University of Bologna

Image Formation and Acquisition (Part 1)

Luigi Di Stefano (luigi.distefano@unibo.it)

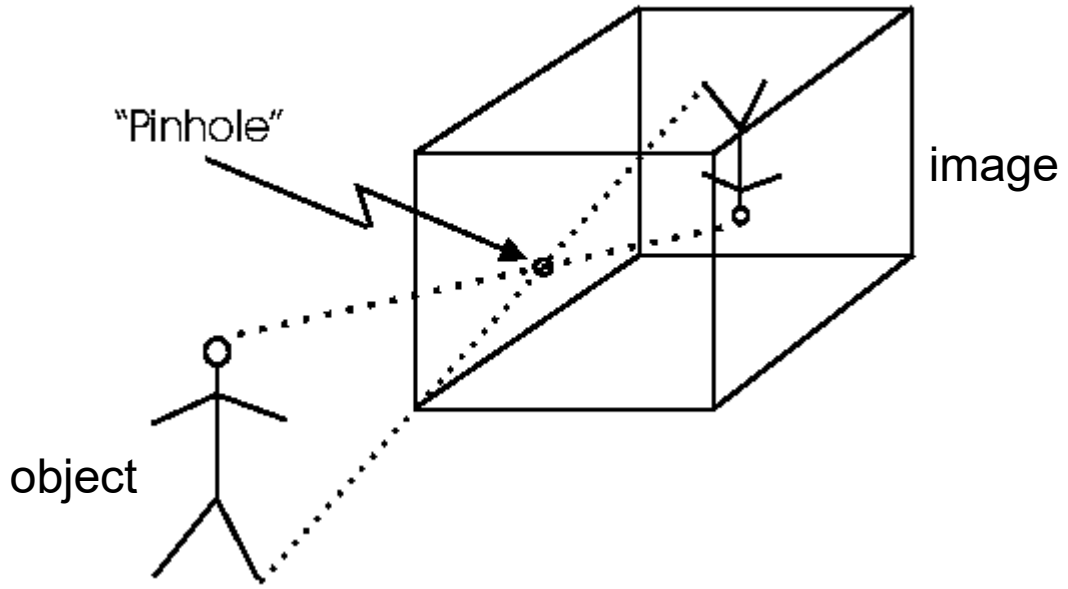
Introduction

- **An imaging device gathers the light reflected by 3D objects to create a 2D representation of the scene (i.e. the image). In computer vision we basically try to invert such a process, so as to infer knowledge on the objects from one or more digital images. It is therefore worth to understand the image formation and acquisition process. We will study:**
  - **The geometric relationship between scene points and image points.**
  - **The image digitization process.**

Pinhole Camera

The “pinhole camera” is the simplest imaging device: light goes through the very small pinhole and hits the image plane.

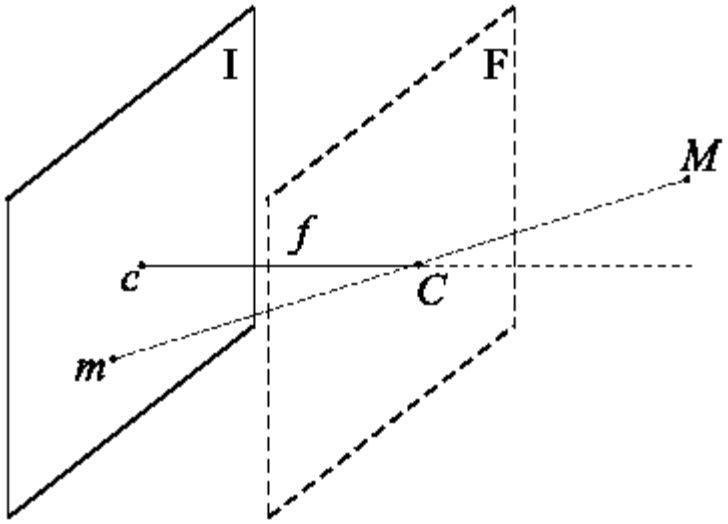
Geometrically, the image is achieved by drawing straight rays from scene points through the hole up to the image plane.



Although useful images can hardly be captured by means of a pinhole camera, its remarkably simple geometrical model turns out a good approximation of the geometry of image formation in most modern imaging devices.

Perspective Projection (1)

The geometric model of image formation in a pinhole camera is known as Perspective Projection.

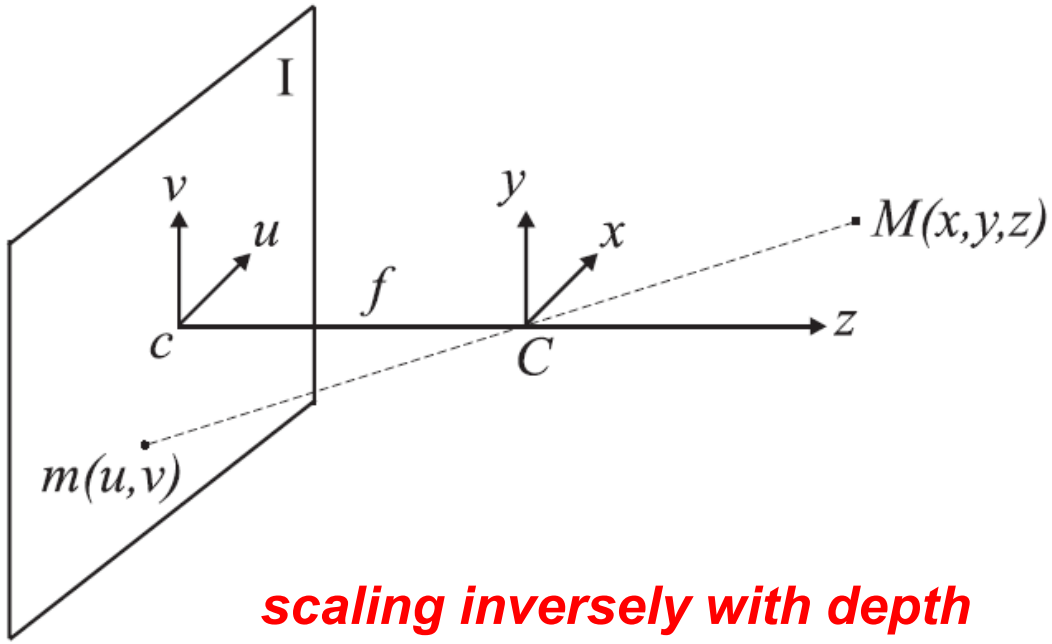


M : scene point
m : corresponding image point
I : image plane
C : optical centre
Line through *C* and orthogonal to *I* : optical axis
c : intersection between optical axis and image plane (image centre or piercing point)
f : focal length
F : focal plane

Perspective Projection (2)

Considering the reference frames shown in the figure, the equations to map scene points into their corresponding image points are as follows:

$$\frac{u}{x} = \frac{v}{y} = -\frac{f}{z} \rightarrow u = -x\frac{f}{z}; v = -y\frac{f}{z}$$



scaling inversely with depth

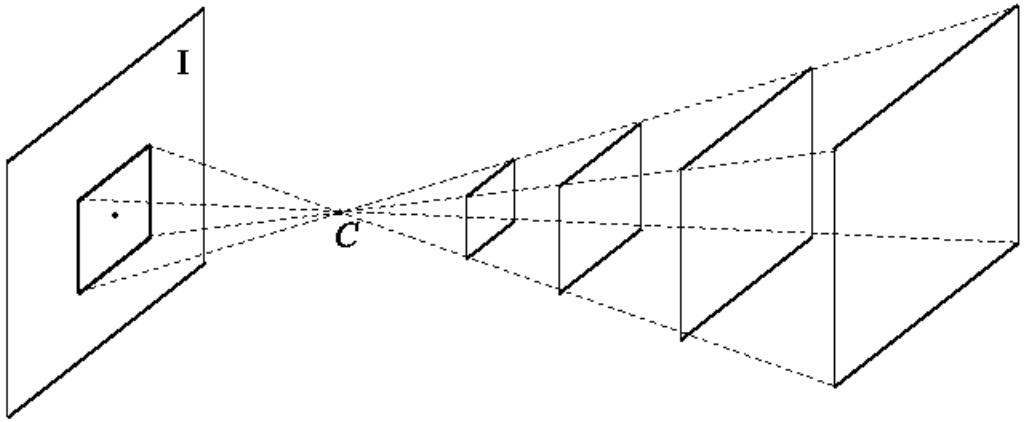
To get rid of the up-down and left right inversions, the image plane can be thought of as lying in front rather than behind the optical centre

$$u = x\frac{f}{z}; \quad v = y\frac{f}{z}$$

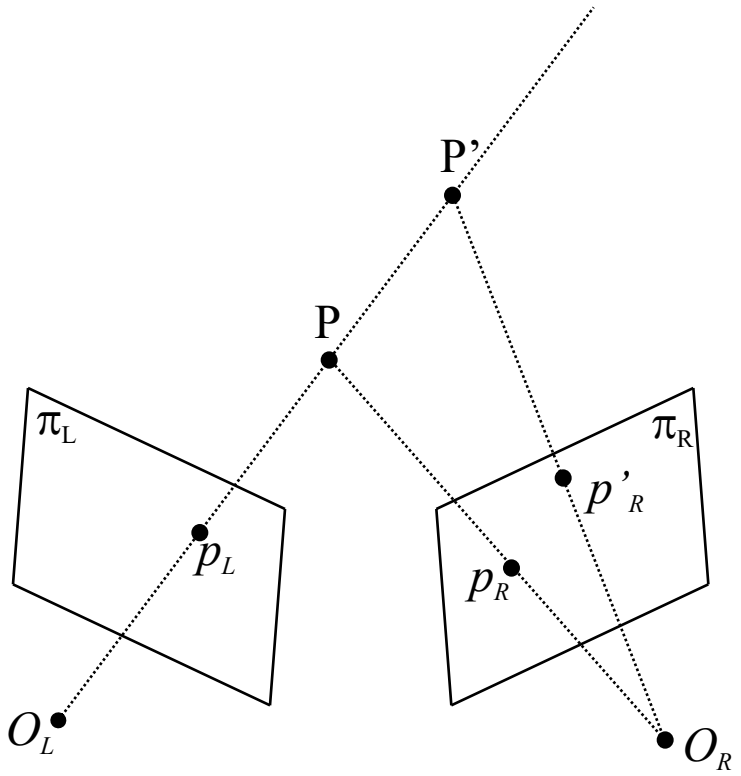
The notion of <u>scale</u> pertains to the apparent size of objects in images, which depends on the true size, the distance from the camera and the focal length. In many image recognition applcations we wish to achieve <u>scale-invariance.</u>

Perspective Projection (3)

- The image formation process deals with mapping a 3D space onto a 2D space, thus leading inevitably to loss of information.
- Indeed, the mapping is not a bijection: a given scene point is mapped into a unique image point, but a given image point is mapped onto a 3D line (i.e. the line through the point, *m*, and the optical centre, *C*).
- Thus, recovering the 3D structure of a scene from a single image is an ill-posed problem (the solution is not unique), as once we take an image point we can only state that its corresponding scene point lays on a line, but cannot disambiguate a specific 3D point along such a line (i.e. we know nothing about the distance to the camera).



Stereo images allow to infer 3D



Given correspondences, 3D information can be recovered by triangulation

Standard stereo geometry

- **Parallel (x,y,z) axes: the transformation between the two reference frames is just a translation (*b*), usually horizontal:**

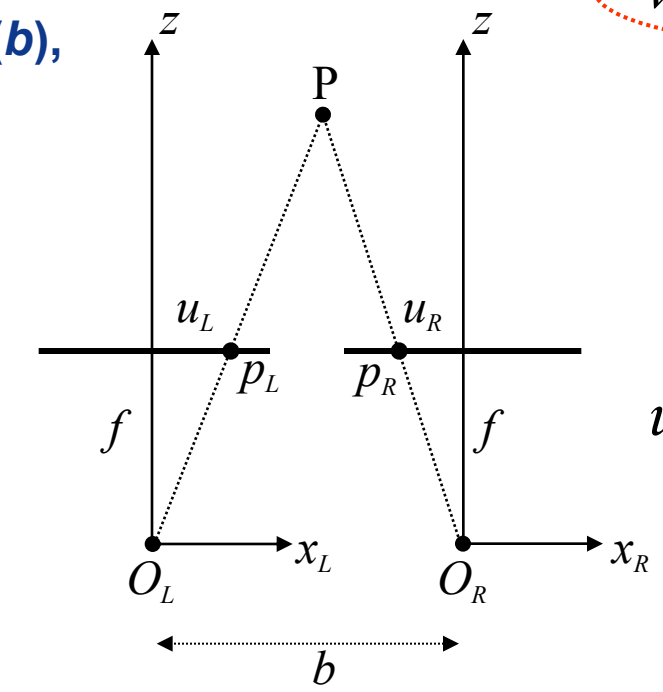
$$\mathbf{P}_L - \mathbf{P}_R = \begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix}$$

- **Same focal length (coplanar image planes)**

$$x_L - x_R = b$$

$$y_L = y_R = y$$

$$z_L = z_R = z$$



$$v_L = v_R = y \cdot f/z$$

$$u_L = x_L \cdot f/z$$

$$u_R = x_R \cdot f/z$$

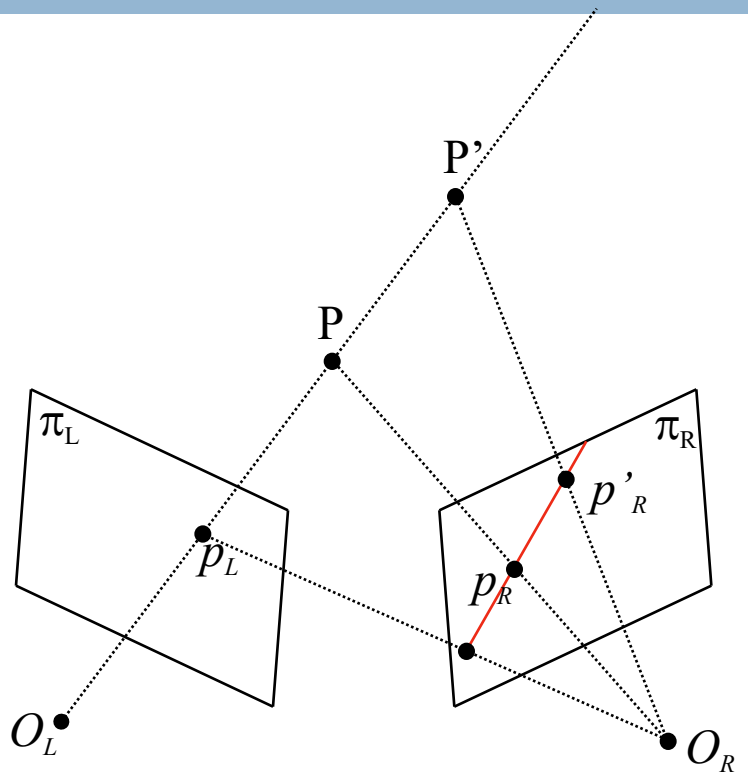
$$u_L - u_R = b \cdot f/z$$

$$u_L - u_R = d \quad (disparity)$$

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

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

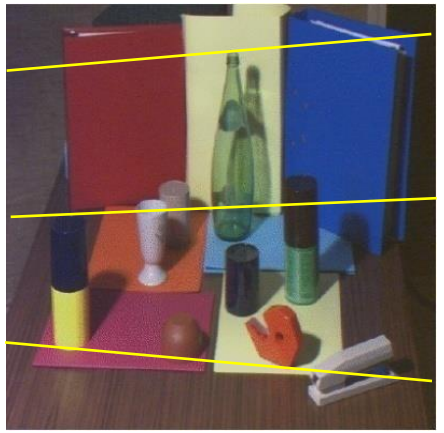
Epipolar Geometry



Epipolar line
(associated with p_L in Π_R)

The search space of the stereo correspondence problem is always 1D !

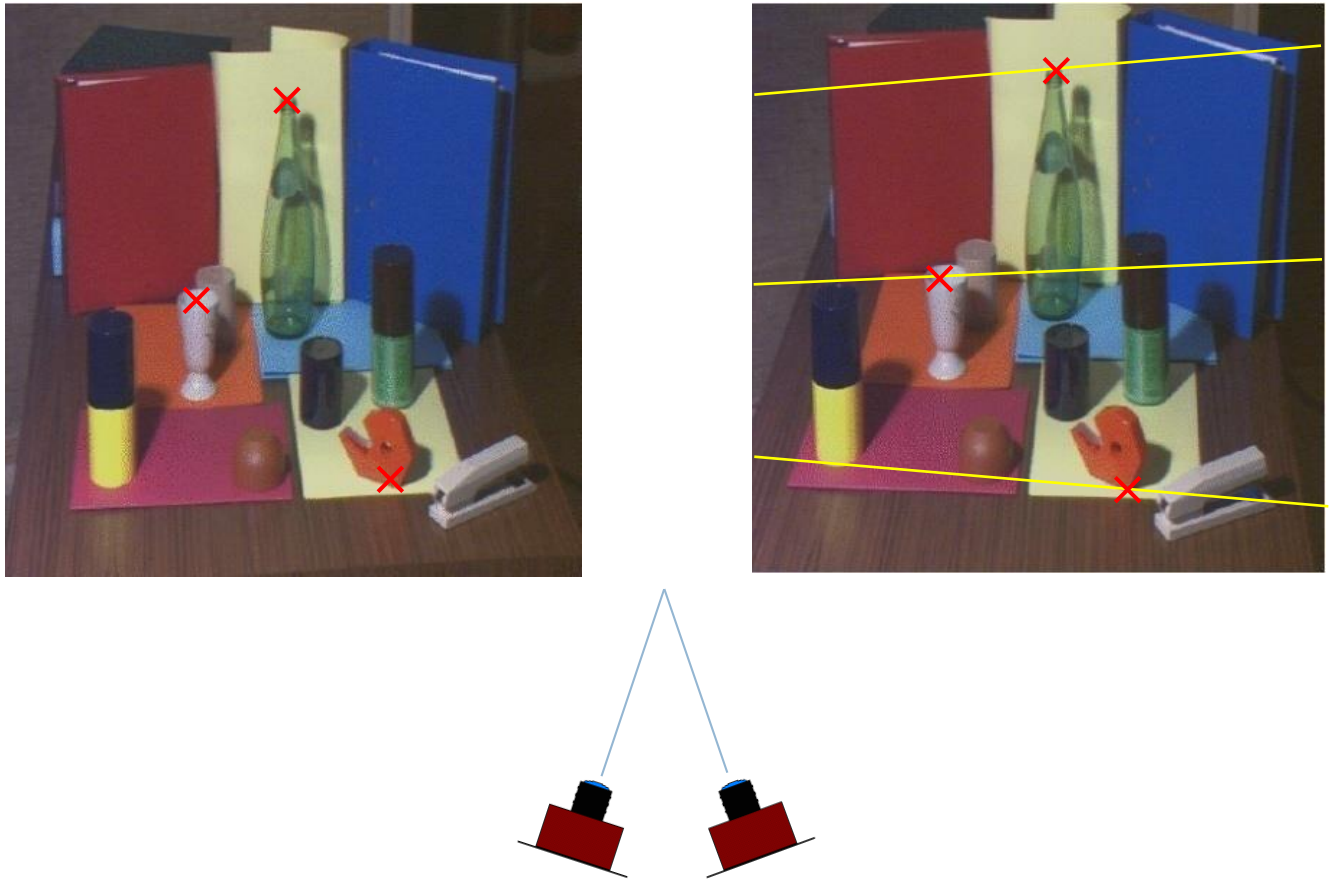
All the epipolar lines in an image meet at a point called *epipole* (i.e. the projection of the optical center of the other image)



However, searching through oblique epipolar lines is awkward !

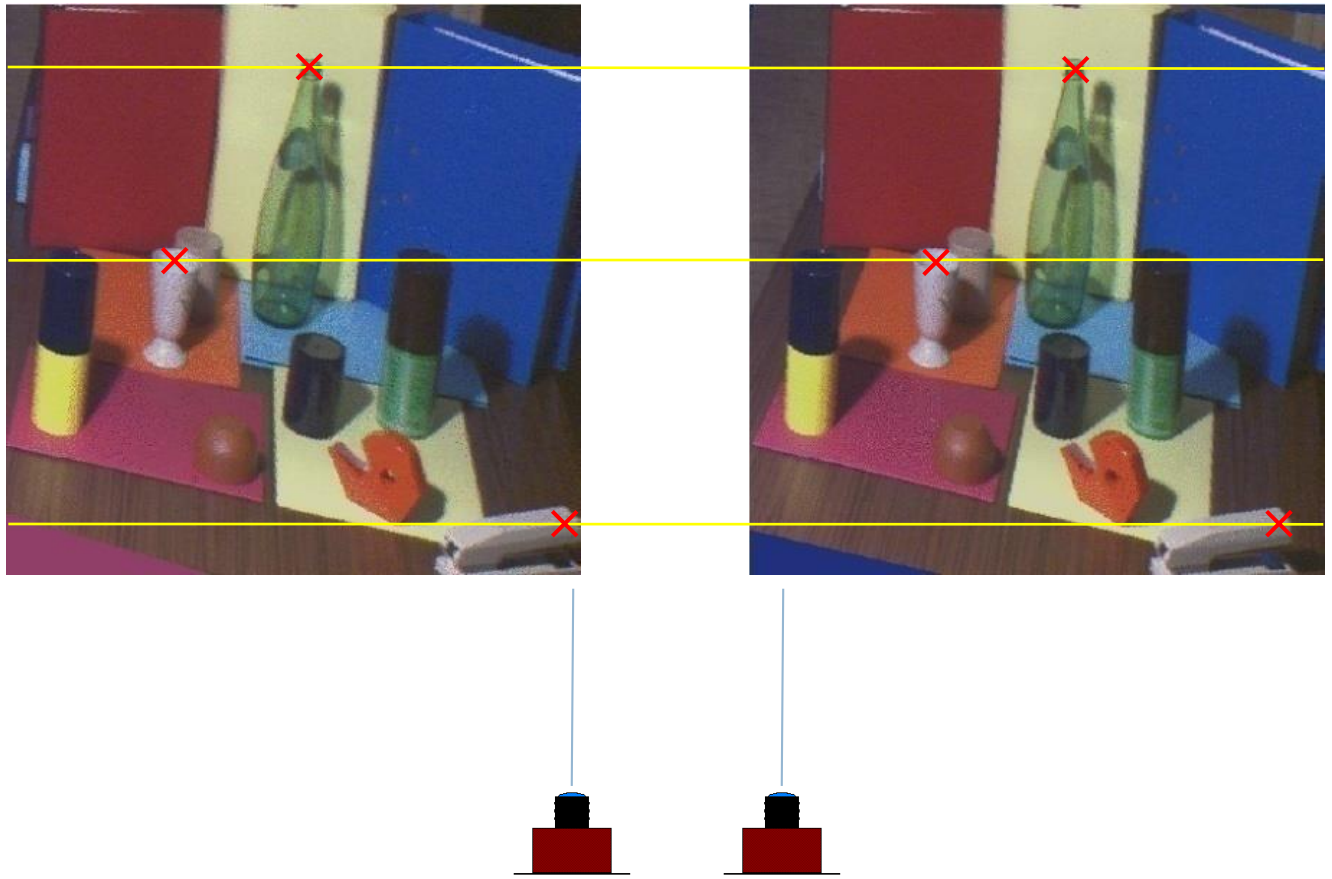
Rectification

We can always warp the images as if they were acquired through a standard geometry (horizontal and collinear conjugate epipolar lines) by computing and applying to both a transformation (i.e. homography) known as *rectification*.



Rectification

We can always warp the images as if they were acquired through a standard geometry (horizontal and collinear conjugate epipolar lines) by computing and applying to both a transformation (i.e. homography) known as *rectification*.



The Stereo Correspondence Problem

Given a point in one image (e.g. *L*) find that in the other image (*R*) which is the projection of the same 3D point. Such image points are called *corresponding points*.

KITTI (*Karlsruhe Univ. & Toyota*) Benchmark Suite

Basic cue: corresponding points look *similar* in the two images

Some properties of Perspective Projection

- **The farther objects are from the camera, the smaller they appear in the image. As a matter of example, the image of a 3D line segment of length *L* lying in a plane parallel to the image plane at distance *z* from the optical centre will exhibit a length given by:**

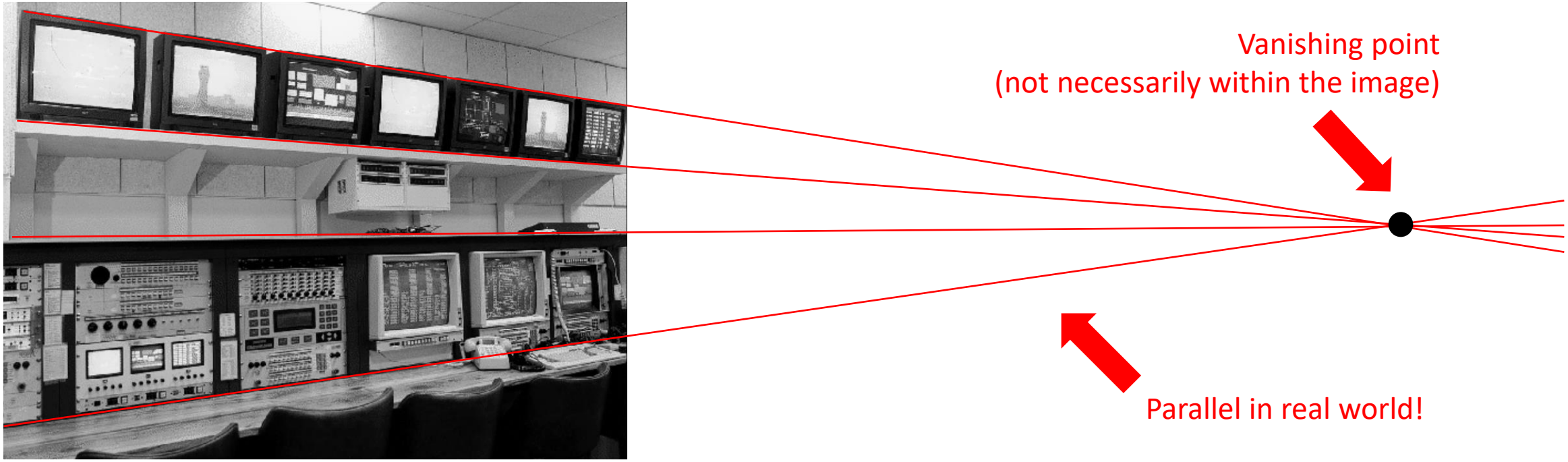
$$l = L\frac{f}{z}$$

The relationship turns out more complicated for an arbitrarily oriented 3D segment, as its position and orientation need to be accounted for as well. Nonetheless, for given position and orientation, length always shrinks alongside distance.

- **Perspective projection maps 3D lines into image lines.**
- **Ratios of lengths are not preserved (unless the scene is planar and parallel to the image plane).**
- **Parallelism between 3D lines is not preserved (except for lines parallel to the image plane)**

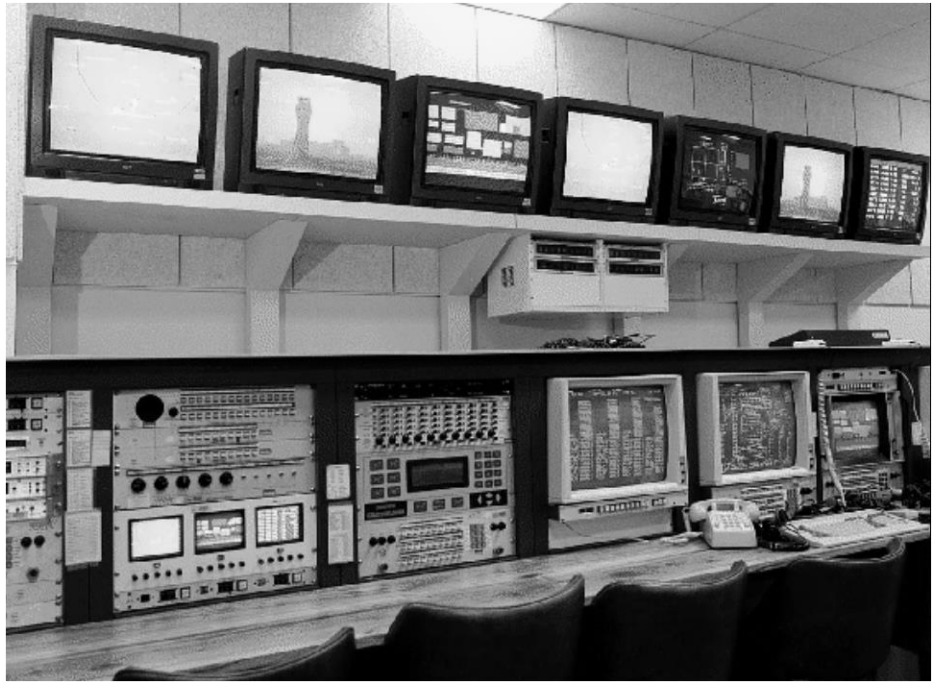
Vanishing Points (1)

The images of parallel 3D lines meet at a point, which is referred to as *vanishing point*.



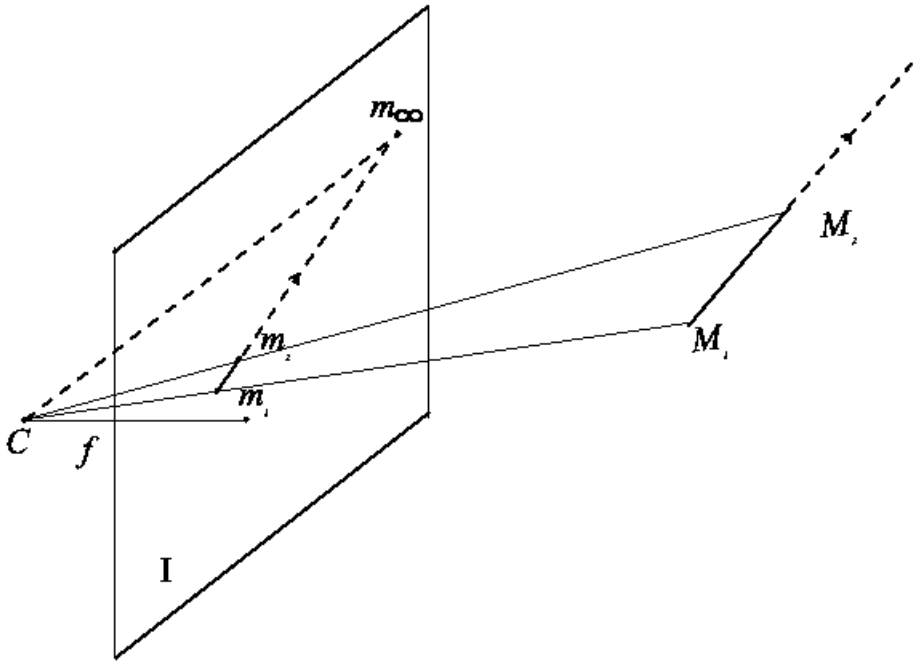
Vanishing Points (1)

The images of parallel 3D lines meet at a point, which is referred to as *vanishing point*.



Vanishing Points (2)

The vanishing point of a 3D line is the *image* of the *point at infinity* of the line (i.e. the image of the point on the line which is infinitely distant from the optical centre).

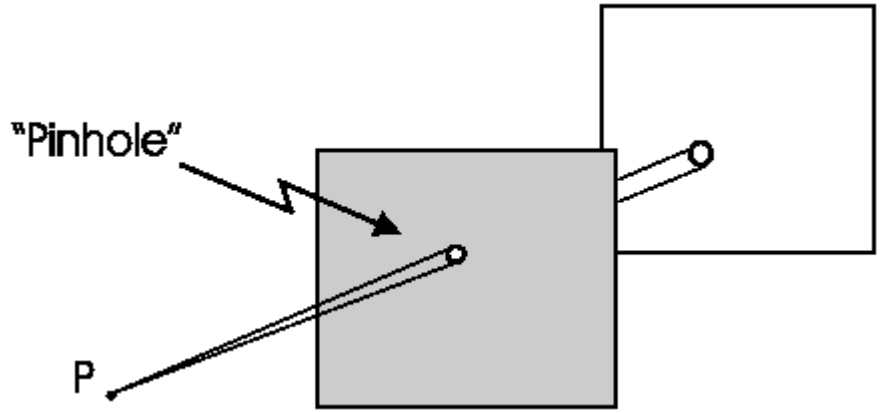


As such, it can be determined by the intersection between the image plane and the line parallel to the given one and passing through the optical centre.

Accordingly, all parallel 3D lines will share the same vanishing point, i.e. they "meet" at their vanishing point in the image, but in the "special case" of such a point being at infinity (i.e. the 3D lines are parallel to the image plane).

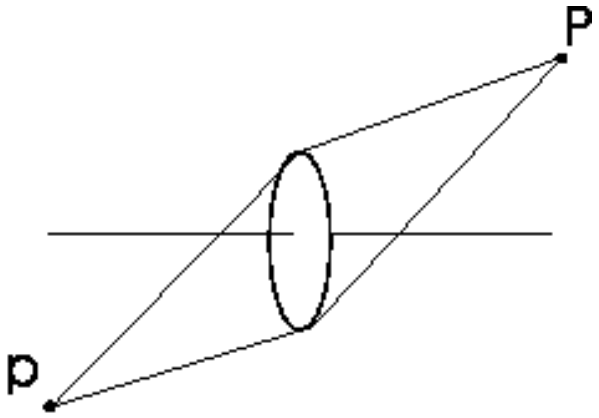
Using Lenses

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 device this happens to all scene points because of the very small size of the hole, so that the camera features an infinite Depth of Field (DOF).



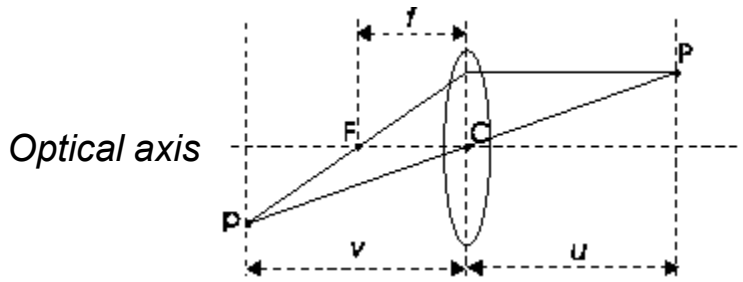
The drawback is that such a small aperture allows gathering a very limited amount of light. Thus, getting sufficiently bright images mandates very long exposure times. As a result, only static scenes can be acquired by a pinhole device to avoid motion blur.

Therefore, cameras rely on lenses to gather more light from a scene point and focus it on a single image point. This enables much smaller exposure times, as required, e.g., to avoid motion blur in dynamic scenes. However, the DOF is no longer infinite, for only points across a limited range of distances can be simultaneously on focus in a given image.



Thin Lens Equation

- **Cameras often feature complex optical systems, comprising multiple lenses. Yet, we will consider here the approximate model known as *thin lens equation*:**



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

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

- **To graphically determine the position of a focused image point we can leverage on the following two properties of thin lenses:**
  1. **Rays parallel to the optical axis are deflected to pass through *F*.**
  2. **Rays through *C* are undeflected.**

- **It is worth pointing out that, if the image is on focus, the image formation process obeys to the perspective projection model, with the centre of the lens being the optical centre and the distance *v* acting as the *effective* focal length of the projection (a different concept than the focal length of the lens !)**

Circles of Confusion (1)

- **Due to the thin lens equation, choosing the distance of the image plane determines the distance at which scene points appear on focus in the image:**

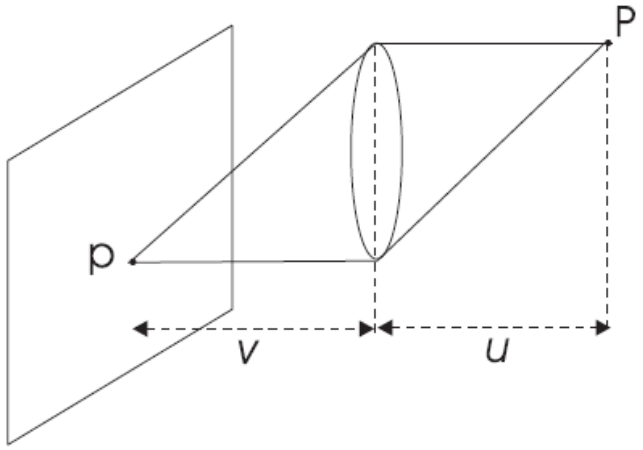
$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \rightarrow u = \frac{vf}{v - f}$$

- **Likewise, to acquire scene points at a certain distance we must set the position of the image plane accordingly:**

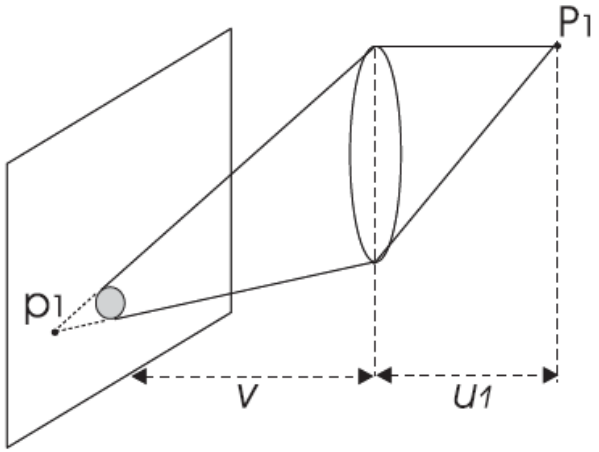
$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \rightarrow v = \frac{uf}{u - f}$$

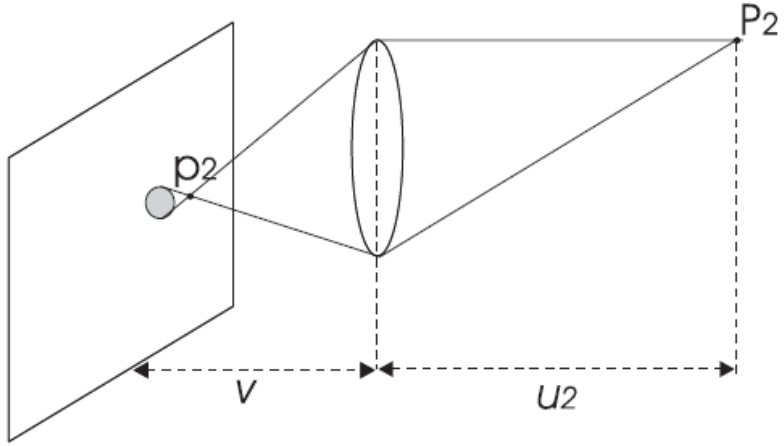
- **Given the chosen position of the image plane, scene points both in front and behind the focusing plane will result out-of-focus, thereby appearing in the image as circles, known as Circles of Confusion or Blur Circles, rather than points.**

Circles of Confusion (2)



P belongs to the focusing scene plane
P_1 lies closer to the lens than P ($u_1<u$)
P_2 is farther away to the lens than P ($u_2>u$)





Diaphragm, DOF and F-number

- In theory, when imaging a scene through a thin lens, only the points at a certain distance can be on focus, with all the others appearing blurred into circles. However, as long as such circles are smaller than the size of the photosensing elements, the image will still look on-focus. The range of distances across which the image appears on focus - due to blur circles being small enough - determines the DOF (Depth of Field) of the imaging apparatus.

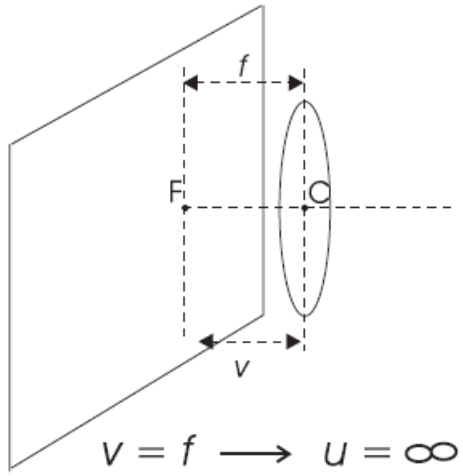
- Cameras often deploy an adjustable diaphragm (iris) to control the amount of light gathered through the *effective aperture* of the lens. The smaller the diaphragm aperture is, the larger turns out the DOF as a result of the smaller size of blur circles.

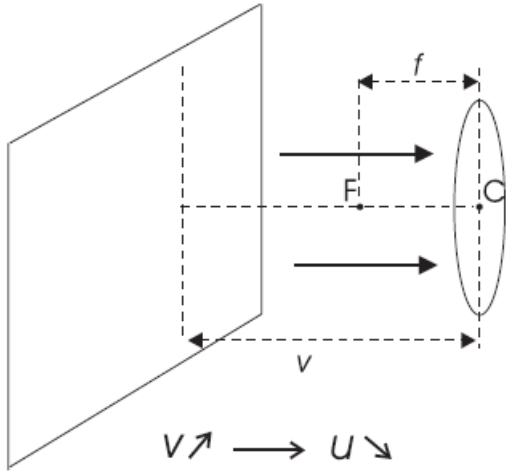
- The *F-number* is the ratio of the focal length to the effective aperture of the lens (f/d). F-number discrete units (also known as *stops*) are usually reported on the diaphragm (e.g. 1.4, 2, 2.8, 4, 5.6, 8, 11, 16…) to allow the user to adjust the effective aperture. The higher the chosen *stop*, the smaller is the diaphragm aperture and thus the larger is the actual DOF.

Focusing Mechanism

- **To focus on objects at diverse distances, another mechanism allow the lens (or lens subsystem) to translate along the optical axis with respect to the –fixed- position of the image plane (in nowadays cameras, a solid state sensor mounted on a PCB).**

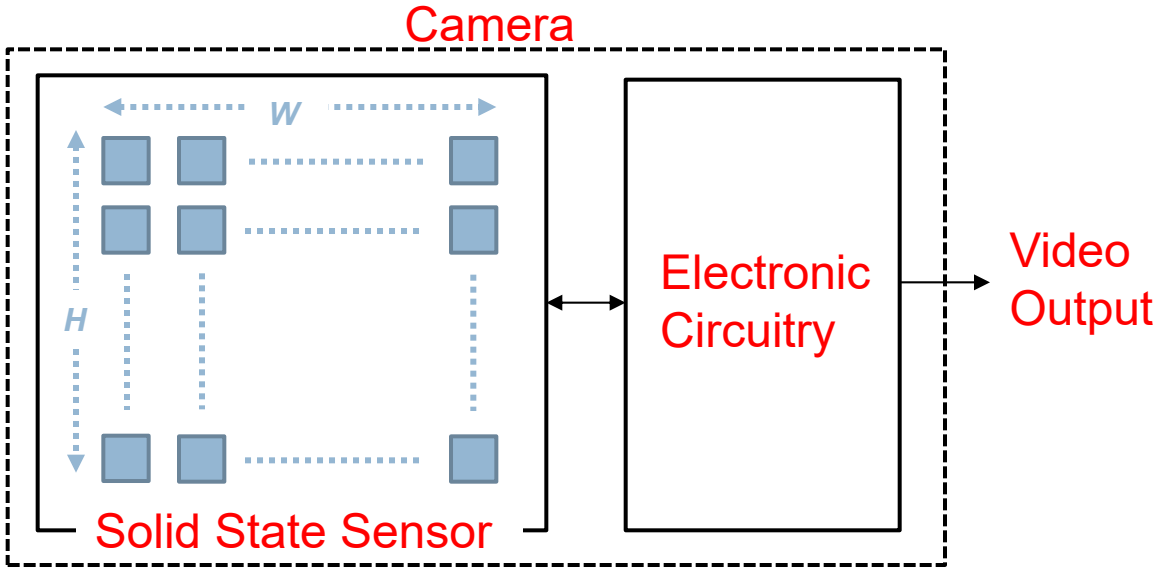
$$\frac{1}{u}+\frac{1}{v}=\frac{1}{f}$$





- **At one end position (*v=f*) the camera is focused at infinity, then the mechanism allow the lens to be translated farther away from the image plane up to a certain maximum value (the second end position), which determines the minimum focusing distance.**

Image Acquisition and Digitization



Thus, a gray-scale digital image with 256 gray-levels is stored in a computer as a 2D array of size *W*×*H* (*columns* × *rows*) where each element is a byte.

For example, storing a gray-scale image in VGA format (*640×480*) requires 300 KB, while a 4K Ultra HD format (3840×2160) requires approximately 8.3 MB.

Today, the two main image sensor technologies are CCD (Charge Coupled Devices) and CMOS (Complementary Metal Oxide Semiconductor)

The image sensor is a 2D grid of photodetectors (photogates or photodiodes), referred to as pixels, of size *W*×*H*. During exposure time, each detector converts the incident light focused by the lens into a proportional electric charge (i.e. photons into electrons). Then, the companion electronic circuitry reads-out the charge to generate an analog voltage signal which is amplified and digitized by an ADC.

In a gray-scale camera the analog voltage - proportional to the amount of light incident at a pixel- is quantized into a digital value, referred to as *intensity* or *brightness,* expressed by *m* bits. Typically, *m=8*, such that the intensity of a pixel is represented with 256 gray-levels (0=black….255=white).

The digital output produced by the camera is transmitted to a computer based on several *Digital Video Standards*, such as - in machine vision applications- Camera Link, CoaxPress, IEEE 1394 (FireWire), USB 2.0, USB3 Vision, GigE Vision.

Camera Parameters (1)

We provide here an intuitive explanation of some of the main camera parameters. Formal definitions and measurement procedures can be found in the EMVA Standard 1288 (v4, 2021).

- **Signal-to-Noise Ratio (SNR) – The intensity observed at a pixel of a digital image under perfectly stationary conditions does vary due to the presence of random noise (i.e. it is not deterministic but rather a random variable). The main noise sources are as follows.**

  - ***Photon Noise* – A photodetector converts the number of photons incident on a pixel during the exposure time (n_p) into a corresponding number of electrons (n_e) according to its quantum efficiency (η), which depends on the *fill-factor* and material properties:**

$$\eta = \frac{n_e}{n_p}$$

However, n_p is not constant but a random variable that follows a Poisson distribution:

$$P(n_p = k) = \frac{{\mu_p}^k e^{-\mu_p}}{k!}$$

where μ_p is the mean of the random variable, i.e. the expected photon count during the exposure time, which depends on the amount of incident light. Indeed, $\mu_p = \lambda t$, where λ is the expected number of photons per unit time and t is the exposure time.

Camera Parameters (2)

Due to n_p being a Poisson-distributed random variable, its variance, σ_p^2, is equal to the mean, μ_p. This means that light has an inherent SNR given by

$$SNR_p = \frac{\mu_p}{\sigma_p} = \frac{\mu_p}{\sqrt{\mu_p}} = \sqrt{\mu_p}$$

Thus, while in absolute terms photon noise grows with the amount of incident light ($\sigma_p^2 = \mu_p$), it is relatively weaker at high light levels: gathering more light yields better image quality (higher SNR_p). It is worth noticing that $n_e = \eta\, n_p$ is also a Poisson distributed random variable, with $\mu_e = \eta\, \mu_p = \sigma_e^2$ and therefore: $SNR_e = \sqrt{\eta \mu_p}$

- *Dark Current Noise* – a random amount of charge due to thermal excitement is observed at each pixel even though the sensor is not exposed to light.
- *Electronic Noise* – It is generated by the electronics which reads-out the charge and amplifies the resulting voltage signal.
- *Quantization Noise* – related to the final ADC conversion.

The other noise sources are also modelled and measured according to standard procedures, in order to estimate the overall SNR of the camera, often expressed in *decibels* ($20 \cdot \log_{10} SNR$) or *bits* ($\log_2 SNR$). In general:

$$\lambda \nearrow, t \nearrow, \eta \nearrow \implies SNR \nearrow$$

Camera Parameters (3)

- **Dynamic Range (DR) – If the sensed amount of light is too small, the "true" signal cannot be distinguished from noise: let's call μ_{p_min} the *minimum detectable amount of light*. On the other hand, the charge stored at each pixel cannot exceed a certain quantity: let's call μ_{p_sat} the amount of light that would fill up (*saturate*) the capacity of a photodetector. The DR of a sensor (often specified in *decibels* or *bits)* is defined as**

$$\mathrm{DR} = \frac{\mu_{p\_sat}}{\mu_{p\_min}}$$

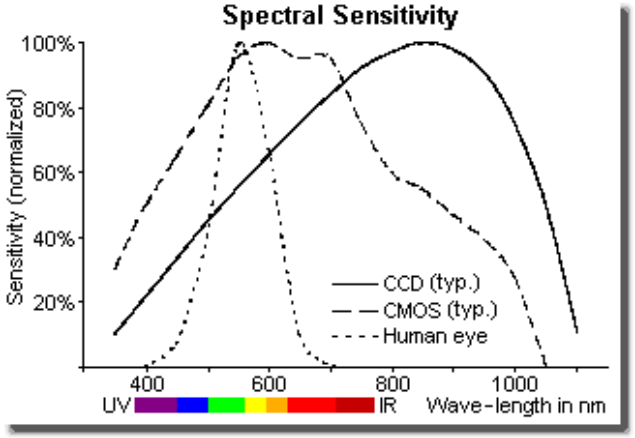
As it is the case of the SNR, also for the DR the higher is the better. Indeed, the higher the DR the better is the ability of the sensor to simultaneously capture in one image both the dark and bright structures of the scene.

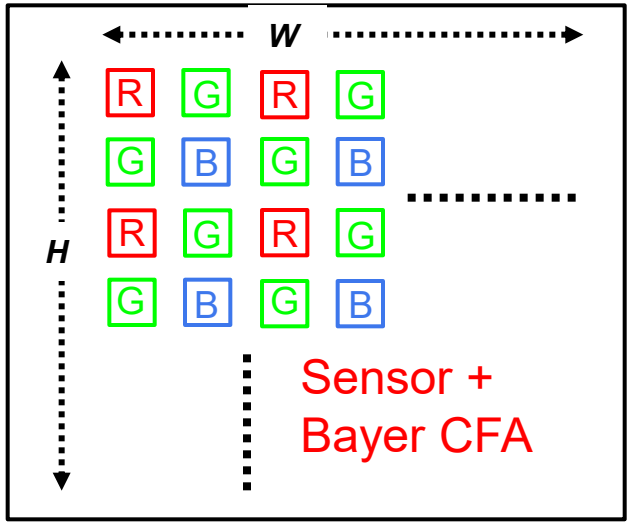
An active research field in image processing deals with creating High Dynamic Range (HDR) images by combining together a sequence of images of the same subject matter taken under different exposure times (see e.g. http://www.hdrsoft.com/index.html).

- **Non-uniformity (spatial or pattern noise) – due to manufacturing tolerances both the responsitivity to light and the amount of dark noise (dark current + electronic noise) vary across pixels.**

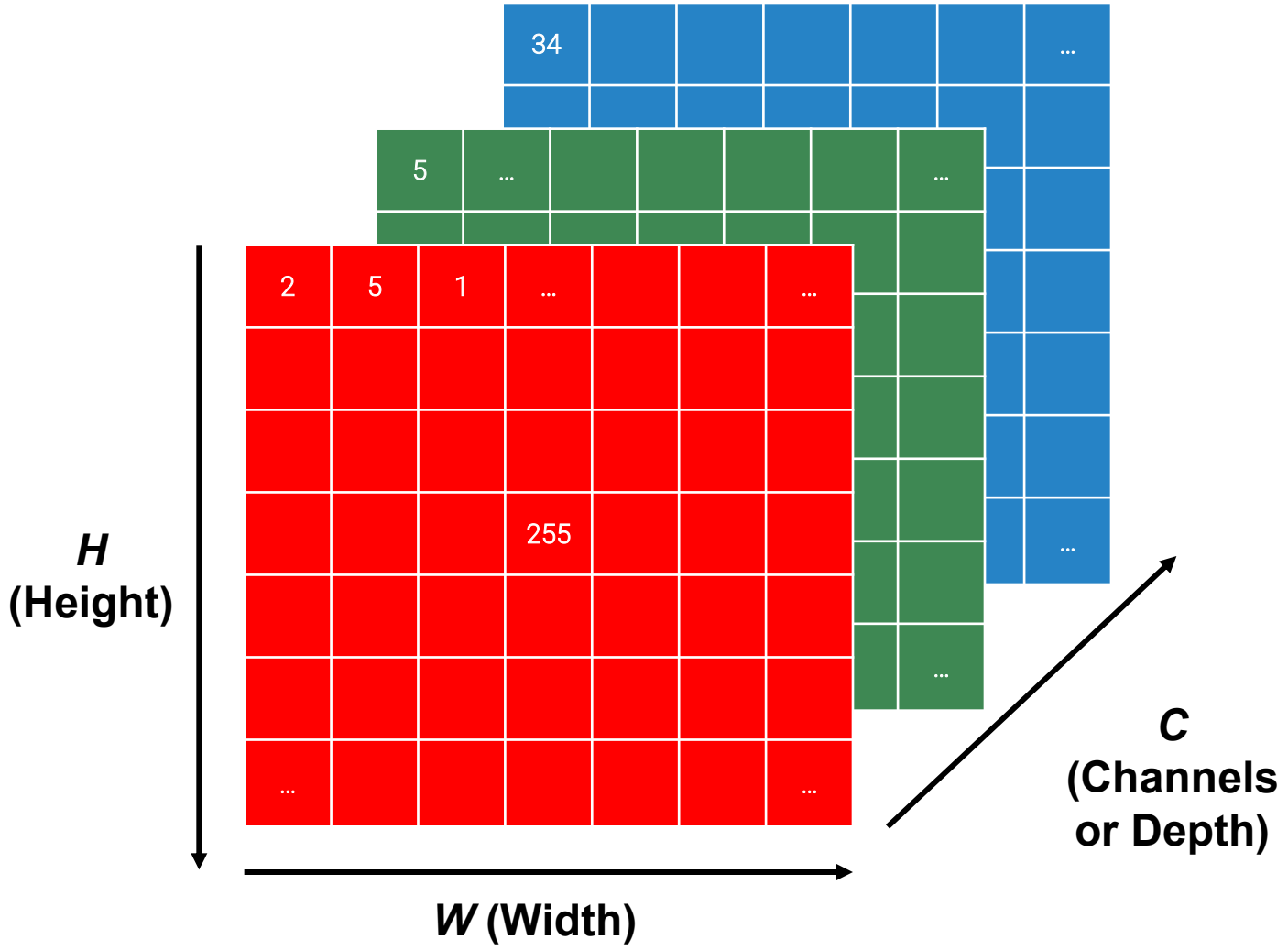
Colour Cameras

- CCD/CMOS sensors are sensitive to light ranging from near-ultraviolet (200 nm) through the visible spectrum (380-780 nm) up to the near infrared (1100 nm). The sensed intensity at a pixel results from integration over the range of wavelengths of the spectral distribution of the incoming light multiplied by the spectral response function of the sensor. As such CCD/CMOS sensor cannot sense colour.
- To create a colour sensor, an array of optical filters (Colour Filter Array) is placed in front of the photodetectors, so as to render each pixel sensitive to a specific range of wavelengths. In the most common Bayer CFA, green filters are twice as much as red and blue ones because the human eye is more sensitive to high-frequency intensity changes than to colour changes and the perceived intensity is mainly determined by the green component. To obtain an RGB triplet at each pixel, missing samples are interpolated from neighbouring pixels (*demosaicking*). However, the true resolution of the sensor is smaller due to the green channel being subsampled by a factor of 2, the blue and red ones by 4. A colour camera based on a CFA may be affected by *aliasing* as well as *artifacts* due to demosaicking. More expensive designs based on optical light splitters and 3 separate sensors (R,G,B) can avoid the above issues.
- In a colour camera each of the three *colour channels* (R,G,B) is quantized into a digital value expressed by *m* bits. Typically, *m=8*, such that the intensity of each channel is represented with 256 values (0=lowest….255=highest). Thus, a digital colour image with 256 gray-levels per channel is stored in a computer as a 3D array (aka *tensor*) of size $3 \times W \times H$ (*channels* × *columns* × *rows*) where each element is a byte.



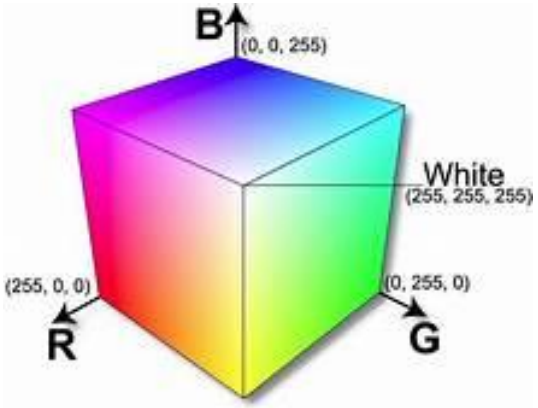


A Digital Colour Image



Colour Spaces (1)

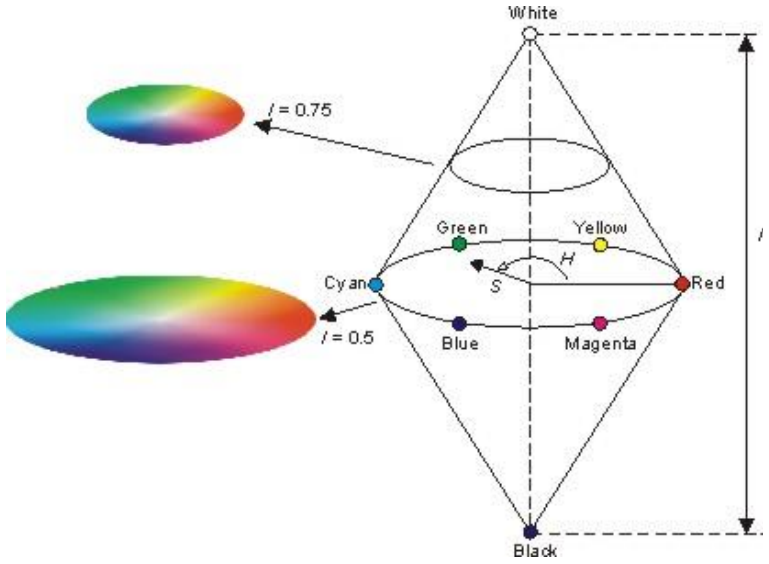
RGB (Red, Green, Blue)



This is the colour space used by cameras and display devices. A colour is expressed as a combination of the three so called *primary* colours (R,G,B).

It mimics the human visual system, which, according to the trichromatic theory, features three main kinds of photoreceptors (*cones*) responsible for detecting colour, roughly sensitive to red, green and blue light.

HSI (Hue, Saturation, Intensity)



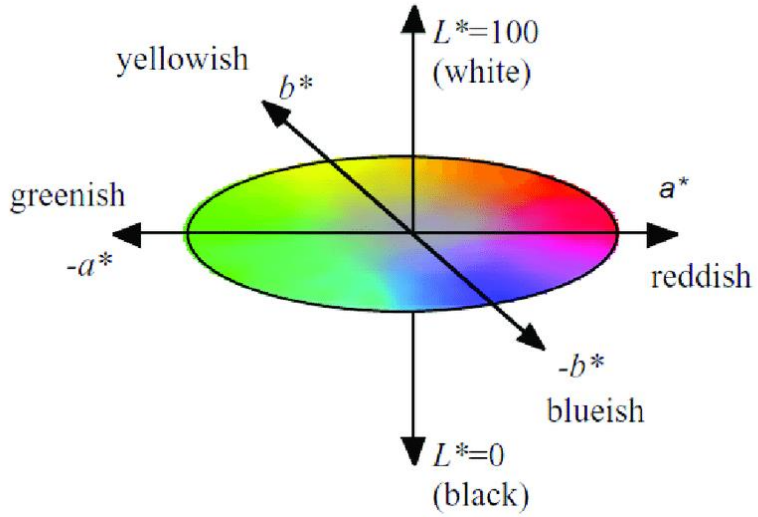
HSI decouples the intensity (I) and colour-carrying information (H,S).

Similar colour spaces are HS*V* (Value) and HS*L* (Lightness)

Hue (Angle in [0,360]): dominant colour (i.e. wavelength) perceived by an observer (e.g. red=0, green=120, blue=240).
Saturation ([0,1]): purity of the main colour, inversely proportional to the amount of white light.
Intensity ([0,1]): achromatic notion of intensity (black=0, white=1)

Colour Spaces (2)

L*a*b* (CIELAB)



L* ([0,100]): Luminance, related to the achromatic intensity, with black=0 and white=100.

a* (-128, 127): first chromatic component, i.e. green vs. red. Negative values means that the colour looks more greenish than reddish, and vice versa.

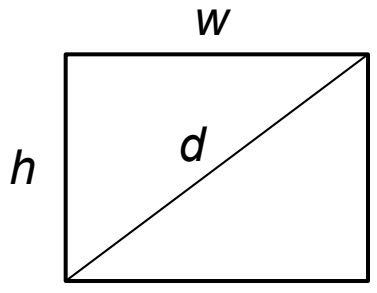
b* (-128, 127): second chromatic component, i.e. blue vs. yellow. Negative values means that the colour looks more blueish than yellowish, and vice versa.

Difference between colours in the RGB space do not match human perception of colour differences. Conversely, the L*a*b* has been designed to be more *perceptually uniform*: a uniform change in the colour space corresponds to a uniform change in the perceived colour (CIEDE2000 colour difference formula).

L*a*b* is based on *opponent process theory*, which conjectures that the human visual system perceive colour through pairs of *opponent colours* (*green vs red*, *blue vs yellow*) which inhibits each other. Thus, to perceive a colour, we rely on the two chromatic components thar are dominant in each pair.

Sensor Sizes

- **CCD/CMOS sensors come in different sizes, which are specified in inches for the sake of legacy wrt old cameras based on cathode ray tubes. In such old cameras the size (in inch) was the outer diameter of the tube, the effective image plane size being roughly 2/3 of the diameter. Nowadays, the size of the diagonal of a solid state sensor is roughly 2/3 of its size. The table below reports some typical sensor sizes, together with the corresponding (square) pixel sizes for a VGA format ($w \times h = 640 \times 480$) sensor. In case of a higher resolution format, pixel sizes shrink proportionally (e.g. by a factor of 2 for a 1280 × 960 sensor).**



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

Size (inch)	Width (mm)	Height (mm)	Diagonal (mm)	VGA Pixel Size (µm)
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

Main References

1) V. S. Nalwa, “A Guided Tour of Computer Vision”, Addison-Wesley Publishing Company, 1993.
2) Richard Szeliski “Computer Vision: Algorithms and Applications”, 2nd Edition, Springer, 2021.
3) Andreas Geiger and Philip Lenz and Raquel Urtasun, “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite”, CVPR 2012.
4) C. Steger, M. Ulrich, C. Wiedemann, “Machine Vision Algorithms and Applications”, 2nd Edition, Wiley, 2018.
5) “Standard for Characterization of Image Sensors and Cameras”, EMVA Standard 1288, European Machine Vision Association, Rel. 4.0, June 2021.

https://www.emva.org/standards-technology/emva-1288/