Lecture Notes: Image Formation

Subhransu Maji

February 19, 2025

Contents

1 Overview

This lecture will introduce the fundamentals of image formation. Key topics include the pinhole projection model, lens-based cameras, depth of focus, field of view, and lens aberrations. Understanding how images are formed, how lenses work, and their limitations is essential for image processing and computer vision. We will then explore how images are captured and represented. Topics include early color photography and the three-color technique, analog image capture using color film, digital sensors and the process of creating a color image via demosaicing.

2 Pinhole Camera Model

Light reflects off a surface in all directions, causing repeated scattering. As a result, simply placing a film in front of an object does not produce a clear image. A pinhole camera controls this by allowing only a single path for light from each point in the scene to reach the film. This prevents blurring caused by overlapping light rays, resulting in a sharp image. However, this comes at the cost of reduced light efficiency, as most of the incoming light is blocked—more on this later.

The geometry of a pinhole camera causes light passing through a small aperture to project an inverted image onto a surface. The position of the projected point on the screen depends on the camera's parameters, such as the focal length (f), and the location of the point in the 3D scene. Figure 1 illustrates this concept: a point $P = (x, y, z)$ in the scene is projected to a point $P' = (x', y', z')$ whose

coordinates are given by:

$$x' = -f\left(\frac{x}{z}\right) \tag{1}$$

$$y' = -f\left(\frac{y}{z}\right) \tag{2}$$

$$z' = -f \tag{3}$$

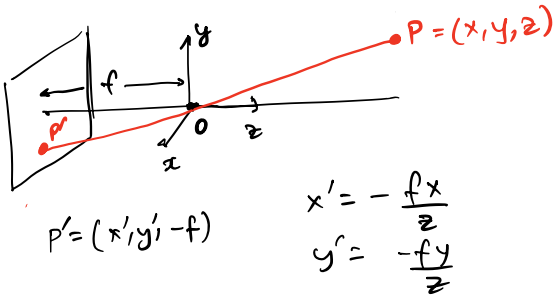


Figure 1: The pinhole camera model and perspective projection equations.

2.1 Qualitative Properties

The basic perspective projection formulas lead to several important phenomena:

1. The projection process transforms a 3D scene into a 2D image. It preserves straight lines and incidences but distorts angles and lengths.

2. The apparent size of an object is inversely proportional to its distance, as seen in the $\frac{1}{z}$ dependence in the projection equations.

3. *Vanishing points* – Lines extending to infinity appear to converge at a single point in the image, known as the vanishing point. All lines parallel in a given direction share the same vanishing point. A useful exercise is to prove this mathematically.

4. *Vanishing lines* – Planes extending infinitely appear to converge to a line. Another way to express this is that all vanishing points of lines lying in a plane form a vanishing line. A special case is the horizon line, which represents the vanishing line of the ground plane.

5. The projection and its associated distortions serve as useful cues for estimating the relative heights and distances of objects.

6. *Orthographic projection* (or parallel projection) is a special case of perspective projection where the object is infinitely far from the image plane. This type of projection is commonly used in games and design visualizations.

3 Lens and Optics

Shrinking the aperture to a very small point reduces the brightness of the image because the amount of light reaching the film is proportional to the aperture's area. One way to compensate for this is by increasing the exposure time. However, this introduces motion blur if either the scene or the camera moves during exposure. This presents a dilemma: reducing the aperture improves sharpness but makes the camera less efficient, while increasing the aperture allows more light but results in a blurry image. The solution to this problem lies in using a lens.

By placing a convex lens at the optical center of the camera, aligned with its axis, we allow multiple paths for light from a single point in the scene to reach the same point on the film. In contrast, a pinhole camera only permits light to travel along a single straight-line path between the source and destination. Mathematically, a thin convex lens has the following properties:

- Light rays passing through the lens center remain undeviated. Thus, the lens center behaves like a pinhole camera.
- *Focal Point and Focal Length*: Parallel rays converge at a specific point known as the *focal point*. The distance between the focal point and the optical center of the lens is called the *focal length*.

3.1 Thin Lens Formula

For a given point in the scene, there exists a specific distance at which the point appears in focus. This relationship is governed by the thin lens equation, which relates the object distance (D), the image distance (D'), and the focal length (f):

$$\frac{1}{D'} + \frac{1}{D} = \frac{1}{f} \tag{4}$$

The formula can be derived by applying similar triangles to the geometric setup shown in Figure 2.

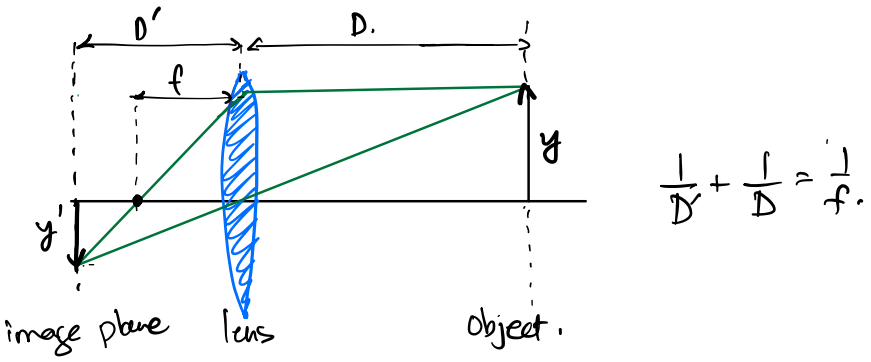


Figure 2: Thin lens geometry.

Points that do not satisfy the thin lens formula project to a circle, also known as the *circle of confusion*. The size of this circle (or blur) depends on how far the point deviates from the ideal depth that satisfies the thin lens formula.

3.2 Lens Phenomenon

The *depth of field* of a camera refers to the range of depth values for which objects appear acceptably sharp in the image. A pinhole camera has an *infinite depth of field*, meaning that objects at any depth appear perfectly sharp. However, a camera with a lens has a *finite depth of field*. The depth of field is inversely related to the *focal length*—cameras with a small focal length (e.g., microscopes) have a narrow depth of field, while those with a large focal length (e.g., telescopes) have a large depth of field.

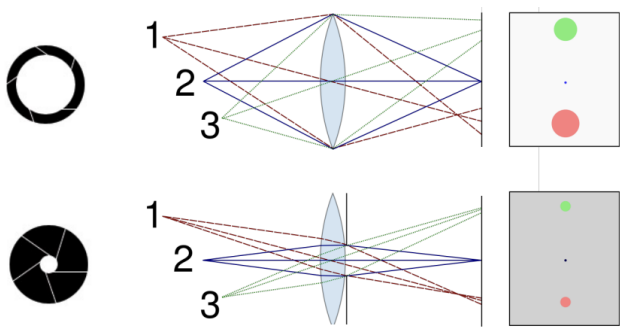
Independently of the focal length, one can control the depth of field by adjusting the *aperture size* (Figure 3a). Reducing the aperture size increases the depth of field, and in the limiting case, it approximates a pinhole camera, which has an *infinite depth of field*. A fun fact is that *pinhole glasses* can be used as an alternative to prescription glasses. They remove the lens of your eye from the equation, allowing for clear vision (on a bright day).

Another important concept is the *field of view* (FoV), which is the angular extent of the scene captured by the camera. Cameras with a large focal length have a small FoV. By adjusting the focal length while moving the camera (Figure 3b), one can maintain a constant FoV, creating the *Dolly Zoom* effect.

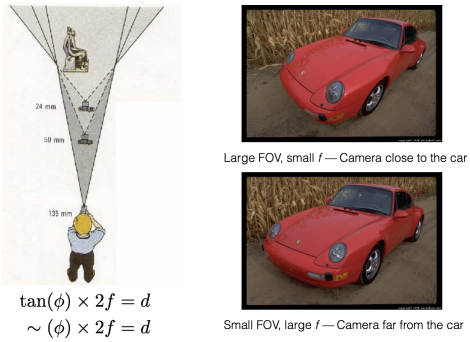
3.3 Lens Flaws

Simple lenses do not perfectly obey the thin lens formula and exhibit several optical flaws, including:

- *Chromatic Aberration*: Color fringing occurs due to different refractive indices for different wavelengths of light. For example, in Figure 4a, blue light bends more than red light.
- *Spherical Aberration*: Blur results from imperfections in the lens shape, causing light rays to focus at different points (Figure 4b.



(a) changing the aperture



(b) changing the field of view and focal length

Figure 3: Effect of changing the aperture, field of view, and focal length in a camera. (a) Making the aperture smaller increases the depth of field of the camera, and the camera behaves like a pinhole camera in the limit when the aperture size goes to zero. (b) By adjusting the the field of view θ so that the object roughly occupies the entire image as the camera moves we get the "dolly zoom" effect `https://en.wikipedia.org/wiki/Dolly_zoom`.

- *Radial Distortion*: Warping effects such as barrel, pincushion, or mustache distortion alter the image geometry (Figure 4).

High-quality camera lenses (e.g., Carl Zeiss Tessar) use a complex arrangement of concave and convex lenses, along with mirrors, to approximate a thin lens with a fixed focal length. Adjustable zoom lenses incorporate even more intricate optical designs. These compound lenses are expensive to manufacture but are essential in applications such as cameras, telescopes, and microscopes.

4 Early Color Photography and the Three Color Technique

By the early 19th century, technology had been developed to record images. The process typically involved coating a flat surface with a light-sensitive material (e.g., silver halide), which was then exposed to light to form a negative image. Later advancements made the process more compact, such as the introduction of roll film. These negatives had to be chemically developed to produce a photograph. Early photographic plates and films could only capture intensity levels, resulting in black-and-white images.

One of the first advancements in color photography was the three-color technique pioneered by Sergey Prokudin-Gorskii. His method involved capturing three separate black-and-white images of the same scene, each through a different color filter—red, green, and blue. These images were then combined to reconstruct a full-color photograph.

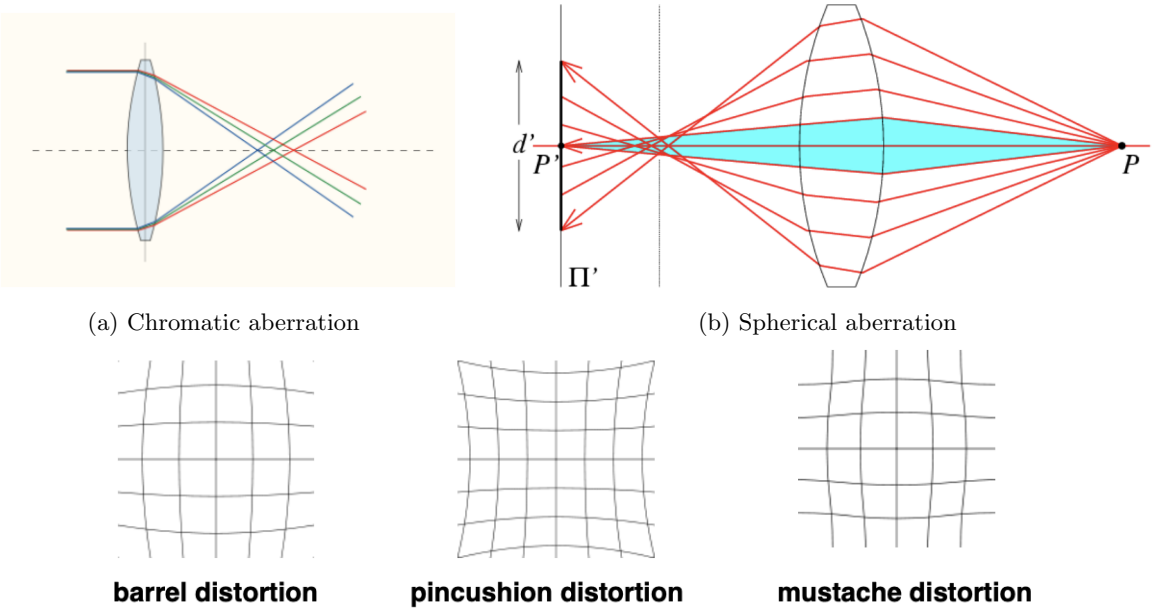
Prokudin-Gorskii used a special camera to take three consecutive black-and-white photographs, with each exposure passing through a red, green, or blue filter. A color filter allows only a specific range of light wavelengths to pass through. As a result, each of the three negatives contained the brightness information corresponding to a different primary color channel. The negatives were later aligned and projected using colored light filters to recreate a full-color image as shown in Figure 5. Alternatively, modern digital techniques can align the negatives and reconstruct the color image for display on screens.

Prokudin-Gorskii used this technique to document the Russian Empire in the early 1900s, creating one of the first extensive color photography archives. His work was a precursor to modern color photography and influenced later innovations, such as the introduction of Kodachrome film in 1935.

4.1 Digital Reconstruction by Alignment

Two main drawbacks of the three-color technique are that it requires a long exposure time, as three separate images must be captured, and that the negatives must be spatially aligned. It is quite possible that the camera or the scene moves slightly between captures. As a result, the three negatives may not align perfectly. Simply stacking the brightness components of the three primary channels together results in an incorrect image filled with color artifacts.

Fortunately, we can address the alignment problem by exploiting the spatial correlation between the color channels. Although the three brightness images are different, they share many similarities because they depict the same scene—bright regions in one image often correspond to bright regions in other



(a) Chromatic aberration (b) Spherical aberration

Figure 4: Real lenses exhibit flaws including (a) chromatic aberration as different wavelengths have different refractive indices; (b) spherical aberration leading to lack of a single point of focus; as well as non-linear distortions.

Figure 5: The three-color technique combines black-and-white images obtained through different color filter to reconstruct a full photograph. However the three images may be misaligned due to camera motion between the shots.

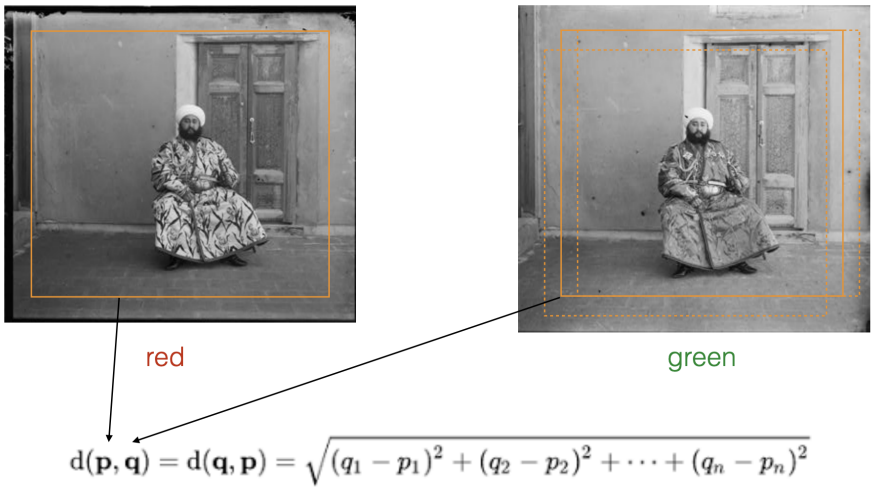


Figure 6: Aligning images by searching over horizontal and vertical translations and measuring similarity using Euclidean distance. The optimal translation is the one that maximizes similarity.

channels. While this holds true statistically, it is not always the case. For example, a bright blue pixel will have a high value in the blue channel but low values in the red and green channels.

We can leverage the similarity between color channels to solve the alignment problem. The basic idea is to search for transformations between channels that maximize their similarity. While the space of possible transformations can be quite large (e.g., including rotation, zoom, translation, etc.), for simplicity, let's assume that the transformation is restricted to horizontal and vertical translations within a small range, given by: $\delta x \in [-k, k]$ and $\delta y \in [-k, k]$.

For each possible value of $\delta x, \delta y$ we compute the similarity between the red and green channels and select the shift that results in the highest similarity as shown in Figure 6. This process aligns the green channel with the red channel. We can then repeat the process for the red and blue channels to align the blue channel with the red. Finally, we construct the color image by stacking the aligned red, green, and blue channels with the estimated shifts applied.

4.2 Measuring Similarity Between Two Black-and-White Images

There are many ways to measure the similarity between two black-and-white images. One simple approach is to compute the sum of squared differences (SSD) between corresponding pixels. Since SSD is a distance metric, smaller values indicate higher similarity. Mathematically, given two images $I(x, y)$ and $J(x, y)$, the SSD is computed as:

$$\mathrm{SSD}(I, J) = \sqrt{\sum_{x,y} \left(I(x, y) - J(x, y)\right)^2} \tag{5}$$

Another common similarity measure is the cosine similarity, defined as:

$$\mathrm{COSINE}(I, J) = \frac{\sum_{x,y} I(x, y) J(x, y)}{\sqrt{\sum_{x,y} I(x, y)^2} \sqrt{\sum_{x,y} J(x, y)^2}} \tag{6}$$

Cosine similarity computes the dot product between two images, treating them as unit vectors. Its value ranges between 0 and 1, where higher values indicate greater similarity.

A related metric is the normalized cross-correlation (NCC), which first subtracts the mean intensity before computing the cosine similarity:

$$\mathrm{NCC}(I, J) = \frac{\sum_{x,y} (I(x, y) - \mu_I)(J(x, y) - \mu_J)}{\sqrt{\sum_{x,y} (I(x, y) - \mu_I)^2} \sqrt{\sum_{x,y} (J(x, y) - \mu_J)^2}} \tag{7}$$

where μ_I and μ_J represent the mean intensity values of images I and J, respectively.

4.3 Sensitivity to Image Shifts

These similarity measures are highly sensitive to image shifts. For example, if an image is shifted a few pixels to the right, the SSD with the original image will increase significantly in most cases.

For image alignment, this sensitivity is actually desirable, as it helps estimate pixel-level shifts. However, for image recognition, we often prefer similarity measures that are less sensitive to slight shifts. A cat remains a cat even if it is shifted a few pixels to the right.

Later in this course, when we discuss image recognition, we will introduce alternative image representations that are more robust to pixel shifts.

5 Modern Color Films

Color photographic film, pioneered by Kodak in the 1930s and further refined in the 1940s, contained multiple layers of light-sensitive emulsions and dye couplers to capture light of different wavelengths simultaneously. Each layer was sensitive to a different primary color—red, green, or blue—allowing the film to record a full-color image in a single exposure. This advancement solved the alignment problem inherent in the three-color technique, by eliminating the need to capture separate images for each color channel. However, this required a complex film design and a sophisticated chemical development process to properly render colors.

Kodak and Fujifilm were among the leading companies that pioneered and commercialized color films for producing paper prints, with Kodak introducing Kodachrome in 1935 and Fujifilm later developing Fujicolor in the mid-20th century. These films became widely popular for both amateur and professional photography. Over time, improvements in film sensitivity, color reproduction, and development techniques led to higher-quality images with more vibrant and accurate colors. Despite the dominance of digital photography today, color film remains in use among photography enthusiasts and professionals.

6 Digital Camera Sensors

A digital camera replaces traditional photographic film with a sensor array that captures incoming light electronically. Each cell in this array is a light-sensitive photodiode that converts incoming photons into electrical signals, which are then processed to form a digital image. The two most common types of sensor arrays used in modern cameras are:

- Charge-Coupled Device (CCD): Known for high image quality and low noise, but more power-intensive and expensive.
- Complementary Metal-Oxide-Semiconductor (CMOS): More power-efficient, cheaper to manufacture, and widely used in consumer and professional cameras today.

To capture color information, a color filter array (CFA) is placed over the sensor, ensuring that each pixel records only one of the primary colors—red, green, or blue. One widely used CFA is the Bayer filter, which consists of a grid where 50% of the pixels are green, 25% are red, and 25% are blue as shown in Figure 7. This mosaic arrangement allows the sensor to approximate human vision, which is more sensitive to green light. However, because each pixel only captures a single color, the raw sensor output is a mosaiced image, where each pixel corresponds to only one color value.

6.1 Demosaicing: Estimating Missing Color Information

To reconstruct a full-color image, the missing values for each color channel must be estimated through a process called demosaicing. This process leverages the spatial correlation between pixels in natural images—meaning that neighboring pixels in a given color channel tend to have similar values, as natural images usually exhibit smoothly varying structures. This assumption, however, breaks down at edges and discontinuities, where intensity values change abruptly. Nevertheless, such discontinuities are relatively rare compared to smooth regions. In contrast, images containing random noise have little to no spatial correlation, making them more challenging to reconstruct.

The simplest demosaicing method is the nearest-neighbor interpolation, where the missing pixel values are copied from the nearest known pixel of the same color channel. A slightly more refined approach is bilinear interpolation, which averages the values of surrounding known pixels to estimate the missing

ones. More advanced methods, such as adaptive gradient-based interpolation, take into account edge structures in the image. In this approach, the algorithm selects whether to interpolate using vertical or horizontal neighbors, depending on which direction exhibits greater similarity. This reduces artifacts and improves the accuracy of color reconstruction. Figure 8 illustrates different interpolation methods for demosiacing.

Figure 7: Illustration of the Bayer grid (right) and the concept of demosaicing (right)

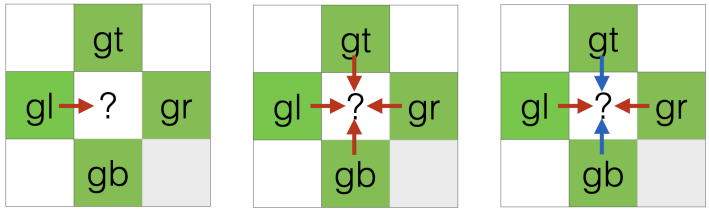


Figure 8: Illustration of nearest neighbor, linear, and adaptive gradient interpolation

For the blue and red channels, the interpolation schemes are more complex and less reliable as there are more missing pixels with the Bayer filter pattern. You could apply different techniques for different channels, and different interpolation schemes for different types of pixel neighborhoods within the same channel as well.

More sophisticated demosaicing techniques, including edge-aware interpolation, frequency-based methods, and machine learning approaches, can be used to improve color reconstruction further. Some techniques apply different interpolation strategies for different color channels, while others adjust the interpolation method based on local pixel neighborhoods within the same channel.

7 Optional readings

- Modern cameras are highly sophisticated, integrating both optical elements and image processing techniques to capture the best possible image.
- *Light field* cameras capture not only the intensity of light but also its direction, unlike traditional cameras that integrate light from all directions. This enables post-capture adjustments to the *depth of field* and other image properties. The field of *computational photography* explores techniques that combine optics, sensors, and computational methods to enhance imaging capabilities.