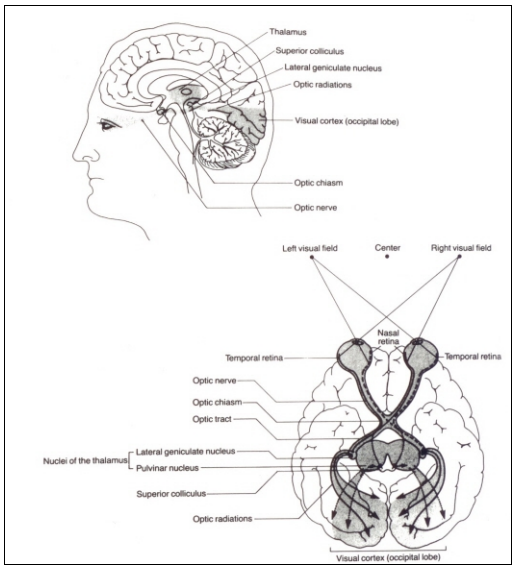
Table of Contents

Representing the World

The problem posed by understanding how humans build a representation of the world from visual input was posed in terms of the mind, not the brain, during the middle ages and the renaissance. Hence, the field of experimentation remained closed for an extended period of time. It is not until neuroscience emerged that an understanding of the visual system could be developed.



Sensory Transducers

Sensory transducers constitute the interface between the world and the brain. In the visual system, these are the rods and cones on the retina. The human visual system has 10^8 transducers. However, only 10^6 axons leave the retina for the visual cortex, suggesting

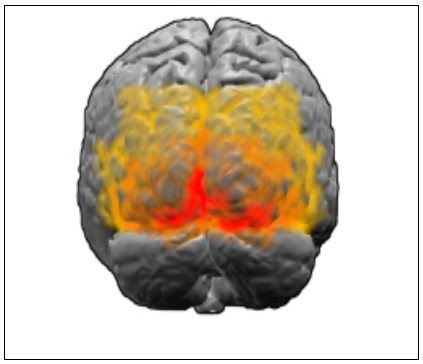
signal compression is occurring at the transducer level. This represents a ratio of 100:1 in the number of photoreceptors and the number of axons in the optic nerve.

Ganglion cells make up the optic nerve. They largely are center-surround in their response pattern. The response delay from photon onto the retina to ganglion cells firing is around 25 milliseconds. Ganglion cells carry information about color, motion, contours, and location of visual stimuli. The visual cortex is involved with the recognition of shapes, speed and direction of motion, and depth.

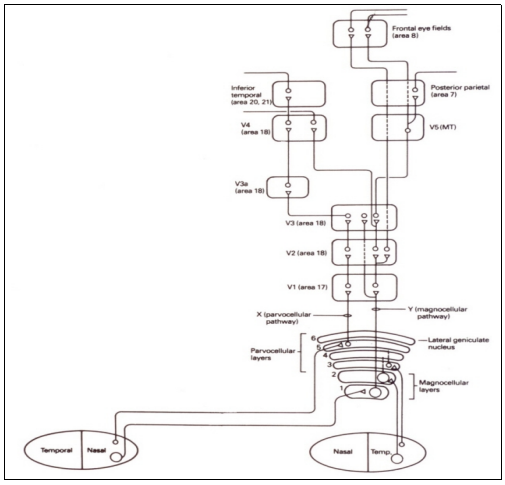
The Lateral Geniculate Nucleus (LGN)

In this area of the brain, cells are segregated according to eye of origin, a property known as ocularity. From the LGN, all cells project to the visual cortex. In monkeys, they all project to area V1 of the visual cortex, in layer 4c. Among various physiological properties, the cells in LGN display a center-surround firing-response pattern.

Areas V1 to V5 the Visual Cortex



Cells in V1 are retinotopically mapped. That is to say, neighboring cells have neighboring visual receptive fields. Area 4b of V1 contains cells with excitatory and inhibitory regions of their receptive fields that are divided by straight lines. It is thought that area V1 is similar in function to tightly tiled spatiotemporal filters. The current consensus among scientists is that V1 encodes for local contrast, rather than intensity per se.



In area V2 (prestriate cortex), cells are also tuned to simple properties such as orientation, spatial frequency, and color. The responses of many V2 cells are modulated by more complex properties, such as the orientation of illusory contours and whether the stimulus is part of the figure or the ground. Cells in area V3 are responsible for processing global motion information contained in the visual field. Area V4, like area V1, is tuned for orientation, spatial frequency, and color, but unlike V1, it responds

to such properties of objects displaying intermediate visual complexity. Area V5 (or MT) plays a major role in the perception of motion, the integration of local motion signals into global percepts and the guidance of some eye movements.

Computer Vision

Computer vision is a discipline with the aim of recovering 3D information about visual scenes from sequences of temporally varying images. Closely related fields are: artificial intelligence, robotics, signal processing, pattern recognition, control theory, and neuroscience.

Within computer vision, sub-fields of research are: image feature detection, contour representation, feature-based image segmentation, range image analysis, shape reconstruction, stereo vision, motion analysis, color vision, active/passive vision, and real-time vision.

Intensity Images

The most frequently used type of images in computer vision are intensity (or brightness) images, either color or black and white (monochrome). Cameras are used to capture these images. In general, the characteristics of the image acquisition system are given by: the type of lens, focal length, field of view, and induced optical distortions. Such systems also have geometric parameters, such as: type of projection (perspective, orthographic, etc), relative sensor position, and so on. Photometric parameters also influence the image acquisition process. For instance, the type, intensity, and direction of incident illumination, along with the reflectance properties of surfaces being imaged partly determine the results of the imaging process.

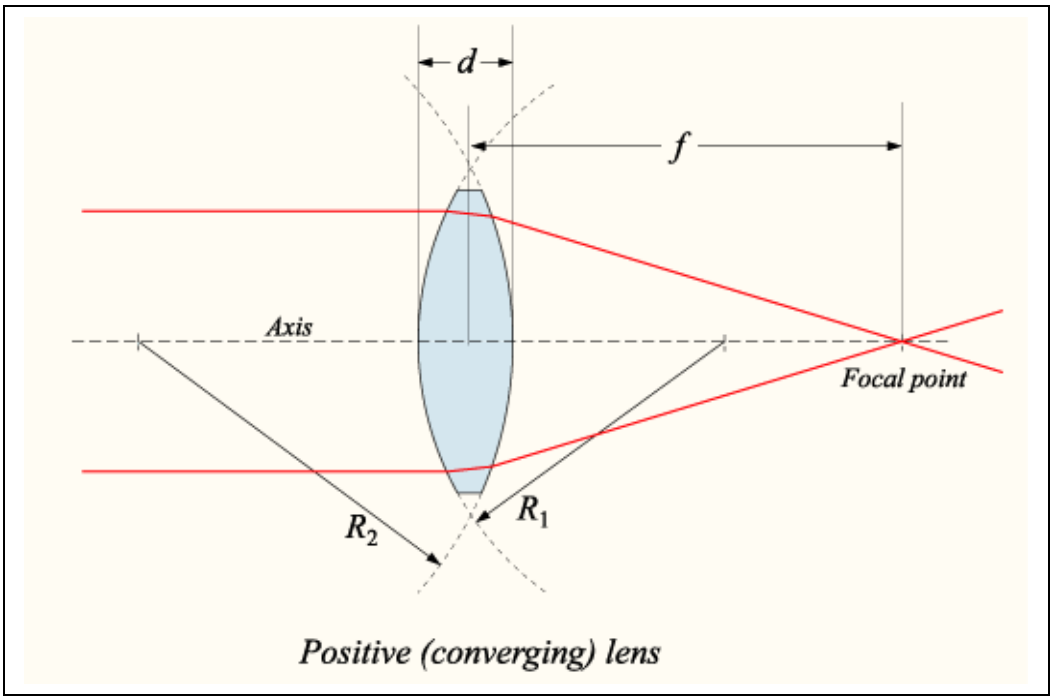
Image Focusing

Focusing is the process by which all light rays coming from a single scene point must converge onto a single image point. Then, the image of the point is said to be in focus. Techniques to focus all rays from a scene point to an image point include pinhole cameras, where the aperture is reduced to the size of a pinhole. In this case, only one ray from any given point enters the sensor, leading to sharp images of objects at different depths. In addition, lenses may be designed to make all rays from a scene point to intersect at a single image point. The simplest optical systems are made with thin lenses. The optical characteristics of thin lenses consist of an optical axis, passing through the center of the lens, and two points on the axis: the left and right foci f_l , and f_r , at a distance f from the center of the lens.

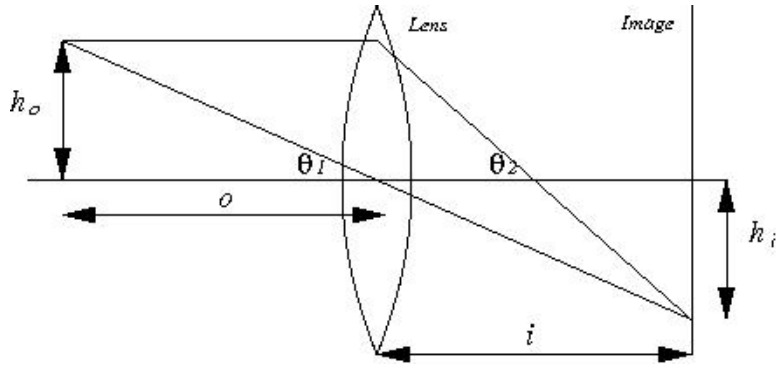
Thin Lenses

The optical properties of thin lenses can be summarized as follows:

- A ray parallel to the optical axis on one side of the lens goes to the focus on the other side
- A ray entering the lens from the focus on one side emerges parallel to the optical axis on the other side



The fundamental equation of a thin lens can be established geometrically with the help of the next figure, and similar triangles:



$$\theta_1 \Rightarrow \frac{h_i}{h_o} = \frac{i}{o}$$

$$\theta_2 \Rightarrow \frac{h_i}{h_o} = \frac{i-f}{f}$$

and hence, we have:

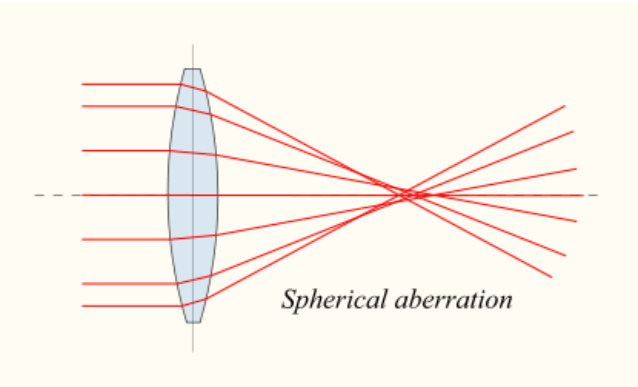
$$\frac{i-f}{f} = \frac{i}{o} \Rightarrow \frac{i}{f} - 1 = \frac{i}{o} \Rightarrow \frac{i}{if} - \frac{1}{i} = \frac{i}{io} \Rightarrow \frac{1}{f} = \frac{1}{o} + \frac{1}{i}$$

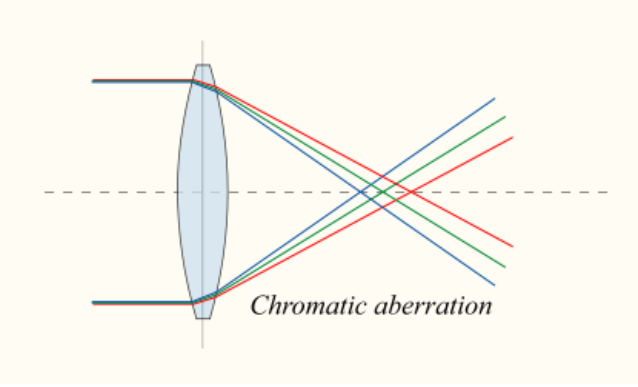
We clearly see from this relation that focusing is a function of depth.

Aberrations

Lenses cannot be perfectly manufactured and, as a result, aberrations may result. There are many types of aberrations, and they all can be alleviated either by corrective optical lenses or effective camera calibration.

Spherical aberrations occur when lenses have imperfect curvature. As a result, incident rays of light which should intersect the optical axis at the focal point do not. The effect can be observed when using inexpensive optics: edges that should be imaged as straight lines at the periphery of the image appear as somewhat curved.

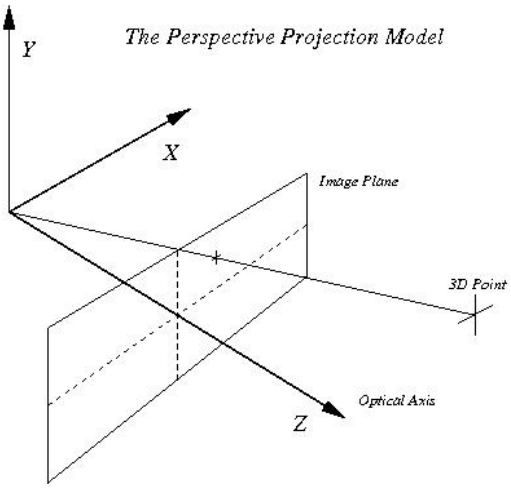




Conversely, chromatic aberrations are not due to manufacturing issues, but to the frequency spectrum of visible light. The diffraction angle of a ray of light depends on the shape and density of the lens, but also on the frequency (color) of the ray. Since different colors have different frequencies, we observe chromatic aberrations. These can be somewhat alleviated with various lens coating methods. There exist other types of aberrations, such as coma, astigmatism, and aperture diffraction which are described in the literature.

Geometric Image Formation

The process of image formation, in the geometric sense, is that of associating a 3D scene point with its 2D image point. The most common model used in computer vision is the perspective model, which as its name implies, uses perspective projection.



The fundamental equations of perspective projection are given by the following equations, which transform the coordinates of a 3D point into its image coordinates in 2D:

$$x = f\frac{X}{Z} \qquad y = f\frac{Y}{Z}$$

where (X, Y, Z) are the 3D coordinates and (x, y, f) are the image coordinates. The perspective equations are non-linear. In addition, depth and angles between lines are not preserved by this transformation.

The field of view is also an important parameter of any imaging device and is defined by both the focal length and the effective diameter of the lens (or aperture) of the device. The equation:

$$\tan\theta = \frac{d}{2f}$$

allows to compute the field of view in angular terms.

Photometric Image Formation

Incident light interacts with surfaces it encounters along its path. The most general model of this interaction is know as BRDF, which stands for Bidirectional Reflectance Distribution Function. Relative to a local coordinate system on a surface, BRDF is a function describing how much of each wavelength arriving with incident direction $\boldsymbol{v}_i$ is emitted in a reflected direction $\boldsymbol{v}_r$. The BRDF is reciprocal, in that the roles of $\boldsymbol{v}_i$ and $\boldsymbol{v}_r$ may be interchanged.

Most surfaces are isotropic in that there is no preferred direction imposed on light transport by surfaces. Consequently, the BRDF can be written as:

$$f_r(\boldsymbol{v}_i, \boldsymbol{v}_r, \boldsymbol{n}, \lambda)$$

where λ is the wavelength of incident light, and $\boldsymbol{n}$ is a surface normal. The amount of light exiting a surface point in a direction $\boldsymbol{v}_r$ under given lighting conditions is given by integrating the product of incoming light with the BRDF over all the incident light directions:

$$L_r(\boldsymbol{v}_r, \lambda) = \int L_i(\boldsymbol{v}_i, \lambda) f_r(\boldsymbol{v}_i, \boldsymbol{v}_r, \boldsymbol{n}, \lambda) \max(0, \cos\theta_i) d\boldsymbol{v}_i$$

If there is a finite number of point-like light sources, the integral is replaced with a summation:

$$L_r(\boldsymbol{v}_r, \lambda) = \sum_i L_i(\lambda) f_r(\boldsymbol{v}_i, \boldsymbol{v}_r, \boldsymbol{n}, \lambda) \max(0, \cos\theta_i)$$

BRDFs usually can be decomposed into their diffuse and specular components.

Diffuse Component

This component, also known as Lambertian reflection, scatters light uniformly in all directions. Consequently, the BRDF is constant and written as:

$$f_d(\boldsymbol{v}_i, \boldsymbol{v}_r, \boldsymbol{n}, \lambda) = f_d(\lambda)$$

The amount of reflected light is a function of the incident light direction and the surface normal. A shading equation thus can be written as:

$$L_d(\boldsymbol{v}_r, \lambda) = \sum_i L_i(\lambda) f_d(\lambda) \max(0, \boldsymbol{v}_i \cdot \boldsymbol{n})$$

A monochromatic approximation of the diffuse component of light, often used in computer graphics, can be written as

$$I_d = I_s \rho_d (\boldsymbol{v_i} \cdot \boldsymbol{n})$$

where I_s is the intensity of the light source, ρ_d is the diffuse reflection coefficient of the surface, $\boldsymbol{v}_i$ is a unit normal vector in the direction of the incident light, and $\boldsymbol{n}$ is a unit surface normal vector. For practical reasons, we write

$$I_d = I_s \rho_d \, max(0, \boldsymbol{v_i} \cdot \boldsymbol{n})$$

Specular Component

Specular reflection is the BRDF component that describes the behavior of incident light reflected by mirror-like surfaces. Unlike diffuse reflection, specular reflection is directional, and the amount of light in the viewing direction depends on the orientation of the surface. Given the direction of incident light, that of reflected specular light is obtained as:

$$\boldsymbol{s}_i = (2\boldsymbol{n}\boldsymbol{n}^T - I)\boldsymbol{v}_i$$

where I is the 3×3 identity matrix. The angle between the viewing direction $\boldsymbol{v}_r$ and the specular direction $\boldsymbol{s}_i$ is given by $\theta_s=\arccos(\boldsymbol{v}_r\cdot\boldsymbol{s}_i)$. Of course not all surfaces are purely specular . Hence, specularity is often modeled with a fallout function centered around the direction of the specular reflection, as in Phong's shading model:

$$f_s(\theta_s,\lambda)=k_s(\lambda)\cos^{k_e}\theta_s$$

where $k_s(\lambda)$ is the color of the specular surface and k_e is the exponent controlling the specular fallout around $\boldsymbol{s}_i$.

A monochromatic approximation of the specular component is given by

$$I_p=I_s\rho_p max(0,\boldsymbol{v_r}\cdot\boldsymbol{s_i})^{k_e}$$

where ρ_p is the specular reflection coefficient of the surface.

Ambient Component

Ambient illumination accounts for the fact that objects are also illuminated by general diffuse illumination as a result of inter-reflection effects. Ambient illumination does not depend on any surface orientation. It is a combination of the colors of the ambient illumination $L_a(\lambda)$ and that of the object $k_a(\lambda)$:

$$f_a(\lambda)=k_a(\lambda)L_a(\lambda)$$

A monochromatic approximation is given by

$$I_a=I_s\rho_a$$

where ρ_a is the coefficient of the ambient light.

Complete Shading Model

Adding the diffuse, specular, and ambient components together yields Phong's shading model, widely used in computer graphics and image rendering:

$$L_r(\boldsymbol{v}_r,\lambda)=k_a(\lambda)L_a(\lambda)+k_d(\lambda)\sum_i L_i(\lambda)\max(0,\boldsymbol{v}_i\cdot\boldsymbol{n})+k_s(\lambda)\sum_i L_i(\lambda)(\boldsymbol{v}_r\cdot\boldsymbol{s}_i)^{k_e}$$

Geometric Primitives

Points, lines and planes are useful geometric representations and the use of homogeneous coordinates simplify many tasks, such as perspective projections, for instance.

2D and 3D Points

2D points are represented as $\boldsymbol{x}=(x,y)^T$, or alternatively as

$$\boldsymbol{x}=\begin{bmatrix}x\\y\end{bmatrix}$$

Homogeneous coordinates are often used to represent 2D points:

$$\tilde{\boldsymbol{x}}=(\tilde{x},\tilde{y},\tilde{\omega})^T=\tilde{\omega}\left(\frac{\tilde{x}}{\tilde{\omega}},\frac{\tilde{y}}{\tilde{\omega}},1\right)^T=\tilde{\omega}\,\bar{\boldsymbol{x}}$$

where $\bar{\boldsymbol{x}}$ is called the augmented vector. Homogeneous points with $\tilde{\omega}=0$ are points at infinity and do not have an inhomogeneous representation.

2D Lines

Equation $\tilde{l}=(a,b,c)$ is a 2D line in homogeneous coordinates. The equation of the line is given by $\bar{x}\cdot\tilde{l}=ax+by+c=0$. We can normalize the line equation so that $l=(n_x,n_y,d)$ where $\boldsymbol{n}^T=(n_x,n_y)$ is a normal vector perpendicular to the line and d is the distance of the line from the origin ($\tilde{l}=(0,0,1)$ is a line at infinity and cannot be normalized).

Alternatively, $\boldsymbol{n}$ can be expressed as a function of a rotation angle:

$$\boldsymbol{n}=\begin{bmatrix}\cos\theta\\\sin\theta\end{bmatrix}$$

with (θ,d) known as polar coordinates.

With the use of homogeneous coordinates, we can compute the intersection of two lines as $\tilde{\boldsymbol{x}}=\tilde{l}_1\times\tilde{l}_2$. Similarly, the line joining two points is given by $\tilde{l}=\tilde{\boldsymbol{x}}_1\times\tilde{\boldsymbol{x}}_2$.

3D points $\boldsymbol{x}^T=(x,y,z)$ can be written in homogeneous coordinates as $\tilde{\boldsymbol{x}}^T=(\tilde{x},\tilde{y},\tilde{z},\tilde{\omega})$ with $\tilde{\boldsymbol{x}}=\tilde{\omega}\,\bar{\boldsymbol{x}}$.

3D Planes

3D planes in homogeneous coordinates can be represented by $\tilde{\boldsymbol{m}}^T=(a,b,c,d)$ with corresponding plane equation $\bar{\boldsymbol{x}}\cdot\tilde{\boldsymbol{m}}=ax+by+cz+d=0$. It can be normalized to yield $\boldsymbol{m}^T=(n_x,n_y,n_z,d)$, where $\boldsymbol{n}$ is a unit vector normal to the plane and d its distance to the origin. $\boldsymbol{n}$ Can be expressed as a function of two angles such that $\boldsymbol{n}^T=(\cos\theta\cos\phi,\sin\theta\cos\phi,\sin\theta)$ in spherical coordinates.

3D Lines

Lines in 3D are less elegant than 2D lines or 3D planes. A parametric equation

can be formed between two 3D points to obtain $\boldsymbol{r}=(1-t)\boldsymbol{p}+t\boldsymbol{q}$. However this equation has six degrees of freedom whereas a line truly only has four. In general, a line is the result of the intersection of two planes.

2D Transforms

2D transforms are those that occur in the 2D plane in image coordinates and are the simplest.

- In homogeneous coordinates, translation is written as

$$\boldsymbol{x}' = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{t}\end{bmatrix}\boldsymbol{x}$$

  where $\boldsymbol{I}$ is the 2 by 2 identity matrix and $\vec{\mathbf{0}}$ is a 1 by 2 zero vector.

- A rigid body transformation is composed of a rotation and a translation. It is also known as the 2D Euclidean transformation:

$$\boldsymbol{x}' = \begin{bmatrix}\boldsymbol{R} & \boldsymbol{t}\end{bmatrix}\boldsymbol{x}$$

  where $\boldsymbol{R} = \begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix}$ is an orthonormal rotation matrix.

- The similarity transform (or scaled rotation) is expressed as

$$\boldsymbol{x} = \begin{bmatrix}s\boldsymbol{R} & \boldsymbol{t}\end{bmatrix}\boldsymbol{x}$$

  where s is a scaling factor. This transform preserves angles between lines.

- The affine transformation is written as

$$\boldsymbol{x}' = \boldsymbol{A}\boldsymbol{x} = \begin{bmatrix}a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12}\end{bmatrix}\boldsymbol{x}$$

  where matrix $\boldsymbol{A}$ is arbitrary. This transform preserves parallelism between lines.

- The projective transformation operates on homogeneous coordinates exclusively and is written as

$$\boldsymbol{x}' = \begin{bmatrix}h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22}\end{bmatrix}\boldsymbol{x} = \boldsymbol{H}\boldsymbol{x}$$

  where $\boldsymbol{H}$ is a 3 by 3 arbitrary matrix. Note that this matrix is homogeneous and therefore only defined up to as scale. Hence two

homogeneous matrices that differ in scale only are equivalent. The resulting homogeneous result $\boldsymbol{x}'$ must be normalized in order to obtain an inhomogeneous result, in the following way:

$$x' = \frac{h_{00}x + h_{01}y + h_{02}}{h_{20}x + h_{21}y + h_{22}} \qquad y' = \frac{h_{10}x + h_{11}y + h_{12}}{h_{20}x + h_{21}y + h_{22}}$$

2D Transform Hierarchy

The following table depicts the hierarchy of 2D transforms. Each transform also preserves the properties listed in the rows below it. For instance, the similarity transform preserves not only angles but also parallelism and straight lines.

2D Transform	Matrix	Degrees of Freedom	Preserves
Translation	$\left[\boldsymbol{I} \mid \boldsymbol{t}\right]_{2\times 3}$	2	Orientation
Rigid Body	$\left[\boldsymbol{R} \mid \boldsymbol{t}\right]_{2\times 3}$	3	Lengths
Similarity	$\left[s\boldsymbol{R} \mid \boldsymbol{t}\right]_{2\times 3}$	4	Angles
Affine	$\left[\boldsymbol{A}\right]_{2\times 3}$	6	Parallelism
Projective	$\left[\boldsymbol{H}\right]_{3\times 3}$	8	Straight Lines

3D Transforms

- Translation: $\boldsymbol{x}' = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{t} \\ \vec{\boldsymbol{0}}^T & 1 \end{bmatrix} \boldsymbol{x}$ where $\boldsymbol{I}$ is the 3 by 3 identity matrix
- 3D Rigid Body Motion: $\boldsymbol{x}' = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \vec{\boldsymbol{0}}^T & 1 \end{bmatrix} \boldsymbol{x}$ where $\boldsymbol{R}$ is a 3 by 3 orthonormal rotation matrix
- 3D Similarity: $\boldsymbol{x}' = \begin{bmatrix} s\boldsymbol{R} & \boldsymbol{t} \\ \vec{\boldsymbol{0}}^T & 1 \end{bmatrix} \boldsymbol{x}$ where s is a scalar
- 3D Rotation around normal vector $\hat{\boldsymbol{n}}$: $\boldsymbol{R}(\hat{\boldsymbol{n}}, \theta) = \boldsymbol{I} + \sin\theta[\hat{\boldsymbol{n}}]_\times + (1-\cos\theta)[\hat{\boldsymbol{n}}]_\times^2$ where

$$[\hat{\boldsymbol{n}}]_\times = \begin{bmatrix} 0 & -\hat{n}_z & \hat{n}_y \\ \hat{n}_z & 0 & -\hat{n}_x \\ -\hat{n}_y & \hat{n}_x & 0 \end{bmatrix}$$

- 3D Affine: $\boldsymbol{x}' = \boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A} = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \end{bmatrix}$ is an arbitrary 3 by 4 matrix
- The 3D Projective transform (perspective transform, homography, collineation), operates on homogeneous coordinates: $\boldsymbol{x}' = \boldsymbol{H}\boldsymbol{x}$, where $\boldsymbol{H}$ is an arbitrary, homogeneous 4 by 4 matrix.

3D Transform Hierarchy

3D Transform	Matrix	Degrees of Freedom	Preserves
Translation	$\begin{bmatrix} \boldsymbol{I} \mid \boldsymbol{t} \end{bmatrix}_{3\times4}$	3	Orientation
Rigid Body	$\begin{bmatrix} \boldsymbol{R} \mid \boldsymbol{t} \end{bmatrix}_{3\times4}$	6	Lengths
Similarity	$\begin{bmatrix} s\boldsymbol{R} \mid \boldsymbol{t} \end{bmatrix}_{3\times4}$	7	Angles
Affine	$\begin{bmatrix} \boldsymbol{A} \end{bmatrix}_{3\times4}$	12	Parallelism
Projective	$\begin{bmatrix} \boldsymbol{H} \end{bmatrix}_{4\times4}$	15	Straight Lines