# Homework 5

EECS/BioE C106A/206A<br>
Introduction to Robotics

Due: October 13, 2020

## Theory

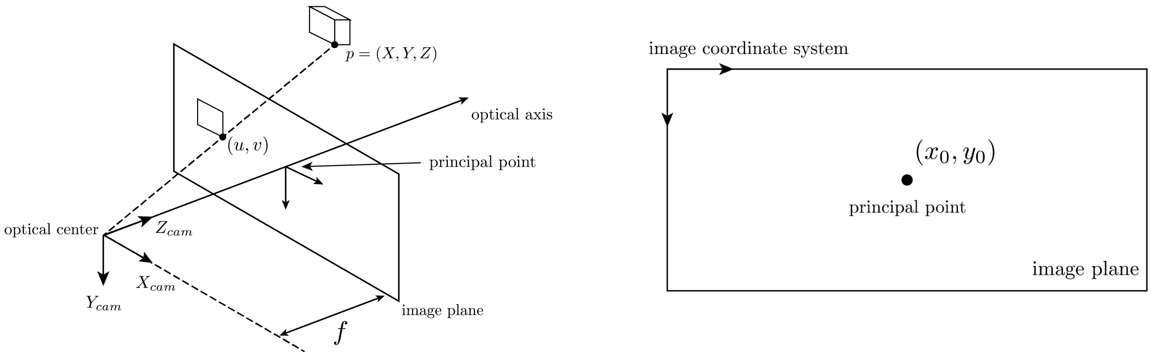

Figure 1: Geometry behind a Pinhole Camera

### Homogeneous coordinates

In class, we have encountered homogeneous coordinates for 3D points, which work by appending a 1 to the end of the 3 vector to get a vector in 4D. When writing the coordinates of a point in the image plane (which is 2 dimensional) we will use *2D homogeneous coordinates.* This means we will represent the point $x = (u, v) \in \mathbb{R}^2$ as $(u, v, 1)$ in homogeneous coordinates. Note that the homogeneous representation of a 2D point is a 3D vector.

### Image formation

We consider a pinhole model of image formation. See Figure 1. We denote the center of the perspective projection (the point in which all the rays intersect) as the optical center or camera center and the line perpendicular to the image plane passing through the optical center as the optical axis. Additionally, the intersection point of the image plane with the optical axis is called the principal point.

We always associate a reference frame with each camera as shown. By convention, we center this reference frame at the optical center, take the $X - Y$ plane of this reference frame to be parallel to the image plane, and take the $Z-$axis to be perpendicular to the image plane, pointing in the direction of viewing. Additionally, there is a 2D reference frame attached to the image plane, with respect to which the "image coordinates" of any point are measured. A discretized version of this reference frame give us the familiar "pixel coordinates" of any points (in columns and rows).

The figure shows a point $p$ with spatial coordinates $\bar{X} = (X, Y, Z)$ in the camera reference frame, and image coordinates $x = (u, v)$. As we stated above, we will default to representing image coordinates in homogeneous form as $x = (u, v, 1)$, and usually we will overload this notation wherever it is obvious if we are using homogeneous or regular coordinates.

The camera parameters (such as the focal length $f$ and others; see Lab 6) are specified in the form of a $3 \times 3$ *camera matrix* $K$. This matrix $K$ is always invertible. The spatial coordinates $\bar{X}$ (in the camera reference frame) and the image (homogeneous) coordinates $x = (u, v, 1)$ of a given point $p$ are related via the $K$ matrix as

$$x = \frac{1}{Z} K \bar{X} \tag{1}$$

where $Z$ is the $Z-$coordinate of the point in the camera reference frame. Observe that this $Z-$coordinate has a significant geometric meaning. It is the distance from the camera's $X - Y$ plane to the point, along the direction of viewing. In other words, it is the "depth" of the point as seen from the camera. So, we give this depth its own symbol $\lambda$ and move it to the LHS to get the less unwieldy expression

$$\lambda x = K \bar{X} \tag{2}$$

Note that given the depth $\lambda$, the camera matrix $K$, and the image coordinates $x$, the spatial coordinates $\bar{X}$ of the point can be recovered by inverting equation (2). On the other hand, without knowing the depth, the spatial coordinates $\bar{X}$ can only be recovered up to a scale factor. This makes geometric sense, since we can see from figure (1) that any point along the line conecting the optical center to $p$ gets projected to the same image coordinates as $p$, and hence knowing only the image coordinates, we can at best specify a line along which $p$ must lie.

**Problem 1. Two-View Triangulation**

Consider two cameras with reference frames $\{1\}$ and $\{2\}$ respectively. As always, the reference frame of each camera is such that the $X-Y$ plane is parallel to the image plane and the $Z$-axis points in the direction of viewing.

Assume we know the relative transform $g_{21} = (R, T) \in SE(3)$. Additionally, assume the cameras are calibrated and normalized, so that the camera matrix $K$ is the identity.

Both cameras are looking at the same point $p$ in 3D space, which has unknown coordinates $X_1 \in \mathbb{R}^3$ in frame $\{1\}$ and $X_2 \in \mathbb{R}^3$ in frame $\{2\}$. We observe their image coordinates $x_1$ and $x_2$, written in 2D homogeneous coordinates.

(a) Write down an expression relating $x_1$ to $X_1$ in terms of an unknown depth $\lambda_1$. Do the same for camera 2.

(b) Write down an expression for $X_2$ in terms of $X_1$.

(c) Find a method for solving for $X_1$ in terms of the known quantities $R \in SO(3), T \in \mathbb{R}^3, x_1, x_2$. Can you deal with the case when the image measurements $x_1, x_2$ are corrupted by some small (white, zero mean, Gaussian) noise?

*Hint: Eliminate $X_1$ and $X_2$ from your expressions, and try to find only the unknown depths $\lambda_1$ and $\lambda_2$. Then, use these depths to recover $X_1$.*

### Problem 2. Epipolar Ambiguities and Structure from Motion

Consider a similar set up as in the previous problem, with two calibrated, normalized cameras, where the transform $g_{21} = (R, T)$ between them is *not* known. Recall that for such a system, we define $E = \hat{T}R \in \mathbb{R}^{3\times 3}$ to be the *essential matrix*. The essential matrix imposes the *epipolar constraint*, which is that whenever $x_1$ and $x_2$ are the (homogeneous) image coordinates of the *same* point, then they must satisfy

$$x_2^T E x_1 = 0$$

Such a pair of image points $x_1, x_2$ that correspond to the same point in 3D space viewed from two different cameras are called *corresponding points*. In this problem, we consider the problem of recovering the relative poses between cameras in a multi-camera setup when we are given a number of corresponding-point pairs.
It turns out that 8 pairs of corresponding points $(x_1^{(1)}, x_2^{(1)}), \cdots (x_1^{(8)}, x_2^{(8)})$ in general position are enough to compute a candidate essential matrix $\tilde{E}$. Each such pair gives us an equation of the form

$$x_2^{(i)^T} E x_1^{(i)} = 0 \tag{3}$$

where the $x$'s are all known. We additionally have the constraint that $E$ should be of the correct form to be written as $\hat{T}R$. i.e. we should be able to write it as the product of a cross product matrix $\in \mathfrak{so}(3)$ and a rotation matrix $\in SO(3)$. We can then solve this system of equations for a nonzero $3 \times 3$ matrix $E$ that satisfies this set of constraints. See chapter 5 from *An Invitation to 3D Vision (Ma, Soatto, Kosecka, Sastry)* for the full details.

(a) Show that we can only recover $E$ up to a scale factor. In particular, show that if $\tilde{E}$ is a matrix that satisfies all the required constraints, then so is $c\tilde{E}$ for any real number $c$.

**Remark:** We can in fact conclude that this ambiguity can be attributed to an unknown scale factor on the translation vector $T$ between the two frames. This means that although we can decompose a computed essential matrix $E$ into rotational and translational components $(R, \tilde{T})$, we can only recover the original translation $T$ up to a scale factor. Typically then, we restrict ourselves to finding a $\tilde{T}$ such that $\|\tilde{T}\| = 1$.

(b) Say we have a system of 3 cameras with reference frames $\{1\}, \{2\}$ and $\{3\}$ respectively, and we are able to recover the transforms $(R_{12}, \tilde{T}_{12}), (R_{23}, \tilde{T}_{23})$ and $(R_{13}, \tilde{T}_{13})$ using point correspondances, where each $\tilde{T}_{ij}$ has norm 1. So there are unknown, nonzero scale factors $\lambda_{ij}$ such that the true $T_{ij} = \lambda_{ij}\tilde{T}_{ij}$. If we could find the three scalars $\lambda_{12}, \lambda_{23}, \lambda_{13}$ then we would have fully recovered the relative poses between the various cameras. Show that in this setting, we can only recover the $\lambda_{ij}$'s up to a single scaling factor.

(c) Consider the same setup as part (c), but now the translation $T_{12}$ is known exactly (i.e. $\lambda_{12}$ is known). Show that now, all $\lambda_{ij}$'s can be recovered and the relative poses between the cameras can be found.