Computer Vision

Matteo Galiazzo

February 21, 2025

Contents

1 Introduction

The difference between computer vision and image processing is the fact that computer vision is the process of extracting information from images, while image processing aims at improving the quality of images. Quite often image processing helps computer vision. The informations we want to extract from the images could be counting, object orientation, object classification, measurements... Computer vision is challenging since we lose depth from the images (3D information becomes 2D), scale and illumination varies, and there's object occlusion (object hiding other objects).

Computer vision started with hand-crafted decision rules that only required few images as example, and evolved to machine learning where the algorithm learns a decision rule, but the training requires hundreds to thousands of images. The big paradigm shift happened with deep learning (e2e learning) which learned both the image representation and the decision rules, but required thousands to millions of words. Deep learning has been enabled by better networks, better hardware and more data.

2 Fundamentals of Image Processing and Computer Vision

2.1 Images

An imaging device gathers the light reflected by 3D objects to create a 2D representation of the scene.

2.2 Pinhole camera model

The "pinhole camera" is the simplest camera model we can define. Light goes through the very small pinhole (to not have saturation) and hits the image plane. Geometrically, the image is achieved by drawing straight rays from scene points through the hold up to the image plane.

This simple geometrical model turns out to be a good approximation of the geometry of image formation. However, useful images can hardly be captured by means of a pinhole camera.

The geometric model of image formation in a pinhole camera is known as **perspective projection**.

Which, by writing the poitns as vectors and the plane's coordinates becomes:

Given the reference frame in 2

- u is the horizontal axis in the image plane.

M : scene point

m : corresponding image point

I : image plane

C : optical centre (pin hole)

Optical axis: line through *C* and orthogonal to *I*

c : intersection between optical axis and image
plane (image centre or piercing point)
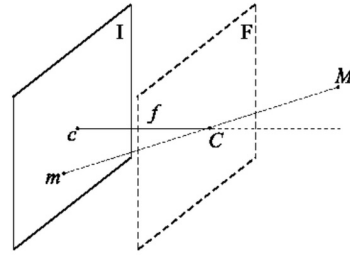
f : focal length
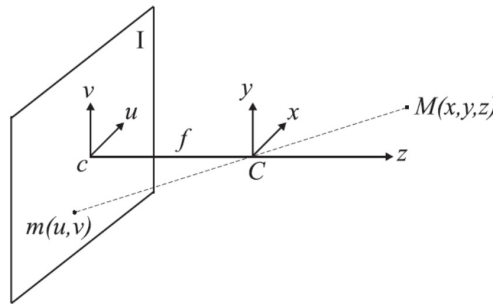
F : focal plane

Figure 1: perspective projection

Figure 2: geometric model of image formation

- v is the vertical axis in the image plane.
- X and Y are the respective axis in the 3D reference system. It's called the **camera reference system** because it is "attached" to the camera.

For the perspective model these axis must be parallel

The equations to map scene points into their corresponding image points are defined as:

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

The minus sign means the axis gets inverted (as we can see in the visualization, and it's what happens in the brain). We can get rid of the sign, since the image plane can be though of as lying in front rather than behind the optical centre.

Image coordinates are a scaled version of scene coordinates (function of depth), this means that if the distance becomes bigger the object gets smaller. When z increases, since it's at the denominator in both the equations the terms gets smaller (object gets smaller). **The larger the focal length the bigger the object is in the image (and viceversa)**.

As we previously said, the image formation process deals with mapping a 3D space onto a 2D space, and so to a loss of information. A given scene point is mapped into an image point, but an image point is mapped onto a 3D line. For an image point we can only state that its corresponding scene point lays on al ine but cannot disambiguate a specific 3D point along such a line.

2.3 Stereo images

The solution is to use multiple images (at least two) to create stereo vision. The human visual system is a stereo vision system. Given correspondences, 3D information can be recovered easily by triangulation. We can use two cameras, or two cameras and an infrared sensor to project guides to align the images.

For standard stereo geometry there are some assumptions we have to make:

- The cameras have parallel (x, y, z) axes.
- The image planes of both cameras are coplanar and aligned.
- Both cameras have identical focal lengths.

Based on this, the transformation between the two reference frames is just a translation, usually horizontal. For stereo vision is also really important to **sense two images at the very same moment**.

The cameras are displaced at a given quantity b called baseline. The **disparity** is the difference between the horizontal coordinates in the left and right images.
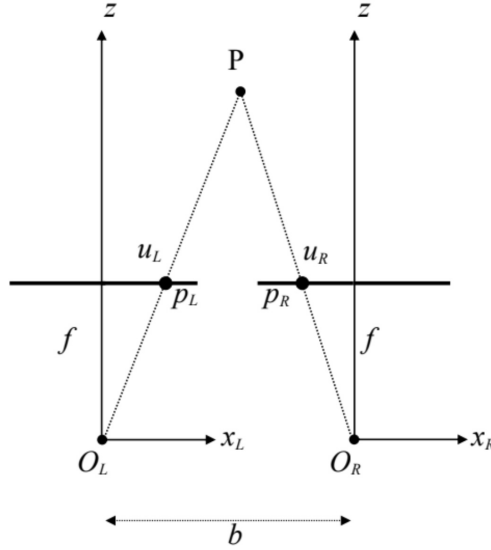
Figure 3: standard stereo geometry

The fundamental relationship in stereo vision is $z = b \cdot \frac{f}{d}$. It's used to calculate the depth of a point in a scene from a pair of stereo images.

In standard stereo geometry since we are given just two 2D images there is no info about the correspondence between two points in the two images. We can recall that the camera have parallel axes, and so we know that we can search for the correspondence alogn the horizontal lines. This task is called stereo matching.

2.4 Stereo correspondence

Given a point in one image find it in the other image which is the projection of the same 3D point. Such image points are called corresponding points. **Points farther away have a smaller disparity, while close poitns have a larger display**

Figure 4: Corresponding points look similar in the two images

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:

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

This relationship is more complicated for an arbitrarily oriented 3D segment, as its position and orientation need to be accounted for as well. For a **given position and orientation, length always shrinks alongside distance**.

Perspective projection maps 3D lines into image lines. **Parallelism between 3D lines is not preserved** (except for lines parallel to the image plane). This is the reason why if we look at a really long road into the distance we have the perception that the road becomes thinner, and the lines of the road intersect in the distance. The images of parallel 3D lines intersect at a point, called **vanishing point**, which isn't necessarily within the image.

If the lines are parallel to the image plane they meet at infinity.

2.5 Epipolar geometry

What if the two cameras are no longer aligned? Do we need to search through the whole image? We can project the line related to point P_L in the right plane and search across that line. The issue is that this projection can be computed only if the transformation between the two cameras is known (the relative mapping between the two cameras).
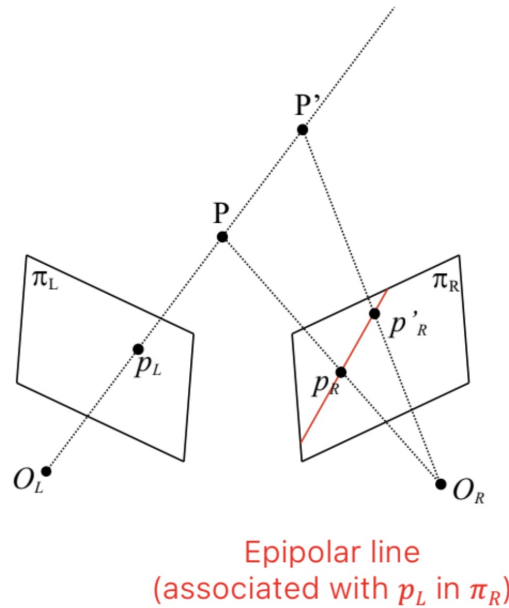

Figure 5: Epipolar line

It is almost impossible to build a stereo rig which is perfectly aligned horizontally. Searching through oblique epipolar lines is awkward, and computationally is less efficient. What people do in practice is to convert epipolar geometry to standard geometry with rectification/warping. We warp the images as if they were acquired through a standard geometry, then we can compute and applyto both images a transformation known as rectification.

2.6 Depth of Field (DOF)

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. The image is really sharp, but has a really low light. If a point is projected onto a circle instead of a point (bigger pinhole) the image is not sharp (not on focus). If we cannot gather through aperture we have to integrate through time, by using a longer exposure time.

2.7 Lenses

Lenses concentrate light, so we use them to gather more light from a scene point and focus it on a single image point. This enables much smaller exposure times. This way Depth Of Field is no longer infinite, and only points across a limited range of distances can be simultaneously on focus in a given image.

We will consider the approximate model known as thin lens equation, which is useful to graphically determine the position of a focused image point:
- Rays parallel to the optical axis are deflected to pass through F.

- Rays through C are undeflected.

If the image is on focus, the image formation process obeys to the perspective projection model:

- The center of the lens is the optical center.
- The distance v acts as the effective focal length of the projection.

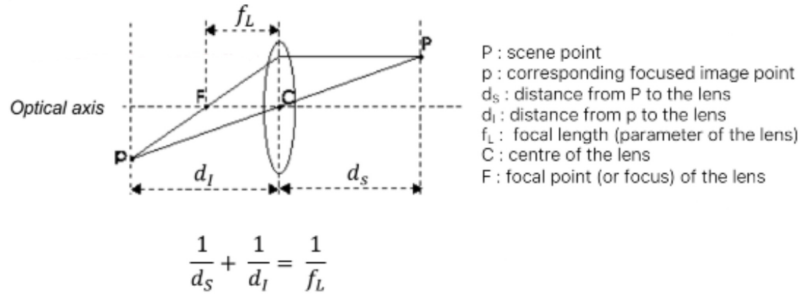

$$\frac{1}{d_S} + \frac{1}{d_I} = \frac{1}{f_L}$$

Figure 6: Scheme of a lens

Choosing the distance of the image plane determines the distance at which scene points appear on focus in the image. Scene points 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.

The advantage of lenses is to have a small exposure time for capturing moving objects but we pay in terms of Depth Of Field.
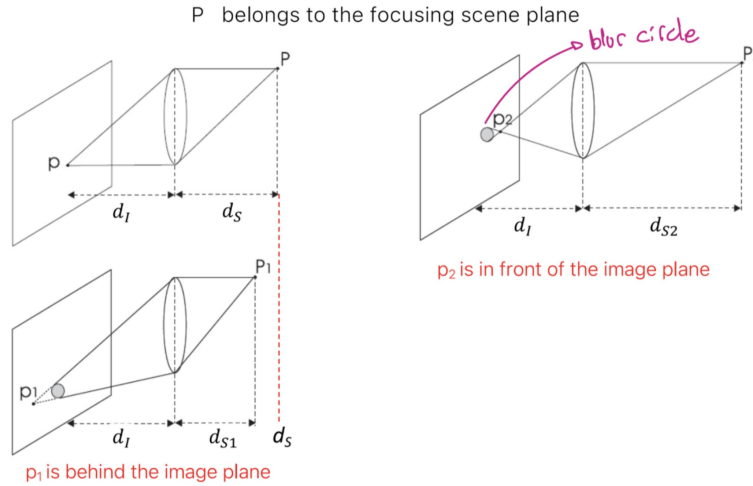

Figure 7: Lens focusing at various distances

Diaphragm

3 Advanced Topics in Deep Learning for Computer Vision