Image Processing and Computer Vision (Module 1)

Last update: 04 March 2024

Academic Year 2023 – 2024
Alma Mater Studiorum · University of Bologna

Contents

1 Image acquisition and formation

1.1 Pinhole camera

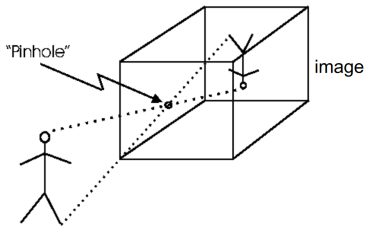
Imaging device Gathers the light reflected by 3D objects in a scene and creates a 2D representation of them. Imaging device

Computer vision Infer knowledge of the 3D scene from 2D digital images. Computer vision

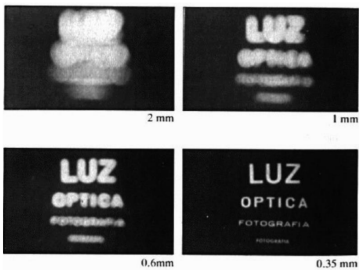
Pinhole camera Imaging device where the light passes through a small pinhole and hits the image plane. Geometrically, the image is obtained by drawing straight rays from the scene to the image plane passing through the pinhole. Pinhole camera

Remark. Larger aperture size of the pinhole results in blurry images (circle of confusion), while smaller aperture results in sharper images but requires longer exposure time (as less light passes through).

Remark. The pinhole camera is a good approximation of the geometry of the image formation mechanism of modern imaging devices.



(a) Pinhole camera model



(b) Images with varying pinhole aperture size

1.2 Perspective projection

Geometric model of a pinhole camera. Perspective projection

Scene point M (the object in the real world).

Image point m (the object in the image).

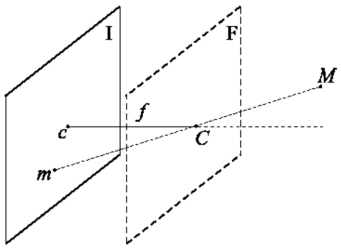
Image plane I.

Optical center C (the pinhole).

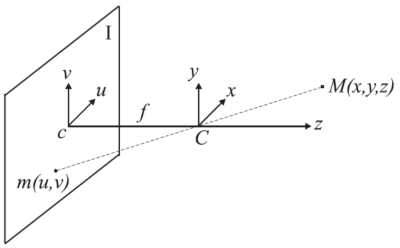
Image center/piercing point c (intersection between the optical axis – the line orthogonal to I passing through C – and I).

Focal length f.

Focal plane F.



- u and v are the horizontal and vertical axis of the image plane, respectively.
- x and y are the horizontal and vertical axis of the 3D reference system, respectively, and form the **camera reference system**.



Camera reference system

Remark. For the perspective model, the coordinate systems (U, V) and (X, Y) must be parallel.

Scene–image mapping The equations to map scene points into image points are the following:

Scene–image mapping

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

Proof. This is the consequence of the triangle similarity theorems.

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

The minus is needed as the axes are inverted

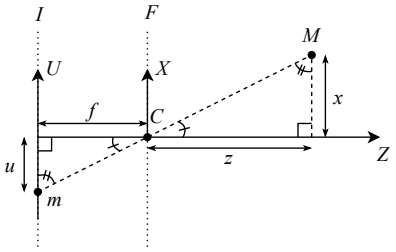


Figure 1.2: Visualization of the horizontal axis. The same holds on the vertical axis.

By inverting the axis horizontally and vertically (i.e. inverting the sign), the image plane can be adjusted to have the same orientation of the scene:

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

□

Remark. The image coordinates are a scaled version of the scene coordinates. The scaling is inversely proportioned with respect to the depth.

- The farther the point, the smaller the coordinates.
- The larger the focal length, the bigger the object is in the image.

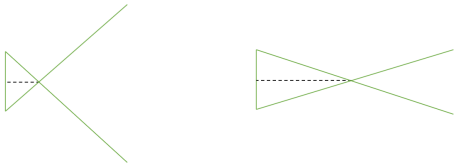


Figure 1.3: Coordinate space by varying focal length

Remark. The perspective projection mapping is not a bijection:

- A scene point is mapped into a unique image point.
- An image point is mapped onto a 3D line.

Therefore, reconstructing the 3D structure of a single image is an ill-posed problem (i.e. it has multiple solutions).

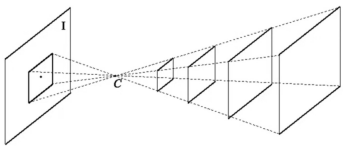


Figure 1.4: Projection from scene and image points

1.2.1 Stereo geometry

Stereo vision

Stereo vision Use multiple images to triangulate the 3D position of an object.

Stereo correspondence

Stereo correspondence Given a point L in an image, find the corresponding point R in another image.

Without any assumptions, an oracle is needed to determine the correspondences.

Standard stereo geometry

Standard stereo geometry Given two reference images, the following assumptions must hold:

- The X, Y, Z axes are parallel.
- The cameras that took the two images have the same focal length f (coplanar image planes) and the images have been taken at the same time.
- There is a horizontal translation b between the two cameras (baseline).
- The disparity d is the difference of the U coordinates of the object in the left and right image.

Fundamental relationship in stereo vision

Theorem 1.2.1 (Fundamental relationship in stereo vision)**.** If the assumptions above hold, the following equation holds:

$$z = b\frac{f}{d}$$

Proof. Let $P_L = \begin{pmatrix} x_L & y & z \end{pmatrix}$ and $P_R = \begin{pmatrix} x_R & y & z \end{pmatrix}$ be the coordinates of the object P with respect to the left and right camera reference system, respectively. Let $p_L = \begin{pmatrix} u_L & v \end{pmatrix}$ and $p_R = \begin{pmatrix} u_R & v \end{pmatrix}$ be the coordinates of the object P in the left and right image plane, respectively.

By assumption, we have that $P_L - P_R = \begin{pmatrix} b & 0 & 0 \end{pmatrix}$, where b is the baseline.

By the perspective projection equation, we have that:

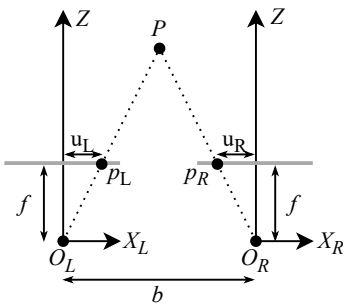
$$u_L = x_L\frac{f}{z} \qquad u_R = x_R\frac{f}{z}$$

Disparity is computed as follows:

$$d = u_L - u_R = x_L\frac{f}{z} - x_R\frac{f}{z} = b\frac{f}{z}$$

We can therefore obtain the Z coordinate of P as:

$$z = b\frac{f}{d}$$



Note: the Y/V axes are not in figure.

□

Remark. Disparity and depth are inversely proportional: the disparity of two points decreases if the points are farther in depth.

Stereo matching If the assumptions for standard stereo geometry hold, to find the object corresponding to p_L in another image, it is sufficient to search along the horizontal axis of p_L looking for the same colors or patterns. Stereo matching

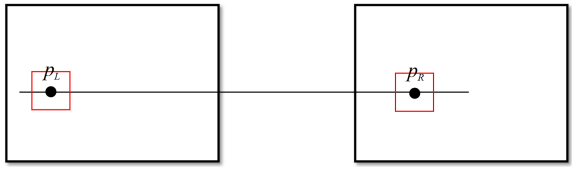


Figure 1.5: Example of stereo matching

Epipolar geometry Approach applied when the two cameras are no longer aligned according to the standard stereo geometry assumption. Still, the focal lengths and the roto-translation between the two cameras must be known. Epipolar geometry

Given two images, we can project the epipolar line related to the point p_L in the left plane onto the right plane to reduce the problem of correspondence search to a single dimension.

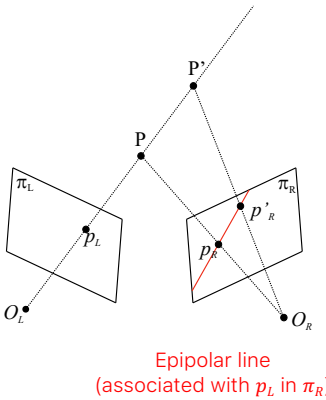
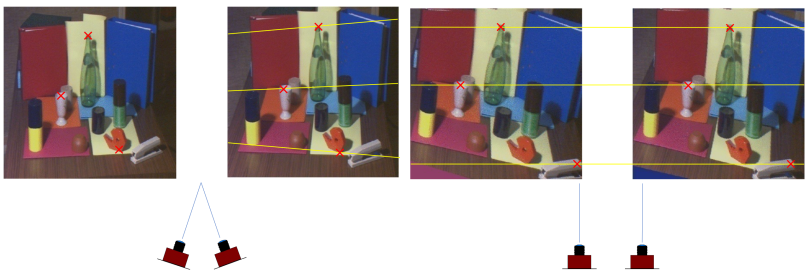


Figure 1.6: Example of epipolar geometry

Remark. It is nearly impossible to project horizontal epipolar lines and searching through oblique lines is awkward and computationally less efficient than straight lines.

Rectification Transformation applied to convert epipolar geometry to a standard stereo geometry. Rectification



(a) Images before rectification (b) Images after rectification

1.2.2 Ratios and parallelism

Given a 3D line of length L lying in a plane parallel to the image plane at distance z, then its length l in the image plane is:

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

In all the other cases (i.e. when the line is not parallel to the image plane), the ratios of lengths and the parallelism of lines are not preserved.

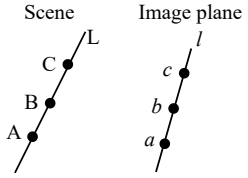


Figure 1.8: Example of not preserved ratios. It holds that $\frac{\overline{AB}}{\overline{BC}} \neq \frac{\overline{ab}}{\overline{bc}}$.

Vanishing point Intersection point of lines that are parallel in the scene but not in the image plane. Vanishing point

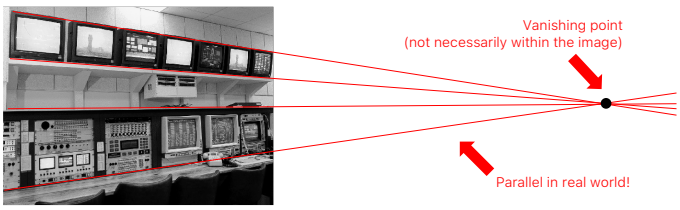


Figure 1.9: Example of vanishing point

1.3 Lens

Depth of field (DOF) Distance at which a scene point is in focus (i.e. when all its light rays gathered by the imaging device hit the image plane at the same point). Depth of field (DOF)

Remark. Because of the small size of the aperture, a pinhole camera has infinite depth of field but requires a long exposure time making it only suitable for static scenes.

Lens A lens gathers more light from the scene point and focuses it on a single image point. Lens

This allows for a smaller exposure time but limits the depth of field (i.e. only a limited range of distances in the image can be in focus at the same time).

Thin lens Approximate model for lenses. Thin lens

Scene point P (the object in the real world).

Image point p (the object in the image).

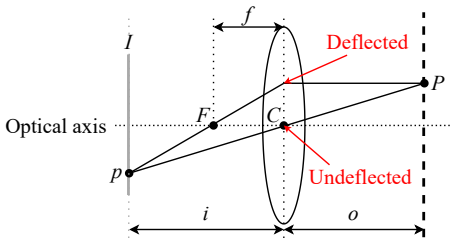
Object–lens distance o.

Image–lens distance i (i.e. focal length of the camera).

Center of the lens C.

Focal length of the lens f.

Focal plane/focus of the lens F.



A thin lens has the following properties:

- Rays hitting the lens parallel to the optical axis are deflected to pass through the focal plane of the lens F.
- Rays passing through the center of the lens C are undeflected.
- The following equation holds: Thin lens equation

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f}$$

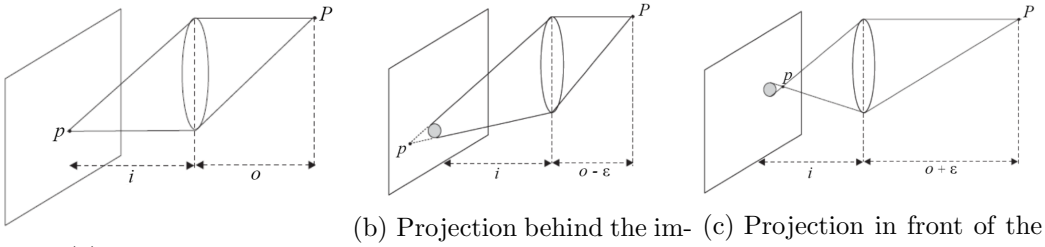
Image formation When the image is in focus, the image formation process follows the normal rules of the perspective projection model where:

- C is the optical center.
- i is the focal length of the camera.

By fixing the focal length of the lens (f), we can determine the distance of the scene point (o) or the image point (i) required to have the object in focus.

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f} \iff o = \frac{if}{i - f} \qquad \frac{1}{o} + \frac{1}{i} = \frac{1}{f} \iff i = \frac{of}{o - f}$$

Remark. Points projected in front or behind the image plane will create a circle of confusion (blur).



(a) Image in focus

(b) Projection behind the image plane

(c) Projection in front of the image plane

Adjustable diaphragm Device to control the light gathered by the effective aperture of the lens. Adjustable diaphragm

Reducing the aperture will result in less light and an increased depth of field.

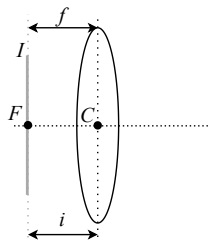
Remark. On a theoretical level, images that are not in focus appear blurred (circles of confusion). Despite that, if the circle is smaller than the photo-sensing elements (i.e. pixels), it will appear in focus.

Focusing mechanism Allows the lens to translate along the optical axis to increase its distance to the image plane. Focusing mechanism

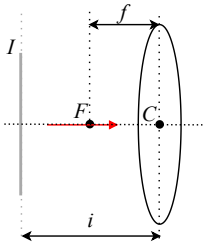
At the minimum extension (Figure 1.11a), we have that:

$$i = f \text{ and } o = \infty \text{ as the thin lens equation states that } \frac{1}{o} + \frac{1}{i} = \frac{1}{f}$$

By increasing the extension (i.e. increase i), we have that the distance to the scene point o decreases. The maximum extension determines the minimum focusing distance.



(a) Minimum extension



(b) Maximum extension

Figure 1.11: Extension of a focusing mechanism

1.4 Image digitalization

1.4.1 Sampling and quantization

The image plane of a camera converts the received irradiance into electrical signals.

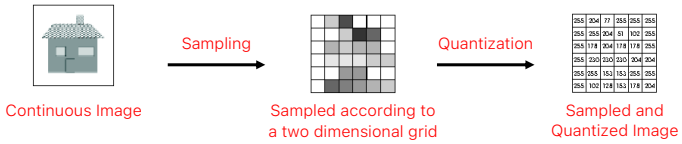


Figure 1.12: Image digitalization steps

Sampling The continuous electrical signal is sampled to produce a $N \times M$ matrix of pixels: Sampling

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

Quantization Let m be the number of bits used to encode a pixel. The value of each pixel is quantized into 2^m discrete gray levels. Quantization

Remark. A grayscale image usually uses 8 bits

An RGB image usually uses $3 \cdot 8$ bits.

Remark. The more bits are used for the representation, the higher the quality of the image will be.

- Sampling with fewer bits will result in a lower resolution (aliasing).
- Quantization with fewer bits will result in less representable colors.

Figure 1.13: Sampling and quantization using fewer bits

1.4.2 Camera sensors

Photodetector Sensor that, during the exposure time, converts the light into a proportional electrical charge that will be processed by a circuit and converted into a digital or analog signal. Photodetector

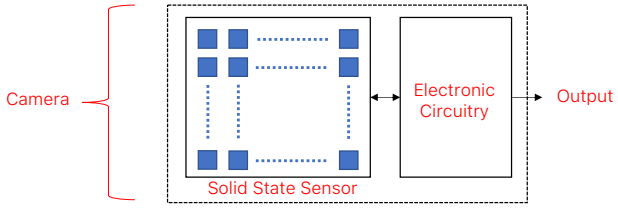


Figure 1.14: Components of a camera

The two main sensor technologies are:

Charge Coupled Device (CCD) Typically produces higher quality images but are more expensive. Charge Coupled Device (CCD)

Complementary Metal Oxide Semiconductor (CMOS) Generally produces lower quality images but is more compact and less expensive. Each sensor has integrated its own circuitry that allows to read an arbitrary window of the sensors. Complementary Metal Oxide Semiconductor (CMOS)

Color sensors CCD and CMOS sensors are sensitive to a wide spectrum of light frequencies (both visible and invisible) but are unable to sense colors as they produce a single value per pixel. Color sensors

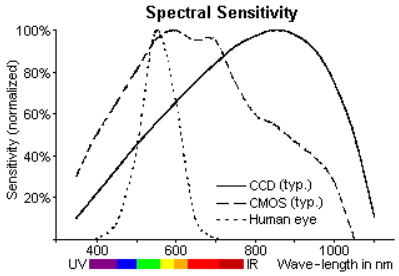


Figure 1.15: CCD and CMOS spectral sensitivity

Color Filter Array (CFA) Filter placed in front of a photodetector to allow it to detect colors. Color Filter Array (CFA)

Possible approaches are:

Bayer CFA A grid of green, blue, and red filters with the greens being twice as much as the others (the human eye is more sensible to the green range). To determine the RGB value of each pixel, missing color channels are sampled from neighboring pixels (demosaicking).

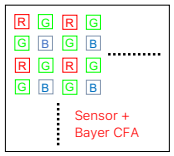


Figure 1.16: Example of Bayer filter

Optical prism A prism splits the incoming light into 3 RGB beams, each directed to a different sensor. It is more expensive than Bayer CFA.

1.4.3 Metrics

Signal to Noise Ratio (SNR) Quantifies the strength of the actual signal with respect to unwanted noise.

Signal to Noise Ratio (SNR)

Sources of noise are:

Photon shot noise Number of photons captured during exposure time.

Elecronic circuitry noise Generated by the electronics that read the sensors.

Quantization noise Caused by the digitalization of the image (ADC conversion).

Dark current noise Random charge caused by thermal excitement.

SNR is usually expressed in decibels or bits:

$$\texttt{SNR}_{\text{db}} = 20 \cdot \log_{10}(\texttt{SNR}) \qquad \texttt{SNR}_{\text{bit}} = \log_2(\texttt{SNR})$$

Dynamic Range (DR) Measures the ability of a sensor to capture both the dark and bright structure of the scene.

Dynamic Range (DR)

Let:

- E_{min} be the minimum detectable irradiation. This value depends on the noise.
- E_{max} be the saturation irradiation (i.e. the maximum amount of light that fills the capacity of the photodetector).

DR is defined as:

$$\texttt{DR} = \frac{E_{\text{max}}}{E_{\text{mix}}}$$

As with SNR, DR can be expressed in decibels or bits.

2 Spatial filtering

2.1 Noise

The noise added to a pixel p is defined by $n_k(p)$, where k indicates the time step (i.e. noise is different at each time step). It is assumed that $n_k(p)$ is i.i.d and $n_k(p) \sim \mathcal{N}(0, \sigma)$. The information of a pixel p is therefore defined as:

$$I_k(p) = \tilde{I}(p) + n_k(p)$$

where $\tilde{I}(p)$ is the real information.

Temporal mean denoising

Temporal mean denoising Averaging N images taken at different time steps.

$$\begin{aligned} O(p) &= \frac{1}{N} \sum_{k=1}^{N} I_k(p) \\ &= \frac{1}{N} \sum_{k=1}^{N} \Big(\tilde{I}(p) + n_k(p) \Big) \\ &= \frac{1}{N} \sum_{k=1}^{N} \tilde{I}(p) + \overbrace{\frac{1}{N} \sum_{k=1}^{N} n_k(p)}^{\mu = 0} \\ &\approx \tilde{I}(p) \end{aligned}$$

Remark. As multiple images of the same object are required, this method is only suited for static images.

Spatial mean denoising

Spatial mean denoising Given one image, average across neighboring pixels.

Let K_p be the pixels in a window around p (included):

$$\begin{aligned} O(p) &= \frac{1}{|K_p|} \sum_{q \in K_p} I(p) \\ &= \frac{1}{|K_p|} \sum_{q \in K_p} \Big(\tilde{I}(q) + n(q) \Big) \\ &= \frac{1}{|K_p|} \sum_{q \in K_p} \tilde{I}(q) + \frac{1}{|K_p|} \sum_{q \in K_p} n(q) \\ &\approx \frac{1}{|K_p|} \sum_{q \in K_p} \tilde{I}(q) \end{aligned}$$

Remark. As the average of neighboring pixels is considered, this method is only suited for uniform regions.