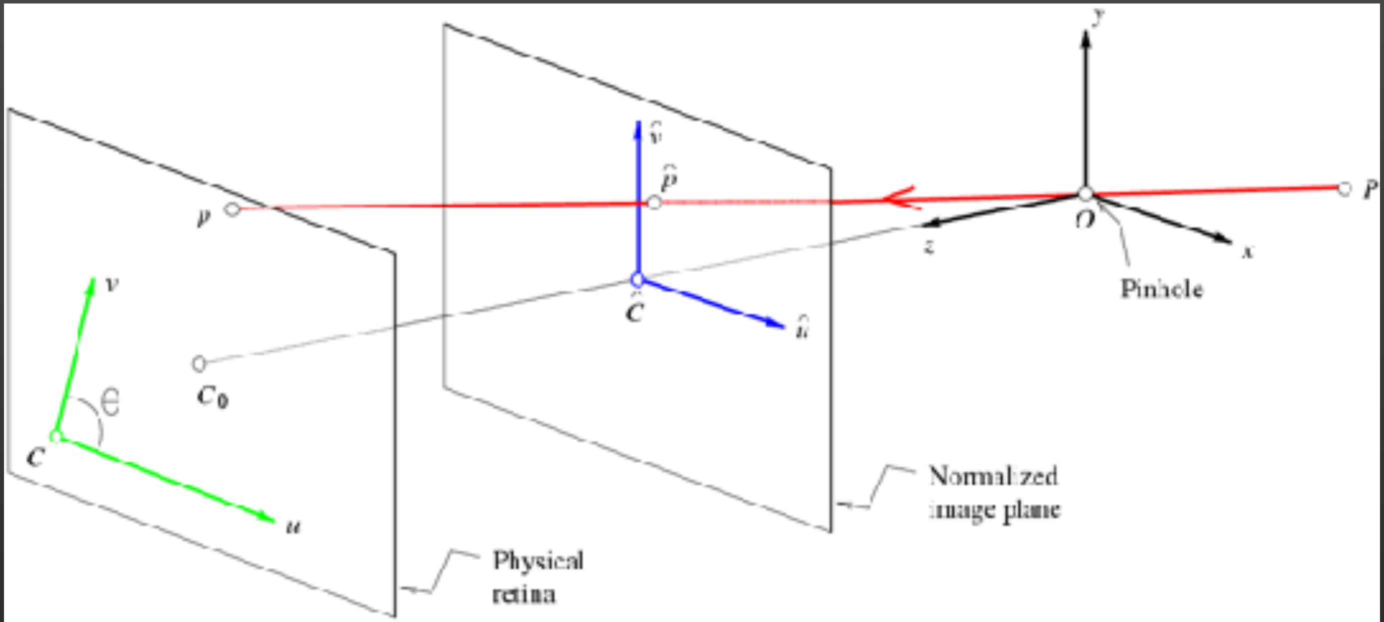
α,β : pixel



Physical image coordinates

Normalized image coordinates

The intrinsic parameters of a camera



Calibration matrix

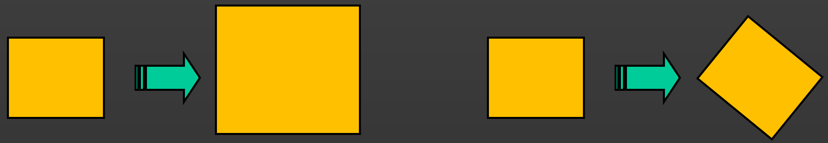
Homogeneous coordinates

The perspective projection equation

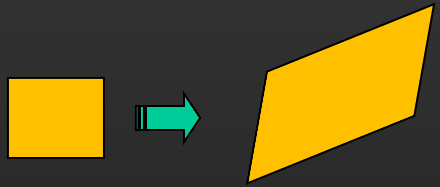
The extrinsic parameters of a camera

2D transformation models

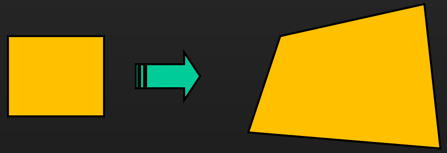
Similarity
(translation,
scale, rotation)



Affine transformation



Projective transformation
(homography)



Why these transformations ???

Weak-perspective projection model

($\boldsymbol{p}$ and $\boldsymbol{P}$ are in homogeneous coordinates)

r

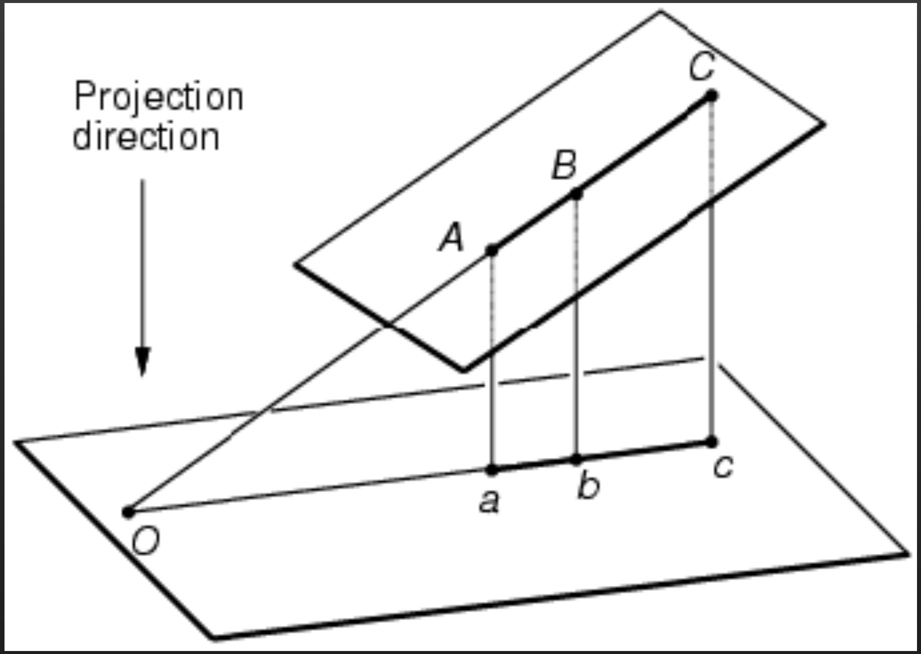
$$\boldsymbol{p} = M\,\boldsymbol{P}$$

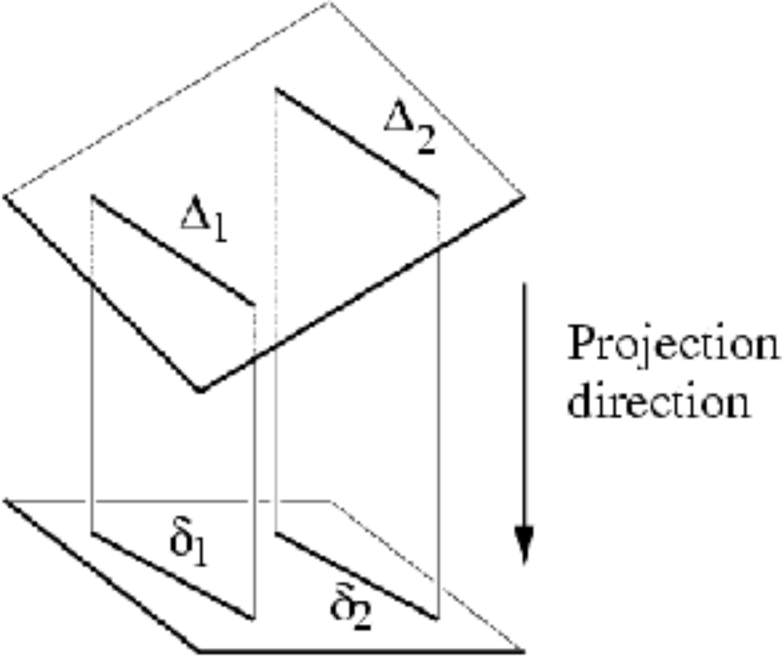
($\boldsymbol{P}$ is in homogeneous coordinates)

$$\boldsymbol{p} = A\,\boldsymbol{P} + \boldsymbol{b}$$

(neither $\boldsymbol{p}$ nor $\boldsymbol{P}$ is in hom. coordinates)

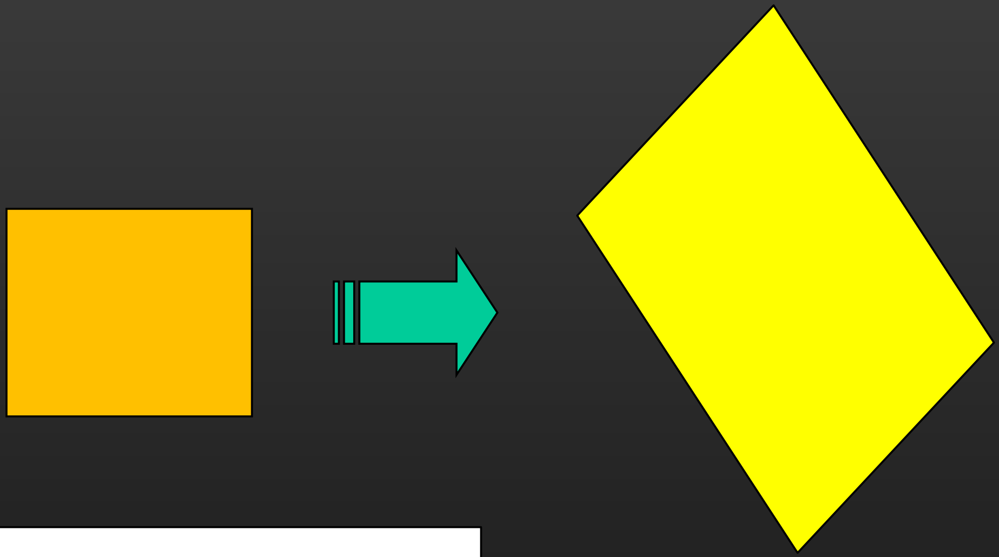
Affine projections induce affine transformations from planes onto their images.





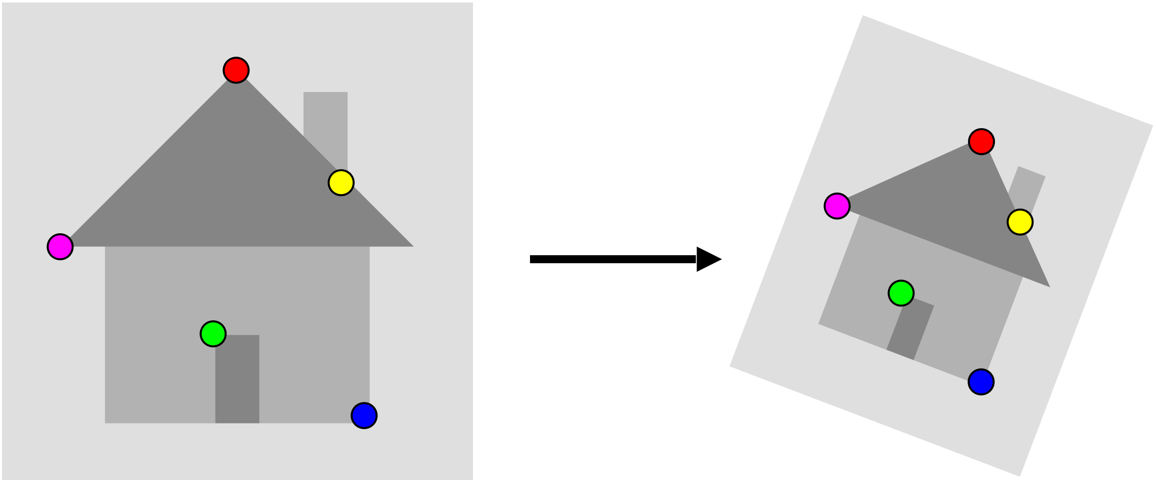
Affine transformations

An affine transformation maps a parallelogram onto another parallelogram



Fitting an affine transformation

Assume we know the correspondences, how do we get the transformation?



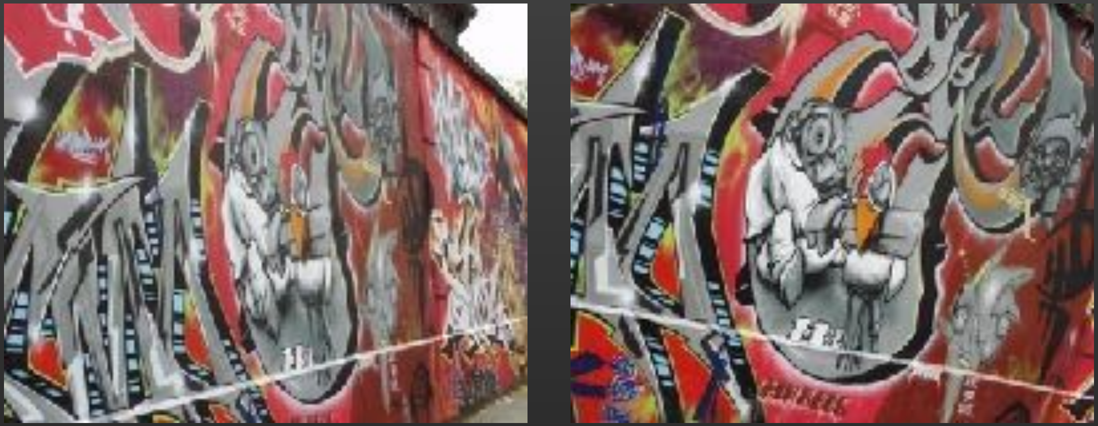
Fitting an affine transformation

Linear system with six unknowns

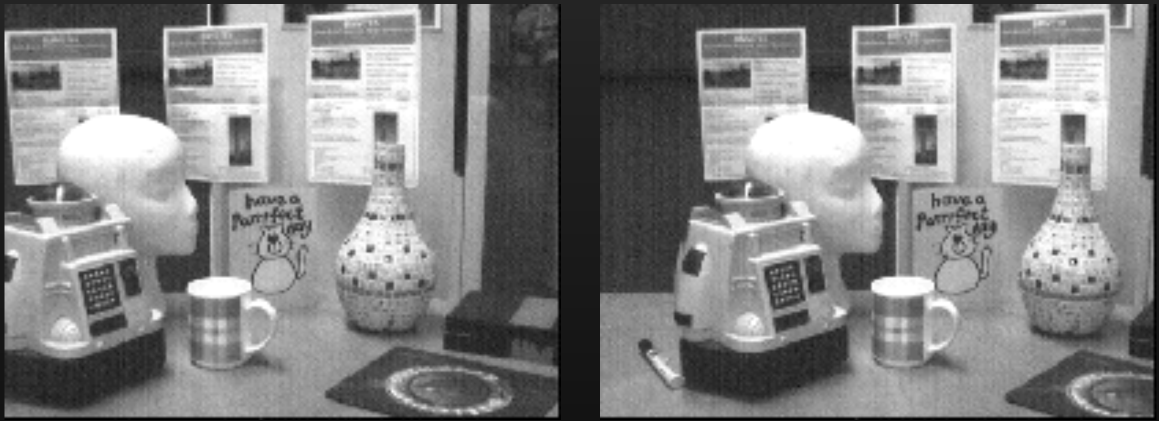
Each match gives us two linearly independent equations: need at least three to solve for the transformation parameters

Beyond affine transformations

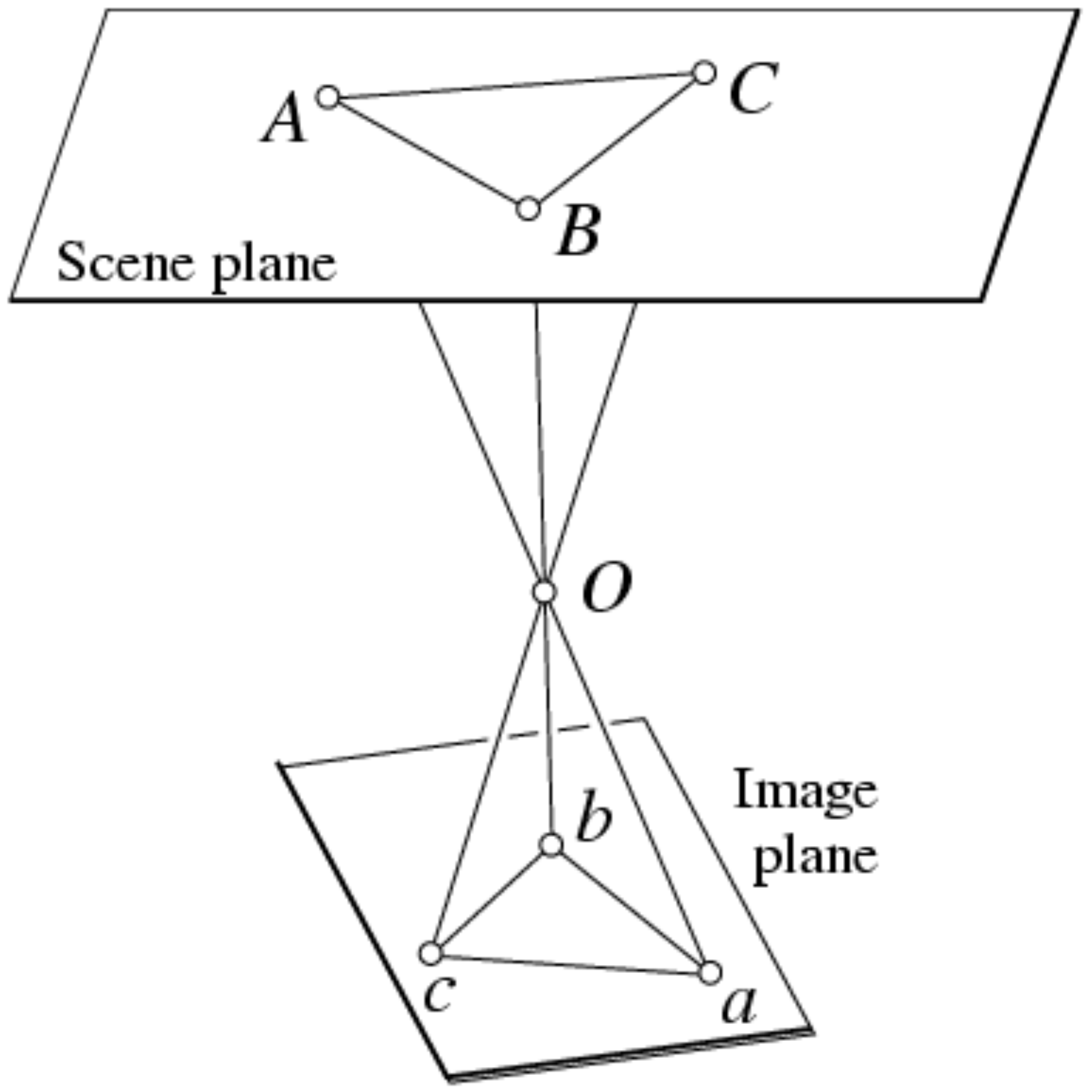
What is the transformation between two views of a planar surface?



What is the transformation between images from two cameras that share the same center?

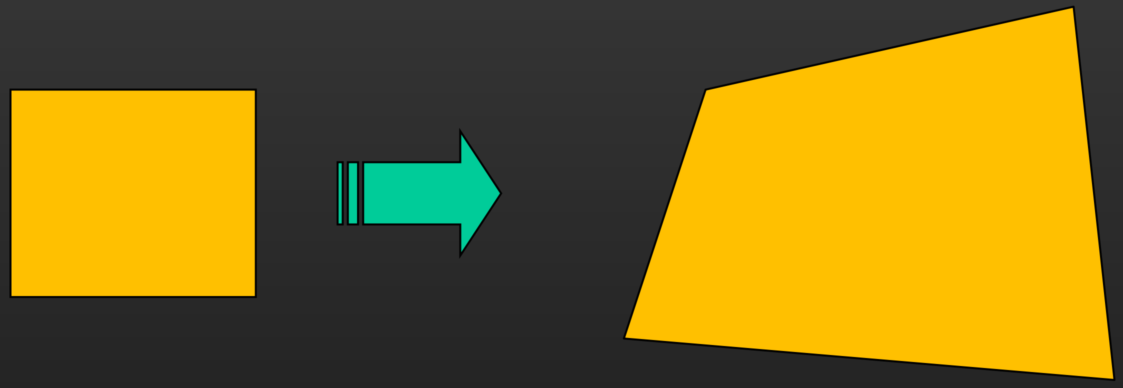


Perspective projections induce projective transformations between planes



Beyond affine transformations

Homography: plane projective transformation (transformation taking a quad to another arbitrary quad)



Fitting a homography

Recall: homogenenous coordinates

$$(x, y) \Rightarrow \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

Converting *to* homogenenous image coordinates

$$\begin{bmatrix} x \\ y \\ w \end{bmatrix} \Rightarrow (x/w, y/w)$$

Converting *from* homogenenous image coordinates

Fitting a homography

Recall: homogenenous coordinates

$$(x, y) \Rightarrow \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

Converting *to* homogenenous image coordinates

$$\begin{bmatrix} x \\ y \\ w \end{bmatrix} \Rightarrow (x/w, y/w)$$

Converting *from* homogenenous image coordinates

Equation for homography:

Fitting a homography

Equation for homography:

9 entries, 8 degrees of freedom
(scale is arbitrary)

3 equations, only 2 linearly
independent

Direct linear transform

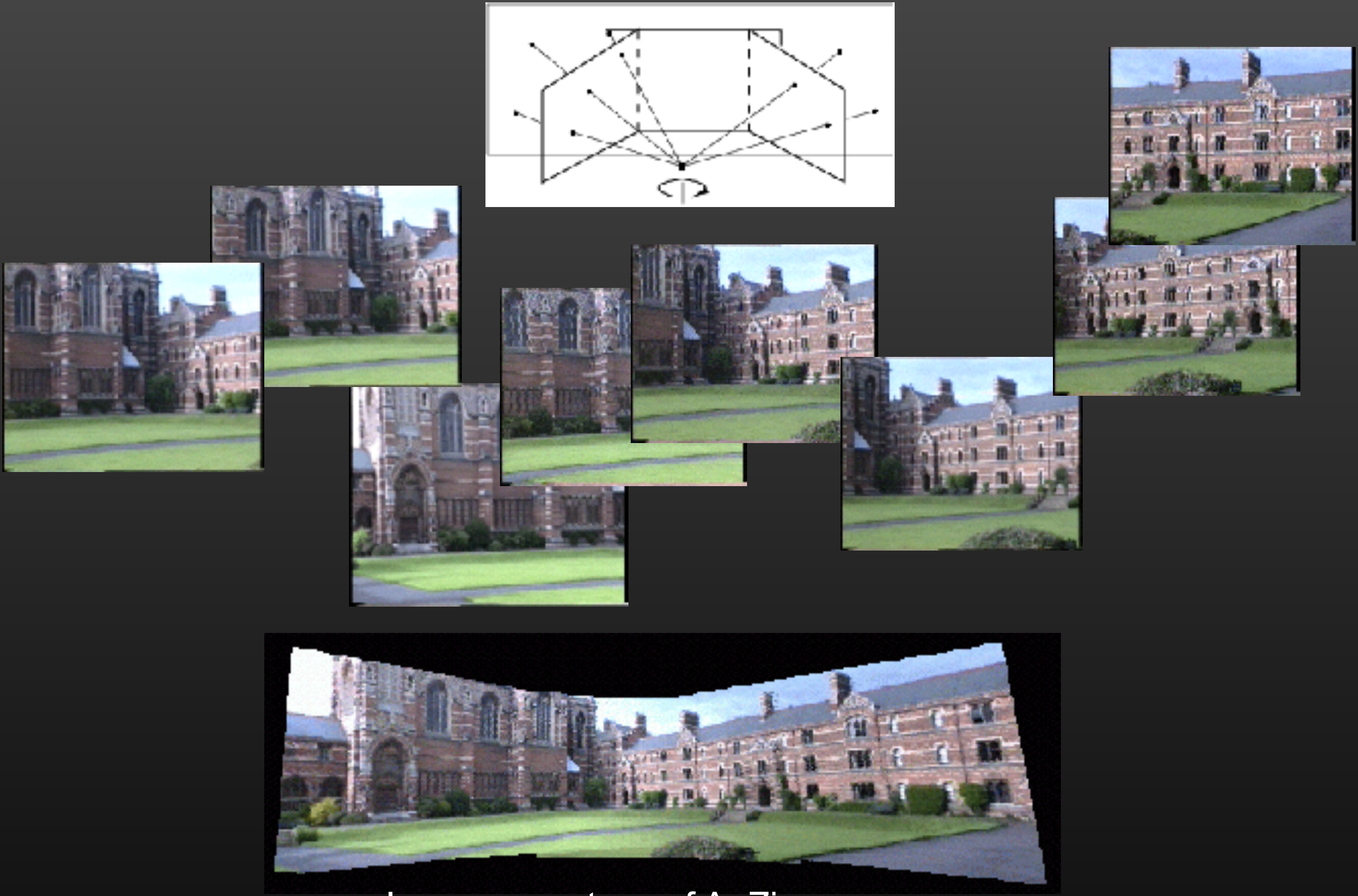
H has 8 degrees of freedom (9 parameters, but scale is arbitrary)

One match gives us two linearly independent equations

Four matches needed for a minimal solution (null space of 8x9 matrix)

More than four: homogeneous least squares

Application: Panorama stitching

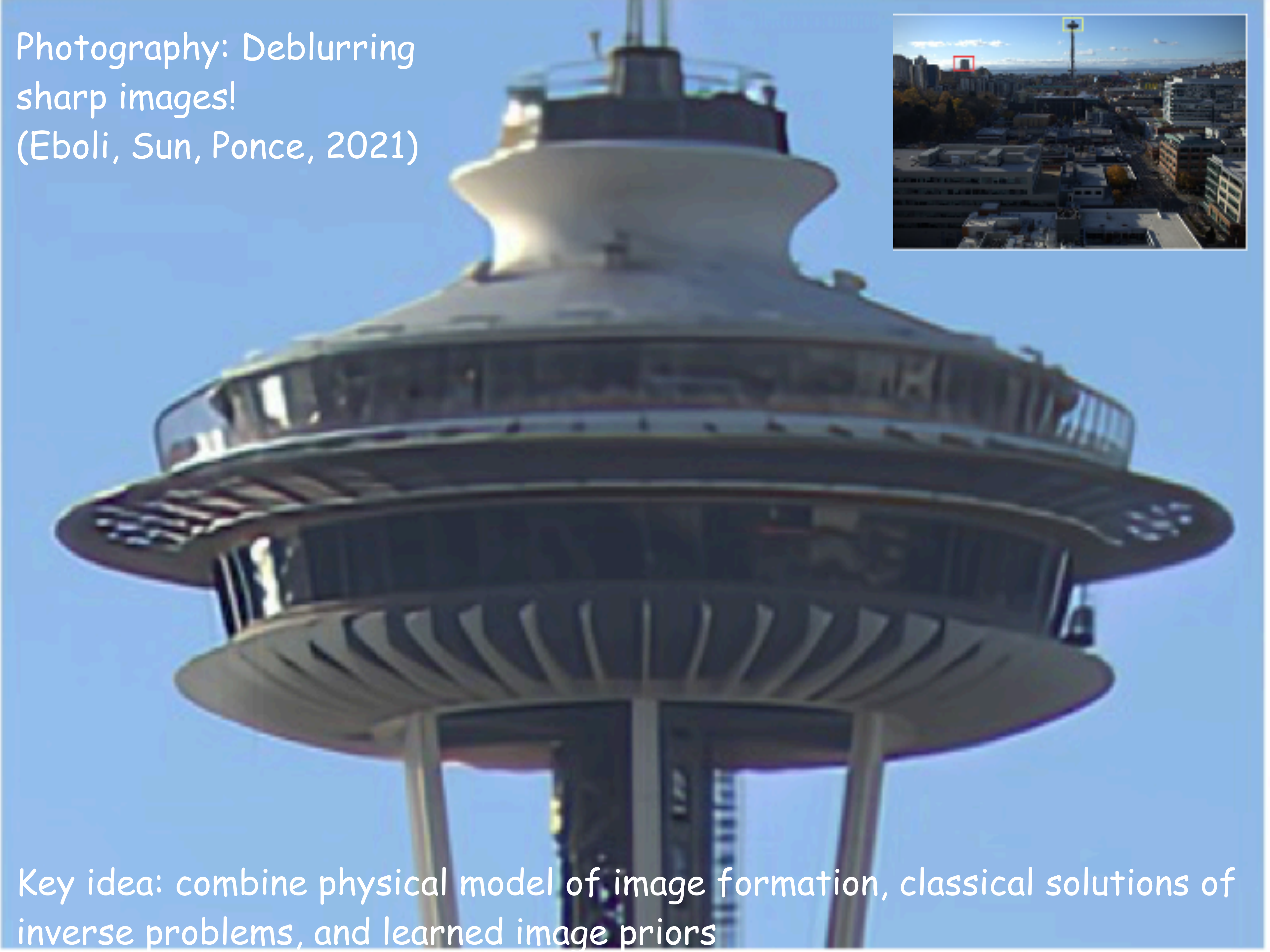


Images courtesy of A. Zisserman.

Outline

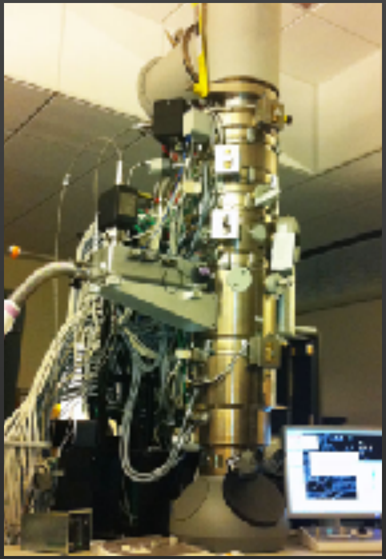
- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry
- Image processing

Photography: Deblurring sharp images!
(Eboli, Sun, Ponce, 2021)



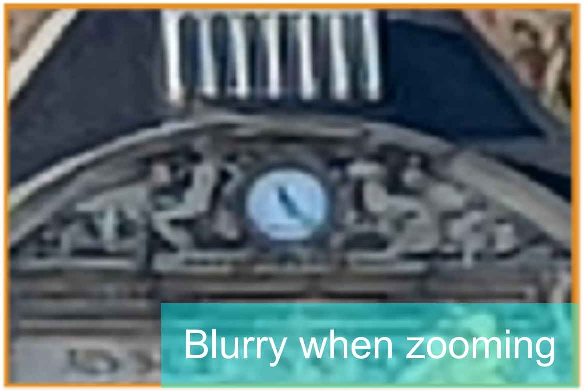
Key idea: combine physical model of image formation, classical solutions of inverse problems, and learned image priors

Image restoration and why



For example, smartphone cameras are great, but..

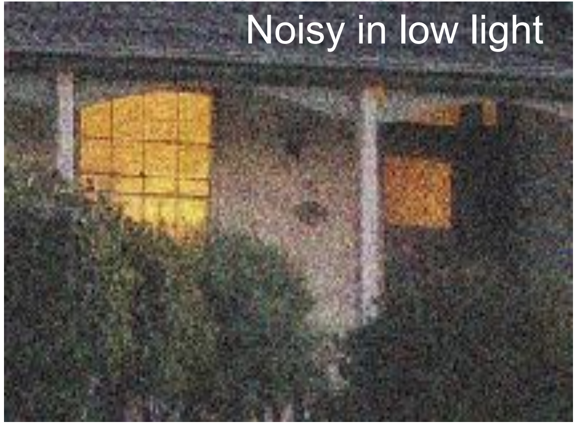
Google Pixel 6, long range tele



Blurry when zooming

Google Pixel 6, crop: limited detail at long range

Poor dynamic range



Noisy in low light

Small aperture and sensor

4.3 mm^2

1900 mm^2

Why machine learning for image restoration?

Reasonable physical models of image corruption

- For example: $y=A(x)+\varepsilon$

- For example: $A(x) = k * x$

➢ One can use prior knowldege

- For example: sparsity, self similarities

➢ Realistic simulated training examples

➢ Interpretable, “functional” architectures

Why machine learning for image restoration?

Reasonable physical models of image corruption

- For example: $y=A(x)+\varepsilon$

- For example: $A(x) = k * x$

➢ One can use prior knowldege

- For example: sparsity, self similarities

➢ Realistic simulated training examples

➢ Interpretable, “functional” architectures

Why machine learning for image restoration?

Reasonable physical models of image corruption

- For example: $y=A(x)+\varepsilon$
- For example: $A(x) = k * x$

➢ One can use prior knowldege

- For example: sparsity, self similarities

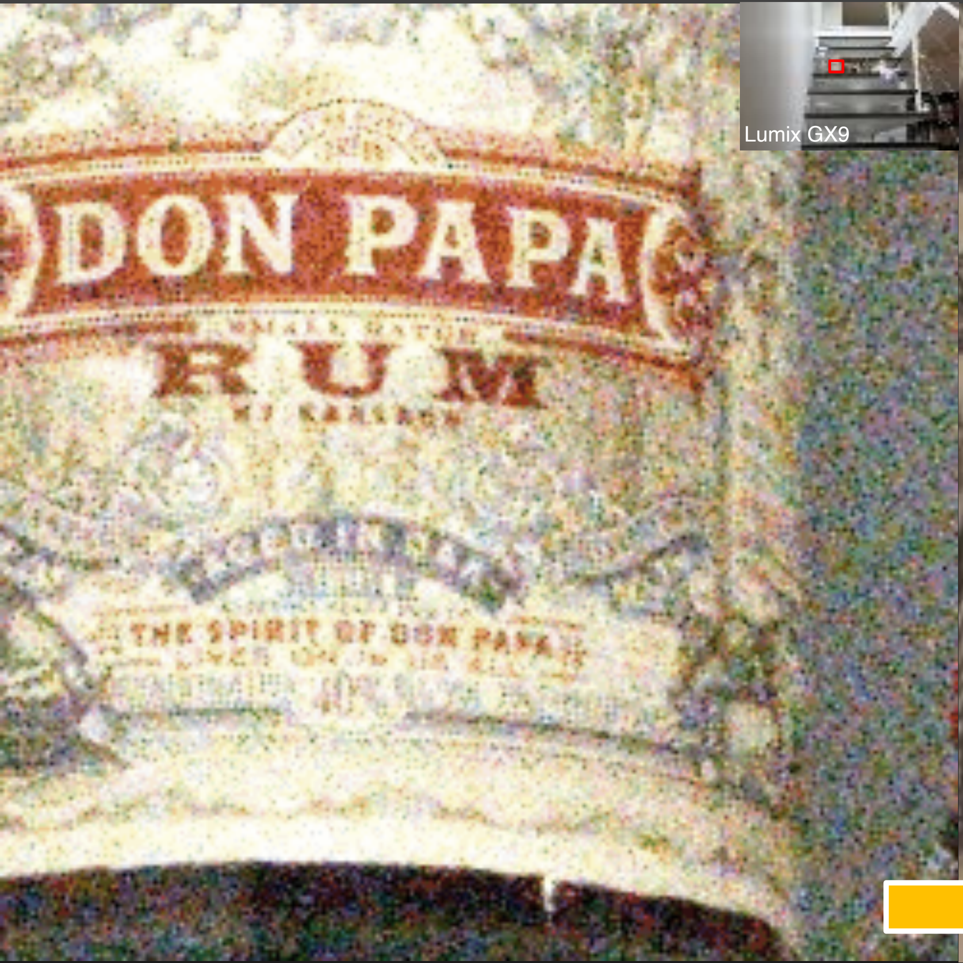
➢ Realistic simulated training examples

➢ Interpretable, “functional” architectures

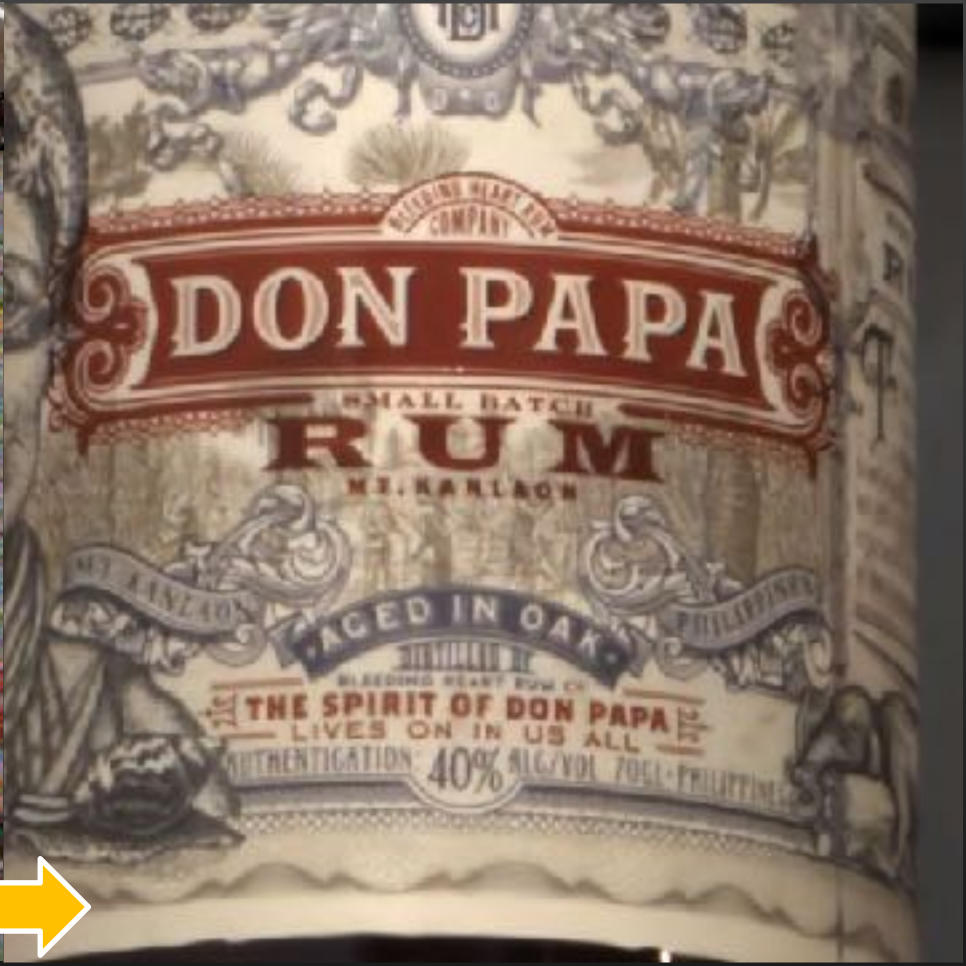
But where does the real ground truth come from, whether for model-based or data-driven methods?

Super-resolution from raw image bursts

Real images, x 4, 12,800 to 25,600 ISO

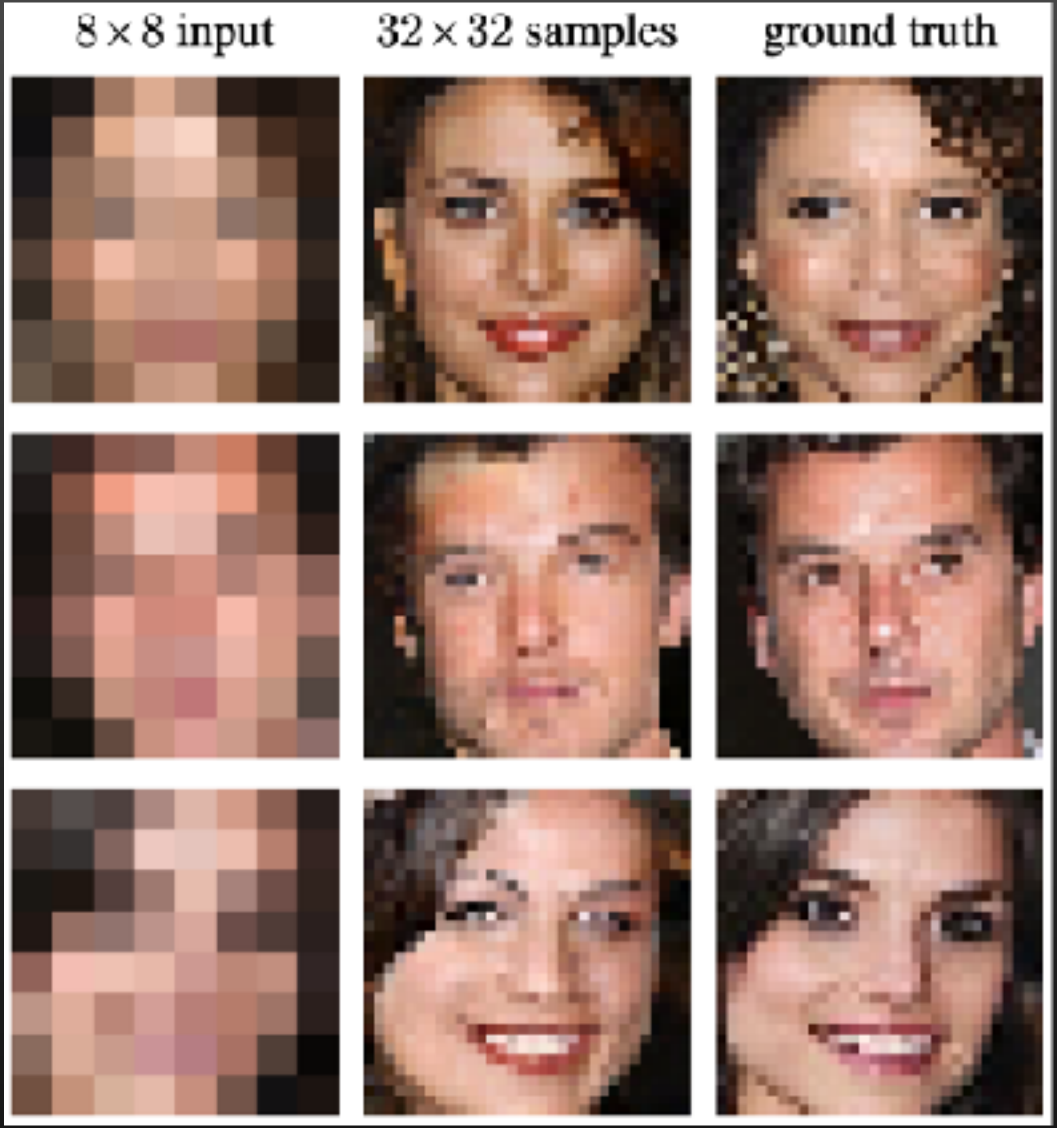


(Small crop of) Burst of raw pictures



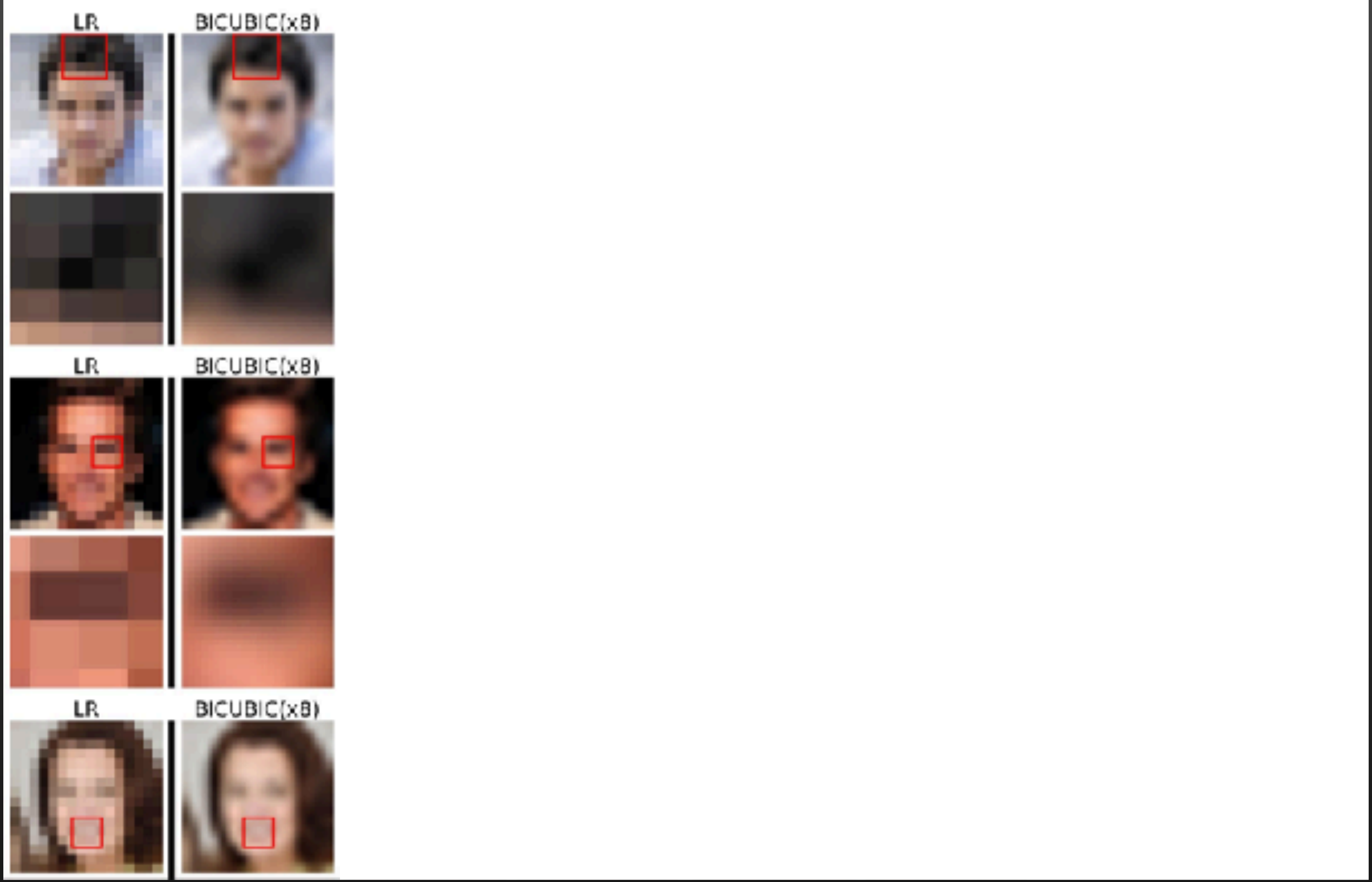
(Lecouat et al., ICCV’21)

Image interpolation (aka "single-image super-resolution")

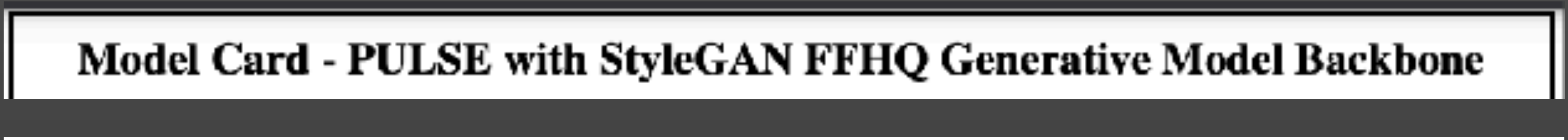


(Dahl et al., 2017)

Image interpolation (aka “single-image super-resolution”)

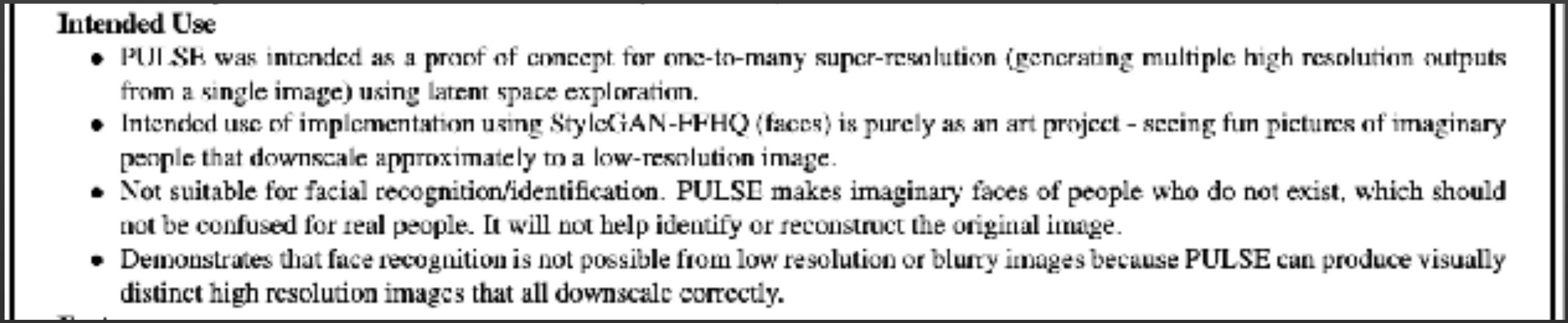


(PULSE, Menon et al., 2020)



Model Card - PULSE with StyleGAN FFHQ Generative Model Backbone

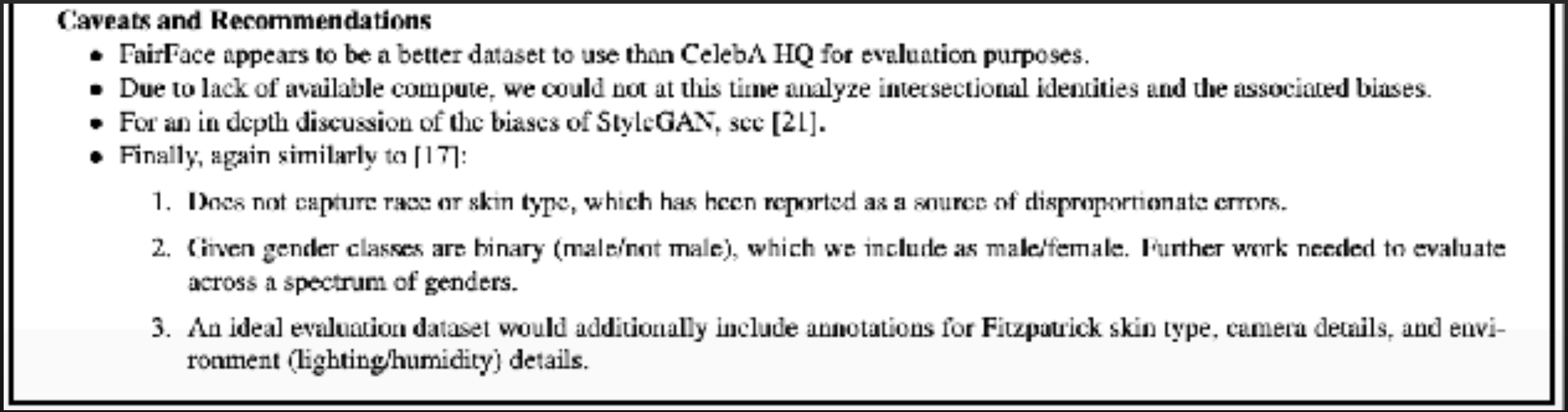
...



Intended Use

- PULSE was intended as a proof of concept for one-to-many super-resolution (generating multiple high resolution outputs from a single image) using latent space exploration.
- Intended use of implementation using StyleGAN-FFHQ (faces) is purely as an art project - seeing fun pictures of imaginary people that downscale approximately to a low-resolution image.
- Not suitable for facial recognition/identification. PULSE makes imaginary faces of people who do not exist, which should not be confused for real people. It will not help identify or reconstruct the original image.
- Demonstrates that face recognition is not possible from low resolution or blurry images because PULSE can produce visually distinct high resolution images that all downscale correctly.

...

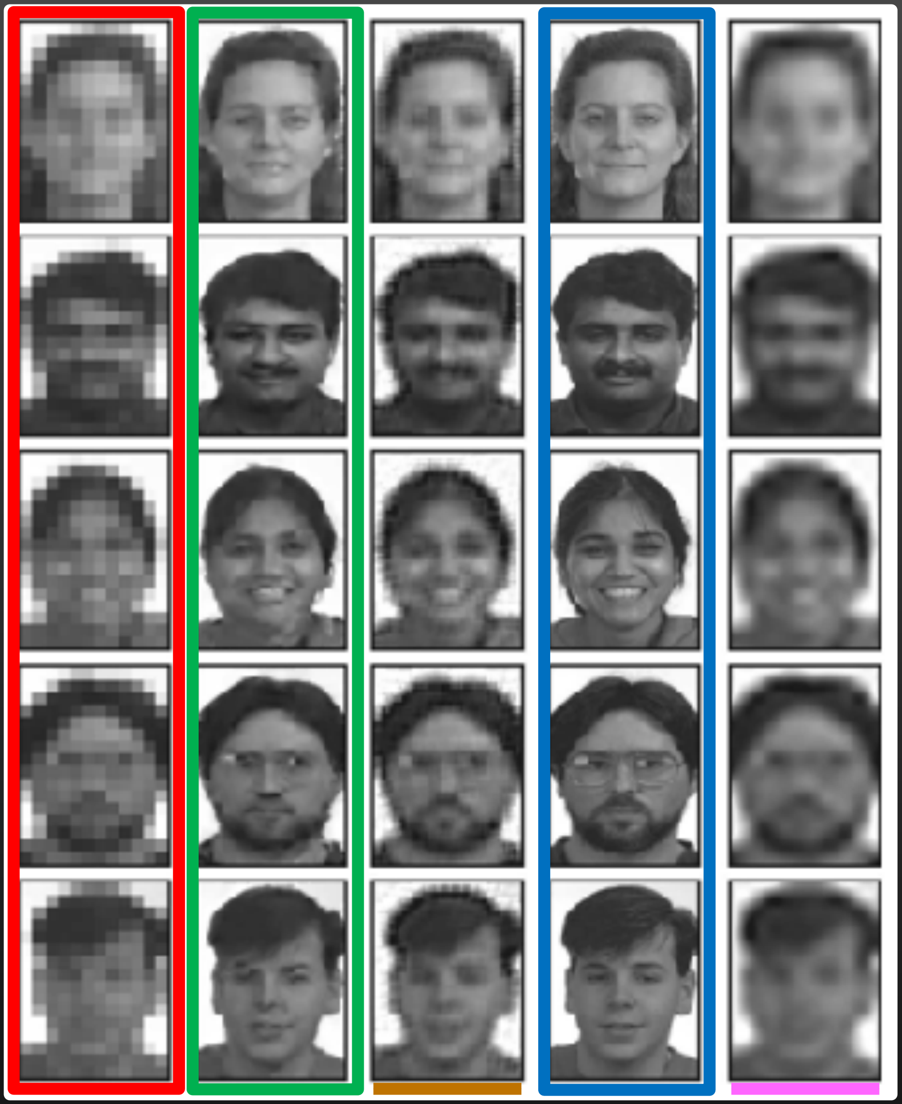


Caveats and Recommendations

- FairFace appears to be a better dataset to use than CelebA HQ for evaluation purposes.
- Due to lack of available compute, we could not at this time analyze intersectional identities and the associated biases.
- For an in depth discussion of the biases of StyleGAN, see [21].
- Finally, again similarly to [17]:

  1. Does not capture race or skin type, which has been reported as a source of disproportionate errors.
  2. Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
  3. An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Super-resolution with “hallucination/recogstruction”



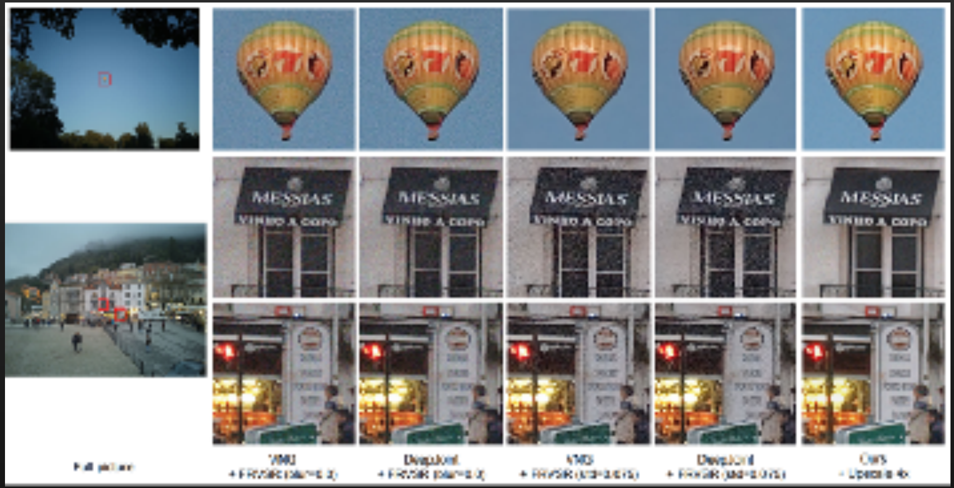
- LR input image (1 of 4)
- Recogstruction
- Ground-truth HR image
- (Hardie et al., 1997)
- Bicubic interpolation

× 4, alignment known exactly

(Baker and Kanade, 2002)

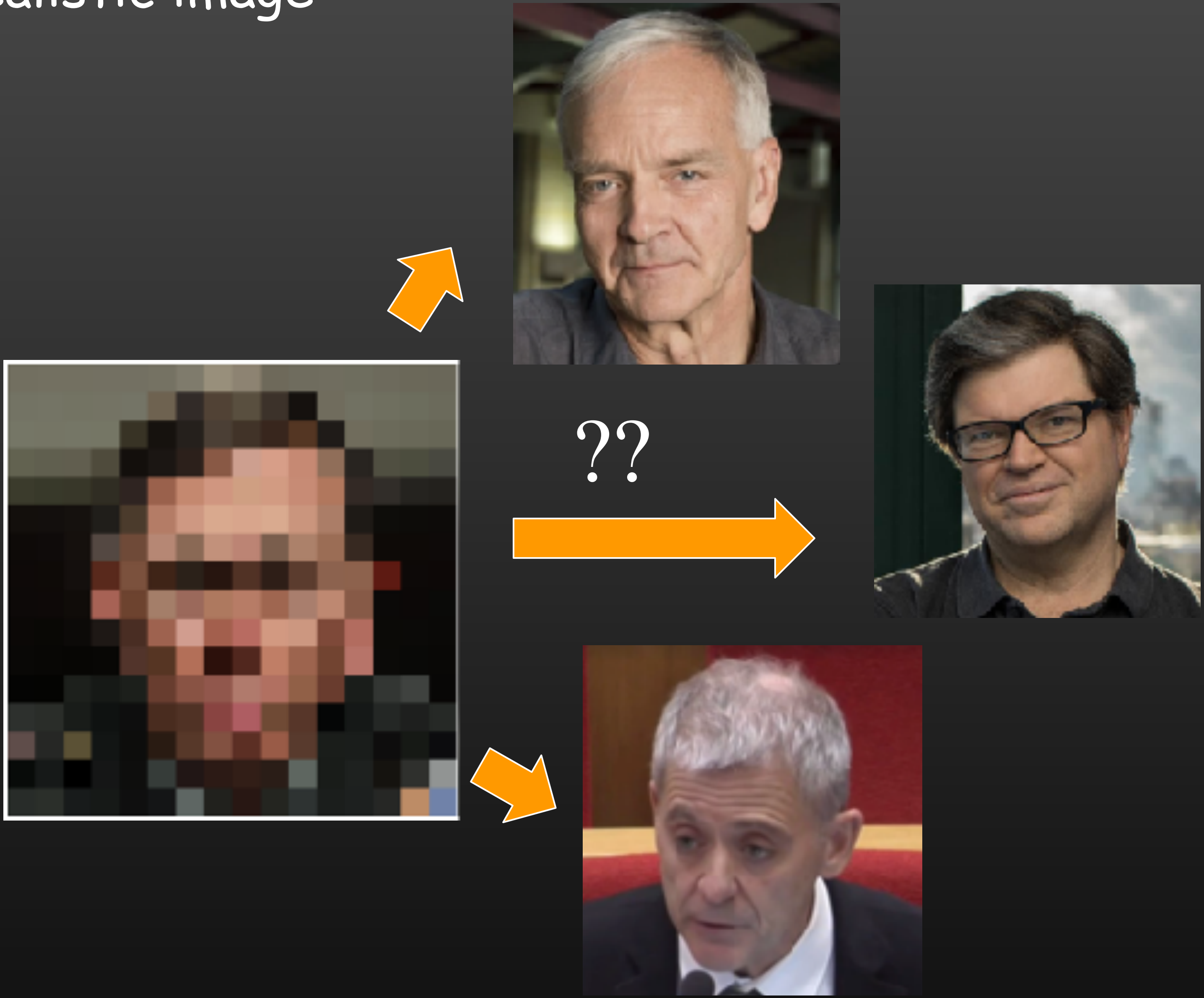
True (multi-frame) super-resolution

(Irani & Peleg, 1991)

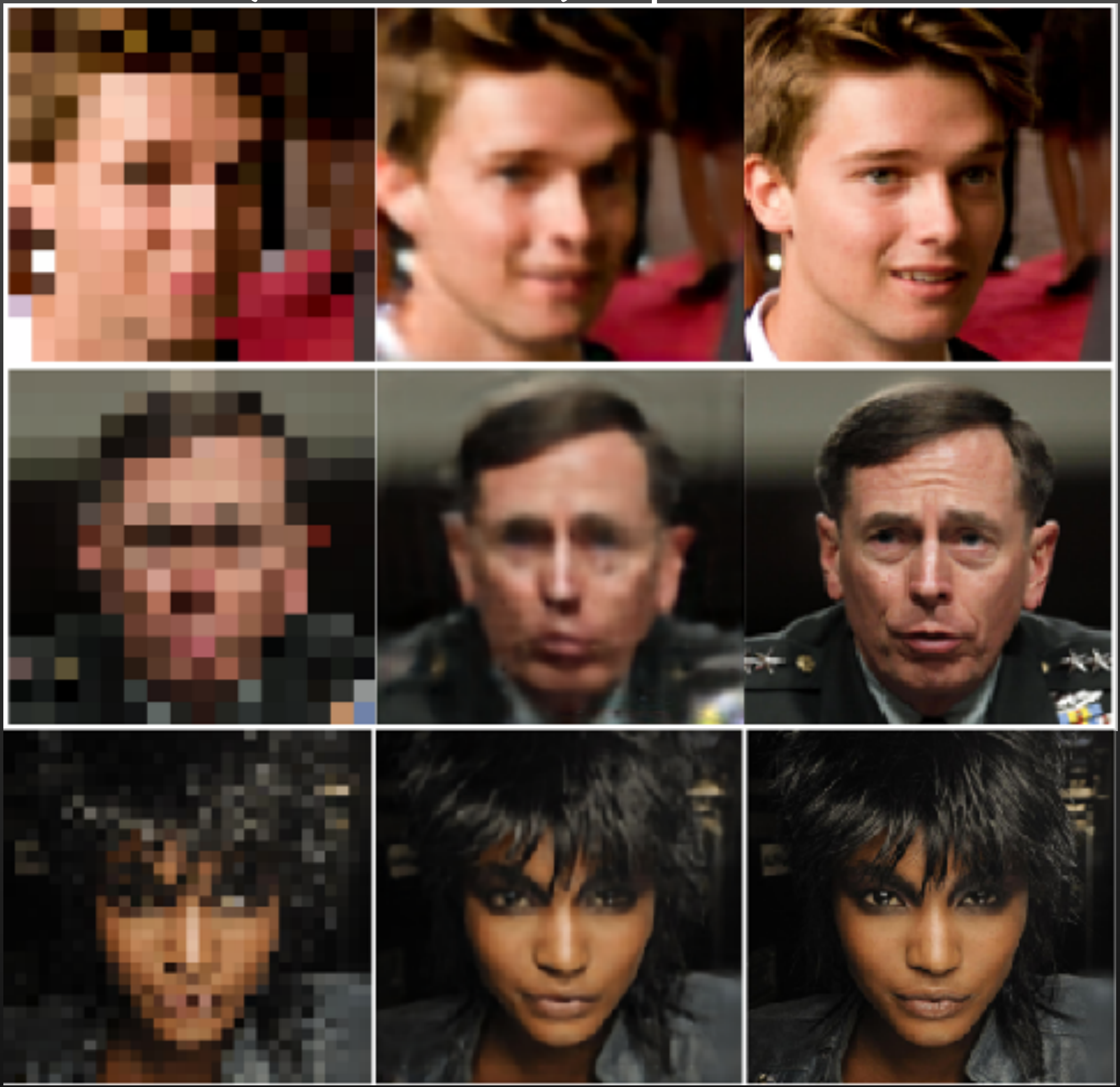


(Wronski et al., 2019)

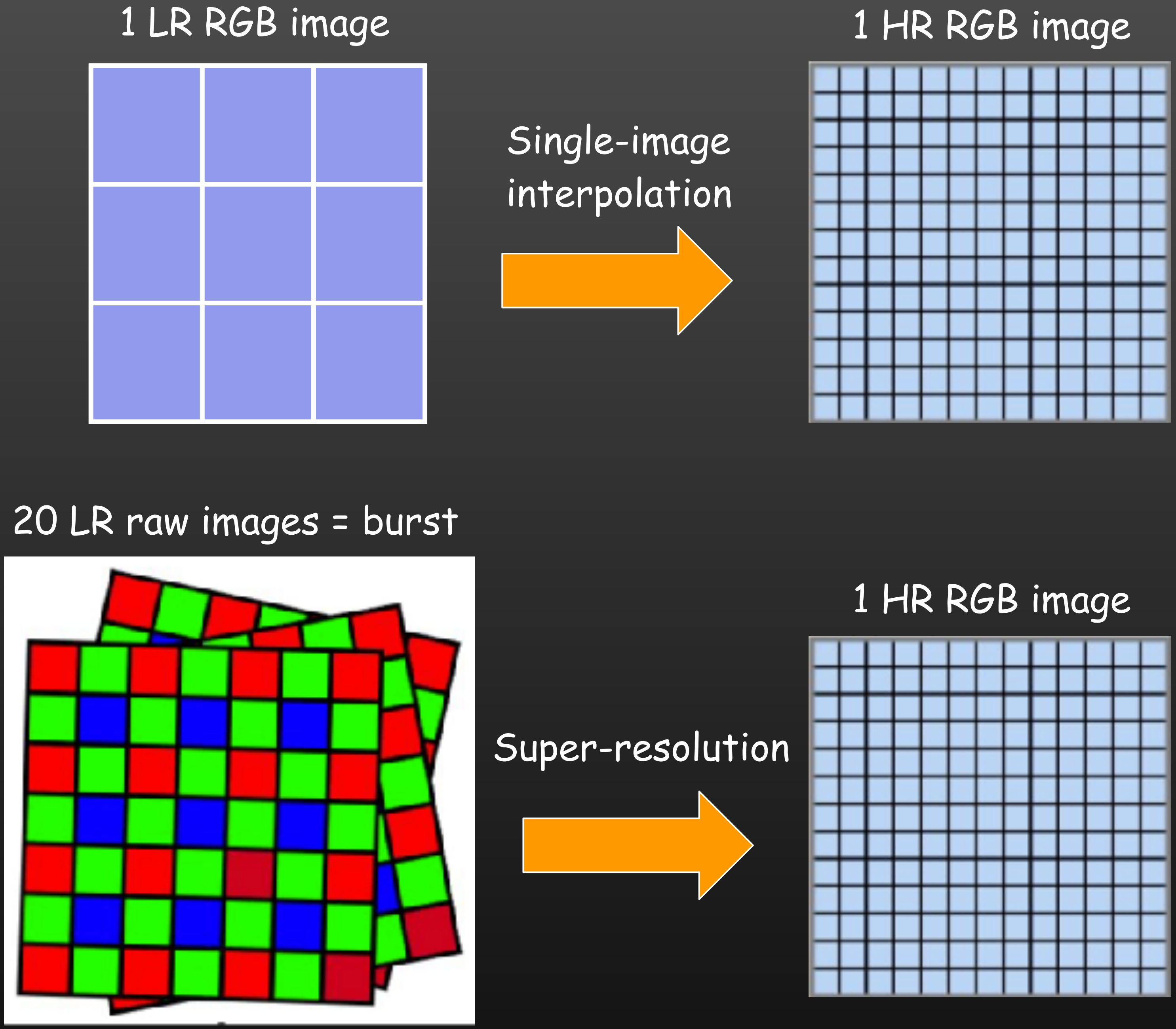
An image interpolation algorithm guaranteed to yield a 100% realistic image



True (multi-frame) super-resolution



(x16 super-resolution on synthetic data)



1 LR RGB image

Single-image interpolation

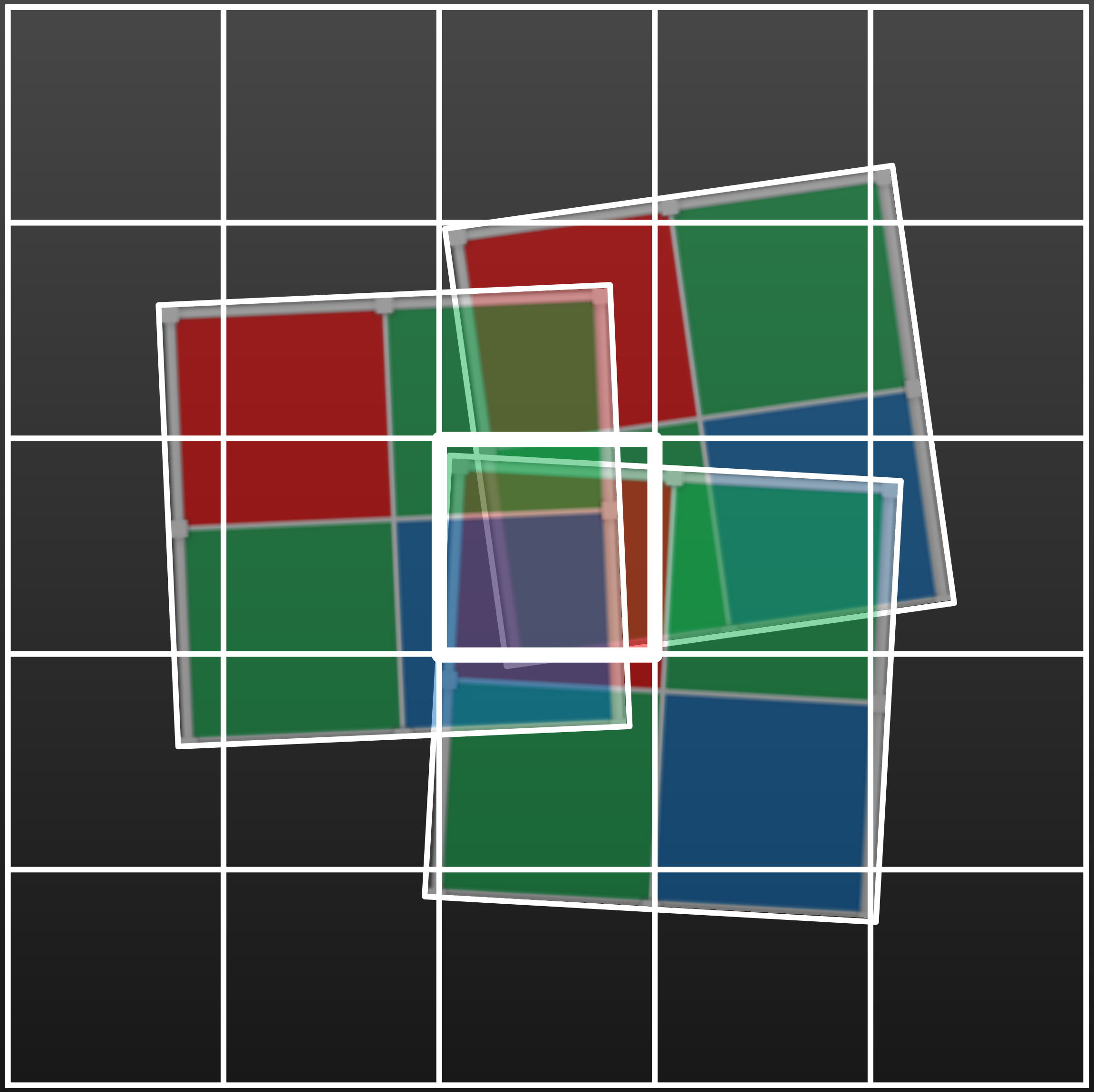
1 HR RGB image

20 LR raw images = burst

Super-resolution

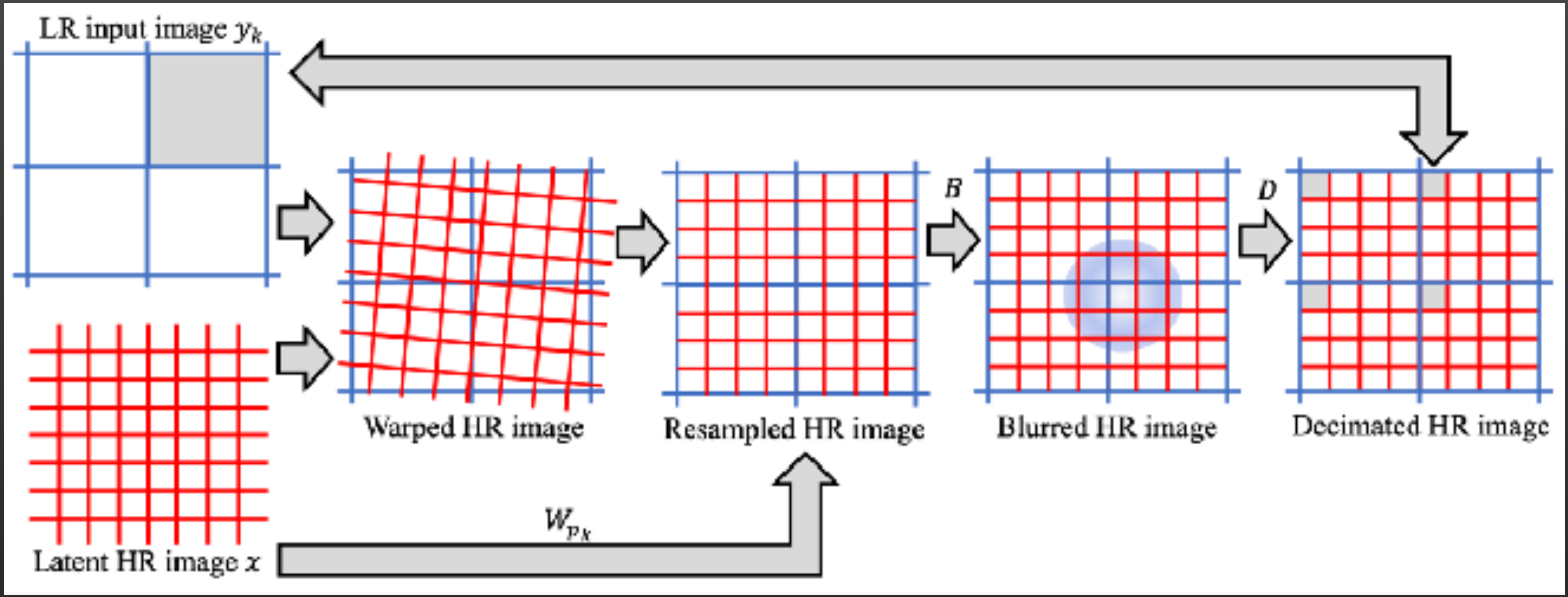
1 HR RGB image

Helps with demosaicking.. (Wronski et al., 2019)



..and denoising too (a la Buades et al., 2005)

Comment ça marche !



- $\boxed{y_k = U_k\, x + \varepsilon_k \ \text{ for } k = 1,\dots, \ K \ \text{ with } \ U_{p_k} = DBW_{p_k}}$ *Physical model*
- Define $x_\theta(y) = \text{argmin}_{x,p} \frac{1}{2} \| y - U_p\, x \|^2 + \boxed{\lambda\varphi_\theta(x)}$ *Learned prior*
- Minimize wrt θ the objective $\frac{1}{n} \sum \| x_i - x_\theta(y_i) \|^1$

Optimization: unrolled iterative algorithm

$$\min_{\mathbf{x},\mathbf{p}} \frac{1}{2}\|\mathbf{y} - U_{\mathbf{p}}\,\mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x})$$

$$\min_{\mathbf{x},\mathbf{p},\mathbf{z}} E_\mu(\mathbf{x},\mathbf{z},\mathbf{p}) = \frac{1}{2}\|\mathbf{y} - U_{\mathbf{p}}\,\mathbf{z}\|^2 + \frac{\mu}{2}\|\mathbf{z} - \mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x})$$

Quadratic penalty (aka HQS) method (three iterations)

- $\mathbf{z}^t \leftarrow \mathbf{z}^{t-1} - \eta_t\left[U^\top_{\mathbf{p}^{t-1}}(U_{\mathbf{p}^{t-1}}\mathbf{z}^{t-1} - \mathbf{y}) + \mu(\mathbf{z}^{t-1} - \mathbf{x}^{t-1})\right]$ — One step of gradient descent (or a few)
- $\min_{\mathbf{p}_k} \frac{1}{2}\|\mathbf{y}_k - DBW_{\mathbf{p}_k}\mathbf{z}^t\|^2$ — Gauss-Newton (aka Lucas-Kanade)
- $\mathbf{x}^t \leftarrow \arg\min_{\mathbf{x}} \frac{\mu_{t-1}}{2}\|\mathbf{z}^t - \mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x})$ — Proximal update
- Increment μ

K. Gregor, Y. LeCun, “Learning fast approximations of sparse coding”, ICML’10

Optimization: unrolled iterative algorithm

$$\min_{\mathbf{x},\mathbf{p}} \ \frac{1}{2}\|\mathbf{y} - U_{\mathbf{p}}\,\mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x})$$

$$\min_{\mathbf{x},\mathbf{p},\mathbf{z}} \ E_\mu(\mathbf{x},\mathbf{z},\mathbf{p}) = \frac{1}{2}\|\mathbf{y} - U_{\mathbf{p}}\,\mathbf{z}\|^2 + \frac{\mu}{2}\|\mathbf{z} - \mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x})$$

Quadratic penalty (aka HQS) method (three iterations)

- ➢ $\mathbf{z}^t \leftarrow \mathbf{z}^{t-1} - \eta_t\left[U_{\mathbf{p}^{t-1}}^\top(U_{\mathbf{p}^{t-1}}\mathbf{z}^{t-1} - \mathbf{y}) + \mu(\mathbf{z}^{t-1} - \mathbf{x}^{t-1})\right]$ — One step of gradient descent (or a few)
- ➢ $\mathbf{p}_k^t \leftarrow \mathbf{p}_k^{t-1} - \left(\mathbf{J}_k^{t\top}\mathbf{J}_k^t\right)^{-1}\mathbf{J}_k^{t\top}\mathbf{r}_k^t$ (3 times) — Gauss-Newton (aka Lucas-Kanade)
- ➢ $\mathbf{x}^t \leftarrow \arg\min_{\mathbf{x}} \frac{\mu_{t-1}}{2}\|\mathbf{z}^t - \mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x})$ — Proximal update
- ➢ Increment μ

K. Gregor, Y. LeCun, “Learning fast approximations of sparse coding”, ICML’10

Optimization: unrolled iterative algorithm

$$\min_{\mathbf{x},\mathbf{p}} \ \frac{1}{2}\|\mathbf{y} - U_{\mathbf{p}}\,\mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x})$$

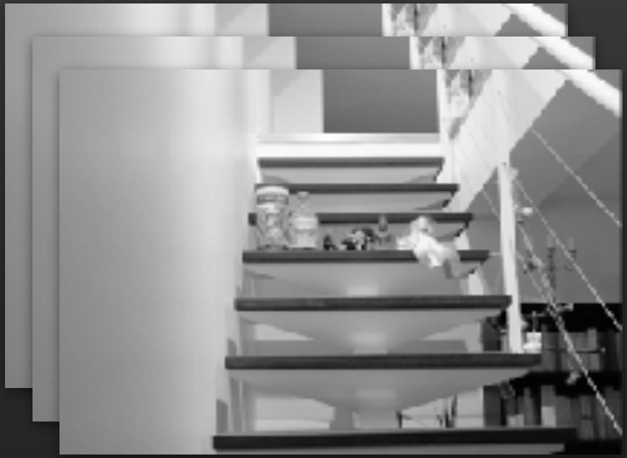
$$\min_{\mathbf{x},\mathbf{p},\mathbf{z}} \ E_\mu(\mathbf{x},\mathbf{z},\mathbf{p}) = \frac{1}{2}\|\mathbf{y} - U_{\mathbf{p}}\,\mathbf{z}\|^2 + \frac{\mu}{2}\|\mathbf{z} - \mathbf{x}\|^2 + \lambda\phi_\theta(\mathbf{x})$$

Quadratic penalty (aka HQS) method (three iterations)

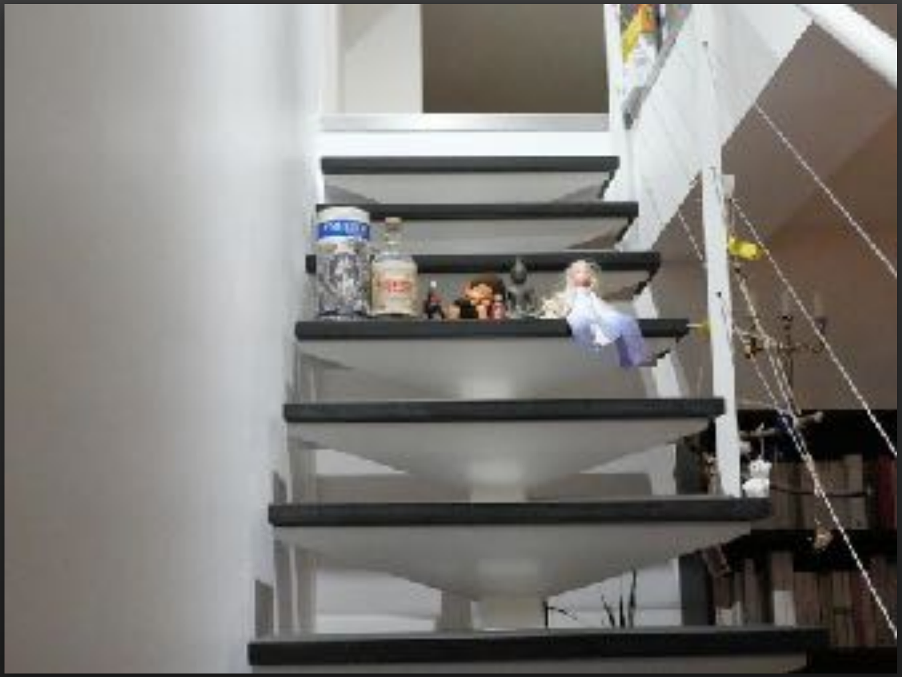
- ➢ $\mathbf{z}^t \leftarrow \mathbf{z}^{t-1} - \eta_t\left[U^\top_{\mathbf{p}^{t-1}}(U_{\mathbf{p}^{t-1}}\mathbf{z}^{t-1} - \mathbf{y}) + \mu(\mathbf{z}^{t-1} - \mathbf{x}^{t-1})\right]$ — One step of gradient descent (or a few)
- ➢ $\mathbf{p}^t_k \leftarrow \mathbf{p}^{t-1}_k - \left(\mathbf{J}^{t\top}_k\mathbf{J}^t_k\right)^{-1}\mathbf{J}^{t\top}_k\mathbf{r}^t_k$ (3 times) — Gauss-Newton (aka Lucas-Kanade)
- ➢ $\mathbf{x}^t \leftarrow f_\theta(\mathbf{z}_t)$ — Plug-and-play approach (small residual U-net)
- ➢ Increment μ

K. Gregor, Y. LeCun, "Learning fast approximations of sparse coding", ICML'10

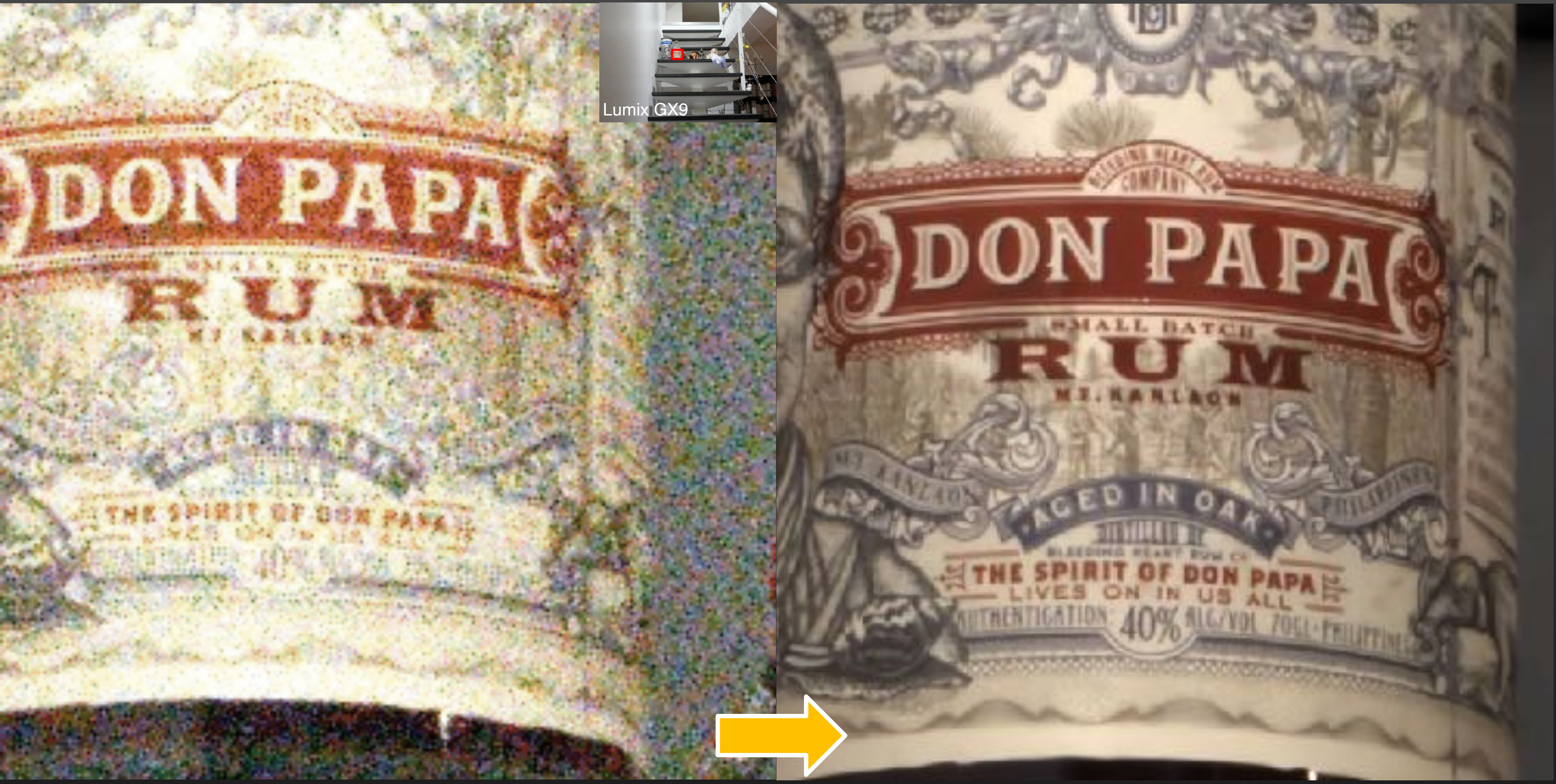
Example



Raw image burst (Lumix GX9)

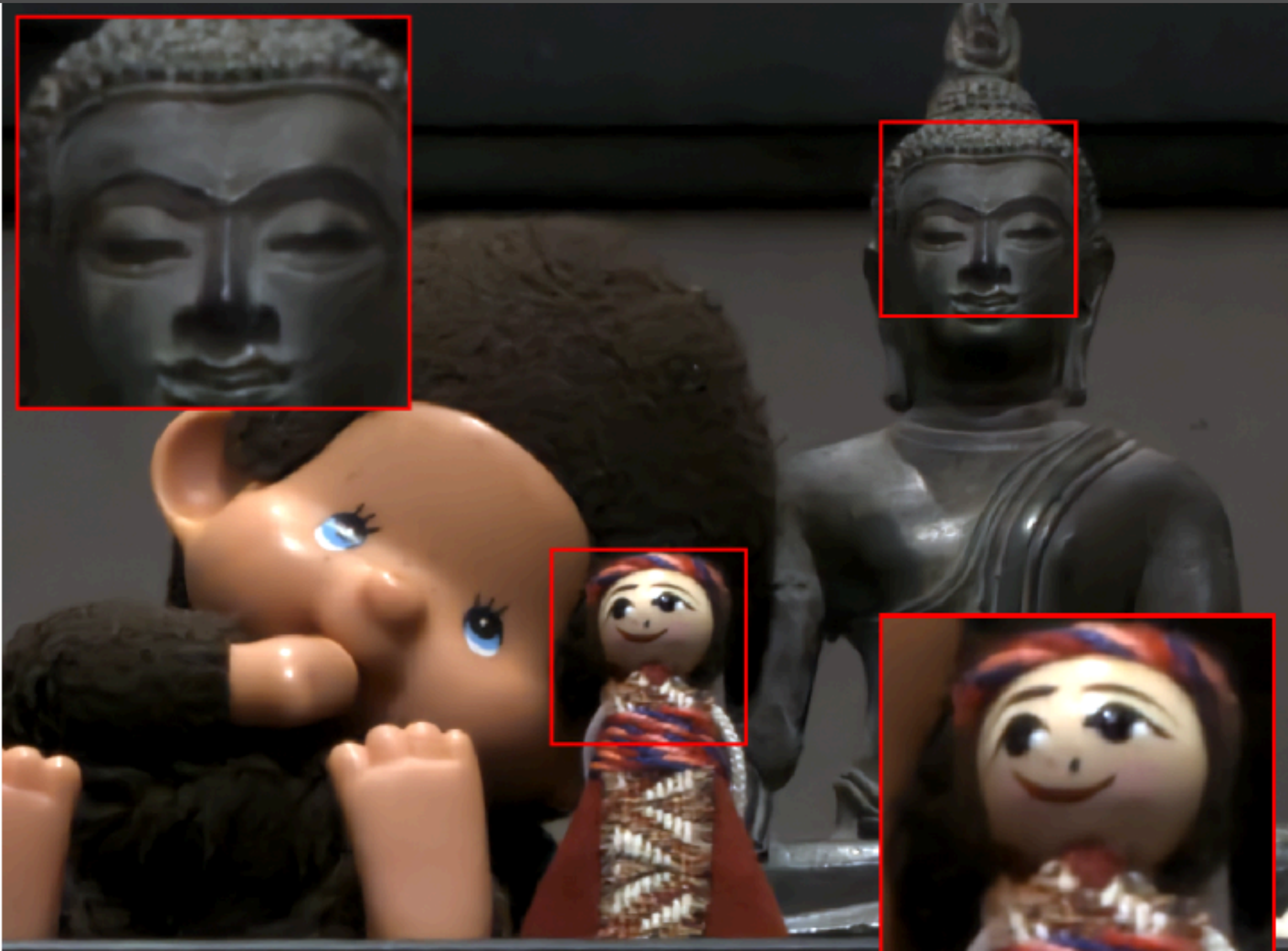


High-quality picture

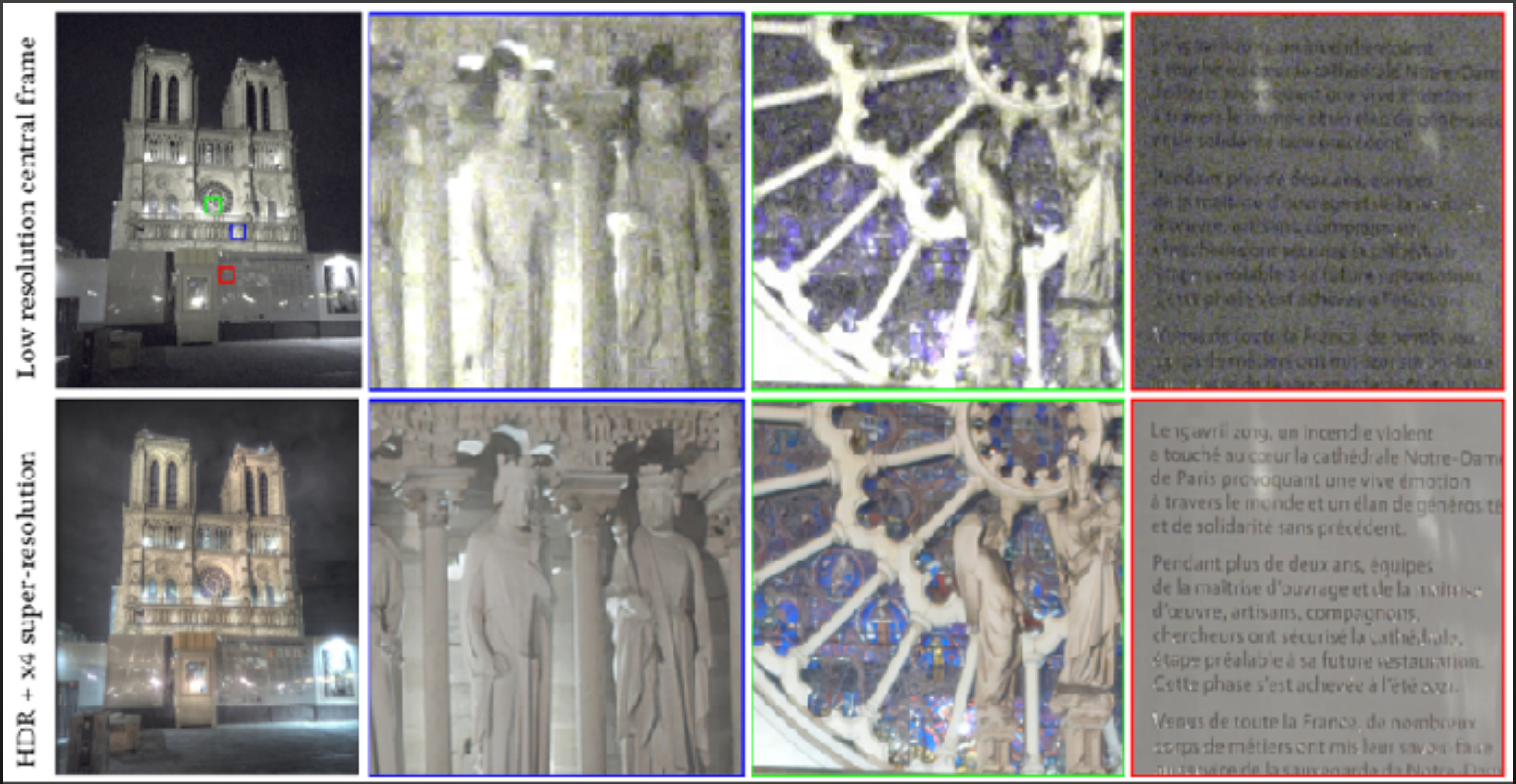


(Small crop of) Burst of raw pictures

(Lecouat et al., ICCV'21)



High-Dynamic Range and Night Imaging (Lecouat et al., SIGGRAPH'22)


Low resolution central frame
HDR + x4 super-resolution
Le 15 avril 2019, un incendie violent
a touché au cœur la cathédrale Notre-Dam
de Paris provoquant une vive émotion
à travers le monde et un élan de généros té
et de solidarité sans précédent.
Pendant plus de deux ans, équipes
de la maîtrise d'ouvrage et de la maîtrise
d'œuvre, artisans, compagnons,
chercheurs ont sécurisé la cathédrale,
étape préalable à sa future restauration.
Cette phase s'est achevée à l'été 2021.
Venus de toute la France, de nombreux
corps de métiers ont mis leur savoir-faire

Google
JAMAICA