TrabHCI Technologies to Reduce the Access Barrier
in Human Computer Interaction
Erasmus Intensive Programme IP29588-1-1731-10

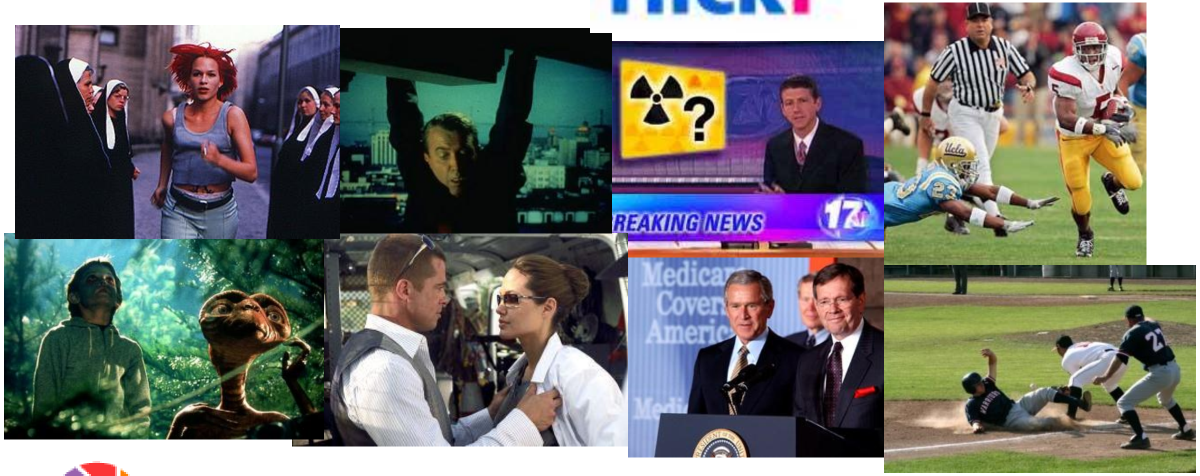
Image and video processing for posture and gesture recognition

Michela Goffredo
University Roma TRE
goffredo@uniroma3.it

Images and video are everywhere!

Personal photo albums

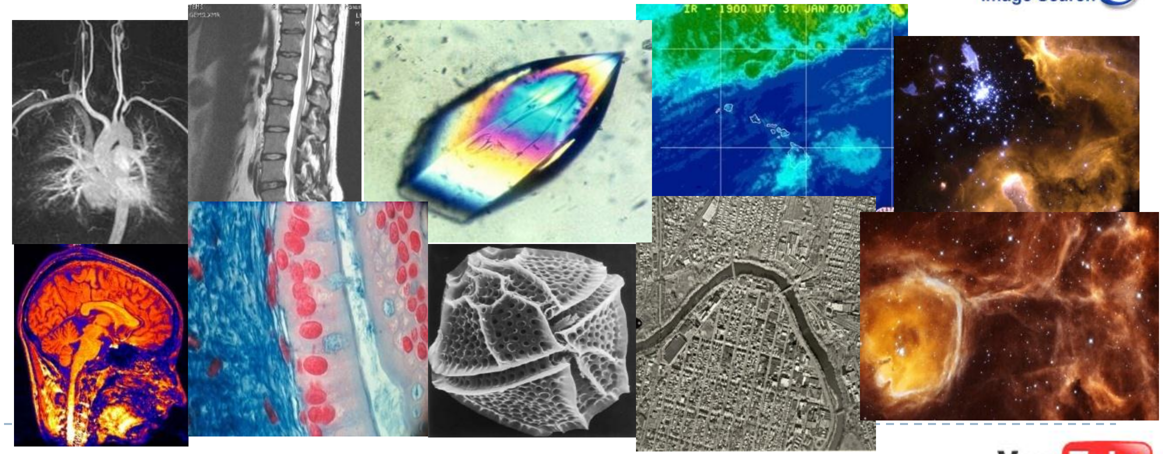
flickr



Picasa™

Movies, news, sports

Google™
Image Search

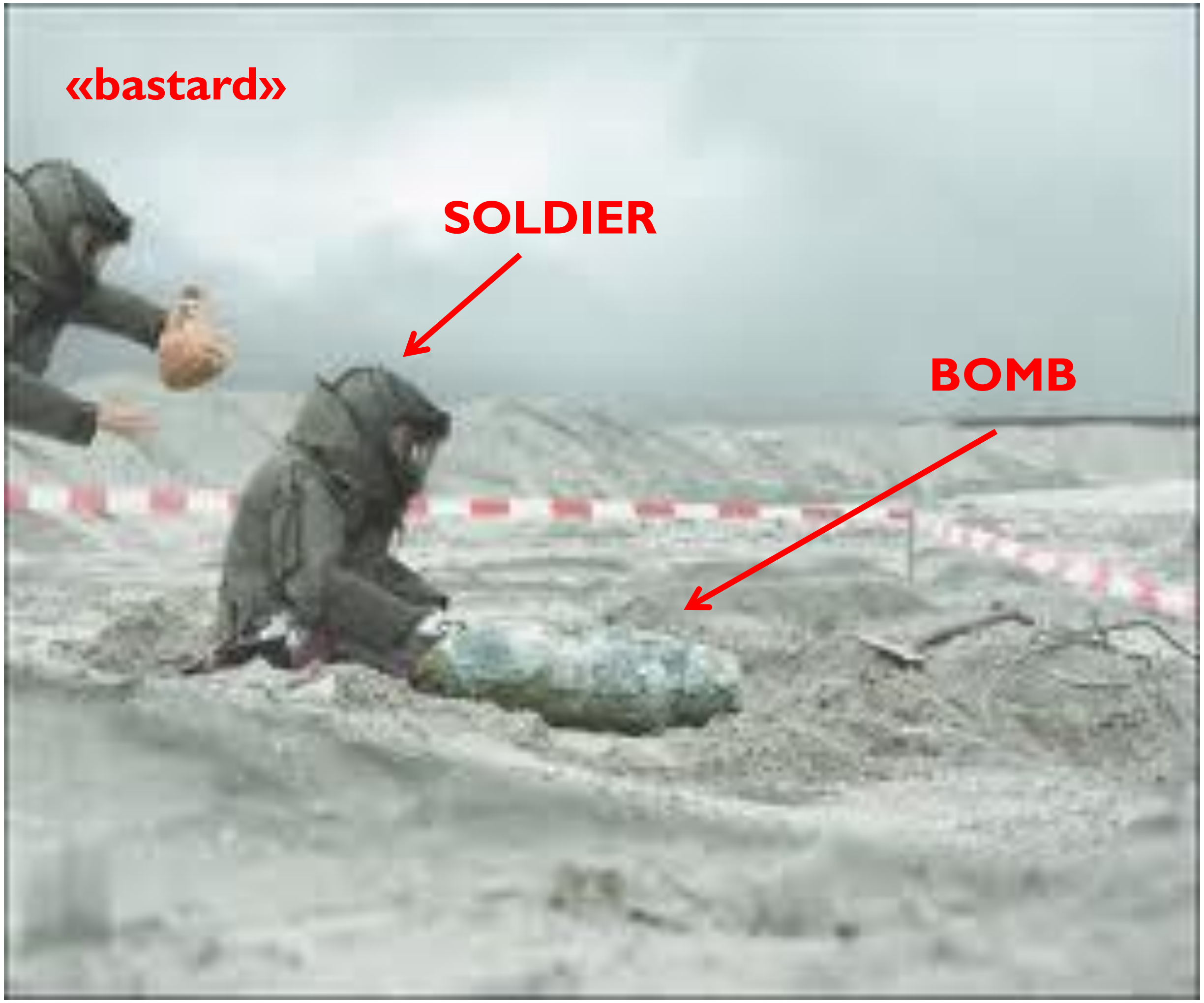


Surveillance and security

Medical and scientific images

YouTube
Broadcast Yourself™

What's Computer Vision?

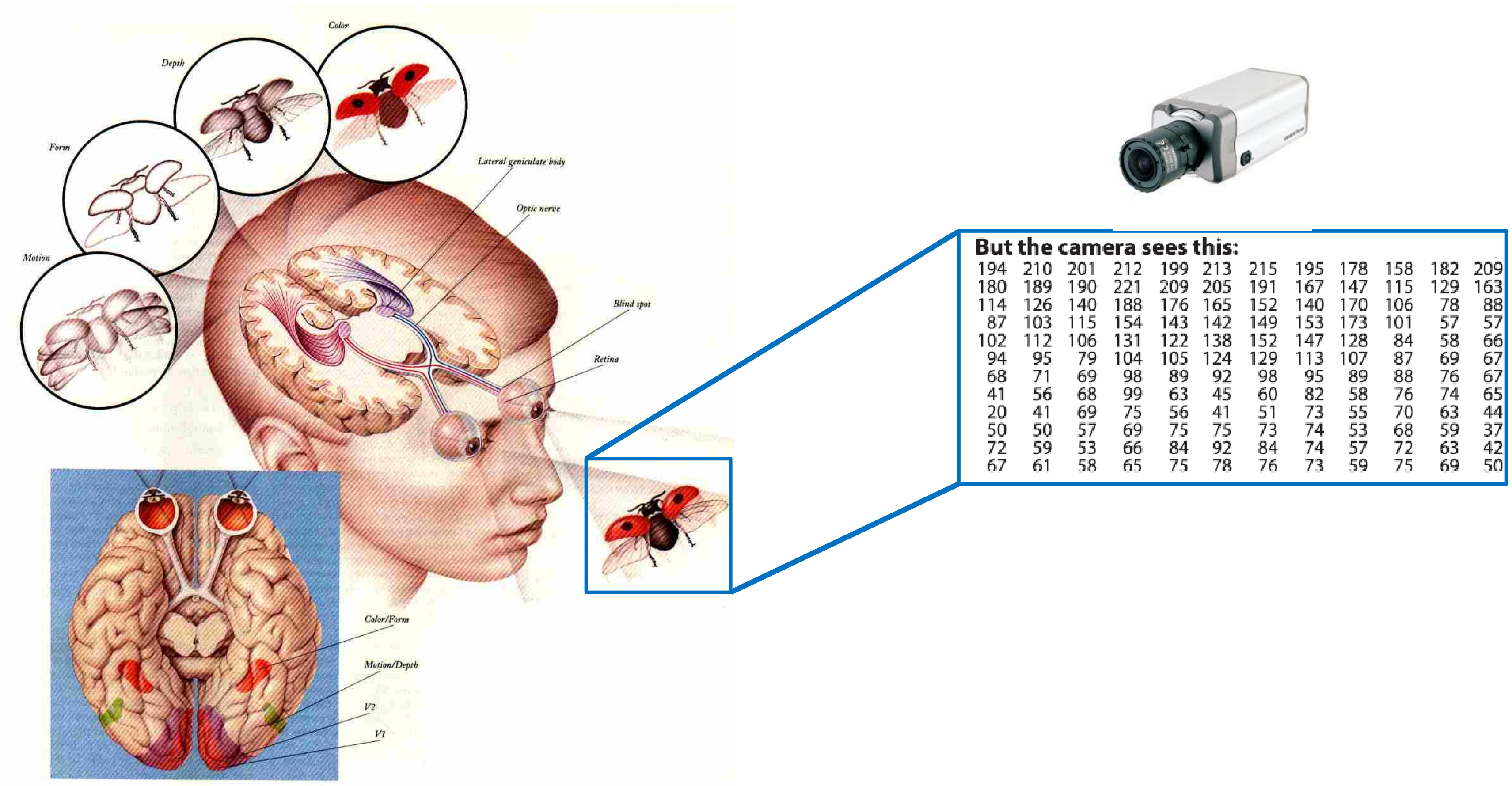


What’s Computer Vision?

- **Data transformation** from a still or video camera into either a decision or a new representation for achieving some particular **goal**.

- The input data may include some **contextual information** such as “the camera is mounted in a car” or “one person is in the scene”.
- The **decision** might be “the person is still” or “there are 5 cars on the road”…

Perceive the “world behind the picture”



But the camera sees this:

194	210	201	212	199	213	215	195	178	158	182	209
180	189	190	221	209	205	191	167	147	115	129	163
114	126	140	188	176	165	152	140	170	106	78	88
87	103	115	154	143	142	149	153	173	101	57	57
102	112	106	131	122	138	152	147	128	84	58	66
94	95	79	104	105	124	129	113	107	87	69	67
68	71	69	98	89	92	98	95	89	88	76	67
41	56	68	99	63	45	60	82	58	76	74	65
20	41	69	75	56	41	51	73	55	70	63	44
50	50	57	69	75	75	73	74	53	68	59	37
72	59	53	66	84	92	84	74	57	72	63	42
67	61	58	65	75	78	76	73	59	75	69	50

In a machine vision system, a computer receives a grid of numbers from the camera or from disk: that’s the digital image.

Perceive the “world behind the picture”

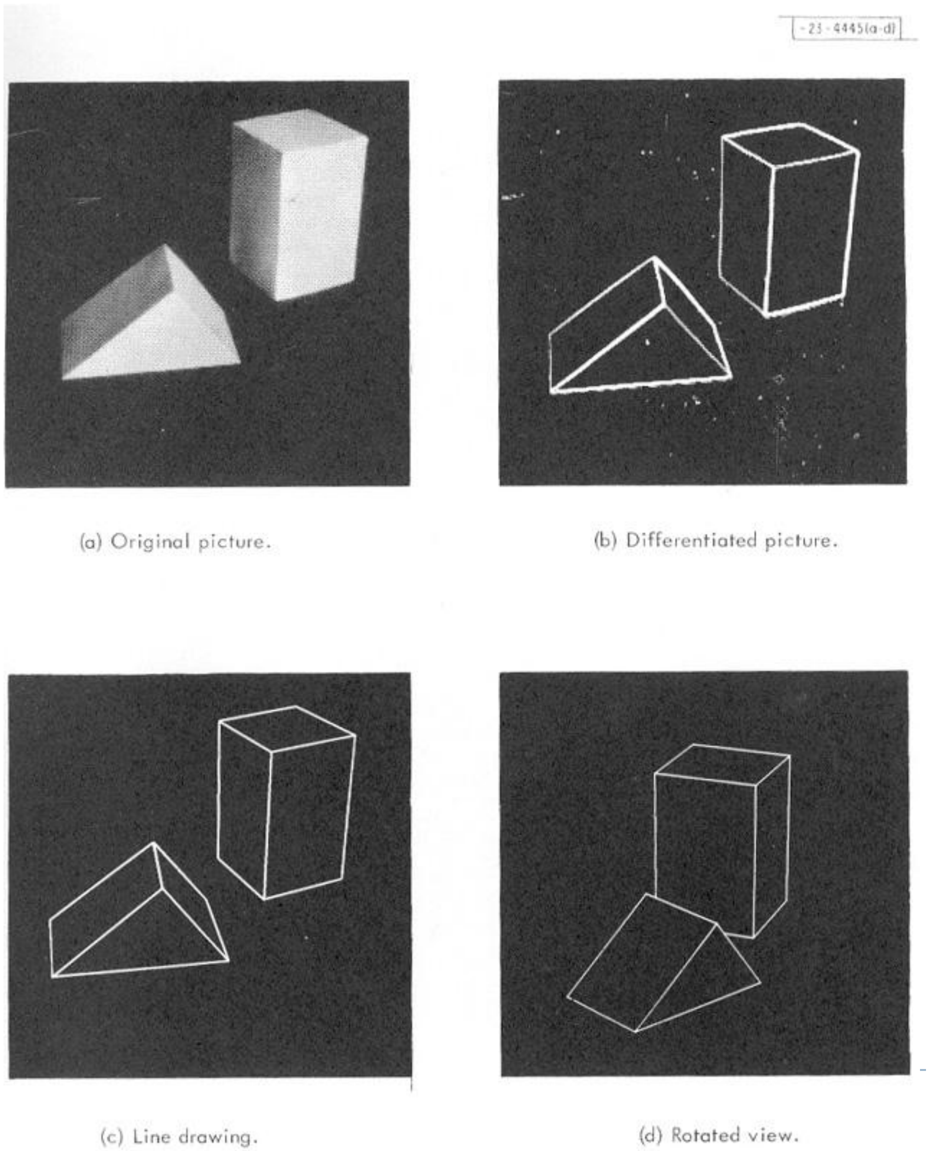
Moreover, **data is corrupted by noise and distortions**:

- from variations in the world (weather, lighting, reflections, movements);
- imperfections in the lens and mechanical setup, finite integration time on the sensor (motion blur);
- electrical noise in the sensor or other electronics;
- compression artifacts…

Additional contextual knowledge can often be used to work around the limitations imposed on us by visual sensors.

General rule: the more constrained a computer vision context is, the more we can rely on those constraints to simplify the problem and the more reliable our final solution will be.

Origins of computer vision



(a) Original picture.

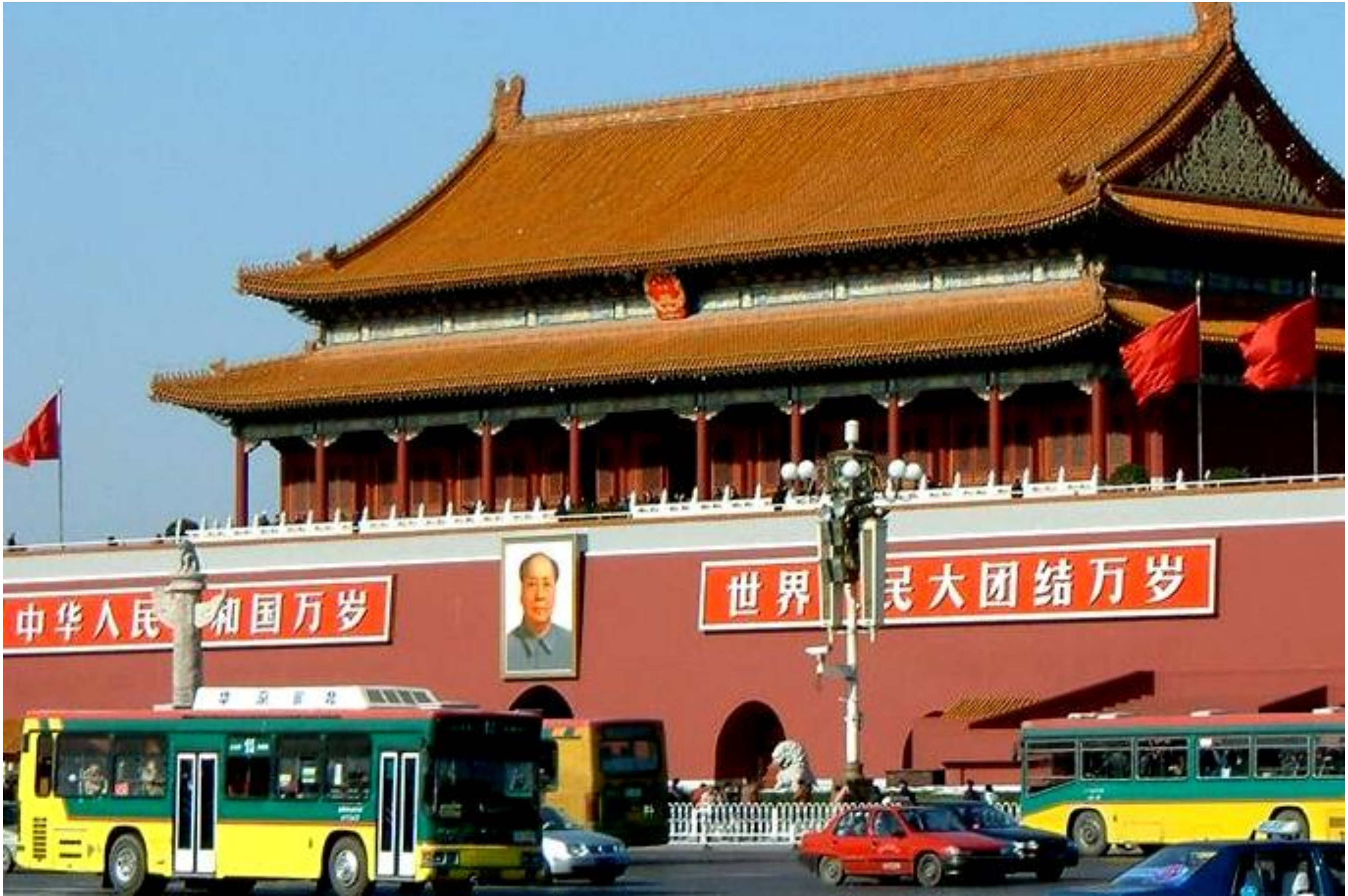
(b) Differentiated picture.

(c) Line drawing.

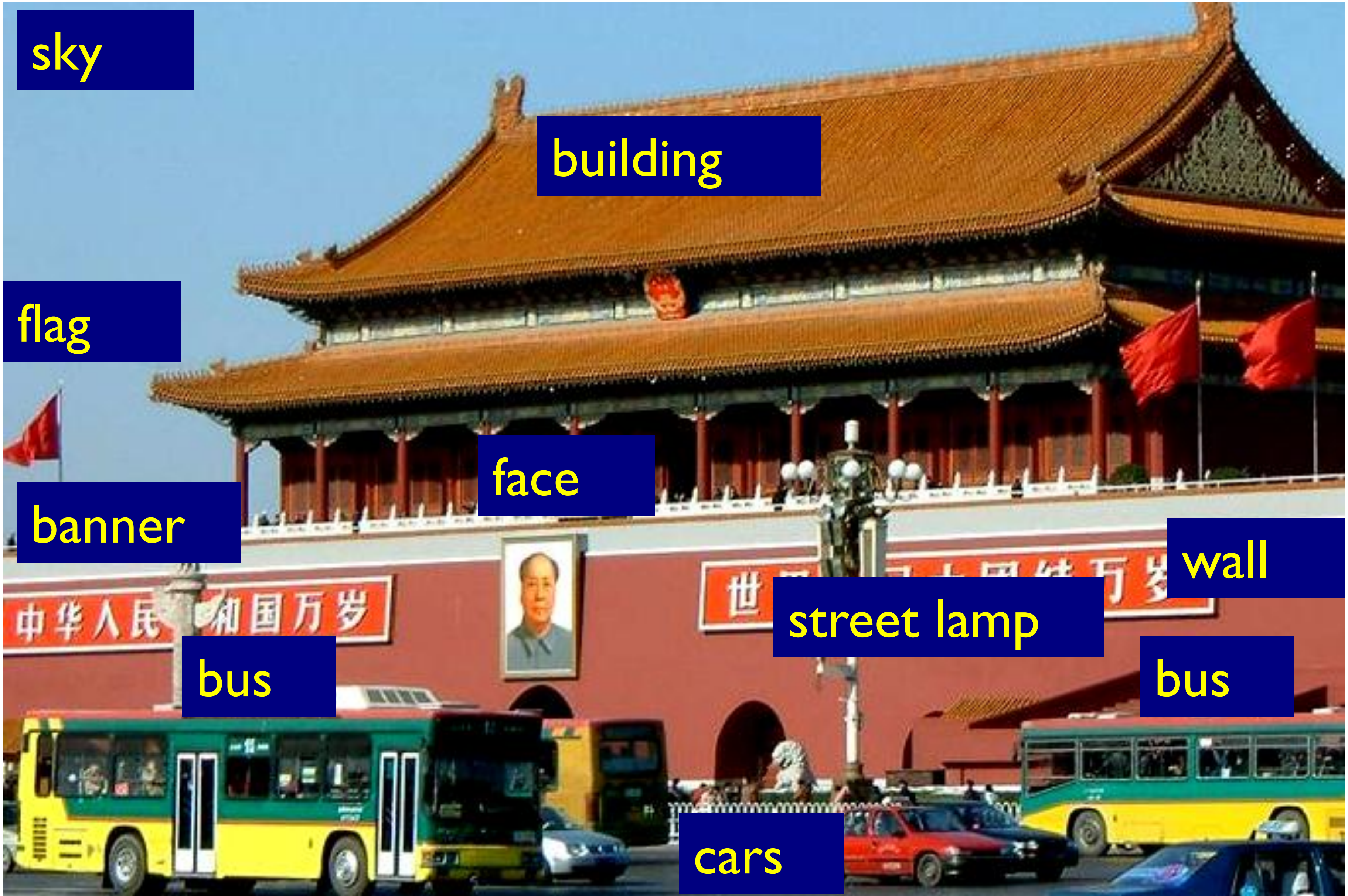
(d) Rotated view.

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

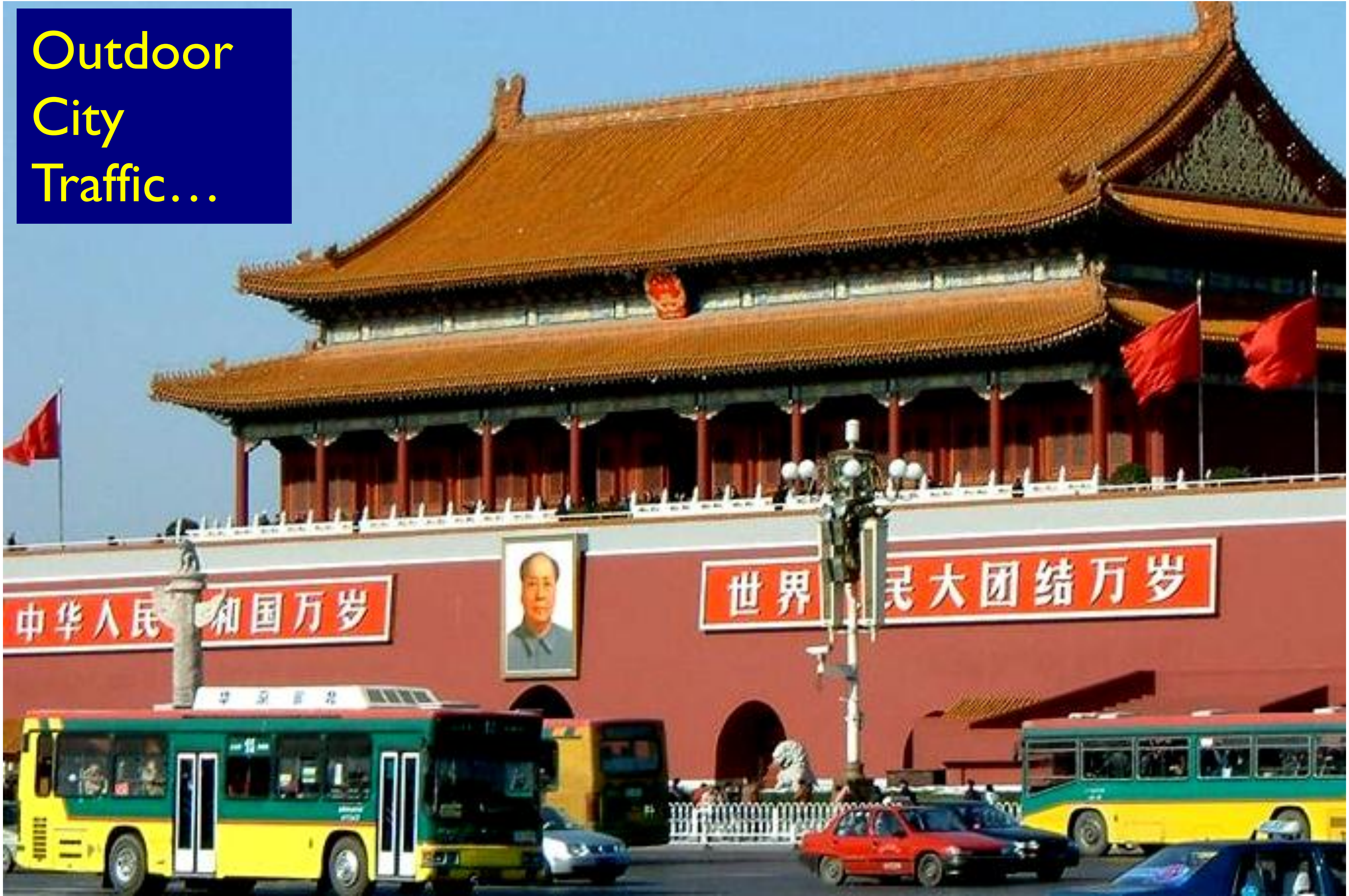
Vision as a source of semantic information



Vision as a source of semantic information

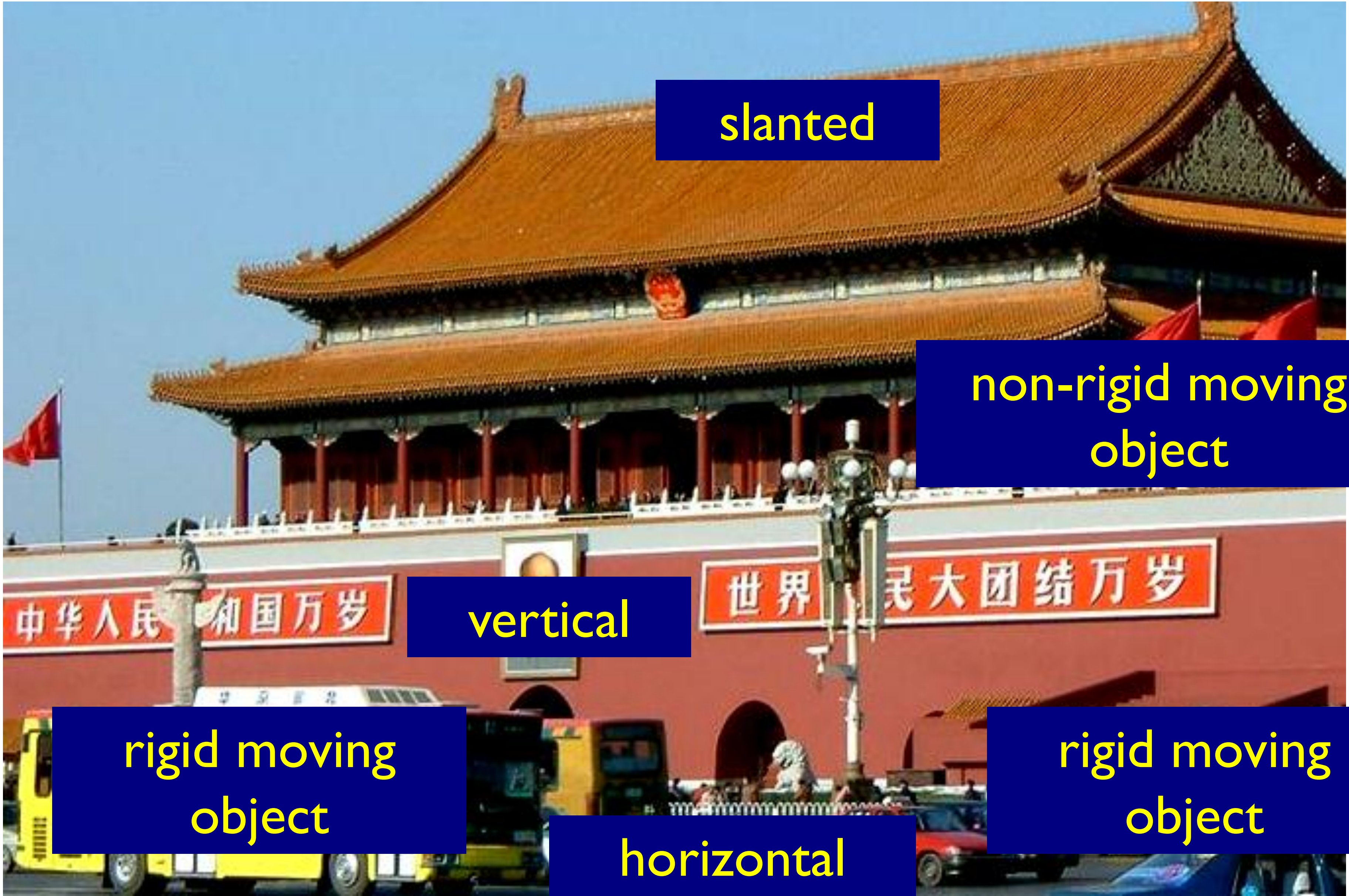


Scene and context categorization



Outdoor
City
Traffic…

Qualitative spatial information



slanted

non-rigid moving object

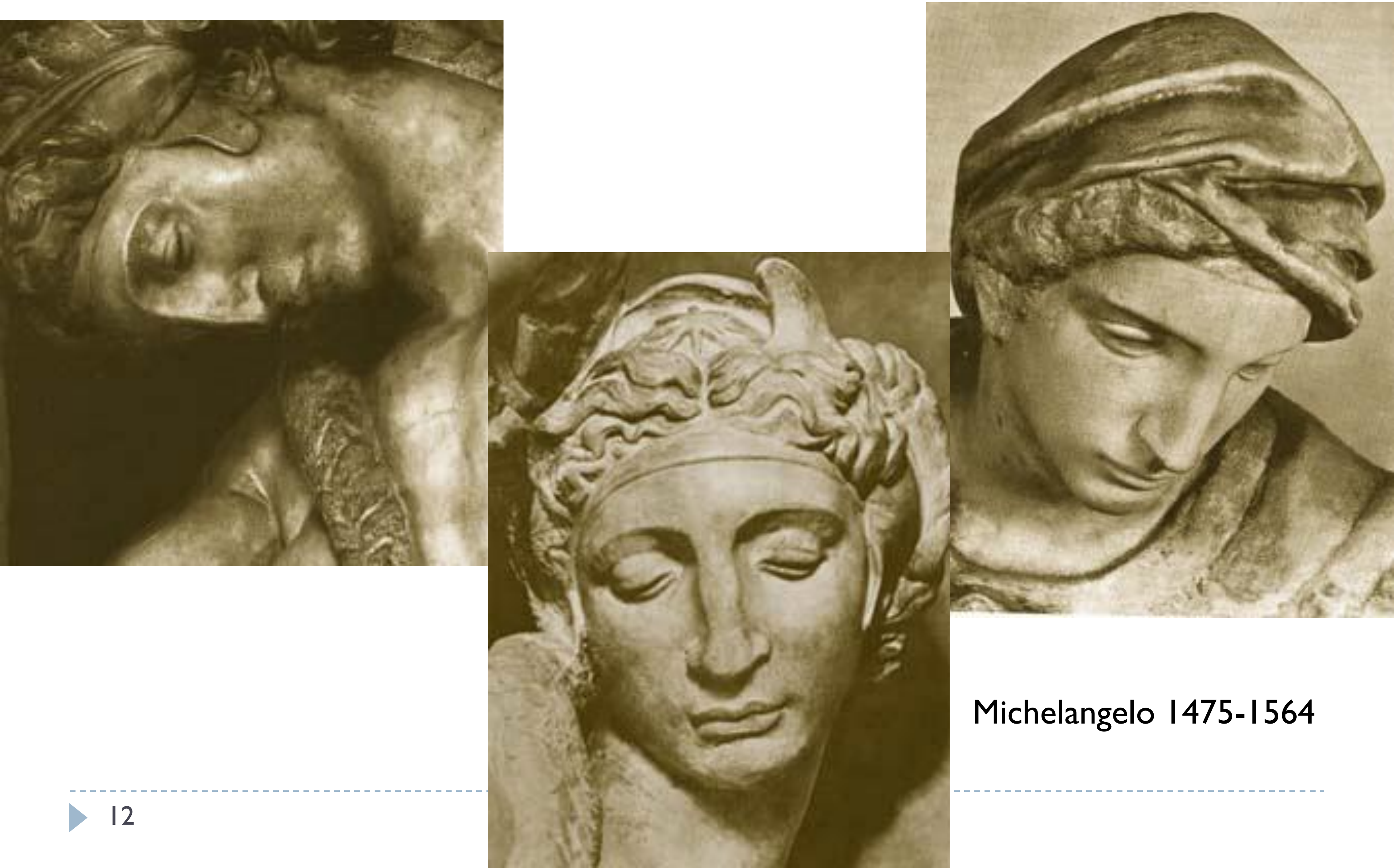
vertical

rigid moving object

rigid moving object

horizontal

Challenges: viewpoint variation



Michelangelo 1475-1564

Challenges: illumination

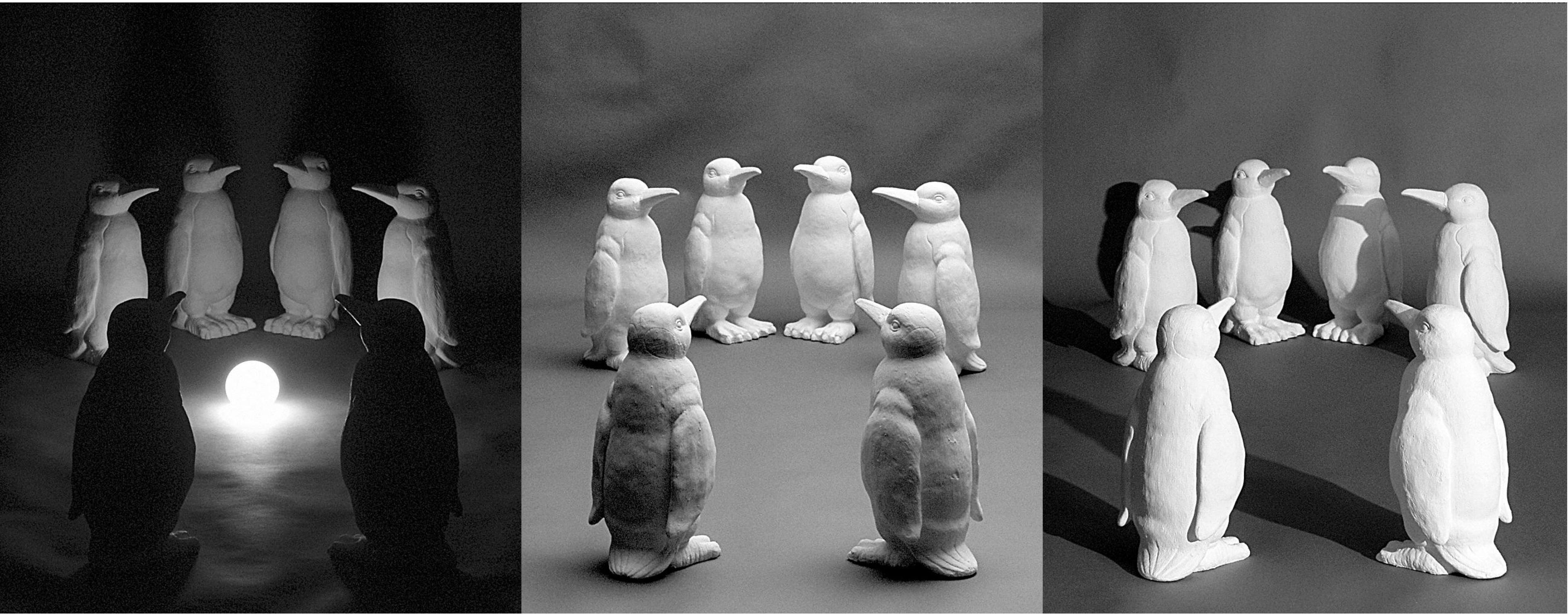
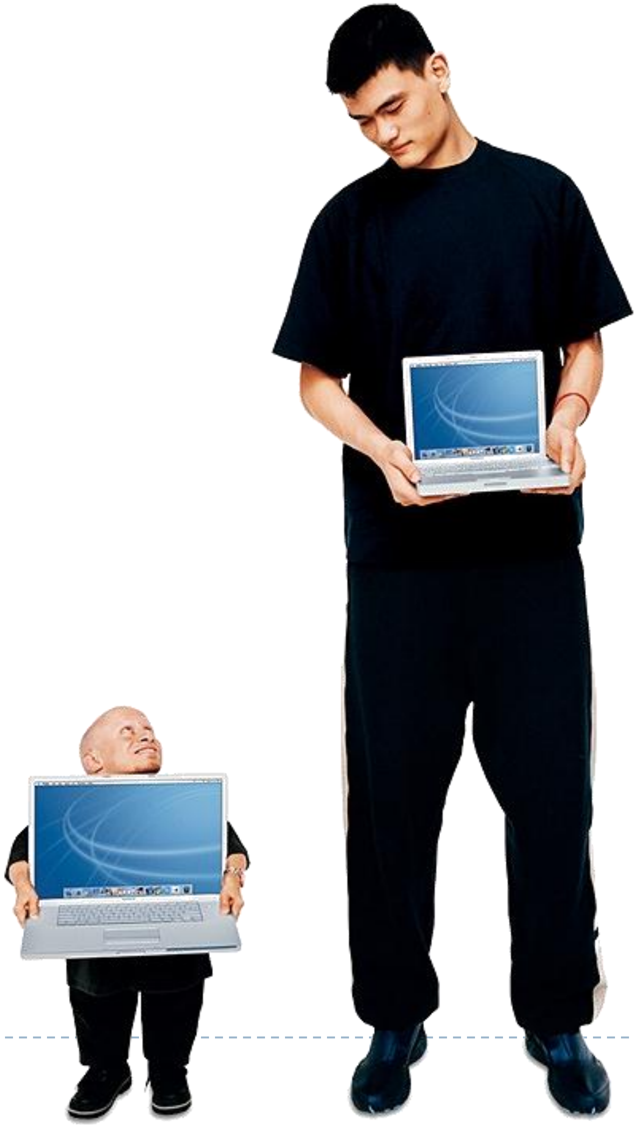


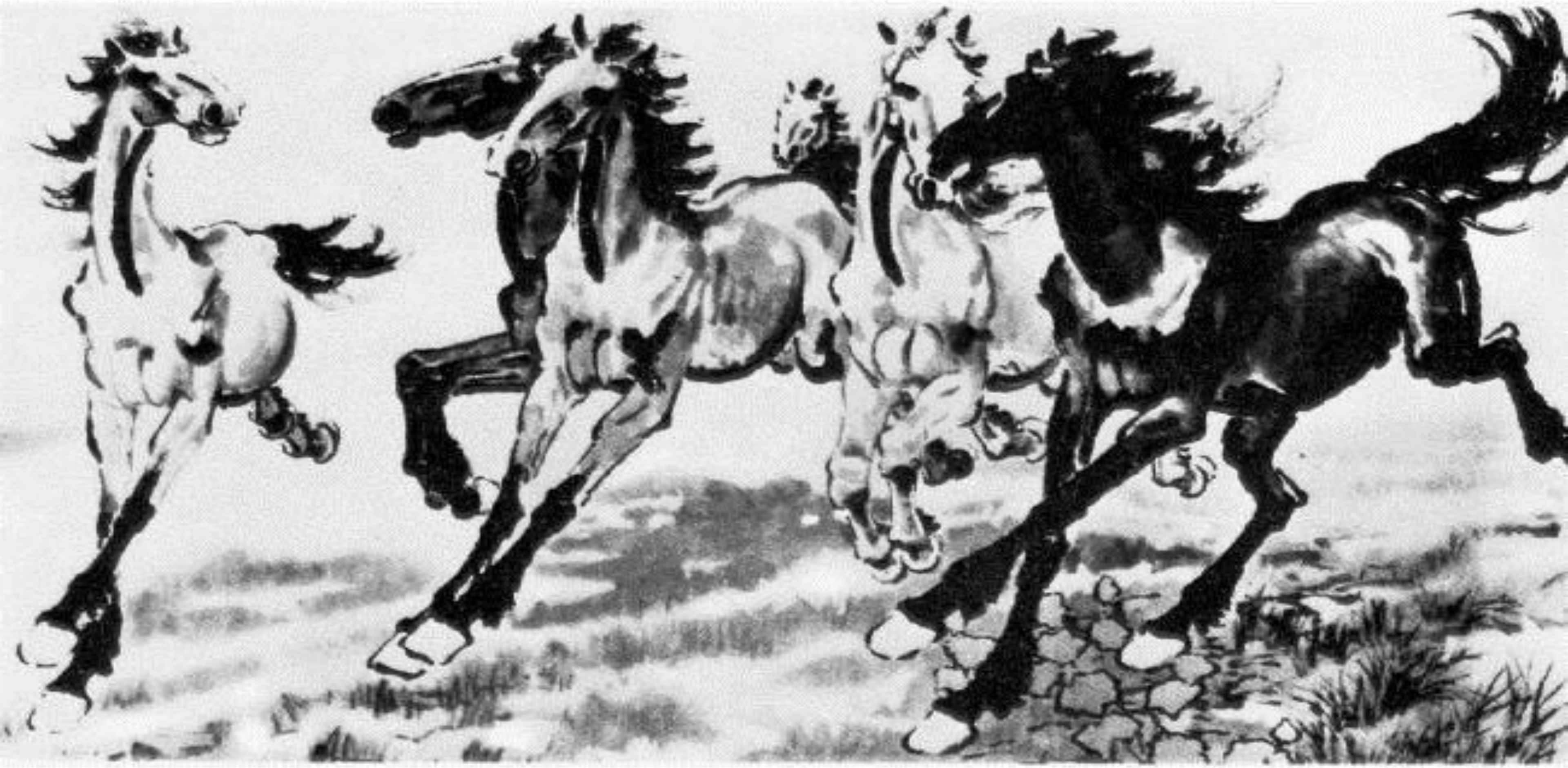
image credit: J. Koenderink

Challenges: scale



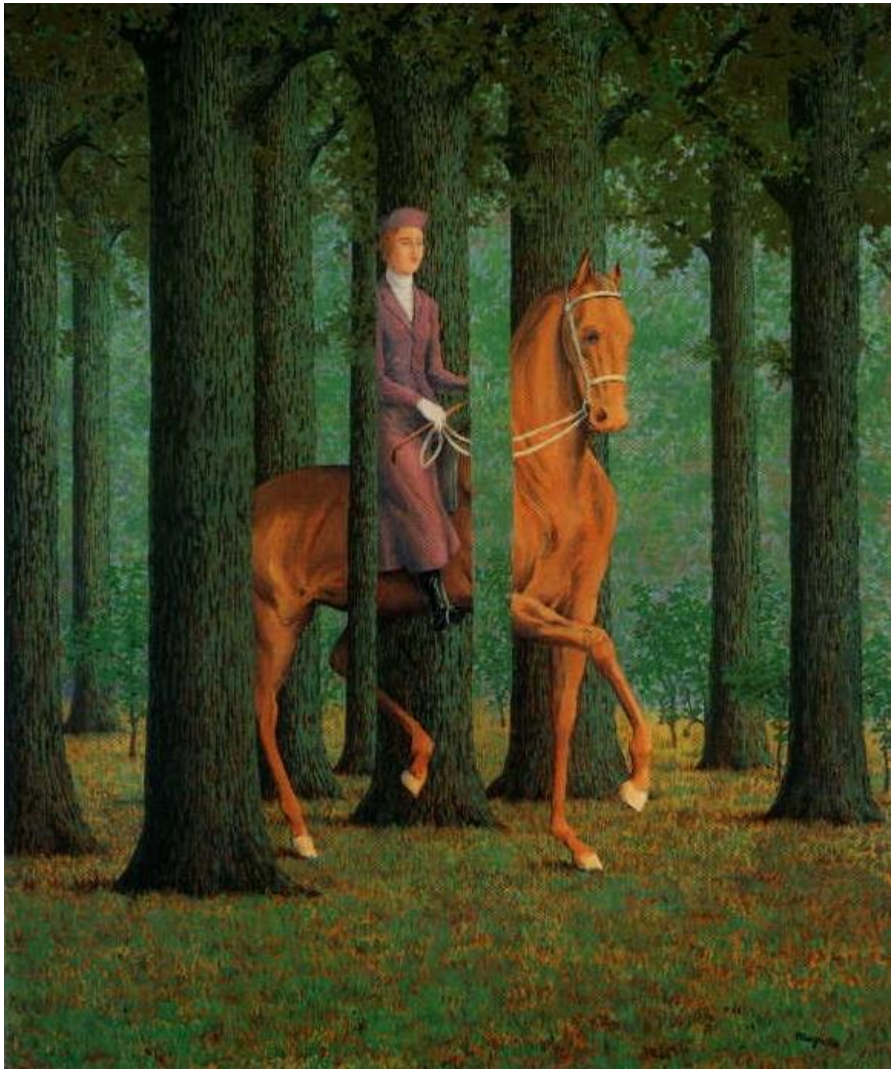
slide credit: Fei-Fei, Fergus & Torralba

Challenges: deformations



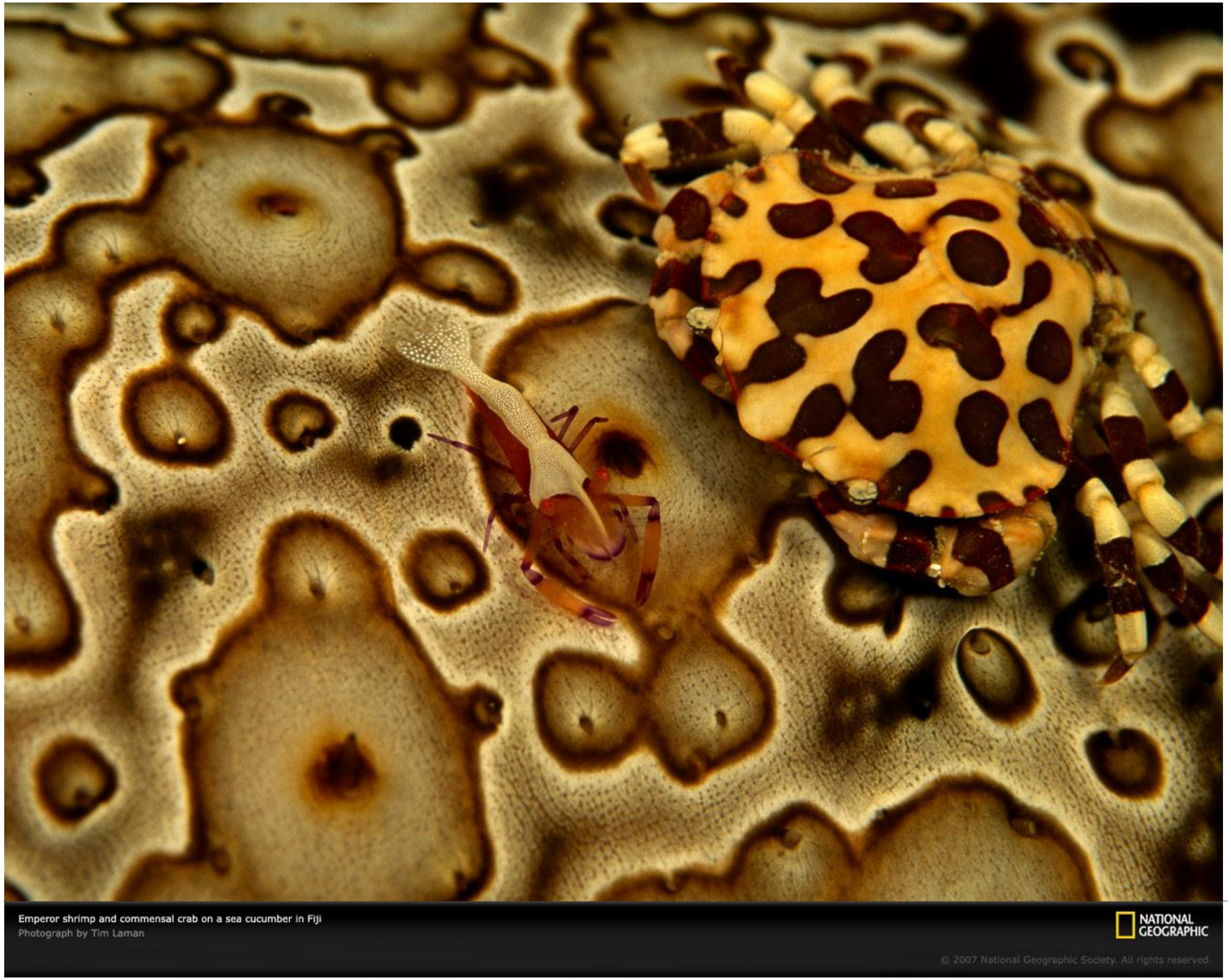
Xu, Beihong 1943

Challenges: occlusions



Magritte, 1957

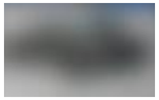
Challenges: background clutter



Emperor shrimp and commensal crab on a sea cucumber in Fiji
Photograph by Tim Laman

Challenges: object intra-class variation

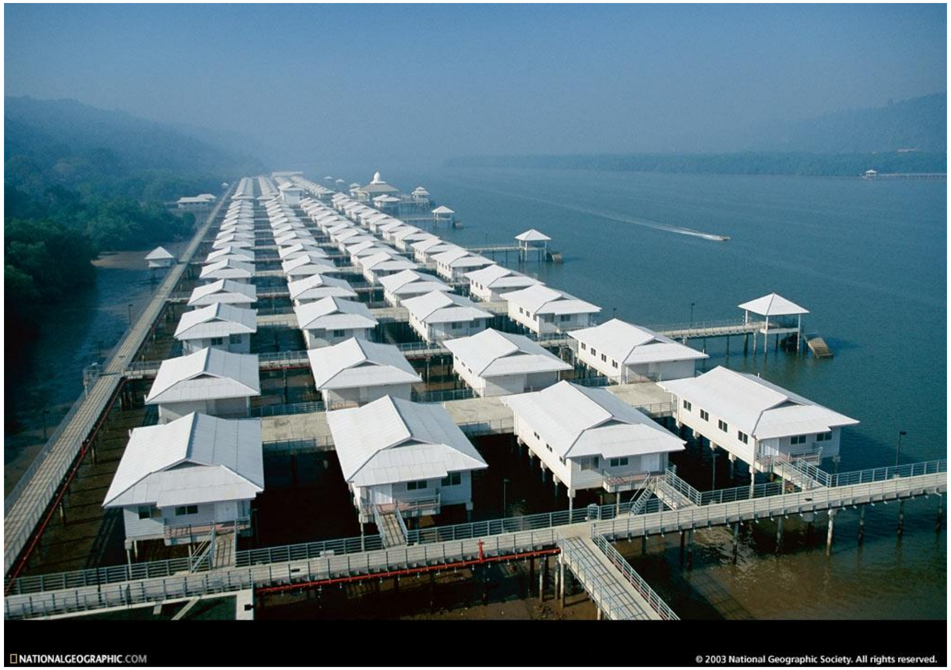
Challenges: local ambiguity



Challenges or opportunities?

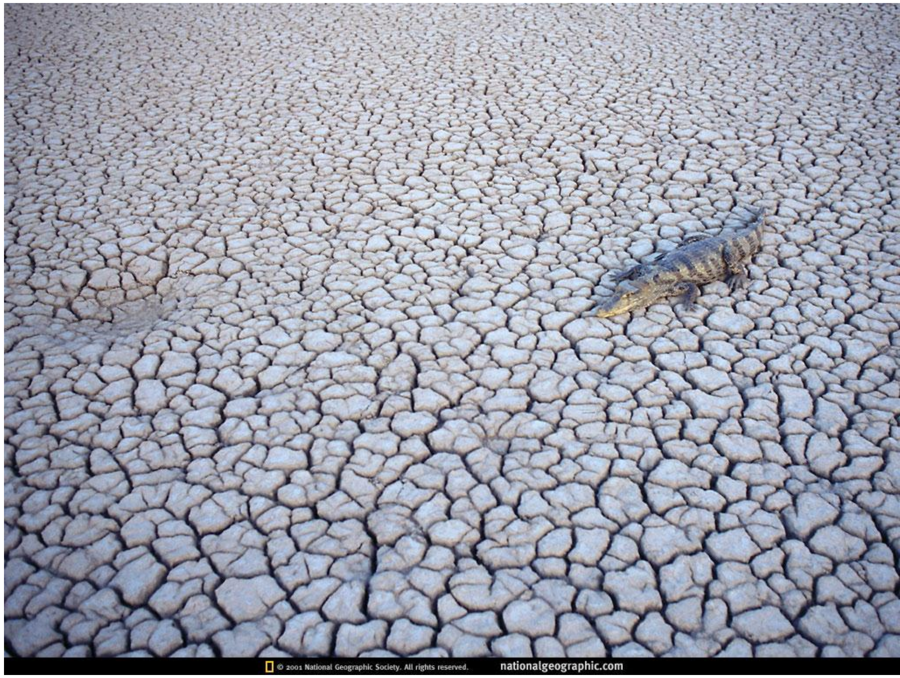
Images are confusing, but they also reveal the **structure** of the world through numerous cues

i.e. Linear perspective, texture gradient,…



NATIONALGEOGRAPHIC.COM

© 2003 National Geographic Society. All rights reserved.

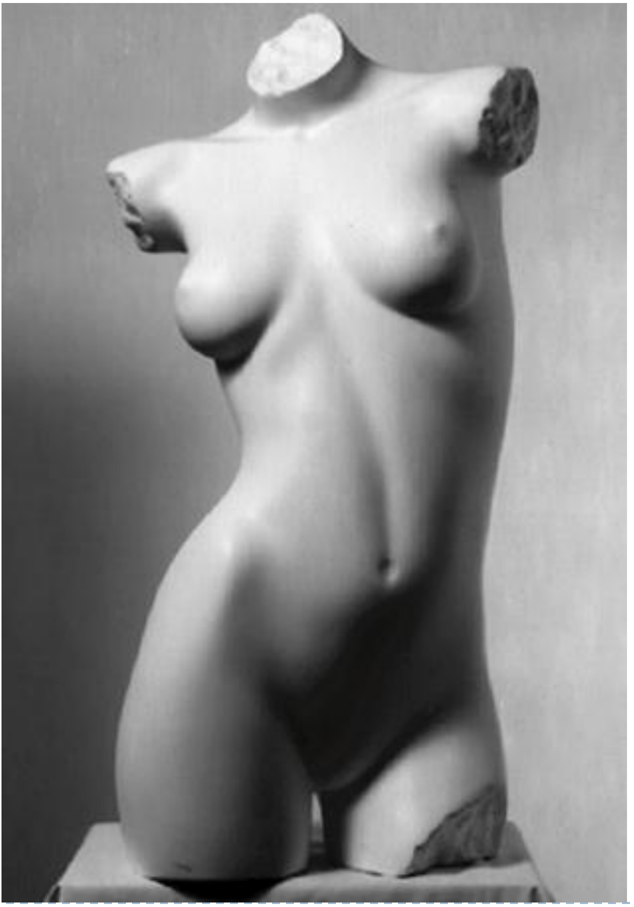


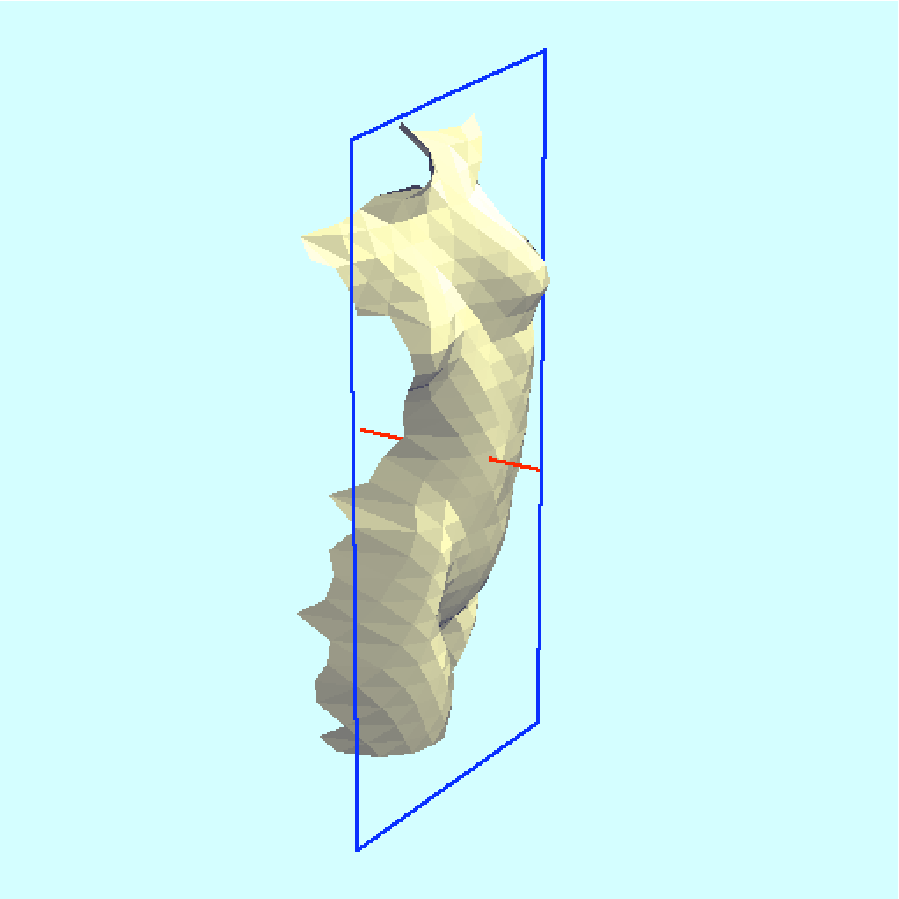
© 2001 National Geographic Society. All rights reserved.

nationalgeographic.com

Challenges or opportunities?

Shape and lighting cues





Challenges or opportunities?

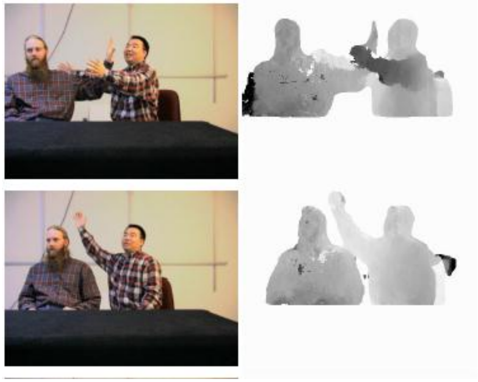
Grouping cues: Similarity

(color, texture, proximity, shape…)

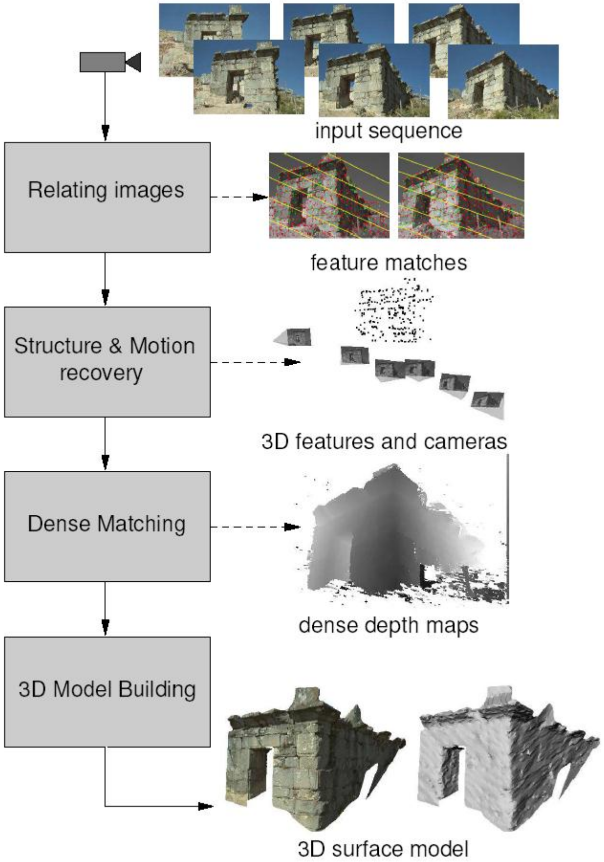
Applications

Real-time stereo

NASA Mars Rover

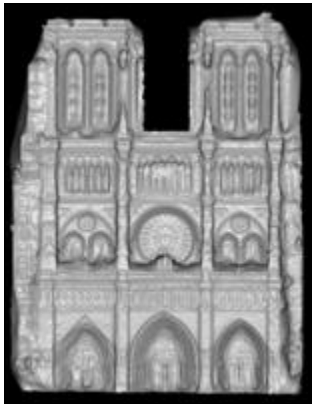


Structure from motion



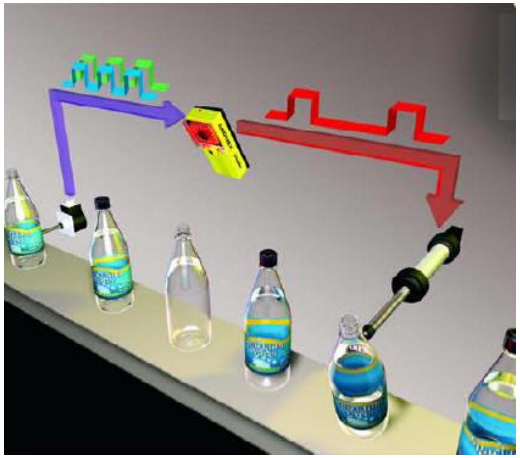
Pollefeys et al.

Multi-view stereo for community photo collections



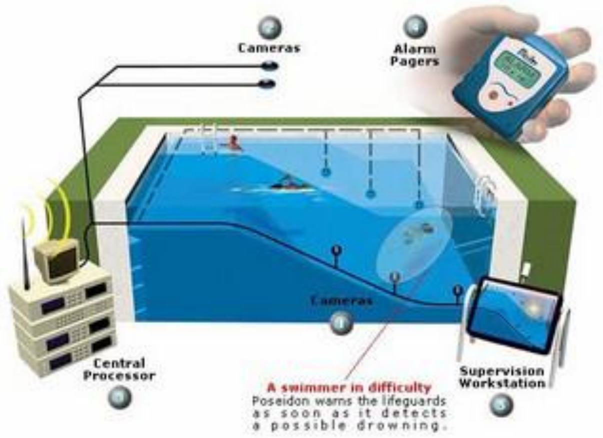
Goesele et al.

Applications

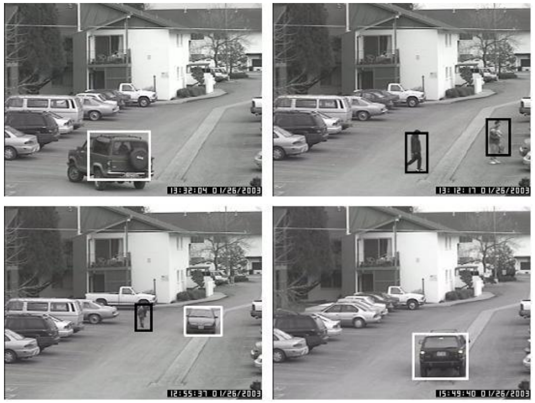


Factory inspection

Reading license plates, checks, ZIP codes



Monitoring for safety (Poseidon)



Surveillance

Autonomous driving, robot navigation

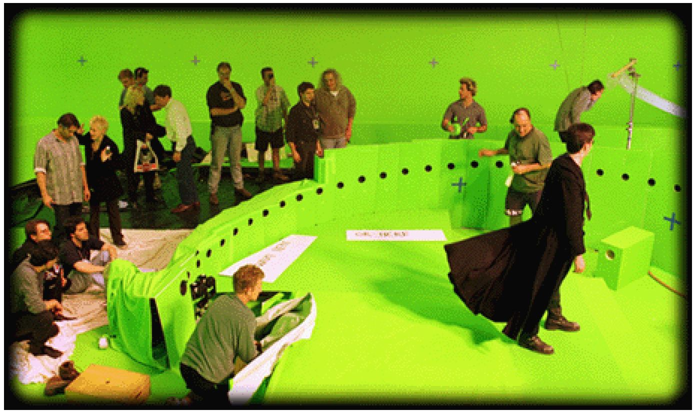
Driver assistance

More info: http://people.cs.ubc.ca/~lowe/vision.html

Applications

Assistive technologies

Entertainment
(Sony EyeToy)



Movie special effects

[Face priority AE] When a bright part of the face is too bright

Digital cameras (face detection for setting focus, exposure)

Visual search
(MSR Lincoln)

More info: http://people.cs.ubc.ca/~lowe/vision.html

Human-computer Interaction

WHY Image and video processing for posture and gesture recognition for Human-computer Interaction (HCI)?

- Computing, communication and display technologies progress quicker & quicker.
- Mechanical devices (keyboard & mouse) for HCI are the bottleneck in the effective utilization of available progresses.
- The future is "natural", i.e. be inspired by the natural human –to-human communication modalities:
  - Speech
  - Gestures

100 %
NATURAL
100 % natural 100 % natural 100 % natural 100 % natural 100 % natural

Human-computer Interaction

Human-computer interfaces inspired by human-human communication.

In 1991, Myron Krueger wrote a pioneering book, “Artificial reality”, where it is reported:

“Natural interaction means voice and gesture. New interface technologies requires tools and features that mimic the principles of human communication.”

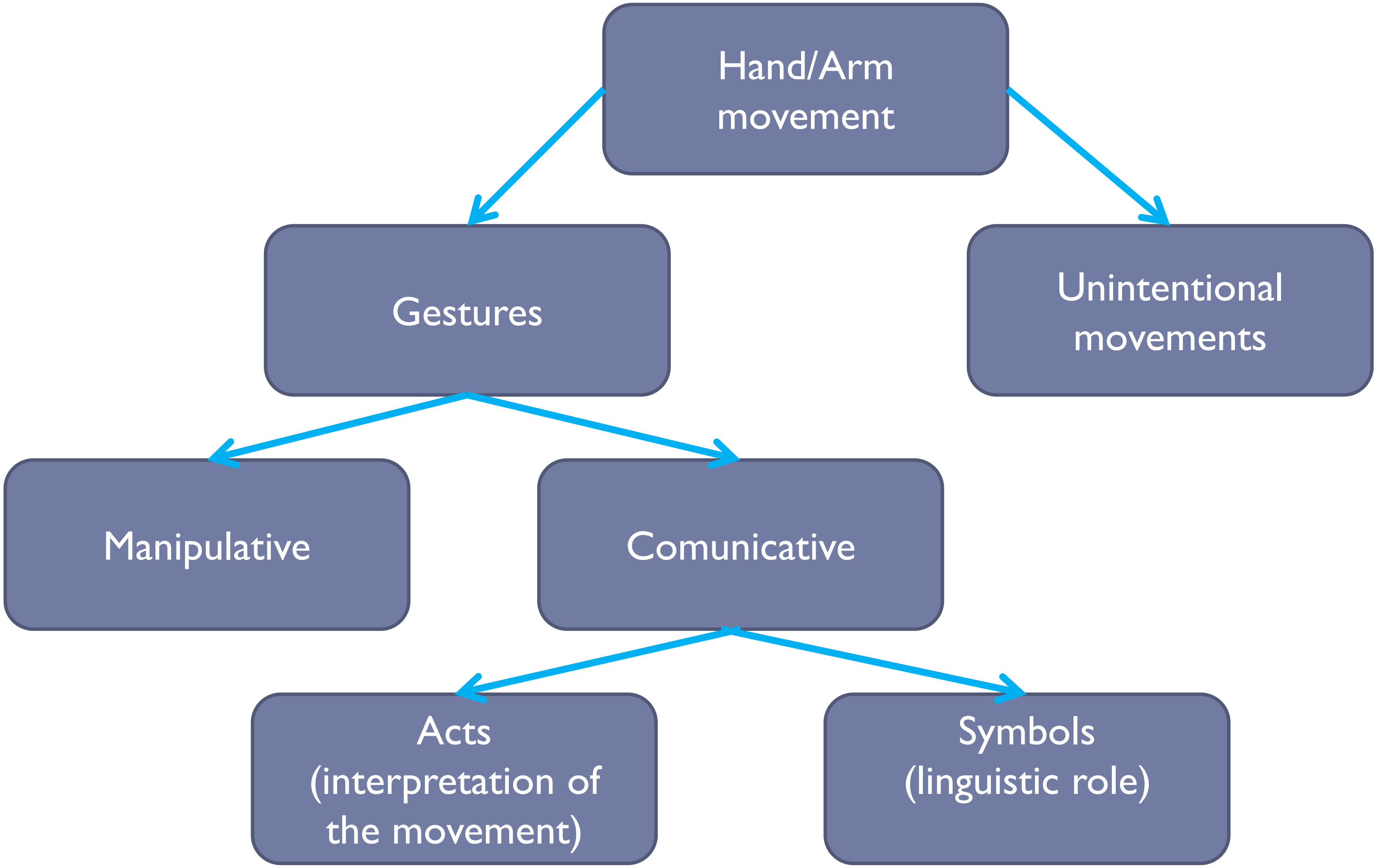
Gesture recognition

- Gestures are useful for computer interaction since they're the most primary and expressive form of human communication.

- Webster's dictionary defines gestures as

"...the use of motions of the limbs or body as a means of expression; a movement usually of the body or limbs that expresses or emphasizes an idea, sentiment or attitude"

Gestural taxonomy


Hand/Arm movement

Gestures

Unintentional movements

Manipulative

Comunicative

Acts (interpretation of the movement)

Symbols (linguistic role)

Gesture recognition

- No single method for automatic gesture recognition is suitable for every application: each gesture-recognition algorithm depends on:
  - user cultural background;
  - application domain;
  - environment.

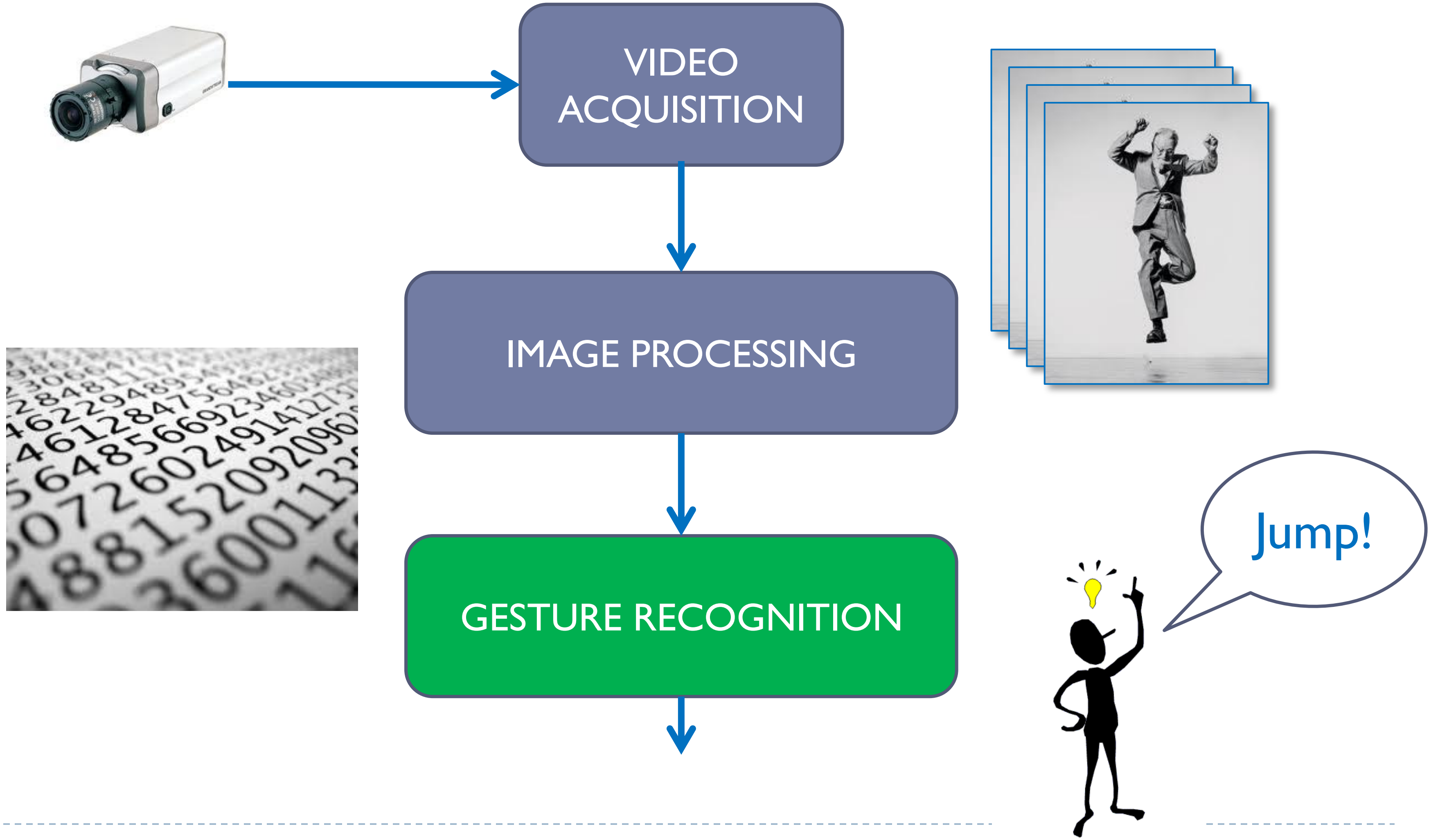
- System requirements vary depending on the aim of the application (i.e. entertainment system vs surgical system..)

Vision-based gesture recognition

- Visual interpretation of:
  - body pose
  - body gesture
  - hand pose
  - hand gesture

Using imaging devices: still/video cameras

In a nutshell



VIDEO ACQUISITION

IMAGE PROCESSING

GESTURE RECOGNITION

Vision-based gesture recognition

- Advantages of video-based gestures applications over conventional human-machine interaction:
  - Access information while maintaining total sterility: touchless interfaces; healthcare environments.
  - Easily explore large and complex data.
  - Provide a source of expressiveness.
  - Overcome physical handicaps: disable people; elderly people.

Differences

Do you know the difference between handicap & disability?

Disability: inability to execute some class of movements, or pick up sensory information of some sort, or perform some cognitive function, that typical unimpaired humans are able to execute or pick up or perform.

Handicap: inability to accomplish something one might want to do because of the environment/situation.

Differences

Do you know the difference between handicap & disability?

We can be handicapped, even when we are not disabled.

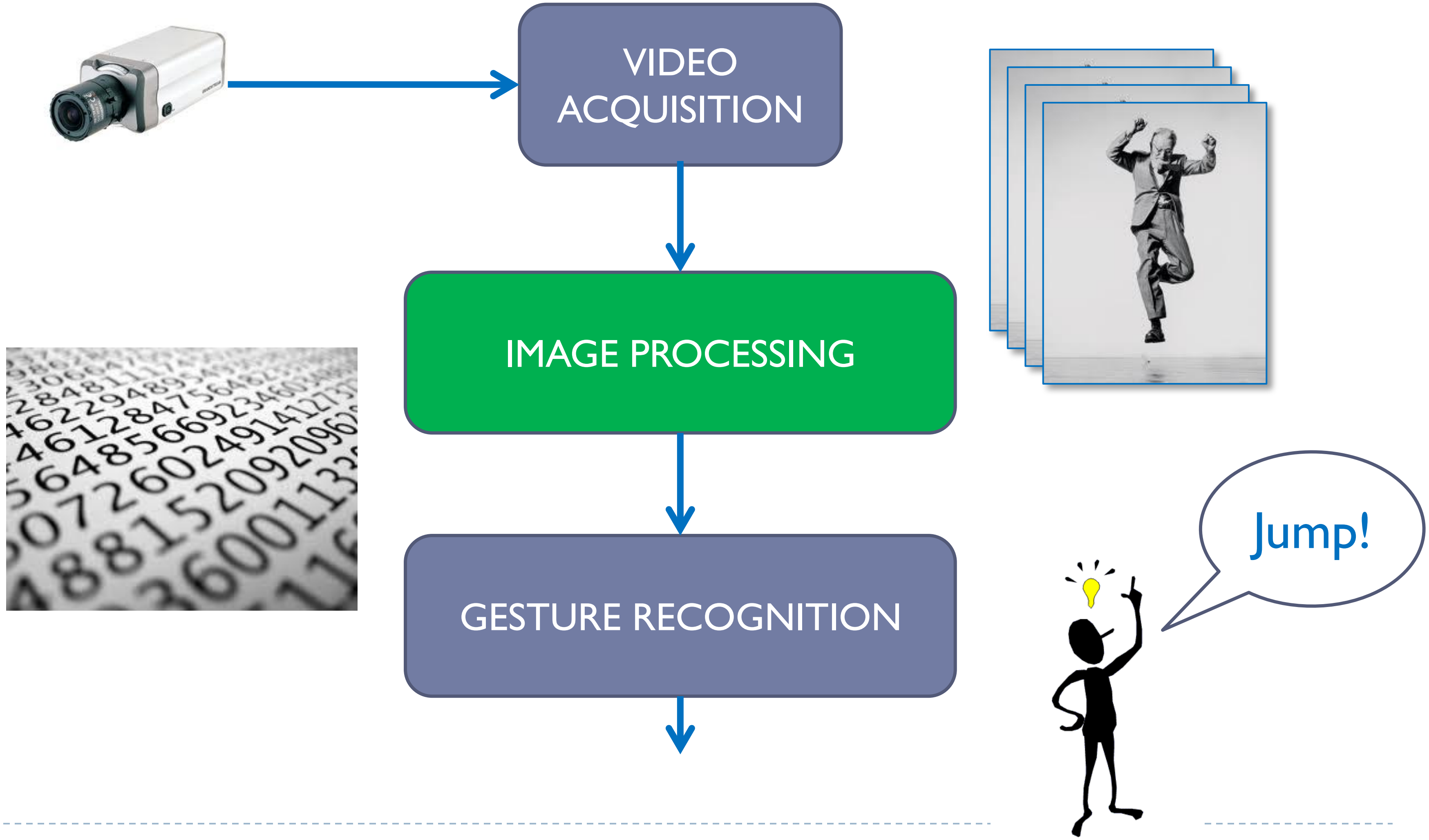
Italians who do not speak Japanese will be handicapped when they visit Tokyo, because while most people will be able to gather important information by reading signs on buildings, they will not.

And one can be disabled, without being handicapped relative to many tasks, if the proper tools and supporting structures are provided.

Vision-based gesture recognition

- Gesture interpretation:
  1. **Definition** of gestures and dictionary
  2. **Temporal modelling** of gestures: set of temporal parameters
  3. **Spatial modelling** of gestures: characterization of spatial properties of limb's trajectories

In a nutshell



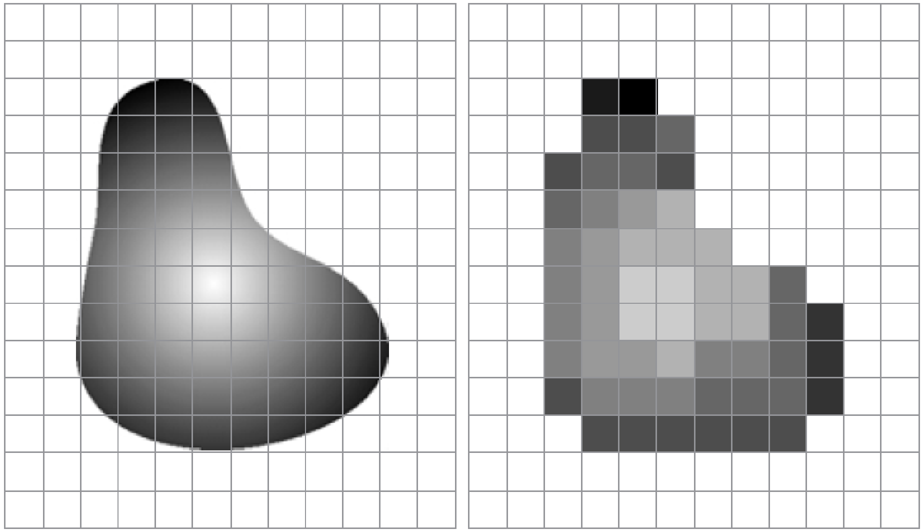
VIDEO ACQUISITION

IMAGE PROCESSING

GESTURE RECOGNITION

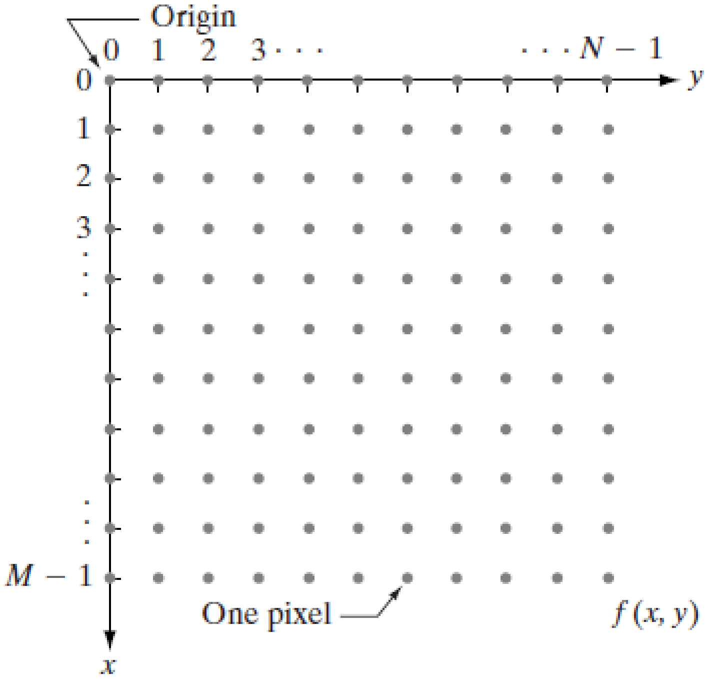
Digital images and processing

- **Image**: 2D function, **I**(x,y), where x and y are spatial coordinates, and the amplitude of f at any pair (x, y) is called the intensity or gray level of the image at that point.
- If x, y, and the amplitude values of **I** are all finite, discrete quantities, we call the image a **digital image**.



Digital images and processing

- **Digital image processing**: processing digital images by means of a digital computer.
- A digital image is composed of a finite number of elements (**pixels**) defined by:
  - location (x,y)
  - intensity value



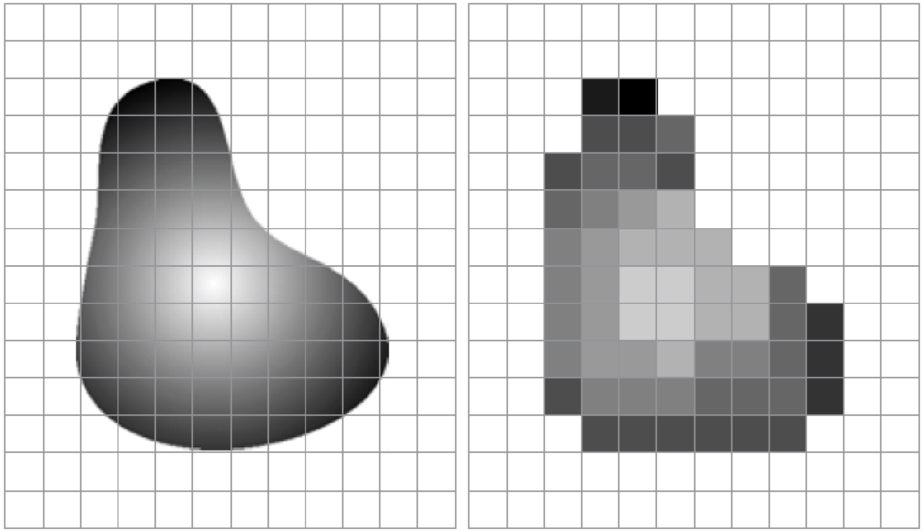
Digital images and processing

An image may be continuous with respect to:

- the x- and y-coordinates;
- in amplitude.

To convert it to digital form, we have to sample the function in both coordinates and amplitude.

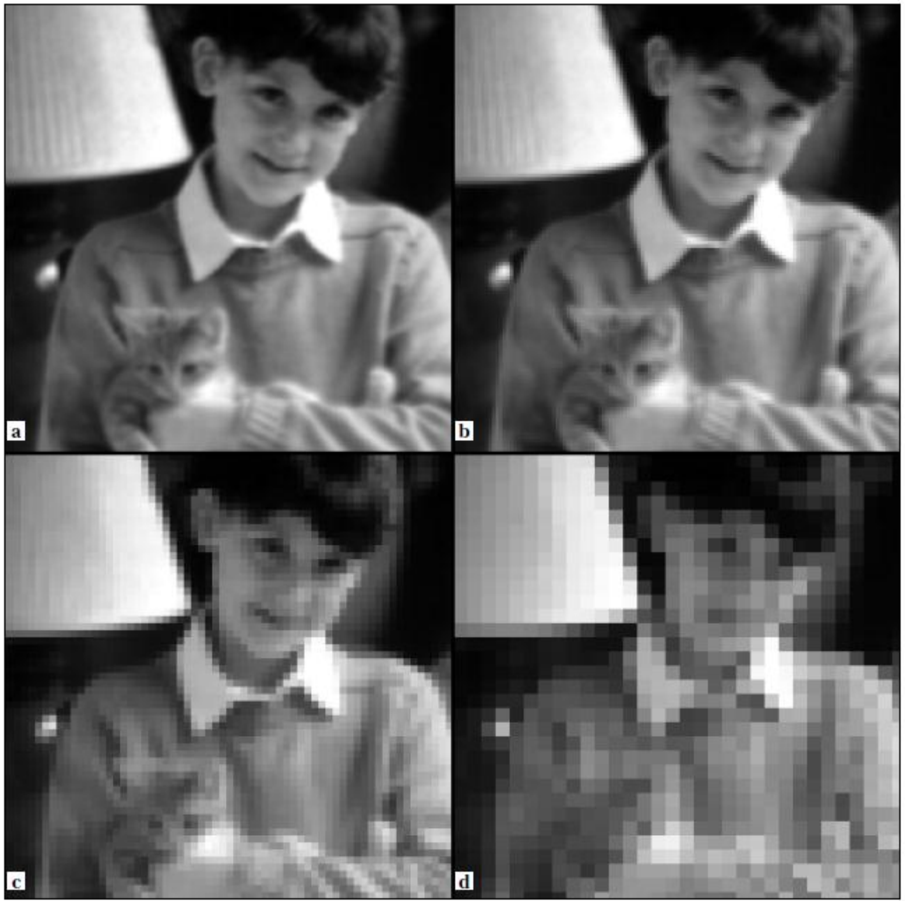
- **sampling**: digitizing the image coordinate values.
- **quantization**: digitizing the image amplitude (black2white)



Digital images and processing

Sampling

Spatial resolution is the smallest discernible detail in an image.



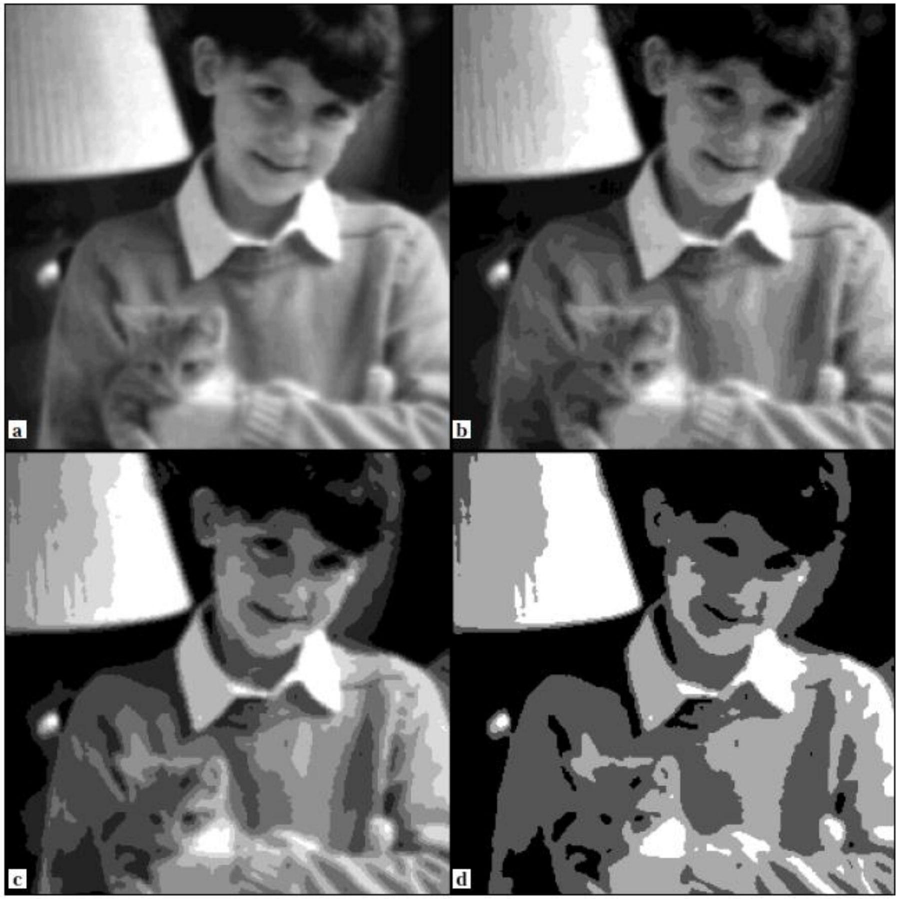
Four representations of the same image, with variation in the number of pixels used:

a) 256x 256;
b) 128 x128;
c) 64x 64;
d) 32 x 32.

Digital images and processing

Quantization

Gray-level resolution refers to the smallest discernible change in gray level.



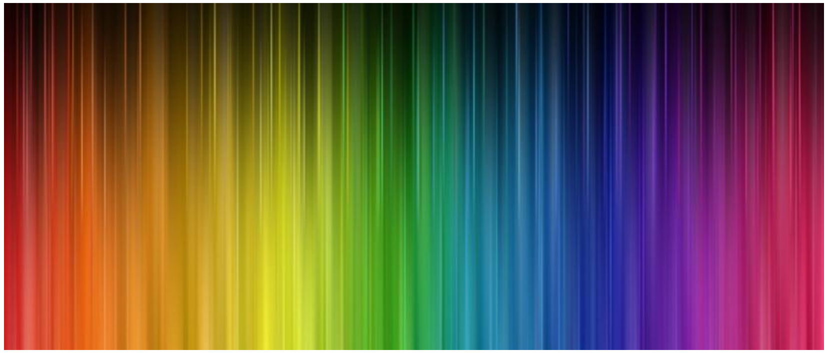
Four representations of the same image, with variation in the number of grey levels used:

a) 32;
b) 16;
c) 8;
d) 4.

Digital images and processing

Color imaging

Most real-world images are not monochrome, of course, but full color!



Images from digital cameras are usually in Red Green Blue (RGB) colorspace.

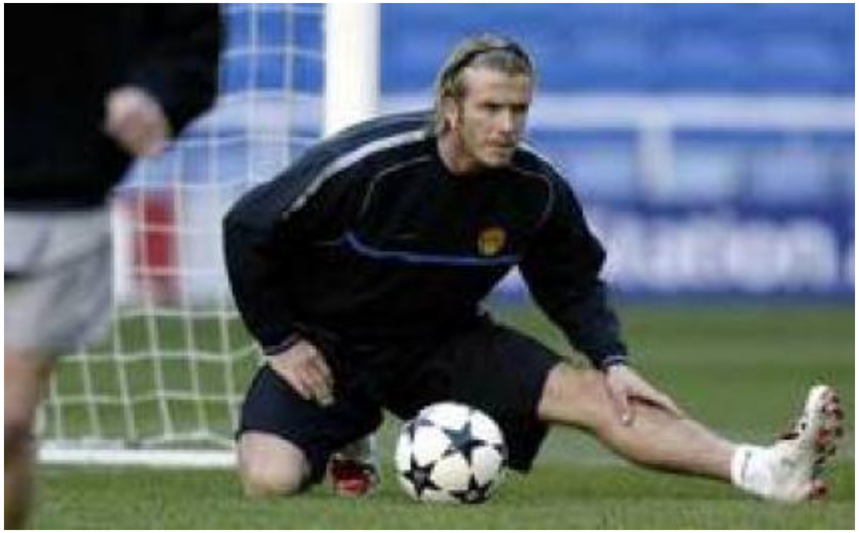
A number of color spaces or color models have been suggested and each one of them has a specific color coordinate system and each point in the color space represents only one specific color.

Each color model may be useful for specific applications.

Digital images and processing

- What is it possible to do with color video frames for motion detection?

Example: skin detection



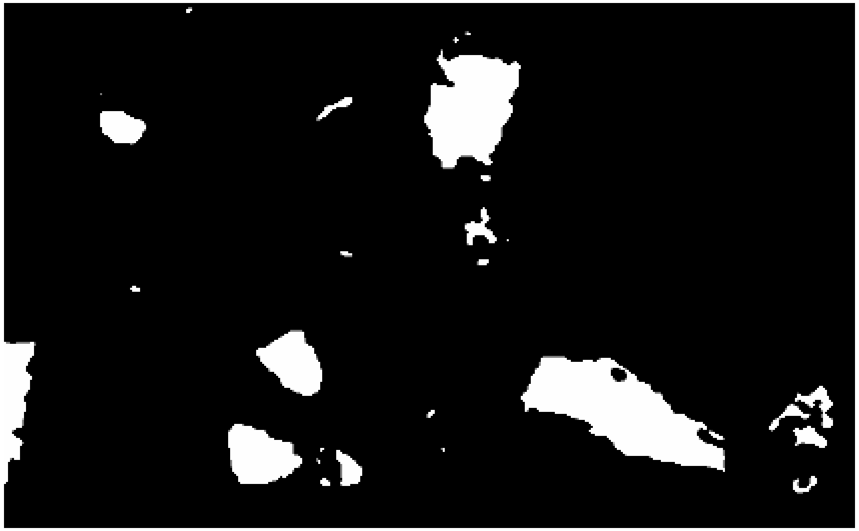
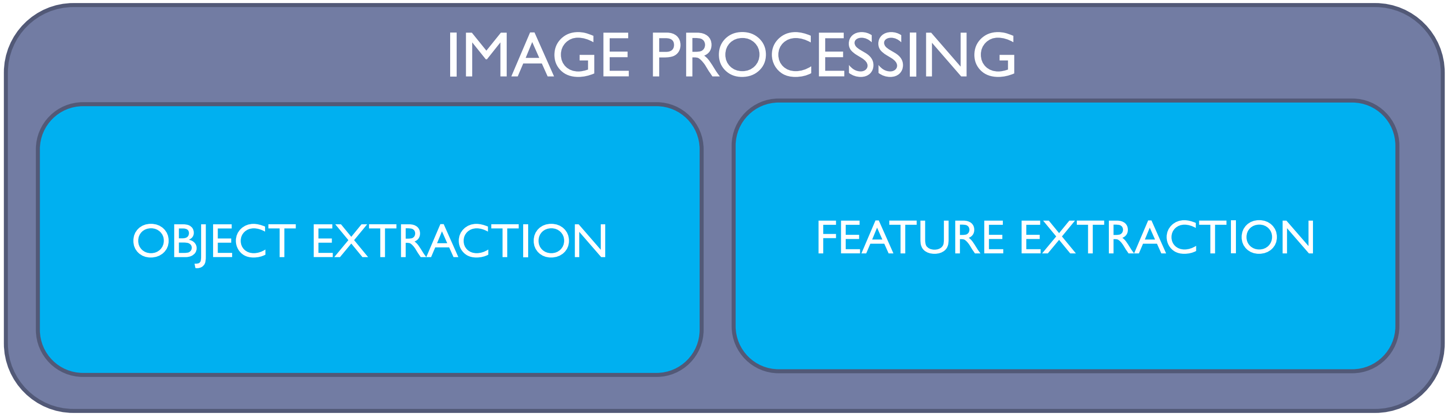


Image processing

- We said that **Computer vision** is the transformation of data from a still or video camera into either a decision or a new representation.

- **Image processing** is part of Computer Vision and aim at **transforming the image** so that information can be extracted.



Global operator: threshold

- Simple and useful image processing method for getting information is segmenting pixels with respect to their values, i.e. **segmenting objects**.
- The basic global threshold algorithm aims at
  - scanning the image pixel by pixel;
  - labelling each pixel whether its value greater or less than a **threshold value T**.

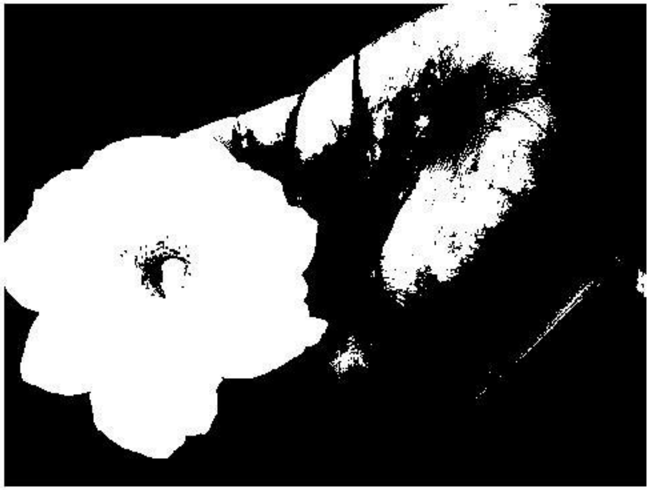
- If the pixel value is >= T, then it's set to a maximum value M (usually 255)
- If the pixel value is < T, then it's set to a minimum value m (usually 0)

Global operator: threshold

- Of course, the result depends on the choice of T…

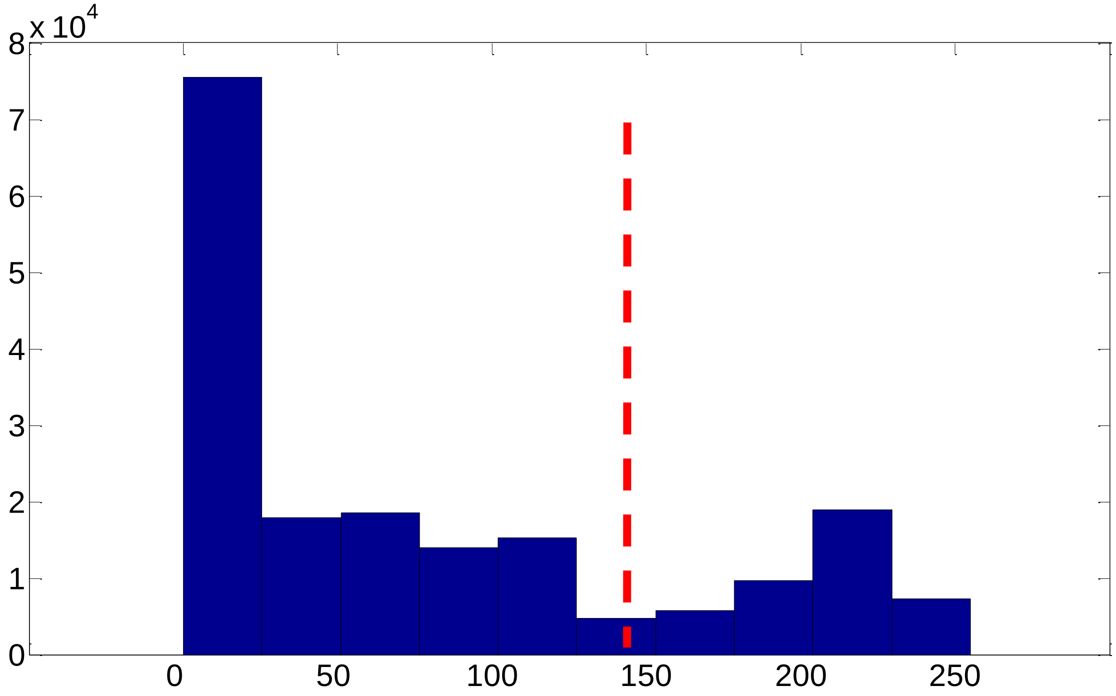
Original

T = 100



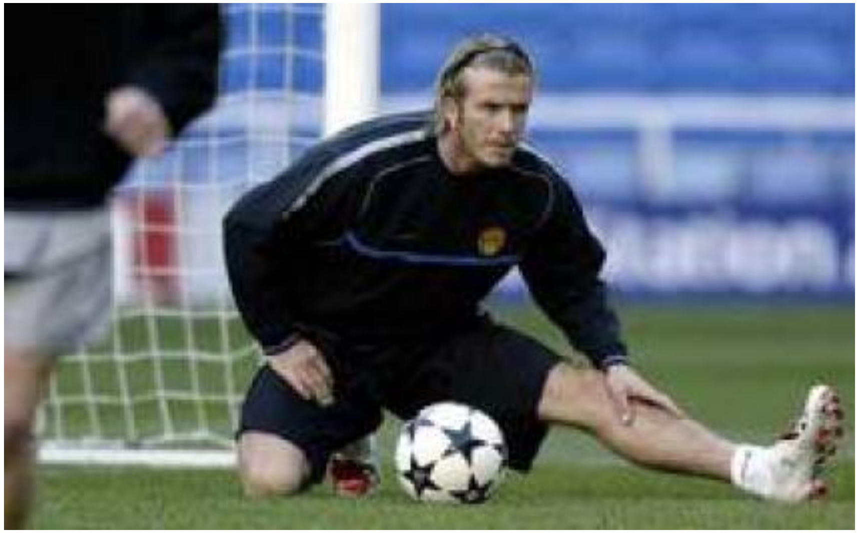
Global operator: threshold

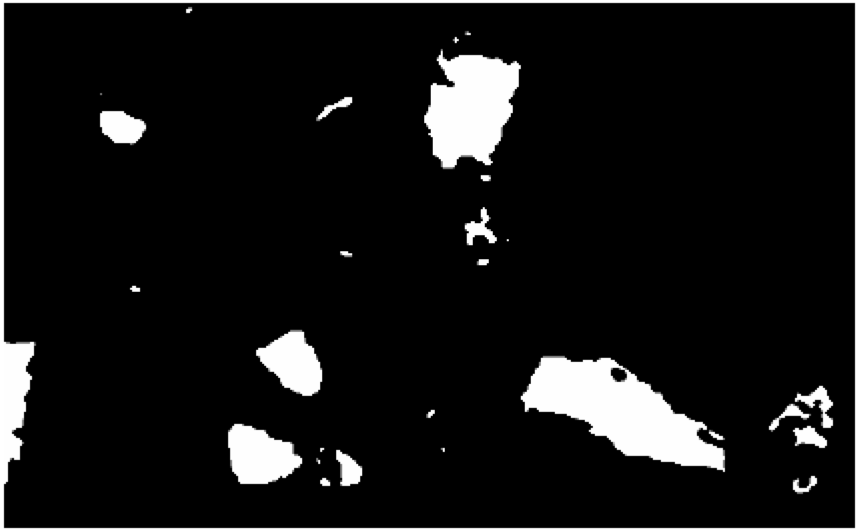
- Choosing the threshold looking to the histogram of the pixel intensity values



Global operator: threshold

- Sometimes pixels belonging to the object of interest are **in the range of 2 values**.
- In this case, we need 2 thresholds
- Remember Beckham?

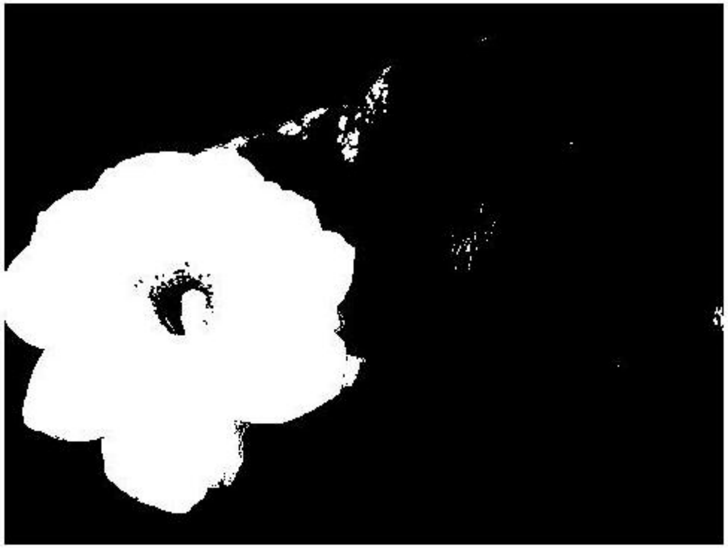




- Thresholds in HSV colorspace: H(0-20); S (30-150); V(80-255)

Morphology

- After image threshold, we have a **binary image** where white pixels (255) correspond to the object of interest (+ noise)



- Image processing include algorithms which allow to **connect** isolated pixels sufficiently close to other and/or **deleting** isolated pixels (**Mophology operators**)

Morphology

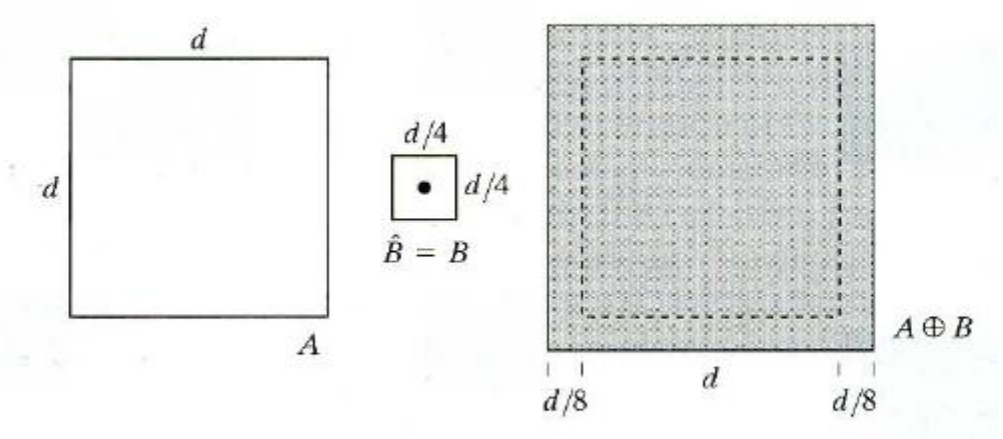
- The basic morphological transformations are called **dilation and erosion**, and they aim at:
  - removing noise;
  - isolating individual elements;
  - joining disparate elements

  in an image.

  Morphology is a local image operator.

Morphology

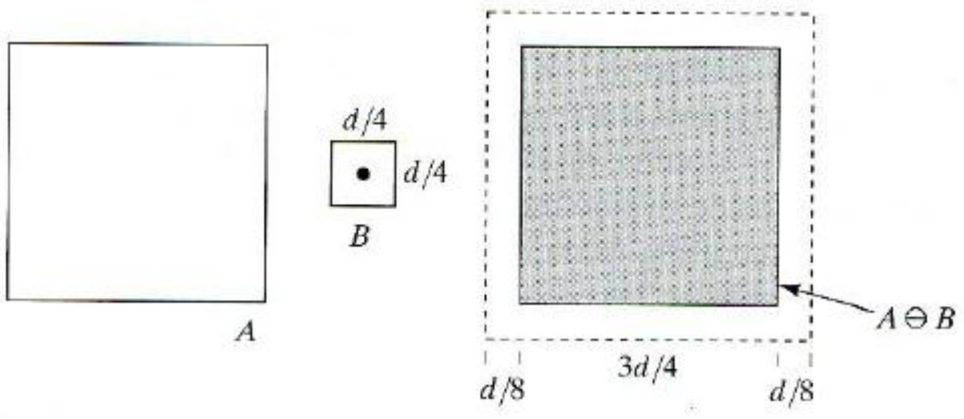
Dilation is a convolution of an image A with a kernel B and causes bright regions within an image to grow.

- As the kernel B is scanned over the image, we compute the maximal pixel value overlapped by B and replace the image pixel under the central point with that maximal value.
- 

- The kernel can be any shape or size (usually small solid square or disk)

Morphology

Erosion is a convolution of an image A with a kernel B and causes dark regions within an image to grow.

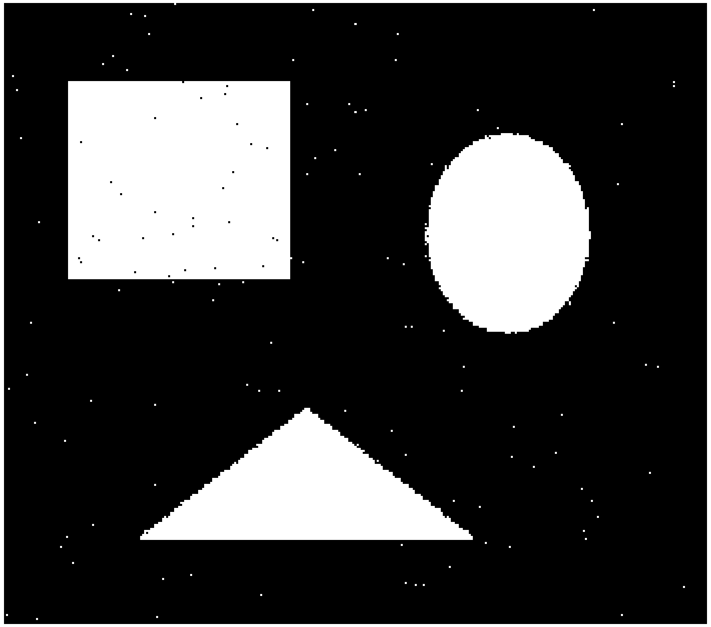
- As the kernel B is scanned over the image, we compute the minimum pixel value overlapped by B and replace the image pixel under the central point with that minimum value.
- 

- The kernel can be any shape or size (usually small solid square or disk)

Morphology

How and when do we need to use them?

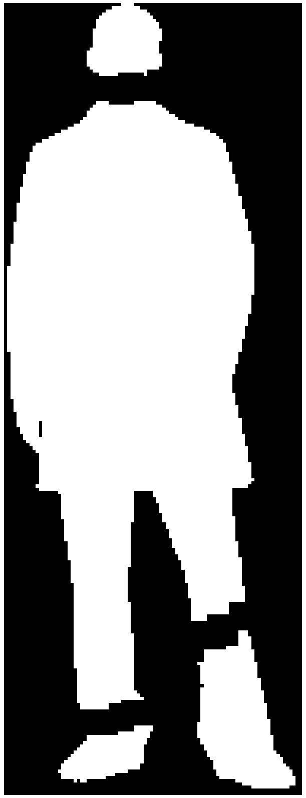
- The erode operation is often used to eliminate "speckle" noise in an image. The idea here is that the speckles are eroded to nothing while larger regions that contain visually significant content are not affected.



Morphology

How and when do we need to use them?

- The dilate operation is often used when attempting to find connected components (i.e., large discrete regions of similar pixel color or intensity). The utility of dilation arises because in many cases a large region might otherwise be broken apart into multiple components as a result of noise, shadows, or some other similar effect. A small dilation will cause such components to "melt" together into one.



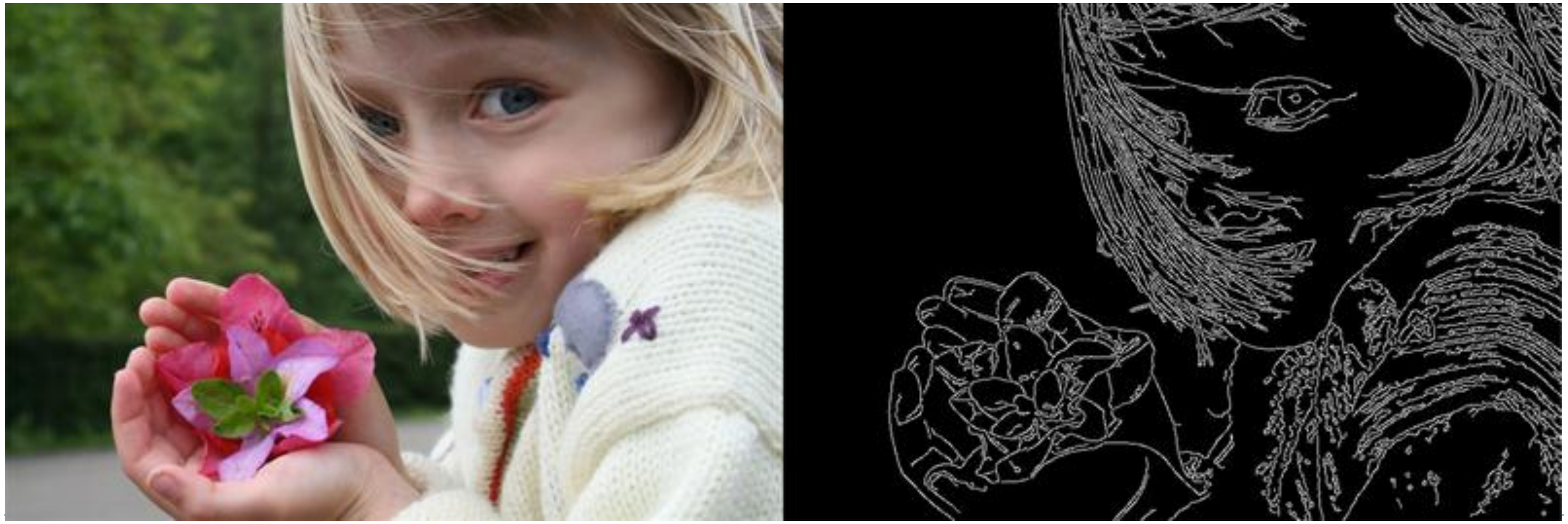
Edge detection

- Tipically, image processing prefer to have information about **features**, rather than object detection.
- The **detection of edges** is a fundamental tool in image processing, in the areas of feature detection and feature extraction, which aim at identifying points in a digital image at which the image brightness changes sharply or, more formally, has discontinuities.
- AIM: to capture important events and changes in properties of the world assuming that **discontinuities in image brightness** are likely to correspond to:
  - discontinuities in depth and surface orientation,
  - changes in material properties and variations in scene illumination.

Edge detection

Set of connected curves that indicate the boundaries of objects.

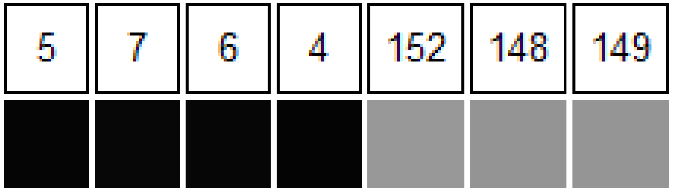
- Applying an edge detection algorithm to an image may significantly reduce the amount of data to be processed and may therefore filter out information that may be regarded as less relevant, while preserving the important structural properties of an image.



Edge detection

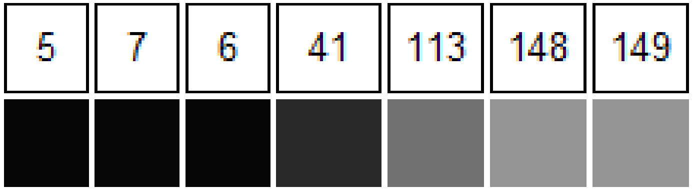
What discontinuities are?

Consider the problem of detecting edges in the following 1Dsignal. We may intuitively say that there should be an edge between the 4th and 5th pixels.



5	7	6	4	152	148	149

BUT here?



5	7	6	41	113	148	149

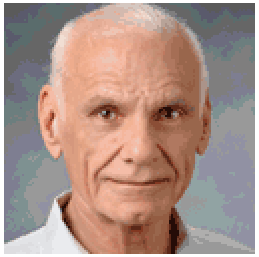
Not trivial at all…

Edge detection

Hopefully…

Mr. Canny

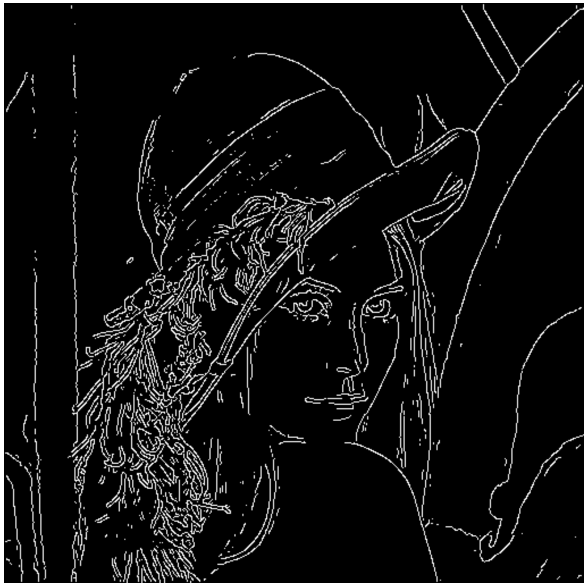
Mr. Sobel



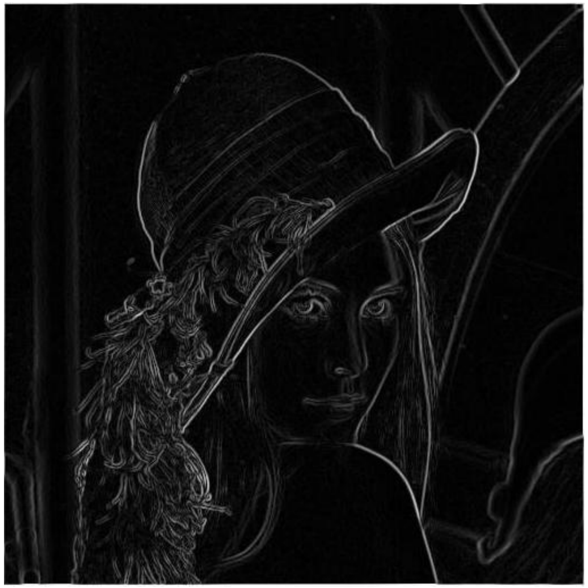
Mr. Roberts

…& Miss Lena

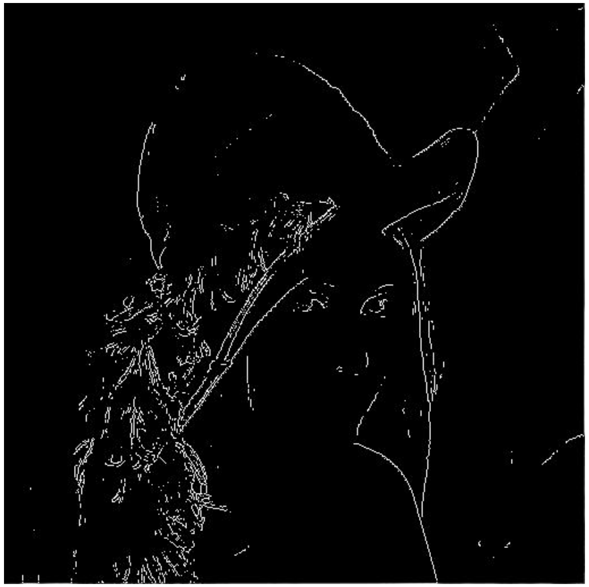
Edge detection



Lena for Mr. Canny



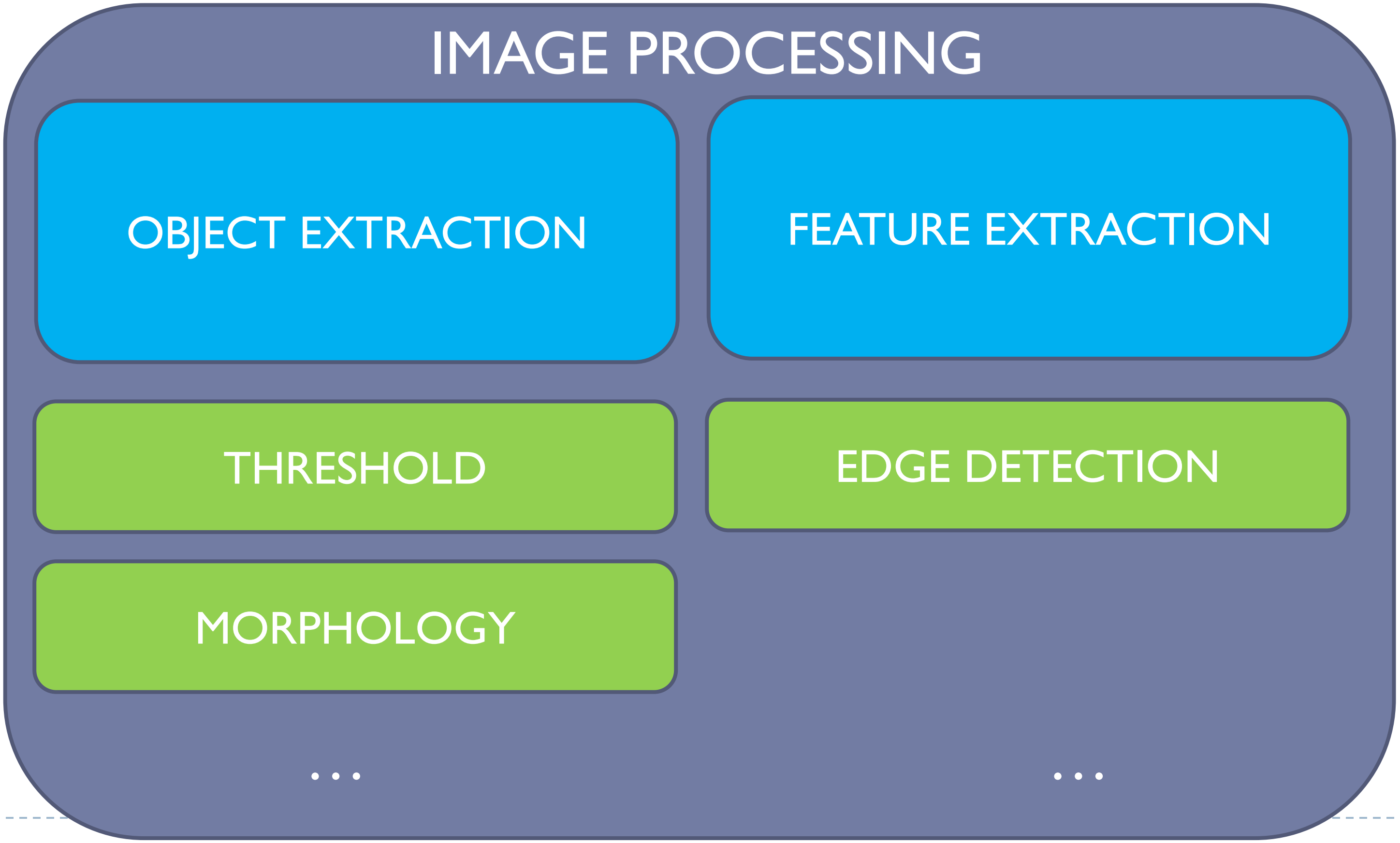
Lena for Mr. Sobel



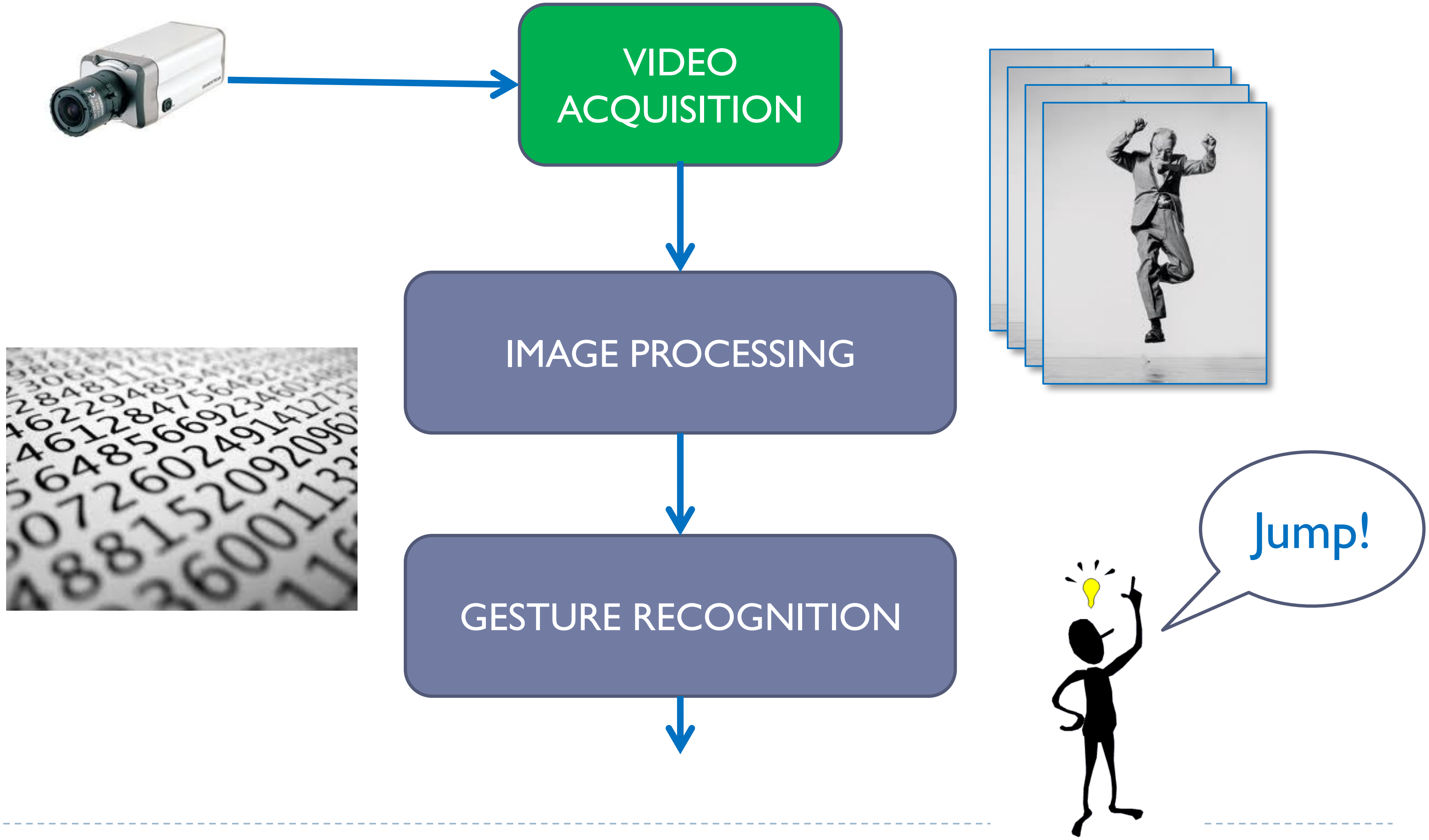
Lena for Mr. Roberts

Image processing: basics

- **Image processing** aims at **transforming the image** so that information can be extracted.



In a nutshell



VIDEO ACQUISITION

IMAGE PROCESSING

GESTURE RECOGNITION

Video camera

- Monochrome
- color
- IR

- CCD
- CMOS

- 1 camera
- 2 or more cameras

- Stereoscopic view

- Depth cameras

- Spatial resolution (#pixels)
- Temporal resolution (fps)