Chapter 8
Robot Vision

The task of robot vision is to recognize the geometry of the robot workspace from a digital image. It is our aim to find the relation between the coordinates of a point in the two-dimensional (2D) image and the coordinates of the point in the real three-dimensional (3D) robot environment.

8.1 System Configuration

The robot vision system is based on the use of one, two or more cameras. If several cameras are used to observe the same object, information about the depth of the object can be derived. In such case, we talk about 3D or stereo vision. Of course, the 3D view can also be achieved with a single camera if two images of the object are available, captured from different poses. If only one image is available, the depth can be estimated based on some previously known geometric properties of the object.

When analyzing the configuration of the robotic vision system, it is necessary to distinguish between possible placements of the cameras. The cameras can be placed in a fixed configuration, where they are rigidly mounted in the workcell, or in a mobile configuration, where the camera is attached to a robot. In the first configuration, the camera observes objects from a fixed position with respect to the robot base coordinate frame. The field of view of the camera does not change during the execution of the task, which means that basically the accuracy of the measurement is constant. In some tasks, it is difficult to prevent the manipulator from reaching into the field of view of the camera and thereby occluding the objects. Therefore, in such case, it is necessary to put a camera on a robot (in a mobile configuration).

The camera can be attached before or after the robot wrist. In the first case, the camera observes the situation from a favorable position and the manipulator generally does not occlude its field of view. In the second case, the camera is attached to the robot end-effector and typically only observes the object that is being manip-

M. Mihelj et al., *Robotics*, https://doi.org/10.1007/978-3-319-72911-4_8

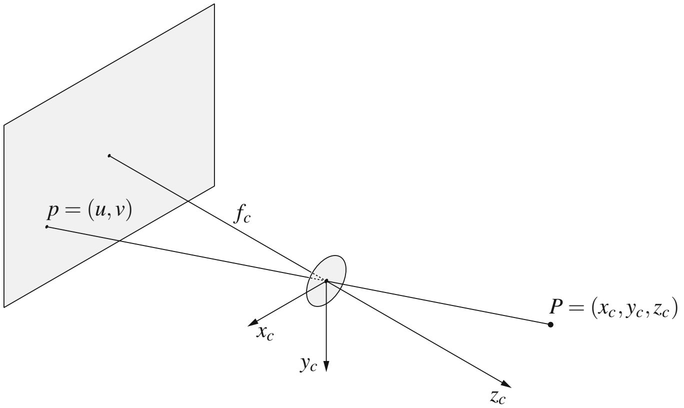

Fig. 8.1 Perspective projection

ulated. In both cases, the field of view of the camera changes with movements of the manipulator. When the manipulator approaches the object, the accuracy of the measurement typically increases.

8.2 Forward Projection

The basic equations of optics determine the position of a point in the image plane with respect to the corresponding point in 3D space (Fig. 8.1). We will therefore find the geometrical relation between the coordinates of the point $P = (x_c, y_c, z_c)$ in space and the coordinates of the point $p = (u, v)$ in the image.

As the aperture of the camera lenses, through which the light falls onto the image plane, is small compared to the size of the objects manipulated by the robot, we can replace the lenses in our mathematical model by a simple pinhole. In perspective projection points from space are projected onto the image plane by lines intersecting in a common point called the center of projection. When replacing a real camera with a pinhole camera, the center of projection is located in the center of the lenses.

When studying robot geometry and kinematics, we attached a coordinate frame to each rigid body (e.g., to robot segments or to objects manipulated by the robot). When considering robot vision, the camera itself represents a rigid body and a coordinate frame should be assigned to it. The pose of the camera will be from now on described by a corresponding coordinate frame. The z_c axis of the camera frame is directed along the optical axis, while the origin of the frame is positioned at the center of projection. We shall choose a right-handed frame where the x_c axis is parallel to the rows of the imaging sensor and the y_c axis is parallel with its columns.

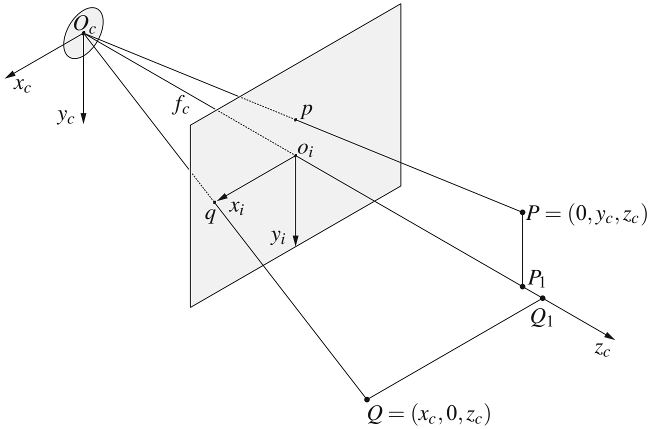

Fig. 8.2 Equivalent image plane

The image plane is in the camera, which is placed behind the center of projection. The distance f_c between the image and the center of projection is called the focal length. In the camera frame the focal length has a negative value, as the image plane intercepts the negative z_c axis. It is more convenient to use the equivalent image plane placed at a positive z_c value (Fig. 8.2). The equivalent image plane and the real image plane are symmetrical with respect to the origin of the camera frame. The geometrical properties of the objects are equivalent in both planes and differ only in the sign.

From now on we shall call the equivalent image plane simply the image plane. Also the image plane can be considered as a rigid body to which a coordinate frame should be attached. The origin of this frame is placed in the intersection of the optical axis with the image plane. The x_i and y_i axes are parallel to the x_c and y_c axes of the camera frame.

In this way the camera has two coordinate frames, the camera frame and the image frame. Let the point P be expressed in the camera frame, while the point p represents its projection onto the image plane. It is our aim to find the relations between the coordinates of the point P and the coordinates of its image p.

Let us first assume that the point P is located in the y_c–z_c plane of the camera frame. Its coordinates are

$$P = \begin{bmatrix} 0 \\ y_c \\ z_c \end{bmatrix}. \tag{8.1}$$

The projected point p is in this case located in the y_i axis of the image plane

$$p = \begin{bmatrix} 0 \\ y_i \end{bmatrix}. \tag{8.2}$$

Because of similarity of the triangles PP_1O_c and poO_c we can write

$$\frac{y_c}{y_i} = \frac{z_c}{f_c}$$

or

$$y_i = f_c \frac{y_c}{z_c}. \tag{8.3}$$

Let us consider also the point Q laying in the x_c–z_c plane of the camera frame. After the perspective projection of the point Q, its image q falls onto the x_i axis of the image frame. Because of similar triangles QQ_1O_c and qoO_c we have

$$\frac{x_c}{x_i} = \frac{z_c}{f_c}$$

or

$$x_i = f_c \frac{x_c}{z_c}. \tag{8.4}$$

In this way we obtained the relation between the coordinates (x_c, y_c, z_c), of the point P in the camera frame and the coordinates (x_i, y_i), of the point p in the image plane. Equations (8.3) and (8.4) represent the mathematical description of the perspective projection from a 3D onto a 2D space. Both equations can be written in the form of perspective matrix equation

$$s \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} f_c & 0 & 0 & 0 \\ 0 & f_c & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix}. \tag{8.5}$$

In Eq. (8.5) s is a scaling factor, while (x_i, y_i) are the coordinates of the projected point in the image frame and (x_c, y_c, z_c) are the coordinates of the original point in the camera frame.

From the matrix Eq. (8.5) it is not difficult to realize that we can uniquely determine the coordinates (x_i, y_i) and the scaling factor s when knowing (x_c, y_c, z_c). On the contrary, we cannot calculate the coordinates (x_c, y_c, z_c) in the camera frame when only the coordinates (x_i, y_i) in the image frame are known, but not the scaling factor. Equation (8.5) represents the forward projection in robot vision. The calculation of (x_c, y_c, z_c) from (x_i, y_i) is called backward projection. When using a single camera and without a priori information about the size of the objects in the robot environment, a unique solution of the inverse problem cannot be found.

For ease of programming it is more convenient to use indices, marking the position of a pixel (i.e., the smallest element of a digital image) in a 2D image instead of metric units along the x_i and y_i axes of the image frame. We shall use two indices which we shall call index coordinates of a pixel (Fig. 8.3). These are the row index and the column index. In the memory storing the digital image the row index runs from the

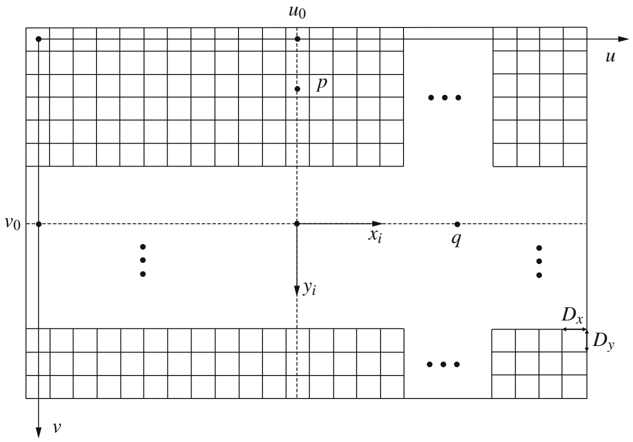

Fig. 8.3 The image plane and the index coordinate frame

top of the image to the bottom while the column index starts at the left and stops at the right edge of the image. We shall use the u axis for the column indices and the v axis for the row indices. In this way the index coordinate frame u–v belongs to each particular image. The upper left pixel is denoted either by (0, 0), or (1, 1). The index coordinates have no measuring units.

In the further text we shall find the relation between the image coordinates (x_i, y_i) and the index coordinates (u, v). Let us assume that the digital image was obtained as a direct output from the image sensor (A/D conversion was performed at the output of the image sensor). In this case each pixel corresponds to a particular element of the image sensor. We shall assume that the area of the image sensor is rectangular.

The origin of the image frame is in the point (u_0, v_0) of the index frame. The size of a pixel is represented by the pair (D_x, D_y). The relation between the image frame x_i–y_i and the index frame u–v is described by the following two equations

$$\begin{aligned} \frac{x_i}{D_x} &= u - u_0 \\ \frac{y_i}{D_y} &= v - v_0. \end{aligned} \tag{8.6}$$

Equations (8.6) can be rewritten as

$$\begin{aligned} u &= u_0 + \frac{x_i}{D_x} \\ v &= v_0 + \frac{y_i}{D_y}. \end{aligned} \tag{8.7}$$

In Eq. (8.7), $\frac{x_i}{D_x}$ and $\frac{y_i}{D_y}$ represent the number of digital conversions along the row and column respectively. Equation (8.7) can be rewritten in the following matrix form

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{D_x} & 0 & u_0 \\ 0 & \frac{1}{D_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}. \tag{8.8}$$

Using the pinhole camera model, we can now combine Eq. (8.5), relating the image coordinates to the camera coordinates, and Eq. (8.8), describing the relation between the image and index coordinates

$$\begin{aligned} s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} &= \begin{bmatrix} \frac{1}{D_x} & 0 & u_0 \\ 0 & \frac{1}{D_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_c & 0 & 0 & 0 \\ 0 & f_c & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \\ &= \begin{bmatrix} \frac{f_c}{D_x} & 0 & u_0 & 0 \\ 0 & \frac{f_c}{D_y} & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix}. \end{aligned} \tag{8.9}$$

The above matrix can be written also in the following form

$$\mathbf{P} = \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{8.10}$$

The **P** matrix represents the perspective projection from the camera frame into the corresponding index coordinate frame. The variables

$$\begin{aligned} f_x &= \frac{f_c}{D_x} \\ f_y &= \frac{f_c}{D_y} \end{aligned} \tag{8.11}$$

are the focal lengths of the camera along the x_c and y_c axes. The parameters f_x, f_y, u_0, and v_0 are called the intrinsic parameters of a camera.

In general the intrinsic parameters of the camera are not known. The specifications of the camera and the lenses are not sufficiently accurate. The intrinsic parameters of the camera are therefore obtained through the camera calibration process. When knowing the intrinsic parameters of the camera we can uniquely calculate the index coordinates (u, v) from the given coordinates (x_c, y_c, z_c). The coordinates (x_c, y_c, z_c) cannot be determined from the known (u, v) coordinates without knowing the scaling factor.

8.3 Backward Projection

The digital image is represented by a matrix of pixels. As the index coordinates (u, v) do not have measuring units, this means that characteristic features of the image are described more qualitatively than quantitatively. If we wish to express the distances in metric units, we must know the relation between the index coordinates (u, v) and the coordinates (x_r, y_r, z_r) in the 3D reference frame. Without knowing the real dimensions or the geometry of the scene it is impossible to recognize the features of the image.

8.3.1 Single Camera

Let us assume that we have a robot vision system with a single camera. The system has the image of the robot workspace as the input and is required to reproduce geometrical measurements as its output. The necessary transformations between the coordinate frames are evident from Fig. 8.4.

Let us suppose that we are now in a position to recognize the point q in the image. It is our aim to determine the coordinates of the real point Q from the coordinates of its image q. This is the problem of backward projection. In order to solve the problem, we must know how the coordinates of the point q are related to the coordinates of the real point Q in the reference frame, which is the problem of forward projection.

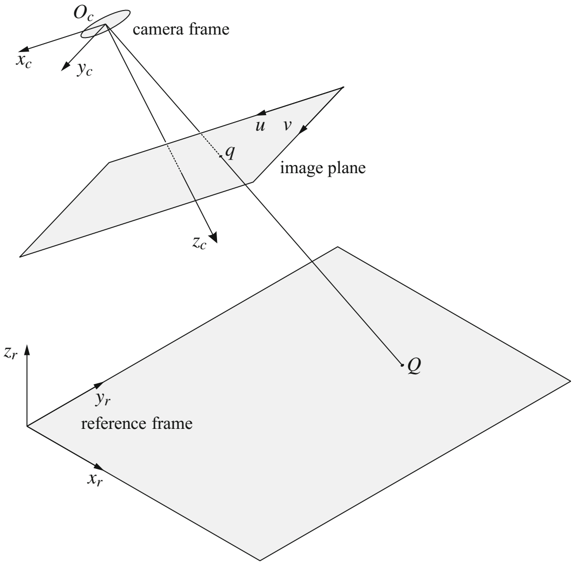

Fig. 8.4 The coordinate frames in a robot vision system

Let us solve first the problem of forward projection. The point Q is given by the coordinates (x_r, y_r, z_r) in the reference coordinate frame. We wish to determine the coordinates of its image $q = (u, v)$, expressed in the index frame. The frame x_c–y_c–z_c is attached to the camera. The matrix $\mathbf{M}$ represents the transformation from the reference into the camera frame

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \mathbf{M} \begin{bmatrix} x_r \\ y_r \\ z_r \\ 1 \end{bmatrix}. \tag{8.12}$$

By combining Eqs. (8.12) and (8.9), we obtain

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{PM} \begin{bmatrix} x_r \\ y_r \\ z_r \\ 1 \end{bmatrix}. \tag{8.13}$$

The relation (8.13) describes the forward projection. The elements of the $\mathbf{P}$ matrix are the intrinsic parameters of the camera, while the elements of the $\mathbf{M}$ matrix represent its extrinsic parameters. The 3×4 matrix

$$\mathbf{H} = \mathbf{PM} \tag{8.14}$$

is called the calibration matrix of the camera. It is used in the calibration process in order to determine both the intrinsic and extrinsic parameters of the camera.

In the further text we shall consider backward projection. It is our aim to determine the coordinates (x_r, y_r, z_r) of the real point Q from the known coordinates of the image point (u, v) and the calibration matrix $\mathbf{H}$. The scaling factor s is not known. In (8.13) we have four unknowns s, x_r, y_r, and z_r and only three equations for a single point in space.

Let us try with three points A, B, and C (Fig. 8.5). We know the distances between these three points. Their coordinates in the reference frame are

$$\left\{(x_{r_j}, y_{r_j}, z_{r_j}), \quad j = 1, 2, 3\right\}.$$

The coordinates of the corresponding image points are

$$\left\{(u_j, v_j), \quad j = 1, 2, 3\right\}.$$

The forward projection can be written in the following form

$$s_j \begin{bmatrix} u_j \\ v_j \\ 1 \end{bmatrix} = \mathbf{H} \begin{bmatrix} x_{r_j} \\ y_{r_j} \\ z_{r_j} \\ 1 \end{bmatrix}. \tag{8.15}$$

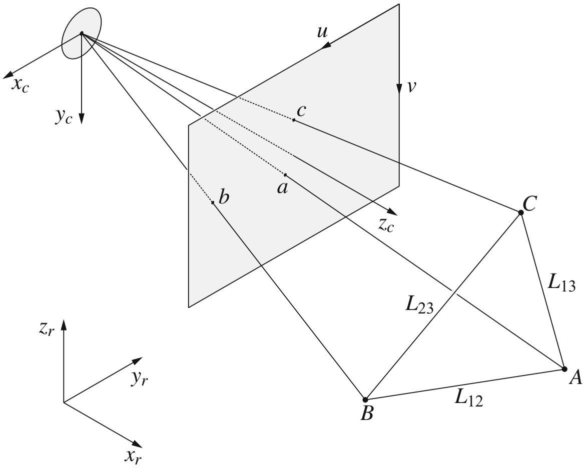

Fig. 8.5 Example of projecting three points

In Eq. (8.15) we have 12 unknowns and 9 equations. To solve the problem we need additional three equations. These equations can be obtained from the size of the triangle represented by the points A, B, and C. We shall denote the triangle sides AB, BC, and CA as the lengths L_{12}, L_{23}, and L_{31}

$$\begin{aligned} L_{12}^2 &= (x_{r_1} - x_{r_2})^2 + (y_{r_1} - y_{r_2})^2 + (z_{r_1} - z_{r_2})^2 \\ L_{23}^2 &= (x_{r_2} - x_{r_3})^2 + (y_{r_2} - y_{r_3})^2 + (z_{r_2} - z_{r_3})^2 \\ L_{31}^2 &= (x_{r_3} - x_{r_1})^2 + (y_{r_3} - y_{r_1})^2 + (z_{r_3} - z_{r_1})^2. \end{aligned} \tag{8.16}$$

Now we have twelve equations for the twelve unknowns. Thus, the solution of the inverse problem exists. It is inconvenient that the last three equations are nonlinear, requiring a computer for numerical solving of the equations. The approach is called model based backward projection.

8.3.2 Stereo Vision

Since the model of the observed object is usually not available or the object changes with time, other solutions to the backward projection problem need to be found. One possible solution is the use of stereo vision: sensing based on two cameras. The principle is similar to human visual perception where the images seen by the left and right eyes differ slightly due to parallax and the brain uses the differences between images to determine the distance to the observed object.

For simplicity we will assume two parallel cameras that observe point Q as shown in Fig. 8.6. Point Q is projected onto the image plane of the left and right cameras. The left camera's image plane contains projection q_l with coordinates $x_{i,l}$ and $y_{i,l}$

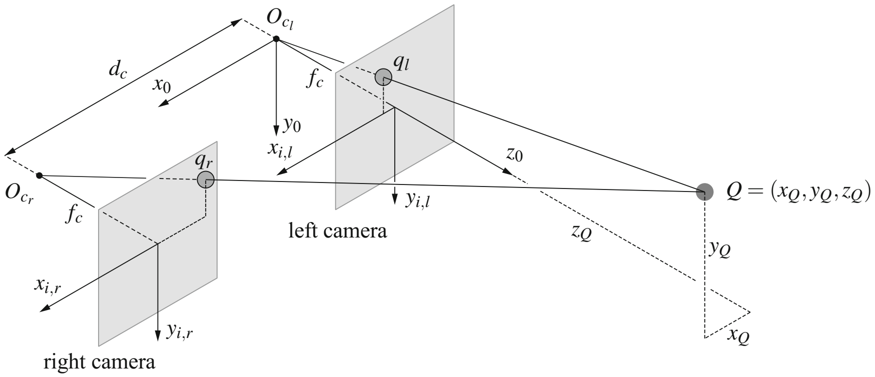

Fig. 8.6 Stereo view of point Q using two parallel cameras

while the right camera's image plane contains projection q_r with coordinates $x_{i,r}$ and $y_{i,r}$. The axes of the vision system coordinate frame x_0–y_0–z_0 have the same directions as the left camera's coordinate frame.

Figure 8.7a shows the top view, while Fig. 8.7b shows the side view of the situation in Fig. 8.6. These views will help us calculate the coordinates of point Q. From the geometry in Fig. 8.7a we can extract the following relations (distances x_Q, y_Q, and z_Q are with regard to the coordinate frame x_0–y_0–z_0)

$$\begin{aligned}\frac{z_Q}{f_c} &= \frac{x_Q}{x_{i,l}}\\ \frac{z_Q}{f_c} &= \frac{x_Q-d_c}{x_{i,r}},\end{aligned} \tag{8.17}$$

where d_c is the distance between the cameras. From the first equation in (8.17) we express

$$x_Q = \frac{x_{i,l}}{f_c} z_Q \tag{8.18}$$

and insert into the second equation to get

$$\frac{x_{i,l}z_Q}{x_{i,r}f_c} - \frac{z_Q}{f_c} = \frac{d_c}{x_{i,r}}. \tag{8.19}$$

We can then determine the distance z_Q to point Q as

$$z_Q = \frac{f_c d_c}{x_{i,l}-x_{i,r}}. \tag{8.20}$$

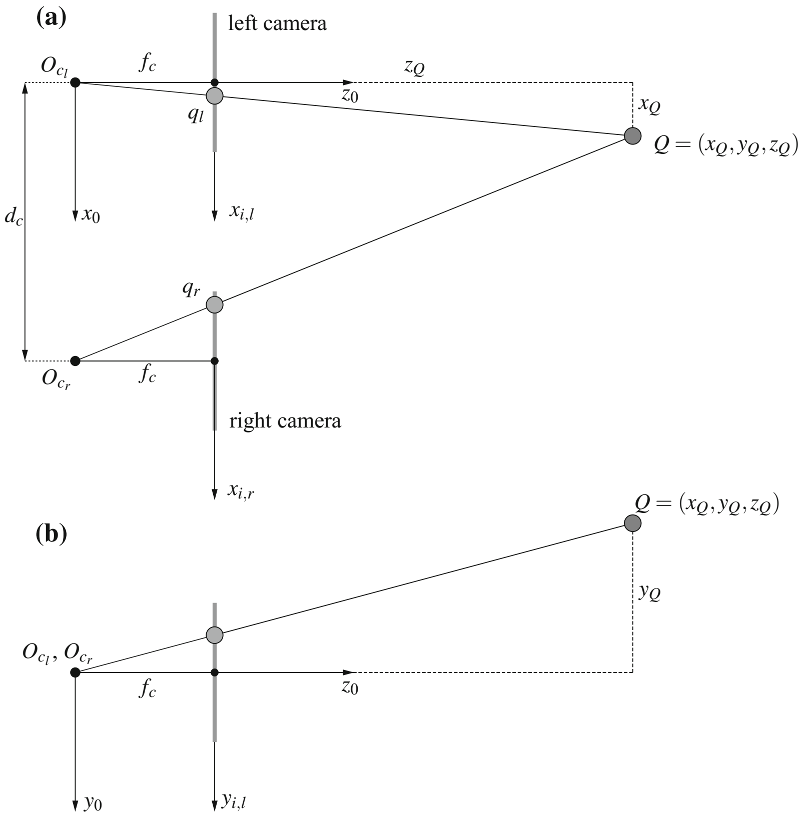

Fig. 8.7 Projections of point Q on the planes of the left and right cameras. The upper figure **a** shows a view of both cameras from above, while the lower figure **b** shows a side view of the cameras

The distance x_Q can be determined from Eq. (8.18). To determine distance y_Q we refer to Fig. 8.7b. From the geometry we can extract relation

$$\frac{z_Q}{f_c} = \frac{y_Q}{y_{i,l}}, \tag{8.21}$$

allowing us to calculate the remaining coordinate

$$y_Q = \frac{y_{i,l}}{f_c} z_Q. \tag{8.22}$$

Use of two cameras enables computation of the position (and orientation) of an object in space without an accurate model of the object.

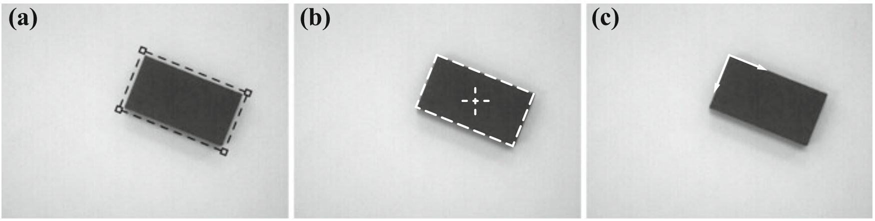

Fig. 8.8 **a** Model definition, **b** recognized object's features, **c** located object

8.4 Image Processing

In contrast to most other sensory systems, visual systems provide very reach information, which requires complex processing algorithms before it can be used for robot control. The goal of image processing is to obtain numerical information from the image, which provides a robust description of the object in the scene. An example of the result of image processing is shown in Fig. 8.8. An object is first identified in the scene and then its pose is determined as marked with the coordinate frame.

Image processing is beyond the scope of this book and it will not be specifically addressed here.

8.5 Object Pose from Image

In order to control the robot relative to the object of interest, the object pose needs to be defined relative to the robot coordinate frame x–y–z. As shown in Fig. 8.8, the pose of the object is known in the image coordinate frame after the image processing. In order to determine its pose in the robot frame, the transformation between the image and the robot coordinate frame must be defined, which is the result of camera calibration. Figure 8.9 presents a simple approach for the calibration problem, where the image plane is parallel to the horizontal plane. For simplicity, the image frame x_i–y_i–z_i is located at the same point as the index frame u–v (the z_i axis was added to the image frame to emphasize the rotation around the vertical axis).

8.5.1 Camera Calibration

Camera is mounted in a fixed position over the robot workspace. The calibration is performed with the calibration pattern (checkerboard), and the calibration tip at the robot end-effector. The calibration pattern can be augmented with a fiducial marker, which appears in the image for use as a point of reference or a measure. The goal of the calibration procedure is to find the transformation matrix $\mathbf{H}_i$ between the image

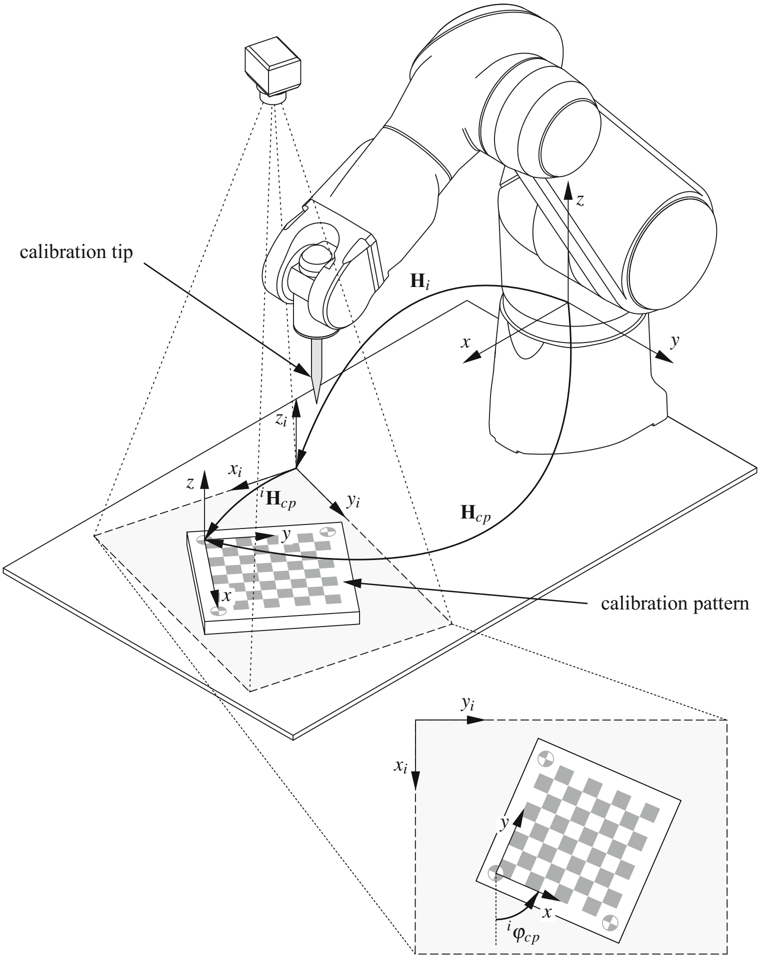

Fig. 8.9 Transformations used for camera calibration

and the robot coordinate frames x_i–y_i–z_i and x–y–z. Based on relations in Fig. 8.9 the following equality can be written

$$\mathbf{H}_{cp} = \mathbf{H}_i {}^i\mathbf{H}_{cp}, \tag{8.23}$$

where $\mathbf{H}_{cp}$ and ${}^i\mathbf{H}_{cp}$ are the poses of the calibration pattern expressed in the robot and the image coordinate frame, respectively.

The pose of the calibration pattern ${}^{i}\mathbf{H}_{cp}$ expressed in the image coordinate frame x_i–y_i–z_i is the result of image processing

$$
{}^{i}\mathbf{H}_{cp} = \begin{bmatrix} \cos {}^{i}\varphi_{cp} & -\sin {}^{i}\varphi_{cp} & 0 & {}^{i}x_{cp} \\ \sin {}^{i}\varphi_{cp} & \cos {}^{i}\varphi_{cp} & 0 & {}^{i}y_{cp} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{8.24}
$$

where ${}^{i}\varphi_{cp}$ and $({}^{i}x_{cp}, {}^{i}y_{cp})$ are the orientation and position of the calibration pattern relative to the image plane, respectively. Position is expressed in metric units as

$$
\begin{bmatrix} {}^{i}x_{cp} \\ {}^{i}y_{cp} \end{bmatrix} = \lambda \begin{bmatrix} u_{cp} \\ v_{cp} \end{bmatrix}, \tag{8.25}
$$

where (u_{cp}, v_{cp}) are the calibration pattern origin coordinates in pixels and λ is the ratio between position expressed in metric units and pixels on the image (the ratio can be obtained from the calibration pattern with the known size of black and white fields). Matrix ${}^{i}\mathbf{H}_{cp}$ represents a rotation around the z_i axis and translation along x_i and y_i axes of the image coordinate frame.

The pose of the calibration pattern $\mathbf{H}_{cp}$ expressed in the robot coordinate frame x–y–z can be determined with the calibration tip at the robot end-effector and the calibration points marked on the calibration pattern. By placing the calibration tip on the calibration point, recording the robot end-effector coordinates and repeating the procedure for the three calibration points, a set of coordinates is obtained that enables the definition of the calibration pattern pose relative to the robot coordinate frame as

$$
\mathbf{H}_{cp} = \begin{bmatrix} \cos \varphi_{cp} & -\sin \varphi_{cp} & 0 & x_{cp} \\ \sin \varphi_{cp} & \cos \varphi_{cp} & 0 & y_{cp} \\ 0 & 0 & 1 & z_{cp} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{8.26}
$$

where φ_{cp} and (x_{cp}, y_{cp}, z_{cp}) are the orientation and position of the calibration pattern relative to the robot frame, respectively.

From Eqs. (8.23), (8.24) and (8.26) the transformation matrix between the image and the robot coordinate frames can be obtained as

$$
\mathbf{H}_i = \mathbf{H}_{cp} {}^{i}\mathbf{H}_{cp}^{-1}. \tag{8.27}
$$

8.5.2 Object Pose

With the known $\mathbf{H}_i$, the object pose $\mathbf{H}_o$ relative to the robot coordinate frame can be determined as shown in Fig. 8.10.

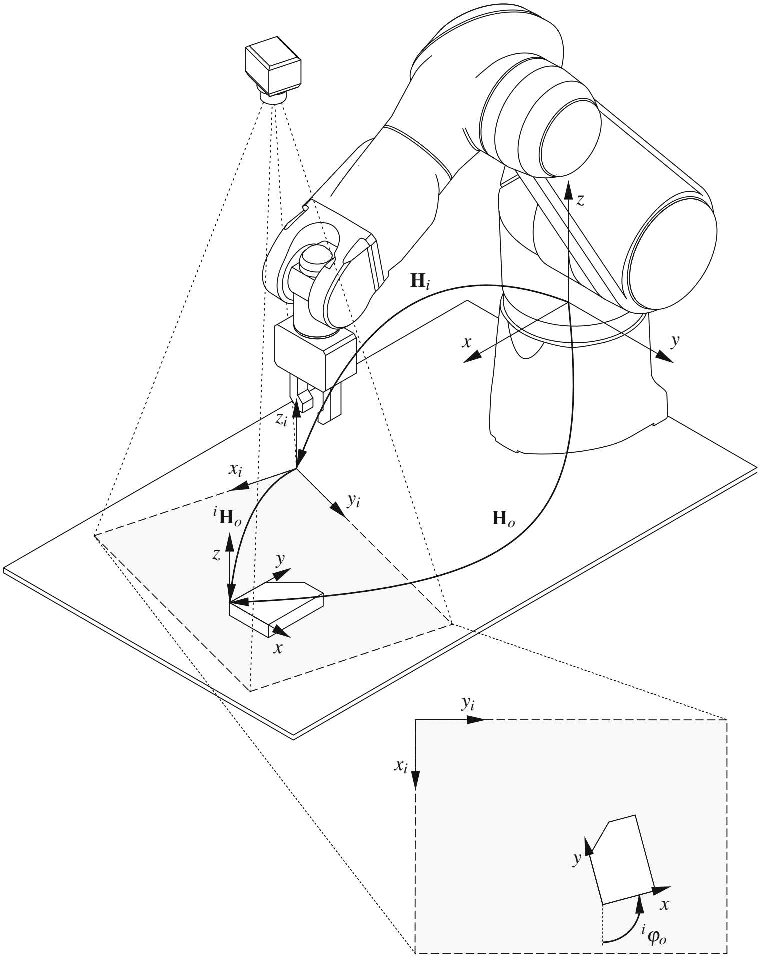

Fig. 8.10 Transformations used for object pose computation

The pose of the object ${}^i\mathbf{H}_o$ expressed in the image coordinate frame x_i–y_i–z_i is the result of image processing

$$
{}^i\mathbf{H}_o = \begin{bmatrix} \cos{}^i\varphi_o & -\sin{}^i\varphi_o & 0 & {}^ix_o \\ \sin{}^i\varphi_o & \cos{}^i\varphi_o & 0 & {}^iy_o \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{8.28}
$$

where ${}^{i}\varphi_o$ and $({}^{i}x_o, {}^{i}y_o)$ are the orientation and position of the object relative to the image plane, respectively. Position is expressed in metric units as

$$\begin{bmatrix} {}^{i}x_o \\ {}^{i}y_o \end{bmatrix} = \lambda \begin{bmatrix} u_o \\ v_o \end{bmatrix}, \tag{8.29}$$

where (u_o, v_o) are the object origin coordinates in pixels.

Finally, $\mathbf{H}_o$ can be determined as

$$\mathbf{H}_o = \mathbf{H}_i {}^{i}\mathbf{H}_o. \tag{8.30}$$