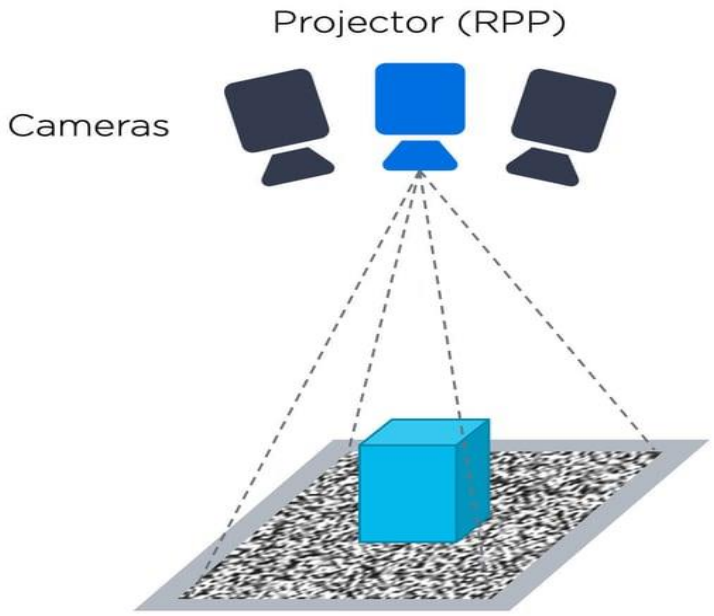
# UNIT-4

**3D Vision Systems**

3D vision system will require stereo-pair image acquisition hardware, usually connected to a computer hosting software that automates acquisition control. Multiple stereo-pairs of cameras might be employed to allow all-round coverage of an object or person, e.g. in the context of whole-body scanners. Alternatively, the object to be imaged could be mounted on a computer-controlled turntable and overlapping stereo-pairs captured from a fixed viewpoint for different turntable positions. Accordingly, sequencing capture and image download from multiple cameras can be a complex process, and hence the need for a computer to automate this process.

The stereo-pair acquisition process falls into two categories, active illumination and passive illumination. Active illumination implies that some form of pattern is projected on to the scene to facilitate finding and disambiguating parallaxes (also termed *correspondences* or *disparities*) between the stereo-pair images. Projected patterns often comprise grids or stripes and sometimes these are even colour coded. In an alternative approach, a random speckle texture pattern is projected on to the scene in order to augment the texture already present on imaged surfaces. Speckle projection can also guarantee that that imaged surfaces appear to be randomly textured and are therefore locally uniquely distinguishable and hence able to be matched successfully using certain classes of image matching algorithm. With the advent of 'high-resolution' digital cameras the need for pattern projection has been reduced, since the surface texture naturally present on materials, having even a matte finish, can serve to facilitate matching stereo-pairs.




For example, stereo-pair images of the human face and body can be matched successfully using ordinary studio flash illumination when the pixel sampling density is sufficient to resolve the natural texture of the skin, e.g. skin-pores. A camera resolution of approximately 8–13M pixels is adequate for stereo-pair capture of an area corresponding to the adult face or half-torso.

The acquisition computer may also host the principal 3D vision software components:

- An image matching algorithm to find correspondences between the stereo-pairs.
- Photogrammetry software that will perform system calibration to recover the geometric configuration of the acquisition cameras and perform 3D point reconstruction in world coordinates.
- 3D surface reconstruction software that builds complete manifolds from 3D *point-clouds* captured by each imaging stereo-pair.
- 3D visualisation facilities are usually also provided to allow the reconstructed surfaces to be displayed, often *draped* with an image to provide a *photorealistic* surface model. At this stage the 3D shape and surface appearance of the imaged object or scene has been captured in explicit digital metric form, ready to feed some subsequent application as described below.

**3D Vision Applications**

A wide variety of applications are now emerging which rely on the fast, efficient and low-cost capture of 3D surface information. The traditional role for image-based 3D surface measurement has been the reserve of *close-range* photogrammetry systems, capable of recovering surface measurements from objects in the range of a few tens of millimetres to a few metres in size. A typical example of a classical close-range photogrammetry task might comprise surface measurement for manufacturing quality control, applied to high-precision engineered products such as aircraft wings.

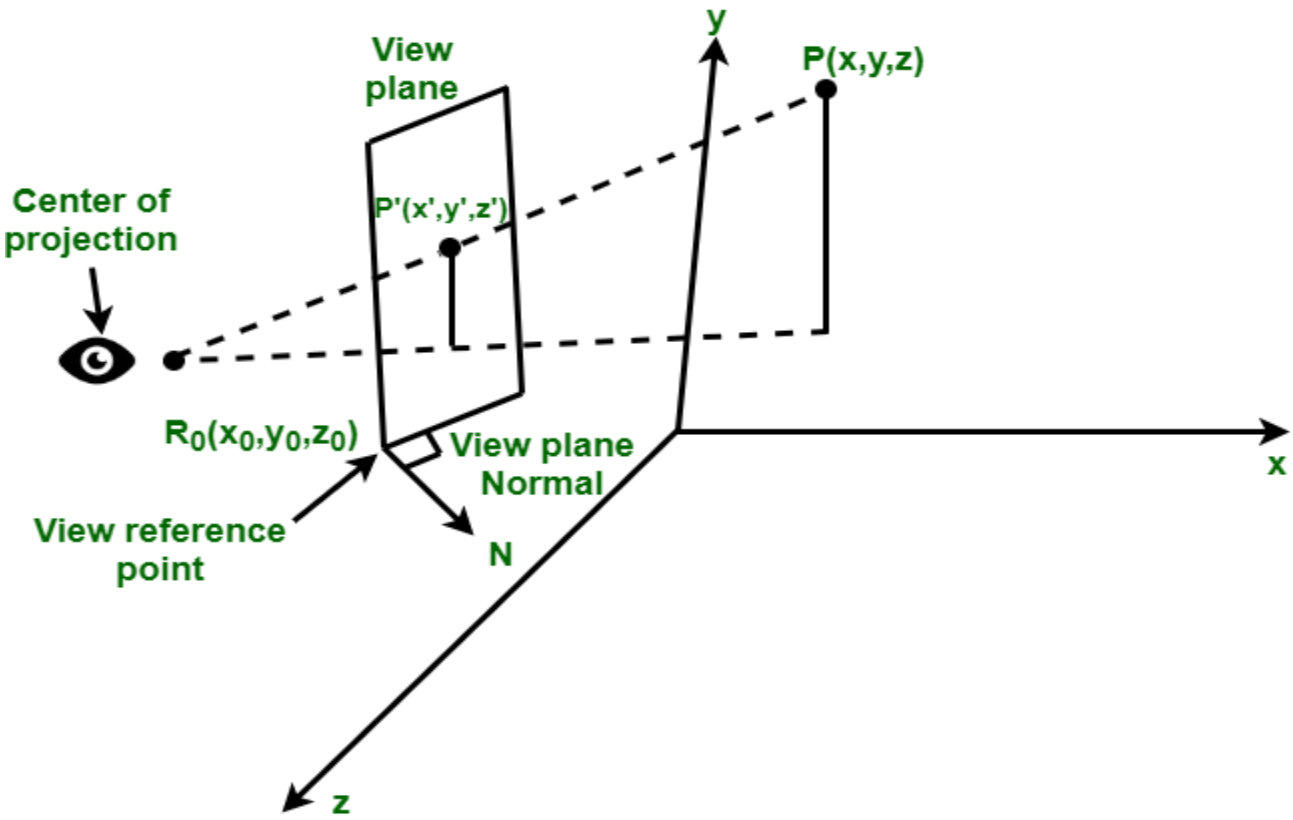
Close-range video-based photogrammetry, having a lower spatial resolution than traditional plate-camera film-based systems, initially found a niche in imaging the human face and body for clinical and creative media applications. 3D clinical photographs have the potential to provide quantitative measurements that reduce subjectivity in assessing the surface anatomy of a patient (or animal) before and after surgical intervention by providing numeric, possibly automated, scores for the shape, symmetry and longitudinal change of anatomic structures. Creative media applications include whole-body 3D imaging to support creation of human avatars of specific individuals, for 3D gaming and cine special effects requiring virtual actors. Clothing applications include body or foot scanning for the production of custom clothing and shoes or as a means of sizing customers accurately. An innovative commercial application comprises a ‘virtual catwalk’ to allow customers to visualize themselves in clothing prior to purchasing such goods on-line via the Internet.

There are very many more emerging uses for 3D imaging beyond the above and commercial ‘reverse engineering’ of premanufactured goods. 3D vision systems have the potential to revolutionize autonomous vehicles and the capabilities of robot vision systems. Stereo-pair cameras could be mounted on a vehicle to facilitate autonomous navigation or configured within a robot workcell to endow a ‘blind’ pick-and-place robot, both object recognition ca- pabilities based on 3D cues and simultaneously 3D spatial quantification of object locations in the workspace.

**Projection Schemes:** A projection scheme in computer vision maps 3D information onto a 2D image plane, with common types including perspective projection (simulating a camera lens and resulting in smaller distant objects) and orthographic projection (a form of parallel projection that maintains parallel lines and equal object sizes). These schemes are fundamental for how computers "see" and interpret visual data, and are often used in the context of cameras, 3D modeling, and analyzing visual input.

**Types of Projection Schemes**

1. **Perspective Projection:** This is the most common type of projection in computer vision, mimicking how a human eye or camera captures a scene. In Perspective Projection the **center of projection** is at finite distance from **projection plane**. This projection produces realistic views but does not preserve relative proportions of an object dimensions. Projections of distant object are smaller than projections of objects of same size that are closer to projection plane. The perspective projection can be easily described by the following figure:

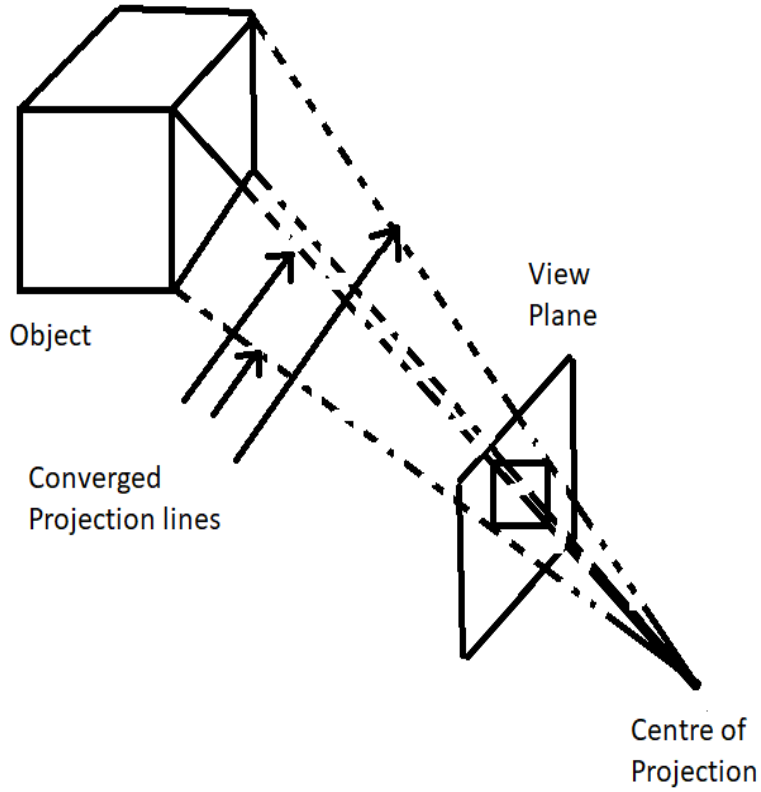



1. **Center of Projection -** It is a point where lines or projection that are not parallel to projection plane appear to meet.
2. **View Plane or Projection Plane -** The view plane is determined by :
   - View reference point $R_0(x_0, y_0, z_0)$
   - View plane normal.
3. **Location of an Object -** It is specified by a point P that is located in world coordinates at (x, y, z) location. The objective of perspective projection is to determine the image point P' whose coordinates are (x', y', z')

The perspective projection, on the other hand, produces realistic views but does not preserve relative proportions.

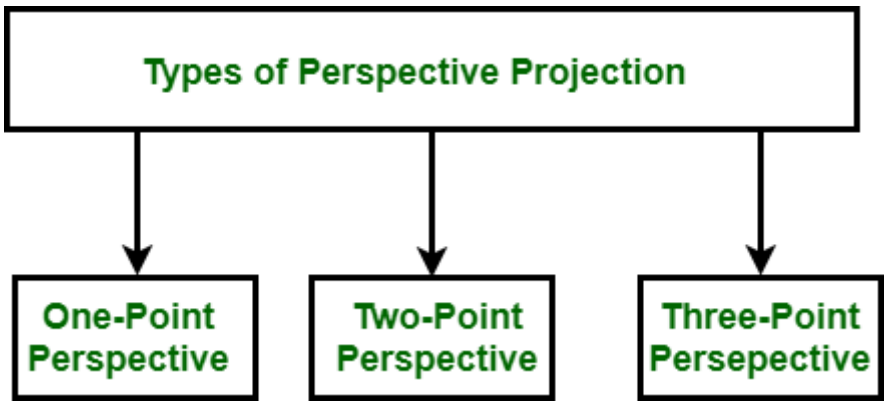
In perspective projection, the lines of projection are not parallel. Instead, they all converge at a single point called the center of projection or projection reference point.

The object positions are transformed to the view plane along these converged projection lines and the projected view of an object is determined by calculating the intersection of the converged projection lines with the view plane,as shown below figure

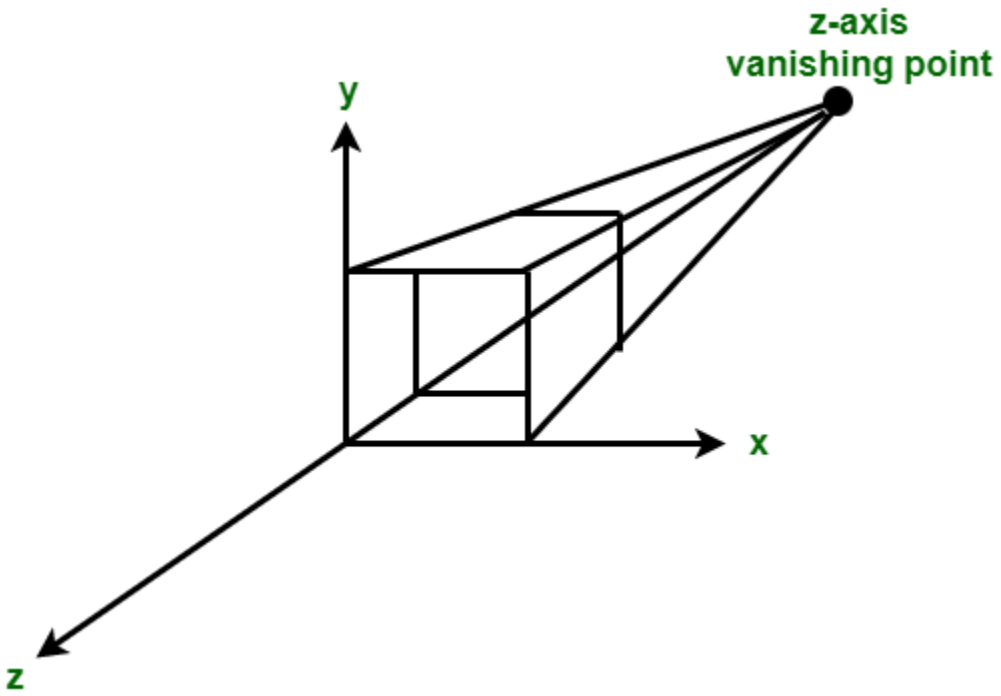



**Perspective Projection of an object to the view plane**

**Types of Perspective Projection:** Classification of perspective projection is on basis of vanishing points (It is a point in image where a parallel line through center of projection intersects view plane.). We can say that a vanishing point is a point where projection line intersects view plane. The classification is as follows:

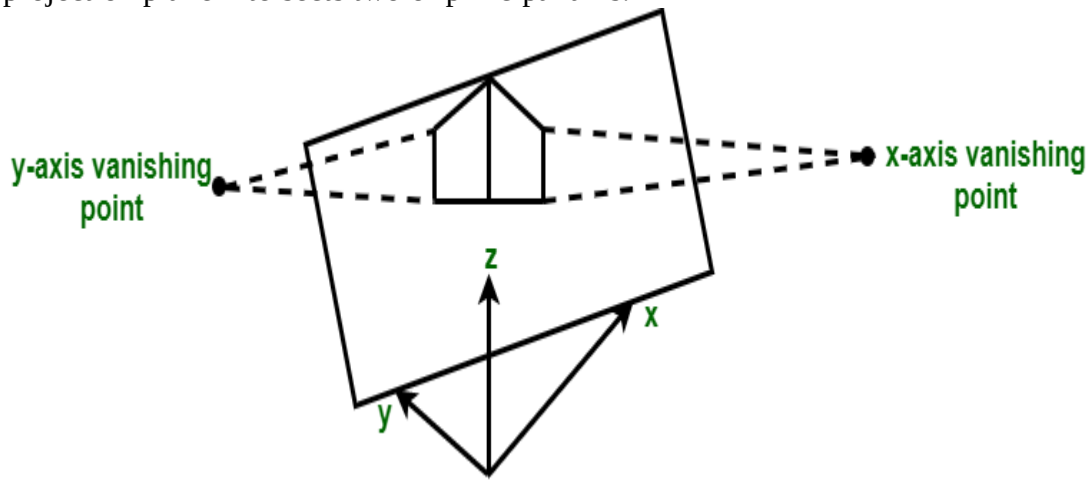


- **One Point Perspective Projection -** One point perspective projection occurs when any of principal axes intersects with projection plane or we can say when projection plane is perpendicular to principal axis.

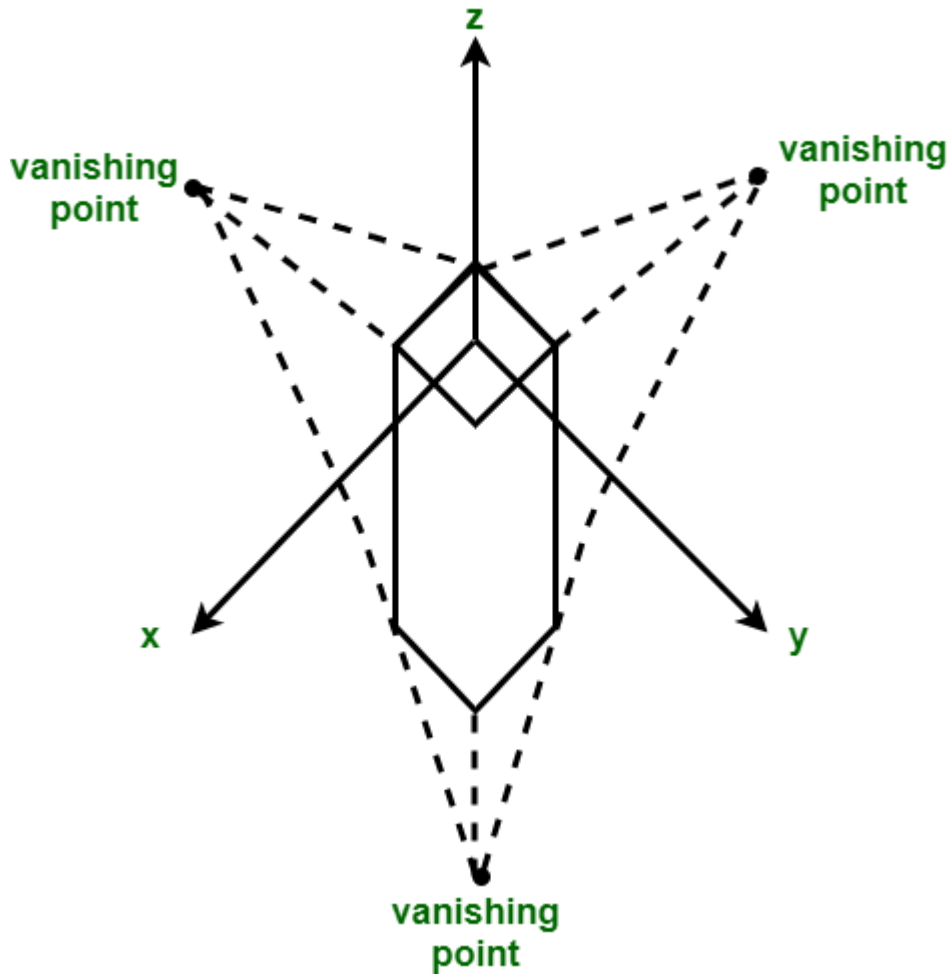



In the above figure, z axis intersects projection plane whereas x and y axis remain parallel to projection plane.

- **Two Point Perspective Projection -** Two point perspective projection occurs when projection plane intersects two of principal axis.



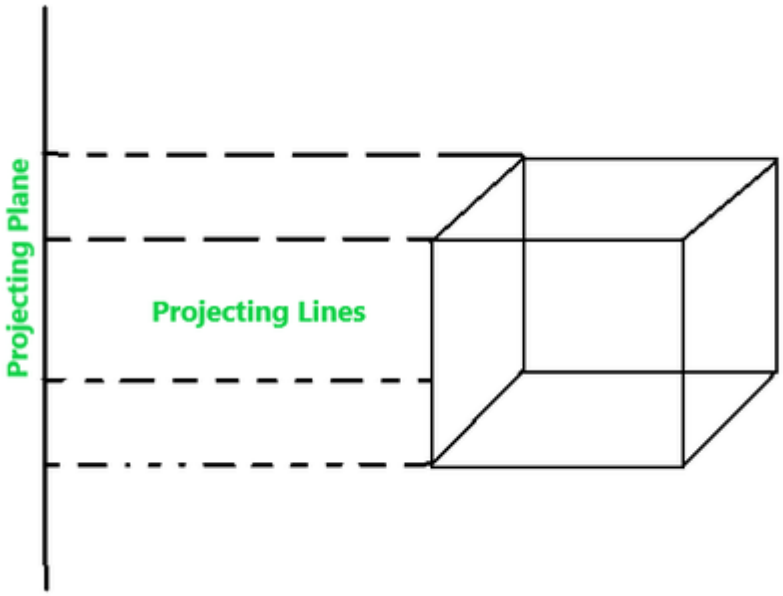

- In the above figure, projection plane intersects x and y axis whereas z axis remains parallel to projection plane.
- **Three Point Perspective Projection -** Three point perspective projection occurs when all three axis intersects with projection plane. There is no any principal axis which is parallel to projection plane.




**Characteristics:** Objects further away from the camera appear smaller, and parallel lines converge to a vanishing point.

**Applications:** Creating realistic 3D scenes, modeling camera behavior, and reconstructing 3D objects from images.

**Parallel Projection:** In contrast to perspective projection, parallel projection does not simulate a viewpoint. It is a kind of projection where the projecting lines emerge parallelly from the polygon surface and then incident parallelly on the plane. In parallel projection, the centre of the projection lies at infinity. In parallel projection, the view of the object obtained at the plane is less-realistic as there is no for-shortcoming. and the relative dimension of the object remains preserves.




**Parallel projection is further divided into two categories:**

**a) Orthographic Projection**

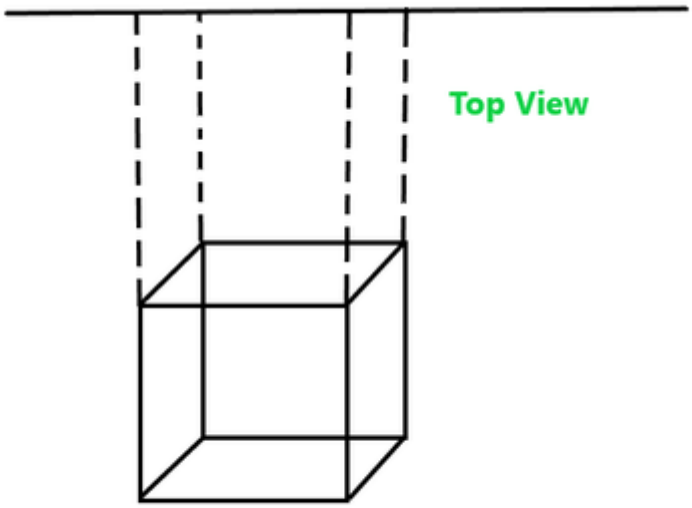
**b) Oblique Projection**

**(a) Orthographic Projection:** It is a kind of parallel projection where the projecting lines emerge parallelly from the object surface and incident perpendicularly at the projecting plane.

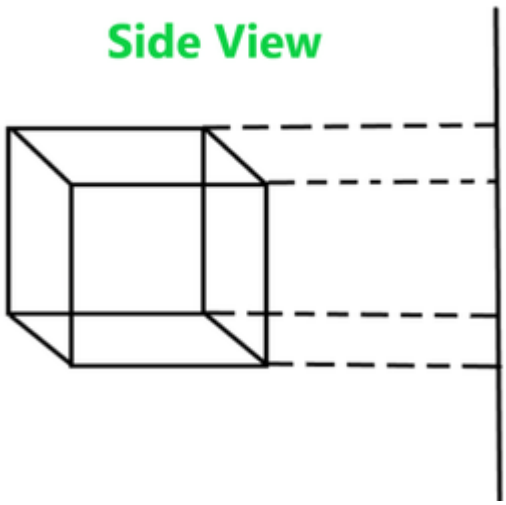
**Orthographic Projection is of two categories :**

**(a).1. Multiview Projection :** It is further divided into three categories -

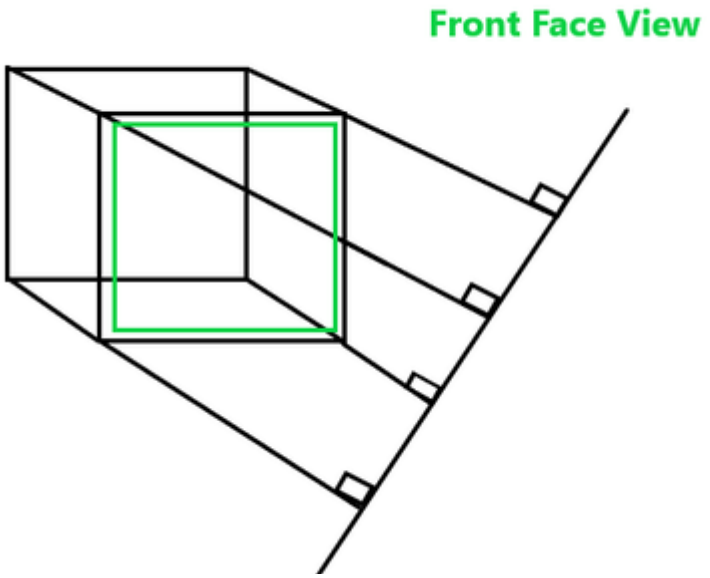
**(1) Top-View :** In this projection, the rays that emerge from the top of the polygon surface are observed.




**2) Side-View :** It is another type of projection orthographic projection where the side view of the polygon surface is observed.



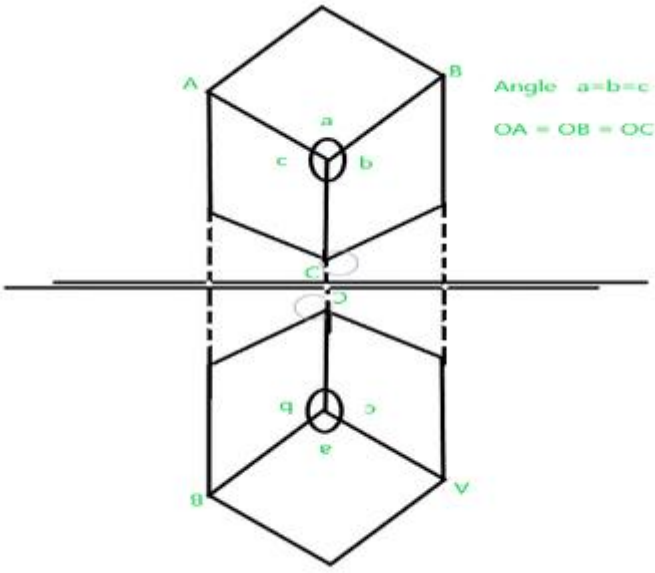

**3) Front-view :** In this orthographic projection front face view of the object is observed.



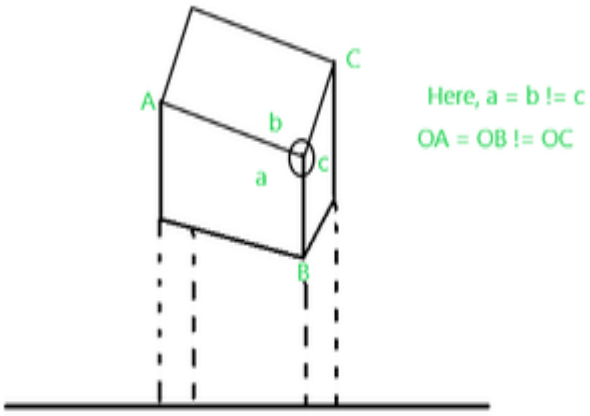

**a.2) Axonometric :** Axonometric projection is an orthographic projection, where the projection lines are perpendicular to the plane of projection, and the object is rotated around one or more of its axes to show multiple sides.

It is further divided into three categories :

**(1) Isometric Projection :** It is a method for visually representing three-dimensional objects in two-dimensional display in technical and engineering drawings. Here in this projection, the three coordinate axes appear equally foreshortened and the angle between any two of them is 120 degrees.

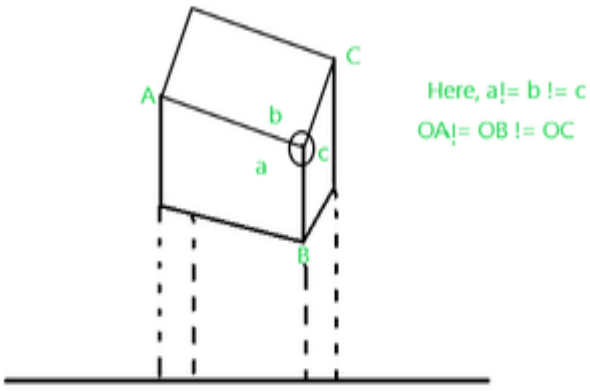


**(2) Dimetric Projection :** It is a kind of orthographic projection where the visualized object appears to have only two adjacent sides and angles are equal.




Dimetric Projection

**(3) Trimetric Projection :** It is a kind of orthographic projection where the visualized object appears to have all the adjacent sides and angles unequal.
\



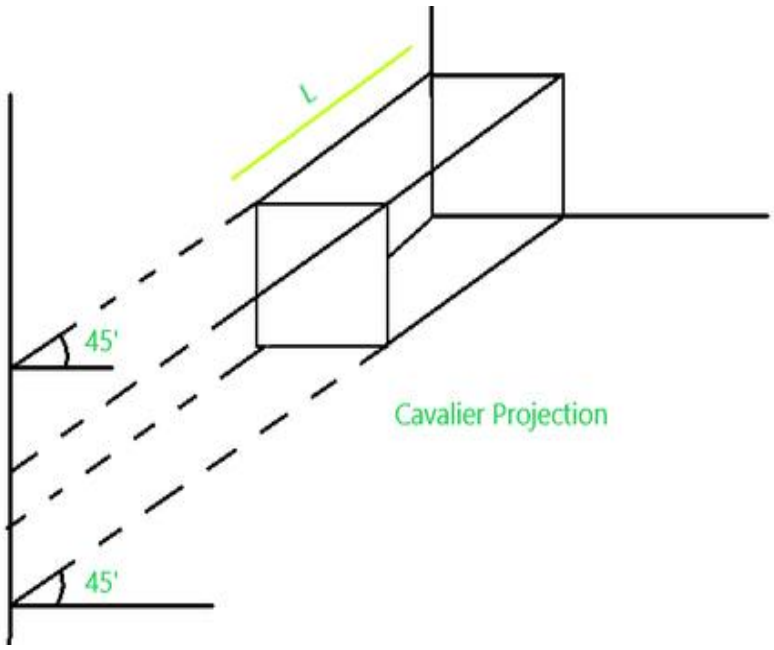

Trimetric Projection

**(b) Oblique Projection :** It is a kind of parallel projection where projecting rays emerges parallelly from the surface of the polygon and incident at an angle other than 90 degrees on the plane.

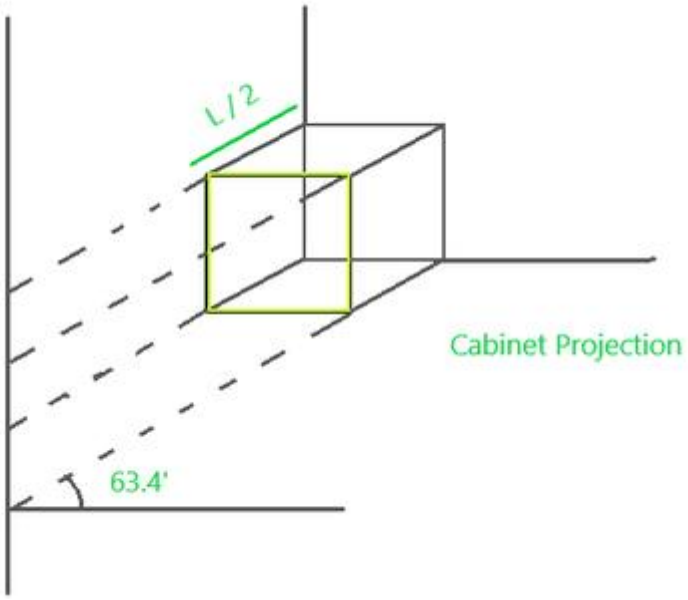
It is of two kinds :

**(b).1. Cavalier Projection :** It is a kind of oblique projection where the projecting lines

emerge parallelly from the object surface and incident at 45‘ rather than 90' at the projecting plane. In this projection, the length of the reading axis is larger than the cabinet projection.

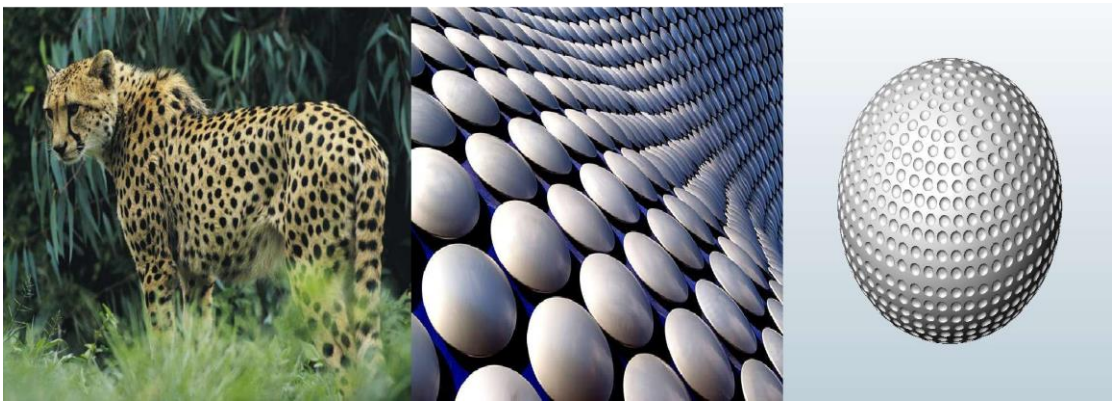



**(b). 2. Cabinet Projection:** It is similar to that cavalier projection but here the length of reading axes just half than the cavalier projection and the incident angle at the projecting plane is 63.4' rather 45'.



**Shape from texture:** Shape from texture is a computer vision technique where a 3D object is reconstructed from a 2D image. Although human perception is capable to realize patterns, estimate depth and recognize objects in an image by using texture as a cue, the creation of a system able to mimic that behavior is far from trivial.

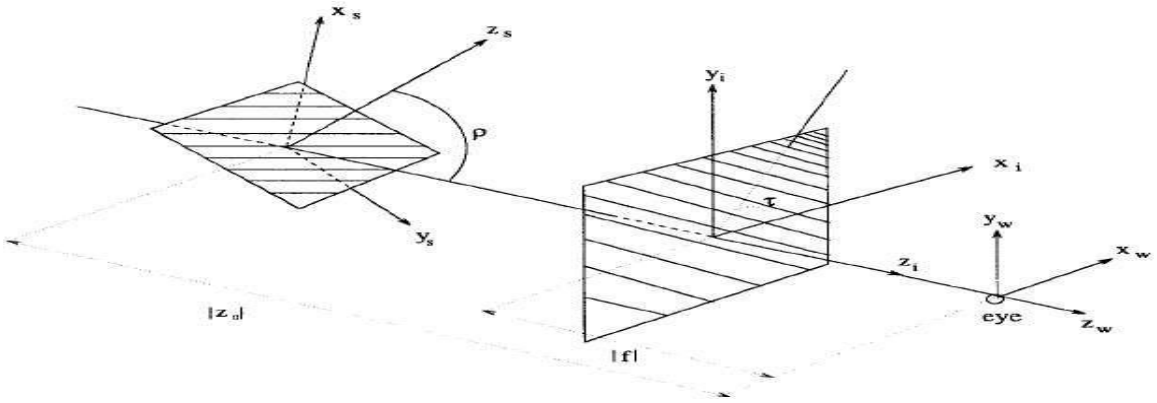
Although texture as a meaning is difficult to describe in our case we mean the repetition of an element or the appearance of a specific template over a surface. Such element or surface is called texel (TEXture ELement). Various textures can be seen in figure .



Some examples of textures

The first person who proposed that a shape can be perceived from a texture was Gibson in 1950. Gibson used the term texture gradient in order to denote that areas of a surface that have similar texture, with other neighbor areas, are perceived differently from the observer due to differences in orientation of the surfaces and the distance from the observer.

In order to measure the orientation of the texels in a texture, we need to find the slant and tilt angles. Slant denotes the amount and tilt denotes the direction of the slope of the planar surface projected on the image plane. In figure the angle ρ between $z_s$ and $z_i$ is the slant angle while the angle τ between $x_i$ and the projection of the surface normal $z_s$ onto the image plane is the tilt angle.




Coordinate relationship between the image plane and the surface plane

**Shape from focus:** It is a method of 3D reconstruction which consists of the use of information about the focus of an optical system to provide a means of measurement for 3D information. One of the simplest forms of this method can be found in most autofocus cameras today. In its most simple form, the methods analyze an image based upon overall contrast from a histogram, the width of edges, or more commonly, the frequency spectrum derived from a fast Fourier transform of the image. That information might be used to drive a servo mechanism in the lens, focusing it until the quantity measured on one of the earlier parameters is optimized.

**Principle of Operation**

1. **Image Sequence Capture:** A camera captures multiple images of the same scene with varying focal lengths.
2. **Sharpness Measurement:** For each pixel, a **focus measure operator** is applied to determine sharpness.

   Example operators:

   Variance of Laplacian

   Gradient-based focus measures

   Wavelet-based methods

3. **Depth Estimation**: The focus level (image index) where a pixel has maximum sharpness → corresponds to its depth.
4. **3D Reconstruction**: Combining the depth values of all pixels → produces a **depth map** and hence 3D shape.

**Advantages of SFF**

- Passive method (no structured light or additional sensors needed).
- Simple setup with a standard camera and lens.
- Provides dense depth maps (depth at almost every pixel).

**Limitations of SFF**

- Requires multiple images → time-consuming.
- Performance depends on texture (works poorly in homogeneous regions).
- Sensitive to noise and illumination changes.
- Depth accuracy limited by camera optics and step size of focus.

**Applications**

- Microscopy imaging (for fine 3D structures).
- Industrial inspection (surface measurements).
- Medical imaging (cell/tissue structure analysis).
- Document analysis (recovering embossed or degraded text).

**Volumetric Representation:** In computer vision, a volumetric representation models a 3D object or scene as a collection of discrete points or values in a 3D space, often using a voxel grid. Each voxel contains information about the property of the object at that 3D location, such as its color, density, or whether it's occupied or empty. This approach allows for the capture and rendering of complex 3D phenomena like smoke, fire, and fluids, and is particularly useful for applications requiring detailed spatial context, such as 3D reconstruction and object recognition.

Key Concepts

- **Voxels:** The fundamental units of a volumetric model, representing a small 3D volume or cell within a 3D grid.
- **Voxel Grid:** A discrete 3D grid where each voxel has a specific location and a value representing a property (e.g., binary value of 0 or 1 for occupied/empty).
- **Properties:** Each voxel can store various properties beyond simple occupancy, including color, density, temperature, or pressure.

How it Works

**Discretization:** A continuous 3D scene is divided into a grid of voxels.

**Data Assignment:** For each voxel, a value is assigned to represent the material or attribute at that location.

**3D Convolution:** Deep learning models can leverage the 3D structure of volumetric representations by applying 3D convolutions to process and understand spatial relationships within the object or scene.

Applications

1. **3D Reconstruction:**

Volumetric representations can be used to build detailed 3D models of objects and environments from multiple input images or depth data.

2. **Volume Rendering:**

This technique creates a 2D image by projecting the entire 3D volume onto a plane, allowing for visualization of complex internal structures without the need for intermediate geometric representations.

3. **Neural Radiance Fields (NeRFs):**

Modern deep learning methods like NeRFs use volumetric representations to learn to generate photorealistic renderings of scenes, capturing appearance and geometry.

4. **Modeling Complex Phenomena:**

Volumetric models are ideal for representing phenomena that are not easily described by surfaces or geometric primitives, such as fire, smoke, clouds, and fluids.

5. **3D Printing:**

V-reps are revolutionizing 3D printing by enabling the creation of objects with internal gradients and complex structures.

**3D Reconstruction:** Three-dimensional (3D) reconstruction is a field of study focused on creating three-dimensional representations of objects, scenes, or environments from two-dimensional (2D) images or other sensor data, with the goal of capturing the spatial structure and geometry of real-world entities in a digital format. 1 This process bridges the gap between the physical world and the digital realm by converting visual or sensor information from single or multiple views into cohesive and accurate 3D models. It is a particularly valuable area within computer vision, computer graphics, and machine learning.

3D reconstruction technology is essential for a wide range of applications, including medical imaging, game development, industrial manufacturing, cultural heritage protection, architectural design, urban planning, and the metaverse.
For example, in the metaverse, 3D reconstruction enables the creation of realistic virtual environments and human avatars by transforming real objects into digital models that capture their shape and appearance. In construction, it is used for progress monitoring, structure inspection, and post-disaster rescue by building as-built 3D models from 2D images or laser point clouds captured on site. 3 Despite advances in deep learning-based algorithms for single-view 3D reconstruction, challenges remain due to difficulties in reconstructing complex shapes, uncertainties in object reconstruction, and limitations in capturing comprehensive 3D information from a solitary view. Multi-view 3D reconstruction offers a more tractable and precise approach by leveraging information from multiple perspectives, enabling more accurate estimation of 3D shape.

**Fundamentals and Representations of 3D Reconstruction**
3D reconstruction involves estimating the 3D geometric properties of an object from its 2D images, including 3D coordinates, orientation, and depth for each point in space. Full 3D reconstruction captures all three degrees of freedom for each point, while 3D shape reconstruction recovers the three-dimensional orientation (two degrees of freedom), and depth reconstruction focuses on the z coordinate (one degree of freedom) for each 3D point. 4 Techniques include single-image methods (shape from X), stereo vision (using two images), and structure from motion (SfM) from image sequences. Stereo-based reconstruction involves point matching and triangulation, with challenges such as ambiguous correspondences and large search spaces, often mitigated by geometric constraints like the epipolar and spatial order constraints.
Common 3D data representations include:

- **Point clouds**: Collections of discrete points in 3D space, each with coordinates (x, y, z), color, and surface normal information. They express spatial distribution and surface characteristics but lack inherent topological relationships, making processing dependent on acquisition accuracy and device characteristics.
- **Meshes**: Composed of vertices and edges forming surfaces, typically triangles or quadrilaterals. Meshes provide higher accuracy, generalization, and expressiveness but can struggle with complex geometry and topological changes such as splitting and merging.
- **Voxels**: Volumetric pixels discretizing 3D space into regular grids, offering good discrimination for simple objects and efficient computation but are memory-intensive at high resolutions and may lose detail as voxel size increases.
- **Depth maps**: 2D arrays where each pixel encodes the distance from the camera to a point in the scene. They require the least storage but often fail to capture correlated spatial information and surface details, with sensitivity to illumination and occlusion.

- **Implicit surfaces**: Use continuous functions, such as signed distance functions, to define surfaces as zero-level sets, enabling high-resolution and flexible modeling.

However, they may require conversion to explicit forms like meshes for rendering and can be computationally expensive for large-scale scenes.

The choice of representation depends on the specific application and scene characteristics, balancing accuracy, efficiency, and expressiveness.

**Introduction to motion:** In computer vision, motion analysis is the study of how objects appear to move by analyzing sequences of images over time. This process involves estimating motion vectors, often in the form of optical flow, which is a dense displacement field describing pixel movement from one frame to the next. Key tasks include motion detection to identify moving objects, motion estimation to quantify motion, and motion tracking to follow objects across frames. Applications range from surveillance and robotics to video compression and augmented reality.

**Motion Estimation:**

The process of determining how objects in a video sequence have moved between frames by estimating motion vectors.

**Optical Flow:**

A dense field of 2D vectors that describes the apparent motion of each pixel in an image, representing the direction and speed of movement from one frame to the next.

**Motion Detection:**

The task of identifying which pixels or regions within a video sequence represent moving objects, often by detecting changes between frames.

**Motion Tracking:**

The process of continuously following a specific object's movement across a series of video frames.

Key Tasks

1. **Frame Differencing:** A simple technique that highlights changes between consecutive frames by subtracting one frame from another, revealing moving areas.

2. **Background Subtraction:** Distinguishes foreground objects by modeling the static background and then identifying pixels that deviate from this model.

3. **Motion Segmentation:** Separating moving objects from the static background to isolate them for further analysis.

4. **Structure from Motion (SfM):** A more complex process that uses sequences of images to simultaneously estimate the 3D structure of a scene and the camera's motion.

Applications

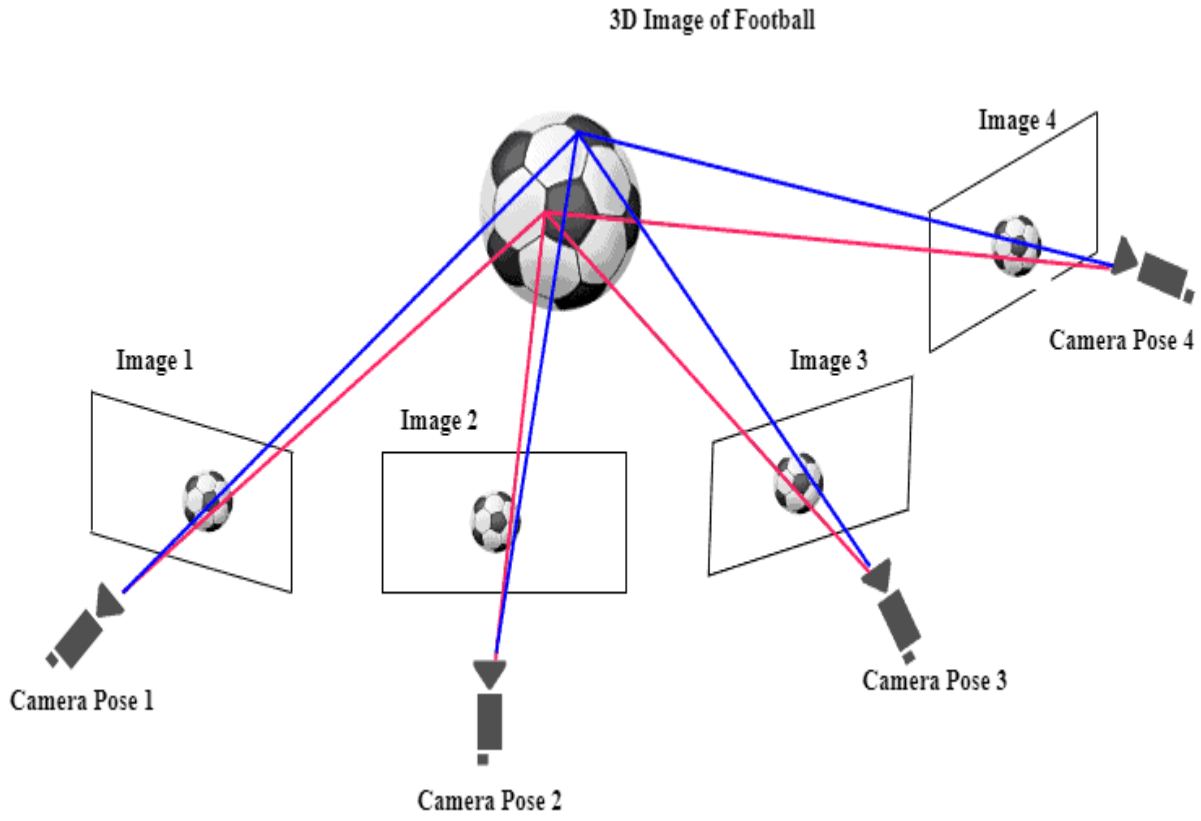
- **Surveillance:** Detecting suspicious activities, tracking vehicles, and recognizing abnormal events.

- **Robotics & Autonomous Systems:** For navigation, obstacle avoidance, and understanding the environment.

- **Video Compression:** Efficiently encoding video by only storing changes between frames rather than entire frames.

- **Augmented Reality (AR):** Separating foreground objects to allow for AR elements to interact realistically with the real world.
- **Human-Computer Interaction:** Tracking gestures and lip movements for gesture control or speech-to-text systems.
- **Video Stabilization:** Compensating for camera movement to produce smoother video footage.

**Bundle adjustment** (BA): It enhances the accuracy and reliability of 3D scene reconstructions from multiple images and camera views:




We use it to **correct errors from the initial 3D reconstruction process,** including inaccuracies in camera pose, scene structure, or feature tracking.

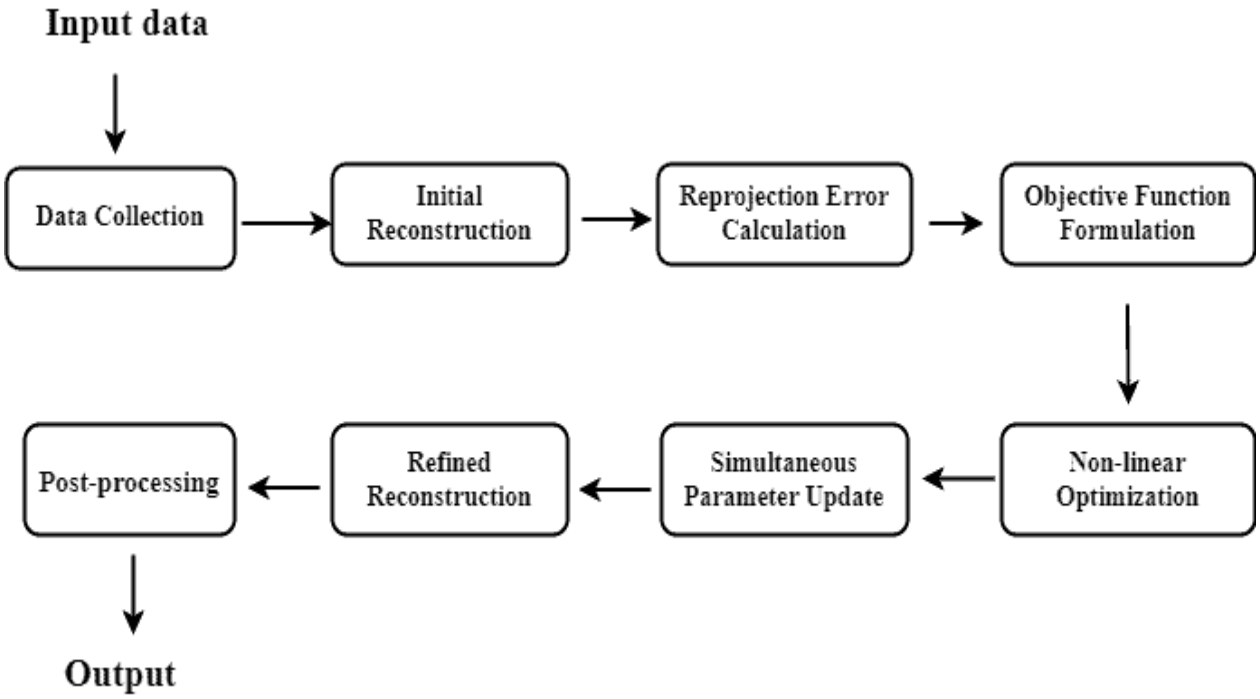
**Applications**

We use it for object tracking, augmented reality, and scene understanding.

Additionally, it contributes to developing lifelike 3D environments, elevating user immersion and interaction in virtual reality and simulation applications.

**Workflow and Key Components of Bundle Adjustment**

The workflow of Bundle Adjustment (BA) consists of a series of steps designed to refine the parameters of a 3D reconstruction model from several 2D images:




**Data Collection and Initial Reconstruction**

We start with a collection of 2D images capturing the scene or object of interest from different viewpoints, ideally with overlapping features or key points recognizable in multiple images.

Translational alignment in computer vision is a geometric transformation technique that involves shifting an image or a set of images in a specific direction without changing its size or shape, allowing for precise pixel movement to align objects or scenes. This process is fundamental in tasks like creating panoramic images, stabilizing video, and aligning sequences from different viewpoints by minimizing misalignment between images. Algorithms for translational alignment typically involve techniques like feature detection and matching, image transformation, and optimization methods, often using a similarity metric to find the optimal translational parameters.

How Translational Alignment Works:

1. **Shifting Pixels**: Translation moves every pixel within an image by a fixed distance along the horizontal (x) and vertical (y) axes.

2. **Mathematical Representation**: The movement can be represented by a 2D or 3D vector, which indicates the distance and direction of the shift.

3. **Alignment Process**:
    - **Feature Detection**: Identify specific, recognizable points or features within the images.

- **Feature Matching**: Match these features between the input image and a reference or database image to find corresponding points.
- **Transformation Estimation**: Estimate the translational parameters (how much to shift) by using the matched features and an optimization method, such as a least-squares minimization technique.
- **Image Transformation**: Apply the estimated translation to the image, effectively shifting its pixels to align with the other image.

## Applications

- **Image Stitching**: Combining multiple images to create a larger, high-resolution image, such as panoramas or satellite photos.
- **Video Stabilization**: Aligning video frames to correct for camera motion, resulting in a stable and smooth video.
- **Optical Flow**: A foundational technique used to track motion between consecutive frames by identifying pixel shifts.
- **Photometric Stereo**: Aligning image sequences captured under varying lighting conditions to better estimate surface normals and details.
- **Object Recognition**: Aligning an object in an image to a standard orientation for recognition purposes.