An introduction to Robot Perception

Octavio Arriaga

December 13, 2022

Table of contents

1. Introduction
2. Color
3. Image formation
4. Image transformations
5. Bibliography

Table of contents

1 Introduction

2 Color

3 Image formation

4 Image transformations

5 Bibliography

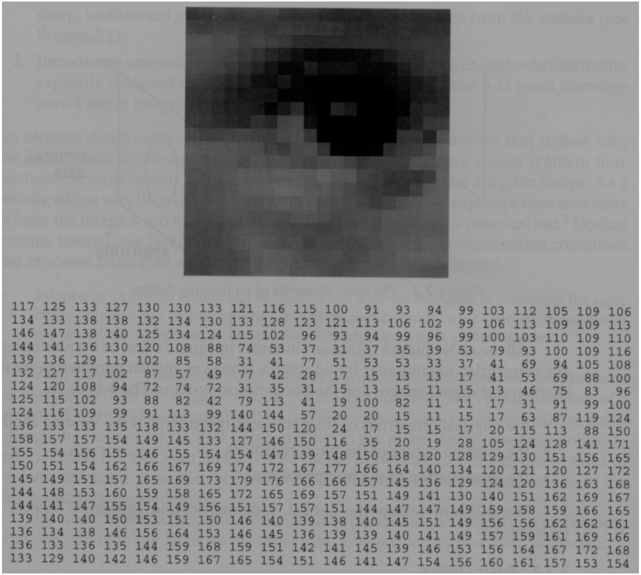


Figure: Why is computer vision hard? [3]

Related disciplines

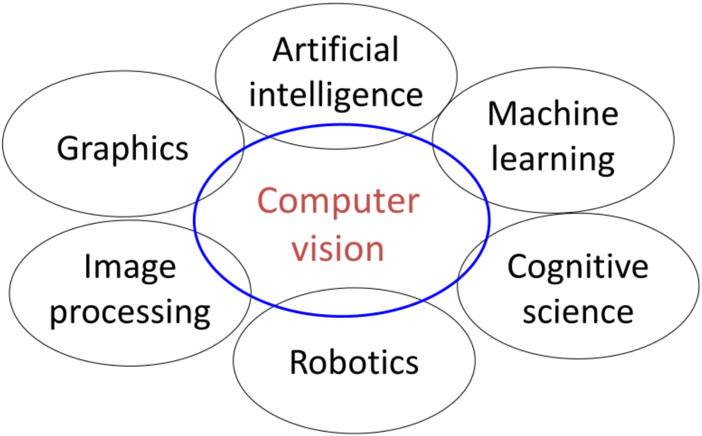


Figure: Related disciplines [4]

Table of contents

1. Introduction
2. Color
3. Image formation
4. Image transformations
5. Bibliography

Light

What is light?

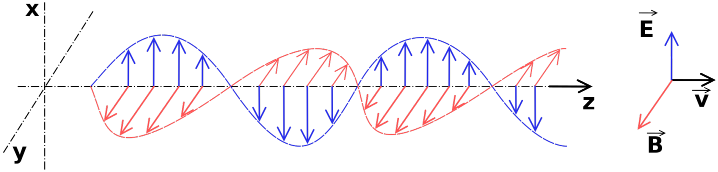


Figure: Electromagnetic wave

- Electromagnetic waves can be created by an oscillating current

Light

What is (visible) light?

- Portion of the electromagnetic spectrum perceived by the human eye.

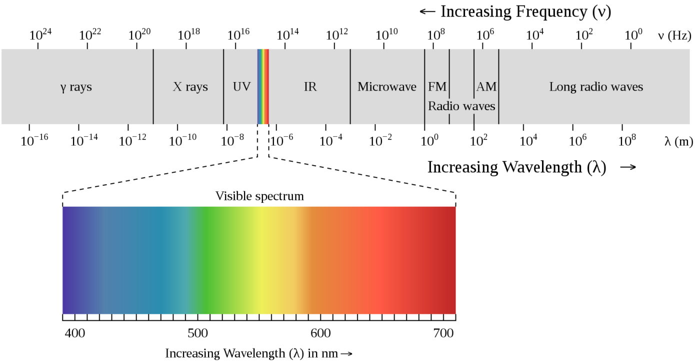


Figure: Electromagnetic spectrum and visible light

Light

- Objects reflect **multiple** wavelengths.

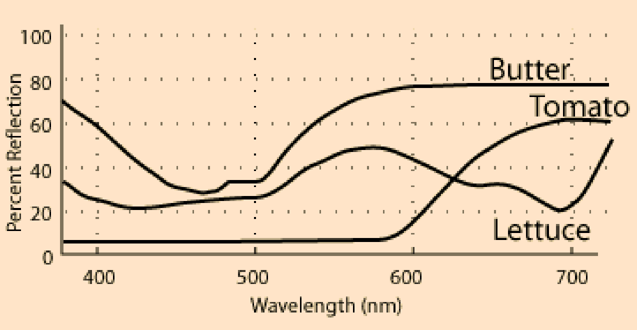


Figure: Examples of spectral power distributions

Color

What is color?

- “Color is a perception that depends on the response of the human visual system to light and the interaction of light with objects.” [2]

- Color is not light. Color is our human interpretation of light.

  It depends on:

  - Physical reality (electromagnetic radiation)
  - The measurement device (the human visual system)

The human eye

The human retina has:

- Cone cells (6M).
  - In charge of bright and medium light conditions.
  - Densely distributed in the fovea.
- Rod cells (120M).
  - In charge of dark conditions.

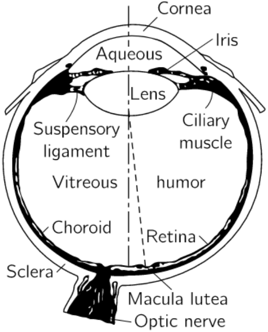


Figure: The human eye

The human eye

The optical nerve:

- Transmits visual information.
- It doesn't contain photo-receptor cells.
  - Therefore it creates a **blindspot**.

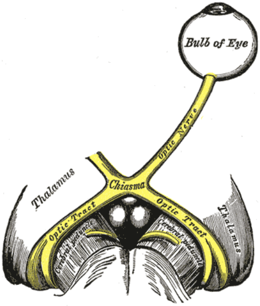


Figure: Optical nerve

The human eye

Testing the blind spot:

- Go to the next slide.
- Cover your right eye.
- Look at the dot (not the *sum*).
- Slowly move closer or farther until the *sum* is not visible.

The human eye

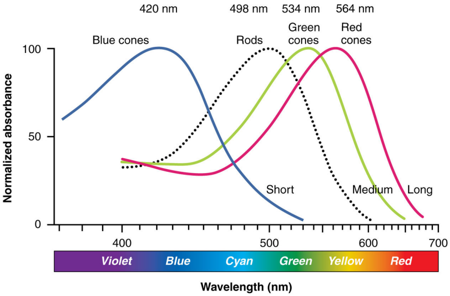


Figure: Sensitivity of cones and rods

- How can someone obtain the graph above?

  "The patient, a male Caucasian aged 46, had his right eye removed ..." [1]
- Our eye has **three** sensors responding to different stimulus.
- What would happen if we only had one?

Color spaces

In 1920 W. David Wright made the following experiment:

- Projecting **three** RGB lights were able to replicate **most** colors.

  $$T = \alpha R + \beta G + \gamma B$$

- For certain colors they needed *negative* light.

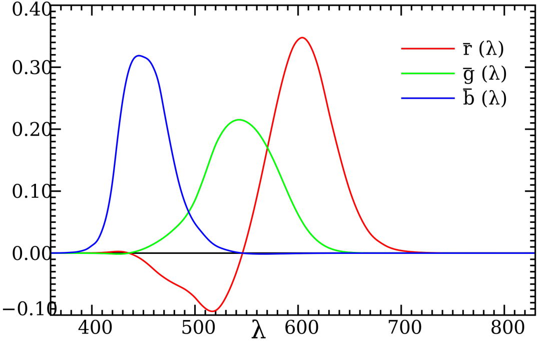


Figure: RGB color matching functions

Table of contents

1 Introduction

2 Color

3 Image formation

4 Image transformations

5 Bibliography

Image formation

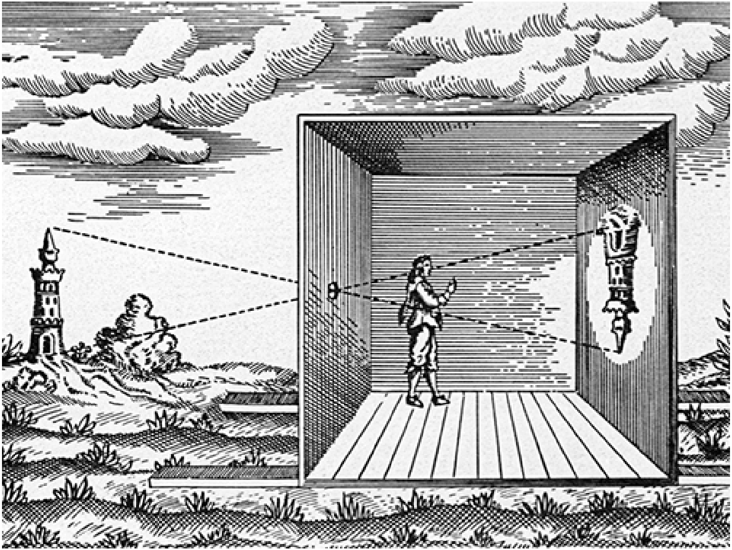


Figure: Camera Obscura

Image formation

A convex lens converges light from the object to the film.

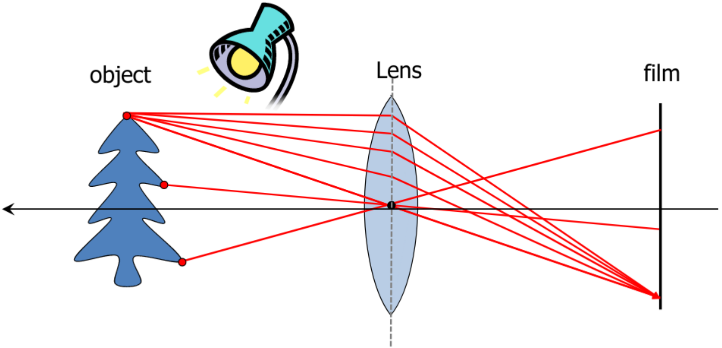


Figure: Image formation [4]

- What would happen if we didn't have the lense?

Image formation

Focal length measures how strongly light is converged or diverged.

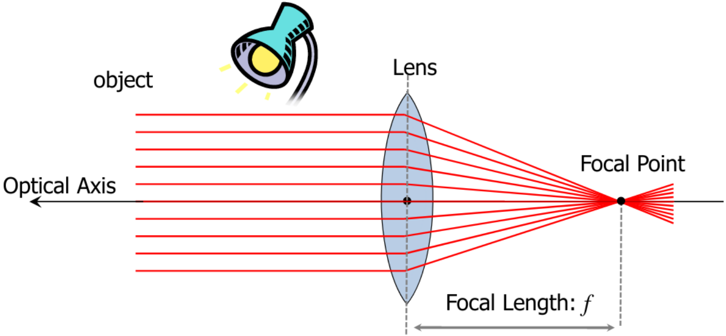


Figure: Image formation [4]

Image formation

We have two similar triangles

$$\frac{B}{A} = \frac{e}{z} \qquad\qquad \frac{B}{A} = \frac{e-f}{f} \tag{1}$$

From 1 we have that:

$$fe = z(e-f) \implies fe\frac{1}{fez} = z(e-f)\frac{1}{fez} \implies \frac{1}{f} = \frac{1}{z} + \frac{1}{e} \tag{2}$$

The last equation on the right of 2 is the *thin lens equation*.

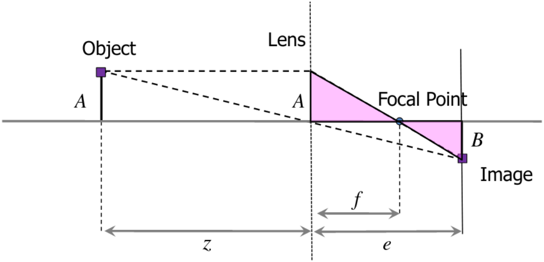


Figure: Image formation [4]

Image formation

- When the *thin lens equation* is satisfied the image is focused. (all rays from one point hit the film on the left on another point)

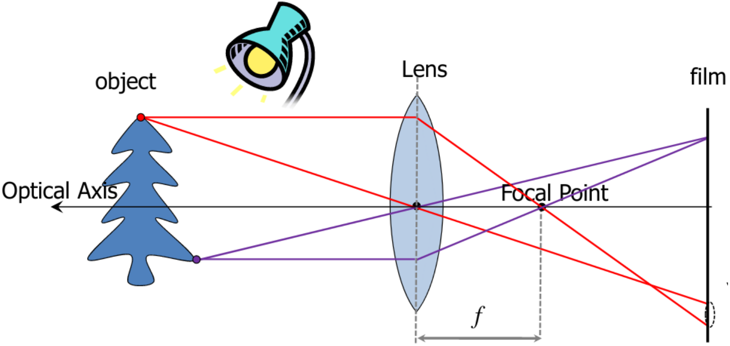


Figure: Image formation [4]

Image formation

Pinhole approximation: $z \gg f$

$$\frac{1}{f} = \frac{1}{z} - \frac{1}{e} \implies \frac{1}{f} \approx \frac{1}{e} \implies f \approx e \tag{3}$$

$$\frac{h^{'}}{h} = \frac{f}{z} \implies h^{'} = \frac{f}{z} h \tag{4}$$

Perspective:
The projected image dimensions are inversely proportional to the distance!

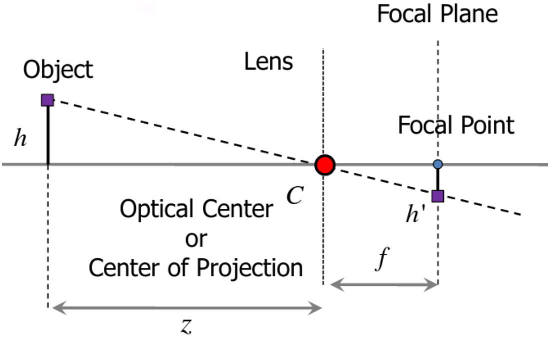


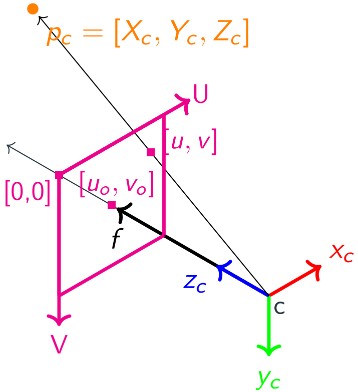
Figure: Pinhole approximation [4]

Image formation

Image formation

Using the pinhole approximation:

- We would like to convert 3D points (p_c) to image coordinates (u, v)



Note that the image plane (focal plane / film) is moved forward. This is physically invalid but convenient for calculations.

Image formation

Using our pinhole approximation 4 we know that:

$$u = f\frac{X_c}{Z_c} \tag{5}$$

$$v = f\frac{Y_c}{Z_c} \tag{6}$$

We would like to use pixels instead of the 3D dimensions (milimiters).

- Conversion to pixels is done using factors k_u and k_v
- k_u and k_v are pixel densities (milimiters / pixels) for u and v.

$$u = k_u f\frac{X_c}{Z_c} = \hat{f}_u\frac{X_c}{Z_c} \tag{7}$$

$$v = k_v f\frac{Y_c}{Z_c} = \hat{f}_v\frac{Y_c}{Z_c} \tag{8}$$

- With $\hat{f}_u := k_u f$ and $\hat{f}_v := k_v f$.

Image formation

Now we translate the origin $[u_o, v_o]$ to the left-top corner (easier indexing).

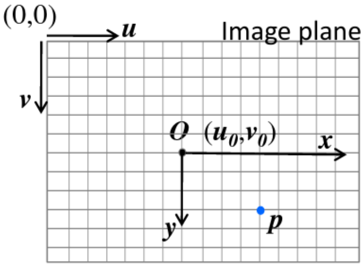


Figure: Perspective projection [4]

Thus we have:

$$u = u_o + k_u f \frac{X_c}{Z_c} \tag{9}$$

$$v = v_o + k_v f \frac{Y_c}{Z_c} \tag{10}$$

Image formation

We can represent equations 9 and 10 in matrix form:

$$\begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} = \begin{bmatrix} k_u f & 0 & u_0 \\ 0 & k_v f & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}$$

If we carry the computation explicitly:

$$\lambda u = k_u f X_c + u_o Z_c \tag{11}$$

$$\lambda v = k_v f Y_c + v_o Z_c \tag{12}$$

$$\lambda = Z_c \tag{13}$$

If we substitute λ and divide by Z_c in both sides we recover 9 and 10. Thus, we can represent our projection equations as a matrix (considering that a division with Z_c is required).

Image formation

This matrix is called the **camera intrinsics matrix** (K).

$$\begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} = \begin{bmatrix} k_u f & 0 & u_0 \\ 0 & k_v f & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = K \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}$$

Why the name intrinsics?

- K represents a physical model of a camera (pinhole camera model).
- This model assumes the pinhole approximation (equation 4)
- This model has the following parameters: k_u, k_v, u_o, u_v, f
- K represents what happens inside the camera thus the name intrinsic.
- Next lecture we will look at what happens outside of the camera.

Camera

- Cameras recreates the process of how our eye *captures* an image.
- Cameras requires multiple lenses to recreate the *thin lens* assumptions.

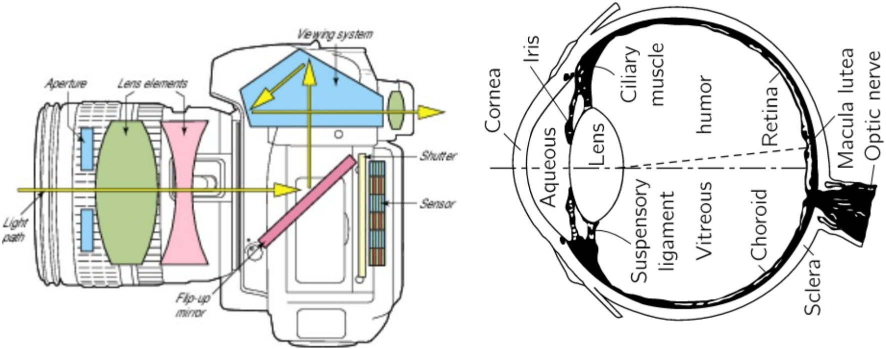


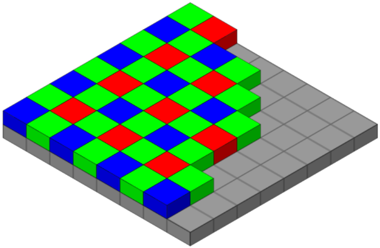
Figure: Side by side optical internals of a camera and a human eye.

Camera

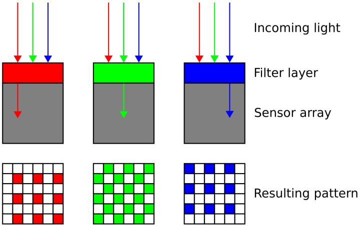
Invented by Bryce Bayer on 1976 while working on Kodak.

Twice as many green sensors

- Imitating human eye sensibility.



(a) Arrangement



(b) Filters

Figure: Bayer sensor arrangement

Table of contents

1 Introduction

2 Color

3 Image formation

4 Image transformations

5 Bibliography

Image transformations

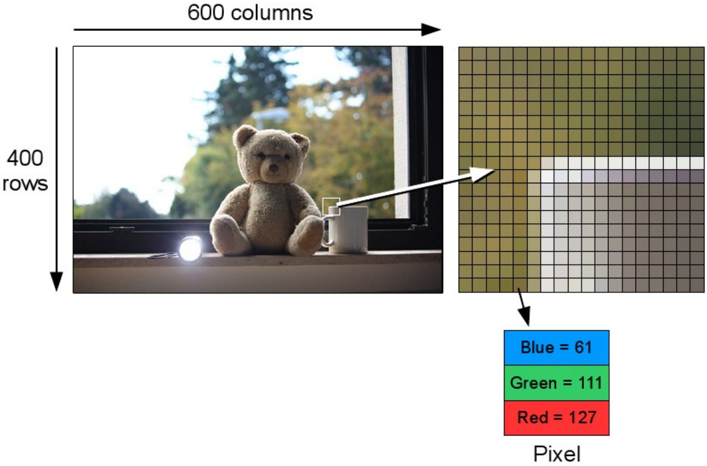


Figure: Image as pixels [5]

Image transformations

Images are often represented in RGB (OpenCV uses BGR) channels.

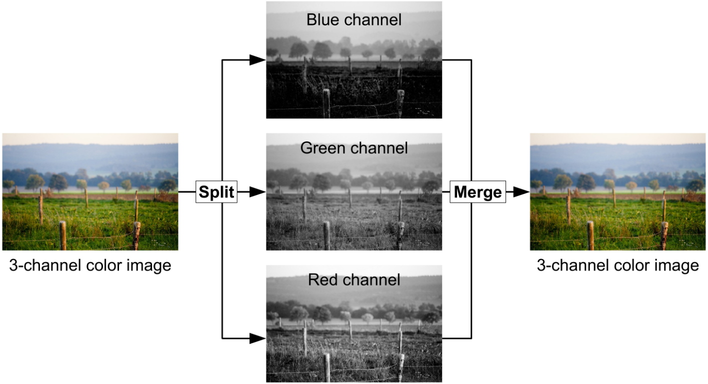


Figure: Image channel split [5]

Image transformations

Figure: Scalar multiplication to images [5]

Image transformations

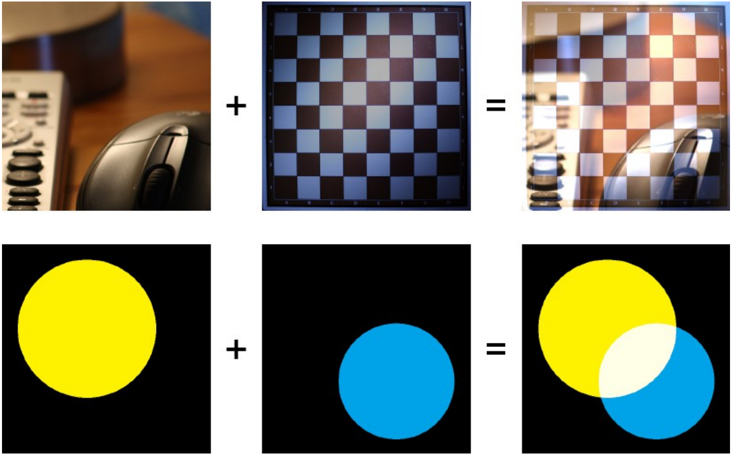


Figure: Sum between two images [5]

Image transformations

- Higher-level image processing might require binary images
  - We can apply a threshold operation
- Per-color threshold operations can help us create classifiers.
  - “How much red does our image has?”

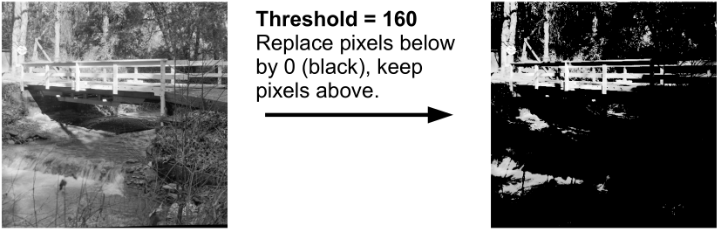
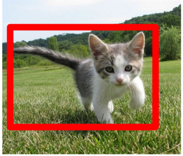


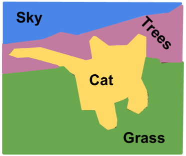
Figure: Threshold filters [5]

Computer vision tasks

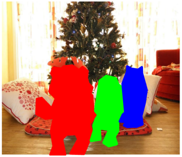
(a) Classification



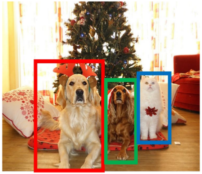
(b) Localization



(c) Segmentation



(d) Instance seg.



(e) Detection

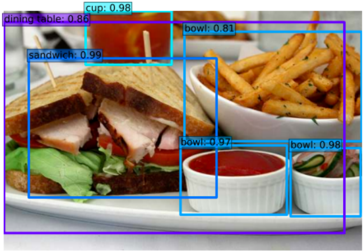
(f) Keypoints (HRNet)

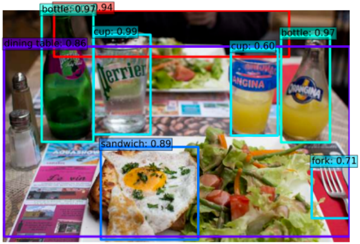
(g) 6D pose

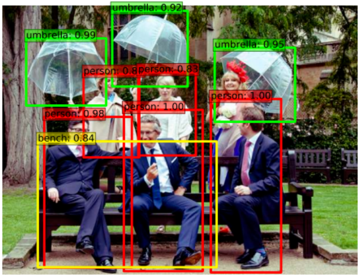
Figure: High-level computer vision tasks

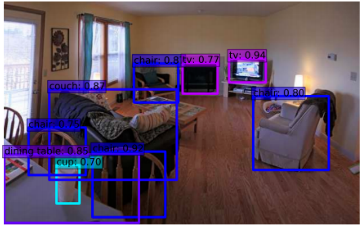
Object detection

Why is object detection hard?





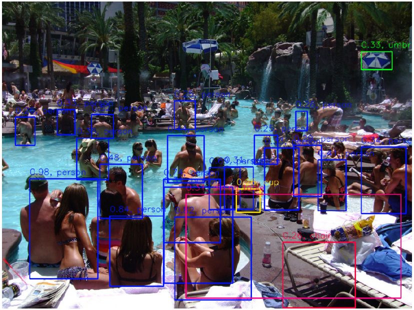




Object detection

Why is object detection hard?

- Intraclass variability
- Variable scales of classes
- Variable output of boxes
- Real-time capabilities



Object detection

Why is object detection hard?
Previous methods consisted on multi-scale sliding windows

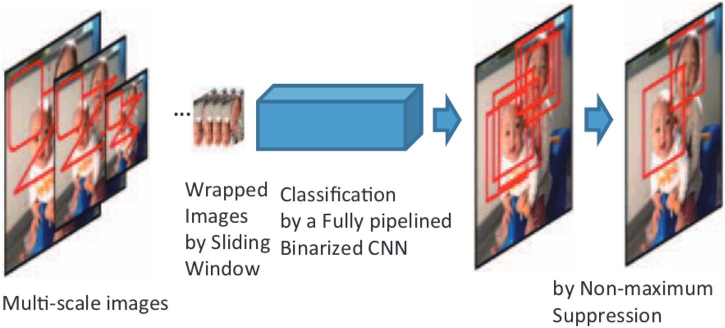
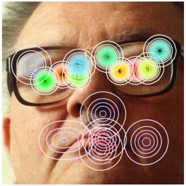
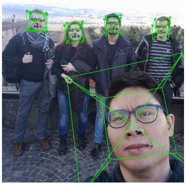


Figure: Multi-scale sliding window

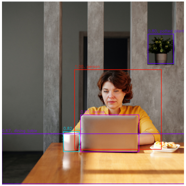
Examples implemented in github/oarriaga/paz



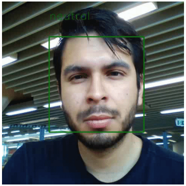
(a) Probabilistic kps



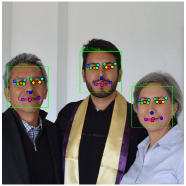
(b) Head-pose est.



(c) Object detection



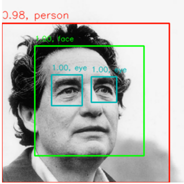
(d) Emotion recog.



(e) Keypoint est.

(f) Inst. segmentation

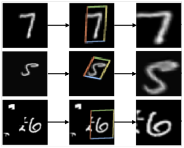
(g) Keypoint discovery



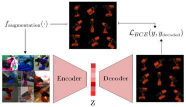
(h) Haar Cascades

(i) Pose estimation

(j) Face recognition



(k) Attention



(l) Implicit pose

Thank you for your attention :)

Questions?

Bibliography I

James K Bowmaker and HJk Dartnall.
Visual pigments of rods and cones in a human retina.
The Journal of physiology, 298(1):501–511, 1980.

Fair Child.
What is color?

Yi Ma, Stefano Soatto, Jana Kosecka, and S Shankar Sastry.
An invitation to 3-d vision: from images to geometric models, volume 26.
Springer Science & Business Media, 2012.

Davide Scaramuzza.
Vision algorithms for mobile robotics, 2019.

David Scherfgen.
Introduction to opencv.
University Lecture, 2020.