1 Overview of Computer Vision and its Applications

Syllabus

Image Formation and Representation : Imaging geometry, radiometry, digitization, cameras and Projections, rigid and affine transformation.

Contents

1.1 Introduction of Computer Vision and its Applications

Computer Vision

Computer vision is an interdisciplinary field that studies how computers can be programmed to perceive digital images or movies at a high level. It aims to automate things that the human visual system can perform from an engineering standpoint. The automatic extraction, analysis, and understanding of useful information from a single image or a sequence of images is the subject of computer vision. Computer vision is a scientific area that studies the theory behind artificial systems that extract information from images. Video sequences, various camera perspectives, or multi-dimensional data from a medical scanner are all examples of picture data. Computer vision seeks to apply its theories and models for the construction of computer vision systems.

- **Geometry** : Concerned with the relationship between points in the three-dimensional world and their images.
- **Radiometry** : Concerned with the relationship between the amount of light radiating from a surface and the amount incident at its image.
- **Photometry** : Concerned with ways of measuring the intensity of light.
- **Digitization** : Concerned with ways of converting continuous signals (in both space and time) to digital approximations.

Applications of Computer Vision

- **Optical character recognition (OCR)** : Reading handwritten postal codes on letters (and automatic number plate recognition (ANPR).
- **Machine inspection** : Rapid parts inspection for quality assurance using stereo vision with specialized illumination to measure tolerances on aircraft wings or auto body parts or looking for defects in steel castings using X-ray vision.
- **Retail** : object recognition for automated checkout lanes.
- **3D model building** : Fully automated construction of 3D models from aerial photographs used in systems such as Bing Maps.
- **Medical imaging** : Registering pre-operative and intra-operative imagery or performing long-term studies of people's brain morphology as they age.
- **Automotive safety** : Detecting unexpected obstacles such as pedestrians on the street, under conditions where active vision techniques such as radar or lidar do not work well.

- **Match move** : Merging computer-generated imagery (CGI) with live action footage by tracking feature points in the source video to estimate the 3D camera motion and shape of the environment.
- **Motion capture (mocap)** : Using retro-reflective markers viewed from multiple cameras or other vision-based techniques to capture actors for computer animation.
- **Surveillance** : Monitoring for intruders, analyzing highway traffic and monitoring pools for drowning victims.
- **Fingerprint recognition and biometrics** : For automatic access authentication as well as forensic applications. the use of precise matting to insert new elements between foreground and background elements.

1.2 Image Formation and Representation

Image formation

The study of image formation encompasses of the radiometric and geometric processes by which 2D images of 3D objects are formed. In the case of the digital images, the image formation process also includes analog to digital conversion and sampling are studied.

1.2.1 Imaging Geometry

Geometric primitives and transformations

The basic 2D and 3D primitives are explained namely points, lines, and planes are the geometric primitives.

Geometric primitives

Geometric primitives form the basic building blocks used to describe 3D shapes.

2D points

- 2D points (pixel coordinates in an image) is denoted using a pair of values,

$$x = (x, y) \in R^2,$$

or

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

- 2D points is represented using homogeneous coordinates $\tilde{x} = (\tilde{x}, \tilde{y}, \tilde{w}) \in P^2$, where vectors that differ only by scale are considered to be equivalent. $P^2 = R^3 - (0, 0, 0)$ is called 2D projective space.
- A homogeneous vector x can be converted back into an inhomogeneous vector $\tilde{x}$ by dividing through by the last element $\tilde{w}$, i.e.,

$$\tilde{x} = (\tilde{x}, \tilde{y}, \tilde{w}) = \tilde{w}(x, y, 1) = \tilde{w}\bar{x}$$

where, $\bar{x} = (x, y, 1)$ is the augmented vector.

- Homogeneous points whose last element is $\tilde{w} = 0$ are called points at infinity or ideal points and do not have an equivalent inhomogeneous representation.

2D lines

- 2D lines can be represented using homogeneous coordinates $\tilde{l} = (a, b, c)$. The corresponding line equation is,

$$\tilde{x} \cdot \tilde{l} = ax + by + c = 0$$

- Normalization can be done to the line equation vector so that $\tilde{l} = (\hat{n}_x, \hat{n}_y, d) = (\hat{n}, d)$ with $\| \hat{n} \| = 1$.

where, $\hat{n}$ is the normal vector perpendicular to the line and d is its distance to the origin.

- (The one exception to this normalization is the line at infinity $l = (0, 0, 1)$, which includes all (ideal) points at infinity.)
- Express $\hat{n}$ as a function of rotation angle θ, $\hat{n} = (\hat{n}_x, \hat{n}_y) = (\cos\theta, \sin\theta)$. This representation is used in the Hough transform for line-finding algorithm. The combination (θ, d) is also known as polar coordinates.
- Using homogeneous coordinates, compute the intersection of the two lines as,

$$\tilde{x} = \tilde{l}_1 \times \tilde{l}_2$$

where, x is the cross product operator.

- The line joining, two points can be written as,

$$\tilde{l} = \tilde{x}_1 \times \tilde{x}_2$$

- A least squares method can be used to fit an intersection point to multiple lines or conversely, a line to multiple points.

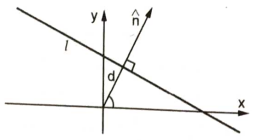


Fig 1.2.1 2D line equation

2D conics : With simple homogeneous polynomial equations is used to express algebraic curves. For example, the conic sections (so called because they arise as the intersection of a plane and a 3D cone) can be written using a quadratic equation.

$$\bar{x}^T Q \bar{x} = 0$$

3D points : The Point coordinates in 3D can be written using inhomogeneous coordinates $x = (x, y, z) \in R^3$ or homogeneous coordinates $\tilde{x} = (\tilde{x}, \tilde{y}, \tilde{z}, \tilde{w}) \in P^3$. To denote a 3D point using the augmented vector $\bar{x} = (x, y, z, 1)$ with $\tilde{x} = \tilde{w}\bar{x}$.

3D planes

- 3D planes can be represented as homogeneous coordinates $\tilde{m} = (a, b, c, d)$ with a corresponding plane equation, $\tilde{x} \cdot \tilde{m} = ax + by + cz + d = 0$
- Normalize the plane equation $m = (\hat{n}_x, \hat{n}_y, \hat{n}_z, d) = (\hat{n}, d)$ as with $\| \hat{n} \| = 1$. In this case, $\hat{n}$ is the normal vector perpendicular to the plane and d is its distance to the origin. (Refer Fig. 1.2.2). As with the case of 2D lines, the plane at infinity $\tilde{m} = (0, 0, 0, 1)$, which contains all the points at infinity, cannot be normalized (i.e., it does not have a unique normal or a finite distance).

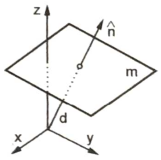


Fig. 1.2.2 3D plane equation, expressed in terms of the normal $\hat{n}$ and distance to the origin d

- Express $\hat{n}$ as a function of two angles (ϑ, φ),

$$\hat{n} = (\cos\vartheta \cos\varphi, \sin\vartheta \cos\varphi, \sin\varphi)$$

- i.e., using spherical coordinates but these are less commonly used than polar coordinates since, they do not uniformly sample the space of possible normal vectors.
- The use of spherical coordinates is less common than the use of polar coordinates since it does not uniformly sample the space of possible normal vectors.

3D lines

- Lines in 3D are less elegant than either lines in 2D or planes in 3D. One possible representation is to use two points on the line (p, q). Any other point on the line can be expressed as a linear combination of these two points.

$$r = (1 - \lambda)p + \lambda q$$

- If the restrict of λ, $0 \leq \lambda \leq 1$ the line segment joining p and q. If we use homogeneous coordinates, write the line as,

$$\tilde{r} = \mu\tilde{p} + \lambda\tilde{q}$$

- A special case of this is when the second point is at infinity, i.e., $\tilde{q} = (\hat{d}_x, \hat{d}_y, \hat{d}_z, 0) = (\hat{d}, 0)$, $\hat{d}$ is the direction of the line. Then re-write the inhomogeneous 3D line equation as,

$$V_r = p + \lambda\hat{d}$$

- A disadvantage of the endpoint representation for 3D lines is that it has too many degrees of freedom, i.e., six (three for each endpoint) instead of the 4 degree that a 3D line truly has.
- By fixing the two points on the line to lie in specific planes, we obtain a representation with four degrees of freedom.
- For example, if we are representing nearly vertical lines, then z = 0 and z = 1 form 2 suitable planes, i.e., the (x, y) coordinates in both planes provide the 4 coordinates describing the line.
- This kind of 2-plane parameterization is used in the light field and Lumigraph image-based rendering systems described to represent the collection of rays seen by a camera as it moves in front of an object.
- The 2-endpoint representation is also useful for representing line segments, even when their exact endpoints cannot be seen.
- In order to represent all possible lines without bias towards any particular orientation, we can use Plucker coordinates. These are the six non-zero entries in a 4 by 4 skew symmetric matrix

$$L = \tilde{p}\tilde{q}^T - \tilde{q}\tilde{p}^T$$

- Most applications do not require a minimal representation. 3D lines can be modelled by estimating their endpoints and a point within the visible portion of the line.

- Most applications do not require minimal representation. A good model of 3D line can be obtained by estimating their position and a point within the visible portion of the line or by using the two endpoints.

3D quadrics : Representing a conic section in 3D as an equilateral triangle as shown

$$\bar{x}^T Q \bar{x} = 0$$

2D transformations

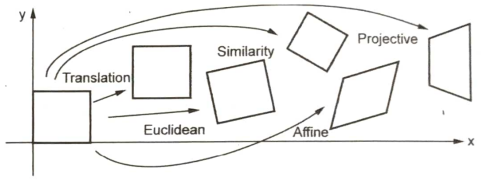


Fig 1.2.3 Basic set of 2D planar transformations

Translation : 2D translations can be written as,

$$x' = x + t \text{ or}$$

$$x' = [I \; t]\,\bar{x}$$

where, I is the (2 × 2) identity matrix or

$$\bar{x}' = \begin{bmatrix} I & t \\ 0^T & 1 \end{bmatrix} \bar{x}$$

where 0 is the zero vector.

- Using a 2 × 3 matrix results in a more compact notation, whereas using a full-rank 3 × 3 matrix (which can be obtained from the 2 × 3 matrix by appending $A[0^T \; 1]$ row) makes it possible to chain transformations using matrix multiplication.

Note : In the equation where an augmented vector such as $\bar{x}$ appears on both sides, it can always be replaced with a full homogeneous vector $\tilde{x}$.

Rotation + translation : This transformation is known as 2D rigid body motion or the 2D Euclidean transformation (since Euclidean distances are preserved). representation be written as,

$$x' = Rx + t$$

or

$$x' = [R \; t]\,\bar{x}$$

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

is an orthonormal rotation matrix with $RR^T = I$ and $|R| = 1$.

$$x' = R(x - c) = Rx - Rc$$

where, c is the center of rotation (often the camera center).

Compactly parameterizing a 3D rotation is a non-trivial task.

Scaled rotation : The 3D similarity transform can be expressed as

$$x' = sRt + t$$

where s is an arbitrary scale factor.

It can also be written as,

$$x' = [sRt]\,\bar{x}$$

$$x' = [sR\ t]\,\bar{x} = \begin{bmatrix} a & -b & t_x \\ b & a & t_y \end{bmatrix} \bar{x}$$

where, we no longer require that $a^2 + b^2 = 1$

The similarity transform preserves angles between lines. This transformation preserves angles between lines and planes.

1.2.1 Hierarchy of 2D Transformations

This represents a set of (potentially restricted) 3 * 3 transformations which operate on 2D homogeneous coordinates. The transformations listed below create a nested set of groups, member they are closed under composition and have an inverse that belongs to the same group.

Co-vectors : The below transformations can be used to transform points in a 2D plane. The homogeneous equation $\tilde{l} \cdot \tilde{x} = 0$. If transform $x' = \tilde{H}x$

$$\tilde{l}' \cdot \tilde{x}' = \tilde{l}^T \tilde{H}\, \tilde{x} = (\tilde{H}^T \tilde{l}^T)\, \tilde{x} = \tilde{l} \cdot \tilde{x} = 0$$

i.e., $\tilde{l}' = \tilde{H}^{-T} \tilde{l}$. Thus, the action of a projective transformation on a co-vector such as a 2D line or 3D normal can be represented by the transposed inverse of the matrix, which is equivalent to the adjoint of $\tilde{H}$, projective transformation matrices are homogeneous.

Transformation	Matrix	# DoF	Preserves	Icon
Translation	$[I \mid t]_{2\times3}$	2	Orientation	
Rigid (Euclidean)	$[R \mid t]_{2\times3}$	3	Lengths	

Similarity	$[sR \mid t]_{2\times3}$	4	Angles	
Affine	$[A]_{2\times3}$	6	Parallelism	
Projective	$[\tilde{H}]_{3\times3}$	8	Straight lines	

Table 1.2.1 Hierarchy of 2D coordinate transformations

Each transformation also preserves the properties listed in the rows below it, i.e., similarity preserves not only angles but also parallelism and straight lines. The 2×3 matrices are extended with a third $[0^T\ 1]$ row to form a full 3×3 matrix for homogeneous coordinate transformations.

Stretch/squash : This transformation changes the aspect ratio of an image,

$$x' = s_x x + t_x$$

$$y' = s_y y + t_y$$

and is a restricted form of an affine transformation. Unfortunately, it does not nest cleanly with the groups listed in Table 1.2.1.

Planar surface flow : This eight-parameter transformation occurs when a planar surface undergoes a small 3D motion.

$$x' = a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy$$

$$y' = a_3 + a_4 x + a_5 y + a_7 x^2 + a_6 xy,$$

arises when a planar surface undergoes a small 3D motion. It can thus be thought of as a small motion approximation to a full homography. Its main attraction is that it is linear in the motion parameters, a_k, which are often the quantity being estimated.

Transformation	Matrix	# DoF	Preserves	Icon
Translation	$[I \mid t]_{3\times4}$	3	Orientation	

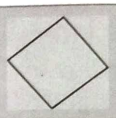
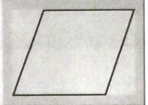
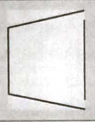
Rigid (Euclidean)	$[R\|t]_{3\times4}$	6	Lengths	
Similarity	$[sR\|t]_{3\times4}$	7	Angles	
Affine	$[A]_{3\times4}$	12	Parallelism	
Projective	$[\tilde{H}]_{4\times4}$	15	Straight lines	

Table 1.2.2 Hierarchy of 3D coordinate transformations

Each transformation also preserves the properties listed in the rows below it, i.e., similarities preserves not only angles but also parallelism and straight lines. The 3 × 4 matrices are extended with a fourth $[0^T\ 1]$ row to form a full 4 × 4 matrix for the homogeneous coordinate transformations. The mnemonic icons are drawn in the 2D but are meant to suggest transformations occurring in a full 3D cube.

Bilinear interpolate : Bilinear interpolation is accomplished by first performing linear interpolation in one direction and then repeating the process in the opposite direction. Despite the fact that each step is linear in terms of sampled values and position, the interpolation as a whole is quadratic in terms of sample location. This eight-parameter transform can be used to interpolate the movement of the four corners of a square.

$$x' = a_0 + a_1x + a_2y + a_6xy$$

$$y' = a_3 + a_4x + a_5y + a_7xy,$$

While the deformation is linear in the motion parameters, it does not generally preserve straight lines.

3D transformations

The set of three-dimensional coordinate transformations is very similar to that available for 2D transformations and is summarized in Table 1.2.2.

Translation : 3D translations can be written as,

$$x' = x + t$$

or $$x' = [I\ \ t]\,\bar{x}$$

where, I is the (3 × 3) identity matrixs and 0 is the zero vectors.

Rotation + Translation : This transformation is also known as 2D rigid body motion or the 2D Euclidean transformation (since Euclidean distances are preserved). It can be written as,

$$x' = Rx + t$$

or $$x' = [R\ \ t]\,\bar{x}$$

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

is an orthonormal rotation matrix with $RR^T = I$ and $|R| = 1$.

$$x' = R(x - c) = Rx - Rc$$

where c is the center of rotation (often the camera center).

- Compactly parameterizing a 3D rotation is a non-trivial task.

Scaled rotation : The 3D similarity transform can be expressed as,

$$x' = sRx + t$$

where, s is an arbitrary scale factor.

It can also be written as,

$$x' = [sRt]\,\bar{x}$$

$$x' = [sR\ t]\ \ \bar{x} = \begin{bmatrix} a & -b & t_x \\ b & a & t_y \end{bmatrix}\bar{x}$$

where we no longer require that $a^2 + b^2 = 1$.

The similarity transform preserves angles between lines. This transformation preserves angles between lines and planes.

Affine : The affine transform is written as,

$$x' = A\bar{x},$$

where, A is an arbitrary 3×4 matrix,

$$x' = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \end{bmatrix}\bar{x}$$

Parallel lines and planes remain parallel under affine transformations.

Projective : This transformation, variously known as a 3D perspective transform, homography or collineation, operates on homogeneous coordinates,

$$\tilde{x}' = \tilde{H}\,\tilde{x}$$

where, $\tilde{H}$ is an arbitrary 4×4 homogeneous matrix.

As in 2D, the resulting homogeneous coordinate $\tilde{x}$ must be normalized in order to obtain an inhomogeneous result x. Perspective transformations preserve straight lines (i.e., they remain straight after the transformation).

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

and

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

3D rotations

The biggest difference between 2D and 3D coordinate transformations is that the parameter- ization of the 3D rotation matrix $\boldsymbol{R}$ is not as straightforward but several possibilities exist.

Euler angles

- A rotation matrix can be formed as the product of three rotations around three cardinal axes,e.g., *x*, *y*, and *z*.
- As the result depends on the order in which the transforms are applied. It is not always possible to move smoothly in the parameter space, i.e., sometimes one or more of the Euler angles change dramatically in response to a small change in rotation.

Note : Some applications, if the rotations are known to be a set of uni-axial transforms, they can always be represented using an explicit set of rigid transformations.

Axis/angle (exponential twist)

- A rotation can be represented by a rotation axis $\hat{n}$ and an angle θ, or equivalently by a 3D vector $\omega = \theta\,\hat{n}$. Fig. 1.2.4 shows how we can compute the equivalent rotation. First, we project the vector v onto the axis $\hat{n}$ to obtain which is the component of v that is not affected by the rotation. Next, we compute the perpendicular residual of v from,

$$v_{\parallel} = \hat{n}\,(\hat{n} \cdot v) = \hat{n}\,\hat{n}^{T} v,$$

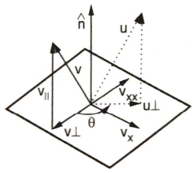


Fig 1.2.4 Rotation around an axis $\hat{n}$ by an angle θ

1.2.2 Radiometry

Radiometry is the field of measurement of any electromagnetic radiations, in particular light. Radiometry is the part of image formation concerned with the relation among the amounts of light energy emitted from light sources, reflected from surfaces, and registered by sensors.

Photometric image formation

In modeling the image formation process, we have described how 3D geometric features in the world are projected into 2D features in an image.

Lighting

- Images cannot exist without light. To produce an image, the scene must be illuminated with one or more light sources.
- Light sources can generally be divided into point and area light sources. A point light source originates at a single location in space (e.g., a small light bulb), potentially at infinity (e.g., the sun).
- In addition to its location, a point light source has an intensity and a color spectrum, i.e., a distribution over wavelengths $L(\lambda)$.
- The intensity of a light source falls off with the square of the distance between the source and the object being lit, because the same light is being spread over a larger (spherical) area.
- A light source may also have a directional falloff (dependence), but we ignore this in our simplified model.

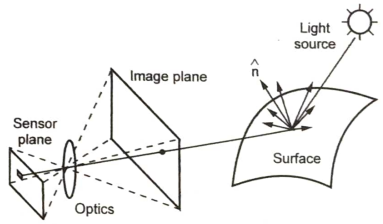


Fig 1.2.5 Photometric image formation

- Model of photometric image formation simplified. Light is emitted by one or more light sources and is then reflected from an object's surface. A portion of this light is directed towards the camera. This simplified model ignores multiple reflections, which often occur in real-world scenes.

World - Reality

Optics - Focus light from world on

Sensor - Sensor converts light to electrical energy

Signal - Representation of incident light as continuous electrical energy

Digitizer - Converts continuous signal to discrete signal

Digital Rep. - Final representation of reality in computer memory

When a fixed-intensity light source is placed at various positions on a frontal plane and viewed through an ideal lens, the apparent intensity of the light in the image is reduced by a factor of $\cos^4 \alpha$ when the source is situated at an angle to the optical axis. As illustrated in Fig. 1.2.6, a drop of $\cos^2 \alpha$ is created because the source location is at distance D from the lens, rather than the distance D cos α for a source on the same frontal plane but along the optical axis. Since brightness decays with the square of distance, this yields a first factor of $\cos^2 \alpha$. Because the cone of light rays from the off-axis source enters the lens at an angle rather than hitting it head-on as light from the on-axis source would, an additional cos factor is introduced. Finally, light from the off-axis source hits the image plane at an angle α, for an additional factor cos α. If the lens has a 90 degree field of view, this drop-off means that the edges of the image will be only one fourth as bright as the center : $\alpha = 90/2 = 45$ degrees, $\cos \alpha = \sqrt{2/2}$ and $\cos^4 \alpha = 1/4$.

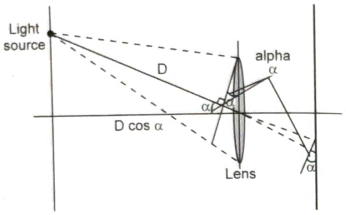


Fig. 1.2.6 Factors that lead to a $\cos^4 \alpha$ image brightness drop-off at an angle α away from the optical axis

Practical aspects : Good images with poor lenses. Real lenses can cause further variation in illumination, called vignetting, for other reasons. Using only narrow-angle lenses with fields of view less than 50 degrees or, even better, using only the core part of an enormous lens with a small sensor, is a frequent technique computer vision researchers have employed to avoid both geometric and radiometric difficulties. Radial distortion and radiometric drop-off become minimal as a result. The absence of peripheral vision, on the other hand, is a handicap for visual searching, navigation, and detection of objects moving towards the observer, all of which require a wide range of view. If intensity values are important in these instances, the $\cos^4 \alpha$ drop-off must be accounted for through calibration.

1.2.3 Image Digitization

A 2D scene can be represented by a 2D function f(x, y) of light intensity at the spatial location(x, y). However, the continuous scene must be digitized in order to be represented and processed digitally in a computer. The digitization process include quantizing the intensity function value as well as sampling the two spatial dimensions. As a result, the image's digital processing can be divided into intensity (gray level) operations that are applied to the pixel values.

- Sampling means measuring the value of an image at a finite number of points.
- Quantization is the representation of the measured value at the sampled point by an integer.

Spatial sampling

The digital image acquisition system must sample the continuous two-dimensional image space to create a raster, a 2D array of pixels (picture-elements) in rows and columns. The sampling theorem applies here as well as in the 1D example, with the exception that the sample is done in two spatial dimensions rather than one temporal dimension.

Fig 1.2.7 Sampling images

If you look closely at these photographs, you will notice a significant difference between high and low resolution photos. However, because their resolutions are diminished when you are far away from the images (or if you squint your eyes), they no longer appear to be significantly different.

Quantization

The continuous range of light intensity $0 \leq x \leq G$ received by the digital image acquisition system need be quantized to gray levels (e.g., $L = 2^8 = 256$). The numbers of gray levels of the eight images are respectively 256, 128, 64, 32, 16, 8, 4, and 2, respectively.

Fig 1.2.8 Quantization images

Pixels are the visual bits that make up digital images. Pixels are usually arranged in an ordered rectangular array. The dimensions of this pixel array define the size of an image. The number of columns in the array determines the picture width, while the number of rows determines the picture height. As a result, the pixel array is a M columns x N rows matrix. We use the coordinates x and y to refer to a specific pixel within the image matrix. Image matrices use a coordinate system in which x increases from left to right and y increases from top to bottom.

1.2.4 Cameras : Pinhole Cameras

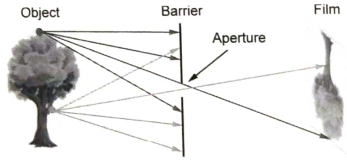


Fig 1.2.9 A simple working camera model : The pinhole camera model

Let's create a simple camera system that can capture an image of an object or scene in three dimensions. A barrier with a small aperture can be placed between the 3D object and a photographic film or sensor to create this camera system. Each point on the 3D object sends several rays of light outwards, as seen in Fig. 1.2.9. Without a barrier, light rays emanating from every point on the 3D object will affect every point on the film. Only one (or a few) of these rays of light travel through the aperture and strike the film due to the barrier. As a result, a one-to-one mapping between locations on the 3D object and the film can be established. As a result of this mapping, the film is exposed to a "picture" of the 3D object. The pinhole camera model is a simple camera model.

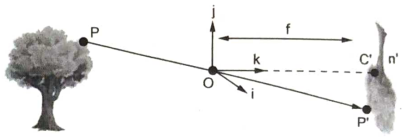


Fig. 1.2.10 A formal construction of the pinhole camera model

Fig. 1.2.10 depicts a more formal build of the pinhole camera. The film is frequently referred to as the image or retinal plane in this arrangement. The pinhole O, often known as the camera's center, is the aperture. The focal length f is the distance between the picture plane and the pinhole O. The retinal plane is sometimes positioned between O and the 3D object at a distance f from O. It is known as the virtual picture or virtual retinal plane in this scenario. Up to a scale (similarity) transformation, the projection of the object in the image plane and the image of the object in the virtual image plane are identical.

Let $P = x\ y\ z^T$ be a point on some 3D object visible to the pinhole camera and P will be mapped or projected onto the image plane Π^I which will result in point1 $P^I = x^I\ y^{I\,T}$. Similarly, the pinhole itself can be projected onto the image plane is given a new point C^I. We can define a coordinate system i j k centered at the pinhole O such that the axis k is perpendicular to the image plane and points toward it. The camera reference system, or camera coordinate system, is another name for this coordinate system. The optical axis of the camera system is the line formed by C^I and O.

Recall that point P^I is derived from the projection of the 3D point P on the image plane Π^I, which therefore derive the relationship between 3D point P and image plane point P^I, We can understand how the 3D world imprints itself upon the image taken by a pinhole camera. Notice that triangle $P^I C^I O$ is similar to the triangle formed by P, O and (0, 0, z). Therefore, using the law of similar triangles we find that :

$$P^I = x^I\ y^{I\,T} = f_z^x\ f_z^{y\,T}$$

Notice that one large assumption we make in this pinhole model is that the aperture is a single point. In most real world scenarios, however, we cannot assume the aperture can be infinitely small. The effects of aperture size on the image. As the aperture size decreases, the image gets sharper, but darker.

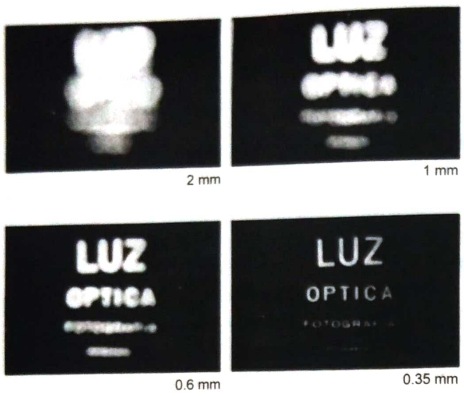


Fig. 1.2.11 The effects of aperture size on the image. As the aperture size decreases, the image gets sharper, but darker

Cameras and lenses

The above contradiction between sharpness and brightness is eased in current cameras by adding lenses, which are devices that can focus or scatter light. When we replace the pinhole with a lens that is suitably placed and sized, we get the following property: all light rays emitted by some point P are refracted by the lens and converge to a single point P'

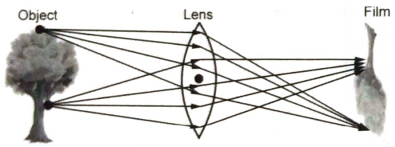


Fig 1.2.12 A setup of a simple lens model

Notice how the rays of the top point on the tree converge nicely on the film. However, a point at a different distance away from the lens results in rays not converging perfectly on the film.

As a result, the problem of the majority of light rays being blocked due to a tiny aperture is no longer an issue (Refer Fig. 1.2.11). Always keep in mind that this property does not apply to all 3D points, but only to a single point P. Take a different point Q that is either closer or further away from the image plane than P. The image's accompanying projection will be blurry or out of focus. As a result, lenses have a "focus distance" at which things are "in focus." This attribute is also linked to the concept of depth of field in photography and computer graphics, which refers to the effective range at which cameras can produce clear photographs.

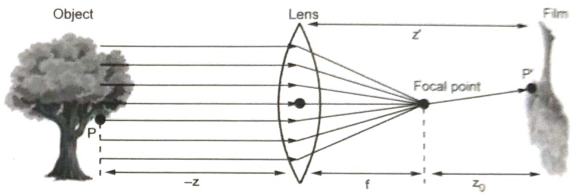


Fig 1.2.13 Lenses focus light rays parallel to the optical axis into the fo-cal point

Furthermore, this setup illustrates the paraxial refraction model, which helps us find the relationship between points in the image plane and the 3D world in cameras with lenses.

Another notable feature of camera lenses is that they focus all light rays moving parallel to the optical axis to a single point known as the focal point (Refer Fig. 1.2.13). The focal length f is the measurement of the distance between the focal point and the lens's center. Additionally, light beams flowing through the lens's center do not deviate. As a result, we can design something similar to the pinhole model, which connects a point P in 3D space with its equivalent point P' in the image plane.

$$P' = \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} z' \frac{x}{z} \\ z' \frac{y}{z} \end{bmatrix}$$

Paraxial refraction model : This model's derivation is outside the scope of the class. Note that $z' = f$ in the pinhole model, but $z' = f + z_0$ in our lens-based approach. This derivation is also known as the paraxial refraction model since it uses the paraxial or "thin lens" assumption.

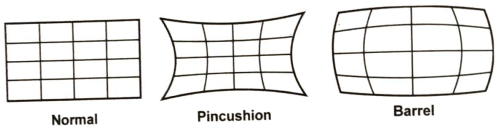


Fig. 1.2.14 Demonstrating how pincushion and barrel distortions affect images

A multitude of aberrations can emerge because the paraxial refraction model approximates using the thin lens assumption. The most prevalent is radial distortion, which causes the picture magnification to change as the distance from the optical axis increases or decreases. When the magnification increases, the radial distortion is classified as pincushion distortion, and when the magnification drops, it is classified as barrel distortion. The fact that different parts of the lens have different focal lengths causes radial distortion. The viewpoint model is flawed. Points are first projected to the reference plane using orthogonal projection and then projected to the image plane using a projective transformation in the weak perspective model.

1.2.5 3D to 2D Projections

- We need to specify how 3D primitives are projected onto the image plane to represent 2D and 3D geometric primitives and how to change them spatially. A linear 3D to 2D projection matrix can be used to do this. Orthography is the simplest model, as it does not require division to obtain the final (inhomogeneous) outcome. It is more common to use a perspective model since it more accurately captures the behavior of real cameras.

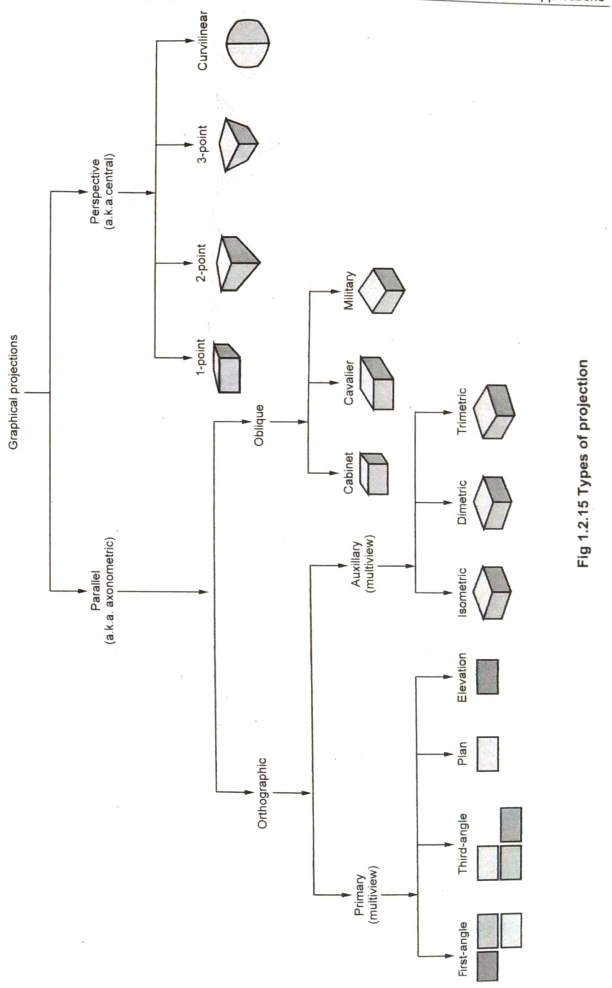
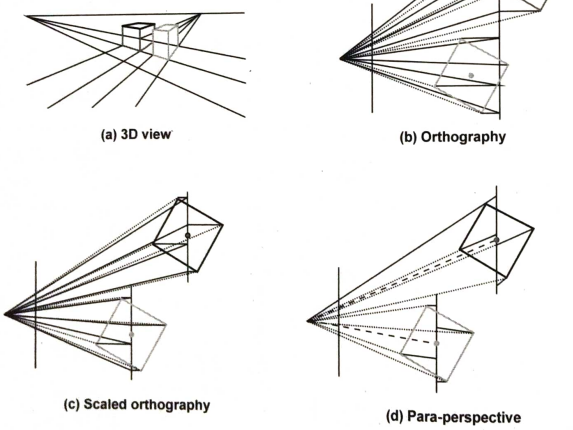


Fig 1.2.15 Types of projection

Orthography and para-perspective

Three-dimensional things are represented in two dimensions. It's a type of parallel projection in which all of the projection lines are orthogonal to the projection plane, resulting in affine transformations on the viewing surface for every plane in the picture. An oblique projection is the inverse of an orthographic projection, which is a parallel projection with projection lines that are not orthogonal to the projection plane. The term orthographic is sometimes used to describe renderings of objects in which the object's main axes or planes are also parallel to the projection plane. In multiview projection, these are referred to as primary views. Furthermore, axonometric visualizations are those in which the major planes or axes of an item in an orthographic projection are not parallel with the projection plane. However, these are most commonly referred to as auxiliary views. Sub-types of primary views include plans, elevations and sections. Sub-types of auxiliary views might include isometric, dimetric and trimetric projections. To obtain the 2D point x, an orthographic projection simply subtracts the z component of the three-dimensional coordinate p. (In this section, 3D points are denoted by p, and 2D points are denoted by x.) This can be expressed as :

$$x = [I_{2\times 2} | 0]\, p$$

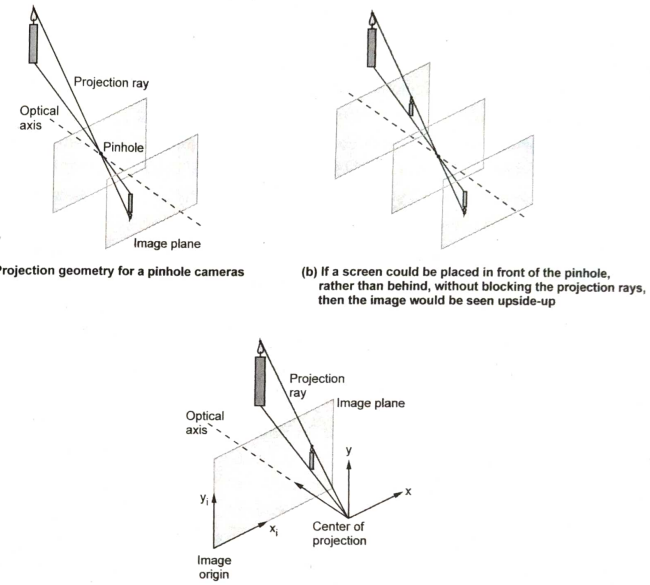


(a) 3D view (b) Orthography

(c) Scaled orthography (d) Para-perspective

Fig. 1.2.16 Commonly used projection models

The camera coordinate frame (x, y, z) is left-handed.



(a) Projection geometry for a pinhole cameras

(b) If a screen could be placed in front of the pinhole, rather than behind, without blocking the projection rays, then the image would be seen upside-up

(c) What is left is the so-called pinhole camera models

Fig 1.2.17 Projection geometry in pinhole camera

The pinhole is referred to as the center of projection in this model. The image plane is the front screen. The focal distance is the distance between the projection center and the picture plane, and it is symbolized by the letter f. The optical axis is a perpendicular to the image plane line that passes through the center of projection. The primary point is where the optical axis pierces the sensor plane. The origin of the picture coordinate system (x_i, y_i) is positioned at the bottom left corner of the image, as is typical in computer graphics. The axes of the camera reference system (x, y, z) are parallel to x_i, y_i, and the optical axis, respectively, and the z axis points toward the scene. The camera reference system is left-handed when the option in Fig. 1.2.17 (c) is selected. The depth of a point in the world is defined by its z coordinate. The units used in the camera reference system to measure

point coordinates (x, y, z) are frequently different from those used in the picture reference system (x_i, y_i). In most camera systems and imaging systems, metric units (meters, centimeters, and millimeters) are used. Pixels are the discrete, rectangular pieces of a digital camera's sensing array, as we'll learn in the section on sensing below. Because pixels are not always square, a millimeter measured horizontally on the array may have a different amount of pixels than a millimeter measured vertically, requiring two independent conversion units to convert pixels to millimeters in both directions. In the camera reference system, every point on the image plane has a z coordinate of f. The third coordinate in the picture reference system, on the other hand, is undefined because the image reference system is two-dimensional.

The other two coordinates differ by a translation and two separate unit conversions :

Let x_0 and y_0 be the pixel coordinates of the image's primary point in the image reference system (x_i, y_i). Then an image point in the camera reference frame with coordinates (x, y, f) in millimeters has image coordinates (in pixels)

$$x_i = s_x x + x_0 \quad \text{and} \quad y_i = s_y y + y_0$$

where, s_x and s_y are scaling constants expressed in pixels per milllimeter.

The projection equations relate a point in space's camera-system coordinates P = (X, Y, Z) to the camera-system coordinates p = (x, y) of P's projection onto the image plane and then to the projection's image-system coordinates p_i =(x_i, y_i) of the projection. These equations can be easily derived for the x coordinate. See that the triangle with orthogonal sides of length X and Z is similar to that with orthogonal sides of length x and f (the focal distance), so that X/Z = x/f. Similarly, for the Y coordinate, one gets Y/Z = y/f. In conclusion under perspective projection, the world point with coordinates (X, Y, Z) projects to the $x = f\frac{x}{z}, y = f\frac{y}{z}$ image point with coordinates.

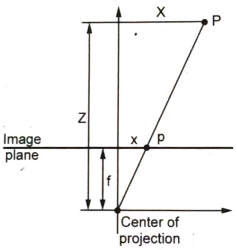


Fig. 1.2.18 A top view of Fig. 1.2.17 (c)

One way to make units of measure consistent in these projection equations is to measure all quantities in the same unit, say, millimeters. In this case, the two constants s_x and s_y in equations have the dimension of pixels per millimeter. It is sometimes more convenient to express x, y and f in pixels (image dimensions) and X, Y, Z in millimeters (world dimensions). The ratios x/f, y/f, X/Z and Y/Z are dimensionless, so they

dimensionally consistent with this choice as well. In this case, the two constants s_x and s_y in equation are dimensionless as well.

Orthographic projection

As the camera recedes and gets farther away from a scene of constant size, the projection rays become more parallel to each other. At the same time, the image becomes smaller, and eventually reduces to a point. To avoid image shrinking, one can magnify the image by $Z_{0/f}$, where Z_0 is the depth of, say, the centroid of all visible points, or that of an arbitrary point in the world. For the magnified coordinates x and y one then obtains,

$$x = x\frac{Z_0}{Z}$$

$$y = y\frac{Z_0}{Z}$$

As the camera recedes to infinity, Z and Z_0 grows at the same rate and their ratio tends to 1. The projection rays are parallel to each other and orthogonal to the image plane, is called orthographic projection.

Under orthographic projection, the world points with coordinates (X, Y, Z) projects to the image point with coordinates x = X, y = Y.

The linearity of these projection equations makes orthographic projection an appealing assumption whenever warranted, that is, whenever a telephoto lens is used.

1.2.6 Rigid and Affine Transformation

Affine : Affine transformation helps to modify the geometric structures of the image, preserving parallelism of lines but not modify the lengths and angles. It preserves and will not modify collinearity and ratios of distances.

The affine transform can be written as

$$x' = A\bar{x},$$

where, A is an arbitrary 3×4 matrix,

$$x' = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \end{bmatrix} \bar{x}$$

Parallel lines and planes remains parallel under affine transformations.

A **rigid transformation** is defined as a transformation that, when acting on any vectors v, produces a transformed vectors T(v) of the form,

$$T(v) = R\,v + t$$

where, $R^T = R^{-1}$ (i.e., R is an orthogonal transformation) and t is a vectors giving the translation of the origin.

A proper rigid transformation has, in addition,

$$\det(R) = 1$$

which means that R does not produce a reflections and hence it represents a rotations (an orientation - preserving orthogonal transformation). Indeed, when an orthogonal transformation matrix produces a reflections, its determinant is −1.

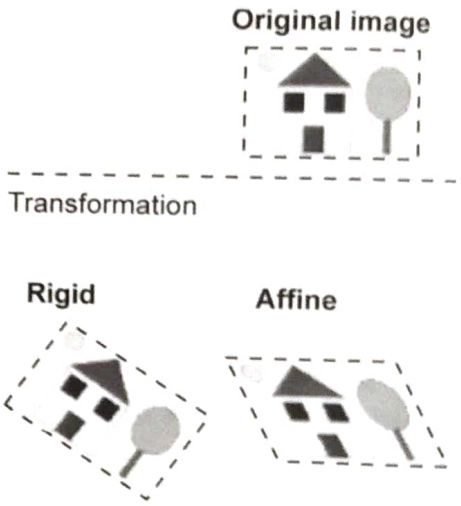


Fig 1.2.19 Rigid and affine transformation

❑❑❑