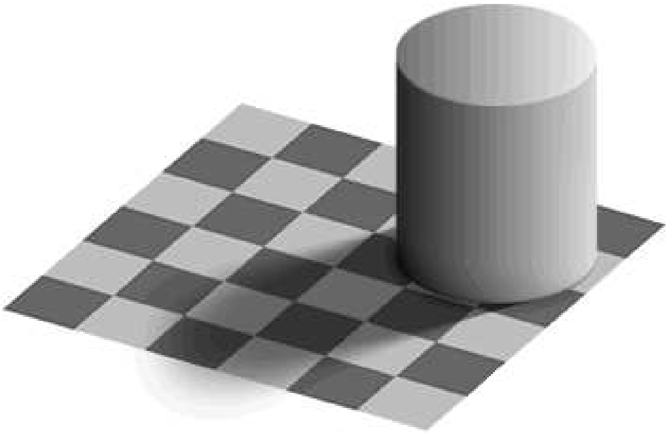
Image Formation

Goal: Introduce basic concepts in image formation and camera models.

Motivation:

Many of the algorithms in computational vision attempt to infer scene properties such as surface shape, surface reflectance, and scene lighting from image image data.



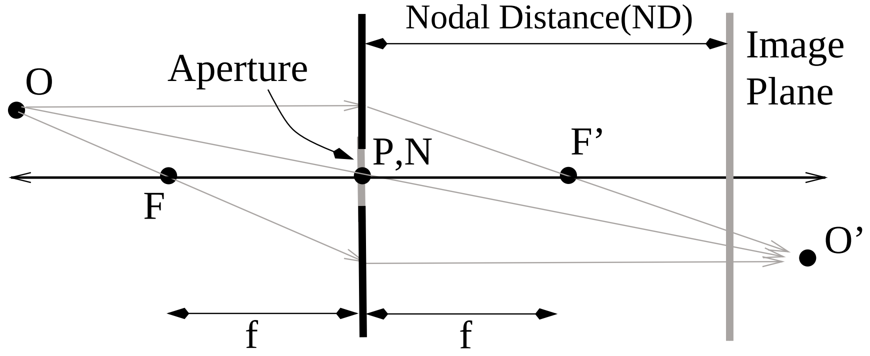
Here we consider the basic components of the "forward" model. That is, assuming various scene and camera properties, what should we observe in an image?

Readings: Part I (Image Formation and Image Models) of Forsyth and Ponce.

Matlab Tutorials: colourTutorial.m (in UTVis)

Thin Lens

Suppose we can neglect the thickness of the lens. And suppose the medium (e.g. air) is the same on both sides of the lens.



A point $\vec{\mathcal{O}}$ in the world is focussed at $\vec{\mathcal{O}}'$ given by the intersection of three rays as shown above:

- A ray from $\vec{\mathcal{O}}$ passing straight through the nodal point N of the camera (N is also called the center of projection).
- Two rays parallel to the optical axis on one side of the lens (aka principal plane), and passing through the front or rear focal points (F and F') on the opposite side of the lens.

Here the point $\vec{\mathcal{O}}$ is not in focus on the image plane, but rather at $\vec{\mathcal{O}}'$.

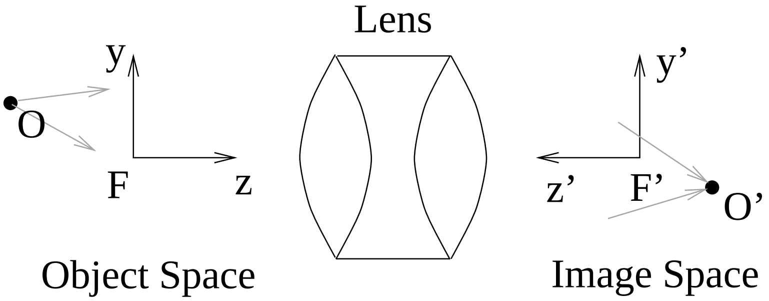
The *aperture* (eg. pupil) of the lens can be modelled as being an occluder within the principal plane.

General Lens Model

Given a general lens, a point at $\vec{\mathcal{O}}$ is imaged to $\vec{\mathcal{O}}'$, where the locations of $\vec{\mathcal{O}}$ and $\vec{\mathcal{O}}'$ are given by the **lens forumla**:

$$\vec{\mathcal{O}}' \equiv \begin{pmatrix} z' \\ y' \end{pmatrix} = \frac{f}{z} \begin{pmatrix} f' \\ y \end{pmatrix}, \quad \vec{\mathcal{O}} \equiv \begin{pmatrix} z \\ y \end{pmatrix} = \frac{f'}{z'} \begin{pmatrix} f \\ y' \end{pmatrix}.$$

Here F, F' are focal points, and f, f' are focal lengths.



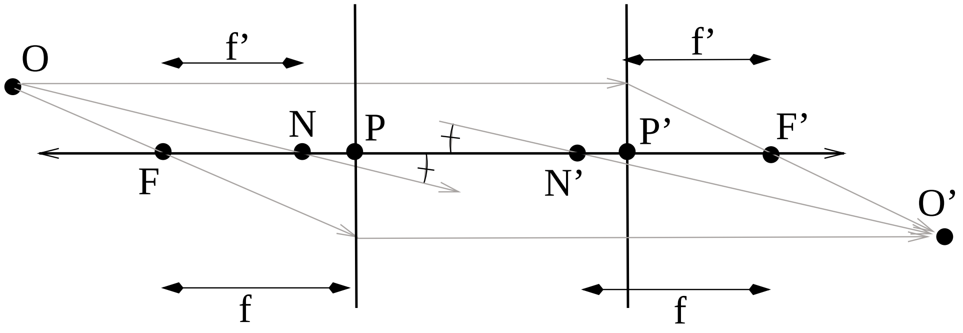
In general the ratio of focal lengths equals the ratio of the indices of refraction of the pre- and post-lens material, that is $f/f' = n/n'$ (eg. $f \neq f'$ for eyes, but $f = f'$ for most cameras). The index of refraction of a material is the ratio of the speed of light in a vacuum over the speed of light in the medium.

As for a thin lens, the formation of the image of $\vec{\mathcal{O}}$ can be interpreted geometrically as the intersection of three canonical rays, which are determined by the **cardinal points** of the lens. The cardinal points are:

Focal Points F, F' provide origins for the object and image spaces.

Nodal Points N, N', are defined using the lens axis, F, F', and focal lengths, f, f'.

Principal Points P, P' are also defined using the lens axis, F, F', and focal lengths, f, f'.



Lens Formula

An alternative coordinate system which is sometimes used to write the lens formula is to place the origins of the coordinates in the object and image space at the principal points P and P', and flip both the z-axes to be pointing away from the lens. These new z-coordinates are:

$$\begin{aligned} \hat{z} &= f - z, \\ \hat{z}' &= f' - z'. \end{aligned}$$

Solving for z and z' and substituting into the previous lens formula, we obtain:

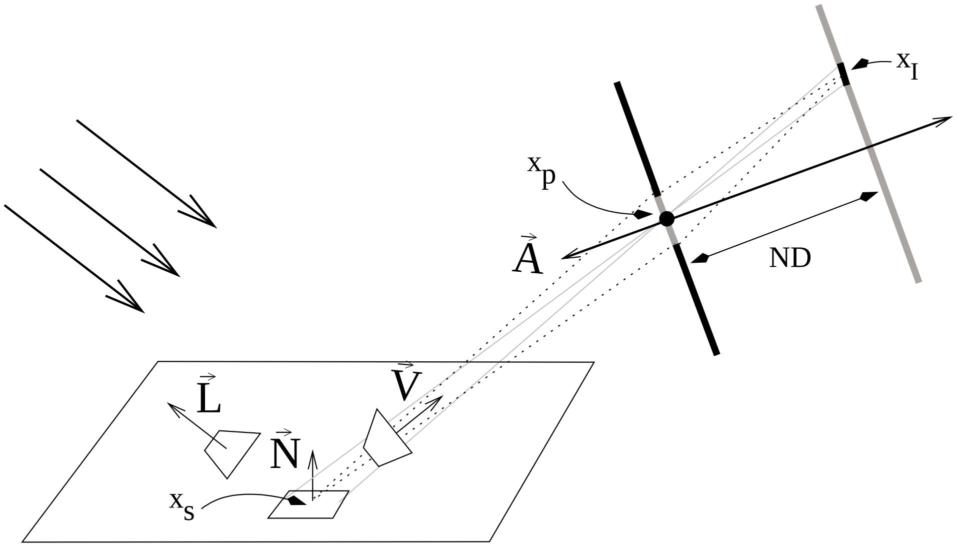
$$\begin{aligned} (f' - \hat{z}) &= ff'/(f - \hat{z}), \\ ff' &= (f' - \hat{z}')(f - \hat{z}) \\ \hat{z}'\hat{z} &= \hat{z}'f + \hat{z}f' \\ 1 &= \frac{f}{\hat{z}} + \frac{f'}{\hat{z}'} \end{aligned}$$

The last line above is also known as the lens formula. As we have seen, it is equivalent to the one on the previous page, only with a change in the definition of the coordinates.

For cameras with air both in front of and behind the lens, we have f = f'. This simplifies the lens formula above. Moreover, the nodal and principal points coincide in both the object and scene spaces (i.e., $N = P$ and $N' = P'$ in the previous figure).

Finally it is worth noting that, in terms of image formation, the difference between this general lens model and the thin lens approximation is only in the displacement of the cardinal points along the optical axis. That is, effectively, the change in the imaging geometry from a thin lens model to the general lens model is simply the introduction of an absolute displacement in the image space coordinates. For the purpose of modelling the image for a given scene, we can safely ignore this displacement and use the thin lens model. When we talk about the center of projection of a camera in a world coordinate frame, however, it should be understood we are talking about the location of the nodal point N in the object space (and not N' in the image space). Similarly, when we talk about the nodal distance to the image plane, we mean the distance from N' to the image plane.

Image of a Lambertian Surface



The irradiance on the image plane is

$$I(\lambda, \vec{x}_I, \vec{n}_I) = T_l \frac{d\Omega_p \, dA_V}{dA_I} r(\lambda) \lfloor \vec{N} \cdot \vec{L} \rfloor I(\lambda, \vec{L})$$

Here

- $\vec{n}_I$ is normal to the image plane;
- $T_l \in (0, 1]$ is the transmittance of the lens;
- dA_I is the area of each pixel;
- $d\Omega_p$ is the solid angle of the aperture from the surface point $\vec{x}_s$;
- dA_V is the area of the cross-section, perpendicular to the viewing direction, of the portion of the surface imaged to the pixel at $\vec{x}_I$.

Derivation of the Image of a Lambertian Surface

From our notes on Lambertian reflection, we have that the radiance (spectral density) of the surface is

$$R(\lambda, \vec{x}_s, \vec{V}) = r(\lambda) \lfloor \vec{N} \cdot \vec{L} \rfloor I(\lambda, \vec{L}).$$

This is measured in Watts per unit wavelength, per unit cross-sectional area perpendicular to the viewer, per unit steradian.

The total power (per unit wavelength) from the patch dA_V, arriving on the aperature, is

$$P(\lambda) = R(\lambda, \vec{x}_s, \vec{V}) d\Omega_p dA_V$$

A fraction T_l of this is transmitted through the lens, and ends up on a pixel of area dA_I. Therefore, the pixel irradiance spectral density is

$$I(\lambda, \vec{x}_I, \vec{n}_I) = T_l P(\lambda)/dA_I,$$

which is the expression on the previous page.

To simplify this, first compute the solid angle of the lens aperature, with respect to the surface point $\vec{x}_s$. Given the area of the aperature is dA_p, we have

$$d\Omega_p = \frac{|\vec{V} \cdot \vec{A}| dA_p}{||\vec{x}_p - \vec{x}_s||^2}.$$

Here the numerator is the cross-sectional area of the aperature viewed from the direction $\vec{V}$. The denominator scales this foreshortened patch back to the unit sphere to provide the desired measure of solid angle. Secondly, we need the foreshortened surface area dA_V which projects to the individual pixel at $\vec{x}_I$ having area dA_I. These two patches are related by rays passing through the center of projection $\vec{x}_p$; they have the same solid angle with respect to $\vec{x}_p$. As a result,

$$dA_V = ||\vec{x}_p - \vec{x}_s||^2 \frac{|\vec{V} \cdot \vec{A}| dA_I}{||\vec{x}_p - \vec{x}_I||^2}$$

The distance in the denominator here can be replaced by

$$||\vec{x}_p - \vec{x}_I|| = ND/|\vec{V} \cdot \vec{A}|.$$

Substituting these expression for $d\Omega_p$, dA_V, and $||\vec{x}_p - \vec{x}_I||$ gives the equation for the image irradiance due to a Lambertian surface on the following page.

Image of a Lambertian Surface (cont.)

This expression for the irradiance due to a Lambertian surface simplifies to

$$I(\lambda, \vec{x}_I, \vec{n}_I) = T_l \frac{dA_p}{|ND|^2} |\vec{A} \cdot \vec{V}|^4 \, r(\lambda) \lfloor \vec{N} \cdot \vec{L} \rfloor I(\lambda, \vec{L})$$

Here, dA_p is the area of the aperture.

Note the image irradiance:

- does not depend on the distance to the surface, $||\vec{x}_s - \vec{x}_p||$;
- falls off like $\cos(\theta)^4$ in the corners of the image. Here θ is the angle between the viewing direction $\vec{V}$ and the camera's axis $\vec{A}$. Therefore, for wide angle images, there is a significant roll-off in the image intensity towards the corners.

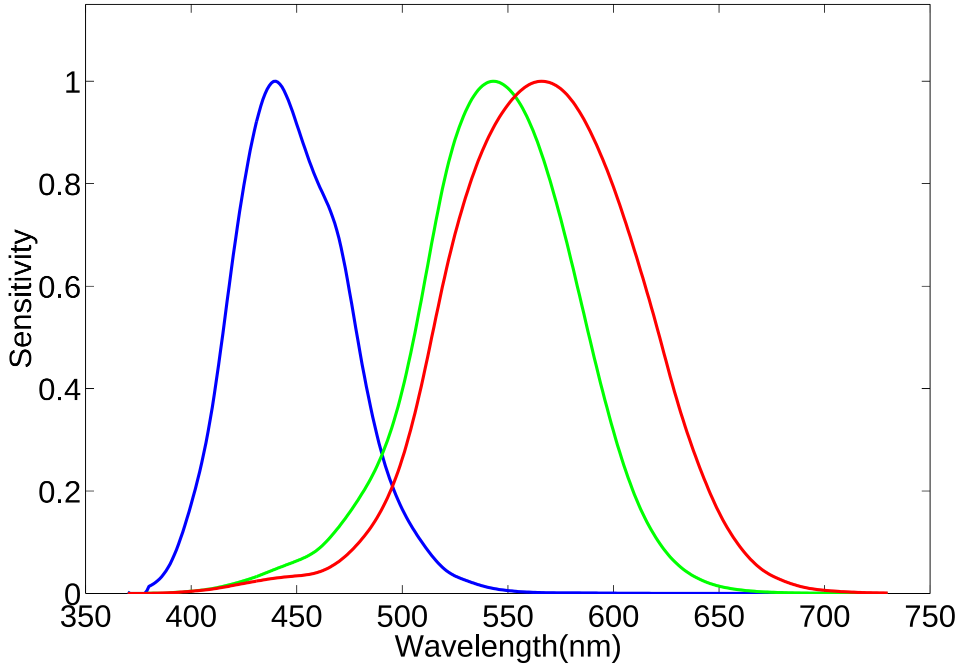
The fall off of the brightness in the corners of the image is called **vignetting**. The actual vignetting obtained depends on the internal structure of the lens, and will vary from the above $\cos(\theta)^4$ term.

Image Irradiance to Absorbed Energy

Spectral Sensitivity. The colour (or monochrome) pixel response is a function of the energy absorbed by that pixel. For a steady image, not changing in time, the absorbed energy at pixel $\vec{x}_I$ can be approximated by

$$e_\mu(\vec{x}_I) = \int_0^\infty S_\mu(\lambda) C_T A_I I(\lambda, \vec{x}_I, \vec{n}_I) d\lambda.$$

Here $I(\lambda, \vec{x}_I, \vec{n}_I)$ is the image irradiance, $S_\mu(\lambda)$ is the spectral sensitivity of the μ^{th} colour sensor, A_I is the area of the pixel, and C_T is the temporal integration time (eg. 1/(shutter speed)).



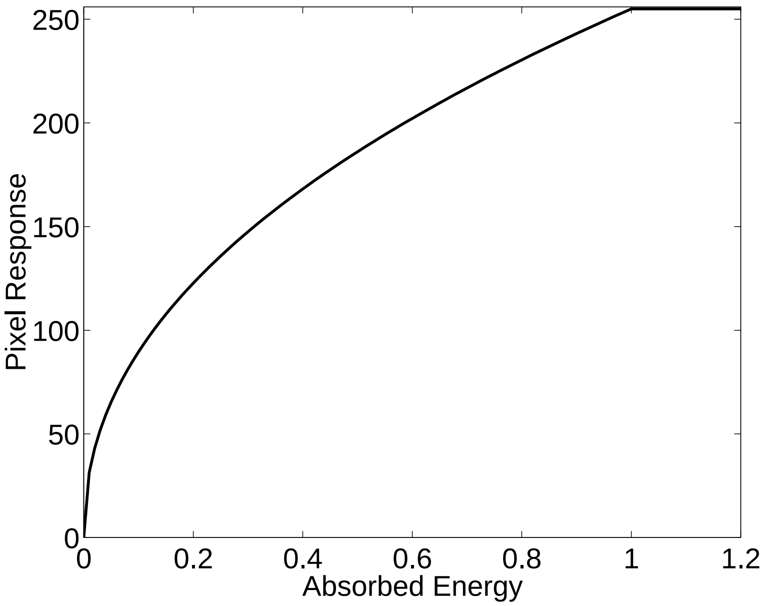
Colour images are formed (typically) using three spectral sensitivities, say $\mu = R, G, B$ for the ‘red’, ‘green’ and ‘blue’ channel. The normalized spectral sensitivities in the human retina are plotted above.

Absorbed Energy to Pixel Response

Gamma Correction. Finally, the absorbed energy e_μ is converted to a quantized pixel response, say r_μ, through a nonlinear function called a gamma correction, for example,

$$r_\mu = \beta \left[e_\mu \right]^{\frac{1}{\gamma}} .$$

Here the value of γ can vary, values between 2 and 3 are common. This response r_μ is quantized, typically to 8 bits.

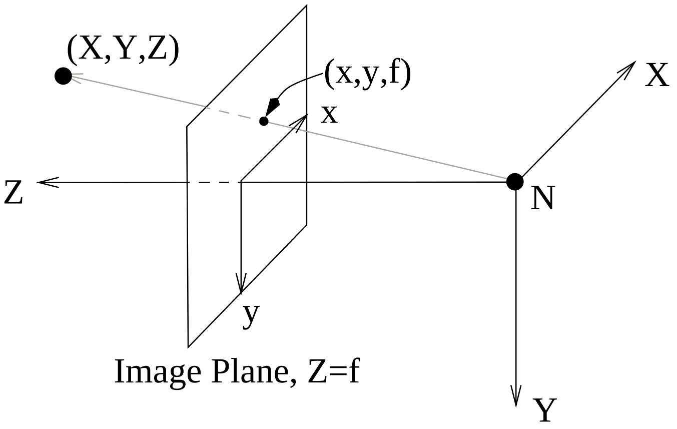


This completes our basic scene and image formation models.

We consider several approximations and simplifications of this model next.

The Pinhole Camera

The image formation of both thick and thin lenses can be approximated with a simple pinhole camera,



The image position for the 3D point (X, Y, Z) is given by the projective transformation

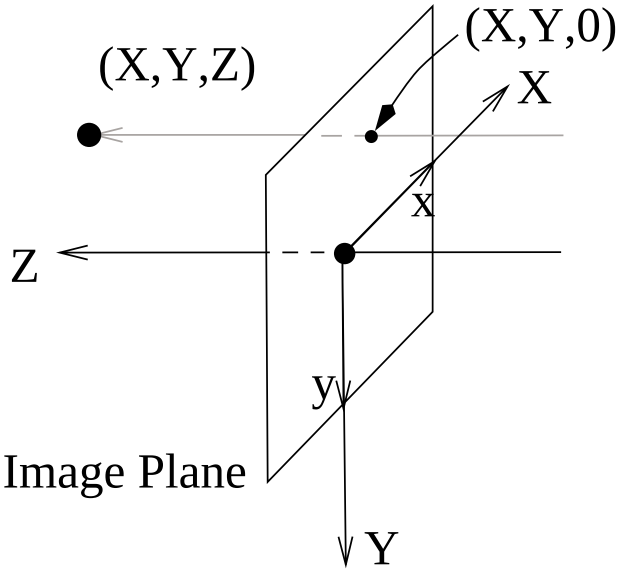
$$\begin{pmatrix} x \\ y \\ f \end{pmatrix} = \frac{f}{Z} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

By convention, the nodal distance $|ND|$ is labelled as f (the "focal length"). Note:

- for mathematical convenience we put the image plane in front of the nodal point (since this avoids the need to flip the image coords about the origin);
- image coordinate x is taken to the right, and y *downwards*. This agrees with the standard raster order.
- the primary approximation here is that all depths are taken to be in focus.

Orthographic Projection

Alternative projections onto an image plane are given by orthographic projection and scaled orthographic projection.



Given a 3D point (X, Y, Z), the corresponding image location under scaled orthographic projection is

$$\begin{pmatrix} x \\ y \end{pmatrix} = s_0 \begin{pmatrix} X \\ Y \end{pmatrix}$$

Here s_0 is a constant scale factor; orthographic projection uses $s_0 = 1$.

Scaled orthographic projection provides a linear approximation to perspective projection, which is applicable for a small object far from the viewer and close to the optical axis.

Coordinate Frames

Consider the three coordinate frames:

- a world coordinate frame $\vec{X}_w$,
- a camera coordinate frame, $\vec{X}_c$,
- an image coordinate frame, $\vec{p}$.

The world and camera frames provide standard 3D orthogonal coordinates. The image coordinates are written as a 3-vector, $\vec{p} = (p_1, p_2, 1)^T$, with p_1 and p_2 the pixel coordinates of the image point.

Camera Coordinate Frame. The origin of the camera coordinates is at the nodal point of the camera (say at $\vec{d}_w$ in world coords). The z-axis is taken to be the optical axis of the camera (with points in front of the camera having a positive z value).

Next we express the transforms from world coordinates to camera coordinates and then to image coordinates.

External Calibration Matrix

The external calibration parameters specify the transformation from world to camera coordinates.

This has the form of a standard 3D coordinate transformation,

$$\vec{X}_c \;=\; M_{ex}[\vec{X}_w^T, 1]^T, \tag{1}$$

with M_{ex} a 3×4 matrix of the form

$$M_{ex} \;=\; \begin{pmatrix} R & -R\vec{d}_w \end{pmatrix}. \tag{2}$$

Here R is a 3×3 rotation matrix and $\vec{d}_w$ is the location of the nodal point for the camera in world coordinates.

The inverse of this mapping is simply

$$\vec{X}_w \;=\; R^T\vec{X}_c + \vec{d}_w. \tag{3}$$

In terms of the camera coordinate frame, the perspective transformation of the 3D point $\vec{X}_c$ (in the camera's coordinates) to the image plane is

$$\vec{x}_c \;=\; \frac{f}{X_{3,c}}\vec{X}_c \;=\; \begin{pmatrix} x_{1,c} \\ x_{2,c} \\ f \end{pmatrix}. \tag{4}$$

Here f is the nodal distance for the camera.

Internal Calibration Matrix

The internal calibration matrix transforms the 3D image position $\vec{x}_c$ to pixel coordinates,

$$\vec{p} \;=\; M_{in} \vec{x}_c, \tag{5}$$

where M_{in} is a 3×3 matrix.

For example, a camera with rectangular pixels of size $1/s_x$ by $1/s_y$, with focal length f, and piercing point (o_x, o_y) (i.e., the intersection of the optical axis with the image plane provided in pixel coordinates) has the internal calibration matrix

$$M_{in} \;=\; \begin{pmatrix} s_x & 0 & o_x/f \\ 0 & s_y & o_y/f \\ 0 & 0 & 1/f \end{pmatrix}. \tag{6}$$

Note that, for a 3D point $\vec{x}_c$ on the image plane, the third coordinate of the pixel coordinate vector $\vec{p}$ is $p_3 = 1$. As we see next, this redundancy is useful.

Equations (1), (4) and (5) define the transformation from $\vec{X}_w$, the world coordinates of a 3D point to $\vec{p}$, the pixel coordinates of the image of that point. The transformation is nonlinear, due to the scaling by $X_{3,c}$ in equation (4).

Homogeneous Coordinates

It is useful to express this transformation in terms of homogeneous coordinates,

$$\begin{aligned}\vec{X}_w^h &= a(\vec{X}_w^T, 1)^T,\\ \vec{p}^h &= b\vec{p} = b(p_1, p_2, 1)^T,\end{aligned}$$

for arbitrary nonzero constants a, b. The last coordinate of these homogeneous vectors provide the scale factors. It is therefore easy to convert back and forth between the homogeneous forms and the standard forms.

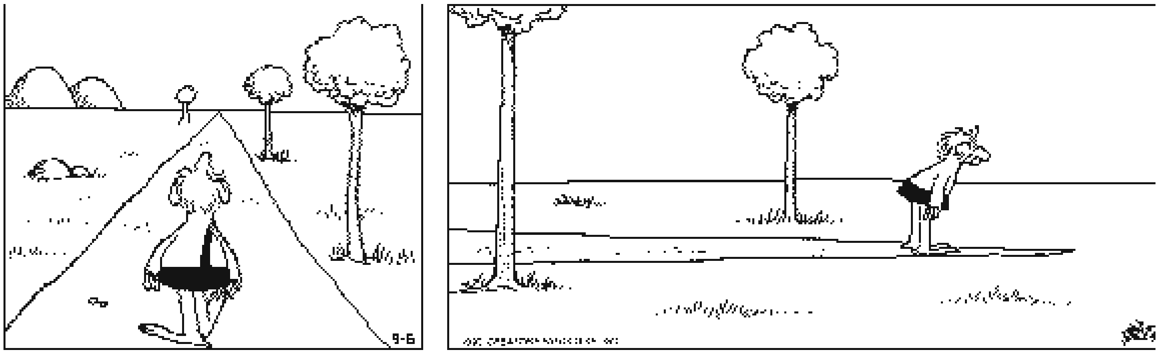
The mapping from world to pixel coordinates can then be written as the *linear* transformation,

$$\vec{p}^h = M_{in} M_{ex} \vec{X}_w^h. \tag{7}$$

Essentially, the division operation in perspective projection is now implicit in the homogeneous vector $\vec{p}^h$. It is simply postponed until $\vec{p}^h$ is rescaled by its third coordinate to form the pixel coordinate vector $\vec{p}$.

Due to its linearity, equation (7) is useful in many areas of computational vision.

Parallel Lines Project to Intersecting Lines



As an application of (7), consider a set of parallel lines in 3D, say

$$\vec{X}_k^h(s) \;=\; \begin{pmatrix} \vec{X}_k^0 \\ 1 \end{pmatrix} + s \begin{pmatrix} \vec{t} \\ 0 \end{pmatrix}.$$

Here $\vec{X}_k^0$, for $k = 1, \ldots, K$, and $\vec{t}$ are 3D vectors in the world coordinate frame. Here $\vec{t}$ is the common 3D tangent direction for all the lines, and $\vec{X}_k^0$ is an arbitrary point on the k^{th} line.

Then, according to equation (7), the images of these points in homogeneous coordinates are given by

$$\vec{p}_k^{\,h}(s) \;=\; M\vec{X}_k^h(s) \;=\; \vec{p}_k^{\,h}(0) + s\vec{p}_t^{\,h},$$

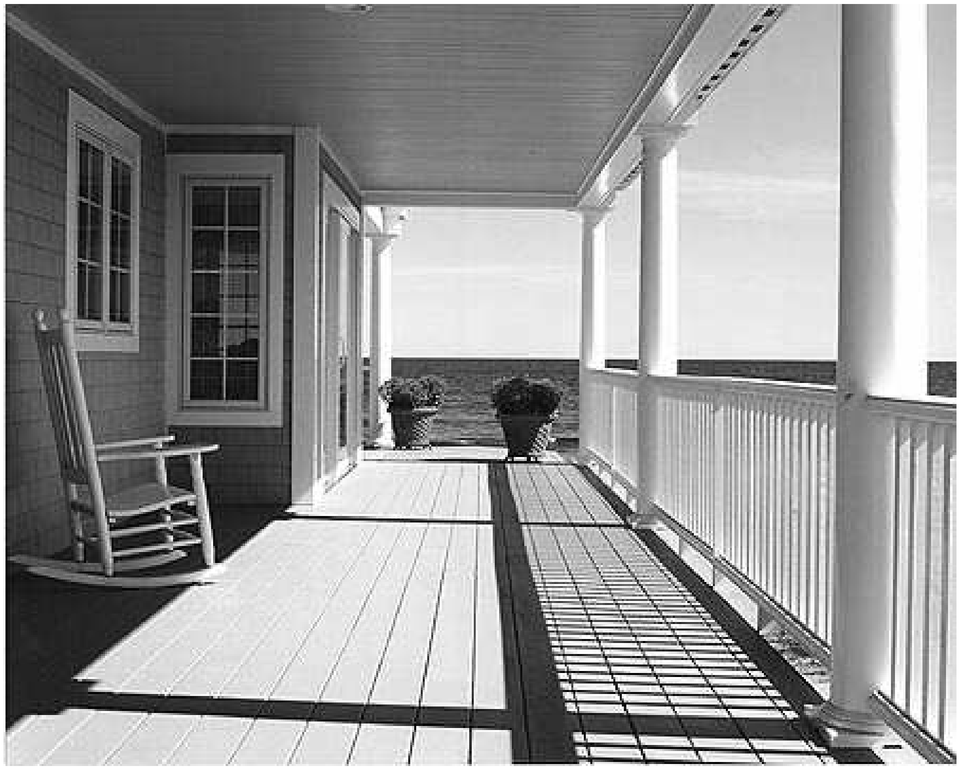
where $M = M_{in}M_{ex}$ is a 3×4 matrix, $\vec{p}_t^{\,h} = M(\vec{t}^T, 0)^T$ and $\vec{p}_k^{\,h}(0) = M((\vec{X}_k^0)^T, 1)^T$. Note $\vec{p}_t^{\,h}$ and $\vec{p}_k^{\,h}(0)$ are both constant vectors, independent of s. Converting to standard pixel coordinates, we have

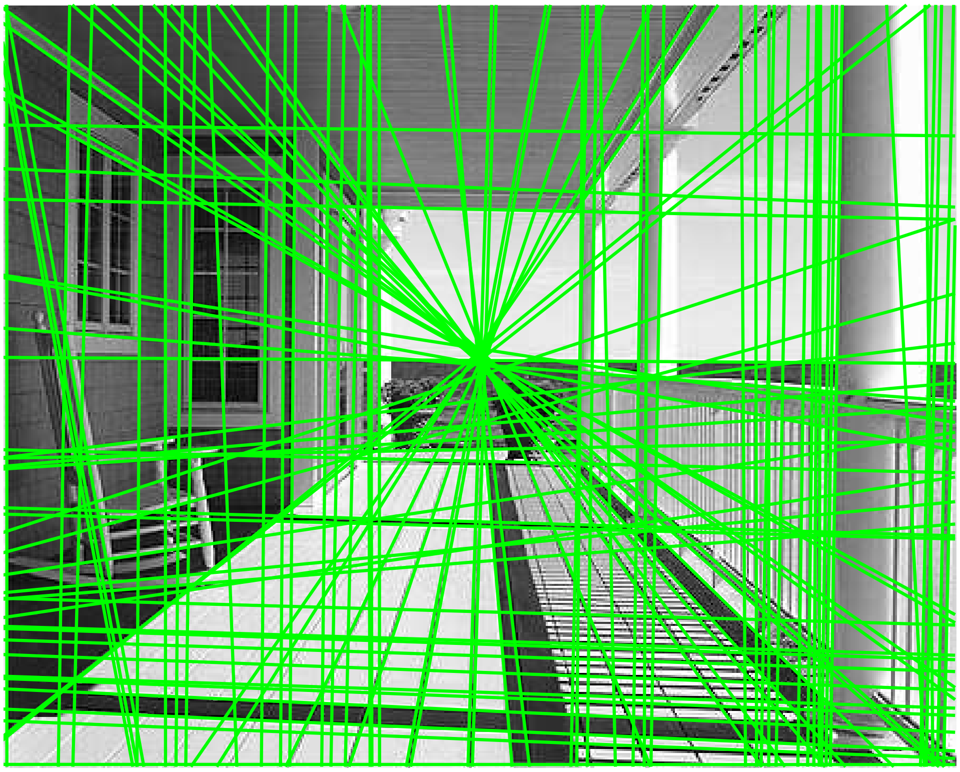
$$\vec{p}_k(s) \;=\; \frac{1}{\alpha(s)}\vec{p}_k^{\,h}(0) + \frac{s}{\alpha(s)}\vec{p}_t^{\,h},$$

where $\alpha(s) = p_{k,3}^h(s)$ is third component of $\vec{p}_k^{\,h}(s)$. Therefore we have shown $\vec{p}_k(s)$ is in the subspace spanned by two constant 3D vectors. It is also in the image plane, $p_{k,3} = 1$. Therefore it is in the intersection of these two planes, which is a line in the image. That is, lines in 3D are imaged as lines in 2D. (Although, in practice, some lenses introduce “radial distortion”, which causes the image of a 3D line to be bent. However, this distortion can be removed with careful calibration.)

In addition it follows that $\alpha(s) = p_{k,3}^h(0) + \beta s$ where $\beta = p_{t,3}^h = (0,0,1)M(\vec{t}^T, 0)^T$. Assuming $\beta \neq 0$, we have $1/\alpha(s) \to 0$ and $s/\alpha(s) \to 1/\beta$ as $s \to \infty$. Therefore the image points $\vec{p}_k(s) \to (1/\beta)\vec{p}_t^{\,h}$, which is a constant image point dependent only on the tangent direction of the 3D lines. This shows that the images of the parallel 3D lines $\vec{X}_k^h(s)$ all intersect at the image point $(1/\beta)\vec{p}_t^{\,h}$.

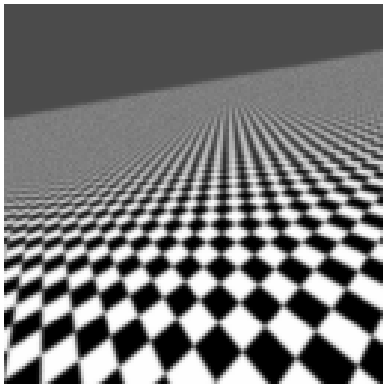
Example of Parallel Lines





The Horizon Line

Suppose the parallel lines discussed on the previous page are all coplanar. Then the analysis above shows that the images of these lines all intersect at the horizon (i.e., the image of points on the plane infinitely far from the camera). This property is depicted in the left panel of the previous cartoon. As another exercise in projective geometry, we will show that the horizon of a planar surface is a straight line in the image.



Consider multiple families of parallel lines in the plane, with each family having the tangent direction $\vec{t}_j$ in 3D. From the previous analysis, the j^{th} family must co-intersect at the image point (in homogeneous coordinates)

$$\vec{p}_j^{\,h} = M(\vec{t}_j^{\,T}, 0)^T,$$

and these points $\vec{p}_j^{\,h}$ must be on the horizon. However, since the tangent directions are all coplanar in 3D, two distinct directions provide a basis. That is, assuming the first two directions are linearly independent, we can write

$$\vec{t}_j = a_j\vec{t}_1 + b_j\vec{t}_2,$$

for some constants a_j and b_j. As a result, we have

$$\vec{p}_j^{\,h} = M([a_j\vec{t}_1 + b_j\vec{t}_2]^{\,T}, 0)^T = a_j\vec{p}_1^{\,h} + b_j\vec{p}_2^{\,h}$$

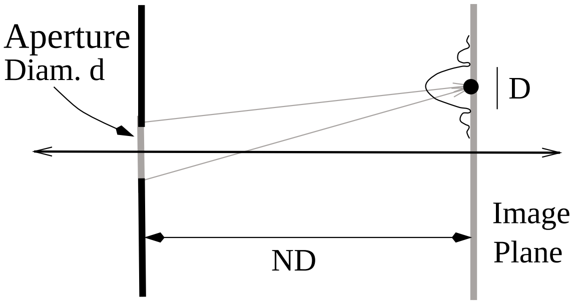
Dividing through by the third coordinate, $p^h_{j,3}$, we find the points of intersection of the j^{th} family of lines is at the image point

$$\vec{p}_j = \left(\frac{1}{p^h_{j,3}}\right)\vec{p}_j^{\,h} == \left(\frac{a_j p^h_{1,3}}{p^h_{j,3}}\right)\vec{p}_1 + \left(\frac{b_j p^h_{2,3}}{p^h_{j,3}}\right)\vec{p}_2 = \alpha_j\vec{p}_1 + \beta_j\vec{p}_2.$$

From the third coefficient of this equation it follows that $\alpha_j + \beta_j = 1$. Hence the image point $\vec{p}_j$ is an affine combination of the image points $\vec{p}_1$ and $\vec{p}_2$. Therefore the horizon must be the line in the image passing through $\vec{p}_1$ and $\vec{p}_2$.

Physical Limitations to Image Resolution

1. **Diffraction**



Even a properly focussed point is *not* imaged to a point. Rather, there is a *point spread function* (PSF).

For diffraction alone, this PSF can be modelled using the 'Airy disk', which has diameter

$$D \approx \frac{1.22\lambda}{n'}\frac{|ND|}{d},$$

where d is the aperture diameter. Lens imperfections and imperfect focusing lead to larger blur diameters.

Diffraction Limit (cont.)

For example, for human eyes (see Wyszecki & Stiles, Color Science, 1982):

- the index of refraction within the eye is $n' = 1.33$;
- the nodal distance is $|ND| \approx 16.7mm$ (accommodated at ∞);
- the pupil diameter is $d \approx 2mm$ (adapted to bright conditions);
- a typical wavelength is $\lambda \approx 500nm$.

Therefore the diameter of the Airy disk is

$$D \approx 4\mu = 4 \times 10^{-6}m$$

This compares closely to the diameter of a foveal cone (i.e. the smallest pixel), which is between 1μ and 4μ. So, human vision operates at the diffraction limit.

By the way, a 2μ pixel spacing in the human eye corresponds to having a 300×300 pixel resolution of the image of your thumbnail at arm's length. Compare this to the typical sizes of images used by machine vision systems, usually about 500×500 or less.

2. **Photon Noise**

   The average photon flux (spectral density) at the image (in units of photons per sec, per unit wavelength, per image area) is

   $$I(\lambda, \vec{x}_I, \vec{n}_I)\frac{\lambda}{\hbar c}$$

   Here $\hbar$ is Planck's constant and c is the speed of light.

   The photon arrivals can be modelled with Poisson statistics, so the variance is equal to the mean photon catch.

   Even in bright conditions, foveal cones have a significant photon noise component (a std. dev. ≈ 10% of the signal, for unshaded scenes).

3. **Defocus**

   An improperly focussed lens causes the PSF to broaden. Geometrical optics can be used to get a rough estimate of the size.

4. **Motion Blur**

   Given temporal averaging, the image of a moving point forms a streak in the image, causing further blur.

Conclude: There is a limit to how small standard cameras and eyes can be made (but note multi-faceted insect eyes). Human vision operates close to the physical limits of resolution (ditto for insects).