Computer Vision

Unit-3

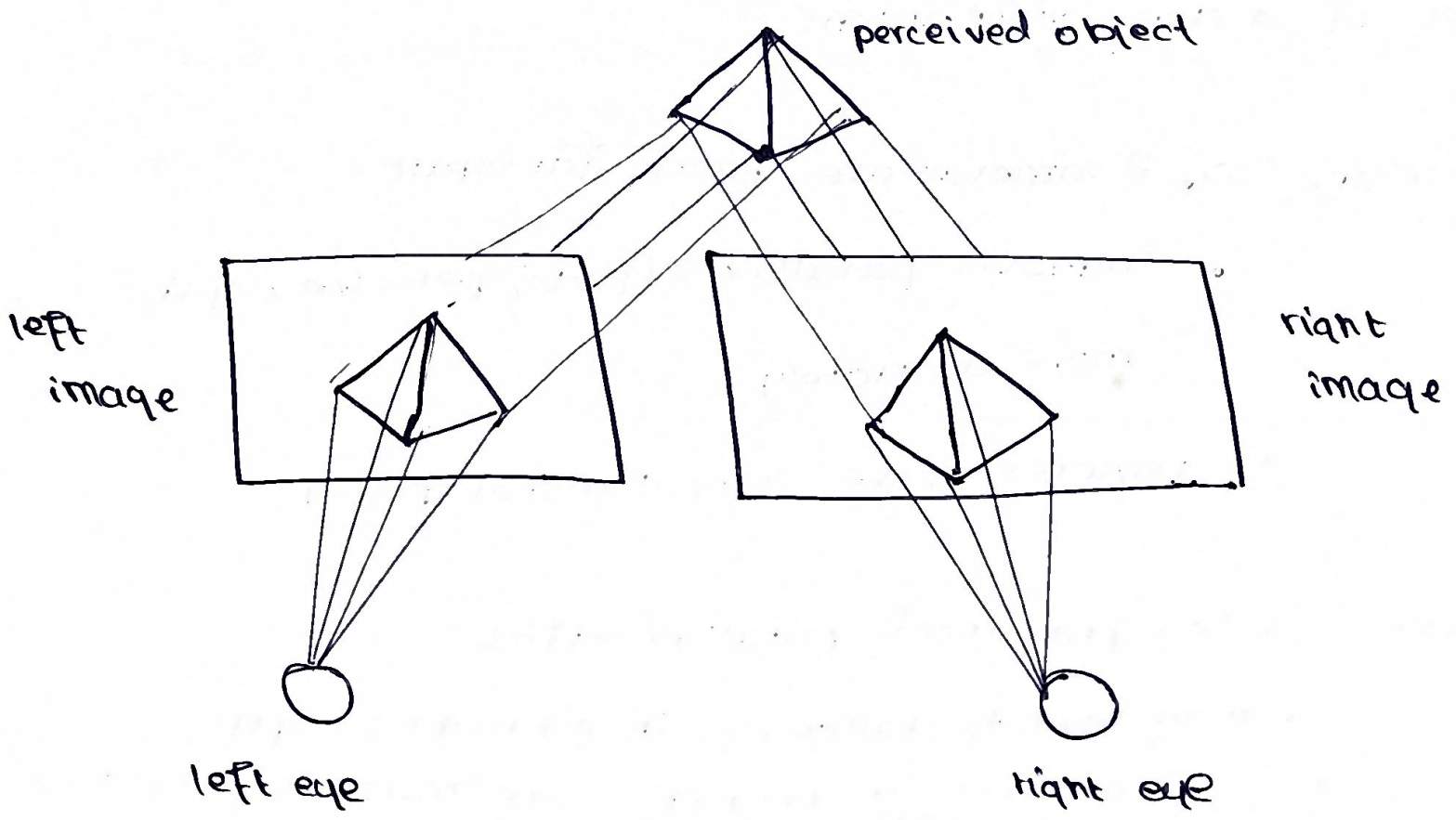
3D Vision- Depth Estimation and 3D Reconstruction

Feature based alignment: 2D and 3D feature-based alignment- pose estimation - geometric intrinsic calibration; structure from motion: triangulation - two-frame structure from motion - Factorization - bundle adjustment; stereo correspondence: epipolar geometry - sparse and dense correspondence - multiview stereo; 3D construction: Shape from X - active range finding - surface representations - point-based representation

3D-Vision / Stereo Vision

→ The recovery of the 3D structure of a scene using two or more images of the 3D scene, each acquired from a different viewpoint in space.

→ When 2 cameras are used, binocular vision is used.



* Terminology

(i) Fixation Point : The point where the two optical axes of the cameras intersect.

→ In a stereo vision setup, each camera captures a slightly different view of the same scene due to their different positions and angles.

→ The fixation point helps align these views and helps calculate depth or distance of objects.

(ii) Baseline - The baseline is the distance between the centers of projection (point at which light rays converge) to form an image of the 2 cameras in a stereo vision.

→ This distance is measured in physical units such as millimeters or meters.

→ The baseline plays a crucial role in determining the depth perception capabilities of a stereo vision system.

Large baseline ⇒
- 2 cameras are spaced far apart
- improves parallax effect, perceive depth more accurately
- requires more computational power

Small baseline ⇒
- less pronounced parallax effect
- may lead to challenges in estimating depth
- differences in perspective are much less evident
- computationally efficient

(iii) Conjugate Pair : A conjugate pair consists of a point that is visible to both cameras. These corresponding points are essential for stereo matching algorithms to calculate the depth or disparity of objects.

(iv) Disparity : a measure of the difference in horizontal position (usually along the x-axis) between corresponding points in the two images captured by the stereo cameras.

Larger disparities → objects are closer to cameras

Smaller disparities → objects are farther away

(v) Disparity Map : a visual representation of the disparities of all points in the scene as captured by the stereo cameras.

* Triangulation

→ Determines the position of a point in space by finding the intersection of the two lines passing through the center of projection and the projection of the point in each image.

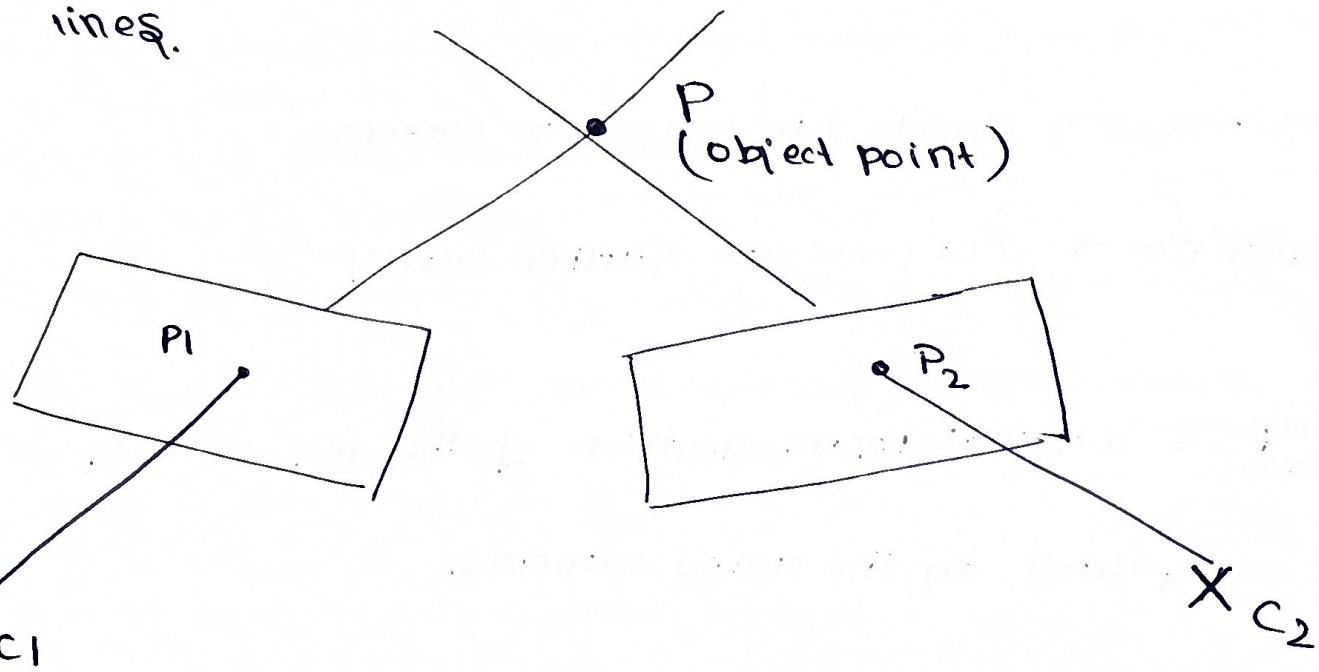
Steps : (i) 2 cameras are used to capture images of a scene.

Each camera has its own center of projection, which is the point where the light rays converge to form an image.

(ii) Consider a point P in space that we want to determine the 3D coordinate of. This point is visible in both camera views and is projected onto image planes P_1 & P_2.

(iii) Draw projection lines from the center of projection of each camera (C_1 & C_2) to the projected points P_1 & P_2

(iv) Triangulation finds the intersection point of those projection lines.



(left center of projection) (right center of projection)

* Stereo Camera Parameters

(i) Extrinsic Parameters : (R, T) describes the relative position and orientation of the two cameras.

$$P_r = R(R_l - T)$$

(ii) Intrinsic Parameters - characterizes the transformation from the image plane coordinates to pixel coordinates, in each camera.

* Problems in Stereo vision

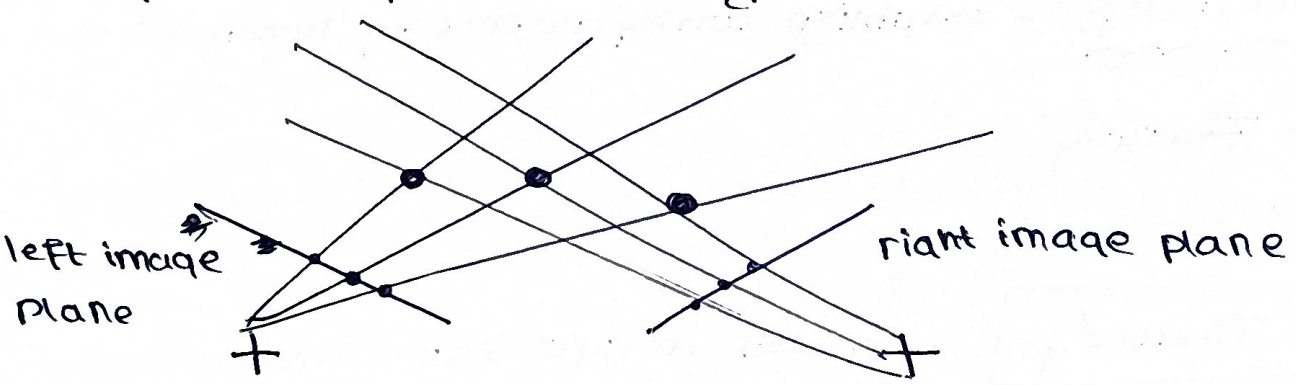
1. Correspondence problem
2. Reconstruction problem

A. Correspondence Problem

→ Refers to the challenge of accurately matching corresponding points between the left and right images captured by stereo cameras.

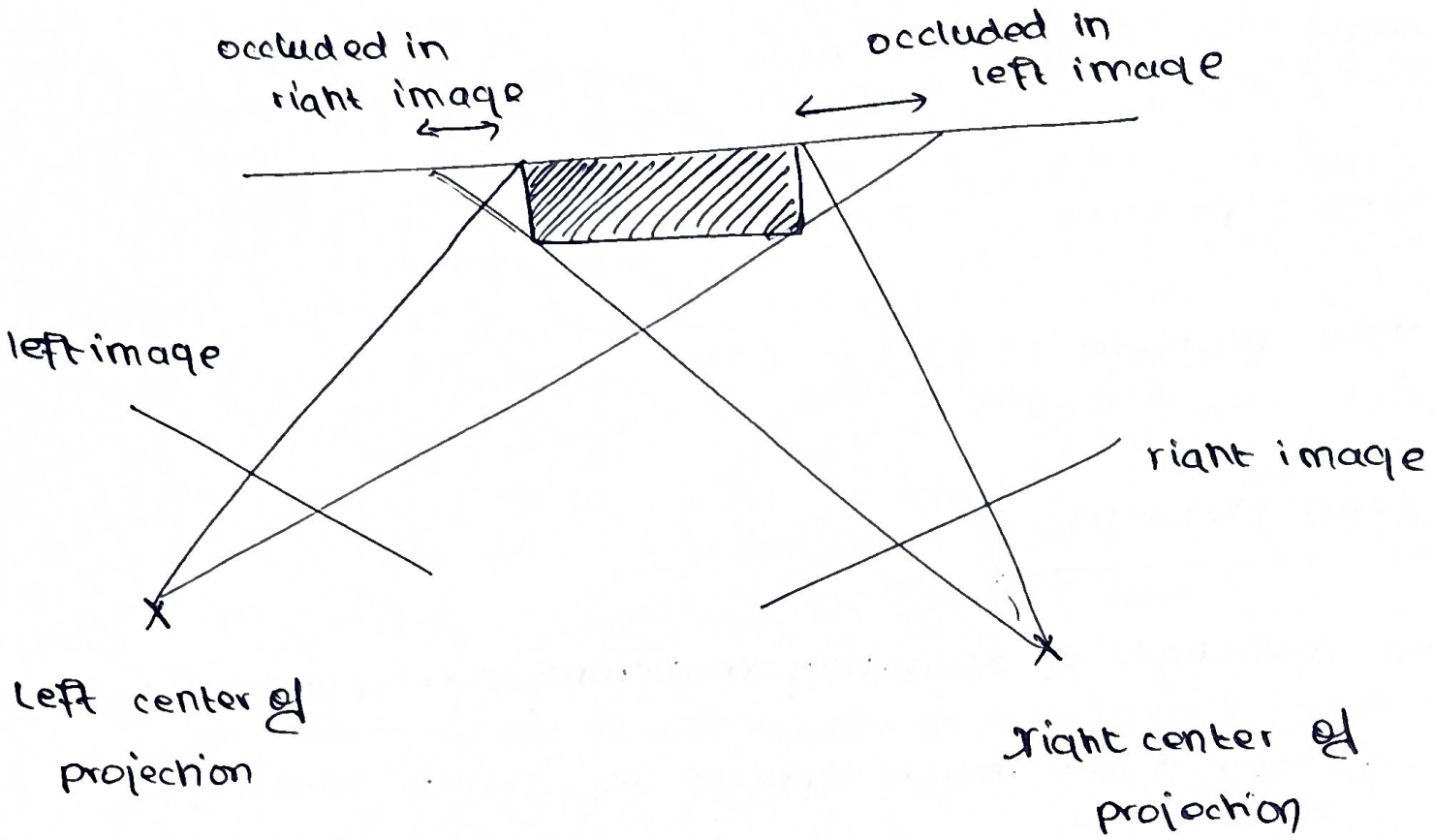
→ Factors contributing to the correspondence problem include variations in lighting, texture, occlusions & noise.

→ Triangulation depends crucially on the solution of the correspondence problem. Ambiguous correspondences may lead to several different representations of the scene.



→ There may also be occlusion - some points in each image will have no corresponding points in the other image, i.e the cameras may have different fields of view.

→ A stereo system must be able to determine the image parts that should not be matched.



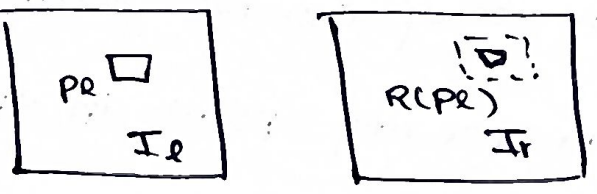
Addressing Correspondence Problems

(1) Intensity-based methods - establish correspondence by matching image intensities

(2) Feature-based methods - establish correspondence by matching a sparse set of image features

(1) **Intensity-based Methods** → match image sub-windows between the two images using correlation

A. Correlation For Establishing Correspondence



Algorithm

Inputs : (i) I_e and I_r

(ii) The width of the subwindow 2w+1

(iii) The search region in the right image R(Pe) associated with a pixel P_e in the left image.

For each pixel $p_l = (i,j)$ in the left image

1. For each displacement $d = (d_1, d_2) \in R(p_l)$ compute:

$$c(d) = \sum_{k=-w}^{w} \sum_{l=-w}^{w} I_l(i+k, j+l)\, I_r(i+k-d_1, j+l-d_2)$$

= cross-correlation

Q. The disparity of p_0 is the vector $\bar{d} = (\bar{d_1}, \bar{d_2})$ that maximizes $c(d)$ over $R(p_r)$

$$\bar{d} = \underset{d \in R}{\operatorname{argmax}} \left[c(d) \right]$$

B. Normalized Cross-Correlation : Normalize $c(d)$ by subtracting the mean and dividing by the standard deviation

$$c(d) = \frac{\sum_{k=-w}^{w} \sum_{l=-w}^{w} \left(I_l(i+k, j+l) - \bar{I_l}\right)\left(I_r(i+k-d_1, j+l-d_2) - \bar{I_r}\right)}{\sqrt{\sum_{k=-w}^{w} \sum_{l=-w}^{w} \left(I_l(i+k, j+l) - \bar{I_l}\right)^2 \sum_{k=-w}^{w} \sum_{l=-w}^{w} \left(I_r(i+k-d_1, j+l-d_2) - \bar{I_r}\right)^2}}$$

* Success of Correlation-based Methods - Depends on whether the image window in one image exhibits a distinctive structure that occurs infrequently in the search region of the other image.

* Choosing parameters for Correlation-based methods

(a) Window size

Too small a window → may not capture enough image structure
may be too noise sensitive

Too large a window ⇒ makes matching less sensitive to noise, but
is also harder to match

(b) Size and Location of R(pe)

Location - If the distance of the fixating point ≫ baseline
choose the location of R(pe) same as that of pe

Size - estimate from maximum range of disparities

② Feature-based Methods

→ Look for a feature in an image that matches a feature in another.

This includes : (i) edge points
(ii) line segments
(iii) corners

A. Line Matching : A line feature descriptor could have:
length: l
orientation: θ
coordinates of midpoint: m
average intensity along the line: i

$$\therefore \quad S = \frac{1}{w_0(\rho_r - \rho_l)^2 + w_1(\theta_l - \theta_r)^2 + w_2(m_l - m_r)^2 + w_3(i_l - i_r)^2}$$

* Intensity - based vs. Feature - based methods

Intensity - based	Feature - based.
1. Provide a dense disparity map	1. Provide a sparse disparity map
2. Sensitive to illumination changes	2. Insensitive to illumination changes
3. Needs textured images to work well.	3. Faster than correlation- based methods.

* Structure from Motion

→ SFM is the process of estimating the 3D structure of a scene from a set of 2D images

→ SFM is used in:

(i) 3D scanning

(ii) augmented reality

(iii) Visual Simultaneous localization and Mapping (VSLAM)

* SFM from Two Views

→ For the case of using 2 stationary cameras or one moving camera, one view = camera1 , and the other is camera2

→ The algorithm assumes that camera1 is at the origin & its optical axis lies along the z-axis.

Steps 1. Point Correspondence

→ Find corresponding either by matching features or tracking points from image1 to image2.

→ Feature tracking techniques such as the Kanade-Lucas-Tomasi (KLT) work well when the cameras are close together.

→ As cameras move further apart, the KLT algo breaks down & feature

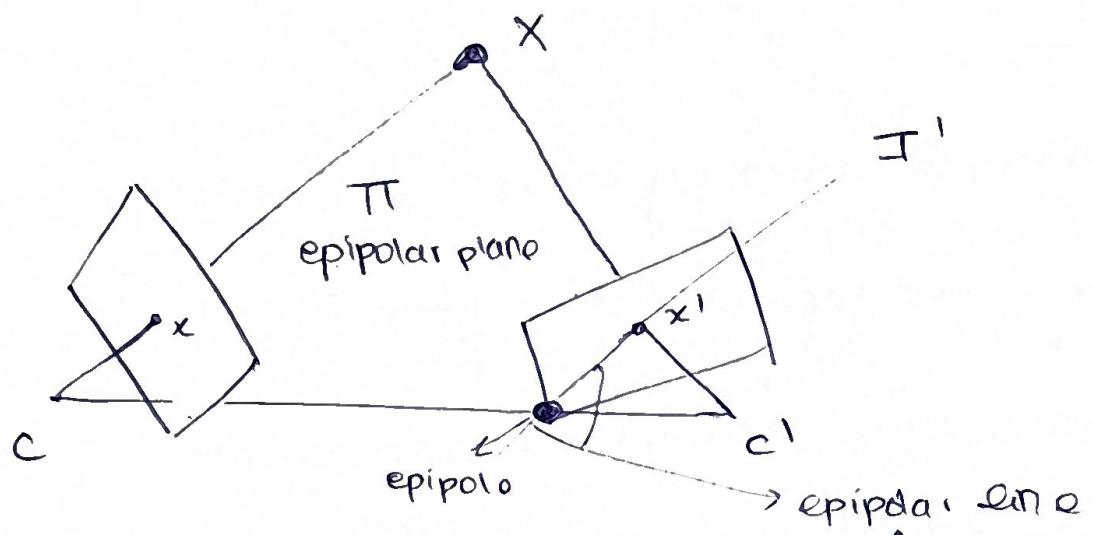
Matching can be used instead.

2. Fundamental Matrix

→ To find the poso of the camera relative to the first camera, the Fundamental matrix must be computed.

→ can be found using the correspondina pts.

→ The fundamental matrix describes the epipolar geometry of the 2 cameras.



→ Say that point X in the 3D-space (viewed as 2 images) is captured as x in the first image and x' in the second.

→ Let C and C' be the camera centers which forms the baseline for the stereo system

→ x, x' and X are coplanar.

→ Rays back-projected from x and x' intersect at X.

→ Consider that only x is known, not x'.

→ The point corresponding to x in the image plane can be searched along / ll^l to I' (the epilor line)

→ The epipole is the point of intersection of the line joining the camera centers with the image plane. The epipolar line is the intersection of an epipolar plane with the image plane. All epipolar lines intersect at the epipao.

$$\begin{bmatrix} x' & y' & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = 0$$

Step3 : Calculate pose

→ Using the fundamental matrix compute the pose of the second camera in the coordinate system of the first camera.

→ Set the distance between the cameras to be 1unit.

* SFM from Multiple Views

→ The approach used for SFM with two views can be extended for multiple views.

→ For multiple views, point correspondences called tracks are needed.

→ Tracks can be computed from pairwise point correspondences

→ Use the approach of SFM from 2 views, find the pose of camera 2 relative to camera 1.

Then find the pose of camera 3 wrt camera 2, and so on.

→ The relative poses must be transformed into a common coordinate system — usually all cameras are calculated wrt camera 1, so that all are in the same coordinate system.

→ The errors increase as the number of views increases — called drift

Some solutions are:

① Non-linear optimization also called bundle adjustment

② Pose graph optimization