4 Perception

One of the most important tasks of an autonomous system of any kind is to acquire knowledge about its environment. This is done by taking measurements using various sensors and then extracting meaningful information from those measurements.

In this chapter we present the most common sensors used in mobile robots and then discuss strategies for extracting information from the sensors. For more detailed information about many of the sensors used on mobile robots, refer to H.R. Everett's comprehensive book *Sensors for Mobile Robots* [20].

4.1 Sensors for Mobile Robots

A wide variety of sensors is used in mobile robots (figure 4.1). Some sensors are used to measure simple values such as the internal temperature of a robot's electronics or the rotational speed of the motors. Other more sophisticated sensors can be used to acquire information about the robot's environment or even to measure directly a robot's global position. In this chapter we focus primarily on sensors used to extract information about the robot's environment. Because a mobile robot moves around, it will frequently encounter unforeseen environmental characteristics, and therefore such sensing is particularly critical. We begin with a functional classification of sensors. Then, after presenting basic tools for describing a sensor's performance, we proceed to describe selected sensors in detail.

4.1.1 Sensor classification

We classify sensors using two important functional axes: *proprioceptive/exteroceptive* and *passive/active*.

Proprioceptive sensors measure values internal to the system (robot), for example, motor speed, wheel load, robot arm joint angles, and battery voltage.

Exteroceptive sensors acquire information from the robot's environment, for example, distance measurements, light intensity, and sound amplitude. Hence exteroceptive sensor measurements are interpreted by the robot in order to extract meaningful environmental features.

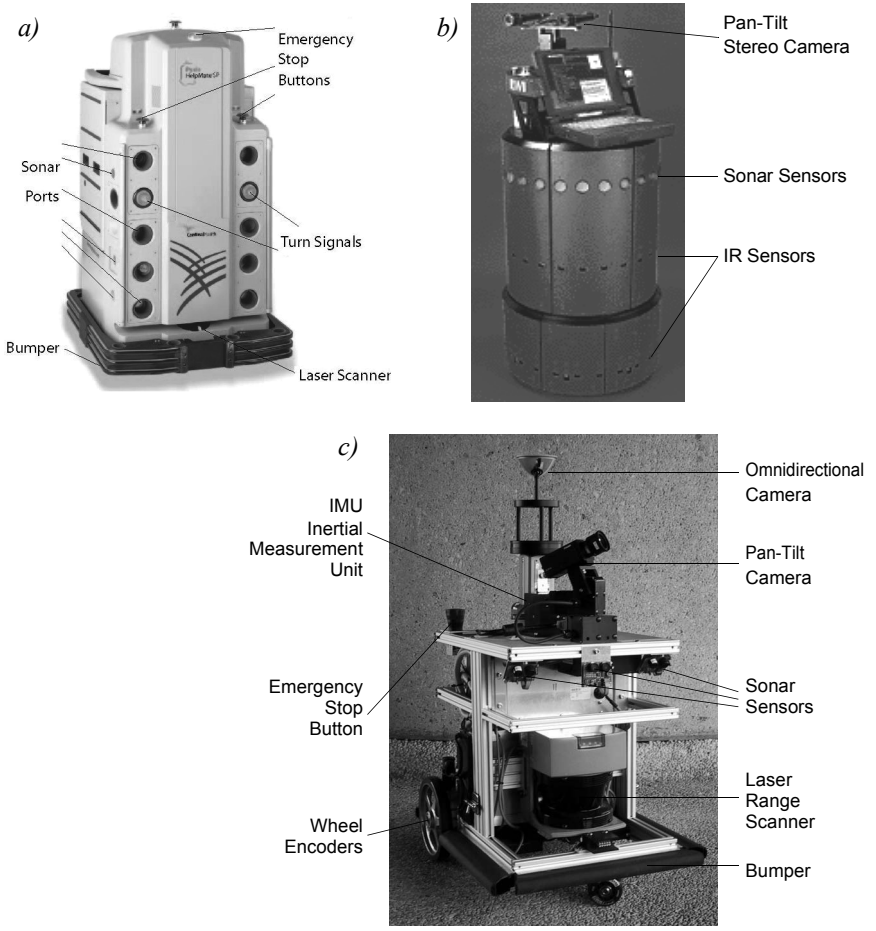


Figure 4.1
Examples of robots with multisensor systems: (a) HelpMate from Transition Research Corporation; (b) B21 from Real World Interface; (c) BIBA Robot, BlueBotics SA.

Passive sensors measure ambient environmental energy entering the sensor. Examples of passive sensors include temperature probes, microphones, and CCD or CMOS cameras.

Active sensors emit energy into the environment, then measure the environmental reaction. Because active sensors can manage more controlled interactions with the environment, they often achieve superior performance. However, active sensing introduces several risks: the outbound energy may affect the very characteristics that the sensor is attempting to measure. Furthermore, an active sensor may suffer from interference between its signal

and those beyond its control. For example, signals emitted by other nearby robots, or similar sensors on the same robot, may influence the resulting measurements. Examples of active sensors include wheel quadrature encoders, ultrasonic sensors, and laser rangefinders.

Table 4.1 provides a classification of the most useful sensors for mobile robot applications. The most interesting sensors are discussed in this chapter. The sensor classes in table 4.1 are arranged in ascending order of complexity and descending order of technological maturity. Tactile sensors and proprioceptive sensors are critical to virtually all mobile robots and are well understood and easily implemented. Commercial quadrature encoders, for example, may be purchased as part of a gear-motor assembly used in a mobile robot. At the other extreme, visual interpretation by means of one or more CCD/CMOS cameras provides a broad array of potential functionalities, from obstacle avoidance and localization to human face recognition. However, commercially available sensor units that provide visual functionalities are only now beginning to emerge [172, 346].

4.1.2 Characterizing sensor performance

The sensors we describe in this chapter vary greatly in their performance characteristics. Some sensors provide extreme accuracy in well-controlled laboratory settings but are overcome with error when subjected to real-world environmental variations. Other sensors provide narrow, high-precision data in a wide variety of settings. In order to quantify such performance characteristics, first we formally define the sensor performance terminology that will be valuable throughout the rest of this chapter.

4.1.2.1 Basic sensor response ratings

A number of sensor characteristics can be rated quantitatively in a laboratory setting. Such performance ratings will necessarily be best-case scenarios when the sensor is placed on a real-world robot, but are nevertheless useful.

Dynamic range is used to measure the spread between the lower and upper limits of input values to the sensor while maintaining normal sensor operation. Formally, the dynamic range is the ratio of the maximum input value to the minimum measurable input value. Because this raw ratio can be unwieldy, it is usually measured in *decibels,* which are computed as ten times the common logarithm of the dynamic range. However, there is potential confusion in the calculation of decibels, which are meant to measure the ratio between *powers*, such as watts or horsepower. Suppose your sensor measures motor current and can register values from a minimum of 1 mA to 20 mA. The dynamic range of this current sensor is defined as

Table 4.1
Classification of sensors used in mobile robotics applications

General classification (typical use)	Sensor Sensor System	PC or EC	A or P
Tactile sensors (detection of physical contact or closeness; security switches)	Contact switches, bumpers Optical barriers Noncontact proximity sensors	EC EC EC	P A A
Wheel/motor sensors (wheel/motor speed and position)	Brush encoders Potentiometers Synchros, resolvers Optical encoders Magnetic encoders Inductive encoders Capacitive encoders	PC PC PC PC PC PC PC	P P A A A A A
Heading sensors (orientation of the robot in relation to a fixed reference frame)	Compass Gyroscopes Inclinometers	EC PC EC	P P A/P
Acceleration sensor	Accelerometer	PC	P
Ground beacons (localization in a fixed reference frame)	GPS Active optical or RF beacons Active ultrasonic beacons Reflective beacons	EC EC EC EC	A A A A
Active ranging (reflectivity, time-of-flight, and geometric triangulation)	Reflectivity sensors Ultrasonic sensor Laser rangefinder Optical triangulation (1D) Structured light (2D)	EC EC EC EC EC	A A A A A
Motion/speed sensors (speed relative to fixed or moving objects)	Doppler radar Doppler sound	EC EC	A A
Vision sensors (visual ranging, whole-image analysis, segmentation, object recognition)	CCD/CMOS camera(s) Visual ranging packages Object tracking packages	EC	P

A, active; P, passive; P/A, passive/active; PC, proprioceptive; EC, exteroceptive.

$$10 \cdot \log\left[\frac{20}{0.001}\right] = 43 \text{ dB}. \tag{4.1}$$

Now suppose you have a voltage sensor that measures the voltage of your robot's battery, measuring any value from 1 mV to 20 V. Voltage is not a unit of power, but the square of voltage is proportional to power. Therefore, we use 20 instead of 10:

$$20 \cdot \log\left[\frac{20}{0.001}\right] = 86 \text{ dB}. \tag{4.2}$$

Range is also an important rating in mobile robot applications because often robot sensors operate in environments where they are frequently exposed to input values beyond their working range. In such cases, it is critical to understand how the sensor will respond. For example, an optical rangefinder will have a minimum operating range and can thus provide spurious data when measurements are taken with the object closer than that minimum.

Resolution is the minimum difference between two values that can be detected by a sensor. Usually, the lower limit of the dynamic range of a sensor is equal to its resolution. However, in the case of digital sensors, this is not necessarily so. For example, suppose that you have a sensor that measures voltage, performs an analog-to-digital (A/D) conversion, and outputs the converted value as an 8-bit number linearly corresponding to between 0 and 5 V. If this sensor is truly linear, then it has $2^8 - 1$ total output values, or a resolution of $5 \text{ V}(255) = 20 \text{ mV}$.

Linearity is an important measure governing the behavior of the sensor's output signal as the input signal varies. A linear response indicates that if two inputs x and y result in the two outputs $f(x)$ and $f(y)$, then for any values a and b, $f(ax + by) = af(x) + bf(y)$. This means that a plot of the sensor's input/output response is simply a straight line.

Bandwidth or *frequency* is used to measure the speed with which a sensor can provide a stream of readings. Formally, the number of measurements per second is defined as the sensor's frequency in *hertz*. Because of the dynamics of moving through their environment, mobile robots often are limited in maximum speed by the bandwidth of their obstacle detection sensors. Thus, increasing the bandwidth of ranging and vision sensors has been a high-priority goal in the robotics community.

4.1.2.2 In situ sensor performance

These sensor characteristics can be reasonably measured in a laboratory environment with confident extrapolation to performance in real-world deployment. However, a number of important measures cannot be reliably acquired without deep understanding of the complex interaction between all environmental characteristics and the sensors in question. This is

most relevant to the most sophisticated sensors, including active ranging sensors and visual interpretation sensors.

Sensitivity itself is a desirable trait. This is a measure of the degree to which an incremental change in the target input signal changes the output signal. Formally, sensitivity is the ratio of output change to input change. Unfortunately, however, the sensitivity of exteroceptive sensors is often confounded by undesirable sensitivity and performance coupling to other environmental parameters.

Cross-sensitivity is the technical term for sensitivity to environmental parameters that are orthogonal to the target parameters for the sensor. For example, a flux-gate compass can demonstrate high sensitivity to magnetic north and is therefore of use for mobile robot navigation. However, the compass will also demonstrate high sensitivity to ferrous building materials, so much so that its cross-sensitivity often makes the sensor useless in some indoor environments. High cross-sensitivity of a sensor is generally undesirable, especially when it cannot be modeled.

Error of a sensor is defined as the difference between the sensor's output measurements and the true values being measured, within some specific operating context. Given a true value v and a measured value m, we can define *error* as $error = m - v$.

Accuracy is defined as the degree of conformity between the sensor's measurement and the true value, and is often expressed as a proportion of the true value (e.g., 97.5% accuracy). Thus small error corresponds to high accuracy and vice versa:

$$\left(accuracy = 1 - \frac{|error|}{v}\right). \tag{4.3}$$

Of course, obtaining the ground truth, v, can be difficult or impossible, and so establishing a confident characterization of sensor accuracy can be problematic. Furthermore, it is important to distinguish between two different sources of error:

Systematic errors are caused by factors or processes that can in theory be modeled. These errors are, therefore, deterministic (i.e., predictable). Poor calibration of a laser rangefinder, an unmodeled slope of a hallway floor, and a bent stereo camera head due to an earlier collision are all possible causes of systematic sensor errors.

Random errors cannot be predicted using a sophisticated model; neither can they be mitigated by more precise sensor machinery. These errors can only be described in probabilistic terms (i.e., stochastically). Hue instability in a color camera, spurious rangefinding errors, and black level noise in a camera are all examples of random errors.

Precision is often confused with accuracy, and now we have the tools to clearly distinguish these two terms. Intuitively, high precision relates to reproducibility of the sensor results. For example, one sensor taking multiple readings of the same environmental state has high precision if it produces the same output. In another example, multiple copies of

this sensor taking readings of the same environmental state have high precision if their outputs agree. Precision does not, however, have any bearing on the accuracy of the sensor's output with respect to the true value being measured. Suppose that the *random error* of a sensor is characterized by some mean value μ and a standard deviation σ. The formal definition of precision is the ratio of the sensor's output range to the standard deviation:

$$precision = \frac{range}{\sigma}. \tag{4.4}$$

Note that only σ and not μ has impact on precision. In contrast, mean error μ is directly proportional to overall sensor error and inversely proportional to sensor accuracy.

4.1.2.3 Characterizing error: The challenges in mobile robotics

Mobile robots depend heavily on exteroceptive sensors. Many of these sensors concentrate on a central task for the robot: acquiring information on objects in the robot's immediate vicinity so that it may interpret the state of its surroundings. Of course, these "objects" surrounding the robot are all detected from the viewpoint of its local reference frame. Since the systems we study are mobile, their ever-changing position and their motion have a significant impact on overall sensor behavior. In this section, empowered with the terminology of the earlier discussions, we describe how dramatically the sensor error of a mobile robot disagrees with the ideal picture drawn in the previous section.

Blurring of systematic and random errors. Active ranging sensors tend to have failure modes that are triggered largely by specific relative positions of the sensor and environment targets. For example, a sonar sensor will produce specular reflections, producing grossly inaccurate measurements of range, at specific angles to a smooth sheetrock wall. During motion of the robot, such relative angles occur at stochastic intervals. This is especially true in a mobile robot outfitted with a ring of multiple sonars. The chances of one sonar entering this error mode during robot motion is high. From the perspective of the moving robot, the sonar measurement error is a random error in this case. Yet, if the robot were to stop, becoming motionless, then a very different error modality is possible. If the robot's static position causes a particular sonar to fail in this manner, the sonar will fail consistently and will tend to return precisely the same (and incorrect!) reading time after time. Once the robot is motionless, the error appears to be systematic and of high precision.

The fundamental mechanism at work here is the cross-sensitivity of mobile robot sensors to robot pose and robot-environment dynamics. The models for such cross-sensitivity are not, in an underlying sense, truly random. However, these physical interrelationships are rarely modeled, and therefore, from the point of view of an incomplete model, the errors appear random during motion and systematic when the robot is at rest.

Sonar is not the only sensor subject to this blurring of systematic and random error modality. Visual interpretation through the use of a CCD camera is also highly susceptible

to robot motion and position because of camera dependence on lighting changes, lighting specularity (e.g., glare), and reflections. The important point is to realize that, while systematic error and random error are well defined in a controlled setting, the mobile robot can exhibit error characteristics that bridge the gap between deterministic and stochastic error mechanisms.

Multimodal error distributions. It is common to characterize the behavior of a sensor's random error in terms of a probability distribution over various output values. In general, one knows very little about the causes of random error, and therefore several simplifying assumptions are commonly used. For example, we can assume that the error is *zero-mean* in that it symmetrically generates both positive and negative measurement error. We can go even further and assume that the probability density curve is Gaussian. Although we discuss the mathematics of this in detail in section 4.1.3, it is important for now to recognize the fact that one frequently assumes *symmetry* as well as *unimodal distribution*. This means that measuring the correct value is most probable, and any measurement that is farther away from the correct value is less likely than any measurement that is closer to the correct value. These are strong assumptions that enable powerful mathematical principles to be applied to mobile robot problems, but it is important to realize how wrong these assumptions usually are.

Consider, for example, the sonar sensor once again. When ranging an object that reflects the sound signal well, the sonar will exhibit high accuracy and will induce random error based on noise, for example, in the timing circuitry. This portion of its sensor behavior will exhibit error characteristics that are fairly symmetric and unimodal. However, when the sonar sensor is moving through an environment and is sometimes faced with materials that cause coherent reflection rather than return the sound signal to the sonar sensor, then the sonar will grossly overestimate the distance to the object. In such cases, the error will be biased toward positive measurement error and will be far from the correct value. The error is not strictly systematic, and so we are left modeling it as a probability distribution of random error. So the sonar sensor has two separate types of operational modes, one in which the signal does return and some random error is possible, and the second in which the signal returns after a multipath reflection and gross overestimation error occurs. The probability distribution could easily be at least bimodal in this case, and since overestimation is more common than underestimation, it will also be asymmetric.

As a second example, consider ranging via stereo vision. Once again, we can identify two modes of operation. If the stereo vision system correctly correlates two images, then the resulting random error will be caused by camera noise and will limit the measurement accuracy. But the stereo vision system can also correlate two images *incorrectly*, matching two fenceposts, for example, that are not the same post in the real world. In such a case stereo vision will exhibit gross measurement error, and one can easily imagine such behavior violating both the unimodal and the symmetric assumptions.

The thesis of this section is that sensors in a mobile robot *may* be subject to multiple modes of operation and, when the sensor error is characterized, unimodality and symmetry may be grossly violated. Nonetheless, as we shall see, many successful mobile robot systems make use of these simplifying assumptions and the resulting mathematical techniques with great empirical success.

4.1.3 Representing uncertainty

In section 4.1.2 we presented a terminology for describing the performance characteristics of a sensor. As mentioned there, sensors are imperfect devices with errors of both systematic and random nature. Random errors, in particular, cannot be corrected, and so they represent atomic levels of sensor uncertainty.

But when you build a mobile robot, you combine information from many sensors, even using the same sensors repeatedly, over time, to build, possibly, a model of the environment. How can we scale up, from characterizing the uncertainty of a single sensor to the uncertainty of the resulting robot system?

We begin by presenting a statistical representation for the random error associated with an individual sensor [14]. With a quantitative tool in hand, the standard Gaussian uncertainty model can be presented and evaluated. Finally, we present a framework for computing the uncertainty of conclusions drawn from a set of quantifiably uncertain measurements, known as the *error propagation law*.

4.1.3.1 Statistical representation

We have already defined *error* as the difference between a sensor measurement and the true value. From a statistical point of view, we wish to characterize the error of a sensor, not for one specific measurement but for any measurement. Let us formulate the problem of sensing as an estimation problem. The sensor has taken a set of n measurements with values ρ_i. The goal is to characterize the estimate of the true value $E[X]$ given these measurements:

$$E[X] = g(\rho_1, \rho_2, \ldots, \rho_n). \tag{4.5}$$

From this perspective, the true value is represented by a random (and therefore unknown) variable X. We use a *probability density function* to characterize the statistical properties of the value of X.

In figure 4.2, the density function identifies for each possible value x of X a probability density $f(x)$ along the y-axis. The area under the curve is 1, indicating the complete chance of X having *some* value:

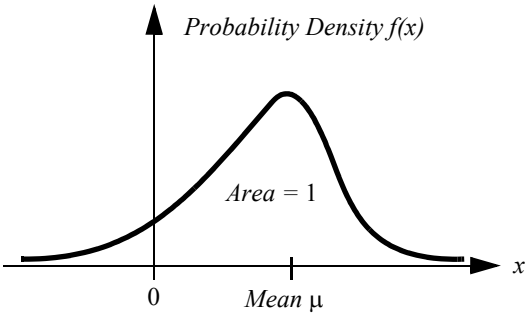


Figure 4.2
A sample probability density function, showing a single probability peak (i.e., unimodal) with asymptotic drops in both directions.

$$\int_{-\infty}^{\infty} f(x)dx = 1 . \tag{4.6}$$

The probability of the value of X falling between two limits a and b is computed as the bounded integral:

$$p[a < X \leq b] = \int_{a}^{b} f(x)dx \tag{4.7}$$

The probability density function is a useful way to characterize the possible values of X because it captures not only the range of X but also the comparative probability of different values for X. Using $f(x)$ we can quantitatively define the mean, variance, and standard deviation as follows.

The *mean value* μ is equivalent to the expected value $E[X]$ if we were to measure X an infinite number of times and average all of the resulting values. We can easily define $E[X]$:

$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx . \tag{4.8}$$

Note in this equation that calculation of $E[X]$ is identical to the weighted average of all possible values of x. In contrast, the *mean square value* is simply the weighted average of the squares of all values of x:

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx . \tag{4.9}$$

Characterization of the "width" of the possible values of X is a key statistical measure, and this requires first defining the *variance* σ^2:

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx . \tag{4.10}$$

Finally, the *standard deviation* σ is simply the square root of variance σ^2, and σ^2 will play important roles in our characterization of the error of a single sensor as well as the error of a model generated by combining multiple sensor readings.

Independence of random variables. With the tools presented here, we often evaluate systems with multiple random variables. For instance, a mobile robot's laser rangefinder may be used to measure the position of a feature on the robot's right and, later, another feature on the robot's left. The position of each feature in the real world may be treated as random variables, X_1 and X_2.

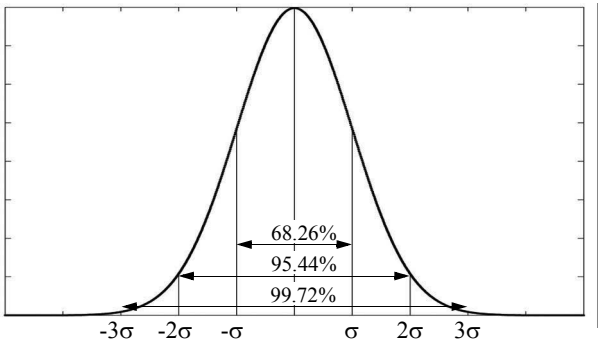
Two random variables X_1 and X_2 are *independent* if the particular value of one has no bearing on the particular value of the other. In this case we can draw several important conclusions about the statistical behavior of X_1 and X_2. First, the expected value (or mean value) of the product of random variables is equal to the product of their mean values:

$$E[X_1 X_2] = E[X_1]E[X_2] . \tag{4.11}$$

Second, the variance of their sums is equal to the sum of their variances:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) . \tag{4.12}$$

In mobile robotics, we often assume the independence of random variables even when this assumption is not strictly true. The simplification that results makes a number of the existing mobile robot-mapping and navigation algorithms tenable, as described in chapter 5. A further simplification, described in the following section, revolves around one specific probability density function used more often than any other when modeling error: the Gaussian distribution.



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Figure 4.3
The Gaussian function with $\mu = 0$ and $\sigma = 1$. We shall refer to this as the reference Gaussian. The value 2σ is often refereed to as the signal quality; 95.44% of the values fall within $\pm 2\sigma$.

Gaussian distribution. The Gaussian distribution, also called the *normal distribution,* is used across engineering disciplines when a well-behaved error model is required for a random variable for which no error model of greater felicity has been discovered. The Gaussian has many characteristics that make it mathematically advantageous to other ad hoc probability density functions. It is symmetric around the mean μ. There is no particular bias for being larger than or smaller than μ, and this makes sense when there is no information to the contrary. The Gaussian distribution is also unimodal, with a single peak that reaches a maximum at μ (necessary for any symmetric, unimodal distribution). This distribution also has tails (the value of $f(x)$ as x approaches $-\infty$ and ∞) that approach zero only asymptotically. This means that all amounts of error are possible, although very large errors may be highly improbable. In this sense, the Gaussian is conservative. Finally, as seen in the formula for the Gaussian probability density function, the distribution depends on only two parameters:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{4.13}$$

The Gaussian's basic shape is determined by the structure of this formula, and so the only two parameters required to fully specify a particular Gaussian are its mean, μ, and its standard deviation, σ. Figure 4.3 shows the Gaussian function with $\mu = 0$ and $\sigma = 1$.

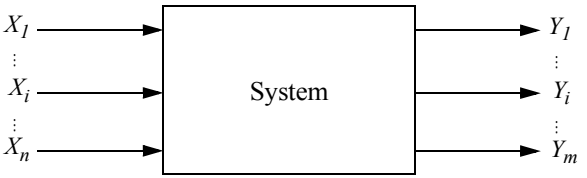


Figure 4.4
Error propagation in a multiple-input multi-output system with *n* inputs and *m* outputs.

Suppose that a random variable X is modeled as a Gaussian. How does one identify the chance that the value of X is within one standard deviation of μ? In practice, this requires integration of $f(x)$, the Gaussian function to compute the area under a portion of the curve:

$$Area = \int_{-\sigma}^{\sigma} f(x)dx . \tag{4.14}$$

Unfortunately, there is no closed-form solution for the integral in equation (4.14), and so the common technique is to use a Gaussian *cumulative probability table*. Using such a table, one can compute the probability for various value ranges of X:

$$p[\mu - \sigma < X \leq \mu + \sigma] = 0.68 ;$$

$$p[\mu - 2\sigma < X \leq \mu + 2\sigma] = 0.95 ;$$

$$p[\mu - 3\sigma < X \leq \mu + 3\sigma] = 0.997 .$$

For example, 95% of the values for X fall within two standard deviations of its mean. This applies to *any* Gaussian distribution. As is clear from the above progression, under the Gaussian assumption, once bounds are relaxed to 3σ, the overwhelming proportion of values (and, therefore, probability) is subsumed.

4.1.3.2 Error propagation: Combining uncertain measurements

These probability mechanisms may be used to describe the errors associated with a single sensor's attempts to measure a real-world value. But in mobile robotics, one often uses a series of measurements, all of them uncertain, to extract a single environmental measure. For example, a series of uncertain measurements of single points can be fused to extract the position of a line (e.g., a hallway wall) in the environment (figure 4.88).

Consider the system in figure 4.4, where X_i are n input signals with a known probability distribution and Y_i are *m* outputs. The question of interest is this: What can we say about the probability distribution of the output signals Y_i if they depend with known functions

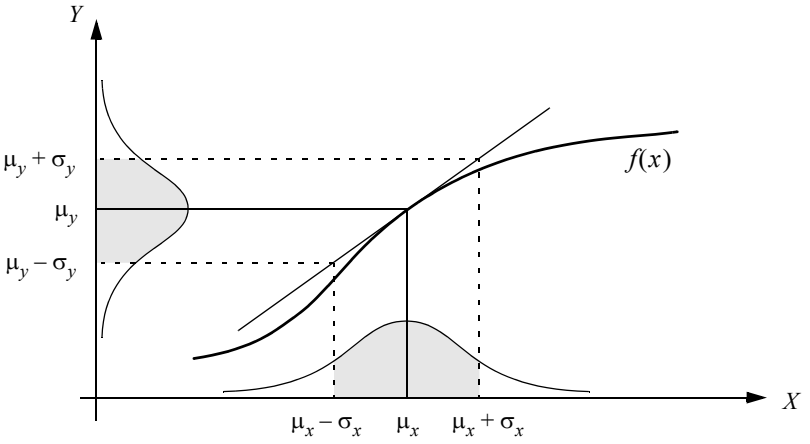


Figure 4.5
One-dimensional case of a nonlinear error propagation problem.

f_i upon the input signals? Figure 4.5 depicts the 1D version of this error propagation problem as an example.

The general solution can be generated using the first-order Taylor expansion of f_i . The output covariance matrix C_Y is given by the error propagation law:

$$C_Y = F_X C_X F_X^T, \tag{4.15}$$

where

C_X = covariance matrix representing the input uncertainties;

C_Y = covariance matrix representing the propagated uncertainties for the outputs;

F_x is the *Jacobian* matrix defined as

$$F_X = \nabla f = \begin{bmatrix} \frac{\partial f_1}{\partial X_1} & \cdots & \frac{\partial f_1}{\partial X_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial f_m}{\partial X_1} & \cdots & \frac{\partial f_m}{\partial X_n} \end{bmatrix}. \tag{4.16}$$

This is also the transpose of the gradient of $f(X)$.

We will not present a detailed derivation here but will use equation (4.15) to solve an example problem in section 4.7.1.

The preceding sections have presented a terminology with which we can characterize the advantages and disadvantages of various mobile robot sensors. In the following sections, we do the same for a sampling of the most commonly used mobile robot sensors today.

4.1.4 Wheel/motor sensors

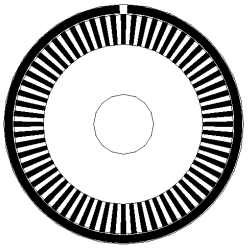
Wheel/motor sensors are devices used to measure the internal state and dynamics of a mobile robot. These sensors have vast applications outside of mobile robotics and, as a result, mobile robotics has enjoyed the benefits of high-quality, low-cost wheel and motor sensors that offer excellent resolution. In the next section, we sample just one such sensor, the optical incremental encoder.

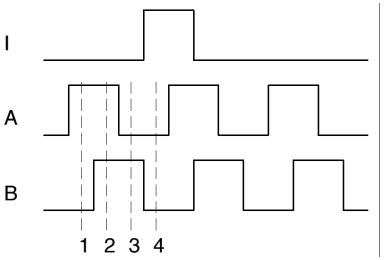
4.1.4.1 Optical encoders

Optical incremental encoders have become the most popular device for measuring angular speed and position within a motor drive or at the shaft of a wheel or steering mechanism. In mobile robotics, encoders are used to control the position or speed of wheels and other motor-driven joints. Because these sensors are *proprioceptive*, their estimate of position is best in the reference frame of the robot and, when applied to the problem of robot *localization*, significant corrections are required, as discussed in chapter 5.

An optical encoder is basically a mechanical light chopper that produces a certain number of sine or square wave pulses for each shaft revolution. It consists of an illumination source, a fixed grating that masks the light, a rotor disc with a fine optical grid that rotates with the shaft, and fixed optical detectors. As the rotor moves, the amount of light striking the optical detectors varies based on the alignment of the fixed and moving gratings. In robotics, the resulting sine wave is transformed into a discrete square wave using a threshold to choose between *light* and *dark* states. Resolution is measured in *cycles per revolution* (CPR). The minimum angular resolution can be readily computed from an encoder's CPR rating. A typical encoder in mobile robotics may have 2000 CPR, while the optical encoder industry can readily manufacture encoders with 10,000 CPR. In terms of required bandwidth, it is of course critical that the encoder be sufficiently fast to count at the shaft spin speeds that are expected. Industrial optical encoders present no bandwidth limitation to mobile robot applications.

Usually in mobile robotics the *quadrature encoder* is used. In this case, a second illumination and detector pair is placed 90 degrees shifted with respect to the original in terms of the rotor disc. The resulting twin square waves, shown in figure 4.6, provide significantly more information. The ordering of which square wave produces a rising edge first identifies the direction of rotation. Furthermore, the four detectably different states improve the res-





State	Ch A	Ch B
s_1	high	low
s_2	high	high
s_3	low	high
s_4	low	low

Figure 4.6
Quadrature optical wheel encoder: The observed phase relationship between channel A and B pulse trains are used to determine the direction of the rotation. A single slot in the outer track generates a reference (index) pulse per revolution.

olution by a factor of four with no change to the rotor disc. Thus, a 2000 CPR encoder in quadrature yields 8000 counts. Further improvement is possible by retaining the sinusoidal wave measured by the optical detectors and performing sophisticated interpolation. Such methods, although rare in mobile robotics, can yield 1000-fold improvements in resolution.

As with most proprioceptive sensors, encoders are generally in the controlled environment of a mobile robot's internal structure, and so systematic error and cross-sensitivity can be engineered away. The accuracy of optical encoders is often assumed to be 100% and, although this may not be entirely correct, any errors at the level of an optical encoder are dwarfed by errors downstream of the motor shaft.

4.1.5 Heading sensors

Heading sensors can be *proprioceptive* (gyroscope, inclinometer) or *exteroceptive* (compass). They are used to determine the robot's orientation and inclination. They allow us, together with appropriate velocity information, to integrate the movement to a position estimate. This procedure, which has its roots in vessel and ship navigation, is called *dead reckoning*.

4.1.5.1 Compasses

The two most common modern sensors for measuring the direction of a magnetic field are the Hall effect and flux gate compasses. Each has advantages and disadvantages, as described below.

The Hall effect describes the behavior of electric potential in a semiconductor when in the presence of a magnetic field. When a constant current is applied across the length of a semiconductor, there will be a voltage difference in the perpendicular direction, across the semiconductor's width, based on the relative orientation of the semiconductor to magnetic flux lines. In addition, the sign of the voltage potential identifies the direction of the mag-

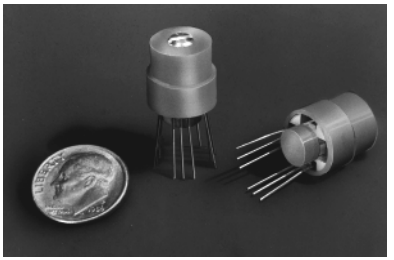


Figure 4.7
Digital compass: Sensors such as the digital/analog Hall effect sensor shown, available from Dinsmore, enable inexpensive (< $ 15) sensing of magnetic fields.

netic field. Thus, a single semiconductor provides a measurement of flux and direction along one dimension. Hall effect digital compasses are popular in mobile robotics, and they contain two such semiconductors at right angles, providing two axes of magnetic field (thresholded) direction, thereby yielding one of eight possible compass directions. The instruments are inexpensive but also suffer from a range of disadvantages. Resolution of a digital Hall effect compass is poor. Internal sources of error include the nonlinearity of the basic sensor and systematic bias errors at the semiconductor level. The resulting circuitry must perform significant filtering, and this lowers the bandwidth of Hall effect compasses to values that are slow in mobile robot terms. For example, the Hall effect compass pictured in figure 4.7 needs 2.5 seconds to settle after a 90-degree spin.

The flux gate compass operates on a different principle. Two small coils are wound on ferrite cores and are fixed perpendicular to one another. When alternating current is activated in both coils, the magnetic field causes shifts in the phase depending on its relative alignment with each coil. By measuring both phase shifts, the direction of the magnetic field in two dimensions can be computed. The flux gate compass can accurately measure the strength of a magnetic field and has improved resolution and accuracy; however, it is both larger and more expensive than a Hall effect compass.

Regardless of the type of compass used, a major drawback concerning the use of Earth's magnetic field for mobile robot applications involves disturbance of that magnetic field by other magnetic objects and man-made structures, as well as the bandwidth limitations of electronic compasses and their susceptibility to vibration. Particularly in indoor environments, mobile robotics applications have often avoided the use of compasses, although a compass can conceivably provide useful *local* orientation information indoors, even in the presence of steel structures.

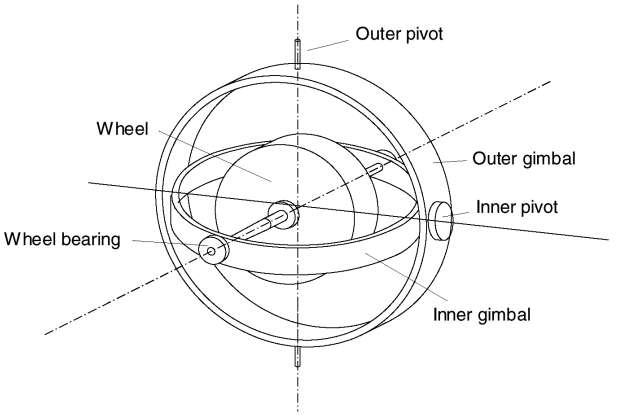


Figure 4.8
Two-axis mechanical gyroscope.

4.1.5.2 Gyroscopes

Gyroscopes are heading sensors that preserve their orientation in relation to a fixed reference frame. Thus, they provide an absolute measure for the heading of a mobile system. Gyroscopes can be classified in two categories, mechanical gyroscopes and optical gyroscopes.

Mechanical gyroscopes. The concept of a mechanical gyroscope relies on the inertial properties of a fast-spinning rotor. The property of interest is known as the gyroscopic precession. If you try to rotate a fast-spinning wheel around its vertical axis, you will feel a harsh reaction in the horizontal axis. This is due to the angular momentum associated with a spinning wheel and will keep the axis of the gyroscope inertially stable. The reactive torque τ and thus the tracking stability with the inertial frame are proportional to the spinning speed ω, the precession speed Ω, and the wheel's inertia I.

$$\tau = I\omega\Omega . \tag{4.17}$$

By arranging a spinning wheel, as seen in figure 4.8, no torque can be transmitted from the outer pivot to the wheel axis. The spinning axis will therefore be space-stable (i.e., fixed in an inertial reference frame). Nevertheless, the remaining friction in the bearings of the gyro axis introduce small torques, thus limiting the long-term space stability and introduc-

ing small errors over time. A high quality mechanical gyroscope can cost up to $100,000 and has an angular drift of about 0.1 degrees in 6 hours.

For navigation, the spinning axis has to be initially selected. If the spinning axis is aligned with the north-south meridian, the earth's rotation has no effect on the gyro's horizontal axis. If it points east-west, the horizontal axis reads the earth rotation.

Rate gyros have the same basic arrangement as shown in figure 4.8, but with a slight modification. The gimbals are restrained by a torsional spring with additional viscous damping. This enables the sensor to measure angular speeds instead of absolute orientation.

Optical gyroscopes. Optical gyroscopes are a relatively new innovation. Commercial use began in the early 1980s when they were first installed in aircraft. Optical gyroscopes are angular speed sensors that use two monochromatic light beams, or lasers, emitted from the same source, instead of moving, mechanical parts. They work on the principle that the speed of light remains unchanged and, therefore, geometric change can cause light to take a varying amount of time to reach its destination. One laser beam is sent traveling clockwise through an optical fiber while the other travels counterclockwise. Because the laser traveling in the direction of rotation has a slightly shorter path, it will have a higher frequency. This principle is known as the *Sagnac effect*. The difference in frequency Δf of the two beams is proportional to the angular velocity Ω of the cylinder. To make the difference measurable, the sensor is a coil consisting of as much as 5 km of optical fiber. New solid-state optical gyroscopes based on the same principle are build using microfabrication technology, thereby providing heading information with resolution and bandwidth far beyond the needs of mobile robotic applications. Bandwidth, for instance, can easily exceed 100 kHz, while resolution can be smaller than 0.0001 degrees/hr.

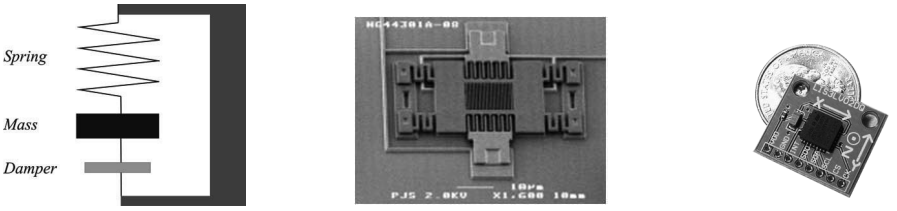
4.1.6 Accelerometers

An accelerometers is a device used to measure all external forces acting upon it, including gravity. Accelerometers belong to the proprioceptive sensors class.

Conceptually, an accelerometer is a spring–mass–damper system (figure 4.9a) in which the three-dimensional position of the proof mass relative to the accelerometer casing can be measured with some mechanism. Assume that an external force is applied on the sensor casing (e.g., gravity) and that we have an ideal spring with a force proportional to its displacement. Then, we can write [118]

$$F_{applied} = F_{inertial} + F_{damping} + F_{spring} = m\ddot{x} + c\dot{x} + kx, \tag{4.18}$$

where m is the proof mass, c is the damping coefficient, k is the spring constant, and x is the equilibrium case relative position. By choosing appropriately the damping material and the mass, the system can be made to converge very quickly to a stable value under the effect



a) *b)* *c)*

Figure 4.9 Accelerometers: (a) Working principle of the mechanical accelerometer; (b) An example MEMS accelerometer produced by Sandia National Laboratories; (c) An example commercial MEMS accelerometer.

of a static force. When the stable value is reached, then $\ddot{x} = 0$ and the applied acceleration can be obtained as

$$a_{applied} = \frac{kx}{m}. \tag{4.19}$$

This is the working principle of a mechanical accelerometer. Modern accelerometers are often small Micro Electro-Mechanical Systems (MEMS) consisting of a springlike structure (*cantilevered beam*) with a proof mass (also known as *seismic mass*). Damping results from the residual gas sealed in the device. When an external force is applied, the proof mass deflects from its neutral position. Depending on the physical principle used to measure this deflection, we can have different types of accelerometers. Capacitive accelerometers measure the deflection by measuring the capacitance between a fixed structure and the proof mass. These accelerometers are reliable and inexpensive (figure 4.9b–c). Another alternative are the piezoelectric accelerometers. They are based on the property exhibited by certain crystals to generate a voltage when a mechanical stress is applied to them. A small mass is positioned on the crystal, and, when an external force is applied, the mass moves, and this induces a voltage that can be measured.

Notice that each accelerometer measures acceleration along a single axis. By mounting three accelerometers orthogonally to one another, an omnidirectional (i.e., three-axis) accelerometer can be obtained.

Also observe that an accelerometer at rest on the Earth's surface will always indicate 1 g along the vertical axis. To obtain the inertial acceleration (due to motion alone), the gravity vector must be subtracted. Conversely, the accelerometer's output will be zero during free fall.

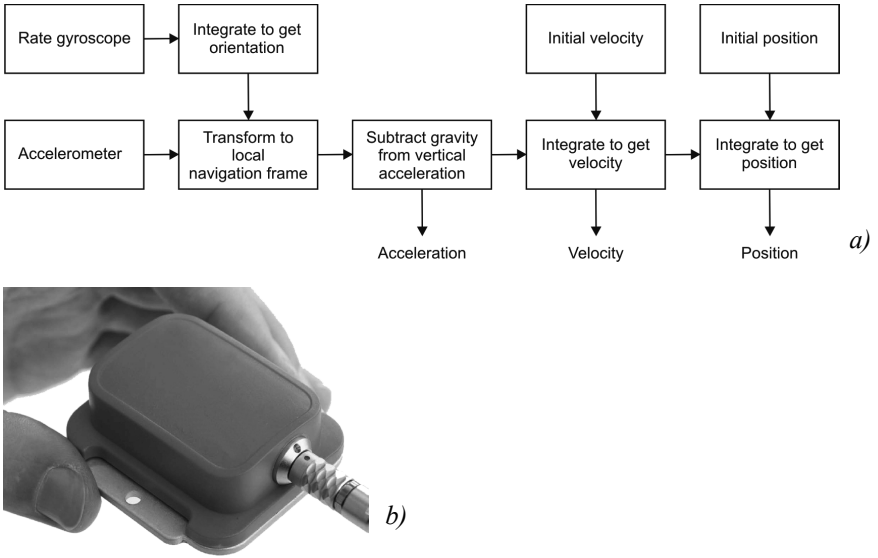


Figure 4.10 (a) IMU block diagram (redrawn from [118]). (b) A commercial IMU produced by Xsens. Image courtesy of Xsens—http://www.xsens.com.

Finally, accelerometers are classified into two categories according to their passband bandwidth: accelerometers for static and dynamic measurements. In the first category are low-pass accelerometers which can measure accelerations from 0 Hz up to usually 500 Hz. This is typical for mechanical and capacitive accelerometers. Typical uses are measurements of the gravitational acceleration or that of a moving vehicle. The second category of accelerometers is used for measuring accelerations of vibrating objects or accelerations during crashes. In this case, the bandwidth ranges between a few Hz up to 50 KHz. Typical accelerometers in this category are those realized with piezoelectric technology.

4.1.7 Inertial measurement unit (IMU)

An inertial measurement unit (IMU) is a device that uses gyroscopes and accelerometers to estimate the relative position, velocity, and acceleration of a moving vehicle. An IMU is also known as an Inertial Navigation System (INS), and it has become a common navigational component of aircraft and ships. An IMU estimates the six-degree-of-freedom (DOF) pose of the vehicle: position (x, y, z) and orientation (roll, pitch, yaw). Nevertheless, heading sensors like compasses and gyroscopes, which conversely only estimate orientation, are often improperly called IMUs.

Besides the 6-DOF pose of the vehicle, commercial IMUs also usually estimate velocity and acceleration. To estimate the velocity, the initial speed of the vehicle needs to be known. The working principle of an IMU is shown in figure 4.10. Let us suppose that our

IMU has three orthogonal accelerometers and three orthogonal gyroscopes. The gyroscope data is integrated to estimate the vehicle orientation while the three accelerometers are used to estimate the instantaneous acceleration of the vehicle. The acceleration is then transformed to the local navigation frame by means of the current estimate of the vehicle orientation relative to gravity. At this point the gravity vector can be subtracted from the measurement. The resulting acceleration is then integrated to obtain the velocity and then integrated again to obtain the position, provided that both the initial velocity and position are a priori known. To overcome the need of knowing of the initial velocity, the integration is typically started at rest (i.e., velocity equal to zero).

Observe that IMUs are extremely sensitive to measurement errors in both gyroscopes and accelerometers. For example, drift in the gyroscope unavoidably undermines the estimation of the vehicle orientation relative to gravity, which results in incorrect cancellation of the gravity vector. Additionally observe that, because the accelerometer data is integrated twice to obtain the position, any residual gravity vector results in a quadratic error in position. Because of this and the fact that any other error is integrated over time, drift is a fundamental problem in IMUs. After long period of operation, all IMUs drift. To cancel this drift, some reference to some external measurement is required. In many robot applications, this has been done using cameras or GPS. In particular, cameras allow the user to annihilate the drift every time a given feature of the environment—whose 3D position in the camera reference frame is known—is reobserved (see sections 4.2.6 or 5.8.5). Similarly, as described in the next section, GPS allows the user to correct the pose estimate every time the GPS signal is received.

4.1.8 Ground beacons

One elegant approach to solving the localization problem in mobile robotics is to use active or passive beacons. Using the interaction of on-board sensors and the environmental beacons, the robot can identify its position precisely. Although the general intuition is identical to that of early human navigation beacons, such as stars, mountains, and lighthouses, modern technology has enabled sensors to localize an outdoor robot with accuracies of better than 5 cm within areas that are kilometers in size.

In the following section, we describe one such beacon system, the global positioning system (GPS), which is extremely effective for outdoor ground and flying robots. Indoor beacon systems have been generally less successful for a number of reasons. The expense of environmental modification in an indoor setting is not amortized over an extremely large useful area, as it is, for example, in the case of the GPS. Furthermore, indoor environments offer significant challenges not seen outdoors, including multipath and environmental dynamics. A laser indoor beacon system, for example, must disambiguate the one true laser signal from possibly tens of other powerful signals reflected off walls, smooth floors, and doors. Confounding this, humans and other obstacles may be constantly changing the envi-

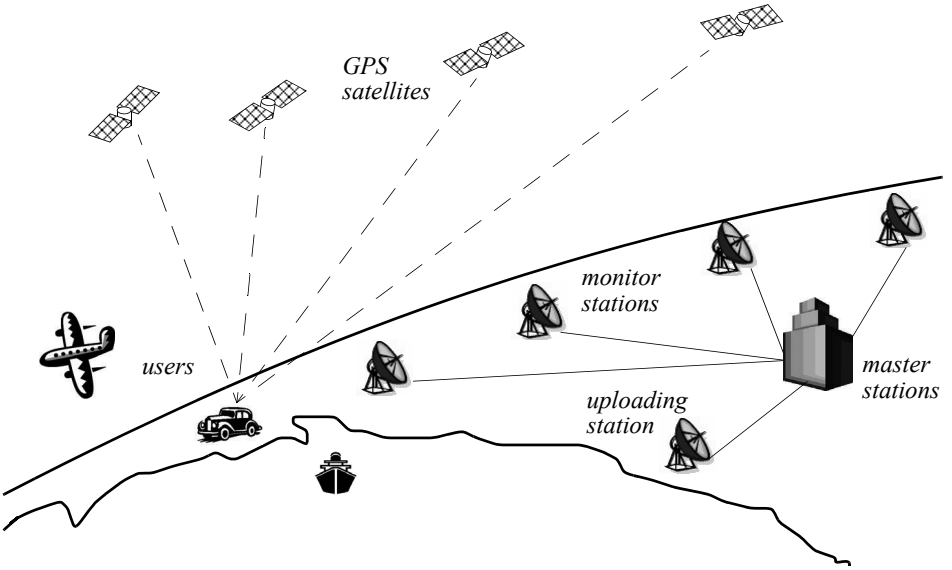


Figure 4.11
Calculation of position and heading based on GPS.

ronment, for example, occluding the one true path from the beacon to the robot. In commercial applications, such as manufacturing plants, the environment can be carefully controlled to ensure success. In less structured indoor settings, beacons have nonetheless been used, and the problems are mitigated by careful beacon placement and the use of passive sensing modalities.

4.1.8.1 The global positioning system

The *global positioning system* (GPS) was initially developed for military use but is now freely available for civilian navigation. There are at least twenty-four operational GPS satellites at all times. The satellites orbit every twelve hours at a height of 20.190 km. Four satellites are located in each of six planes inclined 55 degrees with respect to the plane of the earth's equator (figure 4.11).

Each satellite continuously transmits data that indicate its location and the current time. Therefore, GPS receivers are completely passive but exteroceptive sensors. The GPS satellites synchronize their transmissions so that their signals are sent at the same time. When a GPS receiver reads the transmission of two or more satellites, the arrival time differences inform the receiver as to its relative distance to each satellite. By combining information regarding the arrival time and instantaneous location of four satellites, the receiver can infer its own position. In theory, such triangulation requires only three data points. However,

timing is extremely critical in the GPS application because the time intervals being measured are in nanoseconds. It is, of course, mandatory that the satellites be well synchronized. To this end, they are updated by ground stations regularly and each satellite carries on-board atomic clocks for timing.

The GPS receiver clock is also important so that the travel time of each satellite's transmission can be accurately measured. But GPS receivers have a simple quartz clock. So, although three satellites would ideally provide position in three axes, the GPS receiver requires four satellites, using the additional information to solve for four variables: three position axes plus a time correction.

The fact that the GPS receiver must read the transmission of four satellites simultaneously is a significant limitation. GPS satellite transmissions are extremely low-power, and reading them successfully requires direct line-of-sight communication with the satellite. Thus, in confined spaces such as city blocks with tall buildings or in dense forests, one is unlikely to receive four satellites reliably. Of course, most indoor spaces will also fail to provide sufficient visibility of the sky for a GPS receiver to function. For these reasons, the GPS has been a popular sensor in mobile robotics, but it has been relegated to projects involving mobile robot traversal of wide-open spaces and autonomous flying machines.

A number of factors affect the performance of a localization sensor that makes use of the GPS. First, it is important to understand that, because of the specific orbital paths of the GPS satellites, coverage is not geometrically identical in different portions of Earth and therefore resolution is not uniform. Specifically, at the North and South Poles, the satellites are very close to the horizon, and thus, while resolution in the latitude and longitude directions is good, resolution of altitude is relatively poor as compared to more equatorial locations.

The second point is that GPS satellites are merely an information source. They can be employed with various strategies in order to achieve dramatically different levels of localization resolution. The basic strategy for GPS use, called *pseudorange* and described earlier, generally performs at a resolution of 15 m. An extension of this method is *differential GPS (DGPS)*, which makes use of a second receiver that is static and at a known exact position. A number of errors can be corrected using this reference, and so resolution improves to the order of 1 m or less. A disadvantage of this technique is that the stationary receiver must be installed, its location must be measured very carefully, and of course the moving robot must be within kilometers of this static unit in order to benefit from the DGPS technique.

A further improved strategy is to take into account the phase of the carrier signals of each received satellite transmission. There are two carriers, at 19 cm and 24 cm, and therefore significant improvements in precision are possible when the phase difference between multiple satellites is measured successfully. Such receivers can achieve 1 cm resolution for point positions and, with the use of multiple receivers, as in DGPS, sub-1 cm resolution.

A final consideration for mobile robot applications is bandwidth.The GPS will generally offer no better than 200 to 300 ms latency, and so one can expect no better than 5 Hz GPS updates. On a fast-moving mobile robot or flying robot, this can mean that local motion integration will be required for proper control due to GPS latency limitations.

4.1.9 Active ranging

Active ranging sensors continue to be the most popular sensors in mobile robotics. Many ranging sensors have a low price point, and, most important, all ranging sensors provide easily interpreted outputs: direct measurements of distance from the robot to objects in its vicinity. For obstacle detection and avoidance, most mobile robots rely heavily on active ranging sensors. But the local free space information provided by ranging sensors can also be accumulated into representations beyond the robot's current local reference frame. Thus active ranging sensors are also commonly found as part of the localization and environmental modeling processes of mobile robots. It is only with the slow advent of successful visual interpretation competence that we can expect the class of active ranging sensors to gradually lose their primacy as the sensor class of choice among mobile roboticists.

We next present three *time-of-flight* active ranging sensors: the ultrasonic sensor, the laser rangefinder, and the time-of-flight camera. Then, we present two geometric active ranging sensors: the optical triangulation sensor and the structured light sensor.

4.1.9.1 Time-of-flight active ranging

Time-of-flight ranging makes use of the propagation speed of sound or an electromagnetic wave. In general, the travel distance of a sound of electromagnetic wave is given by

$$d = c \cdot t, \tag{4.20}$$

where

d = distance traveled (usually round-trip);

c = speed of wave propagation;

t = time of flight.

It is important to point out that the propagation speed v of sound is approximately 0.3 m/ms whereas the speed of electromagnetic signals is 0.3 m/ns, which is 1 million times faster. The time of flight for a typical distance, say 3 m, is 10 ms for an ultrasonic system but only 10 ns for a laser rangefinder. It is thus evident that measuring the time of flight t with electromagnetic signals is more technologically challenging. This explains why laser range sensors have only recently become affordable and robust for use on mobile robots.

The quality of time-of-flight range sensors depends mainly on

- uncertainties in determining the exact time of arrival of the reflected signal;
- inaccuracies in the time-of-flight measurement (particularly with laser range sensors);
- the dispersal cone of the transmitted beam (mainly with ultrasonic range sensors);
- interaction with the target (e.g., surface absorption, specular reflections);
- variation of propagation speed;
- the speed of the mobile robot and target (in the case of a dynamic target);

As discussed in the following, each type of time-of-flight sensor is sensitive to a particular subset of this list of factors.

The ultrasonic sensor (time-of-flight, sound). The basic principle of an ultrasonic sensor is to transmit a packet of (ultrasonic) pressure waves and to measure the time it takes for this wave packet to reflect and return to the receiver. The distance d of the object causing the reflection can be calculated based on the propagation speed of sound c and the time of flight t.

$$d = \frac{c \cdot t}{2}. \tag{4.21}$$

The speed of sound c in air is given by

$$c = \sqrt{\gamma R T}, \tag{4.22}$$

where

γ = ratio of specific heats;

R = gas constant;

T = temperature in degrees Kelvin.

In air at standard pressure and 20° C the speed of sound is approximately $c = 343$ m/s.

Figure 4.12 shows the different signal output and input of an ultrasonic sensor. First, a series of sound pulses are emitted, comprising the *wave packet*. An integrator also begins to linearly climb in value, measuring the time from the transmission of these sound waves to detection of an echo. A threshold value is set for triggering an incoming sound wave as a valid echo. This threshold is often decreasing in time, because the amplitude of the expected echo decreases over time based on dispersal as it travels longer. But during transmission of the initial sound pulses and just afterward, the threshold is set very high to suppress triggering the echo detector with the outgoing sound pulses. A transducer will

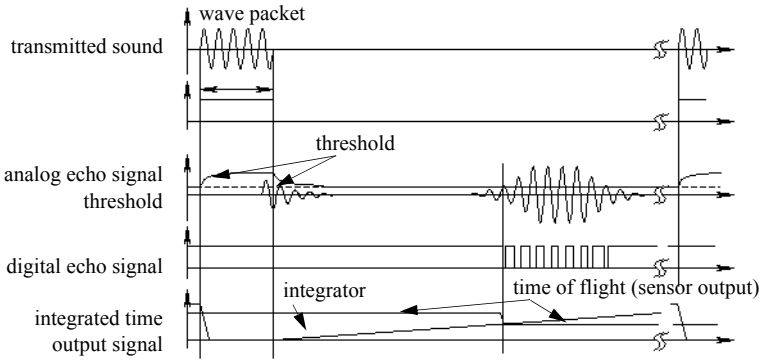


Figure 4.12
Signals of an ultrasonic sensor.

continue to ring for up to several milliseconds after the initial transmission, and this governs the *blanking time* of the sensor. Note that if, during the blanking time, the transmitted sound were to reflect off of an extremely close object and return to the ultrasonic sensor, it may fail to be detected.

However, once the blanking interval has passed, the system will detect any above-threshold reflected sound, triggering a digital signal and producing the distance measurement using the integrator value.

The ultrasonic wave typically has a frequency between 40 and 180 kHz and is usually generated by a piezo or electrostatic transducer. Often the same unit is used to measure the reflected signal, although the required blanking interval can be reduced through the use of separate output and input devices. Frequency can be used to select a useful range when choosing the appropriate ultrasonic sensor for a mobile robot. Lower frequencies correspond to a longer range, but with the disadvantage of longer post-transmission ringing and, therefore, the need for longer blanking intervals. Most ultrasonic sensors used by mobile robots have an effective range of roughly 12 cm to 5 m. The published accuracy of commercial ultrasonic sensors varies between 98% and 99.1%. In mobile robot applications, specific implementations generally achieve a resolution of approximately 2 cm.

In most cases one may want a narrow opening angle for the sound beam in order to also obtain precise directional information about objects that are encountered. This is a major limitation, since sound propagates in a conelike manner (figure 4.13) with opening angles around 20 to 40 degrees. Consequently, when using ultrasonic ranging one does not acquire depth data points but, rather, entire regions of constant depth. This means that the sensor tells us only that there is an object at a certain distance within the area of the measurement cone. The sensor readings must be plotted as segments of an arc (sphere for 3D) and not as

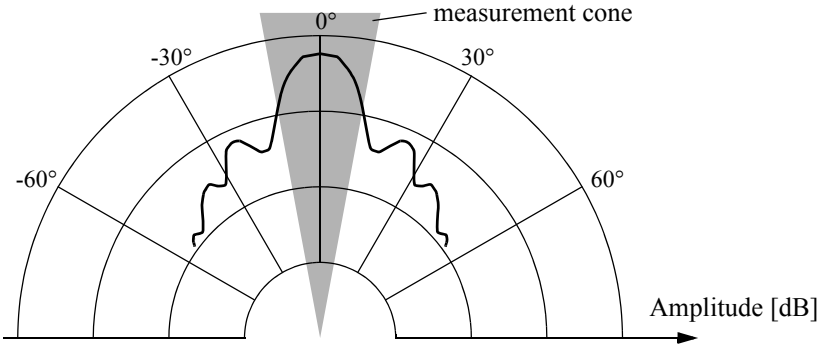


Figure 4.13
Typical intensity distribution of an ultrasonic sensor.

point measurements (figure 4.14). However, recent research developments show significant improvement of the measurement quality in using sophisticated echo processing [149].

Ultrasonic sensors suffer from several additional drawbacks, namely in the areas of error, bandwidth, and cross-sensitivity. The published accuracy values for ultrasonics are nominal values based on successful, perpendicular reflections of the sound wave off an acoustically reflective material. This does not capture the effective error modality seen on a mobile robot moving through its environment. As the ultrasonic transducer's angle to the object being ranged varies away from perpendicular, the chances become good that the sound waves will coherently reflect away from the sensor, just as light at a shallow angle reflects off of a smooth surface. Therefore, the true error behavior of ultrasonic sensors is compound, with a well-understood error distribution near the true value in the case of a successful retroreflection, and a more poorly understood set of range values that are grossly larger than the true value in the case of coherent reflection. Of course, the acoustic properties of the material being ranged have direct impact on the sensor's performance. Again, the impact is discrete, with one material possibly failing to produce a reflection that is sufficiently strong to be sensed by the unit. For example, foam, fur, and cloth can, in various circumstances, acoustically absorb the sound waves.

A final limitation of ultrasonic ranging relates to bandwidth. Particularly in moderately open spaces, a single ultrasonic sensor has a relatively slow cycle time. For example, measuring the distance to an object that is 3 m away will take such a sensor 20 ms, limiting its operating speed to 50 Hz. But if the robot has a ring of twenty ultrasonic sensors, each firing sequentially and measuring to minimize interference between the sensors, then the ring's cycle time becomes 0.4 seconds and the overall update frequency of any one sensor is just 2.5 Hz. For a robot conducting moderate speed motion while avoiding obstacles

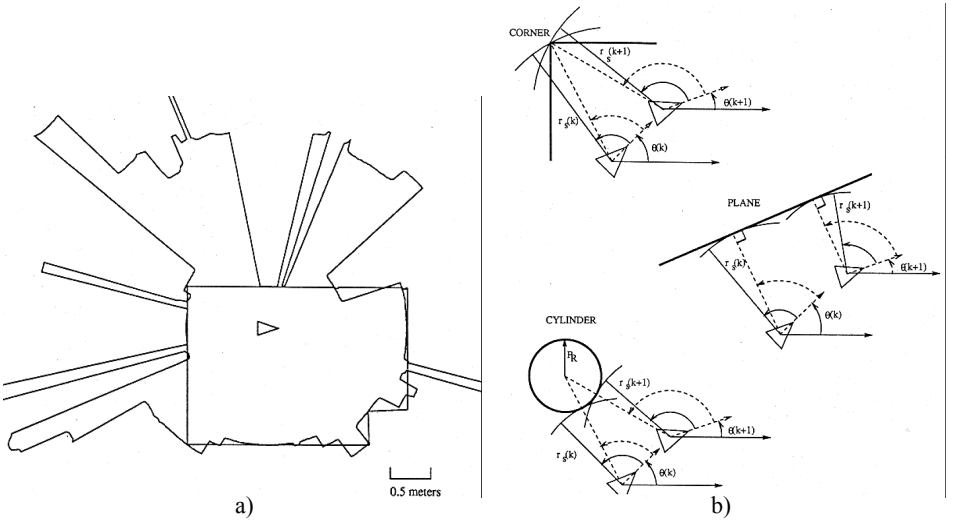


Figure 4.14
Typical readings of an ultrasonic system: (a) 360 degree scan; (b) results from different geometric primitives [35]. Courtesy of John Leonard, MIT.

using ultrasonics, this update rate can have a measurable impact on the maximum speed possible while still sensing and avoiding obstacles safely.

Laser rangefinder (time-of-flight, electromagnetic). The laser rangefinder is a time-of-flight sensor that achieves significant improvements over the ultrasonic range sensor owing to the use of laser light instead of sound. This type of sensor consists of a transmitter that illuminates a target with a collimated beam (e.g., laser), and a receiver capable of detecting the component of light, which is essentially coaxial with the transmitted beam. Often referred to as optical radar or *lidar* (light detection and ranging), these devices produce a range estimate based on the time needed for the light to reach the target and return. A mechanical mechanism with a mirror sweeps the light beam to cover the required scene in a plane or even in three dimensions, using a rotating, nodding mirror.

One way to measure the time of flight for the light beam is to use a pulsed laser and then measure the elapsed time directly, just as in the ultrasonic solution described earlier. Electronics capable of resolving picoseconds are required in such devices and they are therefore

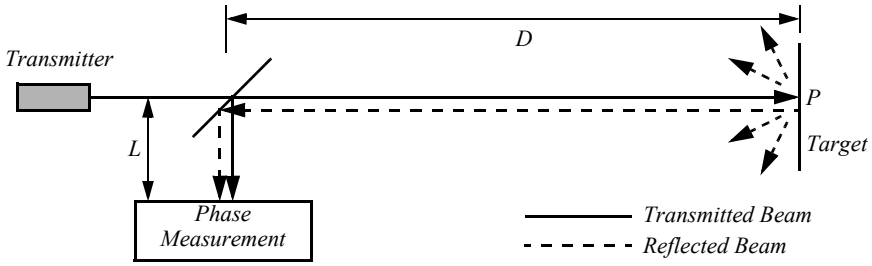


Figure 4.15
Schematic of laser rangefinding by phase-shift measurement.

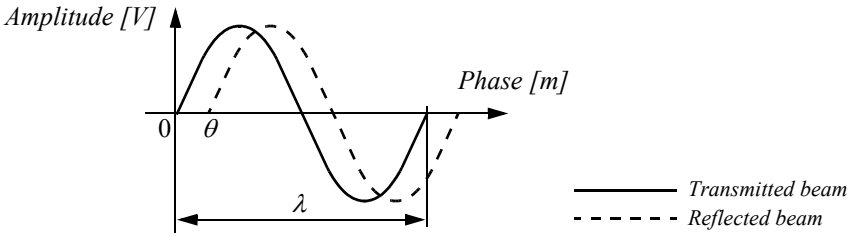


Figure 4.16
Range estimation by measuring the phase shift between transmitted and received signals.

very expensive. A second method is to measure the beat frequency between a frequency-modulated continuous wave (FMCW) and its received reflection. Another, even easier method is to measure the phase shift of the reflected light. We describe this third approach in detail.

Phase-shift measurement. Near-infrared light (from a light-emitting diode [LED] or laser) is collimated and transmitted from the transmitter in figure 4.15 and hits a point *P* in the environment. For surfaces having a roughness greater than the wavelength of the incident light, diffuse reflection will occur, meaning that the light is reflected almost isotropically. The wavelength of the infrared light emitted is 824 nm, and so most surfaces, with the exception of only highly polished reflecting objects, will be diffuse reflectors. The component of the infrared light that falls within the receiving aperture of the sensor will return almost parallel to the transmitted beam for distant objects.

The sensor transmits 100% amplitude-modulated light at a known frequency and measures the phase shift between the transmitted and reflected signals. Figure 4.16 shows how this technique can be used to measure range. The wavelength of the modulating signal obeys the equation $c = f \cdot \lambda$ where c is the speed of light and f the modulating frequency.

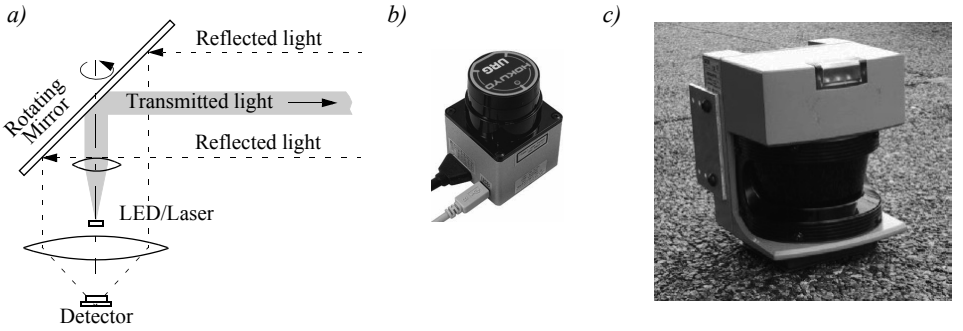


Figure 4.17
(a) Schematic drawing of laser range sensor with rotating mirror; (b) 240-degree laser rangefinder from Hokuyo Ltd.; (c) Industrial 180 degree laser range sensor from Sick Inc., Germany.

For f = 5 MHz (as in the AT&T sensor), λ = 60 m. The total distance D' covered by the emitted light is

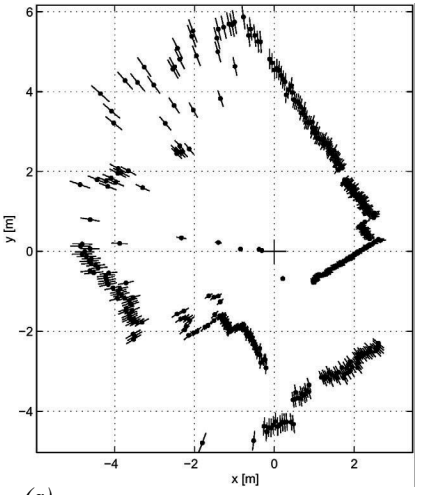
$$D' = L + 2D = L + \frac{\theta}{2\pi}\lambda, \tag{4.23}$$

where D and L are the distances defined in figure 4.15. The required distance D between the beam splitter and the target is therefore given by

$$D = \frac{\lambda}{4\pi}\theta, \tag{4.24}$$

where θ is the electronically measured phase difference between the transmitted and reflected light beams, and λ the known modulating wavelength. It can be seen that the transmission of a single frequency-modulated wave can theoretically result in ambiguous range estimates since, for example, if λ = 60 m, a target at a range of 5 m would give an indistinguishable phase measurement from a target at 35 m, since each phase angle would be 360 degrees apart. We therefore define an "ambiguity interval" of λ, but in practice we note that the range of the sensor is much lower than λ due to the attenuation of the signal in air.

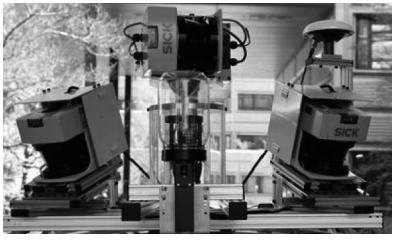
It can be shown that the confidence in the range (phase estimate) is inversely proportional to the square of the received signal amplitude, directly affecting the sensor's accuracy. Hence dark, distant objects will not produce as good range estimates as close, bright objects.



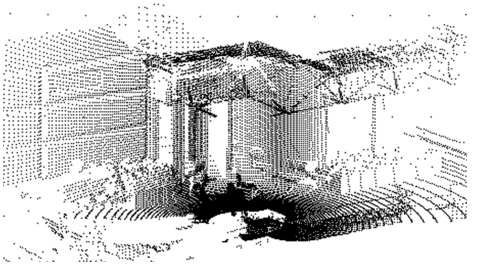
(a)

(b)

(c)



(d)



(e)

Figure 4.18
(a) Typical range image of a 2D laser range sensor with a rotating mirror. The length of the lines through the measurement points indicate the uncertainties. (b) Stanley, the autonomous car from Stanford winning the 2005 Darpa Grand Challenge. (c) The Smarter, the autonomous car developed at the ASL (ETH Zurich). (d) A close view at the Sicks used on the Smarter. (e) A 3D laser-point-cloud built from the rotating Sicks.

In figure 4.17, the schematic of a typical 360-degree laser range sensor and two examples are shown. Figure 4.18a shows a typical range image of a 360-degree scan taken with a laser range sensor.

As expected, the angular resolution of laser rangefinders far exceeds that of ultrasonic sensors. The Sick LMS 200 laser scanner shown in figure 4.17c achieves an angular resolution of 0.25 degree. Depth resolution ranges between 10 and 15 mm and the typical accuracy is 35 mm, over a range from 5 cm up to 20 m or more (up to 80 m), depending on the reflectivity of the object being ranged. This device performs seventy five 180-degrees scans per second but has no mirror nodding capability for the vertical dimension.

As an example of use in mobile robotics, five Sick lasers were used for short range detection on Stanley (figure 4.18b), the autonomous car that won the 2005 DARPA Grand Challenge. In a different configuration, five Sick lasers were also used on the Smarter (figure 4.18c), the autonomous car developed at the ASL (ETH Zurich), which participated in ELROB 2006, the European Land Robot Trial. On the Smarter, one Sick laser at the lower front was used for close obstacle avoidance, while two lasers on the roof (figure 4.18d), slightly canted to the sides, were used for local navigation. Finally, another two lasers, mounted vertically on a turntable (figure 4.18d), were used as a 3D range scanner for 3D mapping.

As with ultrasonic ranging sensors, an important error mode involves coherent reflection of the energy. With light, this will occur only when striking a highly polished surface. Practically, a mobile robot may encounter such surfaces in the form of a polished desktop, file cabinet or, of course, a mirror. Unlike ultrasonic sensors, laser rangefinders cannot detect the presence of optically transparent materials such as glass, and this can be a significant obstacle in environments like, for example, museums, where glass is commonly used

3D laser rangefinders. A 3D laser rangefinder is a laser scanner that acquires scan data in more than a single plane. Custom-made 3D scanners are typically built by nodding or rotating a 2D scanner in a stepwise or continuous manner around an axis parallel to the scanning plane. An example custom-built 3D scanner was developed at the ASL for the Smarter (figure 4.18d). In this case, two Sick lasers were positioned to look into opposite directions. This way, after half rotation of the turntable, a full 3D scan of the environment around the vehicle could be acquired. This data was mainly used to compute a consistent 3D digital terrain model of the environment (figure 4.18e). By lowering the rotational speed of the turntable, the angular resolution in the horizontal direction can be made as small as desired. The advantage of this setting is that the full spherical field of view can be covered (360° in azimuth and 90° in elevation). The drawback is that the acquisition time for a full 3D scan can take up to several seconds depending on the desired resolution. For instance, consider that our Sick scanner acquires 75 vertical plane scans per second and that we need an azimuthal angular resolution of 0.25 degrees. The period for a half-rotation of the turntable necessary to capture a spherical 3D scan with two Sicks is then 360 / 0.25 / 75 / 2 = 9.6 seconds. If one is satisfied with an azimuthal angular resolution of 1 degree, then the acquisition time drops down to 2.4 seconds. This, of course, limits the use of this configuration

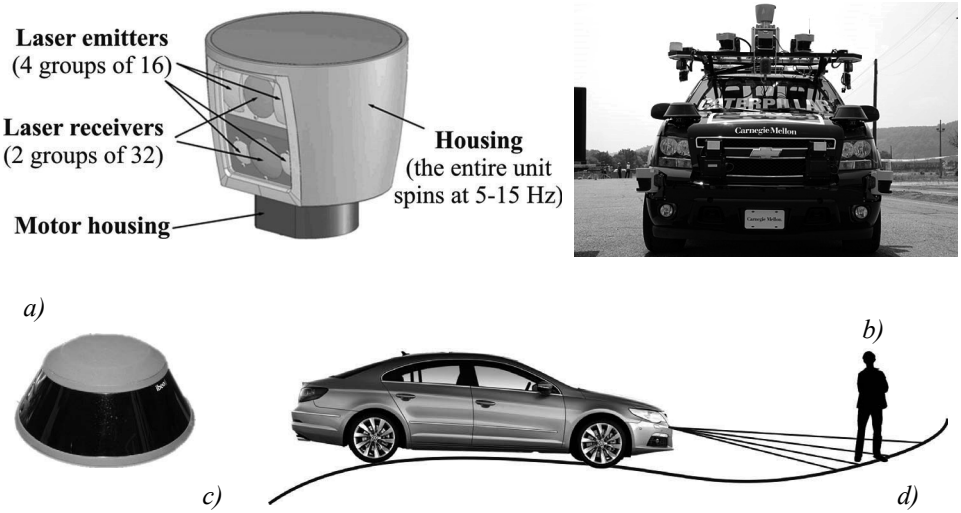


Figure 4.19 (a) The Velodyne HDL-64E unit features 64 laser beams and spins up to 15 Hz to gather data (image courtesy of Velodyne—http://www.velodyne.com/lidar). It delivers over 1.3 million data points per second. (d) The working principle of the Ibeo Alasca XT (c), which uses a four-layer laser beam. The Alasca and Velodyne were used by the CMU Tartan Racing team (b) (image courtesy of the Tartan Racing team).

to static environments. As a matter of fact, the rotating Sicks developed at the ASL (ETH Zurich) were used on an autonomous car running at 10 km/h. In this case, very accurate (up to centimeter) vehicle motion estimation was necessary to correct the errors in the 3D data caused by the movement of the car.

The Velodyne HDL-64E (figure 4.19a) overcomes the drawbacks of custom-made 3D laser range finders. This sensor is a 3D lidar that uses 64 laser emitters instead of the single one used in the Sick. This device spins at rates of 5–15 Hz and delivers more than 1.3 million data points per second. The field of view is 360° in azimuth and 26.8° in elevation and the angular resolution is 0.09° and 0.4° respectively. The distance accuracy is better than 2 cm and can measure depth up to 50 m, with 10% reflectivity, or up to 120 m, with 80% reflectivity. This sensor was the primary means of terrain map construction and obstacle detection for all the top DARPA 2007 Urban Challenge teams. However, the Velodyne is currently still much more expensive than Sick laser rangefinders.

The laser scanner Alasca XT, produced by Ibeo (figure 4.19c–d), on the other hand, splits the laser beam into four vertical layers. Distance measurements are taken independently for each of these layers with an aperture angle of 3.2°. This sensor is typically used for obstacle and pedestrian detection on cars. Because of its multilayer scanning principle, it allows us any pitching of the vehicle (caused by an uneven surface or driving manoeuvres

(a) *(b)* *(c)*

Figure 4.20 (a) The ZCAM produced by the Israeli developer 3DV Systems; (b) the Swiss Ranger SR3000 produced by the Swiss company MESA; (c) Range image of a chair captured with a time-of-flight camera (image courtesy of S. Gächter).

such as braking and accelerating) to be fully compensated. This sensor was used by the Tartan Racing team in the autonomous car from CMU that won the 2007 Urban Grand Challenge. Additionally, it also used the Velodyne sensor (figure 4.19b).

Time-of-flight camera. A Time-of-Flight camera (TOF camera, figure 4.20) works similarly to a lidar with the advantage that the whole 3D scene is captured at the same time and that there are no moving parts. This device uses a modulated infrared lighting source to determine the distance for each pixel of a Photonic Mixer Device (PMD) sensor. As the illumination source is placed just next to the lens (figure 4.20), the whole system is very compact compared to lidars, stereo vision, or triangulation sensors (see below). In the presence of background light, the image sensor receives an additional illumination signal which disturbs the distance measurement. To eliminate the background part of the signal, the acquisition is done a second time with the illumination switched off. As the scene is captured in one shot, the camera reaches up to 100 frames per second and is therefore ideally suited for real-time applications.

The PMD sensor appeared the first time in 1997, but TOF cameras became popular only a few years later, when the semiconductor processes became fast enough for such devices. This sensor typically covers ranges from 0.5 m up to 8 m, but even larger ranges are possible. The distance resolution is about 1 cm. Typical images sizes are quite small: in the two examples shown in figure 4.20, the Swiss Ranger SR3000 by MESA has 174×144 pixels, while the ZCAM by 3DV Systems has 320×240 pixels with 256 depth levels per pixel. The Swiss Ranger has found many robotic applications including map building, obstacle avoidance, and recognition [134, 330]; the ZCAM has been used in video game consoles for player motion detection and activity recognition.

Because it is very easy to extract distance information from the TOF sensor, TOF cameras use less processing power than stereo vision, where complex correlation algorithms are

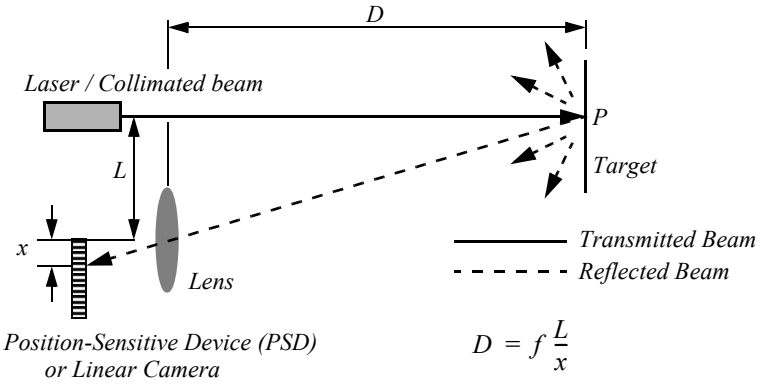


Figure 4.21
Principle of 1D laser triangulation.

used. Additionally, the extracted range information is not disturbed by the patterns on the object as it happens in stereovision.

4.1.9.2 Triangulation active ranging

Triangulation ranging sensors use geometric properties manifest in their measuring strategy to establish distance readings to objects. The simplest class of triangulation rangers are *active* because they project a known light pattern (e.g., a point, a line, or a texture) onto the environment. The reflection of the known pattern is captured by a receiver and, together with known geometric values, the system can use simple triangulation to establish range measurements. If the receiver measures the position of the reflection along a single axis, we call the sensor an optical triangulation sensor in 1D. If the receiver measures the position of the reflection along two orthogonal axes, we call the sensor a structured light sensor. These two sensor types are described in the two sections below.

Optical triangulation (1D sensor). The principle of optical triangulation in 1D is straightforward, as depicted in figure 4.21. A collimated beam (e.g., focused infrared LED, laser beam) is transmitted toward the target. The reflected light is collected by a lens and projected onto a position-sensitive device (PSD) or linear camera. Given the geometry of figure 4.21, the distance D is given by

$$D = f\frac{L}{x}. \tag{4.25}$$

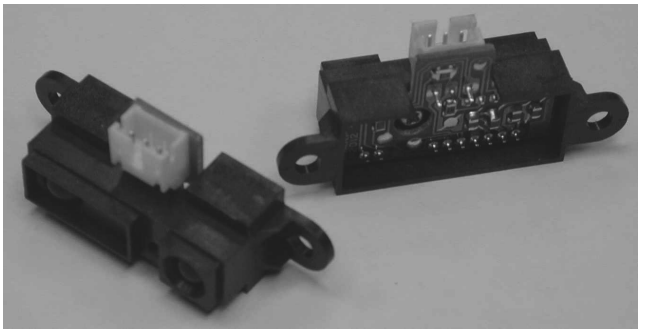


Figure 4.22
A commercially available, low-cost optical triangulation sensor: the Sharp GP series infrared rangefinders provide either analog or digital distance measures and cost only about $15.

The distance is proportional to $1/x$; therefore the sensor resolution is best for close objects and becomes poor at a distance. Sensors based on this principle are used in range sensing up to 1 or 2 m, but also in high-precision industrial measurements with resolutions far below 1 μm.

Optical triangulation devices can provide relatively high accuracy with very good resolution (for close objects). However, the operating range of such a device is normally fairly limited by geometry. For example, the optical triangulation sensor pictured in figure 4.22 operates over a distance range of between 8 and 80 cm. It is inexpensive compared to ultrasonic and laser rangefinder sensors. Although more limited in range than sonar, the optical triangulation sensor has high bandwidth and does not suffer from cross-sensitivities that are more common in the sound domain.

Structured light (2D sensor). If one replaces the linear camera or PSD of an optical triangulation sensor with a 2D receiver such as a CCD or CMOS camera, then one can recover distance to a large set of points instead of to only one point. The emitter must project a known pattern, or *structured light*, onto the environment. Many systems exist which either project light textures (figure 4.23b) or emit collimated light (possibly laser) by means of a rotating mirror. Yet another popular alternative is to project a laser stripe (figure 4.23a) by turning a laser beam into a plane using a prism. Regardless of how it is created, the projected light has a known structure, and therefore the image taken by the CCD or CMOS receiver can be filtered to identify the pattern's reflection.

Note that the problem of recovering depth is in this case far simpler than the problem of passive image analysis. In passive image analysis, as we discuss later, existing features in the environment must be used to perform *correlation*, while the present method projects a

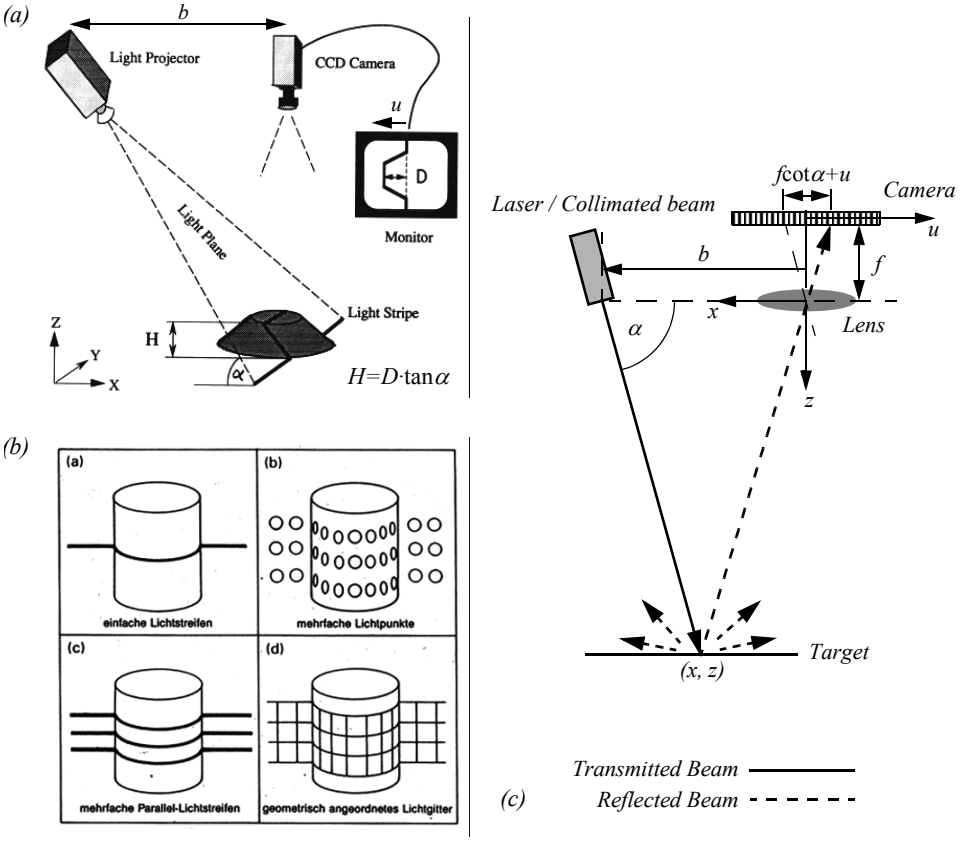


Figure 4.23
(a) Principle of active two dimensional triangulation. (b) Other possible light structures. (c) 1D schematic of the principle. Images (a) and (b) courtesy of Albert-Jan Baerveldt, Halmstad University.

known pattern upon the environment and thereby avoids the standard correlation problem altogether. Furthermore, the structured light sensor is an active device so it will continue to work in dark environments as well as environments in which the objects are featureless (e.g., uniformly colored and edgeless). In contrast, stereo vision would fail in such texture-free circumstances.

Figure 4.23c shows a 1D active triangulation geometry. We can examine the trade-off in the design of triangulation systems by examining the geometry in figure 4.23c. The measured values in the system are α and u, the distance of the illuminated point from the origin in the imaging sensor. Note the imaging sensor here can be a camera or an array of photo diodes of a position-sensitive device (e.g., a PSD).

From figure 4.23c, simple geometry shows that

$$x = \frac{b \cdot u}{f\cot\alpha + u} \; ; \quad z = \frac{b \cdot f}{f\cot\alpha + u}, \tag{4.26}$$

where f is the distance of the lens to the imaging plane. In the limit, the ratio of image resolution to range resolution is defined as the triangulation gain G_p and from equation (4.26) is given by

$$\frac{\partial u}{\partial z} = G_p = \frac{b \cdot f}{z^2}. \tag{4.27}$$

This shows that the ranging accuracy, for a given image resolution, is proportional to source/detector separation b and focal length f, and decreases with the square of the range z. In a scanning ranging system, there is an additional effect on the ranging accuracy, caused by the measurement of the projection angle α. From equation 4.26 we see that

$$\frac{\partial \alpha}{\partial z} = G_\alpha = \frac{b \sin\alpha^2}{z^2}. \tag{4.28}$$

We can summarize the effects of the parameters on the sensor accuracy as follows:

- *Baseline length (b)*: the smaller b is, the more compact the sensor can be. The larger b is, the better the range resolution will be. Note also that although these sensors do not suffer from the correspondence problem, the disparity problem still occurs. As the baseline length b is increased, one introduces the chance that, for close objects, the illuminated point(s) may not be in the receiver's field of view.
- *Detector length and focal length (f)*: A larger detector length can provide either a larger field of view or an improved range resolution or partial benefits for both. Increasing the detector length, however, means a larger sensor head and worse electrical characteristics (increase in random error and reduction of bandwidth). Also, a short focal length gives a large field of view at the expense of accuracy, and vice versa.

At one time, laser stripe structured light sensors were common on several mobile robot bases as an inexpensive alternative to laser rangefinding devices. However, with the increasing quality of laser rangefinding sensors in the 1990s, the structured light system has become relegated largely to vision research rather than applied mobile robotics. However, new possibilities of applications for robotics have recently been opened by Kinect, the sensor released in 2010 within the Microsoft Xbox 360 videogame console, and produced

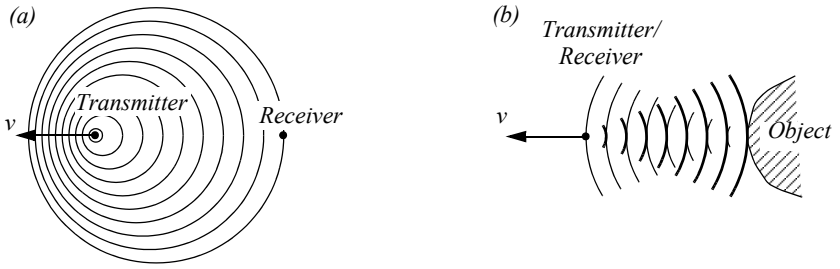


Figure 4.24
Doppler effect between two moving objects (a) or a moving and a stationary object (b).

by the Israeli company PrimeSense. Kinect is a very cheap range camera that uses the structured-light principle explained before. An infrared laser emitter is used to make the projected pattern invisible to the human eye [132].

4.1.10 Motion/speed sensors

Some sensors measure directly the relative motion between the robot and its environment. Since such motion sensors detect relative motion, so long as an object is moving relative to the robot's reference frame, it will be detected and its speed can be estimated. There are a number of sensors that inherently measure some aspect of motion or change. For example, a pyroelectric sensor detects change in heat. When a human walks across the sensor's field of view, his or her motion triggers a change in heat in the sensor's reference frame. In the next section, we describe an important type of motion detector based on the Doppler effect. These sensors represent a well-known technology with decades of general applications behind them. For fast-moving mobile robots such as autonomous highway vehicles and unmanned flying vehicles, Doppler motion detectors are the obstacle detection sensor of choice.

4.1.10.1 Doppler effect sensing (radar or sound)

Anyone who has noticed the change in siren pitch that occurs when an approaching fire engine passes by and recedes is familiar with the Doppler effect.

A transmitter emits an electromagnetic or sound wave with a frequency f_t. It is either received by a receiver (figure 4.24a) or reflected from an object (figure 4.24b). The measured frequency f_r at the receiver is a function of the relative speed v between transmitter and receiver according to

$$f_r = f_t \frac{1}{1 + v/c}, \tag{4.29}$$

if the transmitter is moving and

$$f_r = f_t(1 + v/c), \tag{4.30}$$

if the receiver is moving.

In the case of a reflected wave (figure 4.24b) there is a factor of 2 introduced, since any change x in relative separation affects the round-trip path length by $2x$. Furthermore, in such situations it is generally more convenient to consider the change in frequency Δf, known as the *Doppler shift*, as opposed to the *Doppler frequency* notation above.

$$\Delta f = f_t - f_r = \frac{2f_t v \cos\theta}{c}, \tag{4.31}$$

$$v = \frac{\Delta f \cdot c}{2f_t \cos\theta}, \tag{4.32}$$

where

Δf = Doppler frequency shift;

θ = relative angle between direction of motion and beam axis.

The Doppler effect applies to sound and electromagnetic waves. It has a wide spectrum of applications:

- *Sound waves*: for example, industrial process control, security, fish finding, measure of ground speed.
- *Electromagnetic waves*: for example, vibration measurement, radar systems, object tracking.

A current application area is both autonomous and manned highway vehicles. Both microwave and laser radar systems have been designed for this environment. Both systems have equivalent range, but laser can suffer when visual signals are deteriorated by environmental conditions such as rain, fog, and so on. Commercial microwave radar systems are already available for installation on highway trucks. These systems are called VORAD (vehicle on-board radar) and have a total range of approximately 150 m. With an accuracy of approximately 97%, these systems report range rates from 0 to 160 km/hr with a resolution of 1 km/hr. The beam is approximately 4 degrees wide and 5 degrees in elevation. One of the key limitations of radar technology is its bandwidth. Existing systems can provide information on multiple targets at approximately 2 Hz.

4.1.11 Vision sensors

Vision is our most powerful sense. It provides us with an enormous amount of information about the environment and enables rich, intelligent interaction in dynamic environments. It is therefore not surprising that a great deal of effort has been devoted to providing machines with sensors that mimic the capabilities of the human vision system. The first step in this process is the creation of sensing devices that capture the light and convert it into a digital image. The second step is the processing of the digital image in order to get salient information like depth computation, motion detection, color tracking, feature detection, scene recognition, and so on. Because vision sensors have become very popular in robotic applications, the remaining sections of this chapter will be dedicated to the fundamentals of computer vision and image processing and their use in robotics.

4.2 Fundamentals of Computer Vision

4.2.1 Introduction

The analysis of images and their processing are two major fields that are known as computer vision and image processing. The years between 1980 and 2010 have seen significant advances and new theoretical findings in these fields and some of the most sophisticated computer vision and image processing techniques have found many industrial applications in consumer cameras, photography, defect inspection, monitoring and surveillance, video games, movies, and the like. For more information on the computer vision industry, see [346].

The remaining parts of this chapter are dedicated to these two fields. First, we will introduce the working principle of the digital camera, the imaging sensors, the optics, and the image formation; then, we will present two ways of estimating the depth, which are depth from focus and stereo vision. Next, we will detail some of the most important tools used in image processing. Finally, we will close this chapter by presenting state-of-the-art algorithms for feature extraction and place recognition from digital images.

For an in-depth study of computer vision, we refer the reader to the following books: [21, 29, 36, 49, 53].

4.2.2 The digital camera

After starting from one or more light sources, reflecting off of one or more surfaces in the world, and passing through the camera's optics (lenses), light finally reaches the imaging sensor. How are the photons arriving at this sensor converted into the digital (R,G,B) values that we observe when we look at a digital image?

Light falling on an imaging sensor is usually picked up by an active sensing area, integrated for the duration of the exposure (usually expressed as the shutter speed, e.g., 1/125, 1/60, 1/30 of a second), and then passed to a set of sense amplifiers. The two main kinds of

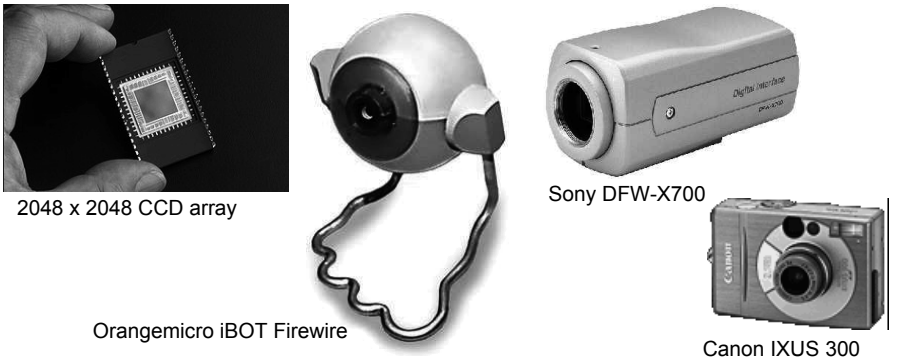


Figure 4.25
Commercially available CCD chips and CCD cameras. Because this technology is relatively mature, cameras are available in widely varying forms and costs.

sensors used in digital still and video cameras today are CCD (charge coupled device) and CMOS (complementary metal oxide on silicon). Below, we review the advantages and drawbacks of these two technologies.

CCD cameras. The CCD chip (see figure 4.25) is an array of light-sensitive picture elements, or pixels, usually with between 20,000 and several million pixels total. Each pixel can be thought of as a light-sensitive, discharging capacitor that is 5 to 25 μm in size. First, the capacitors of all pixels are charged fully, then the integration period begins. As photons of light strike each pixel, they liberate electrons, which are captured by electric fields and retained at the pixel. Over time, each pixel accumulates a varying level of charge based on the total number of photons that have struck it. After the integration period is complete, the relative charges of all pixels need to be frozen and read. In a CCD, the reading process is performed at one corner of the CCD chip. The bottom row of pixel charges is transported to this corner and read, then the rows above shift down and the process is repeated. This means that each charge must be transported across the chip, and it is critical that the value be preserved. This requires specialized control circuitry and custom fabrication techniques to ensure the stability of transported charges.

The photodiodes used in CCD chips (and CMOS chips as well) are not equally sensitive to all frequencies of light. They are sensitive to light between 400 and 1000 nm wavelength. It is important to remember that photodiodes are less sensitive to the ultraviolet end of the spectrum (e.g., blue) and are overly sensitive to the infrared portion (e.g., heat).

The CCD camera has several camera parameters that affect its behavior. In some cameras, these values are fixed. In others, the values are constantly changing based on built-in

feedback loops. In higher-end cameras, the user can modify the values of these parameters via software. The *iris position* and *shutter speed* regulate the amount of light being measured by the camera. The iris is simply a mechanical aperture that constricts incoming light, just as in standard 35 mm cameras. Shutter speed regulates the integration period of the chip. In higher-end cameras, the effective shutter speed can be as brief at 1/30,000 seconds and as long as 2 seconds. *Camera gain* controls the overall amplification of the analog signal, prior to A/D conversion. However, it is very important to understand that even though the image may appear brighter after setting high gain, the shutter speed and iris may not have changed at all. Thus gain merely amplifies the signal, and it amplifies along with the signal all of the associated noise and error.

The key disadvantages of CCD cameras are primarily in the areas of inconstancy and dynamic range. As mentioned earlier, a number of parameters can change the brightness and colors with which a camera creates its image. Manipulating these parameters in a way to provide consistency over time and over environments, for example, ensuring that a green shirt always looks green, and something dark gray is always dark gray, remains an open problem in the vision community. For more details on the fields of color constancy and luminosity constancy, consult [65].

The second class of disadvantages relates to the behavior of a CCD chip in environments with extreme illumination. In cases of very low illumination, each pixel will receive only a small number of photons. The longest possible integration period (i.e., shutter speed) and camera optics (i.e., pixel size, chip size, lens focal length and diameter) will determine the minimum level of light for which the signal is stronger than random error noise. In cases of very high illumination, a pixel fills its well with free electrons. As the well reaches its limit, the probability of trapping additional electrons falls, and therefore the linearity between incoming light and electrons in the well degrades. Termed *saturation*, this can indicate the existence of a further problem related to cross-sensitivity. When a well has reached its limit, then additional light within the remainder of the integration period may cause further charge to leak into neighboring pixels, causing them to report incorrect values or even reach secondary saturation. This effect, called *blooming*, means that individual pixel values are not truly independent.

The camera parameters may be adjusted for an environment with a particular light level, but the problem remains that the dynamic range of a camera is limited by the well capacity (also called *well depth*) of the individual pixels. The well depth typically ranges between 20,000 and 350,000 electrons. For example, a high-quality CCD may have pixels that can hold 40,000 electrons. The noise level for reading the well may be 11 electrons, and therefore the dynamic range will be 40,000:11, or 3600:1, which is 35 dB.

CMOS cameras. The complementary metal oxide semiconductor chip is a significant departure from the CCD. It, too, has an array of pixels, but located along the side of each

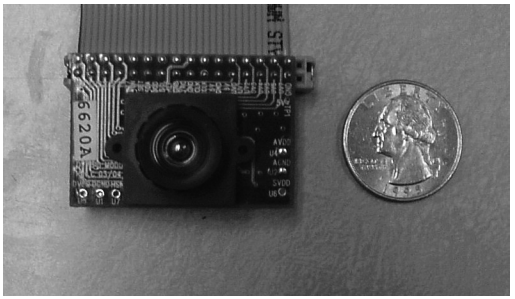


Figure 4.26
A commercially available, low-cost CMOS camera with lens attached.

pixel are several transistors specific to that pixel. As in CCD chips, all of the pixels accumulate charge during the integration period. During the data collection step, the CMOS takes a new approach: the pixel-specific circuitry next to every pixel measures and amplifies the pixel's signal, all in parallel for every pixel in the array. Using more traditional traces from general semiconductor chips, the resulting pixel values are all carried to their destinations.

CMOS has a number of advantages over CCD technologies. First and foremost, there is no need for the specialized clock drivers and circuitry required in the CCD to transfer each pixel's charge down all of the array columns and across all of its rows. This also means that specialized semiconductor manufacturing processes are not required to create CMOS chips. Therefore, the same production lines that create microchips can create inexpensive CMOS chips as well (see figure 4.26). The CMOS chip is so much simpler that it consumes significantly less power; incredibly, it operates with a power consumption that is one-hundredth the power consumption of a CCD chip. In a mobile robot, especially flying, power is a scarce resource and therefore this is an important advantage.

Traditionally, CCD sensors outperformed CMOS in quality sensitive applications such as digital single-lens-reflex cameras, while CMOS was better for low-power applications, but today, CMOS is used in most digital cameras.

Given this summary of the mechanism behind CCD and CMOS chips, one can appreciate the sensitivity of any vision robot sensor to its environment. As compared to the human eye, these chips all have far poorer adaptation, cross-sensitivity, and dynamic range. As a result, vision sensors today continue to be fragile. Only over time, as the underlying performance of imaging chips improves, will significantly more robust vision sensors for mobile robots be available.

Camera output considerations. Although digital cameras have inherently digital output, throughout the 1980s and early 1990s, most affordable vision modules provided analog output signals, such as NTSC (National Television Standards Committee) and PAL (Phase Alternating Line). These camera systems included a D/A converter which, ironically, would be counteracted on the computer using a *frame grabber*, effectively an A/D converter board situated, for example, on a computer's bus. The D/A and A/D steps are far from noise free, and furthermore the color depth of the analog signal in such cameras was optimized for human vision, not computer vision.

More recently, both CCD and CMOS technology vision systems provide digital signals that can be directly utilized by the roboticist. At the most basic level, an imaging chip provides parallel digital I/O (input/output) pins that communicate discrete pixel level values. Some vision modules make use of these direct digital signals, which must be handled subject to hard-time constraints governed by the imaging chip. To relieve the real-time demands, researchers often place an *image buffer chip* between the imager's digital output and the computer's digital inputs. Such chips, commonly used in webcams, capture a complete image snapshot and enable non-real-time access to the pixels, usually in a single, ordered pass.

At the highest level, a roboticist may choose instead to utilize a higher-level digital transport protocol to communicate with an imager. Most common are the IEEE 1394 (Firewire) standard and the USB (and USB 2.0) standards, although some older imaging modules also support serial (RS-232). To use any such high-level protocol, one must locate or create driver code both for that communication layer and for the particular implementation details of the imaging chip. Take note, however, of the distinction between lossless digital video and the standard digital video stream designed for human visual consumption. Most digital video cameras provide digital output, but often only in compressed form. For vision researchers, such compression must be avoided as it not only discards information but even introduces image detail that does not actually exist, such as MPEG (Moving Picture Experts Group) discretization boundaries.

Color camera. The basic light-measuring process described before is colorless: it is just measuring the total number of photons that strike each pixel in the integration period. There are two common approaches for creating *color* images, which use a single chip or three separate chips.

The single chip technology uses the so-called *Bayer* filter. The pixels on the chip are grouped into 2×2 sets of four, then red, green, and blue color filters are applied so that each individual pixel receives only light of one color. Normally, two pixels of each 2×2 block measure green while the remaining two pixels measure red and blue light intensity (figure 4.27). The reason there are twice as many green filters as red and blue is that the

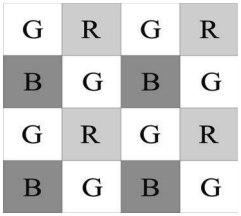


Figure 4.27 Bayer color filter array.

luminance signal is mostly determined by green values, and the visual system is much more sensitive to high frequency detail in luminance than in chrominance. The process of interpolating the missing color values so that we have valid RGB values as all the pixels is known as *demosaicing*. Of course, this one-chip technology has a geometric resolution disadvantage. The number of pixels in the system has been effectively cut by a factor of four, and therefore the image resolution output by the camera will be sacrificed.

The three-chip color camera avoids these problems by splitting the incoming light into three complete (lower intensity) copies. Three separate chips receive the light, with one red, green, or blue filter over each entire chip. Thus, in parallel, each chip measures light intensity for one color, and the camera must combine the chips' outputs to create a joint color image. Resolution is preserved in this solution, although the three-chip color cameras are, as one would expect, significantly more expensive and therefore more rarely used in mobile robotics.

Both three-chip and single-chip color cameras suffer from the fact that photodiodes are much more sensitive to the near-infrared end of the spectrum. This means that the overall system detects blue light much more poorly than red and green. To compensate, the gain must be increased on the blue channel, and this introduces greater absolute noise on blue than on red and green. It is not uncommon to assume at least one to two bits of additional noise on the blue channel. Although there is no satisfactory solution to this problem today, over time the processes for blue detection have been improved, and we expect this positive trend to continue.

In color cameras, an additional control exists for *white balance*. Depending on the source of illumination in a scene (e.g., fluorescent lamps, incandescent lamps, sunlight, underwater filtered light, etc.), the relative measurements of red, green, and blue light that define pure white light will change dramatically. The human eye compensates for all such effects in ways that are not fully understood, but, the camera can demonstrate glaring inconsistencies in which the same table looks blue in one image, taken during the night, and yellow in another image, taken during the day. White balance controls enable the user to change the relative gains for red, green, and blue in order to maintain more consistent color definitions in varying contexts.

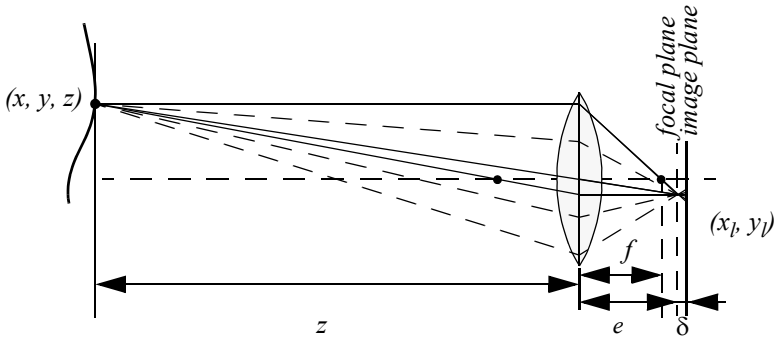


Figure 4.28
Depiction of the camera optics and its impact on the image. In order to get a sharp image, the image plane must coincide with the focal plane. Otherwise the image of the point (x,y,z) will be blurred in the image, as can be seen in the drawing above.

4.2.3 Image formation

Before we can intelligently analyze and manipulate images, we need to understand the image formation process that produced a particular image.

4.2.3.1 Optics

Once the light from the scene reaches the camera, it must still pass through the lens before reaching the sensor. Figure 4.28 shows a diagram of the most basic lens model, which is the thin lens. This lens is composed of a single piece of glass with very low, equal curvature on both sides. According to the lens law (which can be derived using simple geometric arguments on light ray refraction), the relationship between the distance to an object z and the distance behind the lens at which a focused image is formed e can be expressed as

$$\frac{1}{f} = \frac{1}{z} + \frac{1}{e}, \tag{4.33}$$

where f is the focal length. As you can perceive, this formula can also be used to estimate the distance to an object by knowing the focal length and the current distance of the image plane to the lens. This technique is called *depth from focus*.

If the image plane is located at distance e from the lens, then for the specific object voxel depicted, all light will be focused at a single point on the image plane and the object voxel will be *focused*. However, when the image plane is not at e, as is depicted in figure 4.28, then the light from the object voxel will be cast on the image plane as a *blur circle* (or *circle of confusion*). To a first approximation, the light is homogeneously distributed

throughout this blur circle, and the radius R of the circle can be characterized according to the equation

$$R = \frac{L\delta}{2e}. \tag{4.34}$$

L is the diameter of the lens or aperture, and δ is the displacement of the image plane from the focal point.

Given these formulas, several basic optical effects are clear. For example, if the aperture or lens is reduced to a point, as in a pinhole camera, then the radius of the blur circle approaches zero. This is consistent with the fact that decreasing the iris aperture opening causes the *depth of field* to increase until all objects are in focus. Of course, the disadvantage of doing so is that we are allowing less light to form the image on the image plane, and so this is practical only in bright circumstances.

The second property that can be deduced from these optics equations relates to the sensitivity of blurring as a function of the distance from the lens to the object. Suppose the image plane is at a fixed distance 1.2 from a lens with diameter $L = 0.2$ and focal length $f = 0.5$. We can see from equation (4.34) that the size of the blur circle R changes proportionally with the image plane displacement δ. If the object is at distance $z = 1$, then from equation (4.33) we can compute $e = 1$, and therefore $\delta = 0.2$. Increase the object distance to $z = 2$ and as a result $\delta = 0.533$. Using equation (4.34) in each case, we can compute $R = 0.02$ and $R = 0.08$ respectively. This demonstrates high sensitivity for defocusing when the object is close to the lens.

In contrast, suppose the object is at $z = 10$. In this case we compute $e = 0.526$. But if the object is again moved one unit, to $z = 11$, then we compute $e = 0.524$. The resulting blur circles are $R = 0.117$ and $R = 0.129$, far less than the quadrupling in R when the obstacle is one-tenth the distance from the lens. This analysis demonstrates the fundamental limitation of depth from focus techniques: they lose sensitivity as objects move farther away (given a fixed focal length). Interestingly, this limitation will turn out to apply to virtually all visual ranging techniques, including depth from stereo (section 4.2.5) and depth from motion (section 4.2.6).

Nevertheless, camera optics can be customized for the depth range of the intended application. For example, a zoom lens with a very large focal length f will enable range resolution at significant distances, of course at the expense of field of view. Similarly, a large lens diameter, coupled with a very fast shutter speed, will lead to larger, more detectable blur circles.

Given the physical effects summarized by the above equations, one can imagine a visual ranging sensor that makes use of multiple images in which camera optics are varied (e.g., image plane displacement δ) and the same scene is captured (see figure 4.29). In fact, this

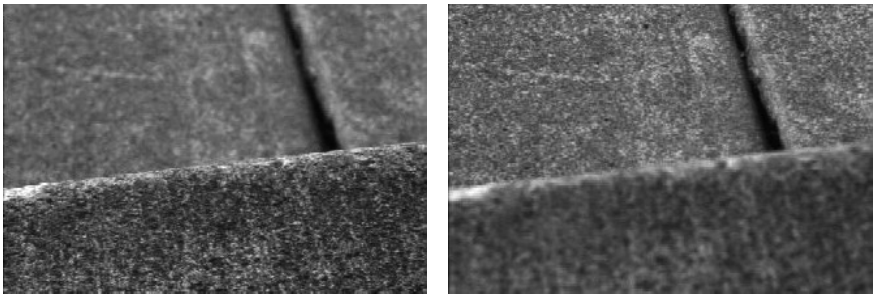


Figure 4.29
Two images of the same scene taken with a camera at two different focusing positions. Note the significant change in texture sharpness between the near surface and far surface. The scene is an outdoor concrete step.

approach is not a new invention. The human visual system uses an abundance of cues and techniques, and one system demonstrated in humans is depth from focus. Humans vary the focal length of their lens continuously at a rate of about 2 Hz. Such approaches, in which the lens optics are actively searched in order to maximize focus, are technically called *depth from focus* [241]. In contrast, *depth from defocus* means that depth is recovered using a series of images that have been taken with different camera geometries, and hence different focusing positions.

The depth from focus method is one of the simplest visual ranging techniques. To determine the range to an object, the sensor simply moves the image plane (via focusing) until maximizing the sharpness of the object. When the sharpness is maximized, the corresponding position of the image plane directly reports range. Some autofocus cameras and virtually all autofocus video cameras use this technique. Of course, a method is required for measuring the sharpness of an image or an object within the image.

An example application of depth-from-focus to robotics has been shown in [250], where the authors demonstrated obstacle avoidance in a variety of environments, as well as avoidance of concave obstacles such as steps and ledges.

4.2.3.2 Pinhole camera model

The pinhole camera, or *camera obscura*, has been the first example of camera in the history, which led to the invention of photography [27]. A pinhole camera has no lens, but a single very small aperture. In short, it is a lightproof box with a small hole in one side. Light from the scene passes through this single point and projects an inverted image on the opposite side of the box (figure 4.30). The working principle of this camera was already known as far back as the 4th century BC by the Greek Aristotle and Euclid and the Chinese Mozi.

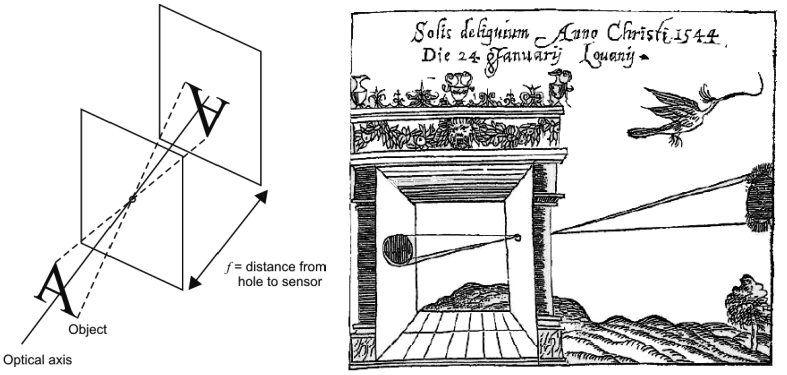


Figure 4.30 (a) When $d \gg f$ and $d \gg L$ the camera can be modeled as a pinhole camera. (b) The *camera obscura* in a drawing from mathematician Reinerus Gemma-Frisius (1508–1555), who used this illustration in his book *De Radio Astronomica et Geometrica* (1545) to describe an eclipse of the sun at Louvain on January 24, 1544. It is thought to be the first published illustration of a camera obscura [27].

The pinhole projection model was also used as drawing aid by artists such as Leonardo da Vinci (1452–1519).

The importance of the pinhole camera is that its principle has also been adopted as a standard model for perspective cameras. This model can be derived directly from equation (4.33). In fact, notice that if we let $z \to \infty$, i.e. we adjust the lens (move the image plane) so that objects at infinity are in focus (i.e. $z \gg f$ and $z \gg L$), we get $e = f$, which is why we can think of a lens of focal length f as being equivalent (to a first approximation) to a pinhole a distance f from the focal plane (figure 4.31a).

When using the pinhole camera model, it is very important to remember that the pinhole corresponds to the center of the lens. This point is also commonly called *center of projection* of *optical center* (indicated with C in figure 4.31). The axis perpendicular to the *image plane* Π, which passes through the center of projection is called *optical axis*.

For convenience, the pinhole camera is commonly represented with the image plane between the center of projection and the scene (figure 4.31(b)). This is done for the image to preserve the same orientation as the object, that is, the image is not flipped. The intersection O between the optical axis and the image plane is called *principal point*.

As shown in figure 4.31b, observe that a camera does not measure distances but angles and therefore it can be thought as a *bearing* sensor.

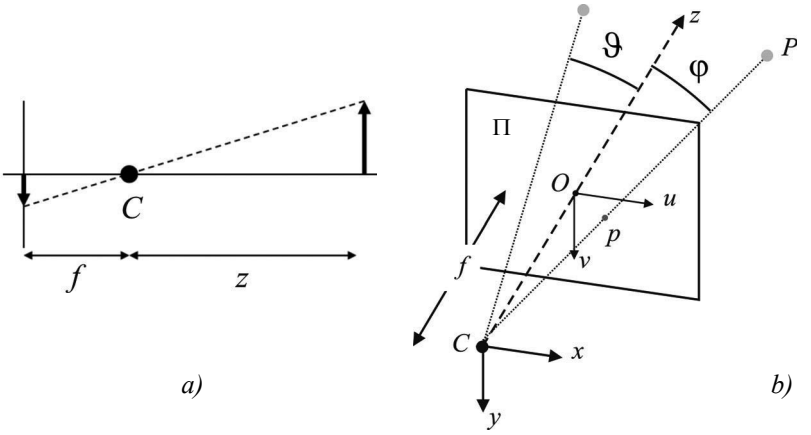


Figure 4.31 (a) Pinhole camera model used for representing standard perspective cameras. (b) The pinhole model is more commonly described with the image plane between the center of projection and the scene for the image to preserve the same orientation as the object.

4.2.3.3 Perspective projection

To describe analytically the *perspective projection* operated by the camera, we have to introduce some opportune reference system wherein we can express the 3D coordinates of the *scene point* P and the coordinates of its projection p on the image plane. We will first consider a simplified model and finally the general model.

Simplified model. Let (x, y, z) be the *camera reference frame* with origin in C and z-axis coincident with the optical axis. Assume also that the camera reference frame coincides with the *world* reference frame. This implies that the coordinates of the scene point P are already expressed in the camera frame.

Let us also introduce a two-dimensional reference frame (u, v) for the image plane Π with origin in O and the u and v axes aligned as x and y respectively as shown in figure 4.31b.

Finally, let $P = (x, y, z)$ and $p = (u, v)$. By means of simple considerations on the similarity of triangles, we can write

$$\frac{f}{z} = \frac{u}{x} = \frac{v}{y}, \tag{4.35}$$

and therefore

$$u = \frac{f}{z} \cdot x \ , \tag{4.36}$$

$$v = \frac{f}{z} \cdot y \ . \tag{4.37}$$

This is the *perspective projection*. The mapping from 3D coordinates to 2D coordinates is clearly nonlinear. However, using *homogeneous coordinates* instead allows us to obtain linear equations. Let

$$\tilde{p} = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \text{ and } \tilde{P} = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \ , \tag{4.38}$$

be the homogeneous coordinates of p and P respectively. We will henceforth use the superscript $\tilde{}$ to denote homogeneous coordinates[3]. The projection equation, in this simplified case, can be written as:

$$\begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} = \begin{bmatrix} fx \\ fy \\ z \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} . \tag{4.39}$$

Note that λ is equal to the third coordinate of P, which—in this special reference frame—coincides with the distance of the point to the plane xy. Note that this equation also shows that every image point is the projection of all infinite 3D points lying on the ray passing through the same image point and the center of projection (figure 4.31b). Therefore, using a single pinhole camera it is not possible to estimate the distance to a point, but we need two cameras (i.e., *stereo camera*, section 4.2.5.2).

General model. A realistic camera model that describes the transformation from 3D coordinates to pixel coordinates must also take into account

3. In homogeneous coordinates we denote 2D points in the image plane as (x_1, x_2, x_3) with $(x_1/x_3, x_2/x_3)$ being the corresponding Cartesian coordinates. Therefore, there is a one to many correspondence between Cartesian and homogeneous coordinates. Homogeneous coordinates can represent the usual Euclidean points plus the points at infinity, which are points with the last component equal to zero that do not have a Cartesian counterpart.

- the *pixelization*, that is, shape (size) of the CCD and its position with respect to the optical center,
- the rigid body transformation between the camera and the scene (i.e., world).

The pixelization takes into account the fact that:

1. The camera optical center has pixel coordinates (u_0, v_0) with respect to the upper left corner of the image, which is commonly assumed as origin of the image coordinate system. Note, the optical center in general does not correspond to the center of the CCD.
2. The coordinates of a point on the image plane are measured in pixels. Therefore, we must introduce a scale factor.
3. The shape of the pixel is in general assumed not perfectly squared and therefore we must use two different scale factors k_u and k_v along the horizontal and vertical directions respectively.
4. The u and v axes might not be orthogonal but misaligned of an angle θ. This models, for instance, the fact that the lens may not be parallel to the CCD.

The first three points are addressed by means of the translation of the optical center and the individual rescaling of the u and v axes:

$$u = k_u \frac{f}{z} \cdot x + u_0 \tag{4.40}$$

$$v = k_v \frac{f}{z} \cdot y + v_0 \ , \tag{4.41}$$

where (u_0, v_0) are the coordinates of the principal point, k_u (k_v) is the inverse of the effective pixel size along the u (v) direction and is measured in $pixel \cdot m^{-1}$.

After this update the perspective projection equations become:

$$\begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} = \begin{bmatrix} fk_u & 0 & u_0 & 0 \\ 0 & fk_v & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \ . \tag{4.42}$$

Observe that we can pose $\alpha_u = fk_u$ and $\alpha_v = fk_v$ which describe the focal lengths expressed in horizontal and vertical pixels respectively.

To take into account the fact that in general the world reference system $(x_w, y_w z_w)$ does not coincide with the camera reference system (x, y, z), we have to introduce the rigid body

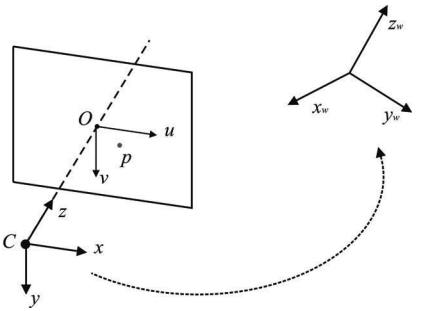


Figure 4.32 Coordinate change between camera and world reference frame.

transformation between the two reference frames (figure 4.32). Let us therefore introduce a coordinate change composed of a rotation R followed by a translation t, therefore

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = R \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} + t \,. \tag{4.43}$$

Using this transformation, equation (4.42) can be rewritten as

$$\begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} = \begin{bmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}, \tag{4.44}$$

or, using the homogeneous coordinates (4.38),

$$\lambda \tilde{p} = A[R|t]\tilde{P}_w \,, \tag{4.45}$$

where

$$A = \begin{bmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.46}$$

is the *intrinsic parameter matrix*.

As anticipated, the most general model also takes into consideration the possibility that the u and v axes are not orthogonal but are inclined of an angle θ. Therefore, the most general form for A is

$$A = \begin{bmatrix} \alpha_u & \alpha_u \cot\theta & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{4.47}$$

where $\alpha_u \cot\theta$ can be absorbed into a single parameter α_c.

α_c, α_u, α_{uv}, u_0, v_0 are called *camera intrinsic parameters.* The rotation and translation parameters R and t are called *camera extrinsic parameters*. The intrinsic and extrinsic parameters can be estimated using a procedure called *camera calibration* that we will shortly describe in section 4.2.3.4.

Radial distortion. The aforementioned image projection model assumes that the camera obeys a linear projection model where straight lines in the world result in straight lines in the image. Unfortunately, many wide-angle lenses have noticeable radial distortion, which manifests itself as a visible curvature in the projection of straight lines. An accurate model of the camera should therefore also take into account the radial distortion of the lens, especially for lenses with short focal length (i.e., large field of view) (figure 4.33).

The standard model of radial distortion is a transformation from the ideal coordinates (i.e., undistorted) (u, v) to the real observable coordinates (distorted) (u_d, v_d). Depending on the type of radial distortion, the coordinates in the observed images are displaced away (*barrel* distortion) or toward (*pincushion* distortion) the image center. The amount of distortion of the coordinates of the observed image is a nonlinear function of their radial distance r. For most lenses, a simple quadratic model of distortion produces good results:

$$\begin{bmatrix} u_d \\ v_d \end{bmatrix} = (1 + k_1 r^2) \begin{bmatrix} u - u_0 \\ v - v_0 \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix}, \tag{4.48}$$

where

$$r^2 = (u - u_0)^2 + (v - v_0)^2. \tag{4.49}$$

and k_1 is the radial distortion parameter, which can be estimated by camera calibration. The radial distortion parameter is also an intrinsic parameter of the camera.

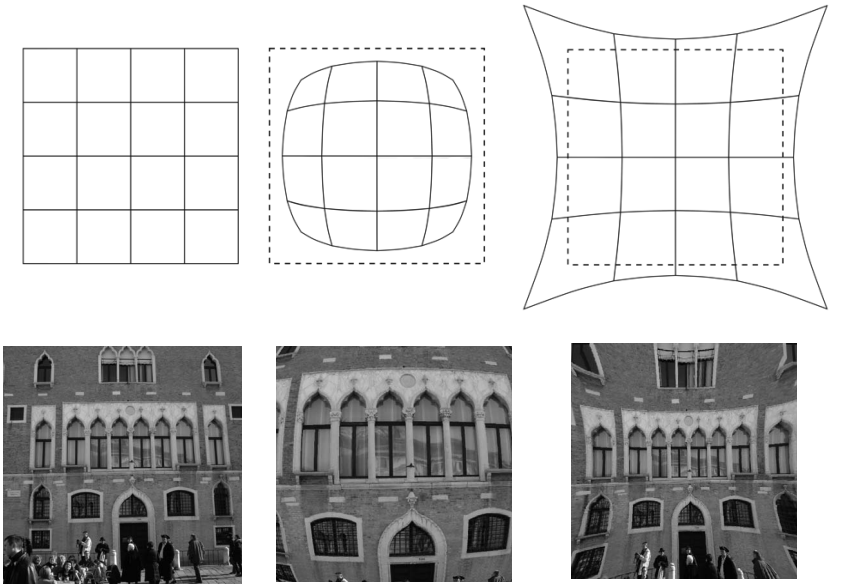


Figure 4.33 Example of radial lens distortion: (a) no distortion, (b) barrel distortion, (c) pincushion.

Sometimes the above simplified model does not model the true distortions produced by complex lenses accurately enough (especially at very wide angles). A more complete analytic model also includes *tangential distortions* and *decentering distortions* [48], but these will not be covered in this book. Fisheye lenses require a different model than traditional polynomial models of radial distortion and will be introduced in section 4.2.4.2.

4.2.3.4 Camera calibration

Calibration consists in measuring accurately the intrinsic and extrinsic parameters of the camera model. As these parameters govern the way the scene points are mapped to their corresponding image points, the idea is that by knowing the pixel coordinates of the image points $\tilde{p}$ and the 3D coordinates of the corresponding scene points $\tilde{P}$, it is possible to compute the unknown parameters A, R, and t by solving the perspective projection equation (4.45).

One of the first and most used camera calibration techniques was proposed in 1987 by Tsai [319]. Its implementation needs corresponding 3D point coordinates and 2D pixels coordinates in the image. It uses a two-stage technique to compute, first, the position and orientation and, second, the internal parameters of the camera.

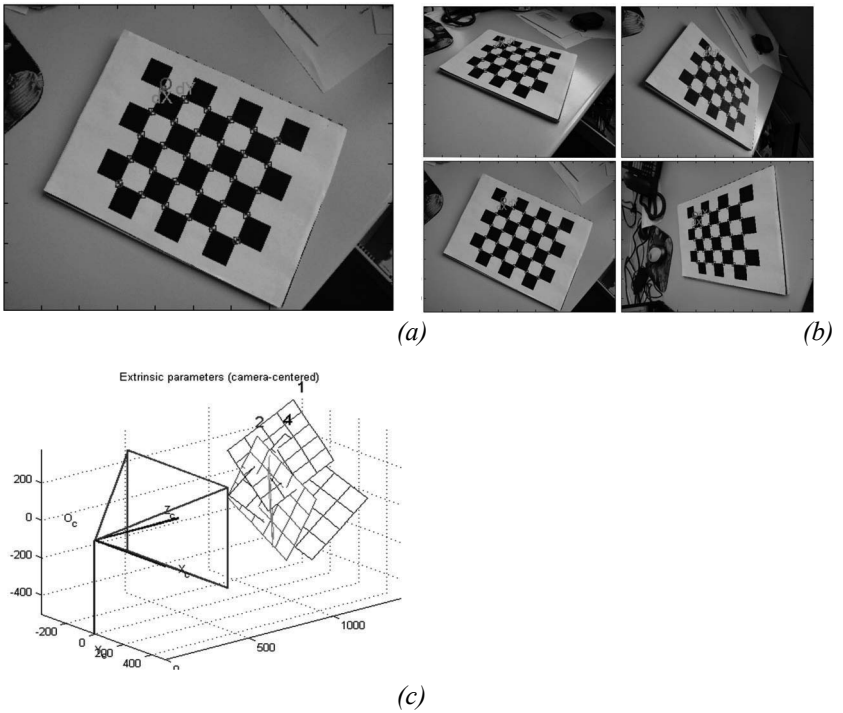


Figure 4.34 Pictures of the Camera Calibration Toolbox for Matlab developed by J. Y. Bouguet. (a) An example checkerboard-like pattern used in camera calibration with extracted corners. (b) Several pictures of the pattern with different positions and orientations. (c) The reconstructed position and orientations of the pattern after calibration.

However, in the last decade an alternative camera calibration technique has been proposed by Zhang [337] that, instead of a three-dimensional calibration object, uses a planar grid. The most common planar grid is a chessboard-like pattern due to the ease of extracting its corners, which are then used for calibration (figure 4.34). This method is known as *calibration from planar grid* and is very simple and practical to execute for both expert and non-expert users. The method requires the user to take several pictures of the pattern shown at different positions and orientations[4]. By knowing the 2D position of the corners on the real pattern and the pixel coordinates of their corresponding corners on each image, the intrinsic and extrinsic parameters (including radial and tangential distortion) are determined simultaneously by solving a least-square linear minimization followed by a nonlin-

4. Note that in this case the number of extrinsic parameters is different for each position of the grid while the intrinsic parameter are obviously the same.

ear refinement (i.e., Gauss-Newton). As pointed out by Zhang, the accuracy of the calibration results increases with the number of images used. It is also important that the images cover as much of the field of view of the camera as possible and that the range of orientations is wide.

This calibration method has been implemented in a very successful open source toolbox for Matlab (which can be downloaded for free [347]) that is also available in C in the open source Computer Vision Library (OpenCV) [343]. This toolbox has been used by thousands of users all around the world and is considered one of the most practical and easy-to-use camera calibration softwares for standard perspective cameras. In this section, we used the same model as the Matlab toolbox. This should facilitate the understanding and implementation of the interested reader. Alternatively, a complete list of all available camera calibration softwares can be found in [348].

4.2.4 Omnidirectional cameras

4.2.4.1 Introduction

In the previous section, we described the image formation of the pinhole camera, which is modeled as a perspective projection. However, there are projection systems whose geometry cannot be described using the conventional pinhole model because of the very high distortion introduced by the imaging device. Some of these systems are omnidirectional cameras.

An omnidirectional camera is a camera that provides wide field of view, at least more than 180 degrees. There are several ways to build an omnidirectional camera. Dioptric cameras use a combination of shaped lenses (e.g., fisheye lenses; see figure 4.35a) and typically can reach a field of view slightly larger than 180 degrees. Catadioptric cameras combine a standard camera with a shaped mirror, like a parabolic, hyperbolic, or elliptical mirror and are able to provide much more than 180 degrees field-of-view in elevation and 360 in the azimuthal direction. In figure 4.35b you can see an example catadioptric camera using a hyperbolic mirror. Finally, polydioptric cameras use multiple cameras with overlapping field of view (figure 4.35c) and so far are the only cameras able to provide a real omnidirectional (spherical) view (i.e., 4π radians).

Catadioptric cameras were first introduced in robotics in 1990 by Yagi and Kawato [333], who utilized them for localizing robots. Fisheye cameras started to spread over only in 2000 thanks to new manufacturing techniques and precision tools that led to an increase of their field of view up to 180 degrees. However, it is only since 2005 that these cameras have been miniaturized to the size of 1–2 centimeters, and their field of view has been increased up to 190 degrees (see, for instance, figure 4.36a).

Thanks to the camera miniaturization, to the recent developments in optics manufacturing, and to the decreasing prices in the cameras' market, catadioptric and dioptric omni-

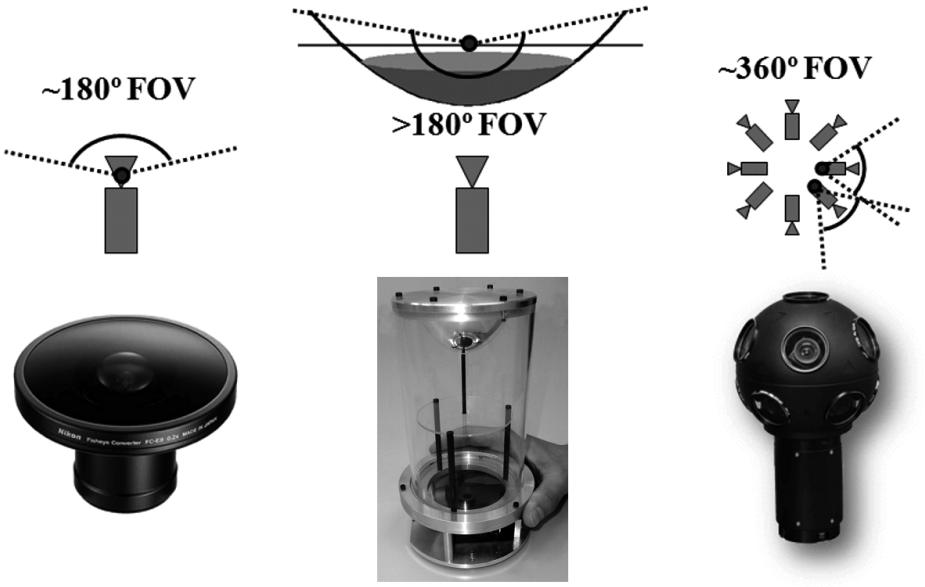


Figure 4.35 (a) Dioptric camera (e.g. fisheye); (b) catadioptric camera; (c) an example polydioptric camera produced by Immersive Media.

directional cameras are being more and more used in different research fields. Miniature dioptric and catadioptric cameras are now used by the automobile industry in addition to sonars for improving safety, by providing to the driver an omnidirectional view of the surrounding environment. Miniature fisheye cameras are used in endoscopes for surgical operations or on board microaerial vehicles for pipeline inspection as well as rescue operations. Other examples involve meteorology for sky observation.

Roboticists have also been using omnidirectional vision with very successful results on robot localization, mapping, and aerial and ground robot navigation [76, 80, 107, 278, 279, 307]. Omnidirectional vision allows the robot to recognize places more easily than with standard perspective cameras [276]. Furthermore, landmarks can be tracked in all directions and over longer periods of time, making it possible to estimate motion and build maps of the environment with better accuracy than with standard cameras, see figure 4.36 for some of examples of miniature omnidirectional cameras used on state-of-the-art micro aerial vehicles. Several companies, like Google, are using omnidirectional cameras to build photorealistic street views and three-dimensional reconstructions of cities along with texture. Two example omnidirectional images are shown in figure 4.37.

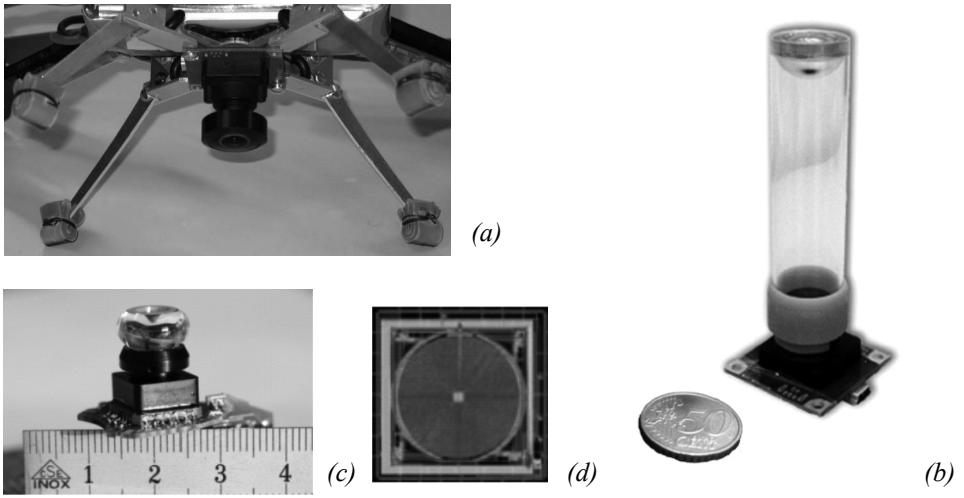


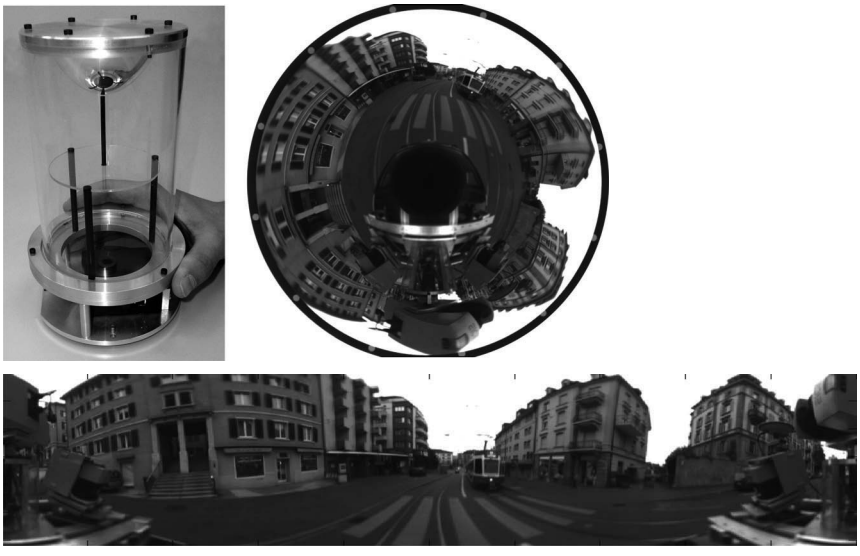
Figure 4.36 (a) The fisheye lens from Omnitech Robotics (www.omnitech.com) provides a field of view of 190 deg. This lens has a diameter of 1.7 cm. This camera has been used on the sFly autonomous helicopter at the ETH Zurich, (section 2.4.3) [76]. (b) A miniature catadioptric camera built at the ETH Zurich, which is also used for autonomous flight. It uses a spherical mirror and a transparent plastic support. The camera measures 2 cm in diameter and 8 cm in height. (c) The muFly camera built by CSEM, which is used on the muFly helicopter at the ETH Zurich (section 2.4.3). This is one of the smallest catadioptric cameras ever built. Additionally, it uses a polar CCD (d) where pixels are arranged radially.

In the next sections we will give an overview of omnidirectional camera models and calibration. For an in-depth study on omnidirectional vision, we refer the reader to [4, 15, 273].

4.2.4.2 Central omnidirectional cameras

A vision system is said to be central when the optical rays to the viewed objects intersect in a single point in 3D called projection center or *single effective viewpoint* (figure 4.38). This property is called *single effective viewpoint property*. The perspective camera is an example of a central projection system because all optical rays intersect in one point, that is, the camera optical center.

All modern fisheye cameras are central, and hence, they satisfy the single effective viewpoint property. Central catadioptric cameras conversely can be built only by opportunely choosing the mirror shape and the distance between the camera and the mirror. As proven by Baker and Nayar [64], the family of mirrors that satisfy the single viewpoint property is the class of rotated (swept) conic sections, that is, hyperbolic, parabolic, and elliptical mirrors. In the case of hyperbolic and elliptical mirrors, the single view point



(a)

(b)

Figure 4.37 (a) A catadioptric omnidirectional camera using a hyperbolic mirror. The image is typically unwrapped into a cylindrical panorama. The field of view is typically 100 degrees in elevation and 360 degrees in azimuth. (b) Nikon fisheye lens FC-E8. This lens provides a hemispherical (180 deg) field of view.

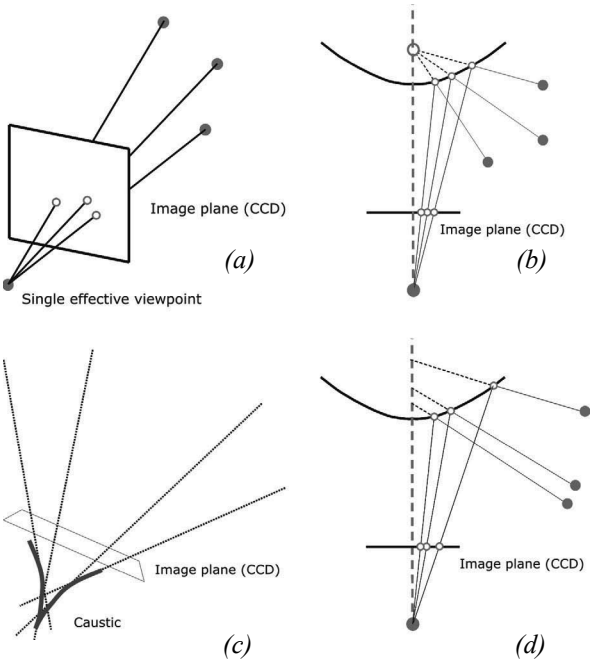


Figure 4.38 (a–b) Example of central cameras: perspective projection and catadioptric projection through a hyperbolic mirror. (c–d) Example of noncentral cameras: the envelope of the optical rays forms a *caustic*.

property is achieved by ensuring that the camera center (i.e., the pinhole or the center of the lens) coincides with one of the foci of the hyperbola (ellipse) (figure 4.39). In the case of parabolic mirrors, an orthographic lens must be interposed between the camera and the mirror, this makes it possible that parallel rays reflected by the parabolic mirror converge to the camera center (figure 4.39).

The reason a single effective viewpoint is so desirable is that it allows us to generate geometrically correct perspective images from the pictures captured by the omnidirectional camera (figure 4.40). This is possible because, under the single view point constraint, every pixel in the sensed image measures the irradiance of the light passing through the viewpoint in one particular direction. When the geometry of the omnidirectional camera is known, that is, when the camera is calibrated, one can precompute this direction for each pixel. Therefore, the irradiance value measured by each pixel can be mapped onto a plane at any distance from the viewpoint to form a planar perspective image. Additionally, the image can be mapped on to a sphere centered on the single viewpoint, that is, spherical projection (figure 4.40, bottom).

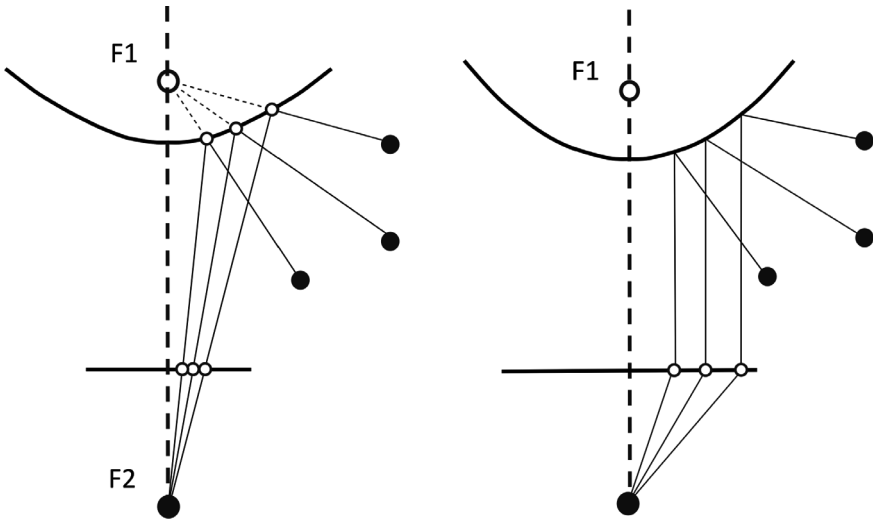


Figure 4.39 Central catadioptric cameras can be built by using hyperbolic and parabolic mirrors. The parabolic mirror requires the use of an orthographic lens.

Another reason why the single view point property is so important is that it allows us to apply the well known theory of *epipolar geometry* (see section 4.2.6.1), which easily allows us to perform *structure from stereo* (section 4.2.5) and *structure from motion* (section 4.2.6). As we will see, epipolar geometry holds for any central camera, both perspective and omnidirectional. Therefore, in those sections we will not make any distinction about the camera.

4.2.4.3 Omnidirectional camera model and calibration

Intuitively, the model of an omnidirectional camera is a little more complicated than a standard perspective camera. The model should indeed take into account the reflection operated by the mirror in the case of a catadioptric camera or the refraction caused by the lens in the case of a fisheye camera. Because the literature in this field is quite large, here we review two different projection models that have become standards in omnidirectional vision and robotics. Additionally, Matlab toolboxes have been developed for these two models, which are used worldwide by both specialists and non-experts.

The first model is known as the *unified projection model for central catadioptric cameras*. It was developed in 2000 by Geyer and Daniilidis [137] (later refined by Barreto and Araujo [66]), who have the merit of having proposed a model that encompasses all three types of central catadioptric cameras, that is, cameras using hyperbolic, parabolic, or elliptical mirror. This model was developed specifically for central catadioptric cameras and is

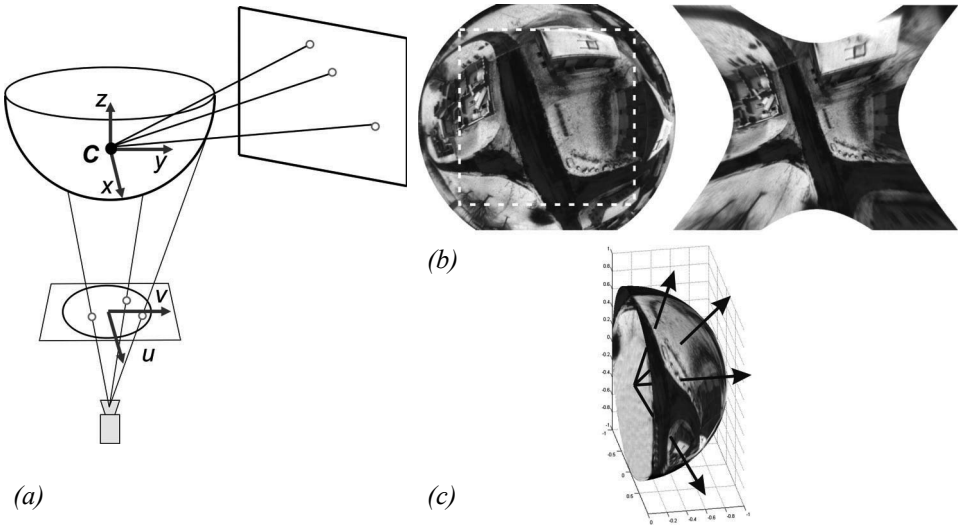


Figure 4.40 Central cameras allows us to remap regions of the omnidirectional image into a perspective image. This can be done straightforwardly by intersecting the optical rays with a plane specified arbitrarily by the user (a). Of course we cannot project the whole image onto a plane but only subregions of it (b). Another possible projection is that onto a sphere (c).

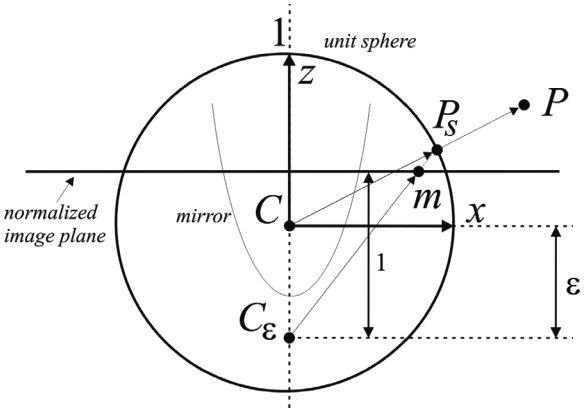
not valid for fisheye cameras. The approximation of a fisheye lens model by a catadioptric one is usually possible, however, with limited accuracy only [335].

Conversely, the second model unifies both central catadioptric cameras and fisheye cameras under a general model also known as Taylor model. It was developed in 2006 by Scaramuzza et al. [274, 275] and has the advantage that both catadioptric and dioptric cameras can be described through the same model, namely a Taylor polynomial.

Unified model for central catadioptric cameras. With their landmark paper from 2000, Geyer and Daniilidis showed that every catadioptric (parabolic, hyperbolic, elliptical) and standard perspective projection is equivalent to a projective mapping from a sphere, centered in the single viewpoint, to a plane with the projection center on the perpendicular to the plane and distant ε from the center of the sphere. This is summarized in figure 4.41.

As we did for the perspective camera, the goal is again to find the relation between the viewing direction to the scene point and the pixel coordinates of the corresponding image point. The projection model of Geyer and Daniilidis follows a four-step process. Let again $P = (x, y, z)$ be a scene point in the mirror reference frame[5] centered in C (figure 4.41).

1. The first step consists in projecting the scene point onto the unit sphere; therefore:



Mirror type	ε
Parabola	1
Hyperbola	$\frac{d}{\sqrt{d^2 + 4l^2}}$
Ellipse	$\frac{d}{\sqrt{d^2 + 4l^2}}$
Perspective	0

Figure 4.41 Unified projection model for central catadioptric cameras of Geyer and Daniilidis.

$$P_s = \frac{P}{\|P\|} = (x_s, y_s, z_s)\,. \tag{4.50}$$

2. The point coordinates are then changed to a new reference frame centered in $C_\varepsilon = (0, 0, -\varepsilon)$; therefore:

$$P_\varepsilon = (x_s, y_s, z_s + \varepsilon)\,. \tag{4.51}$$

Observe that ε ranges between 0 (planar mirror) and 1 (parabolic mirror). The correct value of ε can be obtained knowing the distance d between the foci of the conic and the latus rectum[6] l as summarized in the table of figure 4.41.

3. P_ε is then projected onto the normalized image plane distant 1 from C_ε; therefore,

$$\tilde{m} = (x_m, y_m, 1) = \left(\frac{x_s}{z_s + \varepsilon}, \frac{y_s}{z_s + \varepsilon}, 1\right) = g^{-1}(P_s)\,. \tag{4.52}$$

4. Finally, the point $\tilde{m}$ is mapped to the camera image point $\tilde{p} = (u, v, 1)$ through the intrinsic parameter matrix A; therefore,

5. For convenience we assume that the mirror axis of symmetry is perfectly aligned with the camera optical axis. We also assume that the *x*-*y* axes of the camera and mirror are aligned. Therefore, the camera and mirror reference frames differ only by a translation along *z*.

6. The latus rectum of a conic section is the chord through a focus parallel to the conic section directrix.

$$\tilde{p} = A \cdot \tilde{m}, \tag{4.53}$$

where A is given by (4.47), that is,

$$A = \begin{bmatrix} \alpha_u & \alpha_u \cot\theta & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{4.54}$$

It is easy to show that function g^{-1} is bijective and that its inverse g is given by[7]:

$$P_s = g(m) \sim \begin{bmatrix} x_m \\ y_m \\ 1 - \varepsilon \dfrac{x_m^2 + y_m^2 + 1}{\varepsilon + \sqrt{1 + (1 - \varepsilon^2)(x_m^2 + y_m^2)}} \end{bmatrix}, \tag{4.55}$$

where ~ indicates that g is proportional to the quantity on the right-hand side. To obtain the scale factor, it is sufficient to normalize $g(m)$ onto the unit sphere.

Observe that equation (4.55) is the core of the projection model of central catadioptric cameras. It expresses the relation between the point m on the normalized image plane and the unit vector P_s in the mirror reference frame. Note that in the case of planar mirror, we have $\varepsilon = 0$ and (4.55) becomes the projection equation of perspective cameras $P_s \sim (x_m, y_m, 1)$.

This model has proved to be able to describe accurately all central catadioptric cameras (parabolic, hyperbolic, and elliptical mirror) and standard perspective cameras. An extension of this model for fisheye lenses was proposed in 2004 by Ying and Hu [335]. However, the approximation of a fisheye camera through a catadioptric one works only with limited accuracy. This is mainly due because, while the three types of central catadioptric cameras can be represented through an exact parametric function (parabola, hyperbola, ellipse), the projective models of fisheye vary from camera to camera and depend on the lens field-of-view. To overcome this problem, a new unified model has been proposed, which will be described in the next section.

7. Equation (4.55) can be obtained by inverting (4.52) and imposing the constraint that P_s must lie on the unit sphere and, thus, $x_s^2 + y_s^2 + z_s^2 = 1$. From this constraint you will then get an expression of z_s as a function of ε, x_m, and y_m. More details can be found in [66].

Unified model for catadioptric and fisheye cameras. This unified model was proposed by Scaramuzza et al. in 2006 [274, 275]. The main difference with the previous model lies in the choice of the function g. To overcome the lack of knowledge of a parametric model for fisheye cameras, the authors proposed the use of a Taylor polynomial, whose coefficients and degree are found through the calibration process. Accordingly, the relation between the normalized image point $\tilde{m} = (x_m, y_m, 1)$ and the unit vector P_s in the fisheye (mirror) reference frame can be written as:

$$P_s = g(m) \sim \begin{bmatrix} x_m \\ y_m \\ a_0 + a_2\rho^2 + \ldots + a_N\rho^N \end{bmatrix}, \tag{4.56}$$

where $\rho = \sqrt{x_m^2 + y_m^2}$. As you have probably noticed, the first-order term (i.e., $a_1\rho$) of the polynomial is missing. This follows from the observation that the first derivative of the polynomial calculated at $\rho = 0$ must be null for both catadioptric and fisheye cameras (this is straightforward to verify for catadioptric cameras by differentiating [4.55]). Also observe that because of its polynomial nature, this expression can encompass catadioptric, fisheye, and perspective cameras. This can be done by opportunely choosing the degree of the polynomial. As highlighted by the authors, polynomials of order three or four are able to model very accurately all catadioptric cameras and many types of fisheye cameras available on the market. The applicability of this model to a wide range of commercial cameras is at the origin of its success.

Omnidirectional camera calibration. The calibration of omnidirectional cameras is similar to that for calibrating standard perspective cameras, which we have seen in section 4.2.3.4. Again, the most popular methods take advantage of planar grids that are shown by the user at different positions and orientations. For omnidirectional cameras, it is very important that the calibration images are taken all around the camera and not on a single side only. This in order to compensate for possible misalignments between the camera and mirror.

It is worth to mention three open-source calibration toolboxes currently available for Matlab, which differ mainly for the projection model adopted and the type of calibration pattern.

- The toolbox of Mei uses checkerboard-like images and takes advantage of the projection model of Geyer and Daniilidis discussed earlier. It is particularly suitable for catadioptric cameras using hyperbolic, parabolic, folded mirrors, and spherical mirrors. Mei's toolbox can be downloaded from [349], while the theoretical details can be found in [212].

- The toolbox of Barreto uses line images instead of checkerboards. Like the previous toolbox, it also uses the projection model of Geyer and Daniilidis. It is particularly suitable for parabolic mirrors. The toolbox can be downloaded from [350], while the theoretical details can be found in [67] and [68].
- Finally, the toolbox of Scaramuzza uses checkerboard-like images. Contrary to the previous two, it takes advantage of the unified Taylor model for catadioptric and fisheye cameras developed by the same author. It works with catadioptric cameras using hyperbolic, parabolic, folded mirrors, spherical, and elliptical mirrors. Additionally, it works with a wide range of fisheye lenses available on the market—like Nikon, Sigma, Omnitech-Robotics, and many others—with field of view up to 195 degrees. The toolbox can be downloaded from [351], while the theoretical details can be found in [274] and [275]. Contrary to the other two, this toolbox features an automatic calibration process. In fact, both the center of distortion and the calibration points are detected automatically without any user intervention.

4.2.5 Structure from stereo

4.2.5.1 Introduction

Range sensing is extremely important in mobile robotics, since it is a basic input for successful obstacle avoidance. As we have seen earlier in this chapter, a number of sensors are popular in robotics explicitly for their ability to recover depth estimates: ultrasonic, laser rangefinder, time-of-flight cameras. It is natural to attempt to implement ranging functionality using vision chips as well.

However, a fundamental problem with visual images makes rangefinding relatively difficult. Any vision chip collapses the 3D world into a 2D image plane, thereby losing depth information. If one can make strong assumptions regarding the size of objects in the world, or their particular color and reflectance, then one can directly interpret the appearance of the 2D image to recover depth. But such assumptions are rarely possible in real-world mobile robot applications. Without such assumptions, a single picture does not provide enough information to recover spatial information.

The general solution is to recover depth by looking at *several* images of the scene to gain more information, hopefully enough to at least partially recover depth. The images used must be different, so that taken together they provide additional information. They could differ in camera geometry—such as the focus position or lens iris—yielding depth from focus (or defocus) techniques that we have described in section 4.2.3.1. An alternative is to create different images, not by changing the camera geometry, but by changing the camera viewpoint to a different camera position. This is the fundamental idea behind *structure from stereo* (i.e., *stereo vision*) and *structure from motion* that we will present in the next sections. As we will see, stereo vision processes two distinct images taken at the same time

and assumes that the relative pose between the two cameras is known. Structure-from-motion conversely processes two images taken with the same or a different camera at different times and from different unknown positions; the problem consists in recovering both the relative motion between the views and the depth. The 3D scene that we want to reconstruct is usually called *structure*.

4.2.5.2 Stereo vision

Stereopsis (from *stereo* meaning solidity, and *opsis* meaning vision or sight) is the process in visual perception leading to the sensation of depth from the two slightly different projections of the world onto the retinas of the two eyes. The difference in the two retinal images is called horizontal *disparity*, retinal disparity, or binocular disparity. The differences arise from the eyes' different positions in the head. It is the disparity that makes our brain fuse (perceive as a single image) the two retinal images making us perceive the object as a one and solid. To have a clearer understanding of what disparity is, as a simple test, hold your finger vertically in front of you and close each eye alternately. You will see that the finger jumps from left to right. The distance between the left and right appearance of the finger is the disparity. The same phenomenon is visible in the image pair shown in figure 4.48, in which the foreground objects shift left and right relative to the background.

Computational stereopsis, or stereo vision, is the process of obtaining depth information from a pair of images coming from two cameras which look at the same scene from different positions. In stereo vision we can identify two major problems:

1. the correspondence problem
2. 3D reconstruction

The first consists in matching (pairing) points of the two images which are the projection of the same point in the scene. These matching points are called *corresponding points* or *correspondences* (figure 4.45a). This will be clarified later on. Determining the corresponding points is made possible based on the assumption that the two images differ only slightly and therefore a feature in the scene appears similar in both images. Based only of this assumption, however, there might be many possible false matches. As we will see, this problem can be overcome by introducing an additional constraint which makes the correspondence matching feasible. This constraint is called *epipolar constraint* (section 4.2.6.1) and states that the correspondent of a point in an image lies on a line (called *epipolar line*) in the other image (figure 4.45b). Because of this constraint, we will see that the correspondence search becomes one-dimensional instead of two-dimensional.

Knowing the correspondences between the two images, knowing the relative orientation and position of the two cameras, and knowing the intrinsic parameters of the two cameras, it is possible to reconstruct the scene points (i.e., the structure). This process of reconstruction requires the prior calibration of the stereo camera; that is, we need to calibrate the two cameras separately for estimating their extrinsic parameters, but we also need to determine their extrinsic parameters, i.e. the camera relative position.

Figure 4.42 (Left) The STH-MDCS3 form Videre Design uses CMOS sensors, a baseline of 9 cm, an image resolution of 1280×960 at 7.5 frames per second (fps), or 640×480 at 30 fps. (Right) The Bumblebee2 from Point Grey uses CCD sensors, a baseline of 12 cm, an image resolution of 1024×768 at 20 frames per second (fps), or 640×480 at 48 fps.

The theory of stereo vision has been well understood for years, while the engineering challenge of creating a practical stereo-vision sensor has been formidable [21, 43, 44]. Example of commercially available stereo cameras are shown in figure 4.42.

Basic case. First, we consider a simplified case in which two cameras have the same orientation and are placed with their optical axes parallel, at a separation of b (called *baseline*), shown in figure 4.43.

In this figure, a point on the object is described as being at coordinate (x, y, z) with respect to the origin located in the left camera lens. The image coordinate in the left and right image are (u_l, v_l) and (u_r, v_r) respectively. From figure 4.43a and using equations (4.36) and (4.37), we can write

$$\frac{f}{z} = \frac{u_l}{x}, \tag{4.57}$$

$$\frac{f}{z} = \frac{-u_r}{b-x}, \tag{4.58}$$

from which we obtain

$$z = b\frac{f}{u_l - u_r}, \tag{4.59}$$

where the *difference* in the image coordinates, $u_l - u_r$ is called *disparity*. This is an important term in stereo vision, because it is only by measuring disparity that we can recover depth information. Observations from this equation are as follows:

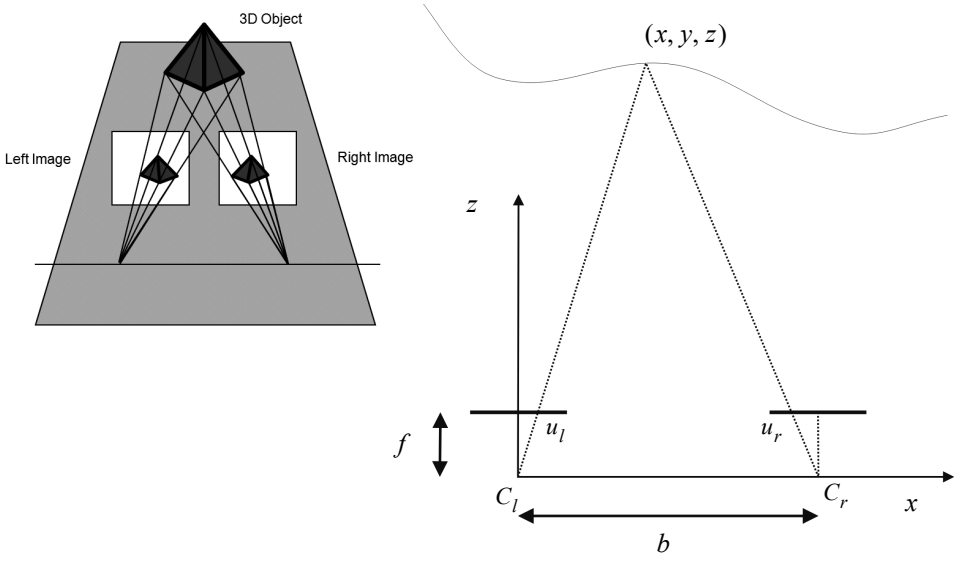


Figure 4.43
Idealized camera geometry for stereo vision. The cameras are assumed be identical (i.e., identical focal lengths and image resolution); furthermore, they are assumed to be perfectly aligned on the horizontal axis.

- Distance is inversely proportional to disparity. The distance to near objects can therefore be measured more accurately than that to distant objects, just as with depth from focus techniques. In general, this is acceptable for mobile robotics, because for navigation and obstacle avoidance closer objects are of greater importance.
- Disparity is proportional to b. For a given disparity error, the accuracy of the depth estimate increases with increasing baseline b.
- As b is increased, because the physical separation between the cameras is increased, some objects may appear in one camera but not in the other. This is due to the field of view of the cameras. Such objects by definition will not have a disparity and therefore will not be ranged.
- If the baseline b is unknown, it is possible to reconstruct the scene point only *up to a scale*. This is the case in structure-from-motion (section 4.2.6).
- A point in the scene visible to both cameras produces a pair of image points known as a *conjugate pair*, or *correspondence pair* (figure 4.44a). Given one member of the conjugate pair, we know that the other member of the pair lies somewhere along a line known

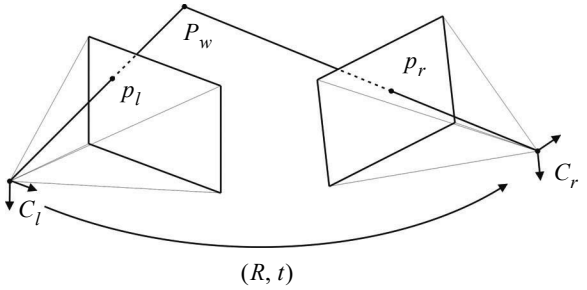


Figure 4.44 Stereo vision: general case.

as *epipolar line*. In the case depicted in figure 4.43a, because the cameras are perfectly aligned with one another, the epipolar lines are horizontal lines (i.e., along the x direction). The concept of epipolar line will be explained later on in this section.

General case. The assumption of perfectly aligned cameras is normally violated in practice. In fact, even the most expensive stereo cameras available in the market do not assume this model. Indeed, two exactly identical cameras do not exist. There will be always differences in the focal length due to manufacturing but, especially, even if such identical cameras could exist, we would never be sure that they are perfectly aligned. The situation is even more complicated by the fact that the internal orientation of the CCD in the camera package is unknown. Ideally it is aligned, but in practice the CCD cannot be considered perfectly aligned. Therefore, the general stereo vision model assumes that the two cameras are different and not aligned (figure 4.44) but requires that the relative position and orientation of the two cameras is known. If the relative position is not known, the stereo camera must be calibrated using the checkerboard-based calibration treated in section 4.2.3.4. Fortunately, the previously mentioned toolbox for calibrating the camera intrinsic parameters [347] allows the user to calibrate stereo cameras as well.

So, let us assume that the two cameras have been previously calibrated. Therefore, the intrinsic parameter matrices A_l and A_r (see equation 4.47) for the left and right camera are known, and the camera extrinsic parameters—i.e. the rotations R_l, R_r and translations t_l, t_r of the two cameras with the respect to the world coordinate system—are also known. In stereo vision, it is a common practice to assume the origin of the world coordinate system in the left camera. Thus, we can write $R_l = I$ and $R_r = R$. This allows us to write the equations of perspective projection for the two cameras as:

$$\lambda_l \tilde{p}_l = A_l[I|0]\tilde{P}_w \text{ (for the left camera)}, \tag{4.60}$$

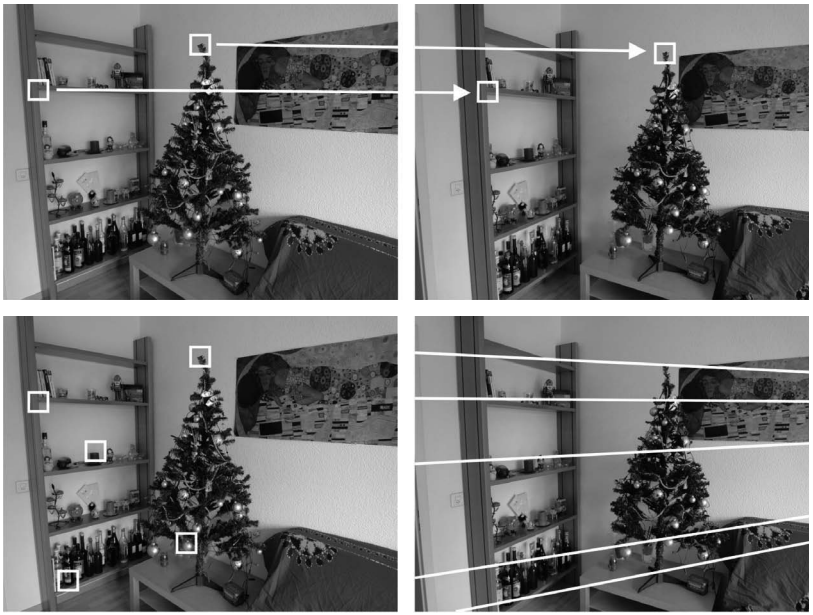


Figure 4.45 A stereo pair. Corresponding points are projections of the same scene point. Because of the epipolar constraint, conjugate points can be searched along the epipolar lines. This heavily reduces the computational cost of the correspondence search: from a two-dimensional search it becomes a one-dimensional search problem.

$$\lambda_r \tilde{p}_r = A_r[R|t]\tilde{P}_w \text{ (for the right camera)}, \tag{4.61}$$

where $\tilde{p}_l = [u_l, v_l, 1]^T$ and $\tilde{p}_r = [u_r, v_r, 1]^T$ are the image points (in homogeneous coordinates) corresponding to the world point $\tilde{P}_w = [x, y, z, 1]^T$ (in homogeneous coordinates) in the left and right camera respectively. λ_l and λ_r are the depth factors. Observe that (4.60) and (4.61) actually contribute three equations each. Therefore, we have a system of six equations in five unknowns, three for the world point $P_w = (x, y, z)$ and two for depth factors, i.e. λ_l and λ_r. The system is overdetermined and can be solved either linearly, using least-squares, or nonlinearly by computing the 3D point that minimizes distances between the two light rays passing through $\tilde{p}_l$ and $\tilde{p}_r$. The solution of these two equations is left as an exercise to the reader in section 4.8.

Correspondence problem. Using the preceding equations requires us to have identified the conjugate pair p_l and p_r in the left and right camera images, which originates from the same scene point $\tilde{P}_w$ (figure 4.45a). This fundamental challenge is called the *correspon-*

dence problem. Intuitively, the problem is: given two images of the same scene from different perspectives, how can we identify the same object points in both images? For every such identified object point, we will then be able to recover its 3D position in the scene.

The correspondence search is based on the assumption that the two images of the same scene do not differ too much, that is, a feature in the scene is supposed to appear very similar in both images. Using an opportune image similarity metric (see section 4.3.3), a given point in the first image can be paired with one point in the second image. The problem of *false correspondences* makes the correspondence search challenging. False correspondences occur when a point is paired to another that is not its real conjugate. This is because the assumption of image similarity does not hold very well, for instance if the part of the scene to be paired appears under different illumination or geometric conditions. Other problems that make the correspondence search difficult are:

- *Occlusions*: the scene is seen by two cameras at different viewpoints and therefore there are parts of the scene that appear only in one of the images. This means, there exist points in one image which do not have a correspondent in the other image.
- *Photometric distortion*: there are surfaces in the scene which are nonperfectly *lambertian*, that is, surfaces whose behavior is partly specular. Therefore, the intensity observed by the two cameras is different for the same point in the scene as more as the cameras are farther apart.
- *Projective distortion*: because of the perspective distortion, an object in the scene is projected differently on the two images, as more as the cameras are farther apart.

Some constraints can, however, be exploited for improving the correspondence search, which are:

- *Similarity constraint*: a feature in the image appears similar in the other image.
- *Continuity constraint*: far from the image borders, the depth of the scene points along a continuous surface varies continuously. This constraint clearly limits the *gradient of disparity*.
- *Unicity*: a point of the first image can be paired only with a single point in the other image, and vice versa (it fails in presence of occlusions, specularities, and transparency).
- *Monotonic order*: if point p_l in the left image is the correspondent of p_r in the right image, then the correspondent of a point on the right (left) of p_l can only be found on the right (left) of p_r. This is valid only for points that lie on an opaque object.
- *Epipolar constraint*: the correspondent of a point in the left image can only be found along a line in the right image, which is called *epipolar line* (figure 4.45b). As a matter of fact, this is the most important constraint and will be explained later on.

The methods for searching correspondences can be distinguished into two categories:

- *Area-based*: these algorithms consider a small patch (window) in one image and look for the most similar patch in the second image by means of an appropriate *correlation* measure. This search is done for every pixel and allows us to obtain a *dense* reconstruction. However, in uniform regions—that is, poor texture—these methods fail. There exist different techniques to measure the similarity between image patches for stereo matching. The most used are the *Sum of Absolute Differences* (SAD), the *Sum of Squared Differences* (SSD), the *Normalized Cross Correlation* (NCC), and the *Census Transform*. An overview of some of these algorithms is given in section 4.3.3. Finally, observe that the search for correspondences is a two-dimensional search: the most similar of a patch in the left image must be searched across all rows and columns of the right image. As we will see in the next section, the search can be reduced to only one line, the epipolar line, thus reducing the dimensionality of the search from two to one (figure 4.45b).
- *Feature-based*: these algorithms extract salient features from the images, which are possibly stable with respect to change of view point. The matching process is applied to the attributes associated to the features. Edges, corners, line segments, and blobs are some of the features that can be used. They do not have to correspond necessarily to a well defined geometric entity. An exhaustive overview on feature extraction is given is section 4.5. Feature-based stereo matching algorithms are faster and more robust than area-based methods but provide only sparse depth maps, which then need to be interpolated.

Epipolar geometry. Given a pixel in one image (say the left image), how can we compute its correspondence with the correct pixel in the other image? As we anticipated in the previous section, one way would be to search the correspondences across all pixels of the second image. In the case of stereo matching, however, we have some information available, namely the relative position and the calibration parameters of the two cameras. This information allows us to reduce the search from two dimensions to an only one dimension. Figure 4.46a shows how a pixel point p_l in one image projects to an epipolar line segment in the other image. The segment is bounded at one end by the projection of P_∞ (the original viewing ray at infinity) and at the other end by the projection of C_l into the second camera, which is known as the epipole e_r. By projecting the epipolar line in the second image back into the first image, we get another line which is bounded by the other corresponding epipole e_l. Notice that two corresponding epipolar lines (figure 4.46b) originate from the intersection of the two image planes with the epipolar plane that passes through the camera centers C_l and C_r and the scene point P_w.

To compute the equation of the epipolar line, we must project the optical ray passing through p_l and C_l to the second image. This is straightforward. The equation of the optical

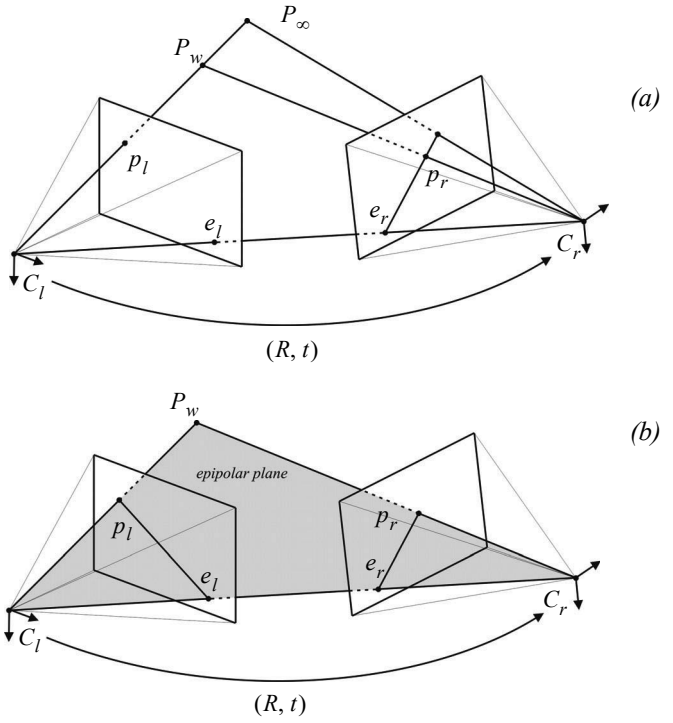


Figure 4.46 Epipolar geometry: (a) epipolar line segment corresponding to one ray; (b) corresponding set of epipolar lines and their epipolar plane.

ray passing through p_l and C_l can be obtained from the perspective projection equation (4.60), which we rewrite here as:

$$\lambda_l \tilde{p}_l = \lambda_l \begin{bmatrix} u_l \\ v_l \\ 1 \end{bmatrix} = A_l [I|0] \tilde{P}_w = A_l P_w = A_l \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{4.62}$$

and therefore the line passing through p_l and C_l has equation:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \lambda_l A_l^{-1} \begin{bmatrix} u_l \\ v_l \\ 1 \end{bmatrix}, \tag{4.63}$$

which we can rewrite in a more compact form as:

$$P_w = \lambda_l A_l^{-1} \tilde{p}_l \,. \tag{4.64}$$

Finally, to find the equation of the epipolar line, we just project this line onto the second image using the perspective projection equation (4.61):

$$\lambda_r \tilde{p}_r = A_r [R|t] \tilde{P}_w = A_r R P_w + A_r t \,, \tag{4.65}$$

and therefore, using (4.64), we obtain the epipolar line

$$\lambda_r \tilde{p}_r = \lambda_l A_r R A_l^{-1} \tilde{p}_l + A_r t \,, \tag{4.66}$$

where $A_r t$ is actually the epipole e_r in the second image, that is, the projection of the optical center C_l of the left camera into the right image.

By applying equation (4.66) to every image point in the left image, we can compute all the epipolar lines in the right image. The correspondence of one point in the left image will then need to be searched only along its corresponding epipolar line. Note that the epipolar lines pass all through the same epipole. However, observe that in computing equation (4.66) we did not take into account the radial distortion introduced by the lens. Although for some narrow-field-of-view cameras the radial distortion is rather small, it is always opportune to take the radial distortion into account when computing the equation of the epipolar line. The reason is that if the epipolar line is not determined precisely, the correspondence search along a non accurate epipolar line can lead to a larger uncertainty in the computation of the disparity as well as in the reconstruction of the scene point P_w.

Instead of taking into account the radial distortion, a common consolidated procedure in stereo vision is that of undistorting first the two images, that is, remapping the left and right image into new images without distortion. Furthermore, the two images can be remapped in such a way that all epipolar lines in the left and right image are collinear and horizontal (figure 4.47d). The process of transforming a pair of stereo images into a new pair without radial distortion and with horizontal epipolar lines is called *stereo rectification* or *epipolar rectification*. We will briefly explain it in the next section.

Epipolar rectification. Given a pair of stereo images, epipolar rectification is a transformation of each image plane such that all corresponding epipolar lines become collinear and parallel to one of the image axes, for convenience usually the horizontal axis. The resulting rectified images can be thought of as acquired by a new stereo camera obtained by rotating the original cameras about their optical centers. The great advantage of the epipolar recti-

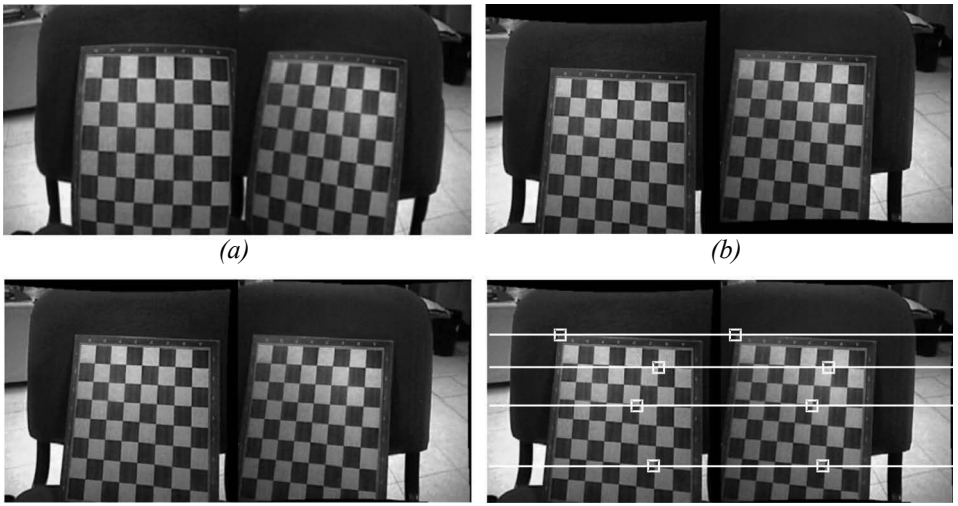


Figure 4.47 Rectification of a stereo pair: (a) original images, (b) compensation of the lens distortion, (c) compensation of rotation and translation, (d) After the epipolar rectification, the epipolar lines appear collinear and horizontal.

fication is the correspondence search becomes simpler and computationally less expensive because the search is done along the horizontal lines of the rectified images. The steps of the epipolar rectification algorithm are illustrated in figure 4.47. Observe that after the rectification, all the epipolar lines in the left and right image are collinear and horizontal (figure 4.47d). The equations for the epipolar rectification algorithm go beyond the scope of this book, but the interested reader can find an easy-to-implement algorithm in [133].

Disparity map. After the calibration of the stereo-rig, the epipolar rectification, and the correspondence search, we can finally reconstruct the scene points in 3D by solving the system of equations (4.60)–(4.61) (see also the problem in section 4.8). Another popular output of stereo vision is the *disparity map*. A disparity map appear as a grayscale image where the intensity of every pixel point is proportional to the disparity of that pixel in the left and right image: objects that are closer to the camera appear lighter, while farther objects appear darker. An example disparity map is shown in figure 4.48. Disparity maps are very useful for obstacle avoidance (figure 4.49). Modern stereo cameras—like those from Videre Design and Point-Grey (figure 4.42)—are able to compute disparity maps directly in hardware.

Left image

Right image

Disparity map

Figure 4.48 An example disparity map computed from the two top images. Every pixel point is proportional to the disparity of that pixel in the left and right image. Objects that are closer to the camera appear lighter, while farther objects appear darker. Image courtesy of Martin Humenberger, AIT Austrian Institute of Technology — http://www.ait.ac.at.

4.2.6 Structure from motion

In the previous section, we described how to recover the structure of the environment from two images of the scene taken from two distinct cameras whose relative position and orientation is known. In this section, we discuss the problem of recovering the structure when the camera relative pose is unknown. This is the case, for instance, when the two images are taken from the same camera but at different positions and at different times,[8] or, alternatively, from different cameras. This implies that both structure and motion must be estimated simultaneously. This problem is known as *Structure from Motion* (SfM). This problem has been studied for long time in the computer vision community, and in this sec-

8. For the sake of simplicity, here we assume that the scene is time invariant (i.e., static). One way to deal with dynamic scenes consists in treating moving objects as outliers.

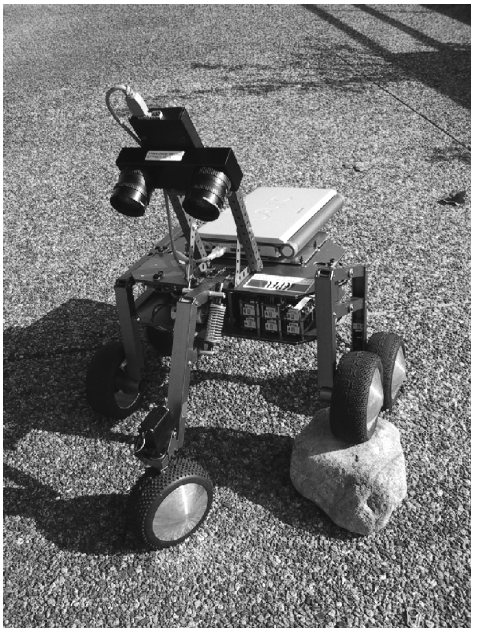


Figure 4.49
A stereo camera from Videre Design on the Shrimp robot developed at the ASL.

tion we provide only the solution to the two-frame structure from motion problem. For an in-depth study of structure from motion, we refer the reader to [21, 22, 29, 36, 53].

Observe that in structure-from-motion, the images do not need to be precalibrated. This allows SfM to work in challenging situations, where, for instance, images are taken by different users using different cameras (for example, images from the Web). The intrinsic parameters can in fact be estimated automatically from SfM itself. A suggestive result of SfM is illustrated in figure 4.50. Here, the scene was reconstructed using dozens of images. Using thousands of images from different viewpoints, SfM can sometimes achieve 3D reconstruction results that are almost comparable in accuracy and density of points to 3D laser rangefinders (page 133). However, this precision is often at the expense of the computation power.

4.2.6.1 Two-view structure-from-motion

Let us start again from the two perspective projection equations (4.60) and (4.61) derived for the stereo vision case, but now remember that R and t denote the relative motion between the first and the second camera position; therefore, we can write: