However, for most computer vision algorithms, it is necessary to know the coordinates of a point in 3D space with respect to the camera reference frame. Thus, it is necessary to know the position and orientation of the camera with respect to the frame $\mathcal{F}_w$. The position and orientation of the camera are given by some affine transformation $(R, \mathbf{T})$ mapping the frame $\mathcal{F}_w$ to the frame $\mathcal{F}_c$, where R is a rotation matrix and $\mathbf{T}$ is a translation vector. Furthermore, the coordinates of an image point are typically known in terms of *pixel coordinates*, and it is also necessary to transform the coordinates of an image point with respect to the camera reference frame to pixel coordinates. In summary, it is necessary to know the transformation that maps a point P in world coordinates (w.r.t. $\mathcal{F}_w$) to pixel coordinates.

This transformation of world coordinates to pixel coordinates turns out to be a projective transformation that depends on the extrinsic and the intrinsic parameters of the camera. The *extrinsic parameters* of a camera are the location and orientation of the camera with respect to the world reference frame $\mathcal{F}_w$. It is given by an affine map (in fact, a rigid motion, see Chapter 13, Section 27.2). The *intrinsic parameters* of a camera are the parameters needed to link the pixel coordinates of an image point to the corresponding coordinates in the camera reference frame. If $\mathbf{P_w} = (X_w, Y_w, Z_w)$ and $\mathbf{P_c} = (X_c, Y_c, Z_c)$ are the coordinates of the 3D point P with respect to the frames $\mathcal{F}_w$ and $\mathcal{F}_c$, respectively, we can write

$$\mathbf{P_c} = R(\mathbf{P_w} - \mathbf{T}).$$

Neglecting distorsions possibly introduced by the optics, the correspondence between the coordinates (x, y) of the image point with respect to $\mathcal{F}_c$ and the pixel coordinates $(x_{\text{im}}, y_{\text{im}})$ is given by

$$\begin{aligned} x &= -(x_{\text{im}} - o_x)s_x, \\ y &= -(y_{\text{im}} - o_y)s_y, \end{aligned}$$

where (o_x, o_y) are the pixel coordinates the principal point $\mathbf{o}$ and s_x, s_y are scaling parameters.

After some simple calculations, the upshot of all this is that the transformation between the homogeneous coordinates $(X_w, Y_w, Z_w, 1)$ of a 3D point and its homogeneous pixel coordinates (x_1, x_2, x_3) is given by

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = M \begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix}$$

where the matrix M, known as the *projection matrix*, is a 3×4 matrix depending on R, $\mathbf{T}$, o_x, o_y, f (the focal length), and s_x, s_y (for the derivation of this equation, see Trucco and Verri [178], Chapter 2).

The problem of estimating the extrinsic and the instrinsic parameters of a camera is known as the *camera calibration* problem. It is an important problem in computer vision.