One can measure thickness on many scales. The galaxy is a spiral disk about 100 Em (10²⁰ m) thick. The solar system is pancake-like, about 1 Tm (10¹² m) thick. The rings of Saturn are about 10 km thick. Closer to home, Earth’s atmosphere is a spherical shell about 40 km thick; the weather occurs in the troposphere, about 12 km thick. The outermost shell of the solid Earth is the crust, about 35 km thick. The ocean has a mean depth of 3.9 km. In the Antarctic, the recently discovered objects believed to be microfossils indicative of ancient Martian life are less than 100 nm thick. In terms of the man-made environment, industry must contend with thickness varying from meters, for construction projects, to millimeters on assembly lines, to micrometers and nanometers for the solid-state, optical, and coatings industries. Perhaps the most familiar way of measuring thickness is by mechanical means, such as by ruler or caliper. Other means are sometimes called for, either because both sides of an object are not accessible, the dimension is either too big or too small for calipers, the object is too fragile, too hot, or too cold for direct contact, or the object is in motion on an assembly line — it may not even be a solid. Thickness may also be a function of position, as either the object may have originally been made with nonuniform thickness, deliberately or not, or the thickness may have become nonuniform with time due either to corrosion, cracking, or some other deterioration. The thickness may also be changing with time due to deliberate growth or etching, as example for thin films. Thus it follows that, in more general terms, measuring thickness might require measuring the topography or height profile of two surfaces and taking the difference. Alternatively, the measurement technique may produce a reading directly related to the difference. Table 7.1 lists some of the many techniques suited to determining thickness, together with the range of usefulness and some comments on accuracy and/or precision.

7.1 Descriptions of the Relatively Mature Measuring Techniques
Mechanical • Electronic Gages • Pneumatic Gaging • Optical: Focusing, Shadowing, Comparing • Weighing • Capacitive Gaging • Inductive Gaging (Eddy Current Sensing) • Magnetic Induction • Hall Effect Gage • Far-Field/Time-of-Flight: Ultrasound, Radar, Lidar • Far-Field/Resonance: Ultrasound, Interferometry, Ellipsometry • Far-Field/Absorption, Scattering, Emission: Beta, Gamma, X-Ray, Infrared • Destructive Techniques

7.2 Future Directions in Thickness Measurement
Concerning Techniques Mentioned Above • THz Technology • Nanoscale-Scanning Probe Microscopy

John C. Brasunas
NASA/Goddard

G. Mark Cushman
NASA/Goddard

Brook Lakew
NASA/Goddard
TABLE 7.1 Thickness Measuring Techniques

<table>
<thead>
<tr>
<th>Technique</th>
<th>Range</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caliper gage, micrometer</td>
<td>1 μm–100 mm</td>
<td>±3 μm accuracy</td>
</tr>
<tr>
<td>Electronic gages: LVDT</td>
<td>0–1 m</td>
<td>Precision depends on noise level</td>
</tr>
<tr>
<td>Pneumatic gaging</td>
<td>50 nm minimum</td>
<td></td>
</tr>
<tr>
<td>Optical/focusing, shadowing, comparing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microscope</td>
<td>5 μm minimum</td>
<td>About 1% accuracy</td>
</tr>
<tr>
<td>Comparators/projectors</td>
<td>25–250 nm</td>
<td></td>
</tr>
<tr>
<td>Laser caliper</td>
<td>100 μm–100 mm</td>
<td>Precision of 6 μm or better</td>
</tr>
<tr>
<td>Weighing</td>
<td>Range depends on area</td>
<td></td>
</tr>
<tr>
<td>Capacitive gaging</td>
<td>From &lt; 1 mm to about 1 cm</td>
<td></td>
</tr>
<tr>
<td>Inductive gaging (eddy current sensing)</td>
<td>0–1.5 mm</td>
<td>Precision of 2.5 μm</td>
</tr>
<tr>
<td>Magnetic induction</td>
<td>0–4 mm</td>
<td>10% accuracy</td>
</tr>
<tr>
<td>Hall effect gage</td>
<td>0–10 mm</td>
<td>1–3% accuracy</td>
</tr>
<tr>
<td>Far-field/time-of-flight</td>
<td>0.5–250 mm</td>
<td>25 μm accuracy</td>
</tr>
<tr>
<td>Radar</td>
<td>0.1 to few hundred km</td>
<td></td>
</tr>
<tr>
<td>Lidar</td>
<td>10 m–5 km</td>
<td></td>
</tr>
<tr>
<td>Far-field/resonance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resonant ultrasound</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interferometry: spectral and spatial</td>
<td>1 nm–100 μm</td>
<td>Accuracy about λ/50</td>
</tr>
<tr>
<td>Ellipsometry</td>
<td>0.3 nm–10 μm</td>
<td>0.1 nm accuracy</td>
</tr>
<tr>
<td>Far-field/absorption, scattering, emission</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gamma-ray backscatter</td>
<td>Range to 25 mm</td>
<td>0.5% precision</td>
</tr>
<tr>
<td>Beta-transmission</td>
<td>2 μm–1 mm</td>
<td>0.2% precision</td>
</tr>
<tr>
<td>Beta-backscatter</td>
<td>100 nm–50 μm</td>
<td>3 to 20% precision</td>
</tr>
<tr>
<td>X-ray fluorescence</td>
<td>0–30 μm</td>
<td></td>
</tr>
<tr>
<td>Infrared absorption</td>
<td>Depends on material</td>
<td></td>
</tr>
<tr>
<td>Scanning techniques: scanning probe microscopy</td>
<td></td>
<td>Precision better than 0.1 nm</td>
</tr>
<tr>
<td>Destructive techniques: electrolytic</td>
<td>15 nm–50 μm</td>
<td></td>
</tr>
</tbody>
</table>

7.1 Descriptions of the Relatively Mature Measuring Techniques

The following descriptions will also refer to some of the relevant vendors, whose addresses are found in Table 7.2. Additional vendor information, with specific price or model number identification, is found in Table 7.3. The words “gage” and “gauge” are used interchangeably.

Mechanical

The fundamental tool for measuring thickness is the line-graduated instrument [1, 2]. It is the only mechanical means to make direct measurements. Graduated spacings that represent known distances are used as direct comparisons to the unknown distance. Instruments include bars, rules, and tapes generically called rulers; caliper gages, which employ a positive contact device for improved alignment of the distance boundaries; and micrometers, which typically have greater precision due to a combination of linear and circumferential scales. Caliper precision can be improved with vernier scales or linear transducers. Fixed gages are often used to measure objects on a pass/fail basis. An object of fixed geometry (length, tapered bore, thread, etc.) is compared to a test piece typically for part inspection. Variations include the master gage, an object used to represent the nominal dimension of the part; the limit gage, an object used to represent the limit condition for tolerance dimensioning; and gage blocks or Johansson blocks, an object of fixed length used as a dimensional reference standard. Dial indicators are used to sense displacement...
from a reference plane and display the deviation thereof. The display can be electronically coupled for amplification and/or display purposes. The range of a measuring instrument may be extended if multiple copies of the object to be measured are available. For example, the thickness of a sheet of paper may be measured by a simple ruler if 500 sheets of paper are stacked. (Vendors: Brown & Sharpe, Starrett, Mitutoyo. Also see [3].)

**Electronic Gages**

A Linear Variable Differential Transformer (LVDT), utilizes multiple toroidal transformers to sense axial displacement of an iron core that is attached to a measuring contact, either directly or by another joint (such as a lever). The displacement has a direct correlation to the distance that other electronics display. Thus, the LVDT serves as a replacement for a lined ruler or micrometer, incorporating an electrical readout. (Vendors: Penny + Giles; Schaevitz; Transicoil Inc.; Trans-Tek.)

---

**TABLE 7.2 Vendor addresses.**

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bomem</td>
<td>Quebec, Canada</td>
</tr>
<tr>
<td>Brown &amp; Sharpe</td>
<td>North Kingstown, RI</td>
</tr>
<tr>
<td>CMI International</td>
<td>Elk Grove Village, IL</td>
</tr>
<tr>
<td>Conductus</td>
<td>Sunnyvale, CA</td>
</tr>
<tr>
<td>deFelsko</td>
<td>Ogdensburg, NY</td>
</tr>
<tr>
<td>Digilab</td>
<td>Cambridge, MA</td>
</tr>
<tr>
<td>Digital Instruments</td>
<td>Santa Barbara, CA</td>
</tr>
<tr>
<td>Electromatic</td>
<td>Cedarhurst, NY</td>
</tr>
<tr>
<td>Fischer</td>
<td>Windsor, CT</td>
</tr>
<tr>
<td>Hewlett Packard</td>
<td>Englewood, CO</td>
</tr>
<tr>
<td>Kta-Tator</td>
<td>Pittsburgh, PA</td>
</tr>
<tr>
<td>Magnetic Analysis Corp.</td>
<td>Mount Vernon, NY</td>
</tr>
<tr>
<td>Mattson</td>
<td>Madison, WI</td>
</tr>
<tr>
<td>Measurex</td>
<td>Cupertino, CA</td>
</tr>
<tr>
<td>Micro Photonics</td>
<td>Allentown, PA</td>
</tr>
<tr>
<td>Midac</td>
<td>Costa Mesa, CA</td>
</tr>
<tr>
<td>Mitutoyo</td>
<td>Plymouth, MI</td>
</tr>
<tr>
<td>Moore Products Co</td>
<td>Spring House, PA</td>
</tr>
<tr>
<td>NDC Systems</td>
<td>Irwindale, CA</td>
</tr>
<tr>
<td>Nicolet</td>
<td>Madison, WI</td>
</tr>
<tr>
<td>Ono Sokki</td>
<td>Addison, IL</td>
</tr>
<tr>
<td>Oxford Instruments</td>
<td>Concord, MA</td>
</tr>
<tr>
<td>Panametrics</td>
<td>Waltham, MA</td>
</tr>
<tr>
<td>Park Scientific Instruments</td>
<td>Sunnyvale, CA</td>
</tr>
<tr>
<td>Penny + Giles</td>
<td>Attleboro, MA</td>
</tr>
<tr>
<td>Perkin Elmer</td>
<td>Norwalk, CT</td>
</tr>
<tr>
<td>Phase-Shift Technology</td>
<td>Tucson, AZ</td>
</tr>
<tr>
<td>Rudolf Instruments</td>
<td>Fairfield, NJ</td>
</tr>
<tr>
<td>Scantron</td>
<td>Dist. by Micro Photonics</td>
</tr>
<tr>
<td>Schaevitz</td>
<td>Pennsauken, NJ</td>
</tr>
<tr>
<td>Sentech</td>
<td>Dist. by Micro Photonics</td>
</tr>
<tr>
<td>SolveTech</td>
<td>Claymont, DE</td>
</tr>
<tr>
<td>Starrett</td>
<td>Athol, MA</td>
</tr>
<tr>
<td>Stresstel</td>
<td>Scotts Valley, CA</td>
</tr>
<tr>
<td>Transicoil Inc.</td>
<td>Valley Forge, PA</td>
</tr>
<tr>
<td>Trans-Tek</td>
<td>Ellington, CT</td>
</tr>
<tr>
<td>Willrich Precision Instrument Co.</td>
<td>Cresskill, NJ</td>
</tr>
<tr>
<td>J.A. Woolam Co., Inc.</td>
<td>Lincoln, NE</td>
</tr>
<tr>
<td>Wyko</td>
<td>Tucson, AZ</td>
</tr>
<tr>
<td>Zygo</td>
<td>Middlefield, CT</td>
</tr>
</tbody>
</table>

© 1999 by CRC Press LLC
Pneumatic Gaging

Pneumatic gages have pressurized air exiting gage orifices. The air velocity differential or backpressure is a function of the separation of the gage and the part. In the direct or open jet method, the pressurized air experiences backpressure due to the impedances posed by the measured part. The typical scenario is that the gage head and the measured part have similar geometry (i.e., a cylindrical gage in a bored hole). By placing two gages on either side of a flat plate, the thickness may be inferred. In the indirect or contact method, the pressurized air pushes on a contact piece that directly contacts the part. Tolerances as small as 50 nm can be measured. (Vendors: Willrich Precision Instrument Co.; Moore Products Co.)

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Model Number</th>
<th>Price</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>KTa-Tator</td>
<td>TI-12</td>
<td>$1595.</td>
<td>General-purpose ultrasonic gage, 0.75 mm to 75 mm range</td>
</tr>
<tr>
<td>NDC Systems</td>
<td>6100TC</td>
<td>$49,300.</td>
<td>Backscatter gamma gage for 60 in. web, 25 mm range</td>
</tr>
<tr>
<td>NDC Systems</td>
<td>25DL</td>
<td>$66,600.</td>
<td>Transmission beta gage, for continuous web products</td>
</tr>
<tr>
<td>Panametrics</td>
<td>25DL Plus</td>
<td>$2200. to $3800.</td>
<td>Single-element ultrasonic gage, 50 mm range</td>
</tr>
<tr>
<td>Panametrics</td>
<td>8000</td>
<td>$1400. to $2500.</td>
<td>Dual-element ultrasonic gage, 250 mm range</td>
</tr>
<tr>
<td>DeFelsko</td>
<td>Positest 1000-N</td>
<td>$1995.</td>
<td>Hall effect magnetic gage, for nonferrous materials, 6 mm range</td>
</tr>
<tr>
<td>Magnetic Analysis</td>
<td>Various</td>
<td>$1500. to $100,000.</td>
<td>Ultrasonic, time-of-flight gages</td>
</tr>
<tr>
<td>Fischer</td>
<td>Deltascope MP2C</td>
<td>$1200.</td>
<td>Magnetic induction gage, measure nonmagnetic coating on ferromagnetic substrate</td>
</tr>
<tr>
<td>Fischer</td>
<td>IsoScope MP1C</td>
<td>$1200.</td>
<td>Eddy current gage, measure nonconducting coating on nonferrous conducting substrate</td>
</tr>
<tr>
<td>Fischer</td>
<td>Fischerscope MMS</td>
<td>$6500.</td>
<td>Beta-backscatter system to measure coating thickness</td>
</tr>
<tr>
<td>Fischer</td>
<td>Fischerscope X-Ray 1020 video</td>
<td>$34,000.</td>
<td>X-ray fluorescence system to measure coating thickness</td>
</tr>
<tr>
<td>Fischer</td>
<td>Couloscope Sx</td>
<td>$2500. + accessories</td>
<td>Electrolytic, destructive system to measure coating thickness</td>
</tr>
<tr>
<td>J.A. Woollam Co. Inc.</td>
<td>M-44</td>
<td>Application specific</td>
<td>Variable angle, multiwavelength spectroscopic ellipsometer</td>
</tr>
<tr>
<td>Rudolf Instruments</td>
<td>431A31WL633</td>
<td>$10,100.</td>
<td>Manual, HeNe wavelength ellipsometer</td>
</tr>
<tr>
<td>Rudolf Instruments</td>
<td>444A12</td>
<td>$34,000.</td>
<td>Automatic, HeNe wavelength ellipsimeter</td>
</tr>
<tr>
<td>Hewlett Packard</td>
<td>HP8712C</td>
<td>$13,500.</td>
<td>RF vector network analyzer, measure transmission/reflection frequency response to 1.3 GHz, optional to 3 GHz</td>
</tr>
<tr>
<td>Stresstel</td>
<td>T-Mike Programmable</td>
<td>$995.</td>
<td>Dual-element ultrasonic system</td>
</tr>
<tr>
<td>Stresstel</td>
<td>TM1D</td>
<td>$1795.</td>
<td>Single-element ultrasonic system</td>
</tr>
<tr>
<td>Measurex</td>
<td>DMC480</td>
<td>Application specific</td>
<td>High-speed X-ray thickness gage</td>
</tr>
<tr>
<td>Bomem</td>
<td>MB series</td>
<td>$20,000. and up</td>
<td>1 cm⁻¹ resolution Fourier transform spectrometer</td>
</tr>
<tr>
<td>Park Scientific Instruments</td>
<td>Autoprobe CP</td>
<td>$65,000.</td>
<td>Ambient scanning probe microscope</td>
</tr>
<tr>
<td>Park Scientific Instruments</td>
<td>Autoprobe VP2</td>
<td>$130,000.</td>
<td>UHV scanning probe microscope</td>
</tr>
<tr>
<td>Digital Instruments</td>
<td>Nanoscope IIIa/D3000</td>
<td>$90,000.</td>
<td>Small sample scanning probe microscope</td>
</tr>
</tbody>
</table>
Optical: Focusing, Shadowing, Comparing

This includes microscopes, which can determine thickness either by comparison with a known reference, or by focusing on the front and rear surfaces of a sample, noting the difference in focus position. Comparators project onto a screen what might be noted through a microscope. Laser calipers retrieve dimensions by measuring the shadowing of a laser beam. (Vendors: NDC Systems for laser caliper; Scantron for laser profilometer.)

Weighing

Given a plate of material with known density, first measure the area with some type of calibrated video system. Then, a measurement of weight can be simply converted to an estimate of the thickness. As is common with this technique and most of the following techniques, estimating the thickness requires knowledge of some other property of the material to be measured — in this case, the density.

Capacitive Gaging

Capacitive gaging is realized by inserting a nonmetallic material into a known electric field. Knowing the gage sensor area and the material’s dielectric constant, the thickness can be determined. Submicron thickness levels can be achieved. (Vendors: Ono Sokki; SolveTech.)

Inductive Gaging (Eddy Current Sensing)

The principle here is that ac currents in a coil induce eddy currents in a nearby conducting plate [4, 5]. These eddy currents can be sensed by a pickup coil, which may be the exciting coil or a second coil. The presence of the eddy currents manifests itself as a modification of the apparent inductance and/or the loss of the pickup coil. This technique is appropriate for nonferrous metals, and is especially sensitive to thickness variations due to flaws such as cracks or corrosion. There is one particular instance in which it is common to measure thickness rather than variations. That would be the thickness of a nonconductive coating on a nonferrous conducting substrate. The coating thickness creates a gap (lift-off) between the exciting coil and the eddy currents, thereby affecting the eddy current signal. The range of this technique would be about 1 mm. Fischer has an instrument designed for measuring the thickness of a newly laid road surface coating to a depth of 40 cm, by burying a conductive plate below the road. (Vendors: Fischer; deFelsko; CMI International.)

Magnetic Induction

This technique is also used to measure coating thickness, in this case a nonmagnetic coating on a ferromagnetic substrate. The nonmagnetic coating creates a gap (lift-off) between the ferromagnetic substrate and a probe. One way to measure the gap and thereby the thickness is by measuring the force required to pull away a magnetic probe. Another technique would be to magnetically couple the ferromagnetic substrate to a transformer core, with a gap between the substrate and the core. This technique would have a range of about 4 mm. CMI International has an informative brochure describing the relative merits of measuring coating thickness via eddy current, magnetic induction, beta-backscatter, microresistance, and X-ray fluorescence; the choice of technique depends, among other things, on the material to be tested. (Vendors: Fischer; CMI International; Electromatic; deFelsko.)

Hall Effect Gage

This sensor measures the thickness of nonferrous materials with 1% accuracy by sandwiching the material being measured between a magnetic probe on one side and a small target steel ball on the other side [6].
It measures up to 10 mm. The Hall effect sensor is used to measure the magnetic field, as a dc measurement; ac Hall effect measurements can be made more precisely because they eliminate bias and are done with less noise. (Vendor: Panametrics.)

**Far-Field/Time-of-Flight: Ultrasound, Radar, Lidar**

Using 1940s sonar principles and today’s microprocessor technology, high-frequency (1–20 MHz) ultrasound waves can be used to measure thickness by sending pulsed sound waves through a material and measuring the transit time of the reflected signal [5, 7]. Knowing the sound velocity of the material, materials from 0.5 mm to 250 mm can be measured, often as fine as 25 μm. Media include metal, glass, ceramic, liquid, rubber, fiberglass, plastic, and concrete. Ultrasound can also be used to measure living tissues, as is often done in the agricultural and medical fields. Fat layers of cattle and pigs can influence marketability. Skin burn depths can direct treatment procedures. The depth of foreign objects in the body is useful for microsurgery. Ultrasonic thickness determination has expanded to include multidimensional echolocation applications, such as imagery and acoustic microscopes that can resolve in the submicron level. The principles behind ultrasound also apply to electromagnetic waves. In the gigahertz range, this is called radar. Radar can be used to estimate the thickness of atmospheric layers such as cloud layers. The light-wave version of radar, called lidar, can be used to measure the thickness of water vapor layers in the lower atmosphere. (Ultrasonic vendors: KTa-Tator; deFelsko; Stresstel; Magnetic Analysis Corp.; Panametrics; Electromatic.)

**Far-Field/Resonance: Ultrasound, Interferometry, Ellipsometry**

The idea here is that when waves such as ultrasound impinge on a plane-parallel slab of material, there will be reflected power from both the front and rear surfaces; depending on whether the slab thickness is an odd or even number of quarter-wavelengths, the reflected beams will be in constructive or destructive interference. If the frequency is swept, the distance in frequency between successive maxima and minima may be related to the slab thickness, if the index of refraction is known. Since the natural, or resonant, modes of an object depend on the properties and dimensions of an object, knowledge of the properties enables estimation of dimensions from the resonant frequencies. Compared with time-of-flight ultrasound, resonant ultrasound is much less common. It has been used to characterize concrete, and is quite sensitive to flaws, as anyone who has heard a cracked bell would know. Resonant techniques are much more common with visible [8], infrared, or microwave [4] radiation. Spectral interferometry would be appropriate to characterize the thickness of transparent substrates with reasonably flat surfaces, sufficiently parallel to one another. A common way to do this would be to measure a transmission spectrum with a spectrometer such as a Fourier transform spectrometer (FTS). The successive maxima and minima are here called the Fabry–Perot effect, and their appearance in a spectrum is called channeling. Thickness can also be measured with spatial interferometry, which is essentially a way of measuring surface topography. An example would be the phenomenon of Newton’s rings, which occur when the surface to be tested is in contact with an optical flat. Using a transparent optical flat, transmit monochromatic light such as a mercury lamp through the flat and onto the interface between the flat and the test surface. If there are variations in the height of the test surface, then the two return beams from the optical flat and the test surface will alternate between constructive and destructive interference, producing fringes or rings. The sensitivity is not limited to the scale of the wavelength λ: with sufficient stability and signal-to-noise, dimensions down to 1/1000 of a fringe can be measured. With sources of longer coherence length, such as lasers, the test surface and the optical flat need not be in direct contact.

Another optical way to measure thickness is with ellipsometry [9], typically used to measure properties of thin, transparent films from a few tenths of nanometers to several hundreds of nanometers thick. This includes metals, as long as the metal is sufficiently thin to be partially transparent. By measuring the change in polarization state for nonnormal incidence light, both the thickness and refractive index of a thin layer may be inferred. Additional information (e.g., the properties of multiple layers) can be obtained.
by varying the angle of incidence and by observing at multiple wavelengths. The ability to estimate both thickness and refractive index is an important advantage of this technique, as often the refractive index of a material in thin film form is not the same as the bulk value, and indeed may be a property of the deposition conditions. (FTS vendors: Bomem; Digilab; Mattson; Midac; Nicolet; Perkin-Elmer. FTS system pricing may range from about $15,000 to over $100,000, depending on the application. Spatial interferometer vendors: Zygo; Wyko, Phase-Shift Technology. Ellipsometer vendors: J.A. Woollam Co., Inc.; Rudolf Instruments; Sentech. The cost of an ellipsometer may range from $10,000 for a manual, single-wavelength system to $200,000 for an automatic, multiwavelength system. Microwave resonance vendor: Hewlett-Packard.)

Far-Field/Absorption, Scattering, Emission: Beta, Gamma, X-Ray, Infrared

These techniques depend on the extinction (scattering or absorption) or emission of photons or massive particles (electrons, protons, neutrons) when transiting the material to be measured. Typically, the extinction or emission shows an exponential dependence on thickness; the dependence becomes linear if the absorption is sufficiently low. These techniques, in particular gamma-ray backscatter and beta-ray transmission, are used to measure continuously moving web materials (paper, metals, fabrics) on assembly lines. Infrared absorption is also suitable if the moisture content is controlled. Beta-backscatter and X-ray fluorescence [10] are used for measuring coatings. In X-ray fluorescence, upon exposure to X-rays, certain elements fluoresce (emit) X-rays at characteristic wavelengths. The strength of this emission is related to thickness. These absorption/emission techniques may sometimes be better suited than time-of-flight ultrasound to the dimensional measurement of objects with complex shapes. (Gamma gage vendor; NDC Systems. X-ray absorption vendor: Measurex. X-ray fluorescence vendors: Fischer; NDC Systems; CMI International. Beta-backscatter vendor: Fischer; Electromatic; CMI International; Measurex. Infrared absorption vendor: NDC Systems. The prices for these systems will depend on the application; a typical system could cost $500,000.)

Destructive Techniques

Fischer markets a system that removes a coating into an electrolyte and then electrolytically deposits the removed coating. The electrical charge required for deposition is related to the coating thickness.

7.2 Future Directions in Thickness Measurement

Concerning Techniques Mentioned Above

Concerning capacitive sensors, the NASA Langley Research Center is developing sensors based on patterns of conductors sandwiched between insulating layers. The presence of ice over the conductors changes the capacitance, providing a way of sensing ice build-up on aircraft wings. With respect to eddy current sensing, one limitation is that a nonsuperconducting sense coil responds best to high-frequency excitations, and not at all to dc magnetic fields. This limits the technique to fairly high frequencies and thus low penetration depths, since the skin depth becomes shallower with increasing frequency. One possibility is to use a SQUID (superconducting quantum interference detector) as the sensor, since the SQUID is probably the most sensitive sensor of dc and low-frequency magnetic fields. One disadvantage of the SQUID has been the need for liquid helium for cooling for low-temperature superconductors; with the recent availability of high-temperature superconductors (HTS, above 90 K) and now HTS SQUIDs, cooling can be done with liquid nitrogen or single-stage mechanical coolers. In the area of spatial interferometry, work at Lawrence Livermore National Laboratory replaces the reference surface with a single-mode fiber in a process called phase-shifting diffraction interferometry. A measurement accuracy of 1.44 nm rms is quoted, with a goal of 0.1 nm rms. (HTS SQUID vendor: Conductus.)
**THz Technology**

With the availability of femtosecond pulsed lasers, Bell Labs has been investigating a technique using 100 fs pulses to pulse an antenna in the range of 0.1 THz to 3.0 THz. The terahertz pulses are sent through the material to be tested, detected, and the received pulse shape is analyzed to extract constituent information. This technique may also provide information on thickness.

**Nanoscale-Scanning Probe Microscopy**

Scanning probe microscopes (SPMs) are used in a wide variety of disciplines, including fundamental surface science, routine surface roughness analysis, and spectacular three-dimensional imaging — from atoms of silicon to micron-sized protrusions on the surface of a living cell [11]. The scanning probe microscope is an imaging tool with a vast dynamic range, spanning the realms of optical and electron microscopes. It is also a profiler with unprecedented 3-D resolution. In some cases, scanning probe microscopes can measure physical properties such as surface conductivity, static charge distribution, localized friction, magnetic fields, and elastic moduli. As a result, applications of SPMs are very diverse. The scanning tunneling microscope (STM), the progenitor of SPMs, utilizes a sharp conductive tip with a bias voltage applied between the tip and the sample. When the tip is within 1 nm of the sample, electrons from the sample begin to tunnel through the 1 nm gap into the tip. If the bias voltage is reversed, the tunneling occurs into the sample. The tunneling current is a function of the separation. Both the tip and the sample must be conductors or semiconductors.

The atomic force microscope (AFM) utilizes a small tip at the end of a cantilever. Forces between the tip and sample cause a deflection in the cantilever, which is translated into a signal. The tip or sample can be scanned covering a large area, producing a topographical map. AFMs can be used on insulators or conductors. AFMs are used in two modes: contact and noncontact. In contact mode, the tip is brought within about 200 pm — about the length of a chemical bond. The electron clouds of the tip and sample atoms interact, netting a repulsive force. For this reason, the contact mode is also called repulsive. Vertical resolution of about 50 pm can be achieved. In noncontact mode, a vibrating cantilever is used in the attractive regime of the van der Waals interactions. The cantilever is typically 2 nm to 20 nm away from the sample surface and has low total force. Noncontact AFM is subsequently less sensitive; thus, sensitive detection systems must be employed. The low force does have the advantage of not contaminating the sample surface and is preferred for applications involving silicon wafers and soft or elastic tissues. In noncontact mode, the cantilever is resonated with a small amplitude. As the tip comes near the sample surface, the resultant force changes the spring constant, translating into a deviation of the resonance frequency. This change in resonance (or vibrational amplitude) reflects changes in the sample topology.

Intermittent-contact mode is a combination of noncontact and contact modes and best suited for soft, adhesive, or fragile samples. Contact mode can damage the tip and the sample due to frictional or shear forces and/or create data artifacts from tip/surface adhesion. Noncontact mode produces lower amplitudes and hence lower resolution. Furthermore, surface monolayers of adsorbed gases such as water vapor can produce erroneous results. Intermittent-contact mode avoids these pitfalls by placing the tip in contact with the surface, providing high resolution and then removing the tip to prevent dragging and/or lateral forces. The cantilever is resonated via a piezoelectric crystal (50 kHz to 500 kHz in ambient, 5 kHz to 40 kHz in fluids) overcoming the tip/sample adhesion forces.

In magnetic force microscopy (MFM), the noncontact mode is employed using a tip coated with a ferromagnetic film. Both magnetic and van der Walls interactions are present, but at larger tip/sample separations, the magnetic forces dominate. Multiple scans as a function of tip/sample distance allow differentiation of magnetic forces and topographic information. Magnetic domain structures are resolved to 50 nm via this technique. Current applications of MFM include data storage devices, imaging of micromagnetic structures, IC analysis, imaging of magnetotactic bacteria, and magnetic geophysics. Lateral force microscopy (LFM) is used to generate profiles of changes in surface friction and/or height variations. The probe tip is deflected laterally, indicating some sort of twist. Electronics measure the cantilever deflection. To differentiate between the two effects, LFM and AFM images should be obtained.

© 1999 by CRC Press LLC
simultaneously. Phase detection microscopy or phase imaging is an extension of intermittent-contact AFM. It utilizes the phase lag between the driving frequency (cantilever) and the output signal frequency, generating a map of specific mechanical properties such as adhesion, elasticity, and friction. Identification of contaminants, composite materials, and regions of hardness and low surface adhesion can be obtained at the nanometer scale. Additional techniques include force modulation microscopy, where a periodic signal is applied to the cantilever, generating a map of the sample's elastic modulus and/or contaminants; electrostatic force microscopy, where a charged tip is scanned over the sample, revealing the locally charged domains generating a map of the charge carrier density; scanning capacitance microscopy, where a charged tip, kept at a constant tip/sample distance, generates a map of capacitance correlated information such as dielectric material thickness and subsurface charge carrier distributions (i.e., dopant profiles of ion implanted semiconductors); thermal scanning microscopy, where the tip in noncontact mode and a bimetal cantilever are used to map the thermal conductivity of the sample. (Vendors: Park Scientific Instruments; Digital Instruments; Oxford Instruments.)

References

The objective of this chapter is to review the state-of-the-art in proximity-sensing technologies for robotics. Special attention is paid to the sensing needs of robotic manipulators for grasping applications, in contrast to the needs of mobile robots for navigation purposes. For a review of the application of proximity sensing to mobile robots, the reader is referred to [1].

Robotic sensors can be categorized into three groups: medium-range (object recognition and gross position/orientation estimation) sensors, short-range (proximity) sensors, and contact sensors. Recent literature [2–6] suggests that robotic end effectors should be equipped with both short-range proximity and contact sensors.

Proximity sensors should be able to measure the position and orientation (pose) of an object’s surface. The range must be sufficiently large to compensate for uncertainties in the medium-range pose-estimation process, while maintaining sufficient accuracy to permit effective grasping of the object.

Transducers used by current proximity sensors vary in sophistication. Despite their great variety, however, these transducers and their accompanying electronic interface circuits (together comprising the proximity sensor) cannot presently meet the stringent robustness requirements of most industrial robotic applications. Novel sensing algorithms and techniques still must be developed in order to improve on their current characteristics, and, furthermore, to control both the sensing and grasping processes.

8.1 Proximity Definition

The term “proximity,” quantified by “pose” in this chapter, refers to three geometrical parameters \( x \), \( u \), and \( v \) as shown in Figure 8.1, where:

- \( x \) = the translation from the origin of the sensor’s reference coordinate frame, \( F_p \), to a target point on the surface of the object measured along \( X_p \). This target point defines the origin of the surface-frame, \( F_r \).
- \( u \) = the vertical orientation of the object’s surface, defined as a rotation around \( Y_p \) (of the translated frame), thereby specifying the new \( Z_r \).
- \( v \) = the horizontal orientation of the object’s surface, defined as a rotation around \( Z_r \), thereby specifying \( Y_r \).
8.2 Typical Sensor Characteristics

Conventionally, proximity sensors should be capable of measuring distances of up to 50 mm, and two degree-of-freedom orientations equivalent to an overall inclination of up to $\pm 30^\circ$. The intended principal application of the sensor is to act as a guide for the robot. Thus, it would be desirable to have higher sensitivity and accuracy as the gripper approaches the object, namely when both the relative orientation and the distance approach near-zero values.

The signals received by the electronic interface circuit should be processed without limiting the required operating range of the sensor. The interface circuit should also minimize the effect of interference from the surroundings. It should therefore employ solutions to reduce background-noise interference and dynamic-range limitations.

The operation of the robot should not be slowed down by the sensor. Namely, a pose of the object should normally be estimated in 1 ms to 10 ms.

8.3 Technologies for Proximity Sensing

Proximity sensors have employed various transduction media, including sound waves, magnetic fields, electric fields, and light. Presently, electro-optical techniques seem to be the most appropriate for robotic-grasping applications. Such sensors are relatively small in size, have a large range of operation, and impose almost no restrictions on the object’s material. However, recently, some new ultrasonic and capacitive proximity sensors have been fabricated directly as ICs, also showing the possibility of very-small-size proximity sensors based on these technologies [7, 8].

Brief descriptions of the principles of the primary technologies used by proximity sensors are given below, with the main emphasis being on optical transducers. A survey of commercial proximity sensors capable of measuring distances can be found in [6].
Electro-Optical Sensors

Many proximity sensors use light, directly scattered from a target surface, to determine the distance and orientation of the target object from the gripper. The mechanism by which light is reflected can be explained by a model that specifies four different reflection phenomena. According to this model, light reflects from the surface primarily as a result of one or more of the following interactions:

1. Single surface reflection: Light waves that reflect specularly a single time off a planar microfacet, whose dimensions are significantly larger than the wavelength.
2. Multiple surface reflection: Light waves that reflect specularly at least twice between multiple microfacets.
3. Reflection after penetration: Light waves that penetrate into the material, refract, and then reflect back out as diffused light.
4. Corner reflection: Light waves that diffract from interfaces with surface details about the same size or smaller than the wavelength (such as from corners of microfacets).

The primary phenomenon (1) usually exists in both dielectrics and metal. However, due to the high conductivity of metal surfaces, most of the light reflects specularly off the interface between the metal and the air, while the portion that penetrates into the metal surface is absorbed. Accordingly, the reflection intensity originating from internal refraction in metals is practically zero. In dielectrics, however, a large portion of the light penetrates into the surface, and then reflects back out as diffused light (3). The secondary phenomena (2) and (4) exist both in metals and dielectrics and add to the diffused reflectance.

Common measurement techniques used in optical proximity sensing utilize one or more of the reflected components to determine the pose of the object in relation to the transducer.

Phase Modulation

A phase-modulated (PM) proximity sensor usually consists of two light sources and one or more photodetectors. The light sources are driven by modulated sinusoidal signals having a 90° phase relationship (Figure 8.2).

The emitter control voltages of the emitters, $V_{em1}$ and $V_{em2}$, have amplitudes of $a$ and $b$, respectively:

$$V_{em1}(t) = a \cdot \cos \omega t$$  
(8.1)

$$V_{em2}(t) = b \cdot \sin \omega t$$  
(8.2)
The signal detected by the receiver is a superposition of the two reflected signals, having corresponding attenuations of \(A\) and \(B\).

\[
V_{\text{rec}}(t) = A \cdot V_{\text{em1}} + B \cdot V_{\text{em2}}
\]  

The signal attenuation is a function of the geometrical and electrical parameters of the sensor, the reflectivity characteristics of the object’s surface, and the surface’s distance and orientation with respect to the sensor. The combined signal at the receiver is therefore:

\[
V_{\text{rec}}(t) = M \cdot \sin(\omega t + \phi)
\]  

where \(M\) = the combined attenuation-function, and \(\phi\) = the combined phase-shift.

Usually, only the phase information \(\phi\) is used, and the amplitude is completely neglected or used only for verifying the likelihood of error and its potential magnitude.

A proximity sensor that uses this technique has been reported in [9]. Figure 8.3 shows a sensor head; it comprises six light sources (LEDs) and a photodetector (phototransistor). This sensor can measure the distance from the sensor’s coordinate frame to the target point on the surface of the object (\(x\)), as well as the horizontal and vertical orientation of the object surface (\(u, v\)).

A simple model for the sensor was developed assuming that the light sources (LEDs) have low directivity, the photodetector (phototransistor) has high directivity, and the surface has diffused reflectivity. Figure 8.4 shows the basic configuration for the measurement of distance (\(x\)) and orientation (\(u, v\)). Table 8.1 shows the combinations of the driving signals in each LED needed for the measurement of distance and orientation.

For the measurement of the distance \(x\), LED1 and LED3 are modulated by \(K_1 \sin \omega t\), and LED2 and LED4 by \(K_2 \cos \omega t\), respectively. The brightness detected by the photodetector can be calculated (using Lambert’s law) to be:

\[
I_y = C \left( G_1 \frac{\cos(\alpha - \nu)}{a^2 + x^2} + G_2 \frac{\cos(\beta - \nu)}{b^2 + x^2} + G_3 \frac{\cos(\alpha + \nu)}{a^2 + x^2} + G_4 \frac{\cos(\beta + \nu)}{b^2 + x^2} \right)
\]  

where \(G_i (i = 1, 2, 3, 4)\) are the intensities of the light sources, and \(C\) is the reflection factor of the surface at point \(P\). Considering that for this case: \(G_1 = G_3 = K_1 \sin \omega t\), \(G_2 = G_4 = K_2 \cos \omega t\), \(\cos \alpha = x/\sqrt{a^2 + x^2}\) and \(\cos \beta = x/\sqrt{b^2 + x^2}\), Equation 8.5 can be rewritten as:
The amplitude $M$ and the phase shift $\phi_x$ are given by:

$$
L_y = 2Cx \left[ \frac{K_1}{\left( a^2 + x^2 \right)^{3/2}} \sin \omega t + \frac{K_3}{\left( b^2 + x^2 \right)^{3/2}} \cos \omega t \right] \cos \nu = M \sin (\omega t + \phi_x) \tag{8.6}
$$

The amplitude $M$ and the phase shift $\phi_x$ are given by:

$$
M = 2Cx \left[ \frac{K_1^2}{\left( a^2 + x^2 \right)^3} + \frac{K_3^2}{\left( b^2 + x^2 \right)^3} \right]^{1/2} \cos \nu \tag{8.7}
$$

and

$$
\phi_x = \tan^{-1} \left[ \frac{K_1 \left( a^2 + x^2 \right)^{3/2}}{K_3 \left( b^2 + x^2 \right)^{3/2}} \right] \tag{8.8}
$$

From Equation 8.8, it can be observed that the distance $(x)$ can be obtained from $\phi_x$. It is important to note that, in theory, the calculation is not affected by the reflection factor $(C)$ of the surface.
Similarly, the orientation angles \( u \) and \( v \) can be obtained by driving the LEDs as indicated in Table 8.1. For example, \( v \) can be determined by modulating LED4 and LED2 by the signals \( K_1 \sin \omega t \) and \( K_2 \cos \omega t \), respectively. For this case, the phase shift associated with the brightness of the object at point \( P \) is given by:

\[
\phi_v = \tan^{-1} \left[ \frac{K_1 (x - b \tan v)}{K_2 (x + b \tan v)} \right]
\]  

(8.9)

Note that, in order to recover \( v \) from Equation 8.9, \( x \) must be known. Accordingly, the distance \( x \) must be determined first. Correspondingly, the orientation angle \( (v) \) can be calculated from the new phase shift \( \phi_v \). The angle \( u \) can be calculated by modulating LED6 and LED5 with \( K_1 \sin \omega t \) and \( K_2 \cos \omega t \), respectively, and then determining the corresponding phase shift of the associate brightness at point \( P \).

The pose-estimation results using this sensor were quite satisfactory and showed a good agreement between the theory and experiment.

In [10], an experimental setup of a PM distance sensor, similar to the one in [9], was reported for investigating the effect of the geometric and electronic parameters on the performance of the sensor. Optimal parameters were obtained for some targeted sensor-operation characteristics.

**Amplitude Modulation**

In amplitude-modulated (AM) sensors, the magnitude of the light reflected from a surface is utilized to determine the pose of the object.

AM transducers usually consist of one light source and several photodetectors (Figure 8.5). They were redesigned and optimized several times over the past decade to yield better measurement accuracy [11–14].

Many AM proximity sensors utilize optical fibers to illuminate and collect light from the surfaces of objects. The use of optical fibers, in a Y-guide configuration (Figure 8.6), facilitates the operation of sensitive low-noise circuitry in a shielded environment appropriately remote from the robot’s electromagnetic interference sources.

AM transducers primarily use variations of the basic Y-guide transducer. Two important parameters can be varied in the design of Y-guides: the distance, \( d \), between the emitting and receiving fibers (referred to hereafter as the emitter and the receiver, respectively), and the inclination angle, \( \theta \), of the receiver fiber with respect to the transducer’s surface. The emitter is usually placed perpendicular to the transducer’s surface, due to symmetry requirements, as will be explained later in this section.

The collection of a sufficient amount of reflected light requires the use of relatively wide-diameter fibers, typically having a 0.3 mm to 2 mm core size. This requirement demands the use of relatively low-grade plastic fibers. Although attenuations of up to 1 dB m\(^{-1}\) are common in such plastic fibers, this loss rate is relatively insignificant for Y-guide applications because of the short length of the cables normally used. The numerical aperture (NA) of the plastic fibers, on the other hand, is an important parameter.
in the transducer design, since lenses are rarely used in conjunction with AM-type transducers. In particular, the acceptance angle of the fiber is fixed and given by  $\alpha = 2\sin^{-1} \text{NA}$.

For a Y-guide, the intensity of the light reflected from the surface is not a monotonic function of the distance. Thus, the minimum operating distance of the transducer ($x_{\text{min}}$) is usually limited to a value that will guarantee a monotonic response (Figure 8.7).

For the measurement of surface orientation, a symmetrical three-fiber constellation (Figure 8.8) can be used. In this Y-guide configuration, the emitter is at the center and the two receivers are positioned symmetrically on either side [12]. The light intensities detected by the receivers, for the transducer shown in Figure 8.8, are illustrated in Figure 8.9 as a function of the surface orientation.

In the usual operating range of an AM transducer, the intensity of the light at the receiver is inversely related to the distance squared. As a result, it is conceptually possible to configure a transducer such that
its sensitivity and accuracy will increase as the gripper nears the contact point, at which both the distance and the orientation of the object’s surface are zero [10, 12]. However, in practice, because of the limited dynamic range of the electronic transducer interface, a trade-off exists between the desired maximum accuracy near contact, and the maximum range of operation. These and other considerations must be taken into account when establishing the geometric features of the transducer.

Another important factor to take into account in the design of an AM sensor is the need to reduce, as much as possible, the effect of the variation in the emitting power of the light source, $P_o$, on the transducer’s measurements. This normally leads to the employment of a pair of receivers. A normalized differential voltage (DV) estimation scheme, such as the following, is then applied to the pair of measurements:

$$DV = \frac{V_{rec_1} - V_{rec_2}}{V_{rec_1} + V_{rec_2}}$$

(8.10)

where $V_{rec_1}$, $V_{rec_2}$ are the voltages measured by receivers 1 and 2. However, in order to eliminate the effect of $P_o$ on $DV$, each receiver must linearly convert the light intensity to a corresponding voltage measurement.

In order to use a DV scheme for the measurement of distance, an asymmetrical transducer configuration can be used (Figure 8.10). However, one must note that, although orientation measurements are not affected by variations in distance, distance measurements are significantly affected by the orientation of the surface, e.g., [15].

Accordingly, in using an AM proximity sensor with a DV scheme, the orientation is first approximated, and subsequently the distance is determined. The accuracies of the measured distance and orientation angle can be further improved by an iterative process.

FIGURE 8.9  The light intensity detected by each receiver as a function of the surface orientation ($\theta$).

FIGURE 8.10  An asymmetrical receiver-pair constellation for distance measurements.
Based on the above issues and observations, the outputs of the three receivers of the basic AM proximity sensor (Figure 8.5) can be paired for measuring both distance and orientation: the pair rec$_1$-rec$_2$ can be used for orientation measurement, while the pair rec$_1$-rec$_3$ can be used for distance measurement.

An experimental AM proximity sensor, capable of estimating the pose of an object with high accuracy, was reported in [14, 16, 17]. The transducer consists of one emitter, placed perpendicularly to the sensor head, and eight inclined receiver elements (Figure 8.11). The receivers of this transducer were paired for the specific measurements of distance, as well as of the vertical or horizontal orientation. However, the pose of the surface was determined with higher accuracy by using a polynomial fit technique (as opposed to the DV scheme described above), that provided relationships between the individual estimated parameters ($x$, $u$, and $v$) and all eight signals received.

The sensor presented in [14], and shown in Figure 8.11, operates in the range of 0 mm to 50 mm and ±20°. It can achieve an accuracy of 6.25 μm in distance and an accuracy of 0.02° in angular measurements in the near-contact region (0 mm to 6 mm range), using a general calibration-per-group strategy for different material groupings. This implies that the measured object’s material belongs to a calibration group, which includes similar object surface characteristics; for example, machined metals. Better accuracies can be achieved using a calibration-per-surface strategy.

A similar configuration to the one shown in Figure 8.11 was reported earlier in [18] for the measurement of distances, where the orientations of each receiver pair relative to the emitter $\theta_1$ and $\theta_3$ were set at 10°. However, in this case, the apertures of the emitter and receiver were severely restricted by a collimating graded index (GRIN) lens. The emitter diameter was larger than that of the receivers in order to transmit more light. The measurements of the transducer were then processed in two phases: (1) the DVs of all the receiver pairs were processed independently to provide four distance estimations; and, (2) the four distance estimations were then averaged to provide a more accurate estimate, eliminating adverse effects due to variations in surface orientation.
Geometrical Techniques

Proximity sensors based on geometrical techniques determine the pose of the object by examining the geometrical attributes of the reflected and incident light beams. Two of these techniques, one based on triangulation and the other based on the Gaussian lens law, are presented here.

Figure 8.12 shows the basic configuration of a proximity transducer for measuring distance (x) based on the triangulation technique [19, 20]. The sensor head consists of a laser light source and a linear array of photodetectors (R_i, with i = 1, 2, …, n). A narrow light beam illuminates point P and the receivers detect the reflected light from the illuminated point through a transmitting slit. The geometry of the ray trajectory provides the basic information for the estimation of the distance (x). While the light source illuminates the surface of the object, the photodetector array is scanned to detect the light path used for making the output signal maximum. The light path obtained by this scanning is called the effective light path [19]. This light path is the one indicated in Figure 8.12. The distance (x) can be determined by accurately detecting the position (y_i) and precisely measuring the dimensions (h) and (y_o),

\[ \frac{x}{x + h} = \frac{y_o}{y_i} \]  \hspace{1cm} (8.11)

or

\[ x = \frac{y_i h}{y_i - y_o} \]  \hspace{1cm} (8.12)

In [26], it is claimed that such a sensor has the following properties: (1) the influence of irregularities, reflectivity, and orientation of the object is negligible; (2) the distance measurement is not affected by illumination from the environment and luminance of the object (their influence is eliminated by comparison of two sensor signals obtained in successive on-and-off states of the light source); and (3) the sensor head is sufficiently small to be used in a robot hand.

An experimental proximity sensor configuration, based on triangulation and capable of measuring both distance and orientation, is shown in Figure 8.13 [21]. The sensor uses six infrared LEDs as light sources, an objective lens, and an area-array detector chip for detecting spot positions. The directions of
the beams are aligned to form a fixed cone of light. This sensor is of a type called scanning-emitter since each LED is sequentially pulsed to perform the measurements. As each LED is sequentially pulsed, the sensor IC detects the position of the spot projected by the reflected light beam from the object’s surface. Knowledge of the spot’s position, together with the camera’s optics and the trajectory of each light beam, can be used to perform a triangulation to determine the three-dimensional coordinates of each light spot on the target surface. A set of six 3-D points are obtained for the six LEDs. Then, by fitting a plane to those points, the distance and orientation of the object’s surface are approximated.

Another scanning-emitter-type proximity sensor was reported in [22]. In this case, a mechanical scanning system was utilized. One notes that inherent problems with sensors that use mechanical scanning devices include lower reliability and increased overall size.

Some recently reported triangulation sensors are sufficiently small in size to be mounted on a gripper [5]. However, they are still susceptible to errors due to distortion and separation of the light beam’s reflection caused by surface irregularity, and can also have blind spots as a result of discontinuities associated with the shape of the sensed object.

Another group of geometrical electro-optical proximity sensors are those based on the Gaussian lens law [23, 24]. The basic configuration of such a transducer is shown in Figure 8.14. A light beam, collinear to the optical axis, forms a spot on the target’s surface. The light scattered from the spot is collected by the lens. In Figure 8.14, PN represents the limiting ray that can be collected by the lens. The target distance \( x \) can be calculated in terms of the focal length of the lens \( f \) and the image position \( w \). Applying the Gaussian lens law, the distance \( x \) can be calculated as:

\[
x = \frac{fw}{w-f}
\]  

(8.13)
A proximity sensor (the HexEye) based on this lens principle is shown in Figure 8.15 [23, 24]. The sensor consists of seven identical "sensing units" arranged in a hexagonal pattern. Each sensing unit in turn comprises four main parts: an objective lens, a conical mirror, six linear receiver arrays, and a laser diode light source.

The light beam generated by a light source forms a spot on the target surface, and the light flux scattered from the spot is collected by the objective lens and projected onto the receiver arrays in each sensing unit. The target distance is determined by the position, size, and shape of an image formed on individual receiver arrays of the unit.

The sensor operates in two modes, either distance or orientation. In the distance measurement mode, the seven light sources are activated one at a time to generate a light spot on the target surface, while all the sensing units receive the light flux scattered from the same light spot. The active unit determines the distance using the principle of the Gaussian lens law, while the nonactive units determine the distance based on triangulation. For the orientation measurement mode, the seven sensing units are grouped into orientation measurement units (OMUs). Each OMU comprises three neighboring sensing units around the center, resulting in a total of six possible OMUs. The local orientation of the surface is estimated by integrating the six orientation measurements.

The mapping between the light distribution on the arrays of light detectors and the target distance is obtained through calibration. Using the light distribution (instead of light intensity as in the case of AM sensors), it is intended that the mapping be independent of surface properties such as color, material, diffusion, and reflectance factors. However, it has been shown that the mapping may be corrupted by "noise" from several sources, including reflection patterns and ambient light.
Using the Gaussian lens law has the following advantages over the triangulation principle: (1) the light source can be located at the center of the objective lens, allowing not only the sensor to be compact, but also the amount of light flux input to the lens to be maximized, and (2) the sensitivity can be optimized for a certain range of measurement distance by controlling $f$.

**Time-of-Flight**

Time-of-flight-measuring electro-optical sensors are radar-type systems. However, unlike regular radar, which transmit a pulse of radio-frequency energy, these sensors normally use a modulated light beam. The distance to the target is extracted from the measured phase shift of the reflected light. Two problems associated with such sensors are: difficulty in measuring short distances (which requires a very high modulation frequency), and the need for a mechanical scanning/switching system to get additional information (such as orientation) [5, 25].

**Photothermal Effect**

The photothermal effect transducer uses a strong light beam directed toward the object’s surface. The distance to the object is extracted from measurements of the thermal wave generated by the light absorbed by the object. The detection scheme and signal processing are similar to those used in an AM sensor. Since the shape of the thermal wave generated at the surface is surface-texture independent, the photothermal sensor does not suffer from the surface robustness problem associated with AM sensors. However, the photothermal sensor is rather slow, and limited to highly absorbing surfaces [26].

**Capacitive Sensors**

Capacitive sensors generate and measure changes in an electric field caused by either a dielectric or conducting object in their proximity.

There are basically two types of capacitive proximity sensor. One type uses the principle of a parallel plate capacitor, the other uses the principle of fringing capacitances [8, 27, 28]. For the parallel plate type proximity sensor, the transducer forms one plate and the object measured forms the other plate. The structure of a parallel plate type proximity sensor and its typical response are shown in Figure 8.16 [8].

The parallel plate type proximity sensor is widely applied in industry. However, this type of a sensor has three major limitations: (1) the object being measured must be conductive; (2) the inverse gap-capacitance relationship is highly nonlinear and (3) the sensitivity drops significantly in the case of large gaps.

The second type of capacitive proximity sensor uses the principle of fringing capacitance [8]. The sensor has two "live" electrodes and the object being measured does not need to be part of the sensor system. The target object could be either conductive or nonconductive. However, the measurement of distances is affected by the type of object material. Therefore, separate calibrations must be carried out for different materials.
In [30], an innovative capacitive microsensor was presented. Using micromachining technology, the electrode thickness can be significantly reduced and the fringing effect increased when compared with other capacitive sensors. Consequently, this sensor yields a better sensitivity. An array of such transducers can be implemented to measure the distance and orientation of an object.

Proximity capacitive sensors have the following general advantages: (1) low energy consumption and (2) simple structure. The major disadvantages, however, are that they are influenced by external signals and a calibration-per-surface technique must be carried out, since their operation directly depends on the object’s material.

**Ultrasonic Sensors**

The basic principle underlying ultrasonic ranging sensors is the measurement of the time required for a sound wave to travel from the emitter to the object’s surface and return to the detector. By using several such emitters and detectors, one can obtain information about the distance and orientation of the surface.

In [19], a novel method is proposed to measure the orientation angles of an object’s surface using the phase differences of reflected echoes. Figure 8.17(a) shows the configuration of a planar sensor head with $n$ receivers, which are equally spaced and located on a circle of radius $R$ around the transmitter T. A linear relationship exists between the difference in lengths, $\delta$, of two reflecting paths for an adjacent pair of receivers and the object inclinations.

In Figure 8.17 (b), the relationship between the inclination of the target surface and $\delta$ is shown. It can be observed that the measuring range of the sensor can be enlarged with an increase in the number of receivers. In [29], it is also shown that the measuring range can be enlarged by reducing $R$. However, it was noted that measurements carried out with a small sensor are potentially less accurate.

Experimental results using a transducer with six ($R = 30$ mm) and eight ($R = 20$ mm) receivers were reported in [29]. With the six-receiver transducer, the measuring range of the orientation angles was $\pm 15^\circ$; while for the eight-receiver transducer, the maximum measuring range was $\pm 30^\circ$. With the six-receiver transducer, the orientation angle could be determined with an accuracy of $0.5^\circ$, in the measuring range of $\pm 15^\circ$, and $0.2^\circ$ when the range was restricted to $\pm 5^\circ$.

One of the major disadvantages of ultrasonic proximity sensors is that they are relatively large in size. However, implementing these sensors using micromachining could solve this problem. In [7], the generation and detection of ultrasound, for proximity sensing, was investigated using micromachined resonant membrane structures.

**Magnetic Sensors**

A magnetic-type sensor creates an alternating magnetic field, whose variation provides information about the object’s position.

The simplest magnetic sensors are reed microswitches or Hall effect switches. However, the most commonly used sensors in robotics are based on the electromagnetic inductive principle, emphasizing...
eddy current generation. The basic principle consists of creating a magnetic field using appropriate coils around a core with high permeability and an oscillator with a frequency excitation high enough to minimize the penetration of the field inside a conductive material. The main problems with magnetic sensors are their high size/range ratio and difficulty in providing reliable distance measurements in varying magnetic environments.

Acknowledgments

The authors would like to thank Martin Bonert for careful review and critique of this chapter. We also acknowledge the financial support of Natural Sciences and Engineering Research Council of Canada.

References

Distance

9.1 Basic Distinctions Between Range Measurement Techniques
Contact or Noncontact • Active or Passive • Time-of-Flight, Triangulation, or Field Based • Form of Energy • Coherent or Noncoherent Detection • Ranging, Range Imaging, or Position Tracking

9.2 Performance Limits of Ranging Systems
Range Accuracy • Depth of Field • Maximum Range • Lateral Resolution • Rate of Acquisition

9.3 Selected Examples of Ranging, Range Imaging, and Motion Tracking Systems
Laser-Based Active Triangulation Ranging and Range Imaging Sensors • Laser-Based Lidar Range Imaging Sensors • Position Tracking with Active Targets

9.4 A Sampling of Commercial Ranging, Range Imaging, and Motion Tracking Products

The tools and techniques of distance measurement are possibly one of humankind’s longest-running inventive pursuits. The scale shown in Figure 9.1 illustrates the enormous range of distances that science and engineering have an interest in measuring [1]. This chapter concerns itself with methods to measure a relatively small segment of this range — from centimeters to kilometers. Even within this limited segment, it would hardly be possible to list, much less describe, all of the distance measurement approaches that have been devised. Nevertheless, the small sampling of technologies that are covered here should be of help to a broad range of readers.

Distance measurement, at its most basic, is concerned with determining the length of a unidimensional line joining two points in three-dimensional space. Oftentimes, a collection of distance measurements is called for, so that the shape, the orientation, or the changes in position of an object can be resolved. Therefore, one must consider not only the measurement of distances, but also their spatial and temporal distributions. The terminology “ranging” will be used in reference to systems that perform single sensor-to-target measurements, “range-imaging” for systems that collect a dense map or grid of spatially distributed range measurements, and “position tracking” for systems that record the time history of distance measurement to one or several targets.

9.1 Basic Distinctions Between Range Measurement Techniques

Range measurement devices may be classified according to some basic distinctions. Generalizations can be made based on these broad classes, thereby facilitating the process of comparison and selection. The following subsections identify the fundamental bases for classification.

W. John Ballantyne
Spar Aerospace Ltd.
A common approach to measuring the distance to a point on an object is through a calibrated mechanical device that simultaneously connects the selected point to a reference position. Any tape measure, feeler gage, or dial gage may be considered an example of a simple contacting measurement device. Mechanical/electronic devices are available that allow a user to “digitize” discrete point positions on a three-dimensional surface. A gimbaled probe on the end of an \( X-Y-Z \) positioner or articulated arm is used to touch a specific point, and sensory information of the linear positions or joint articulations provide an accurate position estimate. Mechanical, contact-based methods are widely used in industry and can be extremely accurate. Some coordinate measuring machines (CMMs), for example, can achieve 1 \( \mu \)m repeatability.

The chief disadvantage of mechanical approaches is that they are usually restricted to distances and work volumes up to a few meters at maximum. This is due to fundamental scaling laws for mechanical structures. As the requirement to span larger distances increases, the mass and mechanical tolerancing requirements on the machine make designs impractical. Also, mechanical approaches are too slow to make multiple measurements in rapid succession, as is typically required in range imaging or position tracking, when the measurement involves large sets of spatially or temporally distributed data.

Noncontact techniques for performing ranging, range imaging, and position tracking are many and varied. Besl [2] reviews and compares several of these. In the centimeters to meters range, most do not approach the accuracy of CMMs; but at larger scales and for large quantities of data, they become a practical necessity. The rest of this chapter will review noncontact approaches only.

Active or Passive

Noncontact distance measurement may be divided into active or passive techniques. Active techniques involve some form of controlled energy (field or wave) linking a known reference location to the unknown target location. The source of energy is typically associated with the reference location, but in some cases the target, or both target and reference, may be active. Passive techniques rely on an externally occurring source of energy (e.g., sunlight or target/background temperature contrast) to make the target detectable.

An active approach can often simplify the distance measurement problem because it allows a greater degree of control over the many factors that can influence a measurement. For example, the choice of the form of energy and the power level of the active source can minimize the effect of uncontrolled variables like ambient illumination, weather, and atmospheric conditions. Furthermore, an active approach provides an opportunity to selectively localize the measurement spatially and temporally, eliminating possible ambiguity about which target point was measured at a given time. In contrast, passive

![Figure 9.1](image-url)
systems (e.g., stereo ranging) sometimes suffer from the so-called “correspondence problem,” which is concerned with how to determine whether a given target point, detected from two or more viewpoints, or over two or more instants, is in fact the same physical point.

A common use of active approaches is to make range measurements “through” materials that are mechanically or optically impenetrable. Examples include medical imaging, where various forms of directed energy (ultrasound, X-rays) are used to build surface or volumetric maps of organs and bones; sonar, which penetrates water better than light does; and ground-penetrating radar, which can detect objects and their depth beneath ground surface.

Passive approaches, while not offering the same range of control and flexibility of active approaches, offer certain advantages. First, because they emit no energy, their existence cannot be detected by another remote detection system. This feature is very important in military applications. Second, passive systems can often collect multiple point range measurements more quickly because they are not limited by the rate at which they can direct an energy source toward a target point, as is the case with most active systems. For example, a stereo ranging system effectively collects all resolvable target points in its field of view simultaneously, while a scanning laser, radar, or sonar ranging system collects each measured point sequentially. Finally, the absence of a directed energy source is a simplification that can significantly reduce the size, cost, and hardware complexity of a device (although at the expense of increased signal processing complexity).

**Time-of-Flight, Triangulation, or Field Based**

There are many different classes and instances of noncontact ranging devices, but with very few exceptions they are based on one of the following three basic principles:

1. Energy propagates at a known, finite, speed (e.g., the speed of light, the speed of sound in air)
2. Energy propagates in straight lines through a homogeneous medium
3. Energy fields change in a continuous, monotonically decreasing, and predictable manner with distance from their source

The techniques associated with these basic phenomena are referred to as time-of-flight, triangulation, and field based, respectively.

**Time-of-Flight**

*Time-of-flight* (TOF) systems may be of the “round-trip” (i.e., echo, reflection) type or the “one-way” (i.e., cooperative target, active target) type. Round-trip systems effectively measure the time taken for an emitted energy pattern to travel from a reference source to a partially reflective target and back again. Depending on whether radio frequencies, light frequencies, or sound energy is used, these devices go by names such as radar, lidar, and sonar. One-way systems transmit a signal at the reference end and receive it at the target end or vice versa. Some form of synchronizing reference must be available to both ends in order to establish the time of flight.

A characteristic of many TOF systems is that their range resolution capability is based solely on the shortest time interval they can resolve, and not the absolute range being measured. That is, whether an object is near or far, the error on the measurement is basically constant.

**Triangulation**

*Triangulation* techniques were known and practiced by the Ancients. Triangulation is based on the idea that if one knows the length of one side of a triangle and two of its angles, the length of the other sides can be calculated. The known side is the “baseline.” Lines of detection extend from either end of the baseline to the target point as shown in Figure 9.2. If the angles formed between these lines and the baseline can be determined, the distance is calculated as:

\[
R = \frac{b \sin \alpha_{left} \sin \alpha_{right}}{\sin(\alpha_{right} - \alpha_{left})}
\]  

(9.1)
Classical surveying is a passive range-finding technique based on the above formula. A surveyor uses a precision pointing instrument to sight a target from two positions separated by a known baseline. Reference [3] notes that the distance to a nearby star may be calculated by observing it through a pointing instrument at 6-month intervals and using the diameter of Earth's solar orbit as the baseline. Stereo ranging, which compares the disparity (parallax) between common features within images from two cameras, is another form of passive triangulation. It is of interest to note that human vision estimates distance using a variety of cues, but two of the most important — stereopsis and motion parallax — are fundamentally triangulation based [4].

Active triangulation techniques use a projected light source, often laser, to create one side of the triangle, and the viewing axis of an optical detection means to create the second side. The separation between the projector and detector is the baseline.

A fundamental issue for all triangulation-based approaches is that their ability to estimate range diminishes with the square of the range being measured. This may be contrasted with TOF approaches, which have essentially constant error over their operating range. Figure 9.3 illustrates how, conceptually, there is a “crossover” distance where TOF techniques become preferable to triangulation techniques.

Field-Based Approaches
Whereas TOF and active triangulation techniques employ the wave propagation phenomena of a particular energy form, field-based approaches make use of the spatially distributed nature of an energy form. The intensity of any energy field changes as a function of distance from its source. Moreover, fields often exhibit vector characteristics (i.e., directionality). Therefore, if the location of a field generator is known and the spatial characteristics of the field that it produces are predictable, remote field measurements contain information that may be used to infer distance from the source.
An interesting distinction between field-based approaches and wave-based approaches is that the former, although they employ energy fields, do not rely on the propagation and conversion (and concomitant losses) of energy. That is, they may employ stationary fields, like those generated by a magnet or static charge. Such fields encode position information by their very shape. Sound and light, although having a wave nature, can be exploited in the same manner as stationary fields because of their distance-dependent intensity.

Field-based techniques must confront some basic issues that limit their range of application. First, the characteristics of most practically exploitable fields are typically influenced by objects or materials in the vicinity, and it is not always possible to ensure that these influences will remain constant. Second, the variation of fields through space is highly nonlinear (typically inverse square or inverse cube), implying that the sensitivity of a measurement is strongly affected by proximity to the source. Notwithstanding these concerns, devices have been developed and are available that perform very well in the situations for which they are intended [7].

**Form of Energy**

As discussed above, all noncontact, active ranging devices employ some form of energy. This is true whether time-of-flight, triangulation, or field-based principles apply. The following subsections describe the various forms of energy employed and some generalizations about the effectiveness of each in various situations.

**Sound**

Ranging systems based on sound energy are usually of the pulsed-echo TOF type and employ carrier frequencies in the so-called "ultrasonic" (beyond audible) range of frequencies. Besides being inaudible...
(an obvious benefit), ultrasonic frequencies are more readily focused into directed beams and are practical to generate and detect using piezoelectric transducers. Ultrasonic signals propagate through air, but long-distance transmission is much more effective in liquids, like water, where higher density-to-viscosity ratios result in higher wave velocity and lower attenuation per unit distance. Ultrasonic ranging techniques (or SONAR, for SOund NAvigation and Ranging) were first developed for subsea applications, where sound is vastly superior to electromagnetic energy (including light) in terms of achievable under-water transmission distances [5]. Low-cost, portable sonar systems are widely used by sport fishermen as "fish finders" [6].

The frequencies typically used in sonic ranging applications are at a few tens of kilohertz to a few hundred kilohertz. A basic trade-off in the choice of ultrasonic frequency is that while high frequencies can be shaped into narrower beams, and therefore achieve higher lateral resolution, they tend to fade more quickly with distance. It may be noted that beam widths narrow enough for range imaging applications (less than 10°) are effective in a fluid medium, but attenuate too quickly to be practical in air. Interestingly, although sound energy attenuates more rapidly in air than in water, useful short-range signals can be generated in air with relatively low power levels because the much lower density of air requires smaller dynamic forces in the transducer for a given wave amplitude.

When comparing sound energy to electromagnetic energy for TOF-based techniques, one needs to remember that sound, unlike light, propagates at not only much lower speeds, but with considerably more speed variation, depending on the type and state of the carrying media. Therefore, factors like air humidity and pressure will affect the accuracy of a TOF ranging device. For underwater applications, salinity and depth influence the measurement. The lower speed of sound has a detrimental impact on the rate at which range samples can be collected. For example, a target 10 m away takes at least 60 ms to measure through an air medium. This may not seem like a long time to wait for a single sample, but it becomes an issue if the application involves multiple sampling, as in motion tracking or collision avoidance sensing.

**Stationary Magnetic Fields**

Stationary or pseudostationary (i.e., low frequency) magnetic fields are only used in field-based approaches. An advantage of such fields is that they are easily and cheaply produced by either a permanent magnet or electrical coil. Since stationary fields do not transmit energy, the targets cannot be passive — they must actively sense the properties of the field at their particular location. A variety of sensing technologies may be used to make measurements of the direction and intensity of a magnetic field, including flux gate, Hall effect, and magnetostrictive type magnetometers. A comprehensive list of such technologies is given in [7].

**Radio Frequencies**

Echo-type TOF ranging systems based on the band of the electromagnetic spectrum between approximately 1 m and 1 mm wavelength are known as RADAR (RAdio Detection And Ranging). Radio waves can be used for long-distance detection in a variety of atmospheric conditions. As in the case of sound waves, there are trade-offs to be addressed in the choice of frequency. Long waves tend to propagate better over long distances, but short waves can be focused into narrow beams capable of better lateral discrimination. An interesting application of short-range radar is ground-penetrating radar, which can be used to locate and image subsurface objects [8]. Here, the frequency vs. range trade-off is particularly acute because of the need to balance reasonable imaging capability (narrow beam) with good depth penetration (long wave).

An example of a TOF one-way (active receiver) system that uses radio frequencies is the global positioning system (GPS). The distance between a receiver on land is determined by each of several orbiting satellites equipped with a transmitter and a very precise Cesium clock for synchronization. A good description of GPS and its use in vehicle navigation is available in [9].
Light Frequencies

Beyond the radio portion of the electromagnetic spectrum are the infrared, visible, and ultraviolet frequencies. These frequencies can be produced by lasers and detected by solid-state photosensitive devices and are useful for both TOF and active triangulation ranging. Echo-type TOF techniques are known as LIDAR (Light Detection And Ranging), in keeping with the terminology introduced earlier.

While light frequencies attenuate more than radio frequencies through cloud and fog, they can have very narrow beam widths, allowing superior lateral resolution and target selectivity.

Coherent or Noncoherent Detection

Echo-type TOF devices, whether sonar, radar, or lidar, can be further classified according to whether the detection approach measures time-of-flight directly (noncoherent) or exploits an inherent periodicity in the emitted energy to ascertain the flight distance (coherent).

Noncoherent techniques face the problem of timing short intervals. This is not a serious challenge in the case of sound waves, where a meter round trip corresponds to 6 ms, but is somewhat more problematic for light and radio waves, where that distance equates to only 6 ns. Accuracy of noncoherent detection typically relies on the averaging of repeated measurements.

Coherent detection is achieved by combining a portion of the emitted signal with the reflected signal to produce a third signal indicating the amount of phase delay. The signals are continuous wave (CW) as opposed to pulsed. Coherent detection techniques are classified as amplitude modulated (AMCW) or frequency modulated (FMCW).

A basic issue with coherent detection techniques is the inability to distinguish between integral multiples of the basic modulation wavelength. Any coherent detection system must employ techniques to resolve the so-called “ambiguity interval.” Noncoherent techniques do not face this problem.

Ranging, Range Imaging, or Position Tracking

Ranging devices are typically pointed toward a target to produce a single range reading. A common example of simple ranging is the feedback sensor used in auto-focus cameras. There are many active ranging devices currently available based on TOF (i.e., radar, sonar, lidar) and active triangulation principles.

Range imaging devices use the same principles as ranging devices, except that they include some form of scanning that is employed to generate an array of spatially distributed range samples. Sometimes, the scanning action is accomplished by means intrinsic to the sensor (e.g., spinning and nodding mirrors, or phased-array antenna) so that the reference location remains fixed. In this case, the data are recorded in the polar form (range, elevation, azimuth) as shown in Figure 9.4. In other cases, the sensor might scan on only one axis internally while the second scan dimension is realized by moving the sensor location through some set pattern. It is not uncommon to record the “intensity” or return energy associated with a range sample as well. The intensity map may be presented as a “gray scale” image and, like a black and white photograph, often contains additional information useful in interpreting a scene. Range images can be used to produce three-dimensional graphic representations of scenes and objects. A common use of range imaging is aerial terrain mapping.

Position tracking devices are used to measure the change in an object’s position and orientation over time. Basic issues in position tracking are the acquiring of, and locking on to, specific target points. These issues can be avoided by employing active targets, and most systems available today are of this type.

9.2 Performance Limits of Ranging Systems

The performance characteristics of available ranging systems vary widely, as do the requirements of the applications for which they are designed. The following subsections review the most basic performance categories and the technical issues of performance limits.
Range Accuracy

As illustrated in Figure 9.3, TOF and active triangulation techniques differ fundamentally in their error vs. distance characteristics. Currently available systems based on active triangulation achieve better repeatability and accuracy in the less than 1 m range than do TOF systems, but are seldom used at distances of several meters. Hymarc Ltd. and Perceptron Inc. each offer laser triangulation systems with $3\sigma$ accuracy of 25 mm and 50 mm, respectively [10, 11].

In principle, TOF systems could achieve accuracy rivaling active triangulation, but the most promising detection technique — a variation of laser interferometry, which solves the ambiguity interval problem [12] — has yet to make its commercial debut.

Depth of Field

Depth of field refers to the interval of distance through which a stationary reference ranging system can measure without resorting to a change in configuration. Large depth of field is often an important characteristic in practical applications. For example, if the distance to the target is poorly known a priori, then a large depth of field is desirable.

Passive optical triangulation approaches like stereography and photogrammetry tend to have restricted depth of field because they rely on camera-type imaging, which is inherently limited by depth of focus. Timed-interval TOF systems have excellent depth of field because they do not rely strongly on optical imaging except to concentrate the collected return energy on the detector. Some active triangulation systems do rely on optical imaging of the projected laser spot, but the design employed by Hymarc Ltd. regains a large depth of field by tilting the detector array with respect to the lens plane [13].

FIGURE 9.4  Range images are typically an array of individual range values sampled while changing the pointing direction (e.g., azimuth and elevation angles) of a ranging device. A digital range image of the polar form shown can be readily transformed into rectangular coordinates if required.
Maximum Range

Any active ranging, range imaging, or position tracking system has a practical maximum distance that it can measure. This is because the controlled energy, whether propagated as a wave or established as a field, must spread before reaching the detector. The spreading inevitably increases with distance and all detectors, no matter what form of energy they measure, require a certain minimum amount to exceed their inherent “noise floor.”

The “classical radar range equation” is introduced in many texts on radar (e.g., [14]). Jelalian [15] points out that the equation is equally applicable to lidar, which, after all, just employs a higher frequency version of electromagnetic wave. In fact, the same idea applies to sonar and to active triangulation systems as well. The equation computes the power of the received signal as:

\[ P_r = P_t G_t / 4\pi R^2 \times \rho A / 4\pi R^2 \times \pi D^2 / 4 \times \eta_{\text{atm}} \eta_{\text{sys}} \]  

(9.2)

where

- \( P_r \) = power at the receiver
- \( P_t \) = power transmitted
- \( G_t \) = transmitter gain
- \( R \) = range to target
- \( \rho \) = reflectivity of target
- \( A \) = effective area of target
- \( D \) = diameter of collecting aperture
- \( \eta_{\text{atm}} \) = atmospheric transmission coefficient
- \( \eta_{\text{sys}} \) = system transmission coefficient

Equation 9.2 applies when the target area is smaller than the footprint of the incident beam, which is often the case for radar and sonar ranging. However, in the case of laser-based systems, the relatively narrow beam usually means that the laser spot is small compared to the target. For a transmitted beam that spreads with a solid angle \( \theta_t \), the illuminated patch area is:

\[ \sigma_{\text{spot}} = \pi R^2 \theta_t^2 \]  

(9.3)

The definition of transmitter gain is based on the notion of the solid angle beam width as compared to an omnidirectional transmitter

\[ G_t = 4\pi / \theta_t^2 \]  

(9.4)

One can substitute for Equation 9.4 for \( G_t \) and Equation 9.3 for the variable \( s \) in Equation 9.2 to produce the range equation for a small spot size.

\[ P_r = P_t / R^2 \times \rho / 4 \times \pi D^2 / 4 \times \eta_{\text{atm}} \eta_{\text{sys}} \]  

(9.5)

The importance of this equation is primarily in the \( 1/R^2 \) dependence. Any ranging system that works by bouncing energy off a diffuse reflective target encounters severe signal attenuation with increasing distance. Given a detector with a fixed noise floor, the only ways to improve maximum range are to increase the transmitted power or the collecting area. In practice, there are design constraints that limit both of these measures. For example, laser power must sometimes be limited for eye-safety considerations, and increased collecting area can imply a proportional increase in sensor packaging volume.
Lateral Resolution

In range imaging applications, it is generally desirable to use the narrowest possible beam width to provide good lateral discrimination of target surface features. Lasers, because of their short wavelength, can be optically collimated to produce much narrower beam widths than are possible with radio sources. However, even lasers cannot produce arbitrarily narrow beams. The interested reader is referred to [13] for a discussion of Gaussian beam propagation and optimal focusing. There are basically two ways to project laser light. The beam can be “focused down” to produce the smallest possible spot at a particular point inside the measurement range, in which case the beam will diverge as the distance from that point increases; or the beam can be focused at infinity or some very distant point so as to minimize the divergence through the entire measurement range. The former approach provides higher lateral resolution at the focus distance, but by implication restricts the practical depth of field. The latter compromises spot size for increased depth of field.

Rate of Acquisition

The rate at which a ranging sensor can acquire range samples is important when the target object is changing shape or position, or when the required sample density of a range image is very high. There are several potential factors that can limit sample acquisition rate: the amount of time required by the detector to integrate the weak return signal to a sufficient level (integration time); the time constant of any filtering or averaging that must be performed to realize an acceptably “clean” signal (smoothing time); the rate at which samples can be transferred through the signal processing stages (transfer time); and the velocity limits of mechanical scanning apparatus (scanning bandwidth). Acquisition rates vary widely: from tens of hertz for acoustic ranging devices to tens of kilohertz for some laser-based systems. It is worth noting that, in general, there is a trade-off between rate of acquisition, accuracy, and maximum range. Some systems permit control over basic parameters so that this trade-off may be optimized for a particular application. The reader should be aware that data sheets may not be clear as to whether stated performance figures for these three specifications are valid in combination.

9.3 Selected Examples of Ranging, Range Imaging, and Motion Tracking Systems

The following sections review selected examples of some specific ranging, range imaging, and position tracking sensor systems. The list is by no means exhaustive, but offers a reasonable sampling of available technologies.

Laser-Based Active Triangulation Ranging and Range Imaging Sensors

Active Triangulation Basics

Figure 9.5 illustrates the basic active triangulation geometry. In this so-called “pinhole camera” model, practical aspects like lenses for projection and detection and mirrors for scanning are eliminated for clarity. It can be shown by means of similar triangles that the range is inversely proportional to the deflection of the imaged spot.

\[ R = \frac{bf}{u} \]  

\( R \) = distance to object  
\( b \) = baseline distance  
\( f \) = lens to detector distance  
\( u \) = detected spot position in the image plane

© 1999 by CRC Press LLC
The sensitivity of the range measurement, or the incremental change in \( u \) with \( R \), is

\[
\frac{du}{dR} = \frac{bf}{R^2}
\]

(9.7)

The significance of Equation 9.7 is that range estimating performance is expected to fall as distance increases. Improvements in accuracy are realized by increasing the baseline or the lens to detector distance (i.e., the focal length).

**Synchronized Scanning Principle**

Lateral scanning of an active triangulation ranging sensor is accomplished by an elegant and effective technique developed at Canada’s National Research Council and now marketed by Hymarc Ltd. under the name “HYSCAN” [10]. A two-sided oscillating mirror simultaneously steers the outgoing beam on one face and directs the collected light to the spot-imaging optics on the opposite face. By synchronously scanning both the beam and the axis of the detection system, rather than the beam only, as conventionally practiced, significant performance improvements are made. *Figure 9.6* is a schematic illustration of the approach. Note also that the detector plane is tilted with respect to the lens plane. This feature increases the depth of focus so that the ranging performance is maximized over the measuring volume. The Hyscan sensor produces a single-axis sweep, or so-called “line scan.” Full-dimension range images are acquired by translating the sensor over a target surface with a controlled motion pattern.

**Light Plane Principle**

Perceptron Inc. offers a similar line-scan system under the name “TriCam” [11]. In this case, the laser is not swept. Instead, the beam is transformed to a focused plane by means of cylindrical lenses. A two-dimensional detector is used to generate range profiles through the analysis of a deformation of the laser line as the sensor is translated over the object surface.

---

*Figure 9.5* A simple pinhole camera model illustrates the basic active triangulation principle. As the distance \( R \) to the target surface changes, the spot position \( u \) on the detector changes, maintaining similarity between the large triangle outside the camera and the small triangle inside. There is an inverse relationship between \( R \) and \( u \).
Perceptron Inc. also offers a scanning lidar under the name "LASAR" that can produce high-resolution range images through a large measurement volume. The device uses a near-infrared laser that is projected through a collimating telescope to form a spot on the first surface encountered. The spot is swept over

Laser-Based Lidar Range Imaging Sensors

AM Lidar (Phase-Based Detection)

The Hymarc laser triangulation line scanner uses the synchronized scanning principle. Both sides of an oscillating mirror are used to sweep both the projected beam and the axis of detection over the target. The detector array is tilted to the lens plane to maximize the depth of focus.

FIGURE 9.6 The Hymarc laser triangulation line scanner uses the synchronized scanning principle. Both sides of an oscillating mirror are used to sweep both the projected beam and the axis of detection over the target. The detector array is tilted to the lens plane to maximize the depth of focus.
a programmable field of view in a raster pattern by means of a spinning polygon mirror and an oscillating "nodding mirror." Some of the backscattered light is collected and directed by means of an adjacent facet of the polygon mirror. The projected laser light is amplitude modulated at a reference frequency by controlling the power to the laser diode source. The return signal, although orders of magnitude weaker than the outgoing signal, is phase-compared to determine the range for a particular azimuth and elevation. The intensity of the return energy is also recorded. The Perceptron sales literature claims a maximum measurement volume of $60^\circ \times 72^\circ \times 40$ m, a range image grid resolution of $1024 \times 2048$ pixels and a maximum acquisition rate of 360,000 pixels/s. A schematic diagram of the LASAR™ system is shown in Figure 9.7. Details of the Perceptron technical approach may be found in [16].

Resonating Lidar (Frequency-Based Detection)
Acuity Research Inc. has developed a laser-based TOF ranging sensor based on a simple but effective idea. The detector controls the laser output such that the absence of a signal drives the laser on and the presence of a signal turns it off. The finite transit time of the light bounce turns this arrangement into a two-state resonator, with the period being proportional to the target distance. Rather than measuring the period, which is extremely short and difficult to time, the frequency is measured using conventional counting techniques for as many cycles as necessary to yield the required accuracy. The AccuRange 4000, as it is named, is also available in a $360^\circ$ line-scanning arrangement suitable for robotic vehicle navigation applications [17]. Details of the technical approach may be found in [18].

FIGURE 9.7 The Perceptron AM Lidar system described in U.S. patent 5,006,721 uses a rotating polygon mirror for synchronized scanning. A "nodding mirror" is also added to sweep at a slower rate in the orthogonal direction, producing a raster scan pattern. Range measurement is determined by comparing the phase of the outgoing and returning AM laser signal.
Position Tracking with Active Targets

Active target approaches are not convenient in some applications, but they are an excellent way to track the changing positions of several target points simultaneously. Active targets are a way of getting around the "correspondence problem" mentioned earlier. The two systems introduced here are interesting to compare. One employs light energy and triangulation; the other uses a magnetic field-based approach. They are both used for real-time tracking and recording of human kinetics, robotics, and other moving objects.

Active Target Triangulation

The "OPTOTRAK" system offered by Northern Digital Ltd. [19] uses infrared light emitting diodes (LEDs) as targets. The LEDs are multiplexed so that only one at a time can be seen by the camera system, avoiding the correspondence problem. The unique form of stereo ranging is based on three line detectors with lenses that transform the point source LED illumination into a focused line. The simplified triangulating geometry is shown in Figure 9.8. It may be shown from this geometry that the target position \((x_p, y_p, z_p)\) can be determined from the detector outputs \(u_{\text{right}}, u_{\text{left}}\), and \(v\) as follows:

\[
x_p = b \left( \frac{u_{\text{right}} + u_{\text{left}}}{2} \right) - \frac{b}{f} \left( u_{\text{right}} - u_{\text{left}} \right)
\]

\[
y_p = \frac{b v}{f} \left( u_{\text{right}} - u_{\text{left}} \right)
\]

\[
z_p = \frac{b f}{f} \left( u_{\text{right}} - u_{\text{left}} \right)
\]

where \(f\) and \(b\) are the lens-to-detector distance and the baseline separation respectively. In practice, the image space to object space mapping is much more complicated than Equations 9.8 to 9.10, and involves a camera model with more than 60 parameters that are determined through a calibration process.

OPTOTRACK offers high sampling rate, large measurement volume, and high accuracy compared to many other position tracking systems.

Magnetic Position Tracking

A position/orientation tracking sensor based on a three-axis magnetic dipole transmitter and a three-axis magnetic loop detector has been developed by Polhemus Inc. [20]. The transmitted fields are alternating current for ease of detection (i.e., transformer coupled) and time-multiplexed so that the field due to each axis can be distinguished from the others. Distance between transmitter and detector is determined by exploiting the \(1/R^3\) relationship between field strength and distance from the source. Orientation of the detector is determined by exploiting the directional sensitivity of magnetic fields and the direction sensitivity of loop detectors.

An issue with respect to the use of ac fields is the distortions in field shape that occur if metal objects are present, and the consequent effect on sensor accuracy. These distortions result from eddy currents in the conducting metal. Ascension Technology Corp. has developed a variation on the Polhemus sensor based on dc magnetic fields. The switching transient due to time-multiplexing does produce an eddy current effect, but it is allowed to die out before measurement is made. Details of the dc technique are available in [21].

An important difference between optical and magnetic tracking technologies is that the former require an unbroken line of sight to the targets while the latter do not. This gives magnetic trackers an advantage in some applications. On the other hand, the \(1/R^3\) field distribution characteristic of magnetic tracking...
implies an extreme sensitivity loss with distance, whereas optical triangulation has a more benign $1/R$ characteristic. This, to some extent, explains why the volume of measurement and accuracy of optical triangulation systems is generally much better than for magnetic systems.

### 9.4 A Sampling of Commercial Ranging, Range Imaging, and Motion Tracking Products

Table 9.1 contains information collected from vendor literature. Be advised when comparing specifications that test conditions, standards, and interpretations can vary significantly. The specifications, therefore, should serve only as a rough guide.
References


---

**TABLE 9.1  Ranging, Range Imaging, and Position Tracking Products and Vendors**

<table>
<thead>
<tr>
<th>Class</th>
<th>Trade Name</th>
<th>Principle</th>
<th>Features</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranging (contact)</td>
<td>MicroScribe-3DX</td>
<td>Instrumented arm</td>
<td>50 in. spherical work volume, 0.3 mm accuracy</td>
<td>Immersion Corp., (408) 467-1900, <a href="mailto:info@immerse.com">info@immerse.com</a></td>
</tr>
<tr>
<td>Ranging (noncontact)</td>
<td>LASERVISION</td>
<td>TOF, laser</td>
<td>50 m range, 4.9 mm accuracy @ 15 m, integrated electronic level</td>
<td>ZIRCON Corp., (408) 866-8600</td>
</tr>
<tr>
<td>Range-Imaging</td>
<td>HYSCAN</td>
<td>Active triangulation laser</td>
<td>40 mm depth of field, 70 mm swath, 0.025 mm accuracy, 10,000 points/s</td>
<td>Hymarc Ltd., (613) 727-1584, <a href="mailto:info@hymarc.com">info@hymarc.com</a></td>
</tr>
<tr>
<td>Range-Imaging</td>
<td>TriCam</td>
<td>Active triangulation laser</td>
<td>120 mm depth of field, 60 mm swath, 0.05 mm accuracy</td>
<td>Perceptron Inc., (810) 478-7710, <a href="mailto:inquiry@perceptron.com">inquiry@perceptron.com</a></td>
</tr>
<tr>
<td>Range-Imaging</td>
<td>ALTM 1020</td>
<td>TOF laser time-interval</td>
<td>330-1000 m range, 15 cm accuracy, 20° swath</td>
<td>Optech Inc., (416) 661-5904</td>
</tr>
<tr>
<td>Range-Imaging</td>
<td>RangeCam 7000</td>
<td>Laser or strobe triangulation</td>
<td>uses standard CCD camera and light plane projector</td>
<td>Range Vision Inc.</td>
</tr>
<tr>
<td>Range-Imaging</td>
<td>LASAR</td>
<td>TOF, AM Lidar</td>
<td>2-40 m range, 60 x 70° max field of view, 360,000 samples/s</td>
<td>Perceptron Inc., (810) 478-7710</td>
</tr>
<tr>
<td>Position Tracking</td>
<td>OPTOTRAK</td>
<td>Active target triangulation</td>
<td>up to 255 targets, submillimeter accuracy, 3000 3 DoF samples/s</td>
<td>Northern Digital Inc., (519) 884-5142</td>
</tr>
<tr>
<td>Position Tracking</td>
<td>Flock of Birds</td>
<td>Magnetic field based</td>
<td>up to 30 position/orientation targets, approx. 10 mm accuracy, 144 6-DoF samples/s</td>
<td>Ascension Technology Corp. (802) 860-6440</td>
</tr>
</tbody>
</table>

© 1999 by CRC Press LLC
10.1 Altitude Measurement

Dimitris E. Manolakis

Accurate monitoring of aircraft cruising height is required in order to reduce vertical separation to a minimum standard. Interest here focuses on the measurement of the distance between aircraft level and the sea surface level. This distance can be estimated onboard via barometric altimeters or it can be measured — either onboard or in ground stations — via electronic radio wave systems. The indication of the first equipment is referred to as pressure altitude, or simply altitude, whereas that of the second category is referred to as geometric height or simply height.

The altitude information at air traffic control (ATC) centers is based on pressure altitude measurement that the aircraft transponder system sends after it receives an appropriate interrogation — known as mode C interrogation — transmitted by a secondary surveillance radar. Actually, the altitude information is an atmospheric pressure measurement transformed to altitude indication through a formula expressing the pressure-altitude relationship. When a flight level is cleared for an aircraft, it actually means that the pilot must keep flying on an isobaric surface. However, the altimetry system may present systematic errors (biases) that are different for each airplane, and that significantly affect safety. Thus, the altimetry...
system performance as well as the aircraft height keeping performance must be monitored by an independent radar or satellite system.

Radar or satellite systems determine the position of an object through algorithms that are fed with range, or range difference, or range sum, or range and bearing measurements, and they estimate the object position vector employing appropriate techniques such as triangulation or trilateration. The primary radar measurements are contaminated by two kinds of errors: random and systematic errors. The effect of random errors can be reduced by the use of appropriate noise rejection filters, such as a Kalman filter. The second kind of error is usually removed by calibrating the instrument if this is possible; otherwise, suitable algorithms must be invented to anticipate it.

The estimation may be derived in ground stations or onboard the aircraft according to where the data acquisition and processing is performed. In the latter case, the vertical position estimation has to be downlinked to the appropriate ATC center. Also, the estimation may be performed off-line or on-line. Ground-based methods or systems are the Navigation Accuracy Measurement System (NAMS), the height estimation method with a single air traffic control radar, the method with one Secondary Surveillance Radar (SSR) and one omnidirectional radar, the Dual synchronized Autonomous Height Monitoring System (DAMS) and methods that use multiple SSRs and estimate the height by quadrilateration, or by the use of pseudorange measurements, or by the use of range difference measurements. On-board height measurement methods derive their estimates by trilateration using the Distance Measuring Equipment (DME) or the Global Positioning System (GPS) signals.

**Ground-Based Height Estimation**

The radars used to derive the original measurements are either primary or secondary surveillance radars. A primary radar sends a signal and scans for the arrival of its reflection. The range to the object reflecting the signal is derived from the time elapsed between transmission and reception. With secondary radar, the radar sends an interrogation to aircraft — to all aircraft or to a selected one — and the appropriate aircraft sends a reply via its transponder. The range to the aircraft is computed from the time elapsed between the signal transmission and the signal arrival, taking into account the nominal delay time of the transponder. Most of the methods estimating the aircraft height make use of the SSR equipment because it is cost effective, the transponder reply signal is stronger than that reflected to a primary radar, and the system can operate more reliably in dense traffic areas.

Any systematic errors in the primary radar and in the ground equipment of the SSR can be corrected by calibration. However, the problem encountered with SSR is that it involves the transponder delay time in the range measurement process. Thus, any systematic error in the transponder delay time causes range bias errors that are different for each aircraft and thus suitable methods must be used to anticipate for it in the subsequent measurement data processing.

**Navigation Accuracy Measurement System**

Nagaoka has proposed an off-line height estimation system [1–3]. It is composed of a primary marine radar located under an airlane and measures range $R$ and depression angle $\beta$. Figure 10.1 shows the geometry of the system. The antenna rotates about a vertical axis and scans the area above it with rate equal to 1 rotation per 3 s. The principle exploited to derive the height estimate is that the range varies as the aircraft passes through the data acquisition area. The rate of change of range is mainly a function of the flight height $z$ and secondarily of depression angle. It is easily derived from Figure 10.1, that the relation between the above quantities and the position $x$ along the x axis at time $t$ is:

$$R(t) = \sqrt{x^2(t) + R^2(t)} = \sqrt{x^2(t) + \frac{z^2}{\cos^2(\beta(t))}}$$  (10.1)

Let $x_0$ denote the position of the aircraft at time $t_0$. Assuming the aircraft flies in straight and level flight, this means that the velocity $V_x$, the depression angle, and the height $h$ remain constant during the data...
collection period. The above quantities and the range measurement at time $t_i$ are related through Equation 10.2.

$$R_i = \sqrt{x_0 + V_x(t_i - t_0)^2 + \frac{z^2}{\cos^2\beta}} = f(t_i, \beta, \mathbf{q}) = f_i(q)$$  \hspace{1cm} (10.2)$$

where $\mathbf{q}$ is the unknown quantities vector, $\mathbf{q} = [x_0, V_x, z]^T$.

Measurements of $R_i$ and $\beta$ are collected at times $t_0$, $t_1$, ..., $t_n$, and their set is briefly expressed as a vector function of the unknown quantities with the following matrix equation:

$$\mathbf{R} = f(\mathbf{q})$$  \hspace{1cm} (10.3)$$

where

$$\mathbf{R} = \begin{bmatrix} R_0 \\ R_1 \\ \vdots \\ R_n \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{bmatrix}$$

Equation 10.3 is nonlinear. Thus, a nonlinear least square method, such as the Gauss-Newton iterative method, must be used to estimate the unknown vector. Let $\hat{\mathbf{q}}_k$ be the estimate at the $k$th iteration. Then, the next estimate is:

$$\hat{\mathbf{q}}_{k+1} = \hat{\mathbf{q}}_k + (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T(\mathbf{R} - f(\hat{\mathbf{q}}_k))$$  \hspace{1cm} (10.4)$$

where $\mathbf{F}$ is the partial derivatives (Jacobian) matrix; that is:
Assuming typical aircraft velocities and flight levels, and taking into account the antenna fan beam and rotation rate, it is determined that the observation time is approximately 1 min and the number of measurements is approximately 20. The standard deviation (SD) of the height estimation errors is \( \sigma_z = 10 \text{ m} \) when the SD of the measurement errors of range and depression angle are \( \sigma_R = 10 \text{ m} \) and \( \sigma_{\beta} = 0.08^\circ \), respectively.

The system has been developed and operates at the Electronic Navigation Institute of Tokyo, Japan. To estimate and anticipate for systematic errors in the depression angle and range measurements, it has been proposed to perform 40 experimental flights above the radar range. The basic assumption of the NAMS system (i.e., that the aircraft flies at constant height) cannot be validated by the NAMS itself. Thus, an SSR is required to confirm that the flight is indeed performed at the same level via the mode C interrogations [3].

### Height Estimation with a Single Air-Traffic Control Radar

Nagaoka at the Electronic Navigation Institute of Tokyo investigated the possibility to derive a height estimation with a very simple and inexpensive method [4]. The idea is to use a single radar normally used for air traffic control. The radar antenna rotates around a horizontal axis once every 4 s and measures range \( R \) and bearing \( \beta \), as it is shown in Figure 10.2.

Let \((x,y,z)\) be the 3-D position of the aircraft. The bearing angle is defined as:

\[
\beta = \cos^{-1}\left(\frac{y}{\sqrt{x^2 + y^2}}\right)
\]  

**Figure 10.2** Geometric configuration of the height measurement system with a single air-traffic control radar. \( OC = x, OD = y, AB = z, \beta = \angle BOD. \)
Thus, the relation between the measurements and the unknown quantities is:

\[ R = \left( z^2 + x^2 + y_i^2 \right)^{1/2} = \left( z^2 + \frac{y_i^2}{\cos^2 \beta_i} \right)^{1/2} \]  \hspace{1cm} (10.7)

The measurement principle is the same as that of the NAMS system, i.e., that during the time the aircraft passes through the data acquisition area, the range varies as a function of both the horizontal position and the height. When straight and level flight is assumed, the unknown quantities are the height \( z \), the y-axis initial position \( y_0 \), and the velocity \( V_y \). The expression relating the measurements \( R_i \), \( \beta_i \) and the unknown quantities vector \( \mathbf{q} = [y_0, V_y, z]^T \) at time \( t_i \) is:

\[ R_i = \left( z^2 + \left( y_0 + V_y (t_i - t_0) \right)^2 \right)^{1/2} = f(t_i, \beta_i, \mathbf{q}) = f_i(\mathbf{q}) \]  \hspace{1cm} (10.8)

The measurements are collected as the aircraft flies close to the radar at a distance of 30 km. Approximately, the data acquisition area is 15 km, the observation time is 1 min and the total number of scans is 15. The total set of measurements is expressed as a nonlinear matrix function in the form of Equation 10.3; that is, \( \mathbf{R} = \mathbf{f}(\mathbf{q}) \) where \( \mathbf{R} \) denotes the measurement vector. The unknown vector is estimated using a nonlinear least squares method like that of the NAMS system expressed by Equation 10.4.

The system performance is evaluated with Monte Carlo simulations. The best accuracy achieved is \( \sigma_z = 100 \text{ m} \) under the assumption that the SD errors of range and bearing measurements are \( \sigma_R = 100 \text{ m} \) and \( \sigma_\beta = 0.06^\circ \). The method can use either primary or secondary radar. However, in the latter case, the method has to include systematic delays by the transponder.

**Height Estimation with one SSR and One Omni Radar (SSROR)**

Manolakis et al. investigated the possibility to estimate the aircraft flight height with a system that consists of one standard SSR plus a secondary radar equipped with an omnidirectional antenna located far away from the SSR; hence, this system could be referred to as SSROR [5]. Figure 10.3 shows a typical geometric configuration of the system. The SSR measures bearing \( \theta \) and slant range \( R_s \), which is transformed to horizontal range \( R \) with correcting look-up tables. The omni radar measures range \( R_a \). Let \((x,y,z)\), \((x_o, y_o, z_o)\) and \((0,0,0)\) be the coordinates of aircraft, omni radar, and SSR, respectively. Then, from Figure 10.3 it follows that the theoretical height function is:
However, the height measurement obtained through the above formula will be biased because of the transponder delay systematic deviation from its nominal value that affects the range measurements. Since this bias cannot be removed by calibration, a suitable algorithm has to be applied to anticipate for it. The approach proposed is to augment the unknown vector by incorporating the bias term. Let $b$ denote the bias in range measurements. The biased measurements $R_b$ and $R_{ab}$ of $R$ and $R_a$, respectively, are expressed as:

\[ R_b = R + b \cos \phi \]  \hspace{1cm} (10.12)

\[ R_{ab} = R_a + b \]  \hspace{1cm} (10.13)

where $\phi$ is the elevation angle. Let $s$ denote the squared height:

\[ s = (z - z_s)^2 \]  \hspace{1cm} (10.14)

and $s_b$ denote the corresponding quantity derived from (biased) range measurements. Then, after some manipulations of the above equations, the following relation is obtained:

\[ s_b = s + b a \]  \hspace{1cm} (10.15)

\[ a = 2 R_a - \cos \phi \left( R - x_s \sin \theta - y_s \cos \theta \right) \]  \hspace{1cm} (10.16)

The term $a$ is the bias multiplying factor determined by the relative geometry of the system. Equation 10.15 is a linear relation between measurement $s_b$ and the unknown quantities $s$ and $b$. Figure 10.4 shows that the effect of the bias varies as the aircraft passes through the surveillance area. Its form is mainly determined by the flight height. Consequently, it is possible to estimate both the height and the bias by collecting data during the period the aircraft remains in the surveillance area. The measurement equation at time $t_i$ is:

\[ s_{bi} = s_i + b a_i + e_i \]  \hspace{1cm} (10.17)

where $e_i$ represents the effect of the random measurement errors; hence, it could be referred to as the equation error. Assuming level flight, $s_i = s$, and the set of collected data is expressed with the following linear matrix equation:

\[ s_b = A q + e \]  \hspace{1cm} (10.18)
The best estimate $\hat{q}$ of $q$, minimizing the weighted sum of the squared errors $e^T W e$, is:

$$\hat{q} = \left( A^T W A \right)^{-1} A^T W s_b$$  \hspace{1cm} (10.19)

where $W$ is the weighting matrix defined as the inverse of the equation error covariance matrix:

$$W^{-1} = E \left\{ e e^T \right\}$$  \hspace{1cm} (10.20)

where the notation $E\{\}$ stands for the expected value operation. Thus, the estimate of the geometric height is:

$$\hat{z} = \sqrt{\hat{q}} = \sqrt{\hat{s}}$$  \hspace{1cm} (10.21)

In the case of nonlevel flights, an augmented unknown vector is used that comprises bias $b$, initial height $z_0$, and vertical velocity $V_z$ [6]. In this case, however, the measurement equation is nonlinear in terms of the unknown quantities; hence, a nonlinear least squares iterative algorithm must be employed.

The system performance is a function of: (1) the range and bearing measurement standard deviation errors, $\sigma_R$ and $\sigma_\theta$, respectively; (2) the aircraft velocity, which affects the number of scans; (3) the omni radar scan rate; and (4) geometric terms such as the flight level and the distance between the radars. By assuming that $\sigma_R = 70 \text{ m}$, $\sigma_\theta = 0.08^\circ$, the SD of the height estimation error will be between 50 m and 100 m when the values of the other factors lie in reasonable ranges.

FIGURE 10.4 Range bias multiplier in squared height measurements as a function of the aircraft distance from the omni radar. Three different flight heights are examined. The omni radar is at position (150 km, 0, 0).
Dual Synchronized Autonomous Monitoring System (DAMS)

The research has been conducted in the National Aerospace Laboratory NLR, the Netherlands [7]. The system is composed of two primary marine radars. Each radar has a rotating antenna. The antennas scan, with different directions, the same volume — that is an area $3 \text{ km} \times 10 \text{ km}$ above the location site. The primary surveillance area is $3 \text{ km} \times 3 \text{ km}$. The two antennas (A and B) are separated by 2.5 m, are mechanically synchronized, and scan once every 2.5 s. Their data extractors produce measurements of slant ranges and elevation angles, $(R_A, \varphi_A)$ and $(R_B, \varphi_B)$, respectively, for each radar. The tracking software derives estimation of aircraft position and trajectory in off-line mode. At each scan, a combination of four measurements $(R_A, \varphi_A, R_B, \varphi_B)$ is available for use in the estimation of the unknown vector $q = [x_0, y_0, z_0, V_x, V_y, V_z]^T$. A local Cartesian frame is used to perform the calculations. Finally, 16 equations are available to be solved for the 6 unknowns with a weighted least squares method. The weight of each measurement is the measured amplitude of the radar pulse. The maths of the tracker are not presented in [7]. However, one approach could be the following. If a coordinate system is defined such that radar A is at the origin (0, 0, 0) and radar B is at $(x_B, 0, 0)$, then the measurements obtained at scan time $t_i$ can be expressed as:

\[
\begin{align*}
R_{Ai} &= \sqrt{x_i^2 + y_i^2 + z_i^2} = \sqrt{(x_0 + V_x(t_i - t_0))^2 + (y_0 + V_y(t_i - t_0))^2 + (z_0 + V_z(t_i - t_0))^2} \\
\varphi_{Ai} &= \tan^{-1}\left(\frac{z_i}{\sqrt{x_i^2 + y_i^2}}\right) = \tan^{-1}\left(\frac{z_0 + V_z(t - t_0)}{\sqrt{(x_0 + V_x(t - t_0))^2 + (y_0 + V_y(t - t_0))^2}}\right) \\
R_{Bi} &= \sqrt{(x_i - x_B)^2 + y_i^2 + z_i^2} = \sqrt{(x_0 + V_x(t_i - t_0) - x_B)^2 + (y_0 + V_y(t_i - t_0))^2 + (z_0 + V_z(t_i - t_0))^2} \\
\varphi_{Bi} &= \tan^{-1}\left(\frac{z_i}{\sqrt{(x_i - x_B)^2 + y_i^2}}\right) = \tan^{-1}\left(\frac{z_0 + V_z(t - t_0)}{\sqrt{(x_0 + V_x(t - t_0) - x_B)^2 + (y_0 + V_y(t - t_0))^2}}\right)
\end{align*}
\]

Notice that $x_B$ has a small value (2.5 m) compared to the magnitude of the other quantities; consequently, it can be neglected. The above set of equations at four different times yields 16 equations that can be solved with a nonlinear least squares method such as the Gauss–Newton iterative method presented in Equation 10.4.

The SD of the height estimation error will be less than 15 m in the primary surveillance area, and 30 m at the edges of the area when the SD of range and elevation angle measurements are smaller than 10 m and 0.1°, respectively.

Height Measurement by Quadrilateration

Rice proposed a system consisting of four synchronized receiving SSR stations $S_i$, $i = 0, 1, 2, 3$ that use SSR transmissions from the aircraft transponder and estimate the height by quadrilateration [8]. Figure 10.5 shows a typical configuration of systems composed of $N$ SSRs. One of them is an active station, which means that this station has both an interrogator and a receiver. Let $(x, y, z)$ and $(x_0, y_0, z_0)$ denote the Cartesian coordinates of aircraft and station $S_i$, respectively. Also let $R_i$ denote the range from station $S_i$ to aircraft, $c$ denote the velocity of the light, and $T_i$ denote the time of signal transmission from aircraft transponder. The stations measure the time of arrival (TOA) $T_i$, $i = 0, 1, 2, 3$, of the aircraft transponder signal at each site. The following relations hold:

\[
\begin{align*}
R_i &= \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} = \sqrt{(x_0 + V_x(t_i - t_0) - x_i)^2 + (y_0 + V_y(t_i - t_0))^2 + (z_0 + V_z(t_i - t_0))^2} \\
\varphi_i &= \tan^{-1}\left(\frac{z_i}{\sqrt{(x_i - x_0)^2 + y_i^2}}\right) = \tan^{-1}\left(\frac{z_0 + V_z(t_i - t_0)}{\sqrt{(x_0 + V_x(t_i - t_0) - x_0)^2 + (y_0 + V_y(t_i - t_0))^2}}\right)
\end{align*}
\]
The above system of eight equations can be solved for the unknown quantities. The unknown quantities used by Rice are \((R_0, R_1, R_2, R_3, x, y, z, T_s)^T\). However, an equivalent approach is to substitute for \(R_i\) in Equation 10.23, which becomes:

\[
T_i = T_s + \frac{R_i}{c}, \quad i = 0, 1, 2, 3
\]  

Equation 10.24

The above system of eight equations can be solved for the unknown quantities. The unknown quantities used by Rice are \((R_0, R_1, R_2, R_3, x, y, z, T_s)^T\). However, an equivalent approach is to substitute for \(R_i\) in Equation 10.23, which becomes:

\[
T_i = T_s + \frac{1}{c} \left[ (x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2 \right] = f_i(q) \quad i = 0, 1, 2, 3
\]  

Equation 10.25

where \(q = [x, y, z, T_s]^T\) is the unknown vector. Thus, there are four nonlinear equations to be solved for \(q\). One suitable method, for example, is the Newton–Raphson method, which iteratively approximates the solution via the following formula:

\[
q_{k+1} = q_k + \frac{F(q_k)}{\dot{f}(q_k)}(T - f(q_k))
\]  

Equation 10.26

where \(T = [T_0, T_1, T_2, T_3]^T\) is the measurement vector and \(F\) is the Jacobian matrix:

\[
F = \frac{\partial f}{\partial q} = \begin{bmatrix}
\frac{\partial f_0}{\partial x} & \frac{\partial f_0}{\partial y} & \frac{\partial f_0}{\partial z} & \frac{\partial f_0}{\partial T_s} \\
\frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} & \frac{\partial f_1}{\partial T_s} \\
\frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} & \frac{\partial f_2}{\partial T_s} \\
\frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} & \frac{\partial f_3}{\partial z} & \frac{\partial f_3}{\partial T_s}
\end{bmatrix}
\]  

Equation 10.27

Notice that the time of interrogation transmission, as well as the transponder nominal delay time, are not involved in the measurements. The measured quantities are only the TOAs at the station sites. Thus, the height estimate is not affected by any transponder bias.

FIGURE 10.5  Typical configuration of the height estimation systems that are based on N SSR stations. One of the stations is active, i.e., it both transmits the interrogations and receives the replies, whereas the other stations are receivers only.
The theoretical and experimental research has been conducted at the GEC Marconi Research Center, U.K. The optimum station arrangement is to locate the three of them equispaced on a circle and the fourth in the middle. The typical circle radius is 35 km. The same magnitude holds for the measurement range. The method may be implemented in on-line or off-line mode. In the first case, there must be transmitters at the stations to transmit their measurements of TOA to the height monitoring center. The Vertical Dilution Of Precision (VDOP) is a performance index defined by the ratio:

\[ VDOP = \frac{\sigma_z}{\sigma_{rte}} \]  

where \( \sigma_{rte} \) is the SD of the relative timing errors. The VDOP expresses the effect of the relative geometry to system performance. The VDOP of this system achieves a typical value of 3.

**Height Estimation with SSRs and Pseudoranges**

This approach has been investigated by Nagaoka at Electronic Navigation Institute of Tokyo [9]. The system consists of \( N \) SSR receiving stations; see Figure 10.5. One of them, say \( S_0 \), must be active to transmit interrogations to aircraft. The time of interrogation transmission, \( T_t \), and the times of signal arrival at the receiving stations \( T_i, i = 0, 1, \ldots, N - 1 \), are measured. Thus, \( N \) pseudorange measurements \( r_i \) are obtained where \( r_i = c \left( T_i - T_t \right) \). Let \( T_{D_i} \) denote the transponder delay and \( D \) denote the distance corresponding to this delay, \( D = c T_{D_i} \). Then, for each pseudorange measurement \( r_i \), the following relation holds:

\[
\begin{align*}
    r_i &= D + R_i + R_{e_i} = D + q \left( x - x_i \right)^2 + \left( y - y_i \right)^2 + \left( z - z_i \right)^2 + \left( x - x_0 \right)^2 + \left( y - y_0 \right)^2 + \left( z - z_0 \right)^2 \\
    &= f(q) \quad i = 0, 1, \ldots, N-1
\end{align*}
\]

where \( q = [x, y, z, D]^T \) is the unknown vector. The set of \( N \) measurements \( \rho = [\rho_0, \rho_1, \ldots, \rho_{N-1}]^T, N \geq 4 \), and the unknown vector are related through Equation 10.30.

\[
\mathbf{r} = f(q) 
\]  

(10.30)

The unknown vector \( q \) can be obtained from the solution of Equation 10.30 with a nonlinear weighted least squares method. Thus, the best estimate of \( q \) is iteratively calculated as:

\[
\hat{q}_{k+1} = \hat{q}_k + \left( \mathbf{F}^T \mathbf{F} \right)^{-1} \mathbf{F}^T \left( \mathbf{r} - f(\hat{q}_k) \right) 
\]

(10.31)

where \( \mathbf{F} \) is the Jacobian matrix

\[
\mathbf{F} = \frac{\partial f}{\partial q} = \begin{bmatrix}
    \frac{\partial f_0}{\partial x} & \frac{\partial f_0}{\partial y} & \frac{\partial f_0}{\partial z} & \frac{\partial f_0}{\partial D} \\
    \frac{\partial f_0}{\partial x} & \frac{\partial f_0}{\partial y} & \frac{\partial f_0}{\partial z} & \frac{\partial f_0}{\partial D} \\
    \cdot & \cdot & \cdot & \cdot \\
    \frac{\partial f_{N-1}}{\partial x} & \frac{\partial f_{N-1}}{\partial y} & \frac{\partial f_{N-1}}{\partial z} & \frac{\partial f_{N-1}}{\partial D}
\end{bmatrix}
\]

(10.32)

The estimate of \( q \) is free of the transponder delay systematic error because the estimation is based not on the nominal delay, but on the actual delay time, which is one of the parameters to be estimated, whereas the rest of the parameters are the aircraft 3-D position coordinates \( x, y, z \).
The station arrangement proposed by Nagaoka, when there are four stations, is an equilateral triangle formed by the three stations, whereas the fourth station is located in the center. The VDOP, defined as the ratio \( \sigma_z / \sigma_R \) (where \( \sigma_R \) is the observation error) has a typical value of 4 when the aircraft is above the center at a height equal to the baseline radius. The VDOP increases as the aircraft flies higher and longer and as the baseline radius becomes smaller.

**Height Measurement with SSRs and Range Differences**

This approach has been proposed by Manolakis and Lefas [10, 11]. The system consists of \( N – 1 \) receiving SSR stations \( S_i, i = 1, N – 1 \), and one station, say \( S_0 \), which is both receiver and interrogator, see Figure 10.5. The stations receive the reply and the time difference of arrival (TDOA) between a reference station, say \( S_0 \), and station \( S_i \) is measured. A set of \( N – 1 \) TDOA or equivalently range difference (RD) measurements is collected at each time the transponder sends a reply signal. The height estimation derived from this set of measurements is not affected by any transponder delay systematic error since this error is inherently subtracted from the measurements used. This system could be referred to as RD height monitoring unit (RDHMU). The systems that derive the position fix based on this kind of measurement are known as TDOA or RD or hyperbolic systems.

Let \( \tau_i \) denote the TDOA between stations \( S_i \) and \( S_0 \), and \( d_i \) denote the corresponding RD measurement, \( d_i = c \tau_i \). The following relation holds:

\[
\begin{align*}
    d_i &= R_i - R_0 = \sqrt{(x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2} - \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2} \\
    &= f(q) \quad i = 1, 2, \ldots, N-1
\end{align*}
\]  

(10.33)

where \( q = [x, y, z]^T \) is the unknown aircraft position vector. The vector of RD measurements \( d = [d_1, d_2, \ldots, d_{N-1}] \) is expressed as:

\[
d = f(q)
\]  

(10.34)

A commonly employed method to solve for \( q \) in this nonlinear equation is the Taylor series method or equivalently the Gauss-Newton iterative method. The best estimate of \( q \) is iteratively approximated as:

\[
\hat{q}_{k+1} = \hat{q}_k + (F^T F)^{-1} F^T (d - f(\hat{q}_k))
\]  

(10.35)

where \( F \) is the Jacobian matrix:

\[
F = \frac{\partial f}{\partial q} = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_{N-1}}{\partial x} & \frac{\partial f_{N-1}}{\partial y} & \frac{\partial f_{N-1}}{\partial z} \end{bmatrix}
\]  

(10.36)

In the case of four stations the best arrangement is an equilateral triangle with the fourth station in the center. The SD of height estimation error \( \sigma_z \) will be 15 m when the baseline radius is 6 km, the flying height is 9 km, and \( \sigma_{\text{TDOA}} \), 10 ns.

Work on proof of principles and system development of a HMU based on the concept of TDOA measurement of SSR signals has been conducted by Roke Manor Research Ltd., U.K. [12].
Onboard Derived Height Estimation

Height Measurement with Distance Measuring Equipment (DME)

This approach for deriving the geometric height onboard the aircraft using DME equipment was first reported by Rekkas et al. [13], whereas more efficient and general techniques have been proposed by Manolakis [14, 15]. Using the DME interrogation equipment, the distance from three DME ground stations is measured onboard (see Figure 10.6). The three stations are located under an airway. The height is then computed from the range measurement vector \( \mathbf{R} = [R_1, R_2, R_3]^T \) by trilateration. An exact and efficient solution of the nonlinear measurement equation was derived in [15]. Specifically, the height is computed from the closed form:

\[
z = g(\mathbf{R}) = \frac{-b(\mathbf{R}) + \sqrt{d(\mathbf{R})}}{2a}
\]  

where \( b(\mathbf{R}) \) and \( d(\mathbf{R}) \) are the following simple polynomial-type functions:

\[
b(\mathbf{R}) = b_0 + b_1 R_1^2 + b_2 R_2^2 + b_3 R_3^2
\]

\[
d(\mathbf{R}) = d_{00} + d_{01} R_1^2 + d_{02} R_2^2 + d_{03} R_3^2 + d_{11} R_1^4 + d_{12} R_2^4 + d_{13} R_3^4 + d_{22} R_2^4 + d_{23} R_3^4 + d_{33} R_3^4
\]

The coefficients \( a, b_i, d_i \) are analytically defined in the Appendix of [15]. An important aspect of these coefficients is that they are completely defined by the ground stations’ coordinates \( (x_i, y_i, z_i) \), which are fixed. Thus, the coefficients are calculated only once at the moment the aircraft enters the data acquisition area. Then, every time a new set of range measurements is available, the height is computed from the above equations using the range measurements and the stored coefficients. Define the ratio \( \sigma_z/\sigma_R \) as the VDOP of this technique, where \( \sigma_z \) is the SD of the ranging error. The VDOP is 1 in the case where the stations form an equilateral triangle inscribed in a circle with 10 km radius and the aircraft is above the triangle center at a height of 8 km.

FIGURE 10.6 Configuration of the on-board height estimation system that utilizes the distance measurements derived from the DME equipment.
Estimation of Vertical Position with the Global Positioning System (GPS)

The research and development of the GPS has been coordinated by the U.S. Department of Defense. Another similar system is the Global Navigation Satellite System (GLONASS) developed by the former Soviet Union. The GPS is a satellite system providing users with accurate timing and ranging information. The system is available with reduced accuracy to civilian users. Many companies, mainly from the U.S., produce GPS receivers. Let \((x, y, z)\) and \((x_i, y_i, z_i)\) be the coordinates of the user and satellite \(s_i\). The GPS receiver of the user derives the pseudorange measurement \(D_i\), and the corresponding measurement equation is:

\[
D_i = R + cT_b = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} + b = f_i(q)
\]

where \(T_b\) is the user clock bias, and \(q = [x, y, z, b]^T\) is the unknown vector that incorporates the bias term \(b\). Thus, in order to estimate the 3-D position of the aircraft, four pseudorange measurements are required at least; consequently, four satellites must be visible from the receiver. The set of \(N\) pseudorange measurements \(D = [D_1, D_2, ..., D_N]^T\) defines the following matrix measurement equation:

\[
D = f(q)
\]

which is solved for \(q\) with the Gauss–Newton least squares iterative method, that is:

\[
\hat{q}_{k+1} = \hat{q}_k + (A^T W A)^{-1} A^T W (D - f(q_k))
\]

where \(A\) is the partial derivatives matrix:

\[
A = \frac{\partial f}{\partial q} = \begin{bmatrix}
\frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} & \frac{\partial f_1}{\partial b} \\
\frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} & \frac{\partial f_2}{\partial b} \\
\vdots & \vdots & \vdots & \vdots \\
\frac{\partial f_N}{\partial x} & \frac{\partial f_N}{\partial y} & \frac{\partial f_N}{\partial z} & \frac{\partial f_N}{\partial b}
\end{bmatrix}
\]

\[
= \begin{bmatrix}
a_{x1} & a_{y1} & a_{z1} & 1 \\
a_{x2} & a_{y2} & a_{z2} & 1 \\
\vdots & \vdots & \vdots & \vdots \\
a_{xN} & a_{yN} & a_{zN} & 1
\end{bmatrix}
\]

The elements \(a_{xi}, a_{yi}, a_{zi}\) of the partial derivatives matrix \(A\) are the direction cosines from the receiver to the satellite \(s_i\). The weighting matrix is the inverse of the covariance matrix of the pseudorange measurement errors, \(W^{-1} = \text{E}(\delta D \delta D^T)\). The weighting is generally used to take into account the possible different performances of each satellite, although usually the same performance is assumed for all satellites; that is, \(W = I\). The VDOP, defined as \(\sigma / \sigma_n\), depends on the geometry which varies continuously, even in the case of a fixed receiver, because the satellites are not geostationary but move in such orbits as to complete a rotation in 12 h. The world mean value of VDOP is about 2 \[16\]. Typical VDOP values range from 1.5 to 7, depending on the area of the receiver and on the time of day. The ranging error for the precision positioning service (available only to U.S. military users) has been specified to be less than 6 m (SD). For the standard positioning service, normally available to civilian users, the specified ranging error is double (12 m, SD), whereas it will be about 40 m when selective availability is activated by the Department of Defense. The corresponding measured ranging errors found to be smaller than the specified ones. Namely, for the three operating conditions mentioned, the corresponding values for ranging errors were found to be 2.3 m, 6 m, and 20 m, respectively \[17\]. The multiplication of the ranging SD error by the VDOP yields the standard deviation of the height estimation error.
To anticipate for the error intentionally induced by the Department of Defense, the Differential GPS (DGPS) method has been developed. A station located at a precisely known position receives the satellite signals, computes its own position on the basis of pseudorange measurements, compares this position with the known position in order to estimate the included error in satellite signals, and finally transmits the appropriate corrections to the receivers in its neighborhood. The achieved accuracy is in the order of a few meters.

**Special Topics**

The performance analysis of the various parameter estimation systems is usually restricted to the variance analysis, and the estimation error is usually assumed to have zero mean value. However, it is proven that in all of the above systems, the estimation error does not have zero mean value due to the nonlinearity of the measurement equation. Another important aspect is that there are cases where it is not possible to obtain a solution due to the relative geometry of aircraft vs. stations that leads to large errors. In these cases, the successive iterations applied to solve the system of nonlinear equations may not converge. Also, even after convergence, the solutions need not necessarily be "the best" or the "correct ones."

**Inherent Bias**

The nonlinearity of the systems, joined with the measurement random errors, causes inherently biased estimations although the measured quantities are unbiased. For example, take the case of height estimation with DME measurements. The function \( g(R) \) in Equation 10.37, which determines that the height \( z \) is nonlinear. In addition, the range measurements will be contaminated by additive zero mean value random errors. Let \( \textbf{R}_m \) denote the noisy measurement vector and \( z_m \) denote the height measurement derived from \( g(\textbf{R}_m) \), i.e., from the measurement function when it is fed with noisy measurements. Extending \( g(\textbf{R}_m) \) in a Taylor series around the actual values of ranges \( R_1, R_2, R_3 \) up to second-order terms and taking the expected values, it is derived that the expected value of \( z_m \) will not be the actual value \( z \), but it will differ by an amount \( b_z \), which is called inherent bias. Specifically, for the DME case, the inherent bias is evaluated as:

\[
b_z = \mathbb{E}[z_m] - z = \frac{1}{2} \left( \frac{\partial^2 g}{\partial R_1^2} \sigma_{R_1}^2 + \frac{\partial^2 g}{\partial R_2^2} \sigma_{R_2}^2 + \frac{\partial^2 g}{\partial R_3^2} \sigma_{R_3}^2 \right)
\]  

(10.44)

Figure 10.7 shows the inherent bias generated in the DME system. The inherent bias becomes larger as the magnitude of the measurement errors and the system nonlinearity becomes stronger. This bias error is inherently generated in all position estimation algorithms and must be taken into account when precise position estimation is required. Biased height estimates have also been reported in [18] for the SSOR system, in [10] for the RDHMU, and in [3] for the NAMS system.

**Existence and Uniqueness of Position Fix**

In some systems, there are singular cases for which it is not possible to achieve a position fix. This fact has been reported by Abel and Chafee for the GPS system in [19], where it is shown that for some satellites/aircraft relative geometries, it is not possible to solve the relevant equation or there is more than one solution. Also, for the RDHMU, it has been shown by Manolakis and Lefas that there are some station arrangements for which it is not possible to derive height estimation when the aircraft is at specific areas [11]. For example, in the case of four stations, when the quadrilateral defined by the stations is inscribed in a circle, it is not possible to estimate the height when the aircraft is above the center. Also, when the quadrilateral is a rectangle or symmetric trapezoid, it is not possible to derive a position fix when the aircraft is above the line that passes from the middle of the parallel sides. From a mathematical point of view, this singularity is expressed by the singularity of the Jacobian matrix; consequently, this matrix cannot be inverted as is required by the relevant position estimation algorithm. The algorithm in this case diverges from the actual height and finally collapses. Notice that height estimation is achieved everywhere except at this singular point. However, when the aircraft is close to the singular region,
although a position fix is achieved, it is not actually reliable since it is affected by large errors; for example, the VDOP could be larger than 600 in regions close to the singular region.

Table 10.1 presents the institutes and companies that either investigate and develop prototype height monitoring units or provide relevant systems in the market.

**Table 10.1  Height Monitoring Systems and Companies/Institutes that Develop and Provide Them**

<table>
<thead>
<tr>
<th>System</th>
<th>Company/Institute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation Accuracy Measurement</td>
<td>Electronic Navigation Research Institute</td>
</tr>
<tr>
<td>System (NAMS)</td>
<td>Ministry of Transport</td>
</tr>
<tr>
<td></td>
<td>6-38-1 Shinkawa, Mitaka</td>
</tr>
<tr>
<td></td>
<td>Tokyo, 181, Japan</td>
</tr>
<tr>
<td></td>
<td>Tel: +81 422 413171</td>
</tr>
<tr>
<td></td>
<td>Fax: 81-422-413176</td>
</tr>
<tr>
<td>DAMS height monitoring unit</td>
<td>National Aerospace Laboratory NLR</td>
</tr>
<tr>
<td></td>
<td>Anthony Fokkerweg 2</td>
</tr>
<tr>
<td></td>
<td>1059 CM Amsterdam</td>
</tr>
<tr>
<td></td>
<td>The Netherlands</td>
</tr>
<tr>
<td></td>
<td>Tel: +31 (0)20 511 3113</td>
</tr>
<tr>
<td></td>
<td>Fax: +31 (0)20 511 3210</td>
</tr>
<tr>
<td>SSR and quadrilateration technique</td>
<td>GEC-Marconi Electronics Ltd.</td>
</tr>
<tr>
<td></td>
<td>Marconi Research Laboratories</td>
</tr>
<tr>
<td></td>
<td>West Hanningfield Road</td>
</tr>
<tr>
<td></td>
<td>Great Baddow, Chelmsford</td>
</tr>
<tr>
<td></td>
<td>Essex, England</td>
</tr>
</tbody>
</table>

© 1999 by CRC Press LLC
References

3. S. Nagaoka, Possibility of detecting a non-level-flight aircraft by the navigation accuracy measurement system (NAMS), ICAO, Review of the General Concept of Separation Panel, 7th Meeting, Montreal, RGCSP-WP/180, 30/10/90.


10.2 Attitude Measurement

Mark A. Stedham, Partha P. Banerjee, Seiji Nishfuji, and Shogo Tanaka

In many practical situations, it is important to determine and measure the attitude of a particular vehicle, such as a ship, an airplane, a piece of mechanical equipment such as a crane lifter, or a spacecraft. For this reason, many attitude sensors have been developed with advanced computer and semiconductor technologies. This section first introduces the various attitude sensors with an explanation of their operating principles and then presents several methodologies for attitude measurement and determination, including ships and crane lifters, aircraft, and spacecraft applications.

Attitude Sensors for Ships, Aircraft, and Crane Lifters

There are many types of gyroscopes that, corresponding to the physical measurement mechanisms used, may be classified as two-axes freedom gyro and single-axis freedom gyro using precession, vibratory gyro using Coriolis' force, and optic gyro using Sagnac's effect. Among them, the two-axes freedom gyro has the longest history. It consists of a high-speed rotating rotor around a spin axis supported by two orthogonal axes. This type of gyro is generally classified as either a free gyro, a vertical gyro (VG), or a directional gyro (DG).

The single-axis freedom gyro has only one output axis in addition to the spin axis. Depending on the specifications (in which) the gyro is designed, there are two types of gyros, the rate gyro and the rate integrating gyro. Related to these rotating-type gyros is another type of gyro known as the electrostatic gyro, which makes use of a high-speed rotating sphere in a vacuum cavity. Because of its resistance-free property, the electrostatic gyro has the highest accuracy among existing gyros. There are also rotorless gyros. The first one is a vibratory gyro that uses Coriolis' force as the measurement principle. The second type is an optical one. Among optical gyros, there are two types: the ring laser gyro and the fiber optic gyro. Both rely on the Sagnac effect in their measurement mechanisms. The performance of gyros is evaluated by their drift rates, and the performance of various gyros is shown in Table 10.2, for reference, with their primary usages.

Recently, with the development of computer technology, many types of three-axes gyros have been developed that can measure not only the tilt angles but also the angular velocities and the accelerations along the three axes by combining several gyros and accelerometers. Accelerometers are often coupled with gyros to provide flight and ship navigation systems as well as attitude sensors for dynamic objects such as crane lifters. Examples include the attitude and heading reference system (AHRS), inertial navigation system (INS), inertial measurement unit (IMU), and gyro compass (GC), as well as the VGs and DGs discussed above [1].
The principle of a servo-type accelerometer is explained below (see Figure 10.8). As soon as the shift of the beam caused by the acceleration $\alpha$ is detected by the deflection pickup, the current $i$ is generated by the servo-amplifier, which produces a torque to keep the beam at the principle axis of the sensor. Since the torque and the current that generates the torque are proportional to $\alpha$, the acceleration can be measured using the current. The measurement process forms a closed-loop system, so that the sensor is not only robust to disturbances, but also achieves a high measurement accuracy (see Table 10.3).

Similarly, an inclinometer is another inertial sensor that measures tilt angle to provide attitude information (see Figure 10.9). The principle of servo-type inclinometers is the same as that of the servo-type accelerometer, except that the beam in the accelerometer is replaced by a pendulum suspended from the supporting point in the sensor. When the sensor is placed on the inclined static surface of tilt angle $\beta$, the pendulum takes the angle $\beta$ against the principle axis of the sensor, assuming the sensor has no force other than gravity acting on it. The sensor can, however, generate a torque $T_c = mg \sin \beta = mg \beta$ to keep the pendulum at the principle axis, then the tilt angle $\beta$ can be accurately measured using the torque (and consequently the current producing the torque), where $m$ and $l$ are the pendulum mass and length of the pendulum to its mass center, respectively. One must note, however, that such a sensor is essentially designed to measure the tilt angles of static inclined surfaces. Thus, when applied to dynamic inclined surfaces, the accelerations will affect the torque, making the sensor unreliable. An intelligent attitude

<table>
<thead>
<tr>
<th>Type of the gyro</th>
<th>Degrees of freedom</th>
<th>Quantities to be detected</th>
<th>Accuracy ($^\circ$ h$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free gyro</td>
<td>2</td>
<td>Angle</td>
<td>1</td>
</tr>
<tr>
<td>Vertical gyro</td>
<td>2</td>
<td>Declination from horizontal plane</td>
<td>1</td>
</tr>
<tr>
<td>Directional gyro</td>
<td>2</td>
<td>Shift from reference direction</td>
<td>1</td>
</tr>
<tr>
<td>Rate gyro</td>
<td>1</td>
<td>Angular velocity</td>
<td>10</td>
</tr>
<tr>
<td>Rate integrating gyro</td>
<td>1</td>
<td>Angle</td>
<td>0.001–1</td>
</tr>
<tr>
<td>Ring laser gyro</td>
<td>1</td>
<td>Angular velocity</td>
<td>0.005</td>
</tr>
<tr>
<td>Fiber optic gyro</td>
<td>1</td>
<td>Angular velocity</td>
<td>0.01</td>
</tr>
<tr>
<td>Electrostatic gyro</td>
<td>2</td>
<td>Angle</td>
<td>0.00001–0.01</td>
</tr>
</tbody>
</table>

TABLE 10.2  Performance of Different Types of Gyros

FIGURE 10.8  Servo-type accelerometer.
A sensing system that overcomes such difficulty will be introduced later. Although application is limited to static inclined surfaces with minute tilt angles, a dielectric-type inclinometer employing electrodes and a bubble kept in an electrolyte can achieve high accuracies on the order of $10^{-4}$.

Attitude Sensors for Spacecraft Applications

Attitude measurement for spacecraft usually requires two or more sensors for detecting the reference sources needed to satisfy attitude requirements. The choice of which sensors to employ is primarily influenced by the direction the spacecraft is usually pointing as well as the accuracy requirements for attitude determination [2]. Table 10.4 summarizes some performance parameters for these sensors as well as typical manufacturers.

Inertial measurement units generally consist of gyroscopes coupled with accelerometers, which together measure both rotational and translational motion. These IMUs may be either gimbal mounted (movement about a gimbal point, independent of the spacecraft) or a strapdown system (rigidly mounted to the spacecraft body), where expansive software is used to convert sensor outputs into reference frame measurements. IMUs tend to suffer gyro drift and other bias errors and, when used for spacecraft attitude measurements, are often used with one or more of the sensors discussed below.

Sun sensors detect the visible light from the sun, measuring the angle between the sun's radiation and the detector's photocell. The sun is a commonly chosen attitude reference source since it is by far the

![Dielectric-type inclinometer (front view)](image)

**FIGURE 10.9** Dielectric-type inclinometer (front view).
visually brightest object in the sky, having a total radiation per unit area of 1353 W m\(^{-2}\) at Earth distances [3]. Also, it is generally accepted as a valid point source for most attitude applications, having an angular radius of 0.25° at Earth distances. Increased measurement accuracy can be obtained by determining its centroid. Even though sun sensors are quite accurate (0.01° to 3.0°), they do require clear fields of view, and sometimes suffer periods of eclipse from both the Earth and the moon [4]. Also, sensitive equipment (such as imaging devices) must be protected from the powerful radiation of direct sunlight. When the sun is available, the angle between it and the sensor’s primary axis is referred to as the sun angle.

For spacecraft in near-Earth orbits, the Earth is the second brightest object in the sky and covers as much as 40% of the sky. Earth horizon sensors detect the interface between the Earth’s edge (or limb) and the space background. Horizon sensors can detect either of the Earth’s visible limb (albedo sensor), infrared limb, or air glow. The infrared limb is the edge between the warm Earth and the cold space background. The air glow is a region of the atmosphere around the Earth that is visible to the spacecraft when it is on the night side of the Earth. Accuracies for horizon sensors are in the 0.1° to 1.0° range.

Increased accuracy requires Earth oblate spheroid modeling [4]. Some problems associated with albedo detection include the distortion effects of the Earth’s atmosphere, falsely identifying the day/night terminator crossing as the true Earth limb, and the considerable variability of the Earth’s albedo in the visible spectrum (varies from land, sea, ice).

Most sensors used to detect the Earth’s horizon are scanning sensors with narrow fields of view that measure the time between horizon crossings. In general, two horizon crossings occur per sensor scan period: one crossing when the sensor scans from the space background onto the Earth, followed by a second crossing when the sensor scans from the Earth back to space. The combination of horizon crossing times, scan rate, and spacecraft altitude allows for the computation of the Earth’s apparent angular radius.

If the spacecraft horizon sensor is pointing exactly nadir, then the apparent angular radius as measured by the sensor will agree with the physical radius given by the above relation for \( \rho \). However, if the horizon sensor is pointed away from nadir, the horizon crossing times will be smaller than when pointing exactly nadir. This results in an apparent angular radius that is smaller than the physical radius by an amount proportional to the angle between the sensor axis and the nadir vector. This angle is referred to as the nadir angle.

**Star sensors** are used when extreme accuracy requirements are necessary. This high degree of sensor accuracy (0.003° to 0.01°) can be attributed mainly to the point source nature and precise fixed location of stars in space. Star sensors may be categorized as either star trackers or star mappers. A star tracker utilizes a wide field of view in order to search for a given star of specific brightness. A star mapper is similar to a tracker, except that it scans over many stars, recording their relative positions and angular separations. By comparing the recorded data with that from a *star catalog* (database), exact spacecraft
orientation can be obtained. The angle between the star line-of-sight and the sensor’s primary axis is referred to as the *star angle*.

The accuracy of star sensors is obtained with higher costs, however. Star sensors are generally heavier and consume more power than other types of attitude sensors. In addition, star sensors are quite sensitive to stray light sources such as sunlight reflected from the spacecraft or the Earth and sunlight scattered from dust particles and jet exhausts [4]. Most rely on optical shielding to reduce the effects of stray light.

Magnetic sensors (called *magnetometers*) measure both the magnitude and direction of the Earth’s magnetic field. The difference in orientation between the measured field and the true field translates into attitude determination. Magnetometer accuracies (0.5° to 3.0°) are usually less than the other sensor types because of the uncertainty in the Earth’s true field, which tends to change or shift over time. In addition, the Earth’s magnetic field decreases with increasing altitude, and magnetometers are generally limited to altitudes of about 6000 km. For this reason, magnetometers are often used with one of the other sensor types already discussed for improved measurement accuracy [2].

**Automatic On-Line Attitude Measurement for Ships and Crane Lifters**

For on-line attitude measurement for ships and crane lifters, the first thing that comes to mind is to use gyros. However, because they often suffer from drifts, accurate attitude measurements might not be achieved using the gyros. Accordingly, one uses attitude on-line measurement systems that do not utilize gyros but servo-type accelerometers and inclinometers. The philosophy of the measurement systems introduced here is to make the best use of the system dynamics of the object and the sensors and to apply Kalman filters or adaptive filters to achieve high measurement accuracy.

**Attitude Measurement for Ships**

On-line accurate measurement of a ship’s attitude is extremely important in exact search of the seabed patterns with sonars [5, 6]. It is also required by high-performance ships like hovercrafts from the viewpoint of suppressing swings by the waves. The measurement of a ship’s attitude can usually be reduced to that of the heaving, rolling, and pitching of the ship. For such a measurement, a heave sensor has been used, whose output is given by double integration of the output of an accelerometer vertically directed with a gyroscope. However, since the initial values of heaving displacement and its velocity are unknown, the output will contain a bias that increases with time, and the accuracy of the sensor deteriorates considerably. From this viewpoint, one introduces a strapdown-type on-line measurement system that adequately processes the outputs of the two servo-type inclinometers and one accelerometer mounted on the ship [7].

**Location of Sensors and Outputs**

The two servo-type inclinometers and one servo-type accelerometer are located on the deck (at the point A) of vertical distance $L$ from O, the intersection of rolling and pitching axes (see Figure 10.10). The two inclinometers are set in such a way that the rolling and the pitching angles are measured respectively. The accelerometer is set upward to the deck to obtain the information on the heaving. Because inclinometers were originally developed for the measurement of the tilt angles of static inclined surfaces, the rigid pendulum inside the sensor is considerably affected by the ship’s acceleration other than the gravitational one. Applying Lagrange’s equations of motion [8, 9] to rigid pendulums and calculating the torques to keep their deflections from the principal axes almost zero yields the sensor outputs [7]:

\[
z_1(t) = 0(t) - \frac{L}{g} \dot{\theta}(t) + v_1(t) \quad (10.45)
\]

\[
z_2(t) = \dot{p}(t) - \frac{L}{g} \ddot{\theta}(t) + v_2(t) \quad (10.46)
\]
where $z_1(t)$, $z_2(t)$, $\theta(t)$, $p(t)$, and $g$ denote, respectively, the outputs of the two inclinometers, the rolling and the pitching angles, and the gravitational acceleration ($v_1(t)$ and $v_2(t)$: noises of the outputs, including the approximation errors in deriving the outputs).

On the other hand, the accelerometer output is expressed as:

$$z_a(t) = (g + \alpha(t)) \cos \theta(t) \cos p(t) + v_3(t)$$  \hspace{1cm} (10.47)

where $\alpha(t)$ and $v_3(t)$ represent, respectively, the heaving acceleration and the accelerometer noise.

### Dynamics of Attitude Signals

It is well known that each of the heaving, rolling, and pitching in inshore seas has two dominant waves in a short interval. That is, a sinusoidal wave of long periodic length (in the range of 6 s to 10 s) and a sinusoidal wave of short periodic length (in the range of 2 s to 3 s) \[10–12\]. Thus, one model each of the signals in a short interval by a composite wave of the two dominant sinusoidal waves. For the heaving (in a short interval), the displacement is modeled by:

$$x(t) = a_1 \sin(\omega_1 t + \phi_1) + a_2 \sin(\omega_2 t + \phi_2)$$  \hspace{1cm} (10.48)

with the parameters $[a_1]$, $[\phi_1]$, and $[\omega_1]$ unknown. From the 4th-order differential equation satisfied by the $x(t)$, we obtain the linear dynamic equation \[7\]:

$$\dot{x}(t) = Ax(t), A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\omega_1^2 & \omega_2^2 & 0 & -\left(\omega_1^2 + \omega_2^2\right) \end{bmatrix}$$  \hspace{1cm} (10.49)

where $x(t) = (x_1, x_2, x_3, x_4)^T$ ($x_n = d^{n-1}x/dt^{n-1}$ ($n = 1, \ldots, 4$)). On the other hand, the rolling and pitching angles can be modeled by:

$$\dot{x}(t) = a_1 \sin(\omega_1 t + \phi_1) + a_2 \sin(\omega_2 t + \phi_2) + b$$  \hspace{1cm} (10.50)

because there are usually some biases associated with them. From the 5th-order differential equation which Equation 10.50 satisfies, we get the similar state variable representation of the model as...
In practice, the heaving, rolling, and pitching signals have many nondominant sinusoidal waves in addition to the dominant ones. Therefore, Equation 10.49 is modified by introducing a white Gaussian noise $w(t)$ with zero mean and adequate variance $\sigma^2$ as follows:

$$\dot{x}(t) = Ax(t) + \Gamma w(t)$$  \hspace{1cm} (10.51)

where $\Gamma = (0,1,0,0)^T$ for the heaving and $\Gamma = (0,1,0,0,0)^T$ for the rolling and pitching. The higher the order of the models, the better the measurement accuracy will be. If we consider the on-line measurement of the signals, Equation 10.51 will be sufficient.

**On-Line Attitude Measurement**

The observation Equations 10.45 and 10.46 are expressed using their own state vector $x(t)$. The observation equations in a discretized form are:

$$y_k = Hx_k + v_k$$  \hspace{1cm} (10.52)

where $H = [1,0,-L/g,0]$ and $y_k$, $x_k$, and $v_k$, respectively, denote $y(t)$, $x(t)$, and $v(t)$ of the corresponding signals at the $k$-th sampling instant [7, 9]. The discretized form of the dynamic Equation 10.51 is:

$$x_{k+1} = Fx_k + w_k$$  \hspace{1cm} (10.53)

where

$$F = \Phi(t) \bigg|_{t=\Delta T}, \quad \Phi(t) = \left(sI - A\right)^{-1}.$$  \hspace{1cm} (10.54)

Here, $L^{-1}$ and $\Delta T$, respectively, denote the inverse Laplace transformation and the sampling period. The discretized transition noise $w_k$ becomes a white Gaussian noise with zero mean and covariance:

$$W = \sigma^2 \int_0^{\Delta T} \Phi(\Delta T - \tau) \Gamma \Gamma^T \Phi^T(\Delta T - \tau) d\tau$$  \hspace{1cm} (10.55)

The measurement of the rolling and pitching can thus be reduced to the state estimation of the linear discrete dynamic systems (Equations 10.52 and 10.53), if the angular frequencies $\omega_1$ and $\omega_2$ are given and $v_k$ is assumed to have a white Gaussian property. The state estimation is achieved by a Kalman filter [7, 13]. However, difficulties in implementing the filter are that the exact values of the two angular frequencies are a priori unknown and also time variant. To overcome the difficulty, adequate candidates $\{\omega_1, \omega_2\}$ for the parameters $\{\omega_1, \omega_2\}$ are set and a bank of Kalman filters is used. Then, the final estimate is obtained as the conditional expectation of the state estimate as follows:

$$\hat{x}_{k/k} = \sum_{i=1}^{M} p_i \hat{x}_{i/k}$$  \hspace{1cm} (10.56)

where $\hat{x}_{i/k}$ represents the state estimate $\hat{x}_{i/k}$ for the $i$-th candidate $\Omega_i = (\omega_1^i, \omega_2^i)$, and $p_i$ denotes the conditional posteriori probability of the $i$-th candidate calculated based on the Bayesian theorem:

$$p_i = \frac{p\left(y_k / \Omega_i, Y_{k-1}^i \right) p_i}{\sum_{j=1}^{M} p_j \left(y_k / \Omega_j, Y_{k-1}^j \right)}$$  \hspace{1cm} (10.57)

© 1999 by CRC Press LLC
Here, \( p(y_k / \Omega, Y^{k-1}) \) represents the conditional Gaussian probability density function of \( y_k \) under \( \Omega \) and \( Y^{k-1} = \{y_j; j \leq k - 1\} \), whose mean and variance are calculated recursively \([7]\).

The proposed measurement system can adaptively and automatically select the most appropriate candidate versus time. It thus enables an accurate on-line measurement of the rolling and pitching whose dominant angular frequencies vary with time. The first, second, and third components of the final estimate \( \hat{x}_n \) represent, respectively, the estimates of the displacement, velocity, and acceleration. The proposed system thus has an advantage in that it can measure not only the displacements, but also the velocities and the accelerations of the three signals. In order to improve the measurement accuracy of the rolling and pitching, one should place the inclinometers near the intersection \( O \) of the rolling and the pitching axes.

Finally, the dynamics of the heaving is given by Equation 10.51 similar to that of the rolling and pitching. Substituting the estimates \( \hat{\theta}(t) \) and \( \hat{p}(t) \) obtained above into Equation 10.47 and subtracting the effect of the gravitational acceleration, one can derive a linear observation equation for \( \alpha(t) \):

\[
y_k = \left[ z_i(t) - g \cos \hat{\theta}(t) \cos \hat{p}(t) \right]_{t=k\Delta T} = H_k x_k + \nu_k
\]

where \( H_k = [0,0,\cos \hat{\theta}(t) \cos \hat{p}(t)]_{t=k\Delta T} \)

\[
\nu_k = \nu_1(k\Delta T)
\]

Thus, the on-line measurement of the heaving is also realized by executing the same procedure as described before. The location of the rolling and pitching axes were assumed to be known; however, even when they are unknown, the attitude measurement system described above is effective, if we introduce the candidates on the location of the axes adding to the angular frequencies.

### Attitude Measurement for Crane Lifters

**Dynamics of Attitude Signals**

An illustrative diagram of a crane lifter system is shown in Figure 10.11. One of the easiest ways to measure the attitude of the lifter is to set up a high-resolution camera on the bottom of the trolley and to track a mark on the top of the lifter. However, it increases the cost and also the difficulty in maintenance. Furthermore, sometimes the scheme does not work because of shadows and light reflection. As previously mentioned, for gyros not offering sufficiently accurate measurement, a high-sensitivity servo-type accelerometer is used to extract the attitude signals. When setting up the sensor on the lifter, however, there is a secondary swing signal adding to the primary one, due to the free suspension of the lifter and the structure of the lifter. Despite its small amplitude, the secondary one has a higher frequency and for this reason has a large magnitude on the sensor output. The important signal for practical applications, such as the attitude control of the lifter, is the primary one, which has a larger amplitude with a lower angular frequency of \( \omega = \sqrt{g/\ell} \) (\( g \): the gravitational acceleration; \( \ell \): the wire length from the primary supporting point to the center of gravity of the pulley). If we try to attenuate the secondary swing signal by passing the output through a low-pass filter, the phase lag is also introduced into the primary swing signal and the signal can no longer be used for the accurate attitude control of the lifter.

For the above reasons, we introduce an autonomous measurement system that measures both the primary and the secondary swings by modeling the lifter system with a double pendulum and applying a Kalman filter to it \([14]\). The dynamics of the trolley-lifter system is derived using Lagrange’s equations of motion \([8, 9]\).

\[
\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{x}} \right) - \frac{\partial T}{\partial x} + \frac{\partial V}{\partial x} = u - z \dot{x}
\]

\[10.59\]
where \( T \) and \( V \) represent, respectively, the kinetic and the potential energies of the trolley-lifter system, and \( \theta_1, \theta_2 \) denote, respectively, the angles that the primary and the secondary pendulums take against the vertical line. The other variables \( x, u, \) and \( a \) represent, respectively, the location of the trolley, the driving force, and the coefficient of friction between the trolley and the rail. Considering that \( \theta_i, \theta_i (1 \leq i \leq 2) \) are small, the dynamic equation of the trolley-lifter system can be expressed as [14]:

\[
\frac{d}{dt} \left( \frac{\partial T}{\partial \dot{\theta}_i} \right) - \frac{\partial T}{\partial \theta_i} + \frac{\partial V}{\partial \theta_i} = 0 \quad (i = 1, 2) \tag{10.60}
\]

where \( T \) and \( V \) represent, respectively, the kinetic and the potential energies of the trolley-lifter system, and \( \theta_1, \theta_2 \) denote, respectively, the angles that the primary and the secondary pendulums take against the vertical line. The other variables \( x, u, \) and \( a \) represent, respectively, the location of the trolley, the driving force, and the coefficient of friction between the trolley and the rail. Considering that \( \theta_i, \theta_i (1 \leq i \leq 2) \) are small, the dynamic equation of the trolley-lifter system can be expressed as [14]:

\[
\dot{x}(t) = Ax(t) + bu(t) \tag{10.61}
\]

where \( x(t) \) is the state vector \( x(t) = x, \dot{x}, \theta_1, \dot{\theta}_1, \theta_2, \dot{\theta}_2 \)\(^{T}\). Taking into account the approximation errors in deriving Equation 10.61, air resistance, friction in the wires, and microscopic swings at the other connection points, it is reasonable to introduce white Gaussian noises \( w(t) \) (1 \( \leq i \leq 3 \)) with zero mean and appropriate variances to the dynamic Equation 10.61 as in Equation 10.49 as follows [14]:

\[
\dot{x}(t) = Ax(t) + bu(t) + \Gamma \omega(t) \tag{10.62}
\]

where

\[
\Gamma = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}^{T}, \quad \omega(t) = \begin{bmatrix}
\omega_1(t) \\
\omega_2(t) \\
\omega_3(t)
\end{bmatrix}^{T} \tag{10.62a}
\]
Sensor Outputs and On-Line Attitude Measurement

When a servo-type accelerometer is set up on the lifter (in the direction of the swing) at the place of the distance $\ell_1'$ from the secondary supporting point, the output of the sensor becomes:

$$\gamma(t) \cong -\ddot{x} - \ell_1' \ddot{\theta}_1 - \ell_2' \ddot{\theta}_2 - g\dot{\theta}_2$$  \hspace{1cm} (10.63)

where $\ell_1'$ is the distance between the primary and the secondary supporting points (see Figure 10.12). Substitution of Equation 10.61 into Equation 10.63 yields an output expressed in terms of the state vector $x(t)$, as in Equation 10.52. Using a rotary encoder to measure the location and the velocity of the trolley, and then combining these three sensor outputs with the dynamic Equation 10.62 and applying a Kalman filter enables the state vector to be estimated on-line. Using this approach, both angular displacement and velocity of the deflections $\theta_1, \theta_2$ of the two pendulums can be measured exactly.

Aircraft Attitude Determination

The determination of aircraft attitude requires the measurement of angles about three independent body axes. These angles are the roll, pitch, and yaw angles. There are two primary means employed today for measuring these angles: the first method uses VGs to measure the roll and pitch angles, and a DG to measure the yaw angle. The second method, more commonly used today, employs an IMU for full three-axis attitude determination coupled with a baro-altimeter to correct for vertical drift errors in the IMU. Both methods are described below.

Vertical and Directional Gyro Analysis

A VG is a two degree-of-freedom gyro with its spin axis mounted nominally vertical. It employs two specific force sensors mounted nominally horizontal on the inner gimbal. The two angles measured by the VG — roll and pitch — require nearly identical analyses [1]. Consider the situation shown in
Figure 10.13, depicting an aircraft with a roll angle of $\phi$ with respect to the local vertical. The sensed roll angle $\phi_s$ is given by the difference in the actual roll angle and the gyro roll drift error $\phi_{\Delta}$:

$$\phi_s = \phi - \phi_{\Delta}$$  \hspace{1cm} (10.64)

In order to compensate for this drift error, gyros employ a specific force sensor such as an electrolytic bubble device, which senses drift error. This correction device senses the angular difference between the specific force vector $f$ acting on the aircraft roll axis and the gyro axis, as shown in Figure 10.13. Thus,

$$\phi_{\Delta} = \tan^{-1} \left[ \frac{f_x}{f_y} - \phi_{\Delta} \right] \equiv \left( \frac{f_x}{f_y} - \phi_{\Delta} \right)$$  \hspace{1cm} (10.65)

where $f_x$ is the side horizontal component of $f$ and $f_y = \text{force of gravity}$ is the vertical component. A similar analysis for the pitch angle $\theta$ yields:

$$\theta_s = \theta - \theta_{\Delta}$$  \hspace{1cm} (10.66)

$$\theta_{\Delta} = \tan^{-1} \left[ \frac{f_y}{f_x} - \theta_{\Delta} \right] \equiv \left( \frac{f_y}{f_x} - \theta_{\Delta} \right)$$  \hspace{1cm} (10.67)

where $f_y$ is the back horizontal component of $f$. Next, define the gyro angular momentum vector by:

$$\mathbf{H}_{\text{VG}} = \left[ J_x \dot{\phi}_{\Delta}, J_y \dot{\theta}_{\Delta}, -h \right]$$  \hspace{1cm} (10.68)

where $J_x$ and $J_y$ are the sensor moments of inertia and $h$ is the gyro spin angular momentum. In addition, define the inner gimbal axes angular velocity vector as:

$$\mathbf{\omega}_{\text{VG}} = \left[ \phi_{\Delta}, \dot{\theta}_{\Delta}, 0 \right]$$  \hspace{1cm} (10.69)

Finally, define the gimbal torque vector by:

$$\mathbf{Q}_{\text{VG}} = \left[ Q_{\text{cx}} + Q_{\text{dx}}, Q_{\text{cy}} + Q_{\text{dy}}, 0 \right]$$  \hspace{1cm} (10.70)

where

$$Q_{\text{cx}} = \text{gimbal roll control torque} = -k_\theta \theta_{\Delta}$$  \hspace{1cm} (10.71a)
\[ Q_{\phi y} = \text{gimbal pitch control torque} = k_c \phi_b \]  \hspace{1cm} (10.71b)

\[ Q_{\phi x} = \text{gimbal roll disturbance torque} = -k_d (\phi_A - \phi) + \text{random torques} \]  \hspace{1cm} (10.71c)

\[ Q_{\phi y} = \text{gimbal pitch disturbance torque} = -k_d (\theta_A - \theta) + \text{random torques} \]  \hspace{1cm} (10.71d)

and the \( k_c \) and \( k_d \) are constant scaling factors related to each torque component.

Using the vectors defined in Equations 10.68 through 10.70, the gyro equations of motion are given by:

\[
\frac{\partial}{\partial t} \left( H_{\nu_G} \right) + (\omega_{\nu_G} \times H_{\nu_G}) = Q_{\nu_G}
\]  \hspace{1cm} (10.72)

Taking the Laplace transform of the expansion of Equation 10.72, with the assumption that \( J_x \equiv J_y \equiv J \), yields the following gyro equations of motion in the Laplace domain:

\[
\begin{pmatrix}
J_x s^2 + k_c s \\
hs + k_c \\
J_y s^2 + k_c s
\end{pmatrix}
\begin{pmatrix}
\phi_A(s) \\
\theta_A(s)
\end{pmatrix}
\approx
\begin{pmatrix}
-k_d \phi + k_c s \phi(s) + \text{random torques} \\
-k_d \theta + k_c s \theta(s) + \text{random torques}
\end{pmatrix}
\]  \hspace{1cm} (10.73)

For normal gyro operation, \( J_x \equiv J_y \equiv 0 \) and \( k_c/h \ll 1 \); so these factors may be ignored in Equation 10.73. Thus, solving for the desired roll and pitch angles under these assumptions gives [1]:

\[ \phi_A = \begin{cases} 
\phi & \omega \gg k_c/h \\
\phi - \phi_b/g & \omega \ll k_c/h
\end{cases} \]  \hspace{1cm} (10.74)

\[ \theta_A = \begin{cases} 
\theta & \omega \gg k_c/h \\
\theta - \theta_b/g & \omega \ll k_c/h
\end{cases} \]  \hspace{1cm} (10.75)

A DG is a two degree-of-freedom gyro with its spin axis mounted nominally horizontal and pointing in the direction of magnetic north. It employs a single specific force sensor mounted on the inner gimbal [1]. The DG measures the third required aircraft angle, yaw, generally denoted by \( \psi \). The sensed yaw angle \( \psi_s \) is given by the difference in the actual yaw angle \( \psi \) (angle between the aircraft \( z \)-axis and true north) and the gyro heading angle drift error \( \psi_A \) (angle between the gyro axis and true north):

\[ \psi_s = \psi - \psi_A \]  \hspace{1cm} (10.76)

Define the gyro angular momentum vector by:

\[ H_{DG} = \begin{bmatrix} I_x \dot{\theta}_A, I_y \dot{\psi}_A, -h \end{bmatrix} \]  \hspace{1cm} (10.77)

and the inner gimbal axes angular velocity vector as:

\[ \omega_{DG} = \begin{bmatrix} \dot{\theta}_A, \dot{\psi}_A, 0 \end{bmatrix} \]  \hspace{1cm} (10.78)
and the gimbal torque vector as

\[
Q_{\text{gim}} = \begin{bmatrix} Q_\gamma + Q_{\delta b} \\
Q_\alpha + Q_{\delta d} \\
0 \end{bmatrix}
\]  

(10.79)

Here, the torque vector components are given by:

\[
Q_\gamma = k \left( M_\Delta - \psi_\Delta \right)
\]  

(10.80a)

\[
Q_\alpha = -k_\delta \theta_b
\]  

(10.80b)

\[
Q_\delta = -k_\theta \left( \dot{\theta}_\Delta - \dot{\theta} \right) + \text{random torques}
\]  

(10.80c)

\[
Q_{\delta d} = -k_\delta \left( \psi_\Delta - \psi \right) + \text{random torques}
\]  

(10.80d)

where \( M_\Delta \) = magnetic compass heading error (from true north). Therefore, the DG equations of motion are given in Laplace domain as:

\[
\begin{bmatrix}
J_\Delta s^2 + k_\delta s \\
hs + k_\delta 
\end{bmatrix}
\begin{bmatrix}
\Theta_\Delta(s) \\
\Psi_\Delta(s)
\end{bmatrix}
\equiv
\begin{bmatrix}
k_\delta \left( M_\Delta + k_\delta s \dot{\theta} \right) + \text{random torques} \\
-k_\theta \left( \dot{\theta}_\Delta + k_\delta s \dot{\psi} \right) + \text{random torques}
\end{bmatrix}
\]  

(10.81)

The desired yaw angle measurement for the DG is thus given as [1]:

\[
\psi_\Delta = \begin{cases}
\psi \\
\psi - M_\Delta
\end{cases} \begin{array}{l}
\omega \gg \frac{k_\delta}{\hbar} \\
\omega \ll \frac{k_\delta}{\hbar}
\end{array}
\]  

(10.82)

As indicated in Table 10.2, the accuracies of both VGs and DGs are approximately 1°. An improvement of over 2 orders of magnitude can be obtained through the use of inertial measurement units, which are described next.

**Inertial Measurement Units (IMUs)**

Inertial measurement units consist of gyroscopes and accelerometers that together provide full three-axis attitude measurements. Most are mounted on stable gimbaled platforms that remain locally horizontal via torquing devices. An IMU aboard an aircraft cannot measure exactly the local vertical due to the fact that the specific force acting on the aircraft has a horizontal component due to vehicle motion. In addition, since the vehicle is moving with respect to the inertial reference frame, the Earth’s magnetic pole cannot be determined precisely [1].

These problems (errors) are minimized by aligning the IMU to be exactly horizontal and north pointing while the aircraft is stationary. Once platform motion begins, the IMU may be constantly realigned by sensing changes in the direction of vertical and north, and then applying appropriate torques to the platform to keep it properly aligned. This realignment is accomplished by integrating the two orthogonal accelerometer outputs to determine the components of horizontal velocity. This data, combined with the Earth’s rotation rate, yields the desired rates of change in local vertical and true north at the vehicle’s current latitude and longitude. Performing a second integration of the sensor outputs yields an estimate of relative position.
Analysis in Bryson et al. [1], has shown that the pitch angle (variation in platform horizontal position) is given by the IMU sensor output as:

$$\theta(t) = -\varepsilon \sin(\omega_s t) - \frac{b}{g}$$  \hspace{1cm} (10.83)

where $\varepsilon = \text{gyro drift rate error}$, $b = \text{specific force sensor error}$, and $\omega_s = \text{Schuler frequency} = \sqrt{g/R}$, $[g = \text{force of gravity}, R = \text{Earth's radius}]$. Thus, the platform root-mean-square pitch angle becomes:

$$\theta_{\text{rms}} = \left[ \frac{1}{2} \left( \frac{\varepsilon}{\omega_s} \right)^2 + \left( \frac{b}{g} \right)^2 \right]^{1/2}$$  \hspace{1cm} (10.84)

Using typical values for $\varepsilon (\approx 0.015^\circ \text{h}^{-1})$, $\omega_s (\approx 0.71^\circ \text{h}^{-1})$, and $b (\approx 0.01)$ yields an rms pitch angle error of $\theta_{\text{rms}} = 0.01^\circ$. Thus, it is apparent that under normal operating conditions the IMU provides a two orders-of-magnitude improvement in sensor accuracy when compared to the VG and DG.

**Spacecraft Attitude Determination**

Most spacecraft attitude determination techniques rely upon finding the orientation of a single axis in space (e.g., the spacecraft z-axis) plus the spacecraft rotation about this axis. This provides a full three-axis attitude solution. In order to achieve this, reference sources that are external to the spacecraft must be used. Specifically, full three-axis spacecraft attitude determination requires at least two external vector measurements. Commonly used reference sources for these external vector measurements include the sun, Earth, moon, stars, planets, and the Earth’s magnetic field. In addition, IMUs are also used to provide the necessary attitude measurements.

**Attitude Determination Methodology**

The first step in attitude determination is to determine the angles between the spacecraft’s primary axis and the two (or more) attitude reference sources. For example, suppose a particular spacecraft is using the sun and the Earth for attitude reference. The two angles in this case are referred to as the sun angle $\beta_s$ and the nadir angle $\Gamma_N$. Since the orientation of even a single spacecraft axis is unknown at this point, these angles establish two cones along which the attitude vector $A$ must lie. Since the attitude vector must lie on both cones, it must lie along the intersection between the two cones [4] (See Figure 10.14). The two vectors, notably $A_s$ and $A_n$, resulting from the intersection of these two cones may be determined by the following method derived by Grubin [15]. Let $S$ represent the sun vector, $E$ the spacecraft nadir vector, and $A$ the desired attitude vector, each defined in Cartesian space as follows:

$$S = (S_x, S_y, S_z)$$  \hspace{1cm} (10.85)

$$E = (E_x, E_y, E_z)$$  \hspace{1cm} (10.86)

$$A = (A_x, A_y, A_z)$$  \hspace{1cm} (10.87)

Let the vectors $S$, $E$, and $N$ define a set of base unit vectors with:

$$N = \frac{S \times E}{|S \times E|} = \left( N_x, N_y, N_z \right)$$  \hspace{1cm} (10.88)
If we introduce a proper set of scaling factors as follows:

\[
I_x = \frac{\cos \beta - (S \cdot E) \cos \Gamma_N}{1 - (S \cdot E)^2}
\]  
\[
I_y = \frac{\cos \Gamma_N - (S \cdot E) \cos \beta}{1 - (S \cdot E)^2}
\]  
\[
I_z = \sqrt{1 - I_x \cos \beta - I_y \cos \Gamma_N}
\]

then the two possible attitude vectors \(A_1\) and \(A_2\) are found to be:

\[
A_{1,2} = \left( I_x S_x + I_y E_y \pm I_z N_z \right), \left( I_x S_x + I_y E_y \pm I_z N_z \right), \left( I_x S_x + I_y E_y \pm I_z N_z \right)
\]  

In Equations 10.88 through 10.90, \(S \times E\) represents the Cartesian vector product, and \(S \cdot E\) represents the Cartesian scalar product. The radicand in Equation 10.89c may be negative, thus producing imaginary values.
values for $I_z$ whenever the two cones do not intersect. Such occurrences are usually attributed to sensor error or random noise fluctuations. In this case, one can add a predetermined sensor bias to both sensors in order to "broaden" the cone angles, thus forcing the cones to intersect.

It should be noted that for most applications involving spacecraft attitude determination, the principle coordinate system used is the celestial sphere coordinate system. This coordinate system has the $z$-axis aligned with the Earth’s polar axis, and the $x$-axis aligned with the intersection of the Earth’s equatorial plane and the Earth’s orbital plane around the sun (i.e., aligned with the vernal equinox). In this coordinate system, all vectors are considered unit vectors and the two principle measurements describing a vector's position are the right-ascension and declination angles, denoted $\Omega$ and $\Delta$, respectively. Thus, the sun vector $S$ and the Earth nadir vector $E$ used in Equations 10.85 and 10.86 will, in general, be given as right-ascension and declination angles that can be converted to Cartesian coordinates via the following set of transformations:

\[
\begin{align*}
x &= \cos(\Omega)\cos(\Delta); \\
y &= \sin(\Omega)\cos(\Delta); \\
z &= \sin(\Delta)
\end{align*}
\]  

\[
\begin{align*}
\Omega &= \tan^{-1}\left(\frac{y}{x}\right); \\
\Delta &= \sin^{-1}(z)
\end{align*}
\]  

The final step in measuring three-axis attitude is to determine which attitude solution is correct, $A_1$ or $A_2$, and then measure the rotation about this axis. The two ambiguous attitude solutions may be resolved by comparison with a priori attitude information, if available, or through the use of trace averaging [4]. Trace averaging is a method of plotting each attitude solution on a right-ascension versus declination plot and choosing the area of greatest concentration as the correct solution, as demonstrated in Figure 10.15. Since the attitude is assumed to change more slowly than the attitude sensor’s sample rate, over short time intervals the data for the correct solution usually form a “cluster” near the correct attitude; the data for the incorrect solution are usually much more scattered.

Once the correct attitude vector has been obtained, the orientation of the remaining two orthogonal axes may be found by measuring the rotation, or phase angle, of the spacecraft about the preferred axis. Any sensor measurement that provides this phase angle may be used. An example of this technique is provided by the panoramic annular lens attitude determination system (PALADS), described in the next section. This imaging system uses a unique “three-dimensional” lens that provides simultaneous detection of two (or more) reference sources [16]. This information, combined with the orientation of the single axis, uniquely determines three-axis attitude.

---

FIGURE 10.15 Method of trace averaging for resolving ambiguous attitude solutions.
The three angles derived above, which are commonly referred to as Euler angles, define the orientation of the three spacecraft axes with respect to the chosen reference frame. A more formal treatment of the attitude solution usually requires specifying the components of the $3 \times 3$ attitude matrix $A$. Each component of the attitude matrix defines the angular relationship between a given spacecraft axis and a reference frame axis. Various methods exist for computing the attitude matrix $A$ (see [4]); the preferred method depends on the particular application at hand.

**PALADS**

The primary component of PALADS is the panoramic annular lens (PAL), a single-element lens made from a high index of refraction glass with certain portions of the lens coated with a mirrored surface. Hence, it relies on both refraction and reflection in forming an image (Figure 10.16). The lens is unique in that it images a three-dimensional object field onto a two-dimensional image plane, whereas a “normal” lens is capable of only imaging two-dimensional object space onto an image plane. The combination of high index of refraction glass and mirrored surfaces provides the PAL with a field of view extending from approximately $65^\circ$ to $110^\circ$ from the optical axis. This $45^\circ$ field of view covers the entire $360^\circ$ surrounding the optical axis [17]. Any ray originating from outside the $45^\circ$ field of view will not form a part of the image. The PAL may be attached to any high-quality imaging system using an appropriate transfer lens. As currently configured, the PALADS imaging system utilizes a Sony XC-73 charged-couple device (CCD), a black and white video camera coupled to the PAL via a Nikon f/1.4 transfer lens.

The hemispherical view provided by PALADS allows for single-sensor detection of multiple attitude reference sources, such as the Earth and the sun or moon. The position of each reference source in the image plane translates into a unique azimuth elevation angle between the PAL’s optical axis and the reference source. Since the PAL has a $360^\circ$ field of view surrounding the optical axis, it may detect several reference sources simultaneously. The data points associated with each source are extracted from the image plane using digital image processing techniques. Thus, it is easy to see how a single image from PALADS (containing two or more reference sources) provides the necessary angle data to determine three-axis spacecraft attitude.

**References**

Inertial Navigation

Halit Eren and C. C. Fung

The Principles

The original meaning of the word navigation is “ship driving.” In ancient times when sailing boats were used, navigation was a process of steering the ship in accordance with some means of directional information, and adjusting the sails to control the speed of the boat. The objective was to bring the vessel from location A to location B safely. At present, navigation is a combination of science and technology. No longer is the term limited to the control of a ship on the sea surface; it is applied to land, air, sea surface, underwater, and space.

The concept of inertial-navigator mechanization was first suggested by Schuler in Germany in 1923. His suggested navigation system was based on an Earth-radius pendulum. However, the first inertial guidance system based on acceleration was suggested by Boykow in 1938. The German A-4 rocket, toward the end of World War II, used an inertial guidance system based on flight-instrument type gyroscopes for attitude control and stabilization. In this system, body-mounted gyro-pendulum-integrating accelerometers were used to determine the velocity along the trajectory. The first fully operational inertial auto-navigator system in the U.S. was the XN-1 developed in 1950 to guide C-47 rocket. Presently, inertial navigation systems are well developed theoretically and technologically. They find diverse applications,
allowing the choice of appropriate navigation devices, depending on cost, accuracy, human interface, global coverage, time delay, autonomy, etc.

Inertial navigation is a technique using a self-contained system to measure a vehicle’s movement and determine how far it has moved from its starting point. Acceleration is a vector quantity involving magnitude and direction. A single accelerometer measures magnitude but not direction. Typically, it measures the component of acceleration along a predetermined line or direction. The direction information is usually supplied by gyroscopes that provide a reference frame for the accelerometers. Unlike other positional methods that rely on external references, an inertial navigation system (INS) is compact and self-contained, as it is not required to communicate to any other stations or other references. This property enables the craft to navigate in an unknown territory.

Inertial navigation can be described as a process of directing the movement of a vehicle, rocket, ship, aircraft, robot, etc., from one point to another with respect to a reference axis. The vehicle’s current position can be determined from “dead reckoning” with respect to a known initial starting reference position. On the Earth’s surface, the conventional reference will be North, East, and Down. This is referred to as the Earth’s fixed axes. A vehicle such as an aircraft or a marine vessel will have its own local axes: roll, pitch, and yaw.

The inertial sensors of the INS can be mounted in such a way that they stay leveled and pointing in a fixed direction. This system relies on a set of gimbals and sensors attached on three axes to monitor the angles at all times. This type of INS is based on a navigational platform. A sketch of a three-axis platform is shown in Figure 10.18. Another type of INS is the strapdown system that eliminates the use of gimbals. In this case, the gyros and accelerometers are mounted to the structure of the vehicle. The measurements received are made in reference to the local axes of roll, pitch, and yaw. The gyros measure the movement of angles in the three axes in a short time interval (e.g., 100 samples per second). The computer then uses this information to resolve the accelerometer outputs into the navigation axes. A schematic block diagram of the strapdown system is shown in Figure 10.19.

The controlling action is based on the sensing components of acceleration of the vehicle in known spatial directions, by instruments which mechanize Newtonian laws of motion. The first and second integration of the sensed acceleration determine velocity and position, respectively. A typical INS includes

---

FIGURE 10.17 In inertial navigation, the movement of a vehicle, rocket, ship, aircraft, robot, etc. with respect to a reference axis is monitored. On the Earth’s surface, the conventional reference is the Earth’s fixed axes — North, East, and Down. A vehicle such as an aircraft or a marine vessel will have its own local axes, known as roll, pitch, and yaw.
a set of gyros, a set of accelerometers, and appropriate signal processing units. Although the principle of
the systems may be simple, the fabrication of a practical system demands a sophisticated technological
base. The system accuracy is independent of altitude, terrain, and other physical variables, but is limited
almost purely by the accuracy of its own components. Traditional INSs mainly relied on mechanical
gyros and accelerometers, but today there are many different types available, such as optical gyroscopes,
piezoelectric vibrating gyroscopes, active and passive resonating gyroscopes, etc. Also, micromachined
gyroscopes and accelerometers are making an important impact on modern inertia navigation systems.
A brief description and operational principles of gyroscopes and accelerometers suitable for inertial
navigation are given below.

Major advances in INS over the years include the development of the electrostatic gyro (ESG) and the
laser gyro. In ESG, the rotor spins at a speed above $200 \times 10^3 \text{ rpm}$ in a near-vacuum environment. The
rotor is suspended by an electrostatic field; thus, it is free from bearing friction and other random torques
due to mechanical supports. Hence, its operation results in a superior performance compared to others,
closely resembling the performance of a theoretical gyro. Although no system can claim to reach perfection,
an ESG requires less frequent updates as compared to other mechanical gyroscopes.

**Gyroscopes**

There are two broad categories: (1) mechanical gyroscopes and (2) optical gyroscopes. Within both of
these categories, there are many different types available. Only the few basic types will be described to
Mechanical gyroscopes: The first mechanical gyroscope was built by Foucault in 1852, as a gimbaled wheel that stayed fixed in space due to angular momentum while the platform rotated around it. They operate on the basis of conservation of angular momentum by sensing the change in direction of an angular momentum. There are many different types, which are:

1. **Single degree of freedom gyroscopes**: include the rate, rate integrating, spinning rotor flywheel, electron, and particle gyros.
2. **Two degree of freedom gyroscopes**: incorporate the external gimbal types, two-axis floated, spherical free-rotor, electrically suspended, gas-bearing free-rotor gyros.
3. **Vibrating gyroscopes**: include the tuning fork, vibrating string, vibrating shell, hemispherical resonating, and vibrating cylinder gyros.
4. **Continuous linear momentum gyroscopes**: incorporate a steady stream of fluid, plasma, or electrons, which tends to maintain its established velocity vector as the platform turns. One typical example is based on a differential pair of hot-wire anemometers to detect the apparent lateral displacement of the flowing air column.

The operating principle of all mechanical gyroscopes is based on the conservation of angular momentum, as shown in Figure 10.20. The angular momentum is important since it provides an axis of reference. From Newton’s second law, the angular momentum of a body will remain unchanged unless it is acted upon by a torque. The rate of change of angular momentum is equal to the magnitude of the torque, in vectorial form as:

\[ T = \frac{dH}{dt} \]  

(10.92)

where \( H \) = angular momentum (= inertia \times angular velocity, \( \dot{\omega} \)).
If a torque acts about the axis of rotation, it changes the angular velocity by:

\[ T = I \frac{d\omega}{dt} = I \alpha \]  

(10.93)

where \( I \) = inertia about the spin axis
\( \alpha \) = angular acceleration

If the torque is orthogonal to the spinning axis, it cannot change the magnitude of the angular velocity vector, but it can change direction in the same direction as the torque. The spin axis always tries to align with the external torque.

Therefore, from Equations 10.94 and 10.92:

\[ T = dH/dt = H \frac{d\theta}{dt} = H \Omega \]  

(10.95)

where \( \Omega \) is the precession rate or the angular velocity of the spinning wheel about the axis normal to the plane of the spin and the input torque. Generally, the spin axis tries to align with the external input torque.

These equations can be elaborated to describe the operating principles of mechanical gyroscopes by taking into account the angular momentum in \( x \), \( y \), and \( z \) directions, nutation, coriolis accelerations, directions of other influencing torques and linear forces, etc. Here, the operation of the well-known flywheel gyroscope will be described as the basis for further discussions on inertial navigation systems.

An example of a double-axis flywheel gyro is shown in Figure 10.21. In this type of gyroscope, an electrically driven rotor is suspended in a pair of precision low-friction bearings at both ends of the rotor axle. The rotor bearings are supported by a circular ring known as an inner gimbal ring, which in turn pivots on a second set of bearings that is attached to the outer gimbal ring. The pivoting action of the inner gimbal defines the horizontal axis of the gyro, which is perpendicular to the spin axis of the rotor. The outer gimbal ring is attached to the instrument frame by a third set of bearings that defines the vertical axis of the gyro that is perpendicular to both the horizontal axis and the spin axis. This type of
suspension has the property of always preserving the predetermined spin-axis direction in inertial space. Equations governing the two degrees of freedom gyroscope can be written using Equations 10.92 to 10.95. The torque with respect to an inertial reference frame can be expressed as:

\[ T = \dot{H}_1 \]  

(10.96)

If the Earth is taken as a moving reference frame, then:

\[ \dot{H}_1 = \dot{H}_E + \omega_{IE} \dot{H} \]  

(10.97)

If the gyroscope itself is mounted on a vehicle (e.g., aircraft) that is moving with respect to the Earth, then:

\[ \dot{H}_E = \dot{H}_B + \omega_{EB} \dot{H} \]  

(10.98)

The case of the gyroscope can be mounted on a platform so that it can rotate relative to the platform; then:

\[ \dot{H}_B = \dot{H}_C + \omega_{BC} \dot{H} \]  

(10.99)

Finally, the inner gimbal can rotate relative to the case, hence:

\[ \dot{H}_C = \dot{H}_G + \omega_{GC} \dot{H} \]  

(10.100)

Substituting Equations 10.97 to 10.100 into Equation 10.96 yields:

\[ T = \dot{H}_G \left( \omega_{GC} + \omega_{BC} + \omega_{EB} + \omega_{IE} \right) \dot{H} \]  

(10.101)
But:

\[
(\omega_{GC} + \omega_{BC} + \omega_{EB} + \omega_{IE}) = \omega_{IG}
\]  

(10.102)

Therefore,

\[
T = H_G + \omega_{IG} \cdot H
\]  

(10.103)

By carefully constructing the gyroscope and maintaining the spin velocity of the rotor constant, \(H_G\) can be made to be zero. Thus, the law of gyroscopes can be written as:

\[
T = \omega_{IG} \cdot H
\]  

(10.104)

This means that if an external torque \(T\) is applied to the gyroscope, the inner gimbal will precess with respect to the inertial frame with a velocity \(\omega\) such that Equation 10.104 is satisfied.

In most designs (e.g., rate gyros), the gimbal is hermetically sealed in a liquid and liquid is floated in the case, to unload the gimbal bearings and to provide viscous damping. A pick-off senses gimbal deflection by means of position transducers and it controls a servo system, with a servomotor driving the case to maintain pick-off null.

**Optical gyroscopes** are based on the inertial properties of light instead of Newton’s law of motion. They operate on the Sagnac effect, which produces interferometer fringe shift against the rotation rate. In this case, two light waves circulate in opposite directions around a path of radius \(R\), beginning at source \(S\). A typical arrangement for the illustration of operation principles is shown in Figure 10.22. When the gyro is stationary, the two beams arrive at the detector at the same time and no phase difference will be recorded. Assume that the source is rotating with a velocity \(\omega\) so that light traveling in the opposite direction to rotation returns to the source sooner than that traveling in the same direction. Thus, any rotation of the system about the spin axis causes the distance covered by the beam traveling in the direction of rotation to lengthen, and the distance traveled by the beam in the opposite direction to shorten. The two beams interfere to form a fringe pattern and the fringe position may be recorded, or the phase differences of the two beams may be sensed. This phase difference is directional and proportional to the angular velocity. Usually, photodetectors are used to measure the phase shift.

Two different types of optical gyros can be categorized: either passive or active, and resonant or nonresonant. In passive gyrosensors, the Sagnac phase is measured by some external means; whereas in active gyros, the Sagnac phase causes a frequency change internal to the gyro that is directly proportional to the rotation rate.

The Sagnac interferometer is the basis of the **interferometric fiber-optic gyro (IFOG)**. A typical fiber-optic gyro is shown in Figure 10.22. However, the most widely used gyro is the active resonant **ring laser gyro (RLG)**, which is applied extensively in aircraft navigation. Two different types of resonant passive gyros, the **resonant fiber-optic gyro (RFOG)** and the **micro-optic gyro (MOG)**, are lower cost devices commonly used and comparable to RLGs.

**Accelerometers**

In inertial navigation, the absolute acceleration is measured in terms of three mutually perpendicular components of the total acceleration vector. Integrating these acceleration signals twice gives the displacement from an initial known starting location. Details of the acceleration and accelerometers are given elsewhere in this book (see Acceleration, Vibration, and Shock). Accelerometers are made from three basic elements: proof mass, suspension mechanism, and pick-off mechanism. Some accelerometers require electric or magnetic force generators and appropriate servo loops. Accelerometers measure not only real vehicular acceleration, but also respond to gravitational reaction forces. Acceleration due to gravity is a function of position — in particular, latitude and altitude — and is compensated by computers.
The most commonly used accelerometer in navigation systems is based on pendulous types. These accelerometers can be classified as:

1. Generic pendulous accelerometer
2. Q-flex type accelerometers
3. Micromachined accelerometers (A typical example of a modern micromachined accelerometer is given in Figure 10.23.)

Accelerations in the three axes are measured by suitably positioned accelerometers. Since accelerometers contain errors, the readings must be compensated by removing fixed biases or by applying scaling factors. The errors may be functions of operating temperature, vibration, or shock. Measurement of time must be precise as it is squared within the integration process for position determination. The Earth's rotation must also be considered and gravitational effects must be compensated appropriately.

**Errors and Stabilization**

**Errors**

In general, inertial navigation is an initial value process in which the location of the navigating object is determined by adding distances moved in known directions. Any errors in the system cause misrepresentation of the desired location by being off-target. The major disadvantage of an inertial guidance system is that its errors tend to grow with time. These errors in the deduced location are due to a number of reasons, including: imperfect knowledge of the starting conditions, errors in computation, and mainly errors generated by gyros and accelerometers.
If the error build-up with time becomes too large, external aids (e.g., LORAN, OMEGA) may be used to reset or update the system. Optimal use of the data from external aids must account for the geometry of the update and also for the accuracy of the update relative to the accuracy of the inertial system. The Kalman filter, for example, is one of the computational procedures frequently applied for optimally combining data from different sources.

Errors can broadly be classified as:

1. **System heading error**: A misalignment angle in the heading of an object traveling with a velocity can cause serious errors. For example, a vehicle traveling with velocity of 500 km h\(^{-1}\) in the same direction with 0.1° initial heading error will be off the target by approximately 873 m at the end of 1 h travel.

2. **Scale error**: Error in scaling can accumulate. In order to minimize scale errors, a scale factor is used. The scale factor is the ratio between changes in the input and output signals. It simply translates the gyro output (counts per second in the case of RLG) into a corresponding angle rotation. Some instruments may have different scale factors for positive and negative inputs, known as **scale factor asymmetry**. (Scale factors are measured in ° h\(^{-1}\) mA\(^{-1}\), ° h\(^{-1}\) Hz\(^{-1}\), or g Hz\(^{-1}\).)

3. **Nonlinearity and composite errors**: In most cases, scale factors are not constant, but they can have second- or higher-order terms relating the output signals to the input. Statistical techniques can be employed to minimize these errors.

4. **Bias errors**: Zero offset or bias error is due to existence of some level of output signal for a zero input. Bias errors exist in accelerometers, gyros, tilt misalignments, etc.

5. **Random drift and random walk errors**: In some cases, the outputs of the devices can change due to disturbances inside the sensors, such as ball bearing noise in mechanical gyros. These disturbances may be related to temperature changes, aging, etc. White noise in optical gyros can cause a long-term accumulation in angle error known as the **random walk**.

6. **Dead band, threshold, resolution, and hysteresis errors**: These errors can be related to inherent operation of accelerometers and gyros. They can be due to stiction, minimum input required for an output, minimum measurable outputs, and nonrepeatability of variations in the output versus variations in the input.

---

**FIGURE 10.23** A typical example of a modern micromachined accelerometer. Multiple accelerometers can be mounted on a single chip, sensing accelerations in the \(x\), \(y\), and \(z\) directions. The primary signal conditioning is also provided in the same chip. The output from the chip is usually read in digital form.
It should be pointed out that this list is by no means exhaustive. Detailed error analysis can be found in the references cited.

**Stabilization**

The inertial navigation sensors must maintain angles within specified limits in spite of the disturbances imposed by the moving object. Accuracy requirements demand that the system must provide reliable and stable information in spite vibrations and other disturbing factors. One way of achieving stabilization is by using a stabilized platform. These platforms are designed to maintain accelerometers and gyroscopes accurately leveled and oriented in the azimuth direction. In some cases, the platform is driven around its axis by servo amplifiers and electric motors. Sensitive pick-offs on the gyroscopes fed error signals are used to maintain a desired stability of the platform in the presence of disturbing torques. The operation of a typical system, in block diagram form, is shown in Figure 10.24.

Unlike platform models, in a strapped-down system, gyroscopes and accelerometers are rigidly mounted to the vehicle structure so that they move with the vehicle. The accelerometers and gyroscopes are manufactured to measure accelerations and angles up to the maximum expected values. As the vehicle travels, the measured values are frequently transmitted to a computer. The computer uses these values to resolve the readings into the navigation axis sets and make deductions on the body axis sets.

**Vehicular Inertial Navigation**

In modern vehicular navigation, computerized maps and mobile communication equipment are integrated together with inertial and/or other electronic navigation systems. In recent years, in the wake of low-cost GPS systems, the vehicular navigation system has attracted much attention due to its large potential markets for consumer as well as business vehicles.

Automobile navigation systems are based on dead-reckoning, map matching, satellite positioning, and other navigational technologies. Map intelligent systems achieve high relative accuracy by matching dead-reckoned paths with road geometry encoded in a computerized map. This is also used to perform other functions such as vehicle routing and geocoding. Satellite-based navigation systems achieve high absolute accuracy with the support of dead-reckoning augmentation.

The capabilities and functions of automobile navigation systems depend on:

- Choosing the necessary technology
- Integrating the overall system
- Resolving driver interface
- Providing map data basis
- Coordinating mobile communications
Digital maps and mobile data communications combine together for full usefulness and effectiveness. The navigation systems are greatly enhanced in conjunction with stored digital maps combined with effective communications.

The usefulness of a navigation system is related to the accuracy in position determination. There are a number of methods available with varying accuracy; these include the following:

**Dead-reckoning**

Dead-reckoning is the process of determining vehicle location relative to an initial position by integrating measured increments and directions of travel. The devices include the odometer, the differential odometer, and a magnetic compass. Gyros and inertial systems prove to have limited applications in harsh automotive environments. Although, dead-reckoning systems suffer from error accumulation, they are widely used inertial navigation systems, particularly in robotics and vehicular applications. Even the most precise navigation system requires periodic reinitialization and continuous calibrations by computers.

**Radiolocation**

In radiolocation, the global positioning system (GPS) is used extensively. Nevertheless, LORAN is gaining popularity as means of tracking land vehicle location from a central location. But its modest accuracy limits its global application in automotive navigation.

**Map Matching**

Artificial intelligence concepts are applied to match dead-reckoned vehicle paths, which are stored in computers. In map matching, sensed mathematical features of the vehicle paths are continuously associated with those encoded in a map database. Thus, a vehicle’s dead-reckoned location can be initialized automatically at every turn to prevent accumulation of dead-reckoning errors.

The first application of map matching technology was in the Automatic Route Control System (ARCS), which used a differential odometer for dead-reckoning. In another system, the Etak map matching system, a solid-state flux gate compass is used as well as a differential odometer to dead-reckon paths for matching with digitized maps and aerial photographs. Further details on these technologies can be found in the references given at the end of this chapter.

In a map matching system, as long as the streets and road connectivities are accurately defined, the process identifies position relative to the road network as visually perceived by the vehicle driver.

Most of the dead-reckoning equipment commercially available is sufficiently robust to support map matching when operating in a defined road network. However, a good dead-reckoning accuracy is required to achieve reinitialization through map matching upon returning to the road network after off-road operations.

**Proximity Beacon**

This approach uses strategically located short-range transmitters, and reception of their location coded signal infers the receiving vehicle’s instantaneous location. There are several variations of the proximity approach; some versions involve two-way communications with the equipped vehicle. Typically, the driver enters the destination code on the vehicle panel, for automatic transmission to the roadside unit, as the vehicle approaches an instrumented intersection. The roadside unit, which can be networked with a traffic management system, analyzes the destination code and transmits route instructions to the display on the vehicle panel. Proximity beacon systems are being tested in Germany and Japan. One of the most popular system is the ALI-SCOUT (see references) proximity beacon system, which uses dead-reckoning and map matching techniques between beacons to download updated map and traffic data in Berlin.

The approach to the interface between an on-board navigation system and a vehicle operator must take into account ergonomics and safety considerations as well as functional requirements. As a result of intensive research, especially in the aerospace industry, display of information for the operator is a well-developed area. In a well-known European system, Philips’ CARIN, a color CRT map display is used to show vehicle location relative to the surroundings. Many other systems use short visual messages, symbolic graphics, and voice.

© 1999 by CRC Press LLC
Major potential roles for data communications in future automobile navigation need to provide current updates (road additions, closures, detours, etc.) for on-board map databases, and also provide real-time information on traffic conditions for on-board route generation.

**Aircraft**

The primary navigation aid for civil aircraft flying in the airspace of most of the developed countries is the VOR/DME system. The VOR (Very high frequency Omni Range) and the DME (Distance Measuring Equipment) enable on-board determination of an aircraft’s bearing relative to North at the fixed ground station and slant range from the station, respectively. Further details can be found in references given at the end of this chapter.

Many commercial aircraft are equipped with precision Inertial Navigation Systems (INS) not only for navigation purposes, but also for stabilization of the aircraft at all times. Suitably located gyroscopes and accelerometers give signals to control the stability of the aircraft, as shown in Figure 10.25. Many aircraft INS utilize a gyro-stabilized platform on which the accelerometers are mounted. The platform is aligned before take-off to the desired orientation. Due to alignment errors and in-flight random disturbances such as gyro drift, scale factor errors, and accelerometer bias errors, the errors in the desired orientation increase with time. These accumulated errors need to be compensated periodically by external information such as Omega navigation systems.

Omega is a 10 kHz to 14 kHz navigation system that was primarily intended for updating submarine inertial navigators because of its very low frequencies penetrating beneath the ocean surface. It is also used by many general-aviation and military aircraft because of its worldwide coverage. Some airlines (e.g., American Airlines) have equipped their entire fleet with Omega receivers. The U.S. Coast Guard maintains two Omega stations and other countries maintain six more. Each station transmits eight consecutive sinusoidal tones with a 10 s repetition rate. Four are navigation tones common to all stations;
the other four tones uniquely identify the station. Each station has a cesium clock that is calibrated within 1 μs by exchanging portable atomic clocks with other Omega stations and with the U.S. Naval Observatory.

GPS systems give all vehicles on or near the Earth unprecedented navigation accuracy. A number of international airlines are equipped with confined GPS-Glonass receivers. Many experiments are now in progress to balance the cost versus accuracy of various combinations of inertial, Omega, Loran, GPS, and Transit equipment. Airborne receivers are designed that combine navigational equipment operating in a common radio band (e.g., GPS, DME, JTID, and IFF).

INMARSAT’s “Marec” communication satellites serve ship traffic but are configured to serve air traffic by directly communicating with aircraft for approximately $10 per call.

**Underwater**

The Ship's Inertial Navigational System (SINS) was originally developed for precision position-finding required by ballistic missile submarines in the late 1950s and early 1960s. The first deployment was on-board U.S. George Washington in 1960, and SINS are used today in submarines, aircraft carriers, and other surface warships. As the cost and size are continually decreasing, the system is also deployed in naval as well as merchant vessels. Another development of INS for underwater application is in the area of the autonomous underwater vehicle (AUV). In this section, a number of such products are described.

AUVs are used extensively for military and civilian purposes. Application examples are mapping, surveillance, ocean exploration, survey, and mining, all of which require precise position determination. The desired features of such systems are: low power, high accuracy, small volume, light weight, and low cost. Two typical examples are the LN family produced by Litton Guidance and Control Systems, and the system developed by the Harbor Branch Oceanographic Institution Inc. (HBOI). The specifications of some of these systems are briefly described below to give examples of underwater INS.

**The Litton LN-100 System**

Litton's LN-100 is an example of the strapdown INS. The LN-100 system consists of three Litton Zero-Lock Gyros (ZLG), a Litton A4 accelerometer triad, power supply, supporting electronics, and a JIWAG standard 80960 computer. The single-board computer performs all the control, navigation, and interface functions.

**The HBOI System**

The HBOI system was developed with Kearfott Guidance and Navigation (KGN) and utilizes a Monolithic Ring Laser Gyroscope (MRLG), motion sensors, GPS input, and control unit. The inertial measurement unit is based on the Kearfott’s T16-B three-axis ring laser gyro and three accelerometers.

**Robotics**

Closely related to the autonomous underwater vehicles, autonomous mobile robots also use INS extensively as a self-contained, independent navigation system. Typical applications are mining, unknown terrain exploration, and off-line path planning. There are many commercially available inertial navigation systems suitable for cost-effective utilization in the navigation of robots. Some of these are: gyrocompasses, rate gyros, gyrochip, piezoelectric vibrating gyros, ring laser gyros, interferometric, and other types of fiber-optic gyros. Three popular systems will be explained here.

**The Honeywell Modular Azimuth Position System (MAPS)**

Honeywell’s H-726 Modular Azimuth Position System (MAPS) is a typical example of an inertial navigation system for land-based vehicles. It consists of a Dynamic Reference Unit (DRU) that provides processed information from the inertial sensor assembly, a Control Display Unit (CDU) that is used for human-machine interface, and a Vehicle Motion Sensor (VMS) that monitors the vehicle’s directional and distance information. The inertial sensor assembly comprises three Honeywell GG1342 ring-laser
gyros and three Sundstrand QA2000 accelerometers mounted to measure movements in three local axes. The inertial processor translates the information to the navigation processor that resolves the vehicle movement information from the VMS. The CDU provides mode selection, data display, waypoint information, and general operator interface.

**Hitachi Fiber-Optic Gyroscopes**

The Hitachi products are significant as they are relatively inexpensive and were designed for automotive applications. The open-loop Interferometric Fiber-Optic Gyros (IFOG) HOFG-4FT received the “Most Technologically Significant New Products of the Year” award in 1993, and is now installed in one of the production models from Toyota. The subsequent models of IFOG are HOFG-X, HOFG-1, and HGA-D. The HOFG-1 has been employed extensively in industrial mobile robots. The output of the system can be in serial form in RS-232 standard or as analog signals. Specifications of HOFG-1 include a range of ±60° s⁻¹, an update rate of 15 ms, and linearity of ±0.5%. The power requirement is 10 to 16 V dc and 0.5 A.

**References**


**Appendix — List of Manufacturers/Suppliers**

<table>
<thead>
<tr>
<th>Company Name</th>
<th>Address</th>
<th>Telephone</th>
<th>Fax</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerodata Co.</td>
<td>5550 Merric Rd. Suit 205 Massapequa, NY</td>
<td>Tel: (616) 798-1873</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGV Products, Inc.</td>
<td>9307-E Monroe Rd. Charlotte, NC 28270</td>
<td>Tel: (704) 825-1110 Fax: (704) 825-1111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American GNC Corporation</td>
<td>9131 Mason Avenue Chatsworth, CA 91311</td>
<td>Tel: (818) 407-0092 Fax: (818) 407-0093 e-mail: <a href="mailto:agnc@kinseyb.com">agnc@kinseyb.com</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Astronautics Co.</td>
<td>P.O. Box 523 518 W Cherry St. Milwaukee, WI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cybermotion, Inc.</td>
<td>115 Sheraton Drive Salem, VA 24153</td>
<td>Tel: (703) 562-7626 Fax: (703) 562-7632</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First State Map Co.</td>
<td>12 Marry Ella Dr. Wilmington, CT 19805</td>
<td>Tel: (800) 327-7992</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hitachi Cable America, Inc.</td>
<td>50 Main Street, White Plains, NY 10606-1920</td>
<td>Tel: (914) 993-0990 Fax: (914) 993-0997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Honeywell, Inc.</td>
<td>Technology Center 3660 Technology Drive Minneapolis, MN 55418</td>
<td>Tel: (612) 951-7715 Fax: (612) 951-7438</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ketema Inc.</td>
<td>790-T Greenfield Dr. P.O. Box 666 El Cajon, CA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NASA Goddard Space Flight Center Robotics Branch, Code 714.1 Greenbelt, MD 20771</td>
<td>Tel: (301) 286-4031 Fax: (301) 286-1613</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naval Command Control Center</td>
<td>RDT&amp;E Division 5303 San Diego, CA 92152-7383</td>
<td>Tel: (619) 553-3672 Fax: (619) 553-6188</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Navigation Science Co.</td>
<td>31127 Via Clnias Suite 807 Westlake Village, CA 91326</td>
<td>Tel: (818) 991-9794 Fax: (818) 991-9896</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Navigations Technologies Corp.</td>
<td>740 Arques Avenue Sunnyvale, CA 94086</td>
<td>Tel: (408) 737-3200 Fax: (408) 737-3280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naval Command Control Center</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NASA Goddard Space Flight Center</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
10.4 Satellite Navigation and Radiolocation

Halit Eren and C. C. Fung

Modern electronic navigation systems can be classified by range, scope, error, and cost. The range classifications are short, medium, and long ranges, within which exact limits are rather indefinite. The scope classifications can be either self-contained or externally supported, and active (transmitting) or passive (not transmitting) mode of operation.

Short-range systems include radiobeacons, radar, and Decca. Medium-range systems include Decca and certain types of extended-range radars. The long-range systems include Loran-C, Consol, and Omega. All these systems depend on active radio frequency (RF) transmissions, and all are externally supported with respect to the object being navigated, with the exception of the radar. In addition to these, there is another category of systems which are called advanced navigation systems; the transit satellite navigation systems, Glonass, and the Global Positioning Systems (GPS) are typical examples.

Utilization of electromagnetic radio waves is common to all navigation systems discussed here. Understanding of their behavior in the Earth’s atmosphere is very important in the design, construction, and use of all kinds of navigation equipment — from satellites to simple hand-held receivers.

When an FM radio wave is generated within the Earth’s atmosphere, the wave travels outward. The waves may be absorbed or reflected from surfaces of materials they encounter. The absorption and scattering of electromagnetic waves take place for many reasons, one of which is caused by excitation of electrons within the molecules in the propagation media. The behavior of an electromagnetic wave is dependent on its frequency and corresponding wavelength. Figure 10.26 shows the frequency spectrum of electromagnetic waves. They are classified as audible waves at the lower end of the spectrum, radio waves from 5 kHz to 300 GHz, and visible light and various other types of rays at the upper end of the spectrum.

For practical purposes, the radio wave spectrum is broken into eight bands of frequencies; these are: very low frequency (VFL) less than 30 kHz, low frequency (LF) 30 kHz to 300 kHz, medium frequency (MF) 300 kHz to 3 MHz, high frequency (HF) 3 MHz to 30 MHz, very high frequency (VHF) 30 MHz to 300 MHz, ultra high frequency (UHF) 300 MHz to 3 GHz, super high frequency (SHF) 3 GHz to 30 GHz, and extremely high frequency (EHF) 30 GHz to 300 GHz.

For easy identification, the frequencies above 1 GHz are further broken down by letter designators, as: L-band (1–2 GHz), S-band (2–4 GHz), C-band (4–8 GHz), X-band (8–12.5 GHz), and K-band (12.5–40 GHz). Since severe absorption of radar waves occurs near the resonant frequency of water vapor at 22.2 GHz, the K-band is subdivided into lower K-band (12.5–18 GHz) and upper K-band (26.5–40 GHz). Most navigation radars operate in the X- and S-bands, and many weapons fire control radars operate in the K-band range.
The radio waves are transmitted as continuous or modulated waves. A carrier wave (CW) is modulated to convey information in three basic forms: amplitude, frequency, and pulse modulation, as shown in Figure 10.27. The amplitude modulation (AM) modifies the amplitude of the carrier wave with a modulating signal. In frequency modulation (FM), the frequency of the carrier wave is altered in accordance with the frequency of the modulating wave. FM is used in commercial radio broadcasts and the sound

FIGURE 10.26 Electromagnetic wave frequency spectrum. Audible range can be heard if converted to sound waves. Radiobeacons operate in the VLF, LF, and MF ranges. Omega operating at VLF covers the entire world with only eight transmission stations. GPS, Transit, and Glonass use UHF frequencies. Wavelengths less than 10 cm are not suitable for satellite systems, but they are used in radars.

FIGURE 10.27 Amplitude, frequency, and pulse modulation of RF carrier waves. Amplitude modulation is suitable for broadcasting radio stations. Frequency modulation is used in commercial radio broadcasts. The pulse modulation is used in satellite systems, radars, and long-range navigation aids.
portion of the television broadcast. Pulse modulation is different from AM and FM in that there is usually no impressed modulation wave employed. In this modulation, the continuous wave is broken up into very short bursts or "pulses," separated by periods of silence during which no wave is transmitted. This is used in satellite navigation systems, surface search radars, and long-range radio navigation aids such as Loran.

When an electromagnetic wave encounters an obstruction, diffraction takes place marked by a weak reception of the signal within the "shadow" zone. Two waves acting on the same point will also result in interference. The degree of interference depends on the phase and frequency relationship. For example, two waves of the same frequency with a 180° phase difference will result in a null at that point. Also, under certain conditions, a portion of the electromagnetic energy in radio waves may reflect back toward the Earth's surface to form the ionosphere. The ionosphere is a layer of charged particles located about 90 to 400 km high from Earth's surface; such reflected waves are called sky waves.

In the study of radio wave propagation, there are four ionosphere layers of importance, as shown in Figure 10.28. The D-layer is located about 60 km to 90 km and is formed during daylight. The E-layer is about 110 km. It persists through the night with decreased intensity. The F1-layer is between 175 km and 200 km; it occurs only during daylight. The F2-layer is between 250 km and 400 km; its strength is greatest in the day but it combines with the F1-layer later to form a weak F-layer after dark. The layers in the ionosphere are variable, with the pattern seeming to have diurnal, seasonal, and sun spot periods. The layers may be highly conductive or may entirely hinder transmissions, depending on the frequency of the wave, its angle of incidence, height, and intensity on various layers at the time of transmission. In general, frequencies in the MF and HF bands are most suitable for ionosphere reflections during both day and night.
Because of the higher resistance of the Earth’s crust as compared to the atmosphere, the lower portions of radio waves parallel to the Earth’s surface are slowed down, causing the waves to bend toward Earth. A wave of this type is termed a ground wave. The ground waves exist because they use the Earth’s surface as a conductor. They occur at low frequency since LF causes more bending in conformity to Earth’s shape. The ultimate range of such ground waves depends on the absorption effects. Sometimes, in the lower atmosphere, surface ducting occurs by multiple hopping, thus extending the range of a ground wave well beyond its normal limits. It is associated with higher radio and radar frequencies. This phenomenon is common in tropical latitudes. Behavior patterns of waves transmitted at various angles are illustrated in Figure 10.28.

Accuracy of Electronic Fix

There are a number of random effects that influence the accuracy of an electronic position determination; atmospheric disturbances along the transmission path, errors in transmitters and receivers, clocks, inaccuracy in electronic circuitry, gyro errors, etc. As a result, a series of positions determined at a given time and location usually results in a cluster of points near the true position. There are two measures commonly used to describe the accuracy: the first is the circular error probable (CEP) — a circle drawn on the true position whose circumference encompasses 50% of all indicated positions, and the second technique, more common, is the root mean square (rms), where:

\[
\text{rms} = \sqrt{\frac{\sum_{n=1}^{N} (E_n)^2}{N}}
\]

(10.105)

where \( E \) = the distance between actual and predicted positions

\( N \) = the number of predicted positions

A circle, shown in Figure 10.29, with one rms value is expected to contain 68% of all the indicated positions. Another circle of radius equal to 2 rms should contain 95% of all the indicated positions, for isotropic scattering, or errors.
In electronic navigation systems, three types of accuracy are important: (1) *predictable or absolute accuracy* — the accuracy of a position with respect to the geographic coordinates of the Earth; (2) *repeatable accuracy* — the accuracy with which the user can return to a position whose coordinates have been determined at a previous time with the same navigation system; and (3) *relative accuracy* — the accuracy with which a user can measure position relative to that of another user of the same system at the same time.

**Radionavigation Systems**

In the 1930s, improved radio direction-finding techniques and better equipment led to the establishment of systems called *radiobeacons*. These systems consisted of small radio transmitters located in fixed places to provide radio bearings that could be used in all weather conditions. Position findings by these beacons became known as *radionavigation*. Continued refinements in the state-of-the-art electronics technology and a better understanding of electromagnetic wave propagation led to the subsequent development of radar and longer-range radionavigation systems.

Essentially, radiobeacons are single transmitters, transmitting a continuous wave at low power, usually modulated by audible Morse code characters for identification. The transmitted signal is received by an on-board receiver incorporating a radio direction finder (RDF) to be processed further.

**Short- and Medium-Range Radiolocation Systems**

Most short- to medium-range navigation systems are designed to provide either a bearing to a fixed transmitter site, as in the case of radiobeacons, or a range and bearing from the transmitter to a natural or manufactured navigation aid, as in the case of radar.

**Long-Range Hyperbolic Navigation Systems**

Long-range electronic navigation systems are based on *hyperbolic systems*. In these systems, the lines of positions yield in segments of hyperbolas rather than as radial lines. The line connecting the master and secondary stations transmitting the same signal simultaneously is called the *baseline*. Hyperbola lines represent the locus of all points of the arrival of master and secondary pulses at specified time differences. This is illustrated in Figure 10.30. Any change in the position of the receiver near the baseline corresponds to a relatively large change in time difference of reception of pulses.

In practice, the secondary stations transmit pulses at fixed intervals, called *coding delays*, after having received a pulse from the master station. Currently, hyperbolic systems employ atomic time standards to regulate both master and secondary station transmissions to increase the system accuracy by eliminating random time errors due to atmospheric conditions.

In some systems, as an alternative to short pulses, continuous waves are propagated with the leading edges of each cycle representing a given time and distance interval. The corresponding hyperbolas then represent loci of constant phase differences, called *isophase lines*. The space between the hyperbolas are referred to as *lanes*. The position of the receiver within a lane is determined by the phase difference between the master and secondary signals. A disadvantage of this system is that it is not possible to distinguish one lane from another by the use of phase comparison alone. Hence, the lanes must be either counted as they are traversed from some fixed position, or they must be identified by some other means. For good accuracy, the user’s distance from the stations can seldom exceed about six times the length of the baseline.

**Radiobeacons**

Most radiobeacon signals are limited to less than 320 km (200 miles), with a majority not receivable beyond about 32 km (20 miles). Often, radiobeacons located in a given area are grouped on a common frequency, such that each transmitter transmits only during a segment of a time-sharing plan. Radio bearings to the site of transmitter are determined by the use of a radio receiver equipped with a Radio Direction Finder (RDF). There are moderately priced, manually operated RDF receivers, and several more expensive fully automated models are also available. As a general rule, RDF bearings are normally
considered accurate only to within ±20°; for example, under 200 km (120 miles) to the transmitter in favorable conditions and ±5° to 10° when conditions are unfavorable.

Information on the locations, ranges, and using procedures of radio beacons are given in a number of publications such as DMAHTC Publication No. 117 *The Radio Navigation Aids.* Correct radiobeacon bearings are plotted on Mercator charts for position estimation. Because of possible inaccuracies, radio-beacons are not used universally. Navigators such as small boats and merchant ships not equipped with other systems use radio beacons.

**Loran-C**

Loran was developed in the 1940s to be one of the first systems implementing a long-range hyperbolic system for both ships and aircraft. The system used master and slave stations transmitting sequential radio waves in the upper MF band with frequencies between 1850 kHz and 1950 kHz. Loran-A featured ground wave coverage out to between 700 km and 1250 km from the baseline by day, and up to 2200 km by night. It was the most widely used electronic navigation system until 1970. Later, a system employing synchronized pulses for both time-difference and phase comparison measurements was developed, known as Loran-C. Loran-C was configured to operate in a *chain* form consisting of more than one slave station usually located in triangles.

All stations in the system transmit a signal on a common carrier frequency in mid-LF band of 100 kHz ± 10 kHz. The ground wave range is about 1900 km. One-hop sky waves have a range of about 3600 km, and two-hop signals were noted to have been received about 6250 km from the ground station. One-hop sky waves are produced both by day and by night, while two-hop sky waves are formed only at night. Present Loran-C chains have baseline distances between 1500 km and 2300 km. The accuracy of the system varies from about ±200 m rms near the baseline to ±600 m rms near the extreme ranges.
of the system. Differential techniques employed in recent years have increased the accuracy. The low frequency and high-power transmitters allow ground waves to penetrate the surface layer of the sea water, enabling submerged submarines to receive them.

To lessen the large power requirements, multipulsed signals are used. The multipulsed signals of each master station and its associated secondary stations are transmitted in a predetermined sequence, as shown in Figure 10.31. Eight of the nine pulses are separated by 1 ms intervals and the ninth one by a 2-ms interval. Integrated master and secondary pulses are compared at a sampling point at exactly 30 μs from their leading edges.

Loran-C receivers are fully automatic, suitable to be employed for marine, land vehicle, and aircraft applications. Most receivers provide direct lat/long digital readout, precise to a tenth of a minute of arc. They also provide auxiliary features such as destination selection, course and speed overground, etc. Once initialized, they automatically select the best chain and the most suitable master/secondary pulses.

There are 13 Loran-C chains worldwide. Each chain uses a different basic pulse repetition rate (PRR) and pulse repetition interval (PRI). The specific PRI used in a given chain is referred to as group repetition interval (GRI), often called rate. This system has enjoyed great expansion since 1970s, attracting many users. It has found applications in ships, aircraft, as well as land vehicles. Land vehicles are equipped with automatic vehicle location systems (AVLS). In one application in the U.S., the AVLS system is integrated with the city’s emergency telephone system, a computer-aided dispatch system to control the position of emergency vehicles such as ambulances and fire trucks.

Nevertheless, the progress made in satellite-based global positioning systems (GPS) calls for termination of Loran-C in the very near future. However, complete system shut-down may not occur immediately, due to the vast number of current users of the system.

Decca

Decca is similar to Loran-C in that each chain is composed of one master and three slave stations arranged in star pattern, at an angle of about 120° between the baselines. It uses unmodulated continuous waves rather than the pulsed waves of Loran. The characteristic hyperbolic grid pattern is formed by phase comparisons of master and slave signals. All stations transmit in the LF-band between 70 kHz and 130 kHz. The nominal range is about 400 km both by day and by night. The system is extremely accurate within the operating range. The signals transmitted by each of the stations are all harmonics of a single fundamental frequency.

There are a wide variety of receivers available for Decca system, including automatic flight logs for aircraft. In general, the systems consist of four separate receivers, each of which can be set to receive one of the four signals transmitted by a given chain. The lane identification is accomplished by a signal transmitted by each master and slave station, transmitted once every 20 s for a duration of 0.6 s. By some
short signal interruptions and transmissions of harmonics simultaneously, zones and lanes are identified in a precise manner.

**Consol**

Consol is limited to the Eastern and Northern Atlantics. It is a hyperbolic system with extremely short baseline lengths, such that a collapsed hyperbolic pattern is formed. It employs three towers located three wavelengths apart. The operational frequencies are in the MF range between 250 kHz and 370 kHz. The system range is about 1900 km by day and 2400 km by night. The minimum range is about 40 km near the stations. One tower transmits a continuous wave, while other towers transmit waves with 180° phase shift by a keying cycle. The signals are modulated by dots and dashes such that receivers determine the position by counting them and printing on Consol grid patterns.

**Omega**

Omega is a hyperbolic navigation system that covers the entire world with only eight transmission stations located 7500 km to 9500 km apart. It transmits on frequencies in the VLF band from 10 kHz to 14 kHz at a power of 10 kW. The signals of at least three and usually four stations can be received at any position on Earth.

The 10 kHz to 14 kHz frequency band was chosen specifically to take advantage of several favorable propagation characteristics, such as: (1) to use the Earth's surface and ionosphere as a waveguide; (2) to enable submerged submarines to receive the signals; and (3) to form long baselines at 7500 km to 9500 km.

The basic frequency at which all eight stations transmit is 10.2 kHz. Each station transmits four navigation signals as well as a timing signal with atomic frequency standards ensuring that all stations are kept exactly in phase. Two continuous waves are in phase but traveling in opposite directions to produce a series of Omega lanes. Within each lane, a phase difference measurement would progress from 0° to 360° as the receiver moves across, as shown in Figure 10.32. Two Omega lanes complete one cycle, giving a wavelength of 25 km and lane of 12 km expanding as the distance from the baseline increases. Lanes are identified by three other signals transmitted by each station on a multiplexed basis. Omega

![Figure 10.32](image)

**Figure 10.32** Three successive phase comparisons for lane resolution in Omega systems. Phase differences are compared in three stages with respect to three different signals transmitted by each station for accurate position finding. One wavelength is 25 km, representing two lanes. Accuracy of the Omega system is limited.
fixes have been accurate to within ±1.5 km rms by day and ±3 km rms by night. Differential techniques can greatly reduce this error.

Omega receivers are fully automated sets that provide direct lat/long readout for a cost of around U.S. $1,000 for marine and aviation use. Nevertheless, since the precision of the system is not as good as others, they are mainly used as back-up systems.

**Radar**

The word is derived from *radio detection* and *ranging*. It works on the basic principle of reflection, which determines the time required for an RF pulse to travel from a reference source to a target and return as an echo. Most surface search and navigation radar high-frequency electromagnetic waves are formed by a parabolic antenna into a beam form, as shown in Figure 10.33. The receiving antenna is rotated to scan the entire surrounding area, and the bearings to the target are determined by the orientation of the antenna at the moment the echo returns. A standard radar is made up of five components: transmitter, modulator, antenna, receiver, and indicator. They operate on pulse modulation.

Radars are extremely important devices for air control applications. Nowadays, airborne beacon radar systems are well developed in traffic alert and collision avoidance systems (TCAS). In this system, each plane constantly emits an interrogation signal, which is received by all nearby aircraft that are equipped appropriately. The signal triggers a transponder in the target aircraft, which then transmits some information concerning 3-D location and identification.

**Satellite Relay Systems**

The use of satellites is a highly developed technology utilized extensively throughout the world. In the past 2 decades, it has progressed from quasi-experimental in nature to one with routine provisions of new services. They take advantage of the unique characteristics of *geostationary satellite orbits* (GSO). The design of satellite systems is well understood, but the technology is still dynamic. The satellites are useful for long-distance communication services, for services across oceans or difficult terrain, and point-to-multipoint services such as television distribution.

Frequency allocation for satellites is controlled by the International Telecommunication Union (ITU). In the U.S., the Federal Communications Commission (FCC) makes the frequency allocations and assignments for nongovernment satellite usage. The FCC imposes a number of conditions regarding construction and maintenance of in-orbit satellites.

There are many satellite systems operated by different organizations and different countries mainly developed for communications and data transmissions; these include: Iridium of Motorola, Globalstar of Loral Corporation, Intelsat, CS-series of Japan, Turksat of Turkey, Aussat of Australia, Galaxy and Satcom of the U.S., Anik of Canada, TDF of France, etc. Some of the communication satellite systems are suitable for navigation purposes. However, satellite systems specifically designed for navigation are limited in number. The most established and readily accessible by civilian and commercial users are the GPS system of the U.S. and the Glonass of Russia.

The first generation of the satellite system was the *Navy satellite system* (Navsat), which became operational in January 1964, following the successful launch of the first transit satellite into polar orbit.
The system was declared open for private and commercial use in 1967. Civil designation of the name of the system is Transit Navigation Satellite System, or simply Transit. Later, this system evolved to become the modern Navsat GPS system, which will be discussed in detail. Most of the operational principles discussed here are inherited by the GPS system.

The Transit system consists of operational satellites, plus several orbiting spares, a network of ground tracking stations, a computing center, an injection station, naval observatory time signals, and receiver-computer combinations. The transit satellites are in circular polar orbits about 1075 km above ground with periods of revolution of about 107 min. Because of the rotation of the Earth beneath the satellites, every position on Earth comes within range of each satellite at least twice a day, at 12 h intervals. As originally intended, if at least five satellites are operational at any given time, the average time between fix opportunities would vary from about 95 min near the equator to about 35 min or less above 70° North and South.

The Transit system is based on the Doppler shift of two frequencies, 150 MHz and 400 MHz, transmitted simultaneously by each satellite moving its orbit at a tangential velocity of about 7.5 km s⁻¹. Two frequencies are used so that the effects of the ionosphere and atmospheric refraction on the incoming satellite transmission can be compensated for by the receivers. Each frequency is modulated by a repeating data signal lasting 2 min, conveying the current satellite time and its orbital parameters and other information. Within the receiver, a single Doppler signal is created by processing the two signals transmitted by the satellite. By plotting the frequency of this signal versus time, a characteristic curve of the type shown in Figure 10.34 is obtained. Since the frequency of the transmitted signal is compressed as the satellite approaches, according to what proportion of the velocity vector is seen by the user receiver, the curve begins at time $T_1$ at a frequency several cycles higher than the transmitted frequency.

Tracking stations record Doppler observations and memory readout received during each satellite pass to relay them to a computer center. Updated orbital position and time data communications are relayed to an "injection" station from the computer center for transmission to satellite in a burst once each 12 h. Enough data is supplied in this 15 s injection message to last for 16 h of consecutive 2 min broadcasts describing the current orbital positions of the satellite.

The system accuracy depends on the accuracy of the satellite orbit computation, the effect of ionosphere refraction, the precision of the receiver speed, and heading determination. Under optimal conditions, the system is capable of producing fixes with a maximum rms error of about 35 m for the stationary receivers anywhere on Earth. Nevertheless, if a site is occupied for several weeks, an accuracy better than 1 m can be achieved. The time signal transmitted as a "beep" at the end of each 2 min transmission cycle coincides with even minutes of Coordinated Universal Time, which can be used as a chronometer check.
There are other satellite systems, either already in existence or in the planning stages, suitable for navigation. Some of these are: Marec satellites operating at VHF and owned by the intergovernment consortium INMARSAT; privately owned Geostar provides services for oil industry; and many other systems offering transcontinental communication and navigation services as well as position sensing; examples include: SATCOM, ARINC’s, Avsat, Starfix, etc.

**Transponders**

Transponders are transducers that respond to incoming signals by generating appropriate reply messages. Recent developments in technology have made the configuration of transponders possible using elaborate and powerful on-board signal processing. This enhanced the capacity by improving the link budgets, by adjusting the antenna patterns and by making the satellite resources available on a demand basis — called the “switch board in the sky concept.”

Increased interest in deep sea exploration has brought acoustic transponders to the forefront as an important navigation tool. They provide three-dimensional position information for subservience vehicles and devices.

Some transponders are based on radar signals that respond to radar illumination. Transponders are programmed to identify friend or foe or, in some cases, simply inform ground stations about the position of aircraft.

Transponders are used for emergency warning. The U.S. and Canadian satellites carry Sarsat transponders, and Russian satellites carry Cospas transponders. They are also used as warning devices in collision avoidance systems in aircraft and land vehicles.

**Global Satellite Navigation Systems**

**The GPS System**

The Global Satellite Navigation Systems are second-generation satellites evolved primarily from the Naval Global Positioning System. They provide a continuous three-dimensional position-finding capability (i.e., latitude, longitude, and altitude), in contrast to the periodic two-dimensional information of the Transit system. Twenty-four operational satellites, as shown in Figure 10.35, constitute the system. Each satellite orbit is circular, about 2200 km high, and inclined at angles of 55° with respect to Earth’s axis.

The position determination using the GPS system is based on the ability of the receivers to accurately determine the distance to the GPS satellites above the user’s horizon at the time of fix. If accurate distances of two such satellites and the heights are known, then the position can be determined. In order to do this, the receiver would need to know the exact time at which the signal was broadcast and the exact time that it was received. If the propagation speeds through the atmosphere are known, the resulting range can be calculated. The measured ranges are called *pseudoranges*. Nowadays, normally, information is received from at least four satellites, leading to accurate calculations of the fix. The time errors plus propagation speed errors result in range errors, common to all GPS receivers. Time is the fourth parameter evaluated by the receiver if at least four satellites can be received at a given time. If a fifth satellite is received, an error matrix can be evaluated additionally.

Each GPS satellite broadcasts simultaneously on two frequencies for the determination and elimination of ionosphere and other atmospheric effects. The Navstar frequencies are at 1575.42 MHz and 1227.6 MHz, designated as L1 and L2 in the L-band of the UHF range. Both signals are modulated by 30 s navigation messages transmitted at 50 bits s⁻¹. The first 18 s of each 30 s frame contain *ephemeris* data for that particular satellite, which defines the position of the satellite as a function of time. The remaining 12 s is the *almanac* data, which define orbits and operational status of all satellites in the system. The GPS receivers store and use the ephemeris data to determine the pseudorange, and the almanac data to help determine the four best satellites to use for positional data at any given time. However, the “best four” philosophy has been overtaken slowly by an all-in-view philosophy.
The L1 and L2 satellite navigation signals are also modulated by two additional binary sequences called the C/A code for acquisition of coarse navigation and the other P-code for precision ranging. The L1 signal is modulated both by the C/A and P-codes, and the L2 only by the P-code. Positional accuracies of about 20 m rms are usual in using C/A codes alone. The P-code, however, is not available for civilian users. The P-code is redesignated to be a Y-code, decipherable only by high-precision receivers having access to encrypted information in the satellite message. Nevertheless, it is fair to comment that civilians have figured out how to benefit the P/Y signals without actually knowing the codes, but at lower SNR. Further, C/A codes are degraded by insertion of random errors such that positional accuracy is limited to 50 m rms for horizontal values and 70 m for vertical values. These errors are intended to be lifted by the year 2006. Civilian users have access to the so-called Standard Positioning Services (SPS) accurate to 50 m rms, while U.S. and NATO military users will use Precise Positioning Service (PPS).

In enhancing SPS accuracy, differential techniques may be applied, as shown in Figure 10.36, to the encrypted GPS signals. Since the reference receiver is at a known location, it can calculate the correct ranges of pseudoranges at any time. The differences in the measured and calculated pseudoranges give the correction factors. Accuracy less than 1 m can be obtained in the stationary and moving measurements. Recently, differential receivers became commonly available, giving higher accuracy in sub-centimeter ranges. They are getting cheaper day by day and finding applications in many areas such as airplanes, common transport vehicles, cars, geological surveying, orienteering, farming, etc.

There are currently three basic types of GPS receivers designed and built to address various user communities. These are called slow sequencing, fast sequencing, and continuous tracking receivers. The least complicated and lowest cost receiver for most applications is the slow sequencing type, wherein only one measurement channel is used to receive sequential L1 C/A code from each satellite every 1.2 s,
Fast sequencing receivers have two channels: one for making continuous pseudorange measurements, and the other collection of the ephemeris and almanac data. This type is used in medium dynamic applications such as ground vehicles.

Continuous tracking receivers employ multiple channels (at least five) to track, compute, and process the pseudoranges to the various satellites being utilized simultaneously, thus obtaining the highest possible degree of accuracy and making it suitable for high dynamic applications such as aircraft and missiles. The parallel channel receivers are so cost effective nowadays that other types are likely to disappear.

A number of companies produce highly sophisticated GPS receivers. EURONAV GPS receivers operate on two and five channels for military applications. They provide features such as precise time, interfacing with digital flight instruments, RS-422 interface, altimeter input, self initialization, etc.

Software implementation satellite management functions, having different features, are offered by many manufacturers. In the case of DoD NAVSTAR GPS receivers, for example, three functional requirements are implemented: (1) database management of satellite almanac, ephemeris, and deterministic correction data; (2) computation of precise satellite position and velocity for use by navigation software; and (3) using satellite and receiver position data to periodically calculate the constellation of four satellites with optimum geometry for navigation. The DoD receivers are divided for three functions as Satellite Manager (SM), Satellite-Data-Base-Manager (SDBM) SV-Position Velocity Acceleration (SVPVA), and Select-Satellites (SS).

Differential navigation is also applied where one user set is navigating relative to another user set via a data link. In some cases, one user has been at a destination at some prior time and is navigating relative to coordinates measured at that point. The true values of this receiver’s navigation fix are compared against the measured values, and the differences become the differential corrections. These corrections are transmitted to area user sets in real-time, or they may be recorded for post-mission use so that position fixes are free of GPS-related biases.

Differential navigation and GPS systems find applications in enroute navigations for commercial and civil aviation, military application, navigation of ships especially in shallow waters, in station keeping of aircraft, seismic geophysical explorations, land surveying, transport vehicles and traffic controls, etc.
The **Glonass**

There are a number of other satellite navigation systems similar to GPS of the U.S., such as Russian **Glonass**. The Glonass consists of 24 satellites orbiting in circular form 1500 km above the ground. The accuracy of the system is about 10 m rms. Glonass satellites transmit details of their own position and a time reference. The carrier frequencies are in L-band, around 1250 MHz (L2) and 1600 MHz (L1). Only the L1 frequency carries the Civil C/A code. The radio frequency carriers used by Glonass are channelized within bands 1240–1260 MHz and 1597–1617 MHz, the channel spacing being 0.4375 MHz at the lower frequencies and 0.5625 MHz at the higher frequencies. The number of channels is 24. Glonass data message is formatted in frames of 3000 bits, with a duration of 30 s. The ephemeris data are transmitted as a set of position, velocity, and acceleration coordinates in a Cartesian Earth-centered, Earth-fixed (ECEF) coordinate system. The new ephemeris data are available every half hour, valid for the following quarter-hour. The data are sent at a 50 baud rate and superimposed on a pseudorandom noise (PRN) code. The low-precision code has length 511 bits as compared to 1023 bits for Navstar. Glonass accuracy is as good as that for the GPS system. Glonass and GPS have different coordinate frames and different time frames that are being coordinated together.

The **Starfix**

Another interesting satellite-based system — privately funded, developed, launched, and maintained — is the Starfix positioning system. This system is designed primarily for oil surveying. The system consists of a master site, which generates satellite ephemeris data, and four satellites in geosynchronous orbits. The system said to have a precision of 2.5 m rms.

**References**


**Appendix — List of Manufacturers/Suppliers**

<table>
<thead>
<tr>
<th>Company Name</th>
<th>Address</th>
<th>Phone</th>
<th>Fax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Videotech Co.</td>
<td>1840 County Line Rd., Dept. G Huntington Valley, PA 19006</td>
<td>(800) 221-8930</td>
<td>(800) 221-8932</td>
</tr>
<tr>
<td>Astroguide</td>
<td>Lasalle, IL 61301</td>
<td>(815) 224-2700</td>
<td>(815) 224-2701</td>
</tr>
<tr>
<td>AlliedSignal</td>
<td>101 Colombia Rd. Dept. CAC Morristown, NJ 07962</td>
<td>(800) 707-4555</td>
<td>(602) 496-1001</td>
</tr>
<tr>
<td>Colombia Elect. Int. Inc.</td>
<td>P.O. Box 960-T Somis, CA 93066</td>
<td>(805) 386-2312</td>
<td>(805) 386-2314</td>
</tr>
<tr>
<td>GE Co.</td>
<td>3135 Easton Tpke. Fairfield, CT 06431</td>
<td>(800) 626-2004</td>
<td>(518) 869-2828</td>
</tr>
<tr>
<td>Orbitron</td>
<td>351-TR-S Peterson St. Spring Green, WI 53588</td>
<td>(608) 588-2923</td>
<td>(608) 588-2257</td>
</tr>
<tr>
<td>American GNC Corp.</td>
<td>9131 Mason Avenue Chatsworth, CA 91311</td>
<td>(818) 407-0092</td>
<td>(818) 407-0093</td>
</tr>
<tr>
<td>Comstream</td>
<td>10180 Barnes Canyon Rd. San Diego, CA 92121</td>
<td>(800) 959-0811</td>
<td>(619) 458-9199</td>
</tr>
<tr>
<td>STI</td>
<td>31069 Genstar Rd. Hayward, CA 94544-7831</td>
<td>(800) 991-4947</td>
<td>(510) 471-9757</td>
</tr>
</tbody>
</table>

© 1999 by CRC Press LLC
10.5 Occupancy Detection

Jacob Fraden

Occupancy sensors detect the presence of people in a monitored area. Motion detectors respond only to moving objects. A distinction between the two is that the occupancy sensors produce signals whenever an object is stationary or not, while the motion detectors are selectively sensitive to moving objects. The applications of these sensors include security, surveillance, energy management (electric lights control), personal safety, friendly home appliances, interactive toys, novelty products, etc. Depending on the applications, the presence of humans may be detected through any means that is associated with some kind of a human body’s property or actions [1]. For example, a detector may be sensitive to body weight, heat, sounds, dielectric constant, etc. The following types of detectors are presently used for the occupancy and motion sensing of people:

1. **Air pressure sensors**: detect changes in air pressure resulting from opening doors and windows
2. **Capacitive detectors**: of human body capacitance
3. **Acoustic detectors**: of sound produced by people
4. **Photoelectric detectors**: interruption of light beams by moving objects
5. **Optoelectric detectors**: detection of variations in illumination or optical contrast in the protected area
6. **Pressure mat switches**: pressure-sensitive long strips used on floors beneath the carpets to detect the weight of an intruder
7. **Stress detectors**: strain gages embedded into floor beams, staircases, and other structural components
8. **Switch sensors**: electrical contacts connected to doors and windows
9. **Magnetic switches**: a noncontact version of switch sensors
10. **Vibration detectors**: react to the vibration of walls or other building structures; may also be attached to doors or windows to detect movements
11. **Glass breakage detectors**: sensors reacting to specific vibrations produced by shattered glass
12. **Infrared motion detectors**: devices sensitive to heat waves emanating from warm or cold moving objects
13. **Microwave detectors**: active sensors responsive to microwave electromagnetic signals reflected from objects
14. **Ultrasonic detectors**: similar to microwaves, except that instead of electromagnetic radiation, ultrasonic waves are used
15. **Video motion detectors**: video equipment that compares a stationary image stored in memory with the current image from the protected area
16. **Laser system detectors**: similar to photoelectric detectors, except that they use narrow light beams and combinations of reflectors
17. **Triboelectric detectors**: sensors capable of detecting static electric charges carried by moving objects

One of the major aggravations in detecting occupancy or intrusion is a false positive detection. The term “false positive” means that the system indicates an intrusion when there is none. In some noncritical applications where false positive detections occur once in a while, for example, in a toy or a motion switch controlling electric lights in a room, this may be not a serious problem: the lights will be erroneously turned on for a short time, which will unlikely do any harm. In other systems, especially those used for security purposes, the false positive detections, while generally not as dangerous as false negative ones (missing an intrusion), may become a serious problem. While selecting a sensor for critical applications, consideration should be given to its reliability, selectivity, and noise immunity. It is often good practice to form a multiple sensor arrangement with symmetrical interface circuits; this can dramatically improve the reliability of a system, especially in the presence of external transmitted noise. Another efficient way to reduce erroneous detections is to use sensors operating on different physical principles [2]; for example, combining capacitive and infrared detectors is an efficient combination as they are receptive to different kinds of transmitted noise.
Ultrasonic Sensors

Ultrasonic detectors are based on transmission to the object and receiving reflected acoustic waves. Ultrasonic waves are mechanical — they cover frequency range well beyond the capabilities of human ears, i.e., over 20 kHz. However, these frequencies may be quite perceptive by smaller animals, like dogs, cats, rodents, and insects. Indeed, the ultrasonic detectors are the biological ranging devices for bats and dolphins.

When the waves are incident on an object, part of their energy is reflected. In many practical cases, the ultrasonic energy is reflected in a diffuse manner. That is, regardless of the direction where the energy comes from, it is reflected almost uniformly within a wide solid angle, which may approach 180°. If an object moves, the frequency of the reflected waves will differ from the transmitted waves. This is called the Doppler effect (see below). To generate any mechanical waves, including ultrasonic, the movement of a surface is required. This movement creates compression and expansion of the medium, which can be a gas (air), a liquid, or a solid. The most common type of the excitation device that can generate surface movement in the ultrasonic range is a piezoelectric transducer operating in the so-called motor mode [3]. The name implies that the piezoelectric device directly converts electrical energy into mechanical energy.

Microwave Motion Detectors

Microwave detectors offer an attractive alternative to other detectors, when it is required to cover large areas and to operate over an extended temperature range under the influence of strong interferences (e.g., wind, acoustic noise, fog, dust, moisture, etc.). The operating principle of the microwave detector is based on radiation of electromagnetic radio frequency (RF) waves toward a protected area. The most common frequencies are 10.525 GHz (X-band) and 24.125 GHz (K-band). These wavelengths are long enough ($\lambda = 3$ cm at X-band) to pass freely through most contaminants, such as airborne dust, and short enough to be reflected by larger objects.

The microwave part of the detector consists of a Gunn oscillator, an antenna, and a mixer diode. The Gunn oscillator is a diode mounted in a small precision cavity that, on application of power, oscillates at microwave frequencies. The oscillator produces electromagnetic waves, part of which is directed through an iris into a waveguide and focusing antenna that directs the radiation toward the object. Focusing characteristics of the antenna are determined by the application. As a general rule, the narrower the directional diagram of the antenna, the more sensitive it is (the antenna has a higher gain). Another general rule is that a narrow beam antenna is much larger, while a wide-angle antenna can be quite small. A typical radiated power of the transmitter is 10 mW to 20 mW.

An antenna transmits the frequency $f_0$, which is defined by the wavelength $\lambda_0$ as:

$$f_0 = \frac{c_0}{\lambda_0}$$  (10.106)

where $c_0$ is the speed of light. When the target moves toward or away from the transmitting antenna, the frequency of the reflected radiation will change. Thus, if the target is moving away with velocity $v$, the reflected frequency will decrease, and it will increase for the approaching targets. This is called the Doppler effect, after the Austrian scientist Christian Johann Doppler (1803–1853). While the effect was first discovered for sound, it is applicable to electromagnetic radiation as well. However, in contrast to sound waves that may propagate with velocities dependent on movement of the source of the sound, electromagnetic waves propagate with speed of light, which is an absolute constant. The frequency of reflected electromagnetic waves can be predicted by the theory of relativity as:

$$f_r = f_0 \frac{\sqrt{1 - \left(\frac{v}{c_0}\right)^2}}{1 + \frac{v}{c_0}}$$  (10.107)
For practical purposes, however, the quantity \((\nu/c_0)^2\) is very small compared with unity; hence, it can be ignored. Then, the equation for the frequency of the reflected waves becomes identical to that for the acoustic waves:

\[
f_r = f_0 \frac{1}{1 + \nu/c_0}
\]

Due to a Doppler effect, the reflected waves have a different frequency \(f_r\). A mixing diode combines the radiated (reference) and reflected frequencies and, being a nonlinear device, produces a signal that contains multiple harmonics of both frequencies.

The Doppler frequency in the mixer can be found from:

\[
\Delta f = f_0 - f_r = f_0 \frac{1}{\nu/c_0 + 1}
\]

and since \(c_0/\nu \gg 1\), the following holds after substituting Equation 10.106:

\[
\Delta f \approx \frac{\nu}{\lambda_0}
\]

Therefore, the signal frequency at the output of the mixer is linearly proportional to the velocity of a moving target. For example, if a person walks toward the detectors with a velocity of 0.6 m s\(^{-1}\), a Doppler frequency for the X-band detector is \(\Delta f = 0.6/0.03 = 20\) Hz.

Equation 10.110 holds true only for movements in the normal direction. When the target moves at angles \(\Theta\) with respect to the detector, the Doppler frequency is:

\[
\Delta f = \frac{\nu}{\lambda_0} \cos \Theta
\]

**Micropower Impulse Radar**

In 1993, Lawrence Livermore National Laboratory developed a micropower impulse radar (MIR), which is a low-cost, noncontact ranging sensor [3]. The operating principle of the MIR is fundamentally the same as a conventional pulse radar system, but with several significant differences. The MIR consists of a noise generator whose output signal triggers a pulse generator. Each pulse has a fixed short duration, while the repetition of these pulses is random, according to triggering by the noise generator. The pulses are spaced randomly with respect to one another in a Gaussian noise-like pattern. It can be said that the pulses have the pulse frequency modulation (PFM) by white noise with maximum index of 20%. In turn, the square-wave pulses cause amplitude modulation (AM) of a radio transmitter. The radio transmitter produces short bursts of high-frequency radio signal that propagate from the transmitting antenna to the surrounding space. The electromagnetic waves reflect from the objects and propagate back to the radar. The same pulse generator that modulates the transmitter, gates (with a predetermined delay) the radio receiver to enable the output of the MIR only during a specific time window. Another reason for gating the receiver is to reduce its power consumption. The reflected pulses are received, demodulated (the square-wave shape is restored from the radio signal), and the time delay with respect to the transmitted pulses is measured. Since the pulses are spaced randomly, practically any number of identical MIR systems can operate in the same space without a frequency division (i.e., they work at the same carrier frequency within the same bandwidth). There is little chance that bursts from the interfering transmitters overlap and, if they do, the interference level is significantly reduced by the averaging circuit.
Capacitive Occupancy Detectors

Being a conductive medium with a high dielectric constant, a human body develops a coupling capacitance to its surroundings. This capacitance greatly depends on such factors as body size, clothing, materials, type of surrounding objects, weather, etc. However wide the coupling range is, the capacitance can vary from a few picofarads to several nanofarads. When a person moves, the coupling capacitance changes, thus making it possible to discriminate static objects from the moving ones. In effect, all objects form some degree of capacitive coupling with respect to one another. If a human (or for that purpose, anything) moves into the vicinity of the objects whose coupling capacitance with each other has been previously established, a new capacitive value arises between the objects as a result of the presence of an intruding body [3]. Figure 10.37 shows that the capacitance between a test plate and Earth is equal to $C_1$.

When a person moves into the vicinity of the plate, it forms two additional capacitors: one between the plate and its own body $C_a$, and the other between the body and the Earth, $C_b$. Then, the resulting capacitance $C$ between the plate and the Earth becomes larger by $\Delta C$.

$$C = C_1 + \Delta C = C_1 + \frac{C_a C_b}{C_a + C_b} \quad (10.112)$$

With the appropriate apparatus, this phenomenon can be used for occupancy detection [3]. What is required is to measure a capacitance between a test plate (the probe) and a reference plate (the Earth). Figure 10.38 illustrates a circuit diagram for detecting variations in the probe capacitance $C_p$ [4]. That capacitance is charged from a reference voltage source $V_{ref}$ through a gate formed by transistor $Q_1$ when the output voltage of a control oscillator goes low. When it goes high, transistor $Q_1$ closes while $Q_2$ opens. The probe capacitance $C_p$ discharges through a constant-current sink constructed with a transistor $Q_3$. A capacitor $C_1$ filters the voltage spikes across the transistor. The average voltage, $e_p$, represents a value of the capacitor $C_p$. When an intruder approaches the probe, the latter’s capacitance increases, which results in a voltage rise across $C_1$. The voltage change passes through the capacitor $C_2$ to the input of a comparator with a fixed threshold $V_T$. The comparator produces the output signal $V_{out}$ when the input voltage exceeds the threshold value.

When a capacitive occupancy (proximity) sensor is used near or on a metal device, its sensitivity can be severely reduced due to capacitive coupling between the electrode and the device’s metallic parts. An effective way to reduce that stray capacitance is to use driven shields [3].

Triboelectric Detectors

Any object can accumulate, on its surface, static electricity. These naturally occurring charges arise from the triboelectric effect; that is, a process of charge separation due to object movements, friction of clothing...
fibers, air turbulence, atmosphere electricity, etc. Usually, air contains either positive or negative ions that can be attracted to the human body, thus changing its charge. Under idealized static conditions, an object is not charged — its bulk charge is equal to zero. In reality, any object that, at least temporarily, is isolated from the ground can exhibit some degree of its bulk charge imbalance. In other words, it becomes a carrier of electric charges that become a source of electric field. The field, or rather its changes, can be detected by an electronic circuit having a special pick-up electrode at its input [3, 5]. The electrode increases the capacitive coupling of the circuit’s input with the environment, very much like in the capacitive detectors described above. The electrode can be fabricated in the form of a conductive surface that is well isolated from the ground.

If a charge carrier (a human or an animal) changes its position — moves away or a new charge carrying an object enters into the vicinity of the electrode — the static electric field is disturbed. The strength of the field depends on the atmospheric conditions and the nature of the objects. For example, a person in dry man-made clothes walking along a carpet carries a million times stronger charge than a wet intruder who has come from the rain.

It should be noted that contrary to a capacitive motion detector, which is an active sensor, a triboelectric detector is passive; that is, it does not generate or transmit any signal. There are several possible sources of interference that can cause spurious detections by the triboelectric detectors. That is, the detector may be subjected to transmitted noise resulting in false positive detection. Among the noise sources are 60 Hz or 50 Hz power line signals, electromagnetic fields generated by radio stations, power electric equipment, lightnings, etc. Most of these interferences generate electric fields that are distributed around the detector quite uniformly and can be compensated for by employing a differential input circuit with a significant common mode rejection ratio.

**Optoelectronic Motion Detectors**

Optoelectronic motion detectors rely on electromagnetic radiation in the optical range, specifically having wavelengths from 0.4 μm to 20 μm. This covers visible, near- and part of far-infrared spectral ranges. The detectors are primarily used for the indication of movement of people and animals. They operate over distance ranges up to several hundred meters and, depending on the particular need, can have either a narrow or wide field of view.
Most of the objects (apart from very hot) radiate electromagnetic waves only in the mid- and far-infrared spectral ranges. Hence, visible and near-infrared light motion detectors must rely on an additional source of light to illuminate the object. The light is reflected by the object’s body toward the focusing device for subsequent detection. Such illumination can be sunlight or the invisible infrared light from an additional near-infrared light source (a projector).

The major application areas for the optoelectronic motion detectors are in security systems (to detect intruders), in energy management (to turn lights on and off), and in the so-called “smart” homes where they can control various appliances such as air conditioners, cooling fans, stereo players, etc. They can also be used in robots, toys, and novelty products. The most important advantage of an optoelectronic motion detector is simplicity and low cost.

**Sensor Structures**

A general structure of an optoelectronic motion detector is shown in Figure 10.39(a). Regardless what kind of sensing element is employed, the following components are essential: a focusing device (a lens or a focusing mirror), a light detecting element, and a threshold comparator. An optoelectronic motion detector resembles a photographic camera. Its focusing components create an image of its field of view on a focal plane. While there is no mechanical shutter like in a camera, in place of the film, a light sensitive element is used. The element converts the focused light into an electric signal. A focusing lens creates an image of the surroundings on a focal plane where the light sensitive element is positioned. If the area is unoccupied, the image is static and the output signal from the element is steady stable. When an “intruder” enters the room and keeps moving, his/her image on the focal plane also moves. In a certain moment, the intruder’s body is displaced for an angle $\alpha$ and the image overlaps with the element. This
is an important point to understand — the detection is produced only at the moment when the object’s image either coincides with the detector’s surface or clears it. That is, no overlapping — no detection. Assuming that the intruder’s body creates an image with a light flux different from that of the static surroundings, the light-sensitive element responds with deflecting voltage $V$. In other words, to cause detection, a moving image will have a certain degree of optical contrast with its surroundings.

Figure 10.39(B) shows that the output signal is compared with two thresholds in the window comparator. The purpose of the comparator is to convert the analog signal $V$ into two logic levels: 0, no motion detected and 1, motion is detected.

To increase area of coverage, an array of detectors can be placed in the focal plane of a focusing mirror or lens. Each individual detector covers a narrow field of view, while in combination they protect larger areas. All detectors in the array should either be multiplexed or otherwise interconnected to produce a combined detection signal; that is, they can be made into a complex sensor shape. An alternative solution is the use of a multiple-element focusing system.

Complex Sensor Shape

If the detector’s surface area is sufficiently large to cover an entire angle of view, it may be optically broken into smaller elements, thus creating an equivalent of a multiple detector array. To break the surface area into several parts, one can shape the sensing element in an odd pattern, like the interdigitized shape shown in Figure 10.40(A) or parallel grid as in Figure 10.40(B). Each part of the sensing element acts as a separate light detector.

The parallel or serially connected detectors generate a combined output signal, for example, voltage $v$, when the image of the object moves along the element surface crossing alternatively sensitive and nonsensitive areas. This results in an alternate signal $v$ at the detector terminals. Each sensitive and nonsensitive area must be sufficiently large to overlap with most of the object’s image. An alternative solution to the complex shape of the sensor is use of the image distortion mask as shown in Figure 10.40(C); however, this solution requires a larger sensor surface area.

Facet Focusing Element

A cost-effective way of broadening the field of view while employing a small-area detector is to use multiple focusing devices. A focusing mirror or a lens may be divided into an array of smaller mirrors or lenses called facets, just like in the eye of a fly. Each facet creates its own image resulting in multiple images, as shown in Figure 10.41. When the object moves, the images also move across the element,
resulting in an alternate signal. By combining multiple facets, it is possible to create any desirable detecting pattern in the field of view, in both horizontal and vertical planes. Positioning of the facet lens, focal distances, number, and a pitch of the facets (a distance between the optical axes of two adjacent facets) can be calculated in every case by applying rules of geometrical optics [3]. In the far-infrared spectral range (thermal radiation sensors), the polyethylene facet Fresnel lenses are used almost exclusively, thanks to their low cost and relatively high efficiency.

For the visible portion of the optical spectrum, a simple, very inexpensive, yet efficient motion detector can be developed for nondemanding applications, like light control or interactive toys, using simple photoresistors and pinhole lenses [3, 6, 7].

**Far-Infrared Motion Detectors**

A motion detector that perceives electromagnetic radiation that is naturally emitted by any object operates in the optical range of thermal radiation, also called far-infrared (FIR). Such detectors are responsive to radiative heat exchange between the sensing element and the moving object. The principle of thermal motion detection is based on the physical theory of emission of electromagnetic radiation from any object whose temperature is above absolute zero (see Chapter 32, Section 6, on *Infrared Thermometers*).

For IR motion detection, it is essential that a surface temperature of an object be different from that of the surrounding objects, so a thermal contrast would exist. All objects emanate thermal radiation from their surfaces and the intensity of that radiation is governed by the Stefan–Boltzmann law. If the object is warmer than the surroundings, its thermal radiation is shifted toward shorter wavelengths and its intensity becomes stronger. Many objects whose movement is to be detected are nonmetals, hence they radiate thermal energy quite uniformly within a hemisphere. Moreover, the dielectric objects generally have a high emissivity. Human skin is one of the best emitters, with emissivity over 90%, while most fabrics also have high emissivities, between 0.74 and 0.95 [3]. Below, two types of far-infrared motion detectors are described. The first utilizes a passive infrared (PIR) sensor, while the second has active far-infrared (AFIR) elements.

**PIR Motion Detectors**

These detectors became very popular for security and energy management systems. The PIR sensing element must be responsive to far-infrared radiation within a spectral range from 4 μm to 20 μm where most of the thermal power emanated by humans is concentrated. There are three types of sensing elements
that are potentially useful for that detector: thermistors, thermopiles, and pyroelectrics; however, pyroelectric elements are used almost exclusively for the motion detection thanks to their simplicity, low cost, high responsivity, and broad dynamic range. A pyroelectric effect is described in Chapter 32, Section 7 on Pyroelectric Thermometers. How this effect can be used in practical sensor design is discussed here.

A pyroelectric material generates an electric charge in response to thermal energy flow through its body. In a simplified way it may be described as a secondary effect of thermal expansion (Figure 10.42). Since all pyroelectrics are also piezoelectrics, the absorbed heat causes the front side of the sensing element to expand. The resulting thermally induced stress leads to the development of a piezoelectric charge on the element electrodes. This charge is manifested as voltage across the electrodes deposited on the opposite sides of the material. Unfortunately, the piezoelectric properties of the element also have a negative effect. If the sensor is subjected to a minute mechanical stress due to any external force, it also generates a charge that in most cases is indistinguishable from that caused by the infrared heat waves. Sources of such mechanical noise are wind, building vibrations, loud sound, etc.

To separate thermally induced charges from the piezoelectrically induced charges, a pyroelectric sensor is usually fabricated in symmetrical form (Figure 10.43). Two identical elements are positioned inside the sensor’s housing. The elements are connected to the electronic circuit in such a manner as to produce the out-of-phase signals when subjected to the same in-phase inputs. The idea is that interferences produced by, for example, the piezoelectric effect or spurious heat signals are applied to both electrodes simultaneously (in phase) and thus will be canceled at the input of the circuit, while the variable thermal radiation to be detected will be absorbed by only one element at a time, thus avoiding a cancellation.

One way to fabricate a differential sensor is to deposit two pairs of electrodes on both sides of a pyroelectric element. Each pair forms a capacitor that can be charged either by heat or by mechanical stress. The electrodes on the upper side of the sensor are connected together forming one continuous electrode, while the two bottom electrodes are separated, thus creating the opposite-serially connected capacitors. Depending on the side where the electrodes are positioned, the output signal will have either a positive or negative polarity for the thermal influx. In some applications, a more complex pattern of the sensing electrodes is required (for example, to form predetermined detection zones), so that more than one pair of electrodes is needed. In such a case, for better rejection of the in-phase signals (common mode rejection), the sensor should still have an even number of pairs where positions of the pairs alternate for better geometrical symmetry. Sometimes, such an alternating connection is called an interdigitized electrode.

A differential sensing element should be mounted in such a way as to ensure that both parts of the sensor generate the same signal if subjected to the same external factors. At any moment, the optical component must focus a thermal image of an object on the surface of one part of the sensor only, which is occupied by a single pair of electrodes. The element generates a charge only across the electrode pair that is subjected to a heat flux. When the thermal image moves from one electrode to another, the current

![Figure 10.42](image-url)
i flowing from the sensing element to the bias resistor $R$ (Figure 10.43(B)) changes from zero, to positive, then to zero, to negative, and again to zero (Figure 10.43(A) lower portion). A JFET transistor Q is used as an impedance converter. The resistor $R$ value must be very high. For example, a typical alternating current generated by the element in response to a moving person is on the order of 1 pA ($10^{-12}$ A). If a desirable output voltage for a specific distance is $v = 50$ mV, then according to Ohm’s law, the resistor value is $R = \frac{v}{i} = 50$ GΩ ($5 \times 10^{10}$ Ω). Such a resistor cannot be directly connected to a regular electronic circuit; hence, transistor Q serves as a voltage follower (the gain is close to unity). Its typical output impedance is on the order of several kilohms.

The output current $i$ from the PIR sensor can be calculated on the basis of the Stefan–Boltzmann law as [3]:

$$i = \frac{2P\sigma a T^4}{\pi \hbar c} b T^3 \frac{\Delta T}{L}$$  \hspace{1cm} (10.113)

where $\Delta T = (T_o - T_i)$ is the temperature gradient between the object and its surroundings, $P$ is the pyroelectric coefficient, $\sigma$ is the Stefan-Boltzmann constant, $a$ is the lens area, $\gamma$ is the lens transmission coefficient, $b$ is the thickness, and $c$ is the specific heat of the pyroelectric element, respectively, and $L$ is the distance to the object.

There are several conclusions that can be drawn from Equation 10.113. The first part of the equation (the first ratio) characterizes a detector, while the rest relates to an object. The pyroelectric current $i$ is directly proportional to the temperature difference (thermal contrast) between the object and its surroundings. It is also proportional to the surface area of the object that faces the detector. A contribution of the ambient temperature $T_i$ is not as strong as it might appear from its third power. The ambient temperature must be entered in kelvin, hence its variations become relatively small with respect to the scale. The thinner the sensing element, the more sensitive the detector. The lens area also directly affects signal magnitude. On the other hand, pyroelectric current does not depend on the sensor’s area as long as the lens focuses an entire image on a sensing element.

AFIR Motion Detectors

The AFIR motion detector is a new class of thermal sensors whose operating principle is based on balancing thermal power supplied to the sensing element [8, 9]. Contrary to a passive motion detector
that absorbs thermal radiation from a warmer object, an AFIR motion detector is active; that is, it radiates heat waves toward the surroundings. The sensor’s surface temperature \( T_s \) is maintained somewhat above ambient. The element is combined with a focusing system, very much like the PIR detector; however, the function of that system is inverse to that of the passive detectors. A focusing part in the AFIR detector projects a thermal image of the warm sensing element into its surroundings. The AFIR sensors have a significant advantage over the PIR: immunity against many interferences (such as RFI and microphonics).

The output voltage from the AFIR motion detector can be described by the following equation [3]:

\[
\Delta V = -\frac{R}{V_0} \frac{\sigma a}{\pi} b T_s^3 \frac{\Delta T}{E}
\]

where \( R \) is the resistance of the sensor’s heater and \( V_0 \) is the heating voltage. The minus sign indicates that for warmer moving objects, the output voltage decreases. There is an obvious similarity between Equations 10.113 and 10.114; however, sensitivity (detection range) of the AFIR sensor can be easily controlled by varying \( R \) or \( V_0 \). For better sensitivity, the temperature increment above ambient can be maintained on a fairly low level. Practically, the element is heated above ambient by only about 0.2° C.

References