Image Procesing and Computer Vision Notes

by Mattia Orlandi

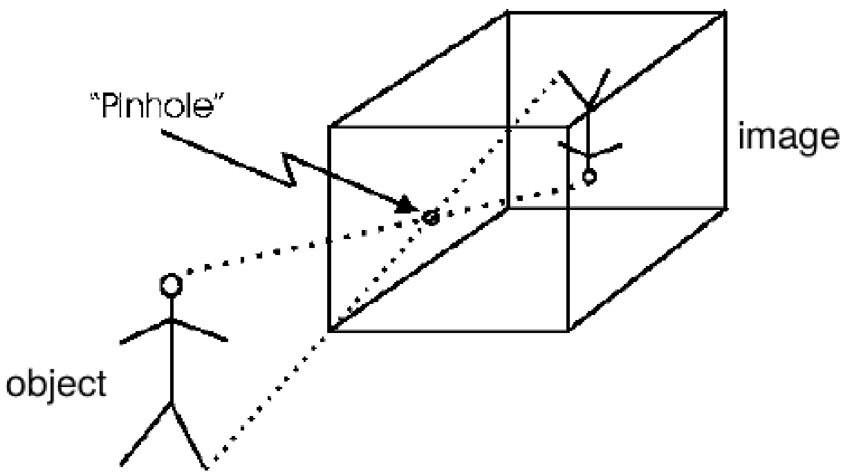
2. Image Formation and Acquisition

An imaging device gathers the light reflected by 3D objects to create a representation in 2D of the scene; Computer Vision tries to invert such a process, so as to infer knowledge on the objects from one or more digital images. This requires studying:

- the geometric relationship between scene points and image points;
- the radiometric relationship between the brightness of image points and the light emitted by scene points;
- the image digitalization process.

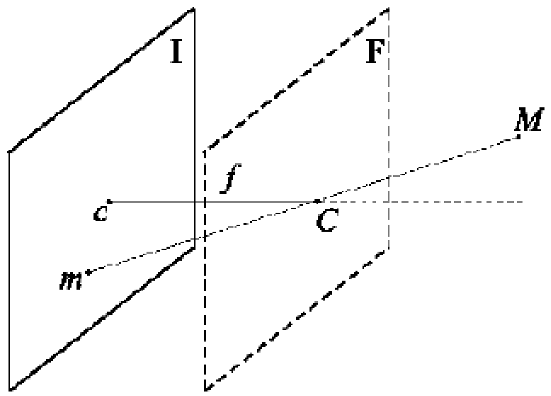
Pinhole camera

It's the simplest imaging device, light goes through the very small pinhole and hits the image plane, in which a film sensible to light captures the image (flipped).



2.1. Perspective Projection

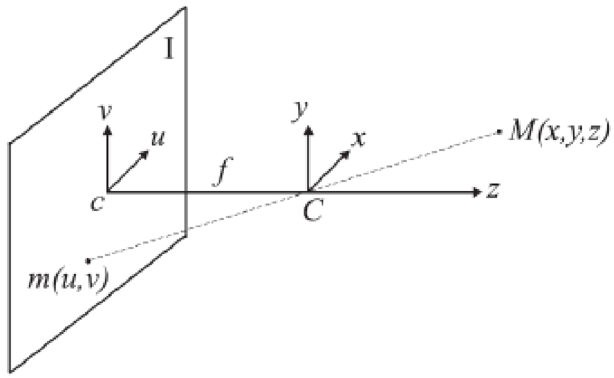
It's the geometric model of image formation in a pinhole camera.



M : scene point
m : corresponding image point
I : image plane
C : optical centre
Line through *C* and orthogonal to *I* : optical axis
c : intersection between optical axis and image plane (image centre or piercing point)
f : focal length
F : focal plane

Using the following reference system the equations mapping scene points into image points are:

$$\frac{u}{x} = \frac{v}{y} = -\frac{f}{z} \Rightarrow \begin{cases} u = -x \cdot f/z \\ v = -y \cdot f/z \end{cases}$$



The image plane can be thought of as lying in front of rather than behind the optical centre, so that the flipping does not happen:

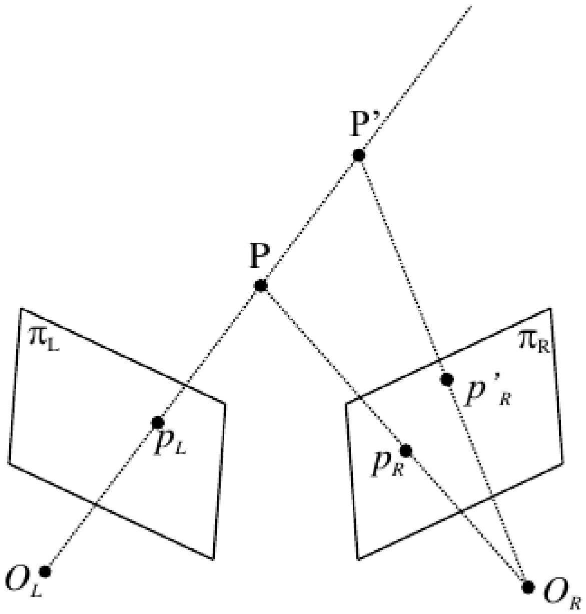
$$\begin{cases} u = x \cdot f/z \\ v = y \cdot f/z \end{cases}$$

where:

- x, y are the lateral coordinates;
- z is the depth coordinate;
- the equations are non-linear;
- points far from the optical centre are more scaled in the resulting image than the points near it;
- the mapping is not a bijection: in fact, in the process of representing a 3D subject in a 2D image there is a loss of information (an image point is mapped into a 3D line, thus it is not possible to recover the 3D structure).

Stereo Images

Stereo Images, captured by two cameras, allow to infer 3D information; in fact, given correspondences between the two resulting images, 3D data can be recovered by triangulation.



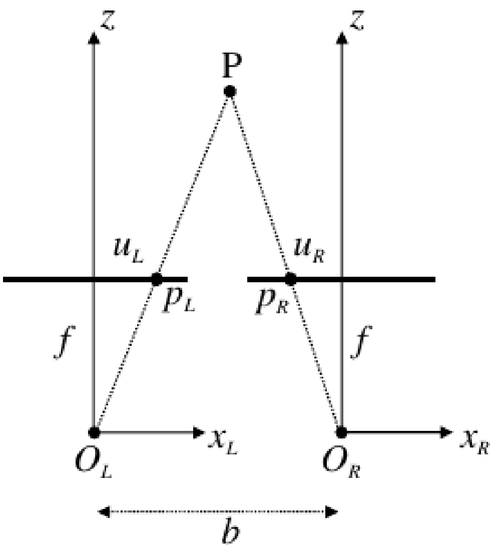
Standard Stereo Geometry

Given:

- two reference systems O_L (left) and O_R (right) with parallel y and z axes;
- same focal length for both reference systems;
- coplanar image planes;

then the transformation is just a translation along the x axis:

$$\begin{cases} p_L = [x_L \quad y_L \quad z_L] \\ p_R = [x_R \quad y_R \quad z_R] \end{cases} \Rightarrow p_L - p_R = [b \quad 0 \quad 0]$$



$$\begin{cases} v_L = v_R = y \cdot f/z \\ u_L = x_L \cdot f/z \\ u_R = x_R \cdot f/z \end{cases} \Rightarrow u_L - u_R = b \cdot f/z = d$$

where d is called **disparity**.
Given the focal length, the translation between the two cameras and the disparity, it's possible to compute the depth of an image:

$$z = b \cdot f/z$$

Given a point P_L in the left image, to compute the disparity one must be able to determine which point P_R on the right image is the projection of the same 3D point P (*stereo correspondence* problem). In case of two cameras with perfectly parallel axes, two corresponding points lie on the same row (1D search space).

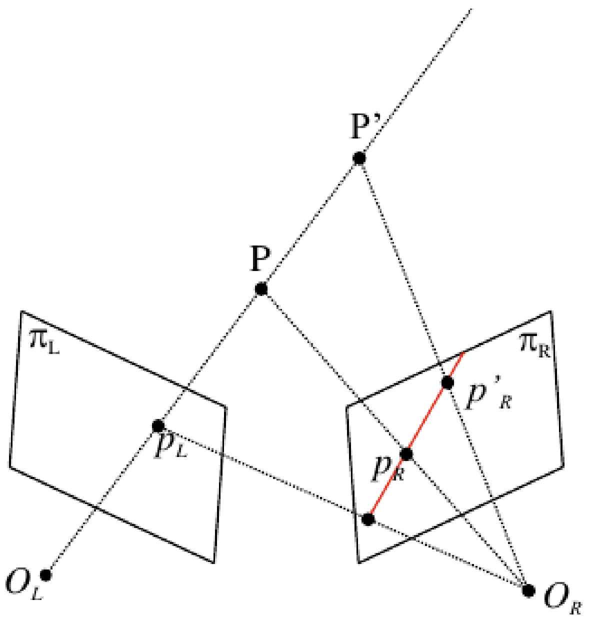
In real-world scenarios, it's impossible to make sure that the cameras have perfectly parallel axes $\Rightarrow$ *epipolar geometry*.

Epipolar Geometry

The two cameras have not parallel axes:

- given a point P, the line connecting it to the optical center of the left camera O_L is seen as a point by the left camera, since it is in line with its optical center;
- that line is seen by the right camera as a line, which is called **epipolar line**;
- the same holds for the opposite case;
- all the epipolar lines in an image meet at the *epipole*, that is the projection of the optical center of the other image.

Therefore, an epipolar line is a function of the position of point P in the 3D space: as P varies, a set of epipolar lines is generated in both image planes.



Given a point P_L in the left image, the corresponding point P_R in the right image lie on the respective epipolar line, so the search space is still 1D, but search would be performed on oblique lines $\Rightarrow$ images are warped as if they were acquired through a standard stereo geometry, i.e. both images have horizontal and collinear conjugate epipolar lines (homography known as **rectification**).

Properties of Perspective Projection

- The farther objects are from the camera, the smaller they appear in the image; a line with real length L parallel to the image plane at distance z will exhibit a length $l = L\frac{f}{z}$.

- Perspective Projection maps 3D lines into image lines.
- Ratios of lengths are not preserved, unless scene is planar and parallel to image plane.
- Parallelism between 3D lines is not preserved (unless lines are parallel to image plane).

Vanishing points

- When parallel 3D lines are projected into the image plane, the corresponding 2D lines meet at a point called *vanishing point*, which is the projection of a 3D point infinitely distant from optical center.
- If the 3D lines are parallel to image plane, then the vanishing point will be at infinity.
- Given the following line:

$$M = M_0 + \lambda D = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + \lambda \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

  where M_0 is a point on the line and D is the direction cosine vector, the vanishing point is calculated by projecting a generic point of the line:

$$m = \begin{bmatrix} u \\ v \end{bmatrix}, \ u = f\frac{x_0 + \lambda a}{z_0 + \lambda c}, \ v = f\frac{y_0 + \lambda b}{z_0 + \lambda c}$$

  and then by considering the limit of the point tending towards infinity:

$$m_\infty = \begin{bmatrix} u_\infty \\ v_\infty \end{bmatrix}, \ u_\infty = \lim_{\lambda \to \infty} u = f\frac{a}{c}, \ v_\infty = \lim_{\lambda \to \infty} v = f\frac{b}{c}$$

  The vanishing point depends on the orientation of the line only and not on its position.
  When the line is parallel to image plane ($c = 0$) it goes to infinity, and in that case the 3D line and the corresponding 2D line have the same orientation:

$$\begin{cases} u_\infty = f\frac{a}{c} \\ v_\infty = f\frac{b}{c} \\ a^2 + b^2 + c^2 = 1 \end{cases}$$

$$\Leftrightarrow c^2(u_\infty^2 + v_\infty^2) = f^2(1 - c^2) \Leftrightarrow c = \frac{f}{\sqrt{u_\infty^2 + v_\infty^2 + f^2}}$$

$$\Leftrightarrow a = \frac{u_\infty}{\sqrt{u_\infty^2 + v_\infty^2 + f^2}}, \ b = \frac{v_\infty}{\sqrt{u_\infty^2 + v_\infty^2 + f^2}}$$

$$\Leftrightarrow \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \frac{1}{\sqrt{u_\infty^2 + v_\infty^2 + f^2}} \begin{bmatrix} u_\infty \\ v_\infty \\ f \end{bmatrix}$$

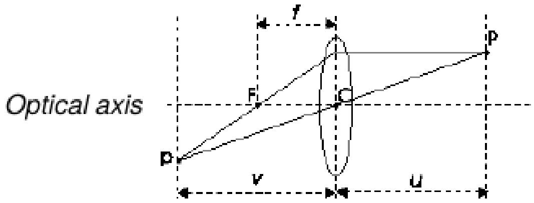
2.2. Lenses

- A scene point is *on focus* when all its light rays gathered by the camera hit the image plane at the same point.
- In a pinhole camera every scene point is on focus because of the very small size of the hole $\Rightarrow$ infinite **Depth of Field** (DoF).

- Small aperture $\Rightarrow$ very limited amount of light $\Rightarrow$ very long exposure times.
- To avoid this, cameras rely on lenses to gather more light from a scene point and focus it on a single image point $\Rightarrow$ smaller exposure times, but limited DoF (only points across limited range of distances are on focus at the same time).

Thin lens equation

- Approximate model featuring only one lens (in real-world scenarios, cameras feature complex optical systems with multiple lenses).



$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$$

P : scene point
p : corresponding focused image point
u : distance from P to the lens
v : distance from p to the lens
f : focal length (parameter of the lens)
C : centre of the lens
F : focal point (or focus) of the lens

- Rays parallel to the optical axis are deflected to pass through F.

- Rays passing through C are undeflected.

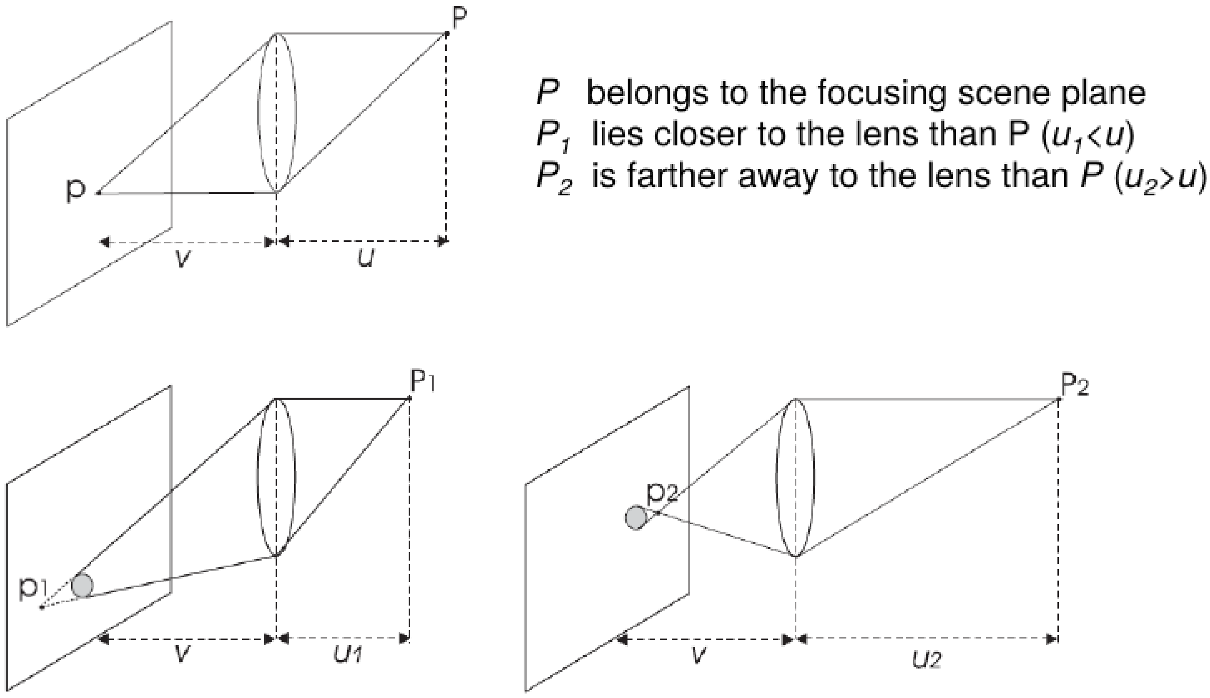
- If image is on focus, image formation process obeys to perspective projection model, with the center of the lens being the optical center and the distance v acting as the *effective focal length of the projection* ($\neq f$, focal length *of the lens*).

Circles of Confusion

- Distance of the image plane v and distance of the focusing plane u are bounded:

  - With v fixed (distance of image plane): $\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \Rightarrow u = \frac{vf}{v-f}$
  - With u fixed (distance of scene points): $\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \Rightarrow v = \frac{uf}{u-f}$

- Given the chosen position of image plane v, scene points in front of the focusing plane or behind it will be out-of-focus, appearing as circles rather than points (*Circles of Confusion* or *Blur Circles*).

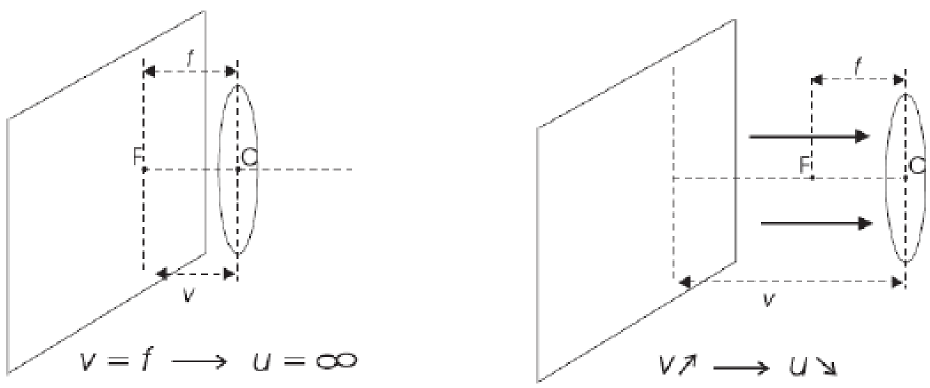


- As long as such circles are smaller than the size of photosensing elements, image will still look on-focus.
- The range of distances across which the image appears on-focus determines the DoF of the lens.
- An adjustable diaphragm (iris) enables to control the amount of light gathered through the *effective aperture* of the lens: closer diaphragm aperture $\Rightarrow$ smaller size of blur circles $\Rightarrow$ larger DoF.
- The *F-number* is the ratio, expressed in discrete units (called *stops*), of the focal length to the effective aperture of the lens: higher stop $\Rightarrow$ closer diaphragm aperture $\Rightarrow$ larger DoF.

Focusing mechanism

- To focus an object at diverse distances, a mechanism allows the lens to translate along the optical axis w.r.t. the fixed position of image plane.

$$\frac{1}{u}+\frac{1}{v}=\frac{1}{f}$$



- At one end position ($v = f$) the camera is focused at infinity, whereas at the other end position (where v is maximum) the focusing distance is minimum.

2.3. Image Digitalization

- The image plane of a camera consists of a planar sensor which converts the amount of light incident to any point (*irradiance*) into an electric quantity (e.g. voltage).
- Such a continuous "electric" image is sampled and quantized to end up with a digital image suitable to visualization and processing by a computer:
  - **Sampling**: planar continuous image is sampled evenly along horizontal and vertical directions to pick up a 2D array (matrix) of $N \times M$ samples known as *pixels*:

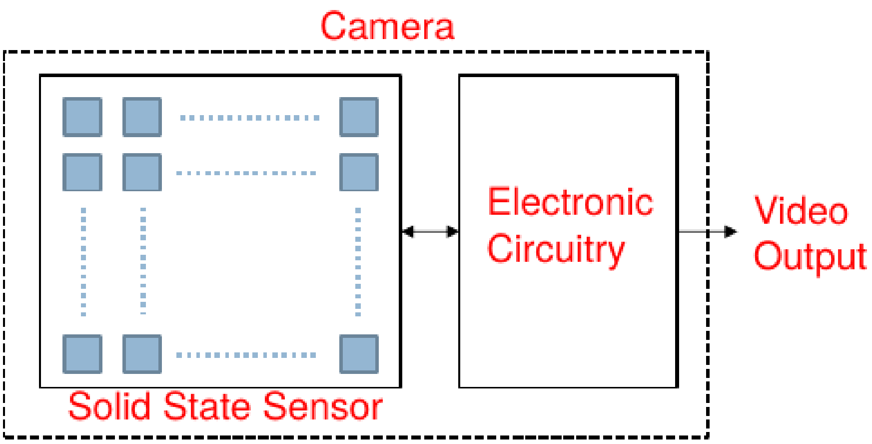
$$I(x,y) \Rightarrow \begin{bmatrix} I(0,0) & \dots & I(0,M-1) \\ \vdots & \ddots & \vdots \\ I(N-1,0) & \dots & I(N-1,M-1) \end{bmatrix}$$

  - **Quantization**: the continuous range of values associated with pixels is quantized into $l = 2^m$ discrete levels known as *gray-levels*, where m is the number of bits used to represent a pixel (usually, $m = 8$); thus, the memory occupancy in bits of a gray-scale image is $B = N \cdot M \cdot m$ (colour digital images are instead represented using three bytes per pixel, one for each RGB channel).
- The more bits are used for its representation, the higher the quality of a digital image.

Digitalization in detail

- The sensor is a 2D array of photodetectors, and during exposure time each of them converts incident light into a proportional electric charge.

- The companion circuitry reads-out the charge to generate the output signal, digital (ADC necessary) or analog (for legacy systems).



- There is never a continuous image since it is sensed directly as a sampled image.

- In analog cameras the native sampling is lost in the generation of the analog output, which is then sampled and quantized by a dedicated circuitry known as *analog frame grabber*: as a result, pixels in digital image coming from analog cameras do not correspond to those sensed by photodetectors.

- The two main sensor technologies are CCD (Charge Coupled Devices) and CMOS (Complementary Metal Oxide Semiconductor).

Camera Parameters

- **Signal-to-Noise Ratio (SNR)**:
  - Intensity measured at a pixel under perfectly static conditions varies due to random noise $\Rightarrow$ pixel value not deterministic but rather a random variable;
  - Main noise sources:
    - *Photon Shot Noise*: time between photon arrivals at a pixel governed by Poisson statistics $\Rightarrow$ number of photons collected during exposure time not constant.
    - *Electronic Circuitry Noise*: generated by electronics which reads-out charge and amplifies resulting voltage signal.
    - *Quantization Noise*: related to final ADC conversion (in digital cameras).
    - *Dark Current Noise*: random amount of charge due to thermal excitement observed at each pixel even though sensor is not exposed to light.
  - It quantifies the strength of the true signal w.r.t unwanted fluctuations induced by noise (the higher, the better).
  - Expressed in decibels or bits:

$$\mathrm{SNR}_{dB} = 20 \cdot \log(\mathrm{SNR});\ \mathrm{SNR}_{bit} = \ln(\mathrm{SNR})$$

- **Dynamic Range (DR)**:
  - If sensed amount of light is too small, true signal cannot be distinguished from noise.
  - Given *minimum detectable irradiation* E_{min} and *saturation irradiation* E_{max} (amount of light that would fill the photodetectors' capacity), the DR is defined as $\mathrm{DR} = \frac{E_{max}}{E_{min}}$, specified in decibels or bits.
  - The higher the DR, the better is the ability of the sensor to simultaneously capture both dark and bright structures of the scene.
  - High Dynamic Range (HDR) combines a sequence of images of the same subject taken with different exposure times.
- **Sensitivity (Responsivity)**: amount of signal that sensor can deliver per unit of input optical energy.
- **Uniformity (spatial or pattern noise**): due to manufacturing tolerances both the response to light and the amount of dark noise vary across pixels.

Sensors

- CCD provides higher SNR, higher DR and better uniformity.

- CMOS provides more compactness, less power consumption and lower system cost (thanks to the fact that electronic circuitry is integrated within the same chip as the sensor $\Rightarrow$ "one chip camera"); moreover, it allows an arbitrary window to be read-out without having to receive the full image (useful to inspect at higher speed a small Region of Interest, or ROI, within the image).

- CCD/CMOS are sensitive to light ranging from near-ultraviolet (200 nm) through visible spectrum (380-780 nm) up to near-infrared (1100 nm).

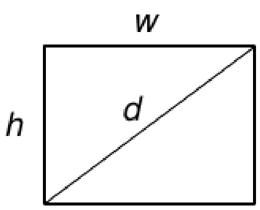
- Sensed intensity at a pixel results from the integration over the range of wavelengths of the spectral distribution of incoming light multiplied by the spectral response function of the sensor $\Rightarrow$ CCD/CMOS cannot sense colour.

- To create a colour sensor, an array of optical filters (Colour Filter Array) is placed in front of photodetectors, so as to render each pixel sensitive to a specific range of wavelengths (in Bayer CFA, green filters are twice as much as red and blue ones to mimic higher sensitivity of human eyes in the green range); to obtain an RGB triplet at each pixel, missing samples are interpolated from neighbouring pixels (*de-mosaicing*).

- True resolution of the sensor is smaller due to the green channel being subsampled by a factor of 2, the red and blue ones by 4.

- A more expensive full resolution colour sensor can be achieved by using an optical prism to split incoming light beam into 3 RGB beams sent to 3 distinct sensors equipped with corresponding filters.

- CCD/CMOS sensors come in different sizes specified in inches for the sake of legacy.



$$\frac{w}{h} = \frac{4}{3}$$

Size (inch)	Width (mm)	Height (mm)	Diagonal (mm)	VGA Pixel Size (m)
1	12.8	9.6	16	20
2/3	8.8	6,6	11	13.8
1/2	6.4	4.8	8	10
1/3	4.8	3.6	6	7.5
1/4	3.2	2.4	4	5