Vision geometry

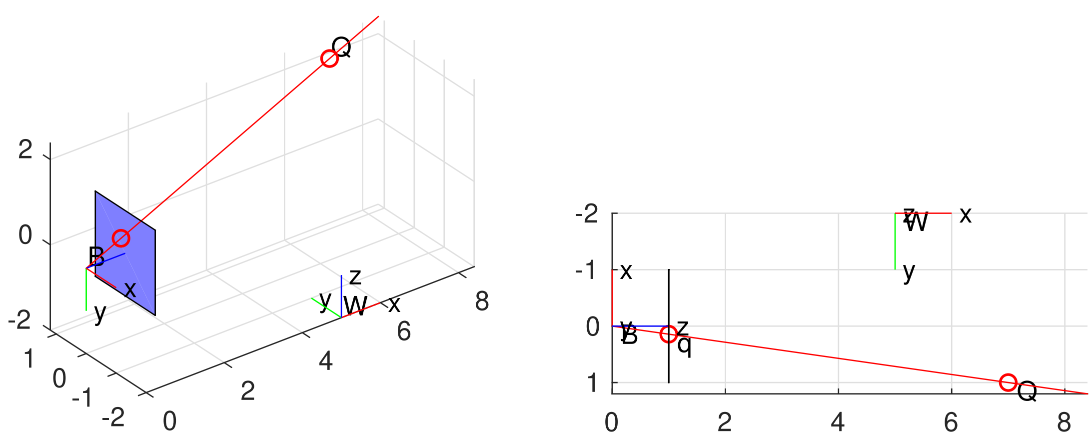
ME 416 - Prof. Tron

Thursday 2nd May, 2019

The perspective camera model is a mathematical description of a 3-D scene projected on an image, that is, a 2-D array of pixels. To begin the explanation, we attach a body frame $\mathcal{B}$ to the camera, and attach a world frame $\mathcal{W}$ to the world. The image plane is a 2-D plane that is defined to be parallel to the xy-plane of $\mathcal{B}$ and is placed at distance f from the origin $O_{\mathcal{B}}$. The value f is called the focal length.

The image formation process model maps a visible 3-D point in the scene, Q, to a 2-D point on the image plane, q. The 2-D point is given by the intersection between a ray joining the point to the origin of the camera ${}^{\mathcal{B}}O$, and the image plane.

See Figure 1 for an illustration of the process. ★★Note that with this conventional choice for the body reference frame, points that are visible to the camera have always positive z coordinates, the x coordinate goes from left to right, and the y coordinate from top to bottom (i.e., it is "upside-down"). The latter two directions are consistent with the coordinate system typically used for images (see the corresponding http://wiki.bu.edu/roboclass/index.php?title=Image_representation_and_color_spaces).



(a) (b)

Figure 1: A 2-D projection q onto an image plane (blue) from a 3-D point Q in space. The focal length, f, is the distance from the origin of the body frame B to the screen. The red line represents the back-projection of the image point q.

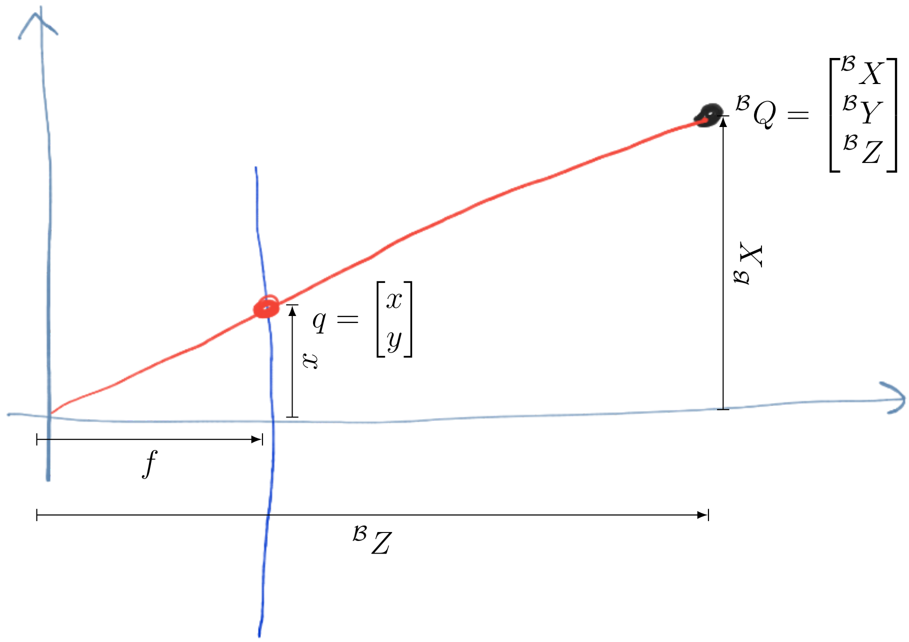


Figure 2: Two-dimensional view of the perspective projection model in camera coordinates (same as Figure 1b, but with annotations)

0.1 Case of aligned body and world frames, point on the vertical axis

To obtain a mathematical expression, we start considering the particular case where the body and world frames are aligned ($\mathcal{W} \equiv \mathcal{B}$). We then consider the projection model restricted to the zx plane, as depicted in Figure 1b.

By the law of similar triangles we have:

$$\frac{y}{f} = \frac{{}^{\mathcal{B}}Y}{{}^{\mathcal{B}}Z}. \tag{1}$$

Rearraging, and writing a similar equation for the y axis, we have:

$$x = f\frac{{}^{\mathcal{B}}X}{{}^{\mathcal{B}}Z}, \qquad y = f\frac{{}^{\mathcal{B}}Y}{{}^{\mathcal{B}}Z}. \tag{2}$$

These two relations can be combined together in a vectorial equation:

$${}^{\mathcal{B}}Z \begin{bmatrix} x \\ y \end{bmatrix} = f \begin{bmatrix} {}^{\mathcal{B}}X \\ {}^{\mathcal{B}}Y \end{bmatrix}. \tag{3}$$

Before generalizing this relation, we introduce the following modifications. First, we define $\lambda = {}^{\mathcal{B}}Z$ to be the *depth* of the point in the camera frame (i.e., the z coordinate). This is the traditional nomenclature used in computer vision. Second, we define the concept of *homogeneous coordinates of a point*, which are obtained by simply appending a “1” at the end of the vector, and are denoted with a bar over the symbol. For instance, in the following we will use:

$$\bar{q} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \qquad {}^{\mathcal{B}}\bar{Q} = \begin{bmatrix} {}^{\mathcal{B}}X \\ {}^{\mathcal{B}}Y \\ {}^{\mathcal{B}}Z \\ 1 \end{bmatrix}. \tag{4}$$

Third, we define the *standard projector matrix*

$$\Pi = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \end{bmatrix}, \tag{5}$$

which can be use to transform 3-D vectors from homogeneous to non-homogeneous coordinates, i.e.:

$${}^{\mathcal{B}}Q = \Pi\, {}^{\mathcal{B}}\bar{Q}. \tag{6}$$

With this, we can rewrite (3) as:

$$\lambda\bar{q} = \lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} {}^{\mathcal{B}}X \\ {}^{\mathcal{B}}Y \\ {}^{\mathcal{B}}Z \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \Pi\, {}^{\mathcal{B}}\bar{Q}. \tag{7}$$

0.2 Introduction of the intrinsic calibration matrix

In the model (7), the images q are expressed in *metric* coordinates (e.g., if ${}^{\mathcal{B}}Q$ is expressed in meters, then also q is expressed in meters), and the point $q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ corresponds to the projection of a point on the z axis of the camera. In practice, however, coordinates for q are expressed in *image* coordinates (pixels). It is then customary to introduce a *calibration matrix* K that transforms from metric to image coordinates, and that usually is taken to have the following form:

$$K = \begin{bmatrix} fs_x & fs_\theta & o_x \\ 0 & fs_y & o_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{8}$$

where:

- s_x and s_y are related to the size (width and height) of the sensor in pixels;
- o_x and o_y represent the position of the z axis of the camera in the image (i.e., the translation of the sensor in the x and y directions of the camera frame);
- s_θ is a *skew* coefficient that accounts for small alignment problems of the sensor with respect to the xy plane of the camera frame;
- f is the focal length as defined previously.

These values are typically called the *intrinsic parameters* of the camera ("intrinsic" because they do not depend on where the camera is positioned with respect to the world reference frame $\mathcal{W}$).

The model (7) is then modified to

$$(\textbf{??})\lambda\bar{q} = K\Pi\,{}^{\mathcal{B}}Q. \tag{9}$$

TODO: explain how the matrix K is usually found or estimated in practice.

0.3 Rigid body transformations in homogeneous coordinates

The model (**??**) is missing one last ingredient: the point Q is expressed in coordinates with respect to the camera frame $\mathcal{B}$ (${}^{\mathcal{B}}Q$) instead of the world reference frame $\mathcal{W}$ (${}^{\mathcal{W}}Q$). The two are related by the rigid body transformation corresponding to the pose of the camera (${}^{\mathcal{B}}R_{\mathcal{W}}, {}^{\mathcal{B}}T_{\mathcal{W}}$):

$${}^{\mathcal{B}}Q = {}^{\mathcal{B}}R_{\mathcal{W}}\,{}^{\mathcal{W}}Q + {}^{\mathcal{B}}T_{\mathcal{W}}. \tag{10}$$

Equation (10) can be written more compactly (with a single matrix-vector multiplication) by using homogeneous coordinates:

$${}^{\mathcal{B}}\bar{Q} = {}^{\mathcal{B}}g_{\mathcal{W}}\,{}^{\mathcal{W}}\bar{Q}, \tag{11}$$

where the *matrix representation of the rigid body transformation in homogeneous coordinates*, ${}^{\mathcal{B}}g_{\mathcal{W}}$, is a 4×4 matrix given by:

$${}^{\mathcal{B}}g_{\mathcal{W}} = \begin{bmatrix} {}^{\mathcal{B}}R_{\mathcal{W}} & {}^{\mathcal{B}}T_{\mathcal{W}} \\ 0\;0\;0 & 1 \end{bmatrix}, \tag{12}$$

that is:

- the upper left 3 block contains the rotation matrix;
- the upper right 3×1 block contains the translation vector;
- the last row is $\begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$.

The transformation (${}^{\mathcal{B}}R_{\mathcal{W}}, {}^{\mathcal{B}}T_{\mathcal{W}}$) is also referred to as the *extrinsic parameters* of the camera (which depend on the pose of the camera in the world frame, but not its internal details).

0.4 Complete model and projection matrices

Combining (12) with (**??**), we have our final, complete model for perspective projection, going from the 3-D coordinates of a point in the world reference frame $\mathcal{W}$ to the 2-D image coordinates of the projection.

$$\lambda\bar{q} = K\Pi\,{}^{\mathcal{B}}g_{\mathcal{W}}\,{}^{\mathcal{W}}\bar{Q}. \tag{13}$$

It is customary to lump together all extrinsic and intrinsic parameters into a single *projection matrix*:

$$P = K\Pi\,{}^{\mathcal{B}}g_{\mathcal{W}}, \tag{14}$$

which allows us to write (13) as:

$$\lambda\bar{q} = P\,{}^{\mathcal{W}}\bar{Q}. \tag{15}$$

As shown here, the matrix P captures all the

This model is "almost" linear, in the sense that it is mainly given by a (linear) matrix-vector multiplication, and the only nonlinear part is the multiplication by λ, which really represents a division by the depth of the point.

1 Triangulation

Model (15) provides a way to go from 3-D coordinates of a point to the corresponding 2-D image coordinates. The process of *triangulation* is concerned with the inverse problem, given the 2-D images coordinates, find the 3-D coordinates of the point.

1.1 Backprojections of points

The first important consideration that needs to be made about triangulation is that it cannot be done with a single image (i.e., the coordinates of a point in a single image plane). In fact, for a given image q, there exist an entire set of points Q that all give the same result under the model (15); this set is called the *back-projection* of the point q. In the example of Figure 1a, this set of points is represented by the red line joining q and Q.

Given a single image q, it is therefore impossible to say which point in its back-projection it really corresponds to. To perform triangulation, it is therefore necessary to use at least two images q_1, q_2 obtained from two different cameras (or the same camera at two different positions).

1.2 Side note: properties of 3×3 skew-symmetric matrices

Given a vector $v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$, we can write a corresponding 3×3 skew-symmetric matrix via the *hat* operator:

$$\hat{v} = \begin{bmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{bmatrix}. \tag{16}$$

This matrix has the following interesting property: given another vector $w \in \mathbb{R}^3$, the cross product $v \times w$ of the two vectors can be written as a multiplication by the matrix obtained by the hat operator: $v \times w = \hat{v}w$. As a consequence, from the properties of the cross product we have:

$$\hat{v}v = 0. \tag{17}$$

1.3 Linear triangulation

Assume we have a series of N images $\{q_i\}$ taken from a series of cameras with projection matrices $\{\P_i\}$. Geometrically, we can solve the triangulation problem, i.e., identify the coordinates of the common point ${}^{\mathcal{W}}Q$, as the point at the intersection of all the back-projections of all the images, see Figure for .

For each one of these, relation (15) holds:

$$P_i {}^{\mathcal{W}}\bar{Q} = \lambda_i \bar{q}_i. \tag{18}$$

Multiplying both sides of these equations by the corresponding matrix $\hat{\bar{q}}_i$, we have

$$\hat{\bar{q}}_i P_i {}^{\mathcal{W}}\bar{Q} = \lambda_i \hat{\bar{q}}_i \bar{q}_i = 0. \tag{19}$$

Equation (19) represents a system of linear equations where the only unknowns are three entries in ${}^{\mathcal{W}}\bar{Q}$, and the known quantities (projection matrices and images) are used to build

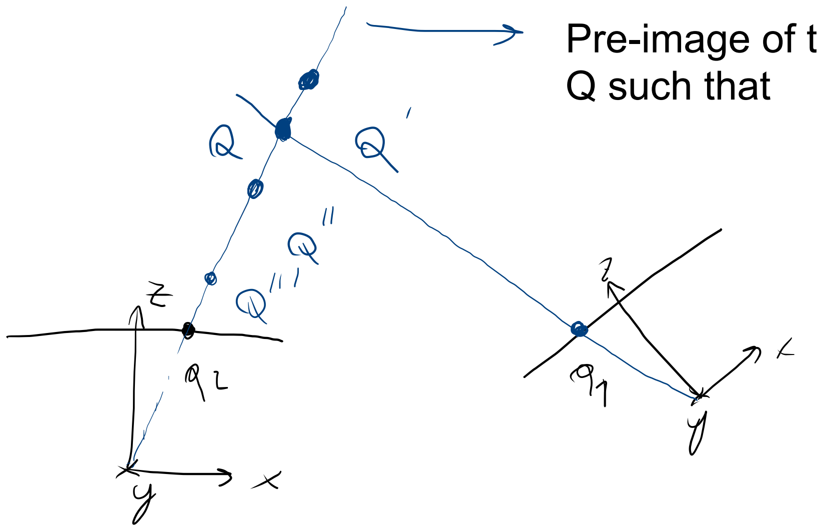


Figure 3: Graphical illustration of triangulation as finding the intersection of back-projections.

the equations. We can stack all the equations from (19) into a single system of the form:

$$\begin{bmatrix} \hat{\bar{q}}_1 P_1 \\ \vdots \\ \hat{\bar{q}}_N P_N \end{bmatrix} {}^{\mathcal{W}}\bar{Q} = A\,{}^{\mathcal{W}}\bar{Q} = 0, \tag{20}$$

where A is a $3N \times 4$.

When $N \geq 2$, we can generally solve the system $A\,{}^{\mathcal{W}}\bar{Q} = 0$ to recover $\bar{Q}$, and hence Q.

Note: the resulting system is *homogeneous*, in the sense that the right hand side is zero; as such, standard solutions will not work.