3D Localisation of Object Using Multiple Cameras

Garima Singh, *Student*, Geereddy Sathwika Reddy, *Student,* and Ishwar Lal Kumawat, *Student,* under Dr. Sandip Ghosh, *Associate Professor,* Department of Electrical Engineering, Indian Institute of Technology (BHU), Varanasi.

Abstract **- This report presents a study on the 3D localization of objects using multiple cameras. The aim of the study was to develop a robust and accurate method for localizing objects in 3D space using a network of synchronized cameras. The approach involved camera calibration, feature extraction and matching, triangulation, and optimization. The study makes a novel contribution to computer vision by providing a robust and efficient approach to 3D object localization, with potential applications in fields such as robotics, surveillance, and augmented reality.**

1. INTRODUCTION

3D localization is needed in many applications where it is essential to accurately determine the position and orientation of objects in three-dimensional space. In robotics, 3D localization is crucial for object recognition, grasping, and manipulation tasks. In surveillance, it can track the movement of people or objects in 3D space, providing more detailed information about their behavior and interactions. 3D localization is essential in virtual reality to create a convincing and immersive virtual environment. It is also useful in localizing electrical faults in power transmission lines, tracking drones or other aerial vehicles, and mapping indoor environments for energy-efficient building design.

This paper proposes a method for 3D object localization and tracking using multiple cameras. The method works by first detecting and tracking the object in each camera view and then using the corresponding points in multiple views to reconstruct the 3D position of the object. The method consists of two main steps: camera calibration and object tracking.

In the camera calibration step, each camera's intrinsic and extrinsic parameters are determined using a calibration pattern. This step is necessary to estimate the 3D position of the object accurately. In the object tracking step, the object is detected and tracked in each camera view using a combination of color and shape features. After this, coordinates are transformed from one plane to another.

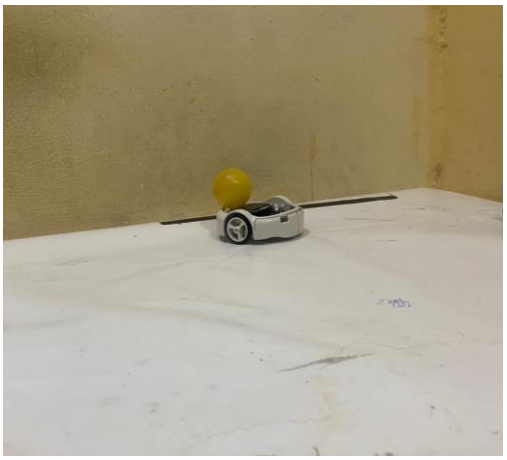

Fig 1. Setup of a yellow ball in the plane

2. CAMERA CALIBRATION

A. INTRINSIC PARAMETERS

The intrinsic properties of a camera are fixed and defined by the manufacturer. These properties include the focal length, sensor size, pixel size, principal point, and lens distortion parameters. They are related to the internal geometry of the camera and are independent of the position and orientation of the camera in the world coordinate system.

The focal length determines the magnification of the camera and is usually given in millimeters. The sensor size and pixel size determine the field of view and resolution of the camera. The principal point is the point on the image sensor where the optical axis of the camera intersects and is usually given in pixel coordinates. The lens distortion parameters account for imperfections in the lens that can cause image distortion, such as radial distortion, tangential distortion, and other types of distortion.

B. EXTRINSIC PARAMETERS

The extrinsic parameters of a camera refer to its position and orientation in the world coordinate system. The extrinsic parameters include:

Rotation matrix (R): This matrix describes the orientation of the camera with respect to the world coordinate system. It is a 3x3 matrix that defines the transformation of the camera coordinate system to the world coordinate system. We can define the matrix by first defining a unit vector that represents the axis of rotation and then applying a rotation about this axis by an angle θ. The resulting rotation matrix will be:

R = [cos(θ) + u_x^2(1-cos(θ)) u_x*u_y*(1-cos(θ)) - u_z*sin(θ) u_x*u_z*(1-cos(θ)) + u_y*sin(θ)]
[u_y*u_x*(1-cos(θ)) + u_z*sin(θ) cos(θ) + u_y^2(1-cos(θ)) u_y*u_z*(1-cos(θ)) - u_x*sin(θ)]
[u_z*u_x*(1-cos(θ)) - u_y*sin(θ) u_z*u_y*(1-cos(θ)) + u_x*sin(θ) cos(θ) + u_z^2(1-cos(θ))]

where u = (u_x, u_y, u_z) is the unit vector representing the rotation axis, and cos(θ) and sin(θ) are the cosine and sine of the rotation angle, respectively.

Translation vector (t): This vector describes the position of the camera with respect to the world coordinate system. It is a 3x1 vector that defines the translation of the camera coordinate system to the world coordinate system. If the position of the camera in the world coordinate system is given by P_cw and the position of the camera coordinate system's origin in the world coordinate system is given by O_cw, then the translation vector t is given by:

$$t = P_{cw} - O_{cw}$$

where $P_{cw} = [X_{cw}, Y_{cw}, Z_{cw}]^T$ and $O_{cw} = [X_{ocw}, Y_{ocw}, Z_{ocw}]^T$

Together, the rotation matrix and translation vector define the transformation matrix that maps points from the camera coordinate system to the world coordinate system. This matrix is typically denoted as [R | t], where "|" represents matrix concatenation.

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

2D Image Coordinates | Intrinsic properties (Optical Centre, scaling) | Extrinsic properties (Camera Rotation and translation) | 3D World Coordinates

Fig 2. 2D to 3D coordinate conversion

3. OBJECT TRACKING

A. OBJECT DETECTION

Object tracking is the process of locating and following a specific object or multiple objects in a video stream over time. The goal of object tracking is to identify and track the movement of an object in a scene, even as it changes position, orientation, size, and appearance. Object tracking has many applications, such as surveillance and security, video analysis, robotics, and self-driving cars.

Object detection is the process of locating instances of objects in an image or video and classifying them into predefined categories. The first step is to acquire the image or video stream using a camera or any other suitable imaging device. The acquired images or video frames may need preprocessing to improve the quality of the image or to extract useful information from the image. This may involve techniques such as color normalization, image enhancement, noise reduction, or image resizing. Once the preprocessing is done, relevant features are extracted from the image. The extracted features should be discriminative enough to distinguish the object of interest from other objects in the image. Now the regions of the image that are likely to contain the object of interest are identified using object proposal techniques. Once the object proposals are generated, the next step is to classify the object of interest into predefined categories.

There are many libraries available that provide pre-trained models for object detection, such as OpenCV, TensorFlow, and PyTorch. OpenCV provides several built-in object detection algorithms that can be used to detect objects in images or videos.

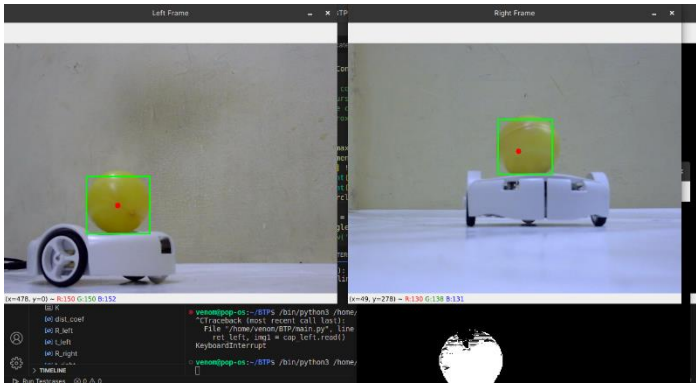

Fig 3. Object Centroid Detection

B. Plane Conversion

Different sensors, cameras, or devices may use different coordinate systems, so it is necessary to convert the coordinates to a common reference frame to perform 3D localization accurately. We need to convert the 2D image coordinates of the point in the camera frame to homogeneous coordinates. We then use the inverse of the camera calibration matrix to convert the homogeneous coordinates from the pixel frame to the normalized camera frame. The camera calibration matrix includes the intrinsic parameters of the camera, such as the focal length and principal point. After that, we use the extrinsic parameters of the camera to perform a coordinate transformation from the camera frame to the world frame. This can be done by multiplying the normalized camera coordinates by the rotation matrix and adding the translation vector. Finally, we convert the resulting 3D homogeneous coordinate to 3D Cartesian coordinates by dividing by the last coordinate.

Once we have the transformation matrix, we can use it to transform a point P from world coordinates to camera coordinates using the following equation:

$$Pc = T * Pw$$

where Pc is the point in camera coordinates, Pw is the point in world coordinates, and T is the 4x4 transformation matrix that describes the position and orientation of the camera.

To perform this transformation, we need to represent the points in homogeneous coordinates by appending a 1 to their coordinates, as follows:

$$Pc = [Xc, Yc, Zc, 1]T$$
$$Pw = [Xw, Yw, Zw, 1]T$$

where Xc, Yc, Zc, Xw, Yw, and Zw are the coordinates of the points in the camera and world coordinates, respectively.

We then multiply the transformation matrix T by the point in world coordinates Pw, as follows:

$$[T11\ T12\ T13\ T14]\ [Xw] = [Xc]$$
$$[T21\ T22\ T23\ T24]\ [Yw] = [Yc]$$
$$[T31\ T32\ T33\ T34]\ [Zw] = [Zc]$$
$$[\ 0\ 0\ 0\ 1\ ]\ [\ 1] = [\ 1]$$

The resulting vector [Xc, Yc, Zc, 1]T is the point in camera coordinates. To convert it back to 3D coordinates, we divide the first three components by the fourth component, as follows:

$$X = Xc / Zc$$
$$Y = Yc / Zc$$
$$Z = 1$$

The resulting vector [X, Y, Z]T is the point in camera coordinates in 3D space.

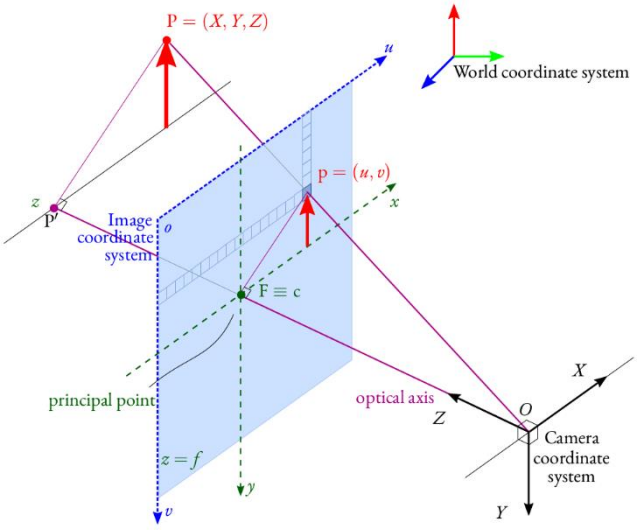

Fig 4. Camera coordinate and World coordinate systems

C. Triangulation

Triangulation is the process of determining the 3D position of a point in space by measuring its projections onto two or more imaging sensors. In 3D localization, triangulation is used to determine the precise location of an object by combining the 2D coordinates of the object in multiple images obtained from different cameras.

The basic principle behind triangulation is to find the intersection point of multiple lines of sight from different

viewpoints. By calculating the intersection point of the lines of sight, we can obtain the 3D position of the object. Firstly, we compute the projection matrices of the cameras. The general form of the projection matrix is given as:

$$P = K[R|t]$$

where P is the 3x4 projection matrix, K is the 3x3 intrinsic matrix, R is the 3x3 rotation matrix representing the orientation of the camera, and t is the 3x1 translation vector representing the position of the camera.

For each pair of cameras that observe the same point of interest, compute the two rays that pass through the corresponding 2D image points and emanate from the camera centers in the direction of the points. We then compute the intersection of the two rays using linear algebra. This intersection point corresponds to the 3D location of the point of interest in the world coordinate system. We repeat it for all pairs of cameras that observe the same point of interest.

4. REFERENCES

[1] T. Svoboda and J. Matas, “Multi-camera tracking and 3D reconstruction of rigid objects”
[2] G. Gallego and A. Gil, “A vision-based multi-camera system for 3D localization of vehicles”
[3] S. Wang, J. Shen, and Y. Wang, “Multi-camera 3D object detection and tracking in warehouse environments”