1

The development of electronic imaging in astronomy

When viewed far from city lights the star-studded night time sky is an awe-inspiring sight that fires the imagination. Even the earliest civilizations realized that careful astronomical observations were important to survival because such observations enabled them to predict seasonal events, such as when to plant and when to harvest. Observational astronomy was also one of the earliest scientific activities.

1.1 OBSERVATIONAL ASTRONOMY

1.1.1 Historical development

The Greek astronomer Hipparchus (c. 127 BC) used astronomical observations to determine the lengths of the four seasons and the duration of the year to within 6.5 minutes. He also derived the distance to the Moon and the Sun, but his most amazing feat was to notice a small westward drift of the constellations which we now call the “precession of the equinoxes”. This effect causes the current Pole Star (Polaris) to move away from the North point and circle back after almost 26,000 years! Chinese astronomers recorded the appearance and fading of an exceptionally bright star in 1054 AD in the constellation we now call Taurus, but it was not until the twentieth century that Edwin Hubble (1889-1953) associated this event with the supernova explosion which gave rise to the Crab Nebula, also known as Messier 1, the first entry in the list of nebulous objects studied by Charles Messier (1730-1817). Following the invention of the telescope in the early 1600's, Galileo Galilei (1564-1642) and others were finally able to enhance the sensitivity of the only light detector available to them — the human eye — and to resolve details such as craters and mountains on the Moon, the rings of Saturn, moons orbiting Jupiter and the individual stars in the Milky Way. By making careful drawings (Fig. 1.1) of what their eyes could detect during moments of minimum atmospheric turbulence — what astronomers today call moments of "good seeing" — the early 17th century scientists were able to convey pictorially those observations to others.

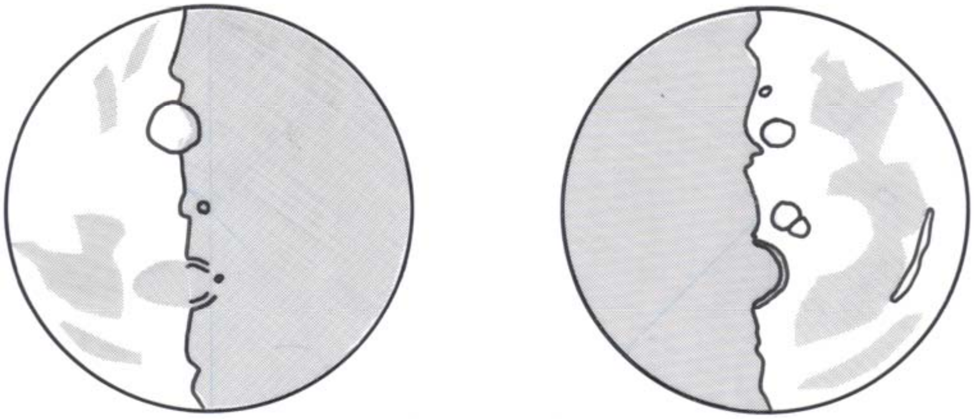


Fig. 1.1 Hand-drawn sketches of features on the surface of the Moon as might have been made by the first astronomers to use telescopes to enhance the power of the eye. Prior to the introduction of photography into astronomy (c.1880), a (subjective) sketch was the only way to preserve a permanent record of the observations.

Better telescopes led to more astronomical discoveries, which in turn stimulated the development of even bigger and better telescopes. Opticians developed color-corrected lenses for telescopes and then, following Isaac Newton (1642-1727), telescopes using reflections from curved mirrors instead of transmission through lenses were gradually introduced. William Herschel (1738-1822), a prolific observer and discoverer of the planet Uranus, pioneered the construction of many reflecting telescopes with long "focal lengths" and large magnifications; in later years the emphasis would move to larger diameter mirrors rather than longer focal lengths. With the invention of the prism spectroscope by Joseph Fraunhofer (1787 - 1826), the chemical constitution of the sun and stars became amenable to physical study. In Fraunhofer's early experiments a beam of sunlight was passed through a narrow rectangular slit in a mask and then through a glass prism to produce a colored spectrum in the manner similar to Newton and others (Fig. 1.2). The critical addition made by Fraunhofer was a small telescope mounted on a moveable arm which could be set to precise angles to view the spectrum. Initially, the light detector was still the human eye. Fraunhofer found that the normal band of colors from violet to red was crossed by numerous dark vertical lines. Eventually the pattern of these Fraunhofer absorption lines (actually images of the entrance slit partially devoid of light) was shown to be characteristic of individual chemical elements. The elements hydrogen, calcium, sodium and iron were recognized in the spectra of the Sun, and later, the stars.

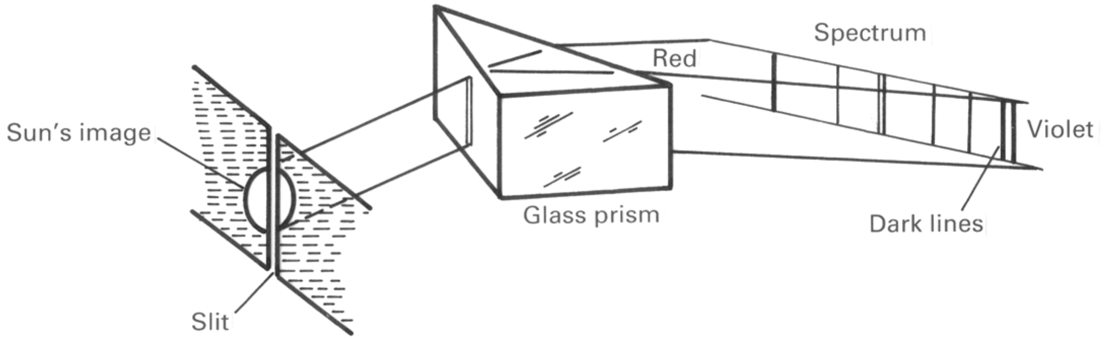


Fig. 1.2 Shown is an illustration of Joseph Fraunhofer's spectroscope and the dark lines in the spectrum of the Sun which now bear his name. This instrument combined with the photographic plate rather than the human eye opened the way to a physical understanding of the universe.

Further spectroscopic observations of the Sun soon led to the discovery by Janssen and Lockyer (in 1868) of an unknown element which we now know to be a major constituent of the universe. This new element, helium, was named after the Greek word for the Sun — *helios*; helium was not discovered on Earth until 1895.

When dry, gelatin-based photographic emulsions became routinely available in the late nineteenth century, astronomers such as Henry Draper (1837-1882) lost no time in putting them to use to catalogue the appearance and properties of a wide range of objects in the night sky. The photographic process was unarguably more accurate and more sensitive than the keenest human eye and the most artistic hand. From planets to stars to galaxies, the new observational tools were applied. Still larger telescopes were constructed, each a technical feat for its era, reaching a mirror diameter of 100 inches (2.54 meters) in 1917 with the completion of the Hooker Telescope on Mount Wilson by George Ellery Hale (1868-1938). Just one of the great discoveries which followed was the expansion of the universe by Edwin Hubble and Milton Humason in 1929.

The history of astronomy is marked by such sporadic progress. Each improvement in scientific apparatus, each new development in technology, helps to provide answers to old questions. Inevitably, the new observational methods uncover a host of new questions, which in turn, drive the quest for even better measuring equipment! Progress in studying the universe has always been related to "deeper" surveys of the cosmos reaching to ever fainter objects, or higher resolution yielding more and more fine detail, or larger statistical samples from which generalizations can be made, or broader spectral response to sample *all* the energy forms passively collected by the Earth. That trend has continued since the Renaissance of the 16th century to the present day in a kind of ever-increasing spiral, with new tools or technologies leading to new discoveries which in turn drive the development of better tools.

A key feature of observational astronomy has been record-keeping, maintaining archives of observations, usually in some pictorial form, for future investigators to compare and consider. In terms of its ability to convert light into a measurable quantity, the photographic plate is actually less sensitive than the human eye. The great advantage of the photographic plate however, is that it can build-up a picture of a faint object by accumulating light on its emulsion for a long period of time. It is therefore called an "integrating" detector. The eye cannot do this to any significant extent. Moreover, the plate provides a permanent record which may be saved for future comparison and study by others.

By using a photographic plate as the recording device in a spectrometer, astronomers could extend their investigations effectively and efficiently into the domain of quantitative astrophysics. Initially, of course, the flood of photographic material was analyzed by human eye, and those eyes were mostly those of a dedicated group of female assistants hired by the director of the Harvard Observatory College, Edward Charles Pickering (1846-1919), toward the end of the last decade of the nineteenth century. Over forty women were employed by the observatory during the period of Pickering's tenure as director, and their efforts in handling the torrent of new astronomical data laid the foundations of modern astrophysics. Stellar spectral classifications led to the understanding that the colors of stars was largely a temperature sequence and that stars shine by the energy released in thermonuclear fusion reactions brought about spontaneously by the enormous temperatures and pressures at their centers. Among the most well-known of the Harvard ladies is Henrietta Leavitt (1868-1921) whose work on the class of stars called Cepheid variables, which pulsate in brightness with a period that is proportional to their true or absolute average brightness, led to a distance-estimator and an appreciation of the true size and shape of our galaxy. During the first half of the 20th century, these tools inevitably resulted in more discoveries (per year) and a massive increase in the "data rate", that is, the amount of information being collected, scrutinized and archived for posterity. But these advances were only the beginning.

Even as the 100-inch Hooker telescope was discovering the expansion of the universe, plans were being laid to build the great 200-inch (5.08-m) reflecting telescope on Mount Palomar in southern California. That telescope, named after George Ellery Hale, went into operation in 1949 and remained the largest telescope in the world until the construction of the Russian (then Soviet) 6-m Bol’shoi Teleskop Azimutal’ny (BTA) in 1976. Construction of both of these large telescopes was challenging. For the 200-inch, Hale secured a grant in 1928 from the Rockefeller foundation, but optical figuring of the Pyrex mirror took from 1936-1947 with four years off for World War II. The telescope was dedicated in June 1948 ten years after Hale’s death, but it was another 16 months before director Ira Bowen (1898-1973) opened the telescope for full time use. Weighing about 1,000 tons, the dome of the Hale telescope stands 41 m (135 ft) high and is 42 m (137 ft) in diameter. Likewise, the BTA on Mount Pastukhov on the northern side of the Caucasus range has a dome that is 58 m high and a primary mirror of Pyrex weighing 42 tons with so much thermal inertia that it can only tolerate a 2 °C change per day if it is to retain its optical figure. Thermal inertia, the large dome and the site turn out to be limitations on the best image quality that can be delivered. In the years that followed astronomers would apply those lessons learned.

Building telescopes larger than 5 meters in diameter was going to be difficult, but observational astronomy received multiple boosts in the 1960's partly by the construction of many new optical observatories with 4-meter class telescopes, that is, with mirror diameters of approximately 3-4 meters. Although the telescopes were slightly smaller, these new facilities were well-equipped and located on excellent but somewhat more remote mountain sites in different parts of the world including the Arizona desert, the mountains of northern Chile, and the summit of Mauna Kea on the Big Island of Hawaii. Another part of the ‘60s expansion was stimulated by the exciting new look at the universe which accompanied the rise of radio astronomy and the discovery of completely new phenomena such as the incredibly luminous and distant quasars—thought to be supermassive black holes at the center of large galaxies—and the remarkable pulsars, now understood to be spinning neutron “stars” embedded in the remnants of a supernova explosion. All of this occurred during the successful development of the Soviet and American space programs which led to satellite astronomy and the opening up of the X-ray, ultraviolet and infrared regions in the sixties and seventies. History shows that the introduction of any new domain results in new discoveries (e.g. Harwit 2003). Other, more subtle, transformations began to occur around this time too through the introduction of electronic computing machines and electronic devices which could be used as detectors of light. Photocells and sensitive "night-vision" TV cameras came first, but the steep rise of consumer micro-electronic products through the seventies was to accelerate the changes rippling through astronomy. Even the telescopes themselves could be improved by the use of electronically-encoded computer-controlled drive systems, thereby enabling much faster set-up times and more reliable tracking across the sky. The newest radio and optical telescopes were remotely controlled, and the concept of converting measurements into an electronic form readily acceptable to a computer became standard practice. Computer power expanded exponentially, and astronomers eagerly used those capabilities to the full.

Construction of larger telescopes stagnated until the mid-1980s when Jerry Nelson of the University of California broke the paradigm by suggesting the concept of a segmented mirror whose shape was controlled by a computer. Around the same time it was also realized that very large thin mirrors with low thermal inertia could be used if computer-controlled force-actuators maintained their shape throughout the night. Consequently, optical telescopes have now reached gigantic proportions with diameters around 10 m (~394 inches) for the twin telescopes of the W. M. Keck Observatory which began operations in 1993 and 1996 respectively. Moreover, there are now telescopes, both on the ground and in space, to cover far more than the visible light our human eyes are designed to see. Today, computers actively control the shape of optical surfaces in the telescope and in associated instruments, performing thousands of calculations per second to correct the image quality. Smaller, highly-automated telescopes survey the entire sky to unprecedented depths and many of these images are immediately available in digital form to all astronomers. This flood of quantitative information is due to strides in the range and sensitivity of electronic detection devices. It is the impact of semiconductor electronic light-sensors attached to the new generation of telescopes (both on the ground and in space) which has had an effect as dramatic as was the introduction of the photographic plate itself over one hundred years ago.

There can be little doubt that we are living in a time of rapid technology development. This is the Digital Age, the age of the "micro-chip". Semiconductor technology, of which the "silicon chip" found in computers is by far the most widely-known example, has touched almost every aspect of our daily lives. The mass production of silicon chips has brought Personal Computers (PCs) of incredible power, at relatively low cost, to almost every environment — homes, schools, offices and industry. The Digital Age is also the age of global electronic communication. There can be few people left who haven't at least heard of the Internet and the World Wide Web! School kids can "down-load" images from the

Hubble Space Telescope web site and "email" messages and pictures to friends half way around the world almost instantaneously by typing at a computer keyboard.

What is a semiconductor? A semiconductor is a crystalline material with some of the properties of a good conductor of electricity (like copper metal), and some of the properties of an electrical insulator (like glass for example). Because of its crystalline (solid-state) structure, a slab of such material behaves the same at all points. Semiconductor crystals can be "grown" in a controlled way from a melt, and moreover, the electrical properties can be tailored by introducing so-called impurity atoms into the crystal structure at the atomic level, so that by microscopic sculpting of the semiconductor material, all sorts of tiny electrical components and circuits can be constructed. The final piece—often not much larger than a thumbnail—is referred to as an integrated circuit or more commonly, as a "chip". Besides silicon, there is germanium, gallium arsenide, indium antimonide, and several other materials with these properties. Semiconductors can be used to manufacture a host of low-power micro-electronic components including amplifiers, all sorts of logic units, computer memory, very complex chips called microprocessors capable of many computational functions, and tiny imaging devices of remarkable sensitivity. Silicon is the most well-developed semiconductor so far, but even for silicon the potential for yet smaller and smaller microchips still exists. Astronomy has benefited in this semiconductor revolution because the apparatus needed for scientific experiments and for complex calculations, which were completely impossible before, are now viable with the aid of the latest electronic imaging devices and powerful high-speed electronic computers.

Almost all modern astronomical research is carried out with photo-electronic equipment, by which we mean instrumentation that converts radiant energy (such as light) into electrical signals which can be digitized, that is, converted into numerical form for immediate storage and manipulation in a computer. Usually highly-automated and remotely-controlled, these instruments, and telescopes to which they are attached, necessarily rely heavily on electronics and computers. Computers play an equally crucial role in helping astronomers assimilate, analyze, model and archive the prodigious quantity of data from the new instruments. The on-going miniaturization of computers and the ever increasing availability of large amounts of relatively cheap computer memory, means that astronomers can employ fairly complex electronic and computer systems at the telescope which speed-up and automate data-gathering. As a result, those astronomical facilities—which may be costly initially—and the data they produce can be available to a much wider range of scientists than would otherwise be possible. Today, a large modern observatory requires an enormous breadth of engineering, scientific and managerial skills to operate efficiently and produce the very best results.

Many readers will be familiar with sources of current and topical astronomical results, whether these are professional journals (e.g. Nature, the Astrophysical Journal) or popular magazines (e.g. Sky & Telescope) or any of the numerous astronomical sites accessible on the World Wide Web. How are such remarkable observations obtained? Most press releases do not describe in detail the apparatus or the technology used in making the discovery. Of course, it would not be easy to do so because of the "jargon barrier" and the complexity of the technology itself. This is unfortunate, because it under-emphasizes an important link between modern technology and the quest for fundamental knowledge embodied in astronomy, a search for answers to the most basic questions about our universe. Our theme throughout this book is to emphasize this link.

1.1.2 What are the observables?

Astronomy is truly an observational science. Unlike in a laboratory experiment, the conditions cannot be changed. That is, we on Earth are passive observers (so far) in almost all astronomical experiments, and we can do nothing other than intercept (observe) the various forms of energy which reach the Earth from the depths of space. Of course, there have been a few notable exceptions for solar system studies involving manned and unmanned spacecraft that have returned samples to Earth, and from time to time we can retrieve rocks from space which have survived passage through the Earth's atmosphere in the form of meteorites. Otherwise, the energy forms that we can intercept passively can be summarized as:

electromagnetic radiation (gamma-rays through radio waves)
cosmic rays (extremely energetic sub-atomic charged particles)
neutrinos (tiny neutral particles with almost immeasurably small mass)
gravitational waves (disturbances in a gravitational field)

Of these, the study of electromagnetic radiation which, as shown by the great Scottish mathematical physicist James Clerk Maxwell (1831-1879) in 1865, incorporates visible light, is still the most dominant. Gravitational waves, ripples in spacetime predicted by Einstein (1879-1955), have not yet been detected directly, but in the USA the Laser Interferometer Gravitational Wave Observatory (LIGO), with sites in the states of Washington and Louisiana, went into operation in 2002 and similar facilities exist in Germany, Italy and Japan. Neutrino detectors and cosmic ray experiments have been developed successfully. Among the most well-known of the neutrino observatories are the Homestake Gold Mine in South Dakota (USA) where Ray Davis (1914-2006; Nobel Prize in Physics 2002) first uncovered the “solar neutrino problem” in which the Sun seemed to be emitting only one-third of the expected number of neutrinos based on the well-understood theory of nuclear hydrogen-helium fusion, and the Sudbury Neutrino Observatory in Ontario (Canada) which resolved the problem by detecting all three neutrino types when it was eventually realized that three kinds of neutrinos existed. The Kamiokande neutrino observatory in Japan was sufficiently sensitive that it detected neutrinos from the supernova explosion (SN1987A) of a star in the Large Magellanic Cloud about 170,000 lightyears away; a lightyear is about 9.5 trillion kilometers (about 5.9 trillion miles) and is the distance light travels in one year. The vast majority of cosmic ray particles are protons, the positively-charged nucleus of the hydrogen atom, although heavier nuclei are also observed. Low energy cosmic rays must be detected from spacecraft, but higher energy rays generate an “air shower” when they impact the Earth’s atmosphere resulting in faint flashes of blue light known as Cherenkov radiation which can be detected by a suitably-designed large telescope on the ground. One of the first telescopes built to detect Cherenkov radiation was the Whipple telescope on Mt. Hopkins in Arizona (1968) but many newer facilities now exist.

Maxwell’s equations are a set of four fundamental relationships that quantify experimental findings about electric and magnetic phenomena, especially those involving the magnetic field due to an electric current (Ampere’s Law modified by Maxwell) and the electric field caused by a changing magnetic flux (Faraday's Law of electromagnetic induction). These two equations can be combined to show that both the electric and magnetic fields satisfy the known form for a wave equation. Maxwell's analysis revealed that light is essentially characterized by oscillations of electric and magnetic fields which give the radiant energy the property of a wave motion. Different regions of the electromagnetic "spectrum" correspond to different "wavelengths" (denoted by the Greek letter lambda, λ; see Appendix for Greek alphabet) and the energy in the wave moves through empty space at a speed of 299,792,458 meters per second (m/s),

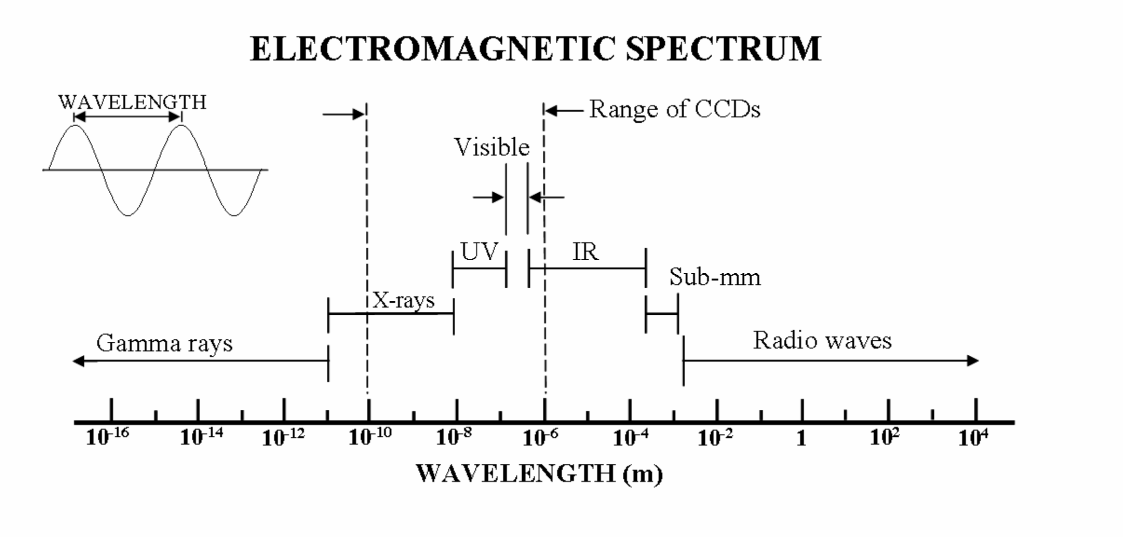


Fig. 1.3 The electromagnetic spectrum: X-rays, light and radio waves are all different forms of electromagnetic radiation. In the vacuum of empty space, each of these forms of radiant energy travel in straight lines with the same speed—the speed of light.

which is of course the speed of light (usually denoted by the letter c); actually, Maxwell derived this number from two electrical constants. Useful approximate values for the speed of light are 300,000 km/s, 186,000 miles per second and 670 million miles per hour. The frequency of the oscillations (denoted by the Greek letter nu, ν) is related to the wavelength by the very simple equation

$$\nu\lambda = c \tag{1.1}$$

In the simplest case of a monochromatic (single wavelength) wave traveling in the x-direction and vibrating in a fixed (x,y) plane, the oscillation can be described by a simple sinusoid, for example, $y = a \sin(\omega t - kx + \phi)$ with $\omega = 2\pi\nu$ and $k = 2\pi/\lambda$, and the average intensity of the light is proportional to the square of the amplitude (or swing) a^2 of the wave and ϕ is the phase. The importance of Equation 1.1 is that it implies no restrictions on the frequencies or wavelengths themselves, only that their product must be the speed of light. Optical measurements show that normal visible light corresponds to wavelengths around 0.5 millionths of a meter and frequencies of 600 trillion cycles per second, but waves of much lower frequency (300 million cycles per second) with huge wavelengths of 1 meter or more should be possible. This result led to the prediction and subsequent discovery of radio waves. The unit of frequency (1 cycle per second) is now called the hertz (Hz) after Heinrich Hertz (1857-1894) who validated Maxwell's predictions by experiments with early radio antennas. Electromagnetic waves can bounce off certain surfaces (reflection), be transmitted through certain materials with a change of direction (refraction), curl around obstacles or through openings by diffraction, and "interfere" with one another to cause cancellation or amplification of the wave. Of these, the phenomenon of diffraction sets a fundamental limit on measurements and we will mention this limit many times in the quest for ultimate perfection in imaging. For now, we note only that the "angular resolution" or ability to separate two closely spaced stars a small angle apart on the sky, for a telescope of diameter D collecting light of wavelength λ, is given approximately by 57.296° λ/D in the diffraction limit. Maxwell's equations, electromagnetic waves and their interactions through interference, reflection, refraction and scattering are described in any good college physics text. More details will be presented as needed in subsequent chapters. Because the electromagnetic oscillations are transverse to the direction of propagation of the energy, these waves can be "polarized" which means they have an associated "plane of vibration."

As shown in Fig. 1.3, all the well-known forms of radiant energy are part of this electromagnetic spectrum. The range in wavelengths is incredibly large. Radio waves are characterized by wavelengths of meters (m) to kilometers (km), whereas X-rays have wavelengths around 1 nanometer (nm) or one billionth (10^{-9}) of a meter, comparable to the size of atoms. Other length units such as the micron (µm, 10^{-6} m) and the angstrom (Å, 10^{-10} m) are commonly used; scientific notation (powers of ten) and prefixes to standard units (such as nano- and micro-) are summarized in the Appendices. Visible light, with wavelengths from about 390 to 780 nm (or 0.39–0.78 µm), occupies only a very small portion of this enormous radiant energy spectrum.

The rate at which the energy flows from a source is called the "radiance" or power, and the power emitted by the Sun for example is about 3.8 x 10^{26} watts; 1 watt is equivalent to 1 joule per second. The power that is *received* by one square meter is the "irradiance" (measured in watts/m^2) and irradiance drops off inversely as the square of the distance from the source. Thus, at the average distance of the Earth from the Sun the solar irradiance is about 1366 watts per square meter above the Earth's atmosphere.

Measurements that can be made on electromagnetic radiation are limited. Basically, we can determine:

the *direction* and *time of arrival* of the radiation
the *intensity* at each wavelength or spectral energy distribution
the *polarization* or degree of alignment of the electric and magnetic fields in the radiation,
the *phase* or relation between waves

Any of these quantities can vary with time and all can be observed with varying amounts of resolution (angular, spectral or time) determined both by the limitations of measuring equipment and the wave nature of light.

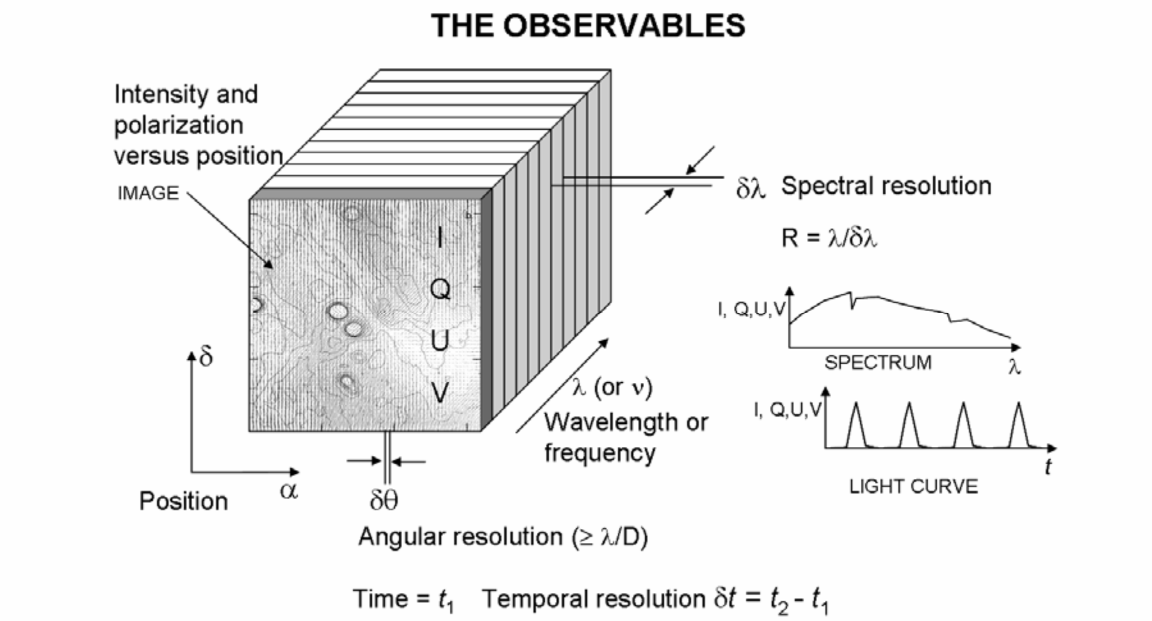


Fig. 1.4 Shown here is a pictorial summary of most of the observables for electromagnetic radiation. Polarization is represented by the Stokes intensities Q, U and V to be defined later.

A map of the distribution of intensity over a given field of view is an “image” of that scene at the given wavelength (see Fig. 1.4). All that we know about the universe must be extracted from measurements of these energy forms. Naturally, astronomy began as an optical science because human beings have built-in optical sensors, our eyes.

1.2 FROM EYES TO ELECTRONIC SENSORS

Looking through a telescope at the stars on a crisp, clear night is usually sufficient to get hooked on astronomy. It certainly was in my case. But the spectacular pictures from the Hubble Space Telescope so familiar to everyone since the early nineties are nothing like what you see when you peer through a telescope with your eye. Why is that? Because, not only have electronic sensors been used to detect light that the eye simply cannot see but also, computers have processed the digital pictures to enhance the appearance of certain features for ease of study. If electronic sensors measure light that the eye cannot see then how can we represent such measurements, other than by a table of numbers? In practice, we assign colors that the eye *can* see to each of the invisible wavelengths in order to create a visualization of the scene. In this case the color is clearly false and does not represent what your eyes would see when looking at this object. Visualization techniques will be explained in later chapters. For now, let’s start by considering the detection of light and the features and limitations of the human eye.

2

Beating the atmosphere

Images of point-like astronomical sources formed on a CCD camera will be represented by a point spread function. In the absence of other degrading effects, the spreading of the image is determined by the diffraction of light. In practice, for ground-based observatories, the light must pass through the atmosphere, which has a major impact on image quality because of turbulence.

2.1 ATMOSPHERIC ABSORPTION AND TRANSMISSION

While the Earth's atmosphere provides the biosphere that we live in and protects life on the planet from harmful radiation from space, it is not so friendly to the pursuit of astronomy. The atmosphere absorbs and scatters incident electromagnetic radiation. It is the scattering of sunlight by air molecules that makes the sky seem blue; Lord Rayleigh (1842-1919) showed that the scattering is inversely proportional to wavelength to the fourth power (λ-4) and so blue photons are scattered much more strongly and reach our eyes from all directions. Scattered sunlight is also highly polarized (90° from the Sun), a fact that is easily demonstrated with Polaroid sunglasses by tilting your head from side to side while looking at the blue sky to see that the intensity changes with the angle of your sunglasses. Under certain conditions the atmosphere also emits radiation. Of more concern is the fact that the atmosphere disturbs the incoming waves through turbulent air motion which in turn limits the ability of a telescope to achieve its ultimate angular resolution.

Figure 2.1 shows a simplified plot of just how opaque the Earth's atmosphere is to electromagnetic energy at each wavelength from gamma rays to radio waves; 100% opacity means that the transmission is zero. There are only two regions of the spectrum that are easily transmitted to the ground by the Earth's current oxygen-rich atmosphere. These regions are the Visible plus Near-Infrared, and the Radio, everything else is opaque. Even within the range of visible light the atmosphere still absorbs in some very narrow bands of wavelengths thus producing absorption effects of terrestrial origin when recording the spectra of astronomical sources. For example, oxygen absorbs strongly in two broad spectral bands called the A band (near 760 nm) and the B band (near 688 nm), well within the CCD spectral range. Each of these bands includes about 40 spectral line transitions with such strong absorption that the atmosphere is essentially opaque at those wavelengths.

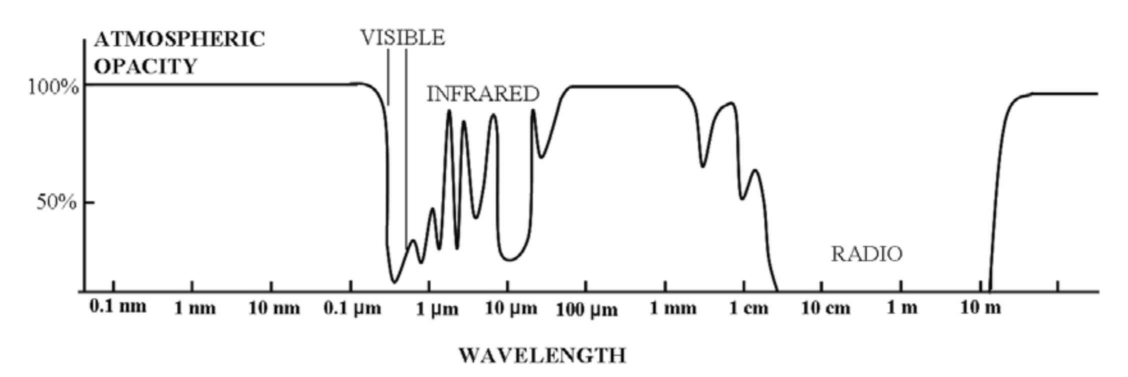


Fig. 2.1 The transmission of the atmosphere at each wavelength is illustrated from gamma ray to radio waves. Except for visible light, some near-infrared light and radio waves, all other forms of electromagnetic radiation are blocked by the atmosphere.

Water vapor absorption occurs weakly at 514, 606, 660, 739 and 836 nm and then more strongly in several bands from 970-1940 nm. Ultraviolet waves, X-rays and gamma-rays are effectively blocked by the atmosphere; mainly by water vapor for X-rays, but also ozone (O_3), oxygen (O_2) and carbon dioxide (CO_2) for UV. Ultraviolet radiation from 400 – 320 nm (also called UV-A) can reach the surface and excellent UV observations can be made from sufficiently high mountain-top observatories. Between 320 and 290 nm is the range called UV-B in the terminology of biological damage from ozone depletion. Attenuation of these UV photons by the ozone layer is something like 350 billion to one compared to the top of the atmosphere. Below 290 nm (UV-C) electromagnetic radiation is completely blocked by ozone at about 35 km. Telescopes operating at very short wavelengths must be in space or carried very high into the atmosphere (above 50 km) by balloons or rockets.

Moving to the infrared, we find that the atmosphere is opaque at some wavelengths and transparent at others. The illustration in Fig. 2.1 is only a guide. In detail there is much more absorption structure throughout the infrared and we will return to this topic in a later chapter on infrared imaging. Again, the main culprits are the greenhouse gases, carbon dioxide (CO2) and water vapor (H2O), which create a series of “windows” for wavelengths out to 20 µm, about 40 times the wavelength of normal visible light. Beyond 20 µm, observations must be done from space (or from the stratosphere) until the wavelength reaches about 2 cm where the atmosphere again becomes transparent and radio astronomy is possible. From very high dry sites however, there is the possibility of sub-millimeter wave observations under good conditions at 450 and 850 µm. At the longest wavelengths, radio waves longer than about 20 m are blocked by the ionosphere.

While water vapor and carbon dioxide do an efficient job of blocking out a lot of infrared and sub-millimeter radiation, the water vapor is sensitive to height in the atmosphere and consequently, high-altitude sites such as Mauna Kea, Hawaii at 4.2 km (13,796 ft) and the high deserts of the Chilean Andes are excellent places for ground-based astronomy. The realization that high-altitude sites might be better suited for astronomy is often credited to Sir Isaac Newton because of this famous statement in his 1730 treatise on Opticks:

"*For the Air through which we look upon the Stars, is in perpetual Tremor ... But, these Stars do not twinkle when viewed through ... large apertures. The only Remedy is a most serene and quiet Air, such as may perhaps be found on the tops of the highest Mountains above the grosser Clouds*."

Charles Piazzi Smyth (1819-1900) was the Scottish Astronomer Royal (from 1846-1888) who experimented with "mountain top" observing in Tenerife, Canary Islands in 1856 with a grant specifically to test Newton's idea. American astronomer Henry Draper (1837-1882) also suggested building observatories in mountainous areas and is known to have mentioned the Andes.

The relative concentrations of the permanent gases in a dry atmosphere are nearly constant: nitrogen (N_2) at 78.1% by volume, oxygen (O_2) at 20.9%, Argon (A) at 0.9% and carbon dioxide (CO_2) at 0.03%. All other permanent constituents are less than 0.002%. The two major variable constituents are ozone and water vapor. The maximum concentration of ozone occurs at high altitudes (10-30 km) and mainly affects the UV transmission. Water vapor is a low-altitude phenomenon and varies strongly with temperature (and hence season) and altitude. Commonly, the amount of water vapor contained in the optical path is called the "precipitable water" and is measured in mm. Precipitable water is defined as the depth of the layer of water that would be formed if all the water vapor (H_2O molecules) along the line of sight was condensed in a container having the same cross-sectional area as the optical beam. It is not necessary to know the cross-sectional area of the beam because, if the area was larger and more water was condensed, it would be spread out over that larger area and its depth would be the same. The amount of precipitable water is usually expressed in mm per km of path length or as the total mm in the air mass above the observatory. Values vary from 1-15 mm at low astronomical sites (below 2 km) but 4 mm is typical for high dry locations like Mauna Kea.

Atmospheric pressure (P) below 120 km altitude is approximately given by an exponential decline $P(h) = P_0 e^{-h/H}$ where h is the altitude and H is called the scale height, and is the value of h where the pressure falls to 1/e = 0.37 (37%) of its value at sea level (P_0). The standard atmosphere (atm) has a pressure of 101.325 kilopascal (kPa) or 14.696 pounds per square inch (psi). Depending on temperature, the typical value for H is ~8 km for the permanent constituents of the atmosphere. Water vapor content falls off much more rapidly with height because it is concentrated close to sea level. A typical value for a site like Mauna Kea is H_{wv} ~1.85 km whereas the mountain top is at 4.2 km. The site of the Atacama Large Millimeter Array (ALMA) telescopes (Chajnantor, Chile) is at a height of 5.06 km with an expected 1-2 mm of precipitable water vapor; for reference, the height of Mt. Everest is 8.848 km (29,029 ft).

The thickness of the atmosphere through which radiation has to pass is measured in terms of airmass, where one airmass is the optical thickness when looking straight up. By treating the atmosphere as plane-parallel slabs, airmass (X) is given fairly well by the secant (= 1/cosine) of the zenith angle (z), also called the zenith distance, the angle between the zenith point overhead and the star; X=sec$z$. For a star of known right ascension and declination (α, δ) on the sky, its zenith angle can be computed from the relation $\cos z = \sin\phi \sin\delta + \cos\phi \cos\delta \cos(\mathrm{LST}-\alpha)$ where ϕ is the latitude of the site and LST is the Local Sidereal Time. (More information on coordinate systems and spherical trigonometry is given in the Appendix.) The atmosphere absorbs preferentially in the blue and therefore both dims and reddens starlight. In fact, the attenuation or extinction as a function of wavelength can be written (approximately) as $E_\lambda = C_\lambda \sec z$ where C_λ is a constant over a limited band of wavelengths. Thus, if m_λ is the true magnitude of a star measured outside the Earth's atmosphere at this wavelength, then the observed magnitude at a given zenith angle becomes $m_\lambda(\mathrm{z}) = m_\lambda + C_\lambda \sec z$ (Bouguer's Law). At the zenith, z = 0° and $\sec z = 1$, while at z = 60° (or an elevation angle of 30° above the horizon) $\sec z = 2$. Bouguer's Law is accurate for values of z up to about 60° after which the plane-parallel atmosphere assumption on which it is based breaks down. A more precise formula is then: $X = \sec z - 0.0018167(\sec z - 1) -0.002875(\sec z - 1)^2 - 0.0008083(\sec z - 1)^3$. Values of the extinction coefficient C_λ have a large range from about 3.7 magnitudes per unit airmass in the UV (300 nm) to about 0.005 magnitudes per airmass in the near-infrared at the silicon CCD limit (1100 nm). This "constant" is also variable from site to site and even during the night at a given site due to subtle changes in the atmosphere.

Stated another way, the measured photon arrival rate (S) is reduced by a transmission factor $T_{\mathrm{atmos}}(\lambda)$ which depends on wavelength and is a number between 0 and 1. The transmittance is typically given by the exponential factor T = $\mathrm{e}^{-\mu_a(\lambda)L}$ where $\mu_\mathrm{a}(\lambda)$ is the absorption coefficient in units of cm^{-1} at a given λ and L is the path length (in cm) for absorption. There will be additional transmission losses through the telescope and instrument, and then a final loss if the quantum efficiency of the detector (η) is less than one. Thus the transmitted signal at a given wavelength will be $S = \tau\eta S_0$, where τ is the product of all transmission factors (atmosphere, telescope, instrument) and S_0 is the incident photon arrival rate.

Finally, the direction of light transmitted through the atmosphere changes because of refraction, and this causes a wavelength-dependent displacement of the star image from its true position. This effect makes a star move apparently towards the zenith from its true position by an amount $\Delta z = (n\text{-}1) \tan z$ where z is the observed zenith angle. For example, taking n=1.00029 (~700 nm) then $\Delta z \sim 59.8'' \tan z$. At shorter wavelengths n is slightly larger. For example, taking n=1.000295 (~480 nm) yields $\Delta z \sim 60.8'' \tan z$ which is a noticeable difference and is known as atmospheric dispersion. This difference would be enough to cause considerable light loss at the slit of a spectrometer if the slit was less than 1 arc second wide. One solution is to orient the slit at the parallactic angle (q), the angle between celestial north and zenith given by $\sin q = \cos\phi\ (\sin H/\sin\alpha)$, where ϕ is the latitude, H is the hour angle and α is the right ascension. It is also possible to eliminate this form of image degradation using a device called an Atmospheric Dispersion Compensator (ADC) as discussed in the next chapter.

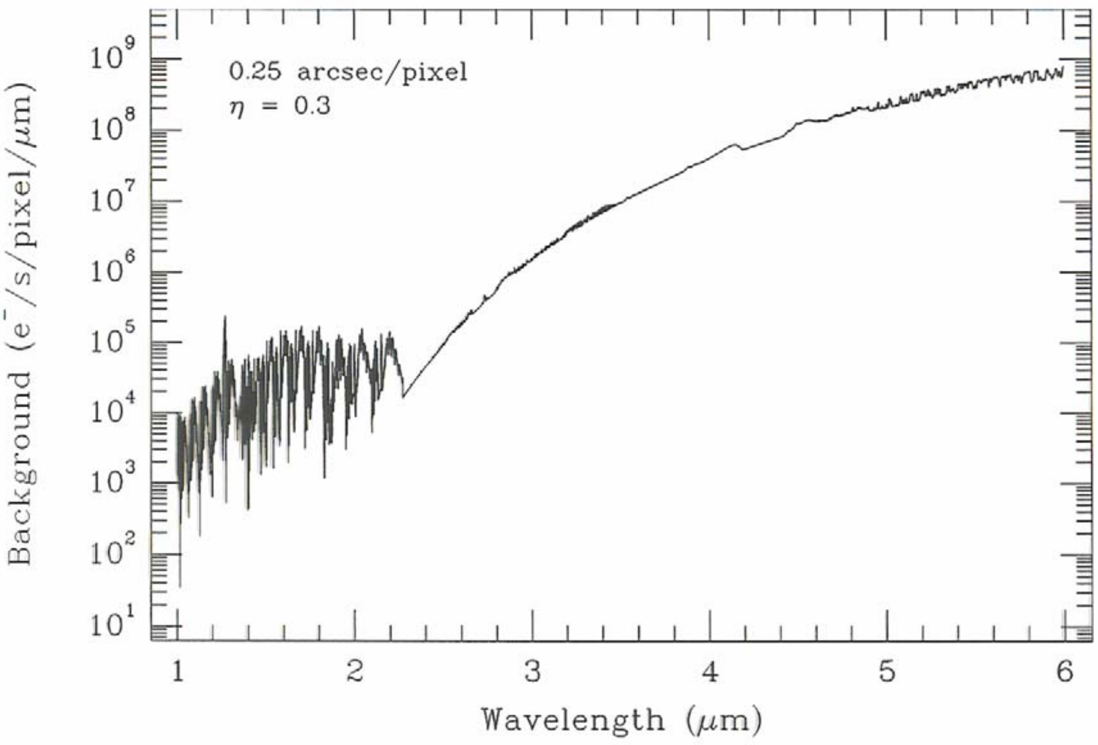


Fig. 2.3 A plot of the combined OH night-sky emission and thermal emission from the telescope in the near-infrared from 1-6 microns showing a dramatic increase in the brightness of the night-time background at these wavelengths.

The intensity scale in Fig. 2.3 is logarithmic, thus the night sky in the infrared is very bright compared to visible wavelengths. As the bulk of the atmosphere is generally ~20 K colder than the telescope mirrors, then for wavelengths less than ~13 μm it is the thermal emission from the telescope itself and any other warm optics in front of the detector that dominates the background. All but high resolution spectroscopy in the infrared will be limited in signal-to-noise ratio by the enormous background emission from the telescope and atmosphere. Mid-infrared (beyond 8 μm) and far-infrared observations (beyond ~20-30 μm) are best done from space or the stratosphere.

Other sources of background light include the zodiacal light caused by sunlight scattered by tiny particles within the solar system; in the visible part of the spectrum zodiacal light is equivalent to about 22-23.5 magnitudes per square arc second and is concentrated in the ecliptic plane. Moonlight of course can be very bright and variable. For example, when the Moon is new the integrated sky background in magnitudes per square arc second in the ultraviolet (U), blue (B), yellow (V), red (R) and near-infrared (I) bands is 22.0, 22.7, 21.8, 20.9 and 19.9 approximately, whereas at full Moon these values become 17.0, 19.5, 20.0, 19.9 and 19.2.

2.3 TURBULENCE

Perhaps the biggest impediment for astronomy created by the atmosphere is turbulence. In the complete absence of turbulence, the image of a "point source" should be determined only by the quality of the telescope's optics and by the diffraction of light, that is, the apparent bending, spreading or interference of the light wave due to disruption by an object in its path, even the mirrors and lenses of a telescope. Diffraction can be thought of as a more complicated example of interference, which in turn is a consequence of the wave nature of light. It is well-known that a plane wave passing through an aperture creates a disturbance that spreads out beyond the aperture; a familiar example would be water waves entering a walled harbor through a single opening and spreading into the protected areas. Long before it was understood that light was an electromagnetic wave, Christiaan Huygens (1629-1695) explained diffraction by the suggestion that new spherical waves (Huygens' wavelets) spread out from all points on the wave front and this new "wavefront" is the shape that envelopes all of these wavelets. Interference between wavelets is now possible. Where the plane wave interacts with the edge of an aperture the symmetry is broken and the emergent wave front can no longer be a plane wave. In a telescope, the aperture (lens or mirror) causes the waves to come to a focus, in effect converting the plane wave to a spherical wave, but interference effects will cause the final pattern to be blurred and complex. The general case leads to the Rayleigh-Sommerfeld diffraction integrals, but there are two well-known simpler cases. The first is known as Fraunhofer diffraction in which the light source and the location of the interference pattern are effectively at infinite distances from the aperture causing the diffraction. This condition, also called the "far field" limit, is achieved using lenses or mirrors. When either the source or the location of the diffraction pattern is at a finite distance from the aperture, the theory is called Fresnel diffraction (after French physicist Augustin-Jean Fresnel (1788-1827) for his major contributions to wave optics). The diffraction pattern formed by plane waves from a point source (a very distant star) passing through the circular aperture of a telescope was solved by astronomer Sir George Airy (1801-1892) in 1983 who obtained the solution in terms of Bessel functions. Details of the derivation can be found in most classic texts on optics. As shown in Chapter 1 (Fig. 1.15) a plot of the intensity across any diameter of the Airy disk reveals a bright central maximum surrounded by fainter rings separated by dark bands. The first dark ring occurs at an angular radius of 1.22 λ/D radians from the center and the Full Width at Half Maximum (FWHM) is given approximately by λ/D radians (or 206265 λ/D seconds of arc) where λ is the wavelength of the light and D is the diameter of the telescope.

If you were wondering how interference and diffraction are related to photons and quantum theory, then perhaps a simple application of the Heisenberg Uncertainty Principle will help (Jenkins and White 1957). Consider a plane wave with wavelength λ incident on an aperture of width D. In the quantum picture the photon has a precise momentum given by $p = h/\lambda$, that is, the uncertainty in p is $\Delta p = 0$. Heisenberg's Uncertainty Principle requires that $\Delta x\ \Delta p \sim h$ and so the uncertainty in the position of the photon must be infinite $\Delta x = h/\Delta p$ where $\Delta p = 0$. This is completely consistent with a plane wave that extends indefinitely along its wavefront. At the aperture however, the location must be limited to $\Delta x = D$ in order for the light to pass. Thus the photon's momentum is now uncertain by an amount $\Delta p = h/\Delta x = h/D$. Momentum is a vector quantity and so an uncertainty in its value means that the direction of the emergent wave can vary by a small angular amount that we can estimate as $\theta = \Delta p/p$. But $\Delta p = h/D$ and $p = h/\lambda$ and so $\theta = \lambda/D$, roughly in accord with our expectations for the diffraction limit.

Let's take an example: for the infrared wavelength of 1 μm (10^{-6} m) on a 10-meter telescope, the value of λ/D is 10^{-7} radians or about 0.02 seconds of arc. This tiny angular resolution corresponds to the size of a small coin 1 cm (~0.4 inches) in diameter at a distance of 100 km (62 miles). In astronomical terms this is the same as the orbital radius of the Earth (about 150 million km) seen at a distance of about 160 lightyears. Alternatively, this angular resolution is only 10 times the separation of the Earth and Moon when viewed from the distance of the nearest star system Alpha Centauri (4.2 lightyears away). These numbers are very intriguing if only they could be realized in practice! Unfortunately, as everyone who has looked through a telescope knows, star images are always much more blurred than this. Time-dependent turbulence blurs the tiny diffraction-limited image by rapid, random shifts of position resulting in a fuzzy "seeing" disk of light that can be 10 to 100 times larger in diameter depending on the site of the telescope and atmospheric conditions. Incoming waves are distorted by randomly moving cells of air with different densities, which in turn arise from temperature variations. American astronomer Horace W. Babcock (1912-2003) performed pioneering studies of this phenomenon from about 1936 onwards, making many hundreds of visual observations at sites in California (including Mt. Wilson) and Chile. Turbulence is characterized by the size of the typical atmospheric cell. It turns out that these cells, even at a very good site, are usually much less than 1 meter across (20 cm is typical), much less than the diameter of a modern large telescope, and it is thus *this* length that determines the size of the fuzzy image or seeing disk. Astronomers compare seeing-limited and diffraction-limited images using the Strehl ratio which is defined as the intensity at the peak of the actual seeing disk divided by the intensity at the peak of the true Airy diffraction pattern. The term comes from a closely related image sharpness criterion defined by Karl Strehl (c. 1895). Typically, the Strehl ratio is ~0.01. If this ratio could be increased to nearer unity, then most of the light would be in the central spike of the Airy diffraction pattern and the contrast against the sky background would be increased enormously. Smaller image sizes also mean that narrower slits can be used in spectrographs, which in turn implies that the whole spectrometer can be made more compact. To achieve such small images is the ultimate goal of adaptive optics, a ground-based method of achieving space-based image quality. Developments in adaptive optics (AO) can be traced to both the astronomical and the military communities. Sustained efforts from the late sixties and throughout the seventies and eighties by astronomy and non-astronomy groups has led to advanced AO systems, such as the pioneering system developed by Bob Fugate at the US Air Force Starfire Optical Range in Albuquerque, New Mexico and the numerous astronomical AO systems on all sizes of telescopes from 3-10 m in diameter located in all parts of the world. References to this already large and rapidly growing field are given at the end of this chapter and the summary below draws on the work of Babcock, Beckers, Fugate, Hardy, Max, Roddier, Thompson, Tyson and many others.

2.3.1 Kolmogorov theory and origin of seeing

What is the effect of the atmosphere on image quality? The Earth intercepts only a tiny fraction of the spherical wave emitted from a distant, point-like source such as a star. When these waves arrive at the Earth the wave fronts are essentially flat and parallel to each other.

3

Telescopes

The first element of any astronomical imaging system is the telescope. Telescopes are of course the means by which the light from distant objects is collected and focused, but telescopes must do more than gather light. Providing excellent telescope optics, good tracking and minimum air turbulence in the telescope dome are important steps in obtaining good images. Following a brief historical review of telescope development, we consider some basic optical properties and their applications to telescope design.

3.1 HISTORICAL DEVELOPMENT

By the end of the 13th century in Europe, artisans in glass-making centers like Venice and Florence had already found practical techniques for grinding and polishing glass relatively cheaply and easily. Moreover, people of that time were aware that the condition known today as *presbyopia,* in which the ageing eye can no longer focus on something held at a comfortable distance, could be helped with a simple magnifying glass. But two smaller disks of glass, convex on both sides and supported in a frame were more convenient. Because these small disks were shaped like lentils, thicker in the middle, they became known as "lentils of glass" or (from the Latin) *lenses*. Concave lenses (inward curving) that correct for *myopia* (near-sightedness) were made in Italy in the middle of the fifteenth century. It was not until the beginning of the 17th century however, that these two types of lenses were combined to make a telescope. The earliest documented record is the 1608 patent application in the Netherlands by Hans Lippershey (c1570-1619), a German-born Dutch citizen, of a device with a convex and a concave lens in a tube with a magnification of about 3 times. On learning of this device Galileo Galilei (1564-1642) made his own version of the telescope in the summer of 1609 and quickly increased the magnification to 8 and then 20 times. The limiting factor of Galileo's telescope was its small field of view of about 15 arc-minutes which meant that only a quarter of the full Moon could be seen. Galileo's telescope produced an upright image. From the early Galilean telescope of 1.52-1.83 m (5 - 6 ft) in length, astronomical telescopes attained lengths of 4.57-6.10 m (15 - 20 ft) by the middle of the 17th century. Typical of this time is the telescope made by Christiaan Huygens (1629-1695), in 1656. It was 7 meters (23 ft) long; its objective lens had an aperture of about 10 cm (~4 inches), it magnified about 100 times, and its field of view was 17 arc-minutes.

In a spherical shaped lens, rays parallel to the optical axis (through the center) fail to converge at one point. Those farther from the optical axis come to a focus closer to the lens than those nearer the optical axis. This effect is called spherical aberration. To eliminate spherical aberration the lens curvatures must be either plane on one side and hyperbolic on the other, or spherical on one side and elliptical on the other. Fabrication of such shapes was beyond the technology of the time. In addition, Sir Isaac Newton (1672) showed that white light is a mixture of colored light and that every color had its own degree of refraction. Consequently, any curved lens will decompose white light into the colors of the spectrum, each of which will come to a focus at a different point on the optical axis. This effect, which became known as chromatic aberration, causes the image of a star to be surrounded by circles of different colors. Thus lenses had limitations for astronomy, but telescopes with long focal lengths helped to reduce both of these effects.

In 1662, at the age of 24, the Scottish mathematician and astronomer James Gregory (1638-1675) wrote a treatise entitled *Optica Promota* describing the concept of a “reflecting” telescope made from two concave mirrors, one parabolic and the other ellipsoidal. Meanwhile, working independently in England, Newton had constructed the first reflecting telescope using a spherical mirror, a version of which was presented to the Royal Society in London in 1672. Gregory’s more difficult design was eventually built successfully by Robert Hooke (1635-1703) and demonstrated in 1673. Around the same time Huygens was made aware of the idea of a similar reflective design to Gregory’s, now attributed to an obscure monk named Laurent Cassegrain (1629-1693). Newton cast a two-inch mirror blank of speculum metal (basically copper with some tin) and ground it into a spherical shape. He placed the mirror at the bottom of a tube and caught the reflected rays on a small flat secondary mirror placed at 45° near the top of the tube which reflected the image into a convex lens outside the tube for easy viewing by eye. When this instrument was shown to the Royal Society it caused a sensation; it was the first working reflecting telescope. Unfortunately, others were unable to grind mirrors of regular curvature, and to add to the problem, the mirror tarnished easily and had to be re-polished every few months. Consequently, the reflecting telescope remained a curiosity for decades. By about 1723 however, John Hadley (1682-1744) the English inventor of the octant, a precursor to the sextant, and others had perfected better polishing techniques and the first parabolic version of the Newtonian telescope was made. By the middle of the 18th century many reflecting telescopes with primary mirrors up to six inches in diameter had been produced. James Short (1710-1769) is said to have made thousands of parabolic and elliptic mirrors around 1740. It was found that for large focal ratios, f/10 or more, the difference between spherical and paraboloidal mirrors was negligible in the performance of the telescope. In the latter half of the 18th century large reflecting telescopes with parabolic ground mirrors came into their own. Sir William Herschel (1738-1822) built a reflector with a mirror diameter of 1.22 m (4 ft) and a 12.2 m (40 ft) focal length which he used to discover moons of Saturn, but by all accounts it was not as easy to use as his 6.1 m (20 ft) long 0.475 m (18.7 inch) reflector. Nevertheless, it remained the largest telescope for over 50 years until Lord Rosse’s 1.83 m (72 inch) reflector was built in 1845. To tackle the serious problem of rapid tarnishing in metal mirrors, Herschel always had a spare ready to exchange when a mirror required re-polishing. Readers are referred to Henry King’s book, *The History of the Telescope* (Dover, 1979) for more historical details.

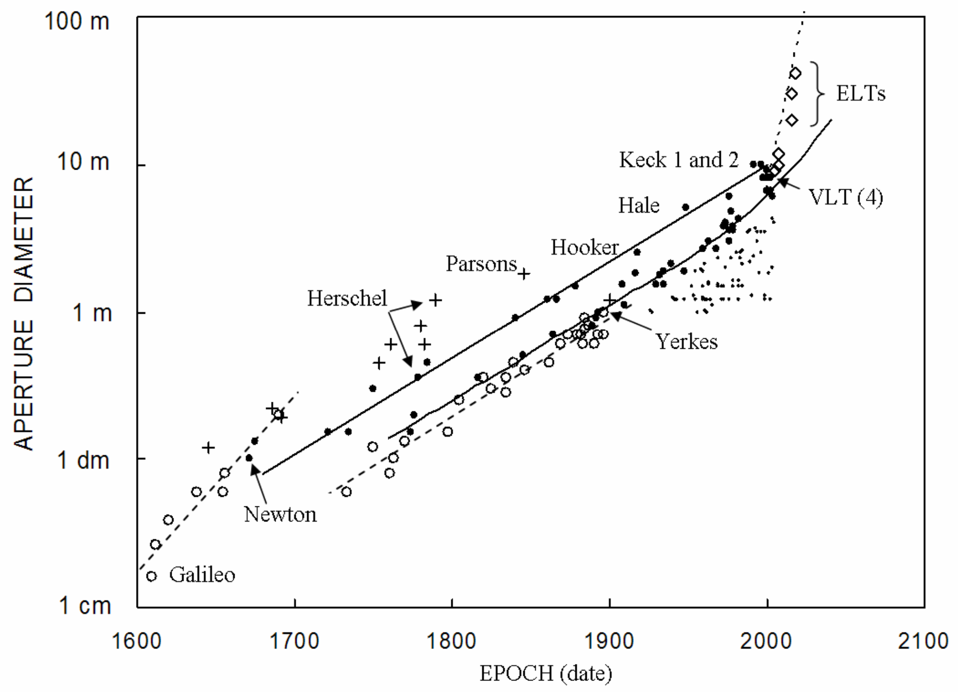


Fig. 3.1 The growth of aperture size with time is plotted from the invention of the telescope to present day. Credit: René Racine.

In the 400 years since Galileo turned his telescope towards the sky and ushered in a revolution, telescope designs have evolved. In classical times the size of a telescope was characterized by its focal length, but modern astronomical telescopes are always identified by the diameter of the primary collecting aperture, or more generally by the diameter of the equivalent circle with the same collecting surface area. The largest telescope used by Galileo had a *diameter* of 4.4 cm (1.75 inches) but as of 2008 there were 10 general-purpose optical telescopes with effective diameters greater than 800 cm (315 inches or 26.25 ft) and several others of special design. In a thorough compilation of astronomical telescopes through the ages, Racine (2004) finds that the doubling time for aperture size was about 50 years up until about 1950, but since then telescopes have been doubling in size at a rate that is at least twice as fast. Figure 3.1 is an updated illustration from Racine’s paper and shows the historical growth of all types of telescopes of modest size. Table 3.1 lists the world's largest telescopes (diameters > 3.5 m) prior to 1993.

Interestingly, after construction of the 5.08-m (200-inch) Hale Telescope on Palomar Mountain in California in 1949, the trend was for smaller telescopes in what has become known by the imprecise term "4 meter class"; in this definition all telescopes with primary mirror diameters between 3.6 m and 4.5 m are lumped together. Exceptions were the Soviet BTA and the Smithsonian MMT, both of which were unique in their design, and both of which may have been somewhat ahead of their time.

Table 3.1 Telescopes larger than 3.5 m in the pre-Keck era

Telescope Name (date opened)/Site	Primary (m)	f/ratio	Mounting
Bol'shoi Teleskop Azimutal'nyi (1976) Mt. Pastukhov, (Russia)	6.0	f/4.0	alt-az
George Ellery Hale Telescope (1949) Mt. Palomar, (California, USA)	5.08	f/3.3	equat
Multiple Mirror Telescope (1977) Mt. Hopkins, (Arizona, USA)	4.5	f/2.7	alt-az
William Herschel Telescope (1982) Roque des los Muchachos, (Canary Is.)	4.2	f/2	alt-az
The Blanco Telescope (1968) Cerro Tololo, (Chile)	4.0	f/2.8	equat
Anglo-Australian Telescope (1974) Siding Spring, (NSW, Australia)	3.9	f/3.3	equat
Nicholas Mayall Telescope (1966) Kitt Peak, (Arizona, USA)	3.8	f/2.8	equat
United Kingdom Infrared Telescope (1979) Mauna Kea, (Hawaii)	3.8 (thin)	f/2.5	equat
Canada-France-Hawaii Telescope (1979) Mauna Kea, (Hawaii)	3.6	f/3.8	equat
European Southern Observatory (1976) Cerro La Silla, (Chile)	3.6	f/3.0	equat

In March of 1993, the largest telescope in the world went into operation and a new era in astronomy was born. It was the first of a pair of ten meter (10 m) telescopes funded by the W. M. Keck Foundation for the California Institute of Technology (Caltech) and the University of California (UC) that employed the unique “segmented mirror” concept championed by Jerry Nelson of the University of California (Nelson 1995). The second telescope was inaugurated in May 1996 (Fig. 3.2). At that time, at least eight other optical/infrared telescopes with collecting apertures larger than 6.5 meters in diameter, and employing different technologies, were also under construction and several more were being contemplated. What drove this remarkable development?

Following the introduction and growth of CCDs, more and more area on the sky could be digitally imaged to deeper levels. In addition, the efficiency of spectroscopy had already been improved by the use of multi-slit devices and optical fibers to observe many objects simultaneously. Once the quantum limits of sensitivity in detectors and instruments has been reached, the only way to gain large factors in efficiency is to construct even larger ground-based telescopes and to develop methods for counteracting the image-blurring effects of turbulence in the Earth's atmosphere.

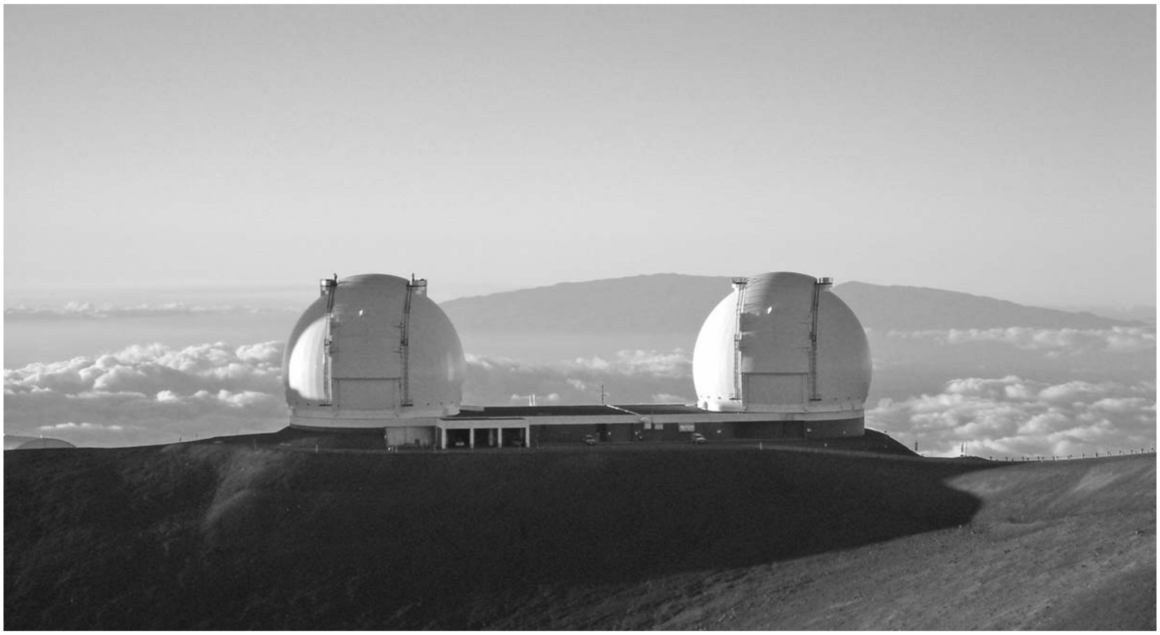


Fig. 3.2 The twin domes of the W.M. Keck Observatory on the summit of Mauna Kea. Each dome encloses a telescope with a segmented mirror having an effective aperture of about 10 meters. The domes are 85 m apart.

There are essentially three fundamental issues:
(1) how to achieve a very large collecting aperture of the required optical performance
(2) how to support and control in the optimum way such a potentially very heavy mechanical structure
(3) how to enclose a very large telescope in a cost-effective way with negligible degradation on image quality due to vibration, air disturbance or inadequate environmental protection (wind, dust).

Moreover, new telescopes must be designed to capitalize on the very best seeing conditions at the world's best sites, and must be designed with remote control in mind. Roger Angel (University of Arizona), one of the pioneers in this field, summed up the situation (Angel 1989), "the problems of building the new generation of telescopes are compounded because they must not only be bigger, but also must give sharper images than their predecessors". Basically, it all comes down to how the mirrors are made and supported. There are three categories of new technology:

- **segmented mirrors** — smaller monolithic disks of thin polished glass are used to form the surface of an efficient rigid "backing" structure of steel or carbon fiber. Position actuators are still required to make the attachment and to correct for thermal and gravitational effects in the backing frame, but the time scale for such corrections is slow. Each segment is individually supported and global changes are sensed at the gaps between segments.

- **meniscus mirrors** — large monolithic disks of solid glass which are so thin that it must be accepted that they will be flexible and therefore they must be actively controlled to maintain the required shape during operation. Bending by unpredictable forces such as wind gusts requires a rapid servo system of precise force actuators.
- **honeycomb mirrors** - a thick mirror is constructed but large pockets of mass are removed from the back to make the mirror lightweight yet very stiff. The method involves a mold and a spinning furnace to form a concave parabolic front face while the ribs and back plate are formed by glass flowing down between the gaps in the mold.

Producing the mirror blank is only the first step. Larger and faster primary mirrors require new polishing methods to achieve their final figure. It is not the deep curvature itself that is the problem, but the asphericity that results in different curvature from place to place and between tangential and radial directions. A rigid pitch lap cannot accommodate the changes of curvature needed as it is stroked over a strongly aspheric surface, unless it is moved only in the tangential direction, but to do this will result in circular grooves or zones. Conventional polishing of 4-meter mirrors is typically limited to focal ratios of *f*/2 or slower. To polish the primary mirror blank for the 4.2-m William Herschel Telescope (Canary Is.), David Brown of Grubb Parsons, England (Brown 1986) used a lap which changed shape as it moved. His method was based on the fact that, when a full-sized lap is used to make polishing strokes across a paraboloid the distortion required to maintain contact is that of coma. For the same reason, the off-axis aberration of a paraboloid is also coma. The general principle of making the lap change shape as it moves is called "stressed lap" polishing.

Each of the three methods described has been applied to build the current generation of very large telescopes for ground-based astronomy at optical and infrared wavelengths. Table 3.2 lists all telescopes operational or nearly so as of March 2008 with primary mirror diameters larger than about 6.5 meters.

The 6.5-m telescope called the MMT in this list is in the same enclosure as the original Multiple-Mirror Telescope and is known as the Monolithic Mirror Telescope in order to preserve the well-known acronym of the observatory. At the time of writing, a copy of the Keck telescope with very slightly larger segments is nearing completion on La Palma for Spain and partners. It is called the Gran Telescopio Canarias (GTC) and will have an effective diameter of about 10.4 m when it begins science operations. The unique Large Binocular Telescope (LBT) is specifically designed to combine the light from two side-by-side 8.4-m primaries giving it an effective aperture of almost 12 m. Both primary mirrors are installed and work is continuing on combining the beams. Before looking at the technology associated with each of the three telescope types it is valuable to review the fundamental issues of telescope design in general.

3.2 TELESCOPE DESIGNS

Telescopes fall into one of three basic types: Refractive (dioptric; using lenses); Reflective (catoptric; using mirrors); Hybrid (catadioptric; using a combination of mirrors and lenses). Hybrid designs are frequently the most popular for amateur astronomy because of their compact design, but all large professional telescopes are reflectors.

4

The discovery power of modern astronomical instruments

Once the light has been collected by the telescope, and perhaps corrected by the adaptive optics system, it then goes through an instrument to a detector. To motivate a more detailed study of detectors and instrumentation, this chapter provides a limited review of the kinds of measurements that can be made. In the space available only a small and incomplete sampling is possible. Subsequent chapters cover the underlying principles of instruments and detectors, and expand the discussion to other wavelength regimes.

4.1 IMAGING THE SKY – MORE THAN PICTURES

Mapping the distribution of celestial sources on the sky at the wavelength of interest, serves not only to locate the position of the source precisely—a practice called astrometry—but also to provide information on its form and that of its local environment. Positional changes of a faint nearby source against the stellar background might locate comets or asteroids or trans-Neptunian objects lying beyond the orbit of Pluto. With sufficient angular resolution the orbit of one star around another can be observed directly for some binary systems. Statistical properties of all kinds of stars and galaxies become practical given large-scale photometric surveys of the sky to great depth.

In 1838 Friederich Bessel (1784-1846) published the first measurement of stellar parallax, the tiny back-and-forth angular shift in position of a foreground object against the more distant stars, that is caused by the Earth's motion (and hence changing viewpoint) in its orbit around the Sun; the average radius of the Earth's orbit is 1 AU. He detected a motion of 0.3 seconds of arc for the binary star 61 Cygni. For such small angles the distance to the object in AU is just d = 206265 /p" where p is the parallax angle in seconds of arc ("). When p = 1" the distance is 206,265 AU and so it is convenient to introduce a new unit of distance called the parsec (pc) where 1 pc = 206,265 AU. A smaller parallax gives a larger distance in parsecs (d = 1/p). For 61 Cygni p = 0.3" and therefore d = 4.33 pc; the modern value is p = 0.287" and d = 4.48 pc. Until the advent of automated plate-measuring machines, computers and then CCDs, astrometric measurements were limited in number. In

1989 the European Space Agency satellite Hipparcos (an acronym chosen to sound like Hipparchus, the discoverer of precession) was launched to perform astrometry from space. By 1993 Hipparcos had obtained the positions, parallaxes and proper motions (non-periodic displacements due to the star's motion within the galaxy) of 118,218 stars with milli-arcsecond (mas) accuracy. Later, additional catalogs called "Tycho" were issued with 1,058,332 stars to positional accuracies of 20-30 mas, and then 2,539,913 stars covering 99% of all stars brighter than 11th magnitude. The detector for these measurements was an image dissector tube scanned at 1200 Hz and photomultiplier tubes, no CCDs were used in this case. The US Naval Observatory B1.0 catalog is a tabulation from digitized plates that gives positions for over 1 billion stellar objects to ~0.2 seconds of arc. There have been major changes in the field of astrometry in recent years (Seidelmann & Kovalevsky 2002). One driving force has been the increased accuracy of measurements using very long baseline interferometers (VLBI) to achieve positional accuracies for extragalactic radio sources to much less than 1 mas, which has now led to a new "space-fixed" reference system. The new International Celestial Reference Frame (ICRF) is defined to be close to the previous dynamical reference frame (the FK5) at J2000.0. In addition, the Global Positioning System (GPS) of satellites has yielded accurate and continuous time transfer and geodesy observations of the polar motion and Earth rotation to the same sub-mas levels. At these levels of accuracy the definitions of the reference systems and the methods of reduction must be based on the theory of relativity. Fundamentals of astrometry are described by Kovalevsky and Seidelman (2004).

Measuring the brightness of a source over relatively broad wavelength bands is called photometry; the term radiometry is used at longer wavelengths. It is also possible to use narrow bands that isolate specific spectral features such as emission from hydrogen gas or other ionized atoms. Narrow band images can be extremely effective in delineating nebulae, supernova remnants and shock fronts. Photometric methods will be discussed later in more detail. Here we introduce only a few basic terms to maintain the flow and connect radiometric measurements with the familiar astronomical magnitude system. The magnitude system for astronomical brightness measurements dates back about 2000 years to the work of Hipparchus (c. 127 BC) and Ptolemy (c. 137 AD) who listed the brightest stars visible to the naked eye and called them stars of the "first magnitude" while those just barely discernible to the eye were designated as stars of "sixth magnitude". After the invention of the telescope efforts were made by the Herschels to develop a better system, but the scale in use today stems from the work of English astronomer Norman Pogson (1829-1891) who in 1856 proposed a mathematically precise logarithmic scale that was in approximate agreement with the ancient system. He noticed that a sixth magnitude star was about one hundred times fainter than a first magnitude star and so he proposed that a difference of five magnitudes be set exactly equal to 100. This means that the ratio between successive magnitudes is given by the fifth root of 100, which is equal to 2.5119. If m_1 and m_2 are the magnitudes of two stars whose fluxes have been measured to be S_1 and S_2 respectively, then $S_1/S_2 = 2.5119^{(\mathrm{m}_2-\mathrm{m}_1)}$. Notice that m_2 comes before m_1 in the exponent because fainter objects (smaller S) means larger magnitude values. A more convenient way of writing Pogson's result is to use logarithms to the base ten as follows.

$$\log\,(S_1/S_2) = \log(2.5119^{(\mathrm{m}_2-\mathrm{m}_1)}) = (m_2 - m_1)\,\log\,(2.5119) = (m_2 - m_1)\,(0.400002)$$

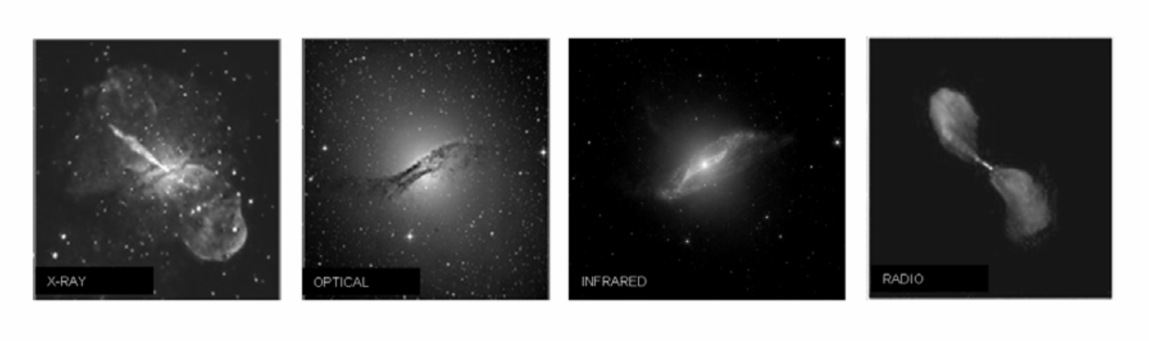


Fig.4.1 Images of the galaxy Centaurus A (Cen A) in X-rays, visible light, infrared and radio illustrate a dramatic change in appearance with wavelength. Credits: NASA/NSF/NRAO/ESO. See book cover credits.

Then, to very good accuracy, we have that 2.5 log (S_1/S_2) = ($m_2 - m_1$) or reversing the order we get the traditional expression:

$$m_1 - m_2 = - 2.5 \log (S_1/S_2)$$

Notice that this well-known equation does not allow either S_1 or S_2 to be determined but only their ratio. In other words, the magnitude of one star is found relative to another and the photometry is relative. Calibration of magnitude systems is discussed later. Brightness ratios (equivalent to magnitude differences) between two wavelength bands (e.g. blue and yellow/visual) define “colors” and can also yield basic physical information, such as the temperature of the object. A plot of magnitude versus color for stars of known distance yields the important Hertzsprung-Russell (HR) diagram of stellar luminosity as a function of temperature. Likewise, plotting one color against another color, especially over the broad optical/IR regime, is a powerful method for separating different classes of objects. For distant galaxies, color information can even provide an estimate of the distance (called a photometric redshift), which would otherwise require a spectrum of the object. Brightness measurements can be made at all wavelengths, and comparisons among the X-ray, optical, infrared and radio fluxes are often diagnostic of the physical process. In fact, the opening up of these other wavelength regimes produced many surprises in terms of which objects were brightest. For example, the front cover of this book shows how different one object can appear (Centaurus A) in radio, X-ray and visible light (Hardcastle *et al*. 2008). Figure 4.1 shows these images again plus an infrared view from the Spitzer space telescope.

4.1.1 Early surveys of the sky

Imaging the entire sky is a daunting task no matter what technology is used. The number of square degrees in the whole sky is about 41,254 while for comparison the area covered by the Moon is only 0.25 square degrees. Most telescopes give high magnification and hence a small field of view (typically < 1 degree across). The exception is the Schmidt telescope which can provide ~42.25 square degrees of field. Even so, this still implies at least 977 images to cover the entire sky. Exposure times need to be long enough to reach faint magnitudes (e.g. 20th magnitude in visible light) and the image quality must be uniformly good. For an all-sky survey two telescopes are needed, one in the northern hemisphere the other in the south and we also need a detector with a large number of pixels to cover this enormous field of view. The advent of Schmidt telescopes and large 14-inch photographic plates in the 1930s made the concept of whole-sky photography possible. Funded by the National Geographic Society, the Palomar Observatory Sky Survey (POSS) of the northern sky was carried out from 1950-1957 with the 1.2-m (48-inch) Schmidt on Mt. Palomar (California) using Kodak 103aO and 103aF plates. The Palomar Schmidt telescope has an f/4.5 1.83-m (72-inch) primary and a 1.2 m aperture covered by a glass corrector. In 1987 this telescope was renamed the Samuel Oschin Schmidt. A second epoch survey POSSII was started in 1985 and ended in 2000, stimulated in part by the availability of finer grain emulsions such as Kodak IIIaJ (blue) and IIIaF (red) in the seventies, and by the need for guide stars for the future Hubble Space Telescope. Meanwhile, similar telescopes were built in the southern hemisphere and new surveys were established by the UK Schmidt Telescope (then part of the Royal Observatory Edinburgh but now operated by the Anglo Australian Observatory) at Siding Springs in Australia, and the ESO Schmidt telescope in Chile. The UK Schmidt had an achromatic corrector and was supported by the COSMOS (later the Super-COSMOS) plate measuring and archive facility in Edinburgh. In the early eighties the Palomar telescope was upgraded with an achromatic corrector plate and POSSII was then designed to be the northern complement of the UK survey. Blue, red and near-infrared (IVN) emulsions were used to obtain images in three colors, hence tripling the numbers of photographs to be taken. While each plate covered 6.5° × 6.5° of sky and a typical plate might easily contain 4.5 billion (potential) picture elements—the grains of emulsion—only 1 out of 50 incident photons was detected (quantum efficiency = 0.02). The product of the number of pixels and the quantum efficiency of a typical (hyper-sensitized) Schmidt plate such as Kodak IIIaJ emulsion is about 60 times better than that of a single 1024 x 1024 (one megapixel) silicon CCD with 80% quantum efficiency. However, the construction of large mosaics of CCDs with hundreds of megapixels has now overtaken even this advantage of large-area plates. Moreover, the CCD provides very large gains in the ultraviolet and far red parts of the spectrum. Higher quantum efficiency implies that fainter limits are reached in the same time. Alternatively, the same detection limit can be reached in a much shorter time, hence allowing several other patches of sky to be measured.

4.1.2 Digitized surveys

Using plate measuring machines (e.g. Fig.4.2), the all-sky photographic surveys produced by the Palomar (first and second epoch), UK and ESO Schmidt telescopes are now in digital form and available on-line as the Digitized Sky Survey (DSS). You can go to the DSS web site and type in the coordinates (Right Ascension and Declination) of the region of interest.

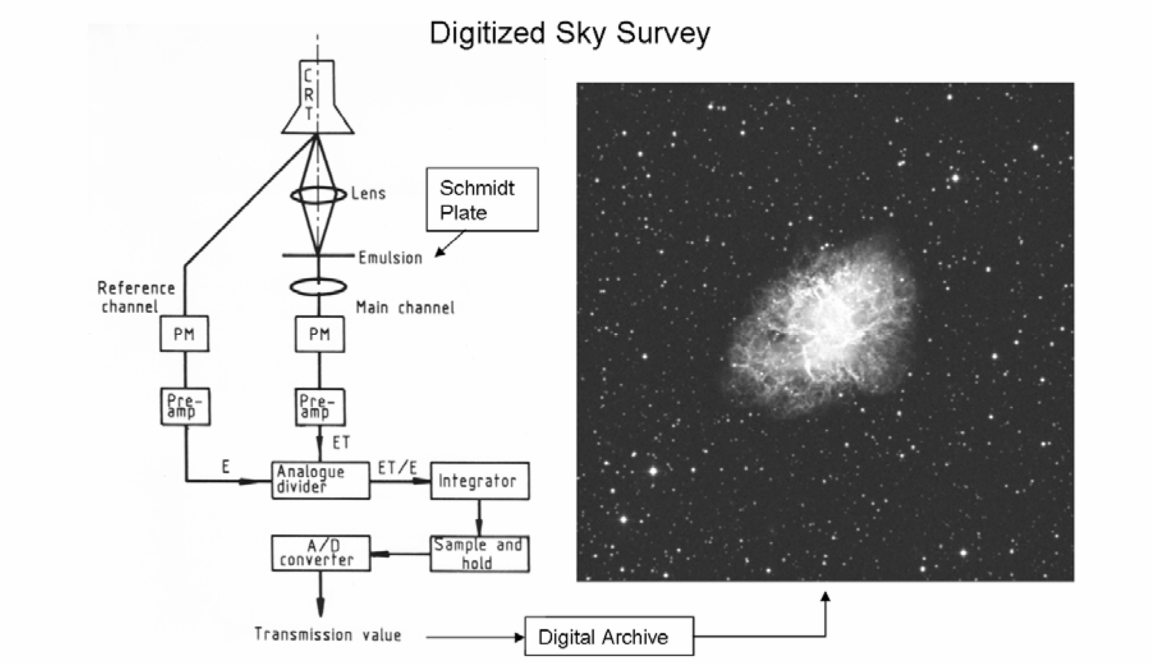


Fig. 4.2 A plate measuring machine used to convert photographic images to digital form and an image of the Crab Nebula (M1) obtained over the internet from the Digitized Sky Survey. Credit: Royal Observatory Edinburgh.

If the object has a name then you can use that instead and the facility will find the coordinates for you. For example, I typed in M1–the Crab Nebula–and the coordinates were returned. You can then select the field size to display (up to 15 x 15 arcmin) and the type of image format, either GIF or FITS. The GIF format is good if you just want a picture for a finding chart or illustration, but the FITS format, the standard among professional astronomers, is better if you want access to the digital data to change the contrast or make measurements on the image. It is easy to obtain display programs for FITS files, including an add-on for the well-known Adobe Photoshop program. A typical plate-scanning device and the result of my simple search of the DSS are shown in Fig. 4.2.

New sky surveys are now digital from the outset. In fact, the Palomar Oschin Schmidt was converted in 2000 to create an all-digital partial sky survey (DPOSS: Djorgovski *et al*. 1999) using a mosaic of 112 CCDs instead of the original 14-inch plates (Fig. 4.3). An enormous mural from part of this digital survey can be seen on a wall at the public Griffith Observatory in Los Angeles. This successful digital survey led to the discovery of many new Kuiper Belt Objects (including Quaoar, Sedna, Orcus and Eris – e.g. Brown *et al*. 2004) and many new distant quasars. DPOSS contains about 3 Terabytes of images plus catalogs of extracted sources. Software using artificial intelligence methods was developed to automate classification and measurement of the 50 million galaxies and 500 million stars.

4.1.3 Drift scanning and the Sloan Digital Sky Survey

A remarkable property of the charge-coupled device that gives it its name is that the electronic charge packet representing the image on a group of pixels can be moved (along a column of pixels) one row at a time at *any* desired rate. Charge-coupling is explained in more detail later. Here, we just need to know that the CCD can be operated at the "sidereal rate", that is, the rate at which the Earth turns on its axis relative to the most distant stars. Suppose a sensitive CCD camera is placed at the focus of a moderate-sized telescope which is pointing at a field on the celestial equator that is just crossing the meridian, but the telescope drive motors are switched off. Stars will drift across the CCD pixels, at the sidereal rate of about 15.4 arcseconds per second of time, producing trails. Now, initiate the electronic process of reading out the CCD charge pattern at the sidereal rate and in the same direction as the star moves, and then open the shutter. There will no longer be star-trails.

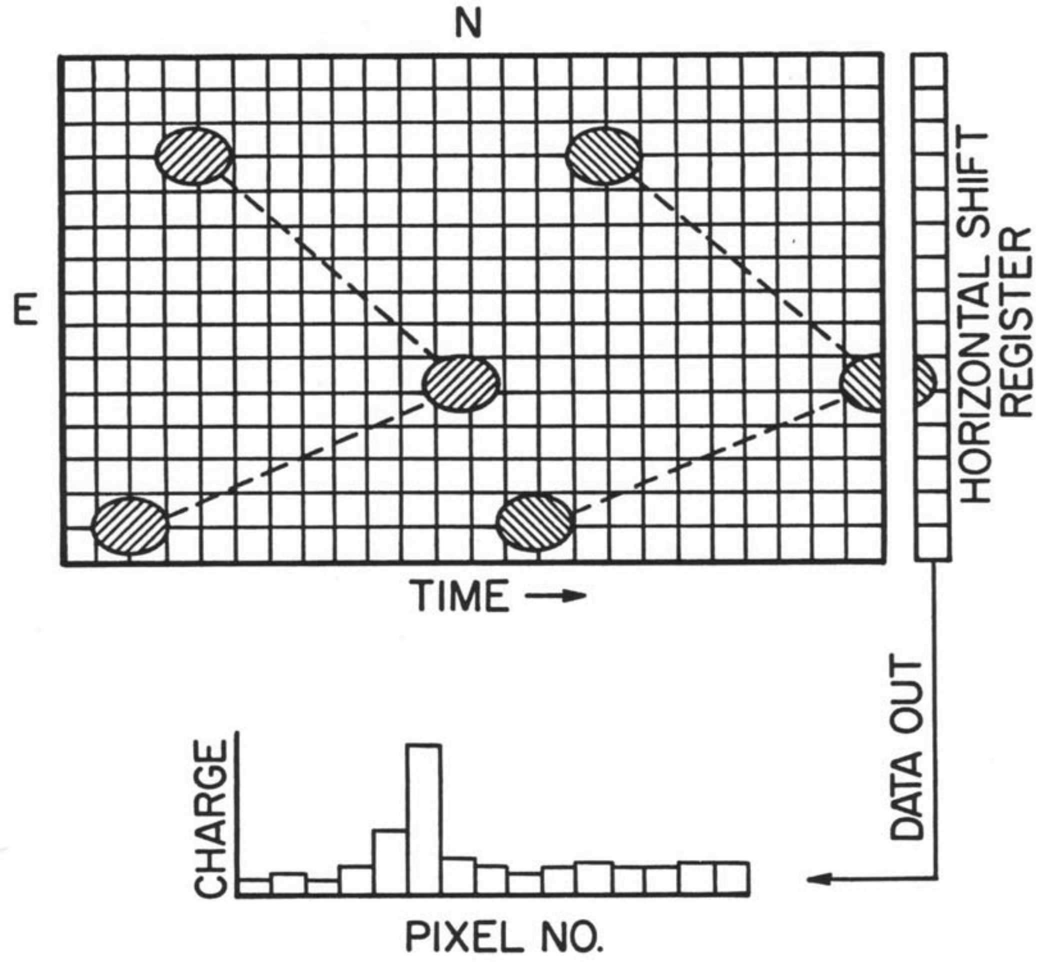


Fig. 4.7 The principle of "drift-scanning" in which the unique charge-coupling property of a CCD allows the image charge to be moved along columns from pixel to pixel at any rate. This technique is used to produce the Sloan Digital Sky Survey.

Instead, the charge image from previous pixels is added to the next one and the current position of the charge pattern will move along the column so as to keep up with the current optical image position (see Fig. 4.7). More and more photons will be collected and, ultimately, the entire column of pixels will be read out and will have contributed to the detection process. Each read out adds a row of pixels to a stored frame of data whose width is that of the CCD but whose length is arbitrary. Carrying on in this way a huge strip of sky can be surveyed systematically to a deep level without actually moving anything except electrons! This extremely important technique is called “drift scanning” and it is used extensively to perform the Sloan Digital Sky Survey (SDSS). Similar methods are used by many other digital sky surveys.

4.1.6 Diffraction-limited imaging

Powerful telescopes in space can provide diffraction-limited images from the near UV to the far-infrared, but the advent of adaptive optics on large ground-based telescopes has changed this advantage because these telescopes are much bigger than those in space, and angular resolution goes as λ/D. Ground-based diffraction-limited imaging in the near-infrared has had an enormous impact on many areas of study ranging widely from non-spacecraft studies of the outer planets and moons of the solar system to the orbital motions of stars close to the black hole at the center of the Milky Way. One of the most impressive AO results to date is the astrometry carried out by two separate teams of the motion of stars located close to the physical center of our Galaxy.

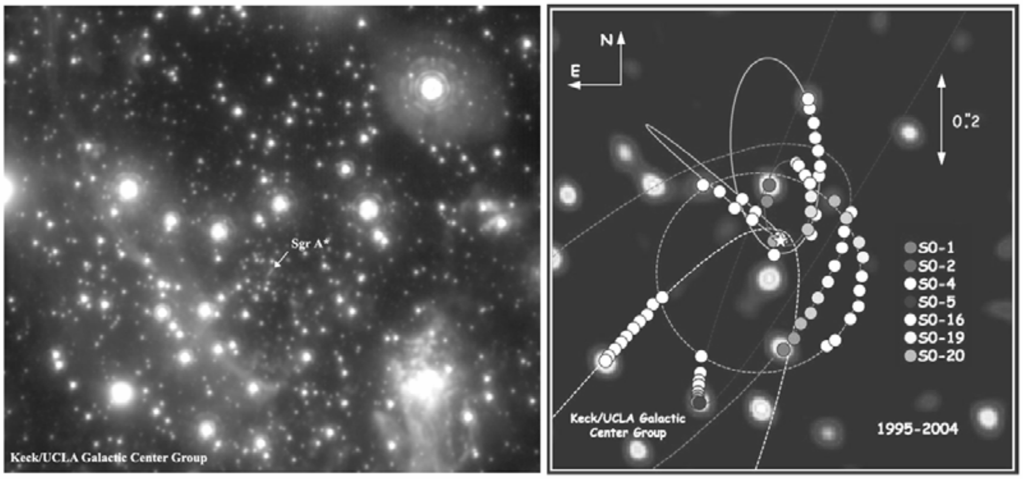

Sgr A*
Keck/UCLA Galactic Center Group
N
E
0".2
S0-1
S0-2
S0-4
S0-5
S0-16
S0-19
S0-20
Keck/UCLA Galactic Center Group
1995-2004

Fig. 4.8 Left: A diffraction-limited infrared image of the Galactic Center obtained using laser guide star adaptive optics on the Keck-2 telescope. Right: The orbits of stars revolving around the central black hole. Credit: Andrea Ghez.

Stunning AO images like the one shown in Fig. 4.8 can be used to track the motions of these stars and then the classical orbital mechanics of Newton can be used to derive the enclosed mass; remember that this is an infrared image and that no visible light reaches us from the Galactic Center. Both groups agree that the mass driving those motions appears to be ~4 million times that of the Sun and yet only a very faint and occasionally variable source is seen there, leading to the conclusion that a “black hole” resides at the center of our galaxy (Genzel et al. 2003; Ghez et al. 2005).

5

Instrumentation and detectors

Thus far, many different astronomical instruments and techniques have been introduced without much explanation of the underlying physical principles. In this chapter each class of instrument is examined in more detail. For optical instruments, typical layouts are shown and the basic relationships involving spatial and spectral resolution are given. Each major type of detector is introduced and the basic properties of semiconductors are also presented.

5.1 PHOTOMETERS AND CAMERAS

Broadly speaking, there are four classes of instruments used in astronomy namely, (1) photometers/cameras: which measure the brightness and direction of radiation; also sometimes called radiometers, (2) spectrometers: which measure the distribution of brightness (or energy) as a function of wavelength, (3) polarimeters: which determine the degree of alignment of wave vibrations in a beam, and (4) interferometers: which rely on coherent phase relationships to achieve interference effects before performing imaging or spectroscopy. Variations of these instruments exist from X-ray wavelengths to radio wavelengths, although the methods of implementation differ considerably. In general, the descriptions which follow are applicable for UV, visible and infrared wavelengths (UV/O/IR).

5.1.1 Photoelectric photometers

A photometer is a device for measuring the apparent brightness of a source, one of the most fundamental observables. Measurements are usually made after the light has been collected by a telescope and after transmission through the atmosphere. Ideally we would like to measure the power received per square meter integrated over all wavelengths, that is, the irradiance (E) or astronomical flux (S). Instead, measurements of brightness are usually limited to a band of wavelengths selected by means of an optical “filter”. Initially, colored glass filters and the detector's own wavelength-dependent response to light determined the wavebands used, but it is now possible to design and make an optical filter to pass any specific band of wavelengths desired. These filters are known as “interference filters” because they utilize destructive interference in multiple, very thin, dielectric (non-conducting) layers deposited on the glass substrate. We will return to their construction later. For photometry of individual stars, a detector with a single cell can be used, such as a photomultiplier tube (PMT). Several systems of brightness measurements have been in use since the introduction of PMTs. The most familiar of these is the UBV system (U = ultraviolet, B = blue and V = visual or yellow). The original UBV system of Johnson and Morgan (1953) was defined by glass filters approximately centered at wavelengths of 360, 440 and 550 nm and the photoelectric response of a CsSb (S-4) photocathode typical of the RCA 1P21 PMT available at that time. Since then many photometric systems have been developed. Careful work is required to relate one filter system to another and this task forms part of the calibration of the instrument.

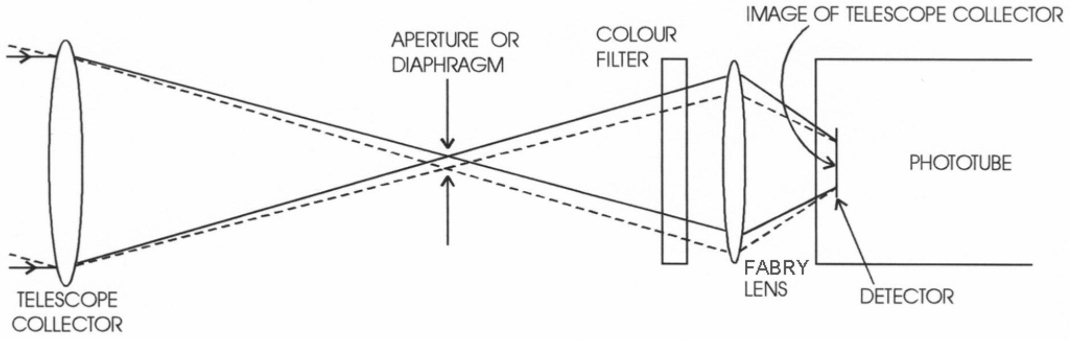


Fig. 5.1 The basic layout of a photometer is shown. Note that the image of the star is not focused onto the detector. By imaging the telescope’s collecting aperture on the detector the signal strength is independent of the position of the star within the diaphragm.

Essential features of a simple photoelectric photometer are shown in Figure 5.1. The instrument is a light-tight box attached to the telescope by means of a flange. At the telescope focus, which falls inside the box, a circular aperture (or diaphragm) isolates a given star. Usually, these apertures will be interchangeable by constructing a series of them in a wheel or slide. The size of the aperture needs to be larger than the image of the star (the "seeing" disk), but not excessively so, otherwise too much "sky" background is included. Another wheel or slide carries a selection of filters such as UBV. The detector is usually a PMT (e.g. a thermoelectrically-cooled GaAs tube). A Fabry lens produces an image of the telescope primary mirror—the collecting aperture—onto the detector and *not* an image of the star. From the Thin Lens Equation, the distance (s) between the Fabry lens and primary mirror is $\sim f_{\mathrm{tel}}$ which is $>> f_{\mathrm{lens}}$, thus the distance between the Fabry lens and PMT is $s' \sim f_{\mathrm{lens}}$. All light rays from the star *must* pass through this pupil image no matter where the star is located in the aperture. This design prevents movement of the illuminated image on the detector which might occur due to drifting of the star image across the diaphragm due to poor tracking. Consequently, the signal is stable and insensitive to variations in detector response over the photocathode surface. As shown in the figure, solid lines trace rays when the star is centered in the focal plane aperture and the dashed lines indicate the light path when the star is at the edge of the aperture. In principle, photometry can be carried out by measuring the current that appears at the anode of the tube as a result of the electron multiplication process, but this is a noisy method at low light levels because of the wide variation in pulse strength (often called pulse height) from identical events at the photocathode. A better solution is to count the pulses emitted by the anode irrespective of their height. The output of the PMT is fed to a pulse amplifier which gives out rectangular voltage pulses of a standard width and with a height proportional to the original signal from the anode. These signals go into a pulse height discriminator which is set to reject the many small pulses associated with amplifier noise. Those pulses that are passed can be counted with digital electronics and supplied to a computer. The PMT sits in a base socket, and a simple resistor chain between the base pins of the PMT can establish the inter-dynode voltages. Usually the photocathode is at a negative potential of about 1600 volts and the anode is at ground potential. Thermoelectric cooling to -20 °C is often sufficient but some PMTs with high dark currents can be cooled with dry ice to -78 °C. Additional information on photoelectric photometers is given in Henden and Kaitchuck (1998).

Until the advent of CCDs, many telescopes were equipped with photoelectric photometers. One of the great advantages of the photomultiplier tube is its speed of response to a change in brightness; typically one thousandth of a second. There are many useful and important applications of high speed photometry. For example, objects such as cataclysmic variables, and pulsars suffer rapid changes in brightness on short time scales. Also, when stars are occulted by the Moon (or a planet) passing in front of them, or satellites of planets are occulted by the planet itself, there is a very rapid dimming which yields the physical dimensions of the sources. PMTs are also ideal for polarization measurements which require very accurate differential photometry.

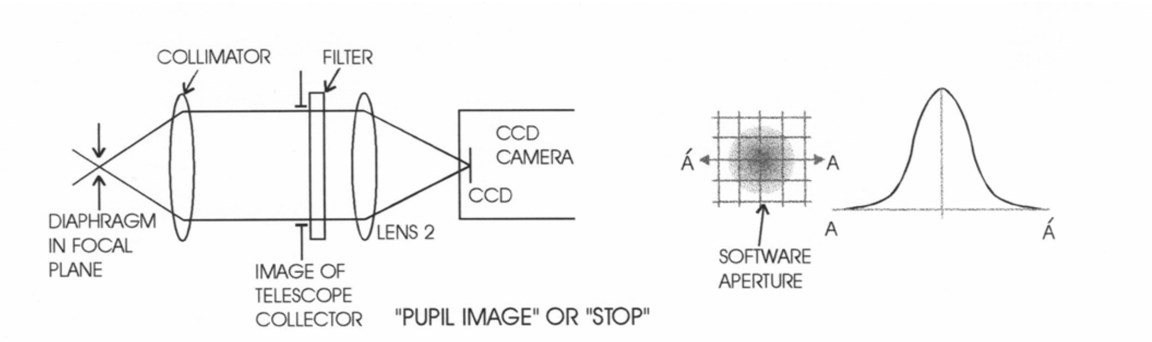


Fig. 5.2 The basic layout of a camera system in which optics are used to collimate the diverging beam from the telescope focus and re-image the field at a different magnification. Filters can be placed in the collimated beam.

5.1.2 Camera systems

Figure 5.2 shows the basic layout of a camera. In the simplest design, the detector (CCD or other array detector) is placed directly in the focal plane of the telescope behind a light-tight shutter. Filters are therefore located in a wheel or slide in the converging beam from the telescope. Care is required to ensure that all filters have the same "optical path", i.e. the product of refractive index and thickness, in order to avoid refocusing the telescope after each filter change. This method works well when the image scale of the telescope is well-matched to the pixel size of the detector, but tends to become challenging for very large telescopes as we will see shortly. An alternative approach, shown in the figure, is to collimate the beam by placing a lens after the focal plane at a distance equal to its focal length ($s = f_{\mathrm{coll}}$) and to re-image the field onto the detector with a camera lens (or mirror) with $s' = f_{\mathrm{cam}}$. This design has many advantages. First, by selecting the focal lengths of the collimator and camera sections one can either magnify or reduce the plate scale; $m = f_{\mathrm{cam}}/f_{\mathrm{coll}}$. Filters of arbitrary thickness can be located in the parallel (collimated) beam and the filters can be placed near the "pupil" image of the primary mirror. In addition, a circular aperture or "stop" can be placed at the pupil image to reject stray light from outside the beam. A stop is extremely important in infrared cameras where the pupil is at cryogenic temperatures and so it becomes a "cold" stop. Of course, in this design, the image of the star may drift by a small amount due to tracking or pointing errors, but photometry is performed "after the fact" on the digital image by selecting an appropriately sized "software aperture" and summing up all the signal. An annulus around the summed region is used to construct an estimate of the sky flux contained in the summed aperture. Thus no separate measurement of the sky is required. Because the star image is spread over many pixels, and as different pixels are used for the sky image, it is essential to have a good procedure to normalize all the pixels to the same sensitivity or gain. This is a general requirement with array detectors which is covered in considerable detail later.

5.1.3 Pixel sampling and matching to the plate scale

There are two issues to be considered when matching the spatial or spectral resolution element to the physical size of the detector pixels; (1) maximizing observing efficiency, meaning more light onto a pixel and therefore keeping the required integration time to a minimum, and (2) accomplishing this task without compromising the ability of the camera system to obtain very accurate brightness measurements (photometry). The spatial resolution element may be determined by seeing conditions or by optical constraints. In general, the image is either critically sampled, meaning that there will be about 2 pixels (also known as the Nyquist limit) across the resolution element, or it will be over-sampled which implies that there may be about five pixels across the resolution element. It is very rare to design a system which is under-sampled deliberately. In a spectrometer, the width of the entrance slit is usually the determining factor. A narrow slit implies higher spectral resolution, but the highest efficiency is achieved when the slit is wide enough to accept the full image diameter.

Using the discussion of Chapter 3, consider first the plate scale of the telescope which is given in seconds of arc per mm (″/mm) by:

$$(ps)_{tel} = \frac{206265}{f_{tel}} \tag{5.1}$$

Here, f_{tel} is the focal length of the telescope in millimeters ($f_{\mathrm{tel}} = D_{\mathrm{tel}} \times F$ where F is the focal ratio or f/number) and the numerical factor is the number of seconds of arc in 1 radian. Plate scales vary considerably. For instance, at the prime focus of the 3.6-m Canada-France-

Hawaii (CFHT) telescope the scale is 13.70 ″/mm, whereas at the Cassegrain focus the scale is 7.33 ″/mm. With an infrared telescope however, the focal ratio is usually larger (slower) so that at the Cassegrain focus of the 3.8 m UK Infrared Telescope (UKIRT) the scale is only 1.52 ″/mm. Our f/16 24-inch reflector at UCLA which, might be typical of many campus telescopes, gives 21.1 ″/mm. For direct imaging, the angle on the sky subtended by the detector pixel is,

$$\theta = (ps)_{tel}\, d_{pix} \tag{5.2}$$

where d_{pix} is the physical pixel size in mm; the pixels are usually square. For CCDs and near infrared array detectors, values range from about 0.009 mm (9 μm) up to about 0.030 mm (30 μm); detector pixels on mid-infrared arrays may be significantly larger. For 20 μm detector pixels we would get 0.27 and 0.15 ″/pixel at the prime and Cass foci of the CFHT respectively, 0.42 ″/pixel on the 24-inch at UCLA and only 0.03 ″/pixel on UKIRT. We need to compare these values with the image quality to determine whether or not some optical magnification is required. For example, for our "roof-top" conditions on the UCLA campus we use 3″ for the average seeing disk, whereas for the instruments on the CFHT and other telescopes on Mauna Kea, Hawaii one might adopt 0.3 - 0.5″ for the seeing! Calculating the required magnification factor can proceed as follows:

- choose a value for the diameter of the seeing in seconds of arc
- decide on the sampling (p = 2 - 5 pixels)
- divide seeing diameter by sampling factor to get angular size of 1 pixel, $\theta_{\text{pix}} = \theta_{\text{see}}/p$ in arcseconds
- given the size of the detector pixels, derive the plate scale at the detector from $(ps)_{det} = \theta_{pix}/d_{pix}$
- the required magnification (m) is then

$$m = \frac{(ps)_{tel}}{(ps)_{\text{det}}} \tag{5.3}$$

where $m = f_{\text{cam}}/f_{\text{coll}}$ as before.

Note that m also defines an Effective Focal Length (EFL = mf_{tel}) for the entire optical system. If $m > 1$, then the optical components are a magnifier, whereas if $m < 1$ (the usual case), then the optics are called a “focal reducer”. We can also relate the pixel size in seconds of arc to the f-number of the focal reducer optics (or simply, "the camera") by

$$\theta_{pix} = 206265 \frac{d_{pix}}{D_{tel}\,(f/number)_{cam}} \tag{5.4}$$

where $(f/number)_{cam} = f_{cam}/D_{cam} = \text{F}_{\text{cam}}$.

5.5 DETECTORS

5.5.1 Classification

We have tried to group instruments into broad classes and it is possible to do the same with detectors. Detectors of electromagnetic radiation are generally grouped into three broad groups:

(1) **photon detectors** in which individual photons release one or more electrons (or other charge carriers) on interacting with the detector material; photon detectors have wide application from gamma-rays to the far-infrared.

(2) **thermal detectors** in which the photon energy goes into heat within the material, resulting in a change to a measurable property of the device, such as its electrical conductivity; thermal detectors have a broad spectral response but are often used for infrared and sub-mm detection.

(3) **coherent detectors** in which the electric field of the wave is sensed directly and phase information can be preserved. The most common form of coherent detection takes advantage of wave interference with a locally-produced field, either before or after conversion of the electromagnetic radiation to an electrical signal. Coherent detectors are used from the far-infrared to the radio.

To distinguish between photon and thermal detectors consider the following. The response of an ideal thermal detector is independent of the spectral distribution of the photons and depends only on the total power absorbed, and therefore its output per watt per unit wavelength interval is flat, independent of wavelength. On the other hand, a photon detector measures the rate of arrival of photons ($N = P/h\nu$), and as the number per second per watt of incident power ($N/P = \lambda/hc$) is proportional to wavelength, its response increases with wavelength up to some maximum wavelength where the photon energy is no longer sufficient to produce a photoelectric event. Photon detectors can be further subdivided into (1) photoemission devices employing the external photoelectric effect in which the photon causes a charge carrier (electron) to be ejected from the material and (2) photo-absorption devices that use the internal photoelectric effect in a semiconductor to free a charge carrier within the material.

The most well-known detector in the photoemission category is the photocathode of a photomultiplier tube (PMT) already described, in which an electron is emitted from the photocathode surface and subsequently amplified by a cascade of impacts with secondary surfaces before being detected as a charge pulse. Photoemissive materials with large work functions can provide excellent detectors far into the ultraviolet. Most importantly, it is possible to create ultraviolet imaging devices based on this process. For example, long, narrow curved tubes or "microchannels" of lead oxide (PbO) can perform the same function as the secondary surfaces in a PMT resulting in a large pulse of electrons emerging from the end provided there is a potential gradient. Such channels can be packaged very close together (like straws in a box) to make a two-dimensional array of microchannels. We will discuss this important class of devices later under UV imaging methods. Finally in this category, some materials emit other lower energy photons (fluoresce) rather than electrons which enables detection of the more energetic photon by a process called "down-conversion"; we will see that this has applications to CCDs later.

Photo-absorption is the largest category, with many possible outcomes, including chemical change (as in photography). In this book we are concerned mainly with absorption processes in semiconductor devices, and there are essentially two basic types of interactions, the photoconduction effect and the photovoltaic (or photodiode) effect. The photoconductor is composed of a single uniform semiconductor material in which the conductance is changed by the creation of free charge carriers in the material when photons are absorbed. There is usually always an external applied electric field. In the photodiode (or photojunction), internal electric fields and potential barriers are created by suitable junctions between different materials or deliberate variations in the electrical properties of the material so that photo-generated carriers in these regions respond to those fields. Before proceeding further it is important to review the properties of semiconductors in general.

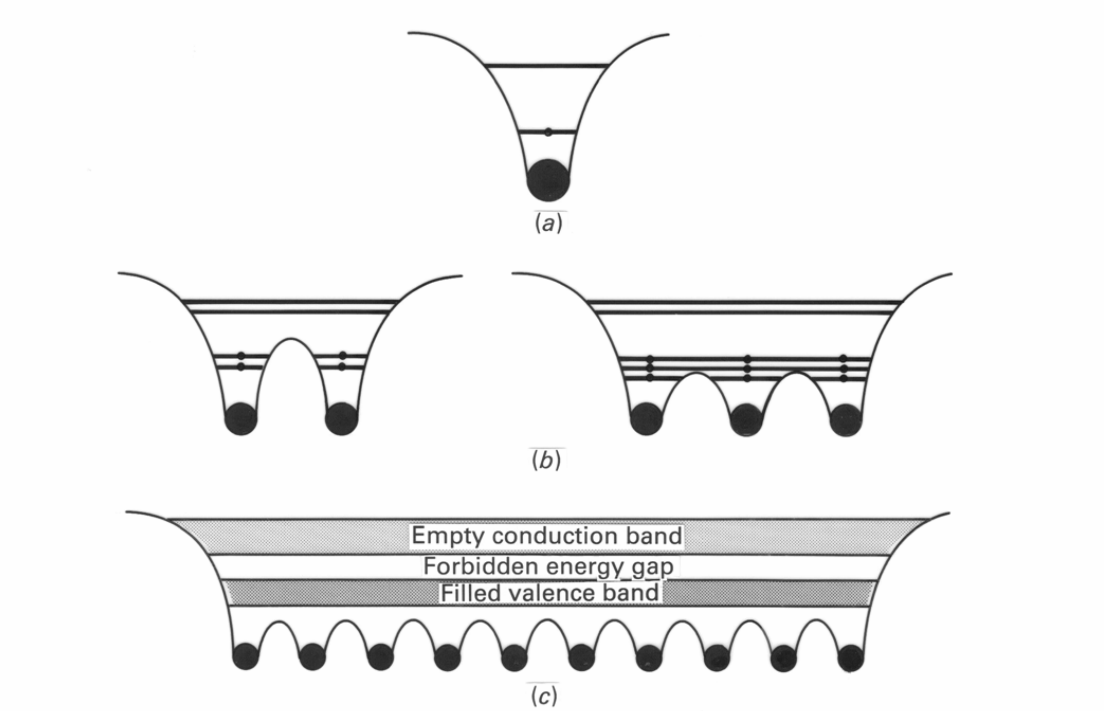


Fig. 5.13 Simplified schematic of the formation of a band gap in a semiconductor crystal.

5.5.2 Semiconductors

The properties of any solid material depend on both the atomic structure of the atoms of the material and the way the atoms are arranged within the solid, that is, the crystal structure. Electrons can exist in stable orbits near the nucleus of an atom only for certain definite values of their energy. When individual atoms come close together to form a solid crystal, electrons in the outermost orbits, or upper energy levels, of adjacent atoms interact to bind the atoms together. Because of the very strong interaction between these outer or "valence" electrons, the outer orbits and therefore the upper energy levels are drastically altered. The result is that the outer electrons are shared between the different atomic nuclei. A simple diagram depicting the "energy levels" of the electrons for a combination of two atoms would therefore have two permitted levels near the core of each atom. A combination of three atoms would have three levels near the core because the outer electrons of all three atoms can be shared. The higher, unoccupied orbits would also split, indicating that they too can in principle take two or three electrons. Even the tiniest sliver of a real crystal will contain many hundreds of millions of atoms, and so there are a huge number of split levels associated with each atom in the crystal because of the sharing of outer electrons. In other words, the energy levels or orbits are spread out into a "band". The lowest band of energies, corresponding to all the innermost orbits of the electrons, is filled with electrons because there is one electron for each atom. This band of allowed, filled energy levels is called the “valence band”. Conversely, the upper energy band is empty of electrons because it is composed of the combined unoccupied higher energy levels or orbits of the individual atoms in the crystal. It is called the “conduction band” for reasons that will become apparent. Thus, the individual atoms have a gap between the permitted inner and outer orbits, that is, a gap in energy between the inner filled levels and the outer unoccupied levels. The energy region between the valence band and the conduction band in the crystal must be a “forbidden energy gap” (E_G). Figure 5.13 summarizes this description. Note that the crystal must be pure and contain atoms of only one kind otherwise additional energy levels corresponding to those atoms will be formed. More importantly, the periodic or repetitive crystalline structure must be unbroken to avoid distortions in the energy levels caused by abnormal sharing of electrons. Of course, in practice both of these conditions are violated in real crystals, and departures from the simplified model presented here contribute to degraded performance of devices such as transistors and CCDs. In metals, the valence and conduction bands overlap and so any of the many valence electrons are free to roam throughout the solid to conduct electricity and heat, and to move in response to the force of an electric field; an electric field could be produced by attaching a battery to both ends of the piece of metal. An insulating material on the other hand, has a highly ordered structure and a very wide forbidden energy gap. The conduction band is totally empty of electrons and so cannot contribute to an electrical current flow. Electrons in the completely filled valence band cannot move in response to an electric field because every nearby orbit is occupied.

In a semiconductor, a few electrons can be elevated from the valence band to the conduction band across the forbidden gap merely by absorbing heat energy from the random, microscopic, jostling motions of the crystal structure at normal "room" temperature. Thermal energy is given approximately by

$$E_{\mathrm{th}}\,(\mathrm{eV}) = kT = 0.026\ (\mathrm{T}/300)\ \mathrm{eV} \tag{5.45}$$

where k is Boltzmann’s constant and T is the absolute temperature. At room temperature (T=300 K) the thermal energy is quite small at 0.026 electron volts. Electrons promoted to the conduction band can then conduct electricity, that is, they are free to move under the influence of an electric force field. Interestingly, the corresponding vacancies or “holes” left in the valence band allow it to contribute to electrical conductivity as well because there is now somewhere for electrons in adjacent atoms to go; descriptions of solid-state devices therefore refer to “electron-hole” pairs.

Because the number of electrical charge carriers (electrons in the conduction band, holes in the valence band) is much less than in the case of a metal, semiconductors are poorer conductors than metals but better than insulators. The width of the forbidden energy gap in semiconductors is an important quantity. Most semiconductor crystals have band gap energies around 1 eV, but the range is from almost 0 to about 3.5 eV. As shown above, 1 eV is roughly 38 times larger than the thermal or heat energy in the crystal atoms at room temperature. Remember also that visible light photons have energies around 2.25 eV (for 550 nm). As the number of electrons which can be promoted to the conduction band by absorbing heat will vary with the temperature of the crystal, typically as exp($-E_G/2kT$), those semiconductors with larger band-gaps are preferred because transistors and other devices made from them will be less sensitive to environmental changes. For this reason silicon is preferred to germanium. If the semiconductor is cooled to a low temperature, random elevation of valence electrons to the conduction band can be virtually eliminated.

Table 5.1 is a section of the periodic table of the elements showing that the primary semiconductors like silicon and germanium belong to the "fourth column" elements, which also includes carbon. Each of these elements has four valence electrons. Compounds of elements on either side of the fourth column can be formed and these alloys will also have semiconductor properties; gallium arsenide (GaAs) and indium antimonide (InSb) are III-IV (or "three-four") compounds and mercury-cadmium-telluride (HgCdTe) is one possible II-VI (or 2-6) compound. Column numbers indicate the number of valence electrons. Small numbers with the symbols are the atomic numbers (number of protons or electrons).

Table 5.1 Part of the periodic table of elements showing the location of semiconductors

IB	IIB	IIIA	IVA	VA	VIA	VIIA
		^{5}B Boron	^{6}C Carbon	^{7}N Nitrogen	^{8}O Oxygen	
		^{13}Al Aluminum	^{14}Si **Silicon**	^{15}P Phosphorus	^{16}S Sulfur	^{17}Cl Chlorine
^{29}Cu Copper	^{30}Zn Zinc	^{31}Ga Gallium	^{32}Ge **Germanium**	^{33}As Arsenic	^{34}Se Selenium	^{35}Br Bromine
^{47}Ag Silver	^{48}Cd Cadmium	^{49}In Indium	^{50}Sn Tin	^{51}Sb Antimony	^{52}Te Tellurium	^{53}I Iodine
^{79}Au Gold	^{80}Hg Mercury	^{81}Tl Thallium	^{82}Pb Lead	^{83}Bi Bismuth		

When a photon is absorbed in the crystalline structure of silicon, its energy is transferred to a negatively charged electron, the photoelectron, which is then displaced from its normal location in the valence band into the conduction band. When the electron reaches the conduction band it can migrate through the crystal. Migration can be stimulated and controlled by applying an electric field to the silicon crystal by means of small metal plates called “electrodes” or “gates” connected to a voltage source. Absorption of photons in silicon is a function of the photon energy (and hence wavelength). The photon flux at depth z in the material is given by

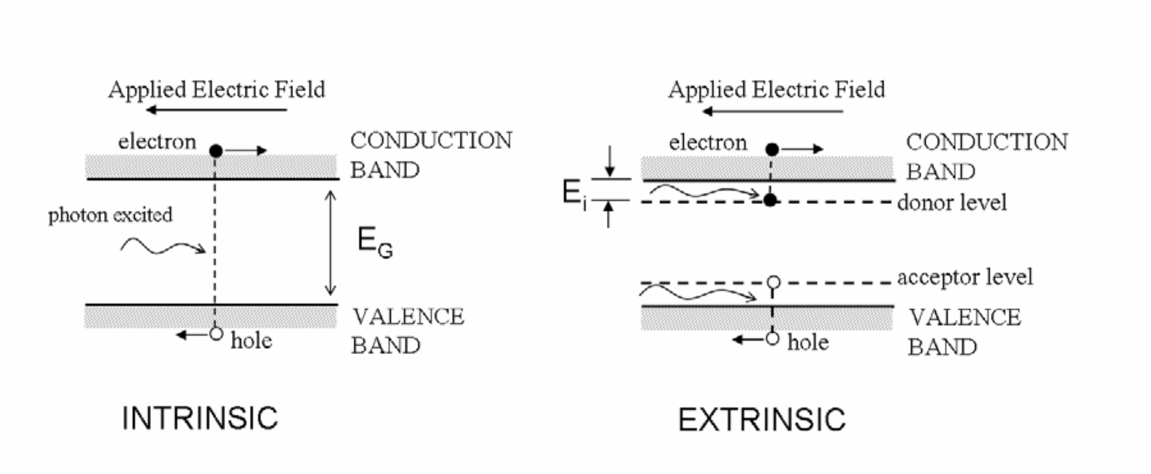


Fig. 5.15 An intrinsic band gap and showing the location of energy levels within the band gap due to doping to form an extrinsic semiconductor.

Table 5.3 Extrinsic semiconductors, doping material and long wavelength cut-off.

Base	:Impurity	λ_c (μm)	Base	:Impurity	λ_c (μm)
Silicon (Si)	:In	8.0	Germanium (Ge)	:Au	8.27
	:Ga	17.1		:Hg	13.8
	:Bi	17.6		:Cd	20.7
	:Al	18.1		:Cu	30.2
	:As	23.1		:Zn	37.6
	:P	27.6		:Ga	115
	:B	28.2		:B	119.6
	:Sb	28.8		:Sb	129

5.5.3 Photoconductors and photodiodes

Photoconductor: This is the simplest application of a semiconductor for detection of photons. A typical photoconductor arrangement is shown in Fig. 5.16. Photons are absorbed and create electron-hole pairs. If the material is extrinsic rather than intrinsic, then E_i must be substituted for E_G. Also, for extrinsic materials there are limits on solubility of the dopants and high concentrations introduce unwanted conductivity modes such as “hopping” which involves conduction between neighboring impurity atoms without raising an electron to the conduction band. In the discussion below we assume that the semiconductor has been cooled to eliminate thermally-generated charges. In practice, both electrons and holes contribute to the photocurrent, but it is usually the electrons that dominate.

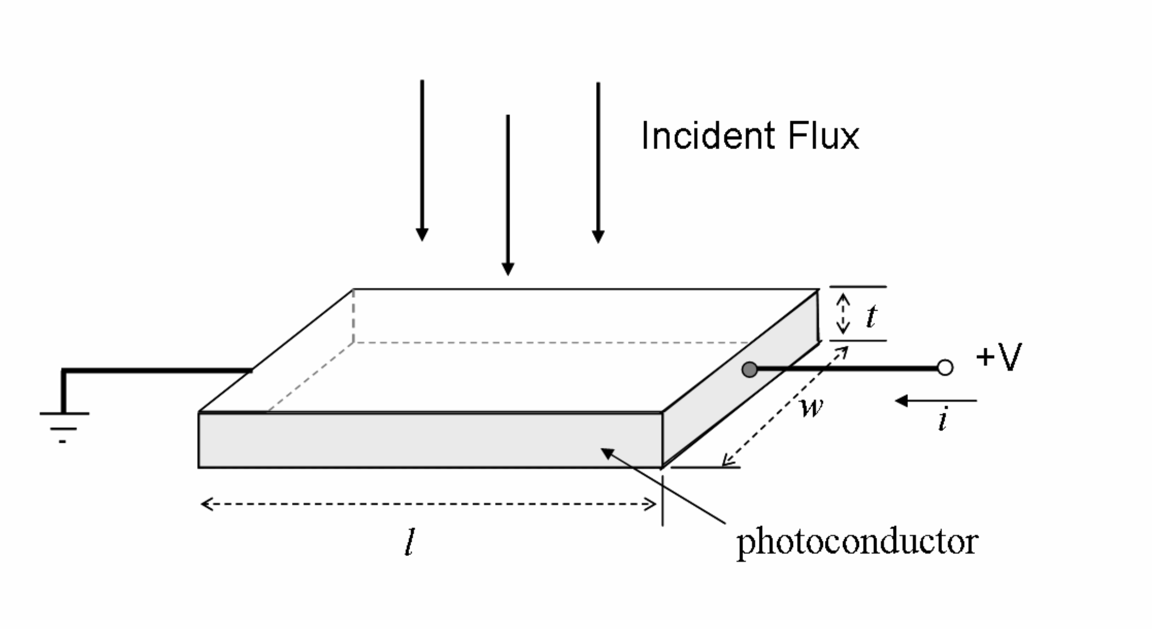


Fig. 5.16 The basic construction and operation of a semiconductor used in photoconduction mode.

The average photocurrent (*I*) between the terminals that is generated by an incident flux with power *P* (watts) is given by

$$I = (e\eta P/hv)\,(\mathrm{v}\tau/l) \qquad (5.48)$$

In this expression η is the quantum efficiency and P/*hv* is just the photon arrival rate. The quantity τ is called the mean carrier lifetime and measures how long the photogenerated charge exists before recombination. Values are usually less than to much less than a few milliseconds but depend on doping and temperature. The average charge carrier velocity is v, which is related to the applied electric field across the photoconductor E = V/*l* by v = μE where μ is called the mobility of the charge carrier. Thus, *l*/v is the transit time across the device from one terminal to the other, and the quantity G = vτ/*l* is just the ratio of mean carrier lifetime to transit time. It is known as the "photoconductive gain". The response of the detector (in amps per watt or volts per watt) is just *I*/*P* or *V*/*RP* where V is the bias voltage across the photoconductor and the resistance *R* due to the photocurrent is $l/\sigma A$ and the conductivity $\sigma = ne\mu$, where *n* is the average density of carriers. It follows that $S = (e\eta G/hc)\lambda$. Finally, the root mean square noise for a photoconductor is given by $\sqrt{}(4eGIB)$ where *B* is the electrical bandwidth of the measurement.

6

Designing and building astronomical instruments

There are many important factors and constraints to be aware of when developing new astronomical instrumentation, whether for small or large telescopes. Of course, complete engineering details are beyond the scope of this book, or any one book, but the following sections will at least provide an appreciation for what is involved and serve as a starting point for newcomers to instrument building.

6.1 BASIC REQUIREMENTS

Understanding the application is the very first step. What are the science goals? Sometimes the goals are fairly general, such as "provide the most sensitive camera with the widest possible field of view consistent with median seeing conditions". This approach is reasonable, on the grounds that the uses of such an instrument are so numerous. On the other hand, the science goals may be quite specific, such as "provide an instrument to search for planets around other stars via Doppler velocities in the 3 m/s range", or "provide an instrument to carry out a survey of redshifts ($z > 1$) of a very large statistical sample of faint galaxies". In these cases the spectrographs involved would be quite different from each other and different from a conventional "workhorse" spectrograph. For the planetary search the spectral resolution needs to be very high and this instrument must provide exceptional long-term stability, whereas for the faint galaxy survey the resolution is much lower, but numerous slit-masks and/or optical fibers are needed to provide a large multi-object advantage. Clearly, the choice of instrument and the details of the design will depend on the kind of science to be done. If it is imaging science, what field of view is required? What are the angular resolution and the wavelength range requirements? Is temporal resolution an issue? Are the measurements going to be read-noise limited or background-limited? The basic requirements must come from the science goals, but be aware of creating a monster with many heads! Too many scientific options in one instrument can be disastrous. In turn, the science requirements are used to generate a “specification” for the instrument. Candidate designs can then be analyzed in a Conceptual Design phase (also called a System Design phase) and the best or most appropriate design selected. There is always more than one way to achieve the desired goals, and changes in technology result in fresh approaches and new ways to improve older methods.

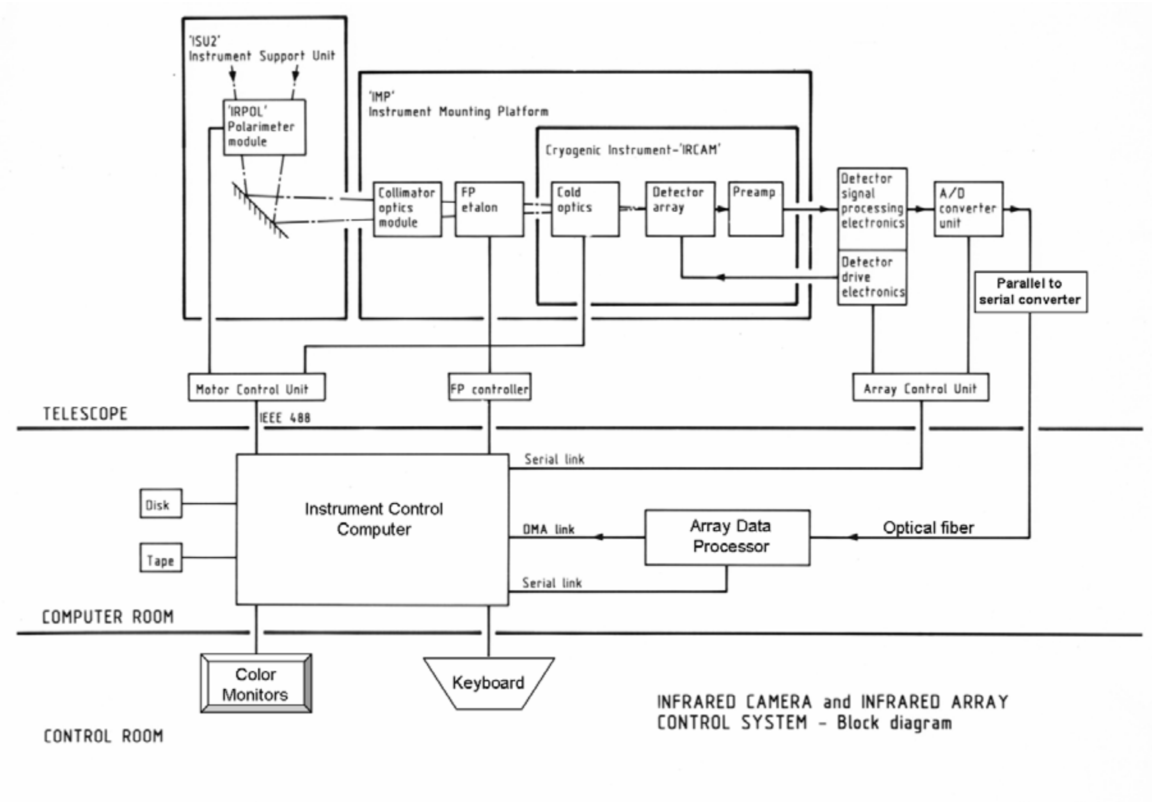


Fig. 6.1 A block diagram layout of an entire camera system for a large telescope. The illustration is for IRCAM, the first infrared camera system developed for the 3.8 m UK Infrared Telescope.

6.2 OVERALL SYSTEM LAYOUT

Laying out all the essential “building blocks” and their interconnections in pictorial form is the next step in the design process. Invariably a modular approach to the instrument works well. A very simple “block” diagram adapted from the author’s first infrared camera is shown in Fig. 6.1. The technology is old but the principle is not, and unlike complex modern instruments this diagram is readable on a small page. In its most basic features this infrared camera system is similar to all other imaging systems irrespective of the detector.

Although it is hard to generalize, certain building blocks are almost always present in an astronomical instrument. These include:

(1) the detector (the photon sensor itself and circuitry packaged close by)
(2) an opto-mechanical system (lenses, mirrors, filters, gratings, fibers, mounts)
(3) an enclosure and cooling system (for the detector and other parts of the instrument)
(4) signal-processing hardware (e.g. amplifiers and noise-suppression circuits) and the analog-to-digital converter (ADC or A/D)
(5) detector "drive" electronics (pulsed and dc bias circuits)
(6) timing logic and synchronization circuits
(7) a "motion control" system & "housekeeping" system (e.g. temperature control)
(8) an electronic interface to a computer (e.g. ethernet, telemetry)
(9) a host computer and peripherals
(10) an image display system and image processing/restoration software

These ten items form the basis of a great many astronomical instruments employing some form of electronic imaging device. In fact, the above list could apply to almost any form of detector system used in astronomy if the items are understood in their most general sense. At the heart of all instruments is the detector. Usually, it is the performance of the detector system that determines whether the instrument is "state-of-the-art" or not.

6.3 OPTICAL DESIGN

In practice, the optical design of a modern astronomical instrument is likely to be carried out by a professional designer with considerable experience of the application. However, some training in optical design for all young observational astronomers and for anyone interested in building astronomical instruments is very valuable. Understanding the issues, recognizing problems and being able to communicate requirements is important. Courses are available which teach practical methods of ray tracing and optical design. Excellent texts on optics include Born & Wolf (1999), Fischer & Tadic (2000), Hecht (2001), Kingslake (1978), Schroeder (2000) and Smith (2000).

6.3.1 First order to ray tracing

A good strategy is to break down the optical design into the following steps and stages:
(1) "first order" requirements
(2) constraints
(3) performance specification
(4) ray-tracing and optimization
(5) tolerance analysis

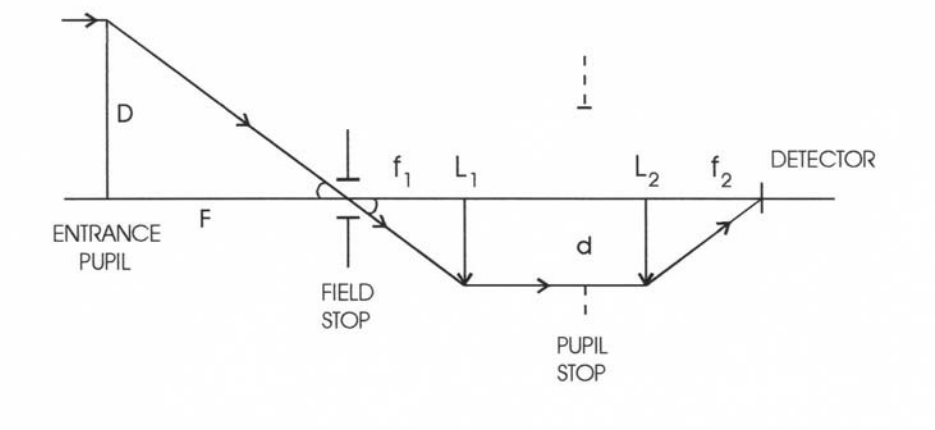


Fig. 6.2 A "first order" optical layout to collimate the beam from the telescope and then re-image the field of view onto a detector is shown. Pupil size and location is determined together with magnification and field of view constraints.

By "first order" requirements is meant a simple design, using "thin lens" formulae, and known facts about the system, such as the f/number and plate scale of the telescope, the object/image distances and the location of any pupils formed, the required field of view, and the desired magnification or image scale at the detector. A typical "first order" layout is shown in Fig. 6.2. An image of the sky is formed at the telescope focal plane, and this "object" is then re-imaged by the instrument optics onto the detector.

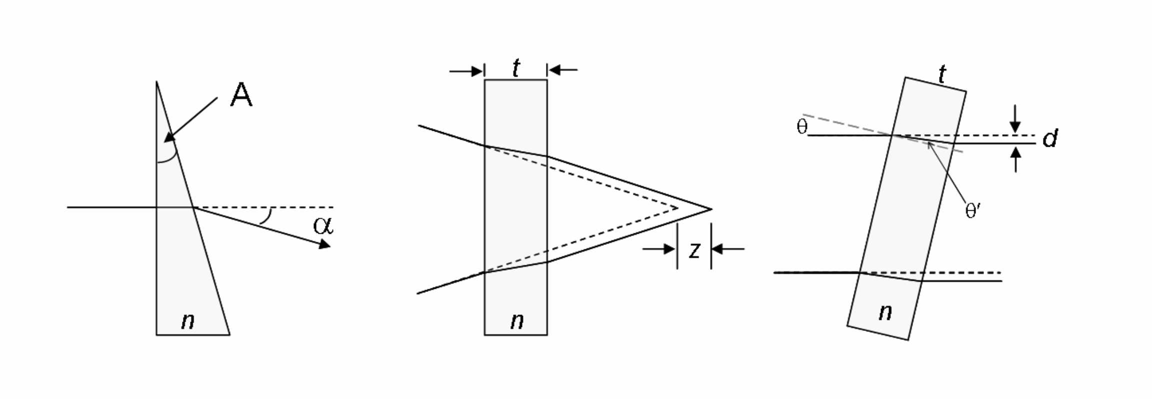


Fig. 6.3 Showing the effect of wedges and tilted plane-parallel plates on the optical beam.

Depending on the ratio of the focal lengths of the transfer optics, there will be a change in magnification of the final image. In combination with all optics upstream, the transfer optics will form an image of the primary mirror of the telescope at some location. This image is the entrance pupil. At this stage, the simple equations of elementary optics given in Chapter 3 can be employed. The "thin lens" equation and the equivalent form for single mirrors can now be applied. It is not likely that simple lenses will be adequate when aberrations are considered, but the required power, beam sizes and detector scales can be estimated. Likewise, the following simple relationships on displacements and deviations illustrated in Fig. 6.3 can prove useful in laying out the initial design:

$$\begin{aligned} &\alpha \approx (n-1)A && \textit{thin wedge} \\ &z = \frac{(n-1)t}{n} && \textit{parallel plate in converging beam} \\ &d = t\sin\theta\left(1 - \frac{\cos\theta}{n\cos\theta'}\right) && \textit{displacement by parallel plate} \end{aligned} \tag{6.1}$$

where α is the angular displacement caused by a wedge of small angle *A* (angles in radians), *z* is the longitudinal (focus) displacement caused by a plane parallel plate of thickness *t* perpendicular to a converging (or diverging) beam and *d* is the lateral displacement caused by a plane parallel plate at angle θ in a parallel (collimated) beam. These formulae are useful when considering the effects of filters, entrance windows to vacuum enclosures, dichroic beam-splitters and polarizing beam-splitters.

Next, identify and list all the known constraints on the design. For example, the wavelength range, the transmittance goals, restrictions or limits on scattered light which are probably driven by the signal-to-noise calculations and the science goals, the desired back focal length and other optical-mechanical packaging issues (size, weight, thermal mass), polarization effects (due to gratings, or birefringence in crystals or boundary conditions), environmental concerns (thermal changes, shock and vibration), ability to test and align the optics and finally, the cost of fabrication.

Except in a few cases, it will not be possible to complete the design of the instrument by purely analytic means. The final step is therefore to enter the prescription into a "ray tracing" program and develop a more sophisticated model. Many excellent programs are available. One of the older packages, and an industry standard, is called Code V ("code five") from Optical Research Associates (ORA). Other very popular packages include OSLO from Sinclair Optics and ZEMAX by Zemax Development Corp.; the latter was the first specifically written for a Windows user-interface. Illumination packages include ASAP from Breault Research Organization (BRO), LightTools from ORA, and ZEMAX (engineering edition). Beware that a ray tracing program cannot design a system for you, it can only trace what you enter, so the first order analysis is very important and it often helps to begin with an existing design and modify it. Most ray tracing programs will provide an algorithm which attempts to optimize a given design or search for different designs within constraints which you can control. In this way you can "explore" some options, but be prepared to use up a lot of computing time. A ray tracing program can assist the designer in studying what the effect of these variations might be, and what compensation techniques (such as refocus) can be applied. It is important to understand the limitations of a given optical design, in order to assess the impact on the astronomical goals, as well as the impact on cost and manufacturability.

6.3.2 Aberrations

We have already alluded to optical imperfections in earlier chapters. A perfect optical system would obey the paraxial equations irrespective of the value of θ in Snell's Law. Imperfect images caused by geometric factors are called aberrations. To see the impact of larger values of θ we can expand the sine functions in Snell's Law in a Taylor series.

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \frac{\theta^9}{9!} - \ldots \tag{6.2}$$

Retaining only the terms in theta to the first power (i.e. sin θ is replaced by θ), we arrive at the familiar "first order" or "paraxial" equations shown previously (Chapter 3). However, by including the third order terms (i.e. replacing sin θ with $\theta - \theta^3/6$) we are led to a useful set of equations for describing lens aberrations as departures from paraxial theory; these equations are called the Seidel or "third-order" aberrations. For monochromatic light, German mathematician and astronomer Philipp von Seidel (1821-1896) classified aberrations as spherical aberration, coma and astigmatism which all affect image quality, and distortion and field curvature which affect the image position. In multi-color (polychromatic) light there is also chromatic aberration and lateral color. For completeness and for their value in recognizing potential problems with astronomical instruments, each of these well-known effects is summarized here.

Spherical aberration is caused by the fact that a spherical surface, whether on a lens or a mirror, is geometrically-speaking the wrong shape to ensure that all light rays converge to a focus at the same point. Spherical aberration is an "axial" aberration because rays at greater and greater radii from the center of a positive lens (the marginal rays) focus closer and closer to the lens; this is positive spherical aberration. Negative lenses have negative spherical aberration. The difference between the marginal and the paraxial focal points along the axis is called the longitudinal spherical aberration, while the difference between the paraxial focus and the marginal ray intercept at the paraxial focal plane is called the lateral or transverse spherical aberration. A focal plane or detector placed on the axis will see a large blurry image instead of a point source. (This effect occurred on the Hubble Space Telescope because the primary mirror was over-polished toward the edge by a mere 2 μm from its designed hyperboloid.) The circular image obtained has a minimum size called the "circle of least confusion" which is located slightly closer to the lens (for a positive lens) than the paraxial focus and roughly half-way between the paraxial and marginal focal points. For three simple cases the angular diameter 2β (in radians) of the blur circle is given by:

$$\begin{array}{ll} \textit{Spherical Mirror}: & 2\beta = 1/64F^3 \\ \textit{Parabolic Mirror}: & 2\beta = 0 \\ \textit{Simple Lens}: & 2\beta = n(4n\text{-}1)/128(n+2)(n\text{-}1)^2F^3 \end{array} \qquad (6.3)$$

Here F is the focal ratio and n is the index of refraction.

Spherical aberration in a lens can be minimized by varying the shape or "bending" of the pair of surfaces, because many different combinations of curvature produce the same focal length. The shape factor of a lens is defined in terms of its radii of curvature (R_1, R_2) by

$$q = (R_2 + R_1)/(R_2 - R_1) \qquad (6.4)$$

and it can be shown that,

$$q = -2(n^2-1)\, p/(n+2) \qquad (6.5)$$

gives the shape factor to produce minimum spherical aberration where p is the position factor $(s' - s)/(s' + s)$ or in terms of focal length f, $p = (2f/s) – 1$; if the lens is used in parallel light then $p = -1$. Alternatively, the lens power (P=1/f) can be "split" between two or more "slower" lenses (larger f/number). Since the angular diameter of the blur circle is inversely proportional to the cube of the focal ratio (Eq. 6.3), then splitting an *f*/2 lens into a pair of *f*/4 lenses reduces the spherical aberration of each by a factor of 8, and the combination has about 0.5 of the original spherical aberration. A doublet with a positive and a negative element can further neutralize spherical aberration because, varying as the *cube* of focal length, the spherical aberration changes sign with the sign of the focal length. If the curvature of the lens or mirror surface departed from that of a sphere in such a way as to compensate for the difference between sin θ and θ, then spherical aberration would be eliminated! Both marginal and paraxial rays would focus at the same point (for an on-axis object placed at infinity). A parabolic (conic) surface achieves this ideal, and although more expensive, non-spherical shapes can now be polished into lens surfaces.

7

Charge-coupled devices

At the heart of all astronomical instruments is some form of detector to convert electromagnetic energy into an electrical signal. Having indicated that the dominant detector in modern astronomy is the charge-coupled device (CCD), it is important to consider these remarkable detectors in more detail. We begin with a brief historical review of CCDs from invention to present day, an amazing story by itself, and then cover the basic principles of how CCDs work. Practical details will follow in the next chapter.

7.1 THE EARLY YEARS

7.1.1 Invention and development

As mentioned in Chapter 1, the charge-coupling principle was invented in 1969 by Willard Boyle and George Smith and demonstrated in a simple one-line nine-electrode device by Gil Amelio, Mike Tompsett and George Smith at the Bell Laboratories in New Jersey, USA. Larger image-forming devices of 100x100 pixels were not introduced until 1973 and Boyle and Smith received the basic patent at the end of 1974. From the original small arrays available around 1973, CCDs have come a long way. Formats of 2048 × 4096 pixels[1] with no wires or structure on three sides are readily available with most observatories using CCD "mosaics" composed of many such "close-butted" devices. Figure 7.1 shows an historical collection of CCDs including some large format devices.

7.1.2 The astronomical push

Many astronomy-related groups familiar with imaging technology—usually with vidicon-type systems—were alert to the potential of CCDs in the early seventies. Gerald M. Smith, Frederick P. Landauer and James R. Janesick of the Advanced Imaging Development

[1] These strange numbers are simply powers of 2; 2^{10} = 1024, 2^{12} = 4096. There is no fundamental reason to use these numbers.

Group at the NASA Jet Propulsion Laboratory operated by the California Institute of Technology (Caltech) in Pasadena, and Caltech scientist James Westphal (1930-2004) were among the first to recognize the potential advantages of such an imaging device for astronomy and space applications.

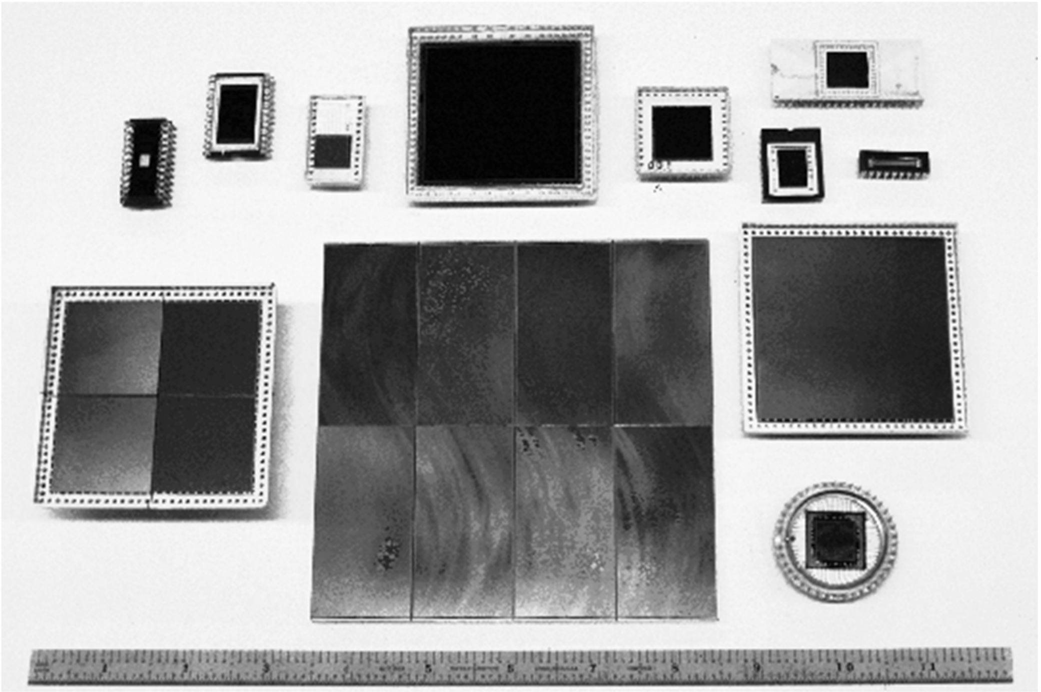


Fig. 7.1 A collection CCDs including eight large format (2Kx4K) devices butted together to form a 64 Megapixel mosaic. Credit: Gerry Luppino.

In 1973, JPL joined with the National Aeronautics and Space Administration (NASA) and with Texas Instruments (TI) Incorporated (Dallas) to initiate a program for the development of large-area CCD imagers for space astronomy, in particular for the proposed Galileo mission to Jupiter. Originally scheduled for 1981, the Galileo spacecraft was finally launched in 1989 and arrived at Jupiter in December 1995. This incredibly successful mission ended on September 21, 2003 when the satellite was directed to plunge into Jupiter's atmosphere.

During the period 1973 to 1979 Texas Instruments (TI) developed CCD arrays of 100 × 160 and 400 × 400 pixels, then 500 × 500 pixels and finally an 800 × 800 pixel array. Testing and evaluation of these devices was carried out at JPL by Fred Landauer and by a young engineer named Jim Janesick, who just happened to be an amateur astronomer. In 1974, Jim attached a 100x100 Fairchild CCD to his small 20.3 cm (8-in) telescope and succeeded in imaging the Moon (Janesick 2001). Having already approached one astronomer at a national institute about testing a CCD on a professional telescope, and having been turned down, Jim luckily met and teamed-up with Dr. Bradford Smith a planetary scientist at the University of Arizona's Lunar and Planetary Laboratory. In early

1976 they obtained the first astronomical imagery with a charge-coupled device on a professional telescope.

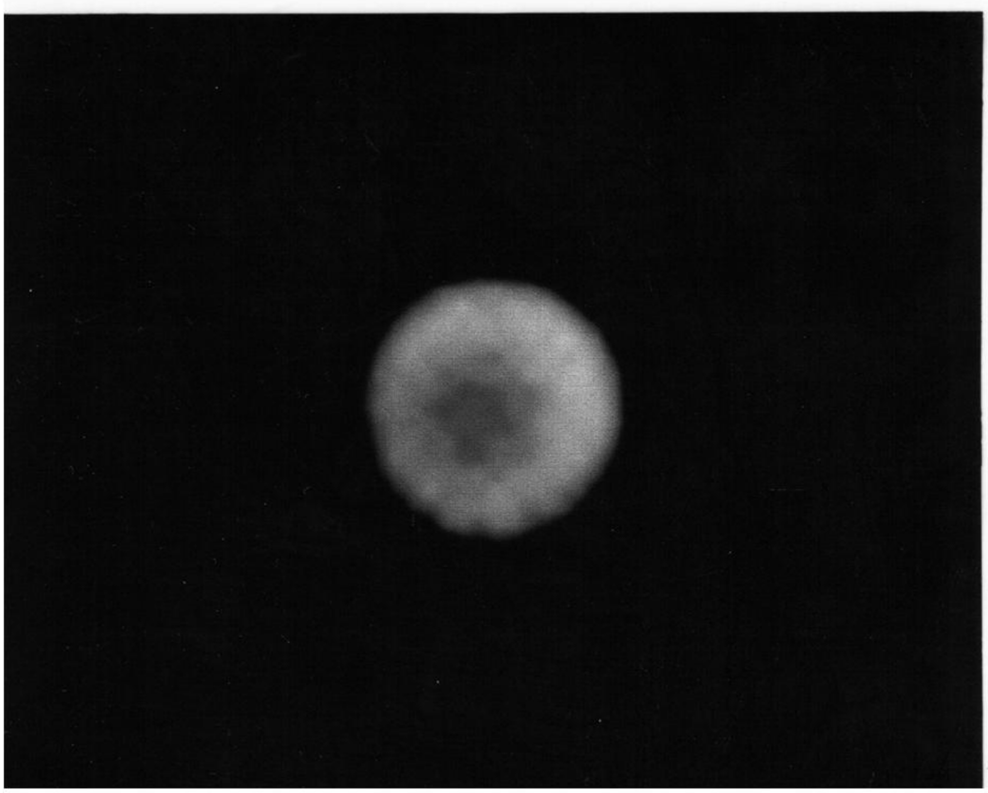


Fig. 7.2 An image of the planet Uranus obtained in 1976 through a methane-band filter with a Texas Instruments CCD. This was the first astronomical CCD image on a professional telescope. Photo courtesy of Brad Smith.

Using the 61-inch telescope designed for planetary imaging on Mt. Bigelow in the Santa Catalina Mountains near Tucson (Arizona), CCD images of Jupiter and Saturn were obtained using a special filter to pick out methane gas in the atmospheres of those giant planets. When the team turned the telescope to Uranus (Fig. 7.2), they immediately thought something had gone wrong. It looked like a "donut"! After checking focus and everything else Brad realized that it must be correct, they were observing "limb-brightening" of Uranus in the methane band for the first time. As Brad Smith recalled vividly to me, "all who participated and who saw those images agreed that the potential of the CCD was superior to other imaging equipment of the time." I joined the Lunar and Planetary Lab in 1977 and got a chance to use one of those early systems with Brad’s post-doctoral researcher, Harold Reitsema (now at Ball Aerospace).

At about the same time, NASA awarded contracts for the procurement of instruments for an earth-orbiting Space Telescope. One of the people awaiting NASA's decision was John Lowrance at Princeton University. John was working with SEC vidicon technology in the 1970's when CCDs were invented. SEC (or secondary electron conduction) vidicons seemed like the most appropriate imaging devices — at least in terms of their wavelength of response — for space ultraviolet/optical experiments such as the Orbiting Astronomical Observatory (OAO). While NASA and various advisory bodies deliberated, John continued to pursue the development of SEC vidicon systems. Finally, after many delays, NASA abandoned the OAO series in favor of the Space Telescope. By that time CCDs had been invented, and JPL had begun their studies. Following a crucial meeting at which CCD results were demonstrated, the initial plan to utilize SEC vidicons for the Space Telescope was dropped and the concept of Principal Investigator instrument teams was introduced. A proposal from a team led by planetary scientist James Westphal (1930-2004) of Caltech in collaboration with JPL was accepted for the inclusion of CCD cameras on the Space Telescope. Westphal had heard about CCDs at a committee meeting a couple of years earlier. When he insisted on knowing more about the JPL results, the chairman (Bob O'Dell) sent him along to JPL to see for himself! There he met Fred Landauer and learned that CCDs were indeed very low-noise devices; 100 electrons (e) noise had been observed and 30-e was predicted. On his return to Caltech he mentioned these numbers to colleague, Jim Gunn of Princeton University, who was getting ready for a major project involving vidicon technology, and whose instant reaction was "that will revolutionize astronomy if it is true !"

John Lowrance, like Jim Westphal, moved away from vidicon technology and began working with CCDs. Luckily there was a key player in the game right on his doorstep, the Electro-optics Division of the Radio Corporation of America (RCA) in nearby Lancaster, Pennsylvania. At the head of the RCA group working on CCDs was Dick Savoye, and Dick was enthusiastic about the astronomical applications, and moreover he believed that the technology possessed by RCA would yield devices extremely sensitive to blue light, as later demonstrated by the superb blue response of the thinned, backside-illuminated 512 × 320 RCA CCD. John Lowrance at Princeton and John Geary at Harvard each established good relations with RCA and began testing these devices in the late seventies. John Geary, having first tried an unthinned device on the 1.5-m and 60-cm telescopes on Mt. Hopkins in April 1980, visited the RCA facility in Lancaster Pennsylvania shortly thereafter to show them the splendid results obtained so far, and urge them to provide him with a thinned CCD. He received one of the very first thinned backside illuminated CCDs manufactured by RCA; this device was considered a reject and was lying in the desk drawer of RCA engineer Don Battson. John put it on the telescope on Mount Hopkins in September 1980 and there it remained for almost a decade!

Meanwhile, the Texas Instruments (TI) chips evolved through a program of systematic development toward the eventual goal of an 800 × 800 array. One of the key figures on that program from the outset was Morley Blouke. Several approaches to the design and fabrication of CCDs were tried. A major constraint was that the device must be able to survive the harsh radiation environment around Jupiter. Therefore, two different constructions evolved called the “buried-channel” and the “virtual phase” CCD. Tens of thousands of CCDs were being manufactured under contract to NASA (the final number was 75,000) and JPL realized—as Jim Janesick stated in a proposal to the Director of JPL in October 1976—that there was "... a need to expose and familiarize astronomers and scientists to the capabilities of the CCD for use in planetary observation and stellar studies."

Around this period (1974-77) other companies were also beginning to develop CCDs. The first company to actually market a high-quality device was a division of Fairchild Semiconductor (Milpitas, CA) which produced a 100 × 100 CCD in 1974; this is the chip

Jim Janesick attached to his own small telescope in 1974 to image the Moon. James Early of Fairchild was a strong advocate for the new technology and Gil Amelio had moved from Bell Labs to Fairchild. At the Kitt Peak National Observatory (KPNO) in Tucson (Arizona), Richard Aikens and Roger Lynds had been working on low-light-level imaging systems for astronomy for many years. Soon the KPNO (now NOAO) began a program of development of CCDs. With Steve Marcus, this team began working on the Fairchild device. The Fairchild CCD201 and CCD202 image sensors were designed for TV applications and, although capable of high performance, they had a serious impediment for astronomical work due to the interline transfer construction (see below) which meant that they had columns of picture elements which were alternately light-sensitive and totally insensitive due to a cover by opaque metal strips; in terms of the image falling on the CCD these devices were half blind! Richard Aikens left the KPNO to set up his own company in 1978, called Photometrics, which played an important role in stimulating the manufacturing of CCDs and the development of scientific camera systems in general. Photometrics became part of Roper Scientific in 1998. Fairchild also changed names several times through successive owner companies including Schlumberger, Weston, Loral, Lockheed Martin and BAE Systems; it is now Fairchild Imaging.

There was a time of great frustration in the late seventies about the lack of access to CCD technology by the main-stream astronomical community. Development of the Wide-Field/ Planetary Camera, abbreviated WFPC but spoken as "wiff pick", was going well and many people were now aware of the sensitivity and the scientific potential of CCDs. Industry too was embracing the new technology, but commercially available products were scarce. During this interlude other forms of less suitable solid-state imagers were tried such as the Charge Injection Device (or CID) from General Electric (America), or the interline transfer device from Fairchild already mentioned. When 512 × 320 RCA CCDs appeared in the late-seventies it was a welcome relief.

The first RCA CCDs were frontside-illuminated which meant they had a poor response to blue light. Soon however, the thinned backside-illuminated CCDs appeared. Clearly, RCA had "the secret" for treating or passivating the thinned backside surface, and these CCDs displayed outstanding sensitivity over a huge spectral range — better even than the TI chips. Unfortunately there was one weakness. The design of the on-chip output amplifier was poor and so the CCD was 5-10 times "noisier" in electrical terms than the TI CCD. Later RCA CCDs were much better. Sadly, in 1985, RCA withdrew from the CCD market for commercial reasons. Detector development work has continued however, at the David Sarnoff Labs (Princeton, NJ).

In early 1980 a somewhat unexpected source of astronomical CCDs appeared. Craig Mackay of the Institute for Astronomy in Cambridge, England had been working on silicon vidicons. Progress was slow due to lack of funds. He had met Jim Westphal on Palomar Mountain in 1975 and was aware of the TI work on charge-coupled devices, but he learned of a British source of CCDs by a curious coincidence. Silicon vidicons had good spectral response, but they were "noisy". Craig needed a very low noise amplifier design. He contacted an eminent designer named Ken Kandiah at the British Atomic Energy Authority at Harwell and asked him to visit Cambridge. Kandiah offered a design based on a Junction Field Effect Transistor (JFET) but recommended Craig to David Burt at the GEC Hirst Research Centre in Chelmsford for a design based on the more readily available

Metal-Oxide-Semiconductor (MOS) transistors. When Craig met David Burt he learned that GEC had a very advanced CCD program. The following year, Craig and his then PhD student Jonathan Wright, put together a CCD drive system based on an existing vidicon camera. The noise associated with typical GEC CCDs was reported as 7 electrons in March 1982 while selected devices gave a mere 3 electrons. Remembering Jim Gunn's excitement on hearing that devices with better than 100 electrons noise were possible, then 3 electrons was a truly amazing result.

By June 1981, the date of the Harvard-Smithsonian conference on solid-state imagers, the number of independent astronomy groups working on CCD systems had already grown from 5 to 20. Devices in use came exclusively from TI, RCA and GEC (UK). Astronomers were clearly pushing the technology as hard as they could in a direction that was good for scientific imaging, yet with only three manufacturers one of which had low noise devices (GEC), one of which had high quantum efficiency devices (RCA) and the other which should have had devices with both properties, but was (a) having problems with blue sensitivity and (b) not available for sale anyway, it was understandable that people began to worry. When production of the GEC devices moved to English Electric Valve (now e2v) there was the inevitable hiccup in supply, and when RCA withdrew from the field, it seemed like the dream had become a nightmare. Eventually, TI CCDs "excess to NASA requirements" started to become available in the USA. Exceptionally detailed studies of the TI chips by Jim Janesick and colleagues at JPL advanced the understanding of CCDs and their optimization, and new devices by companies such as Thomson-CSF (Europe) had been studied in detail by a team at the Royal Greenwich Observatory (RGO including Paul Jorden who later joined e2v Technologies). In 1985 astronomers learned of a most exciting prospect. It was the formation of a team at Tektronix Inc. led by Morley Blouke to produce scientific grade CCDs with large formats and outstanding performance. The initial goal would be a 512 × 512 array with good-sized pixels (0.027 mm) leading to a chip with 2048 × 2048 pixels. Unfortunately, by mid-1986 it became clear that some sort of unexpected fabrication or processing problem was resulting in large numbers of defects called "pockets" thus rendering otherwise excellent low-noise devices unusable and hopes were again dashed. This time the situation for many astronomical groups was serious because new instrument designs and funding for instrument developments had been tied to the expected Tektronix chips.

Morley and his team did not give up and, in collaboration with several interested parties, they valiantly followed every lead in an effort to understand, model and eliminate such problems. The research at Tektronix and at JPL led to an in-depth understanding of the solid-state physics of CCDs which, as Morley remarks, "ought to be much easier to understand than a transistor". Around mid-1988 Tektronix began to ship CCDs to customers with long-standing orders. Later, the Tektronix CCD group was spun off into a company called Silicon Imaging Technologies, Inc. (SITe) which, for example, supplied all the CCDs for the highly successful Sloan Digital Sky Survey.

Among the many people frustrated by the dry spell in CCD supplies during that era was Richard Aikens, founder and president of Photometrics Ltd. In an unprecedented move he contracted with a so-called "silicon foundry" (a division of Ford Aerospace later taken over by Loral Corporation) to produce a custom CCD with 516 × 516 pixels (marketed as 512 × 512 pixels) and this turned out to be an outstanding success. The advantage of this approach is that the silicon foundry, can quote for device production without having to consider the "end-use" product. By the early nineties Dick Bredthauer and his team at Ford Aerospace (later Loral and then Lockheed Martin) had made a 4000x4000 CCD with 15 micron pixels; Dick is now president of Semiconductor Technology Associates (STA). In addition, Photometrics announced the availability of a chemical phosphor coating called Metachrome II which can be applied safely to any CCD by vacuum sublimation and thereby improve its response to blue light. In August of 1988 Lloyd Robinson reported excellent initial experimental results for another brand new device. As the result of a National Science Foundation grant to Lick Observatory, a contract was placed with E G & G Reticon Corporation to construct a large CCD suitable for spectroscopic applications; the format chosen was 400 × 1200. Finally, in the same year, new initiatives at EEV (later called Marconi and now called e2v Technologies) in the UK and at Thomson-CSF in France were announced. A thinning program and a mosaic construction program had begun at EEV. Meanwhile, funded by the European Southern Observatory and the French agency INSU, Thomson-CSF in Grenoble had developed a "buttable" version of their excellent low-noise front-illuminated device and a 2 × 2 mosaic had been constructed at the European Southern Observatory near Munich.

These approaches set a trend that has continued until the present day. Astronomers now work directly with a silicon foundry to obtain customized CCDs. Companies like e2v in England have become major suppliers to astronomical facilities, providing chips for most of the large mosaic cameras. Other manufacturers such as Kodak cater to a mass market and provide CCDs for companies that manufacture complete CCD camera systems. In addition, some US government-funded labs, such as Barry Burke's group at the MIT Lincoln Labs (Lexington, MA) where the orthogonal transfer devices have been developed or the group at the Lawrence Berkeley Lab (Berkeley, CA)) working on deep depletion CCDs, can provide special devices. Most astronomical developments concentrate on forming mosaics of high-yield formats like the 2048 × 4096 chips, and optimizing the response at both long and short wavelengths. The largest single scientific CCD manufactured so far is a 9216 x 9216 device from Fairchild Imaging (Milpitas, CA).

7.2 BASIC PRINCIPLES OF CCDS

7.2.1 Charge storage

A CCD is essentially an array or grid (Fig. 7.3) of numerous individual picture elements (pixels) each one of which can absorb photons of light and utilize the energy to release an electron within the semiconductor. If we are intent on making an imaging device, then we do not want the photon generated electrons to migrate away from the site of impact of the original photons. To confine the electron within a pixel requires a special electrostatic field to attract the charged electron to a specific spot. What happens to the next photon? Clearly we need to create a storage region capable of holding many charges. This can be done by applying metal electrodes to the semiconductor silicon together with a thin (100 nm = 0.1 μm) separation layer made from silicon dioxide, which is an electrical insulator. The resulting structure behaves like a parallel plate capacitor which can therefore store electrical charge. It is called an MOS (metal-oxide-semiconductor) structure. An electric field is generated inside the silicon slab by the voltage applied to the metal electrode.

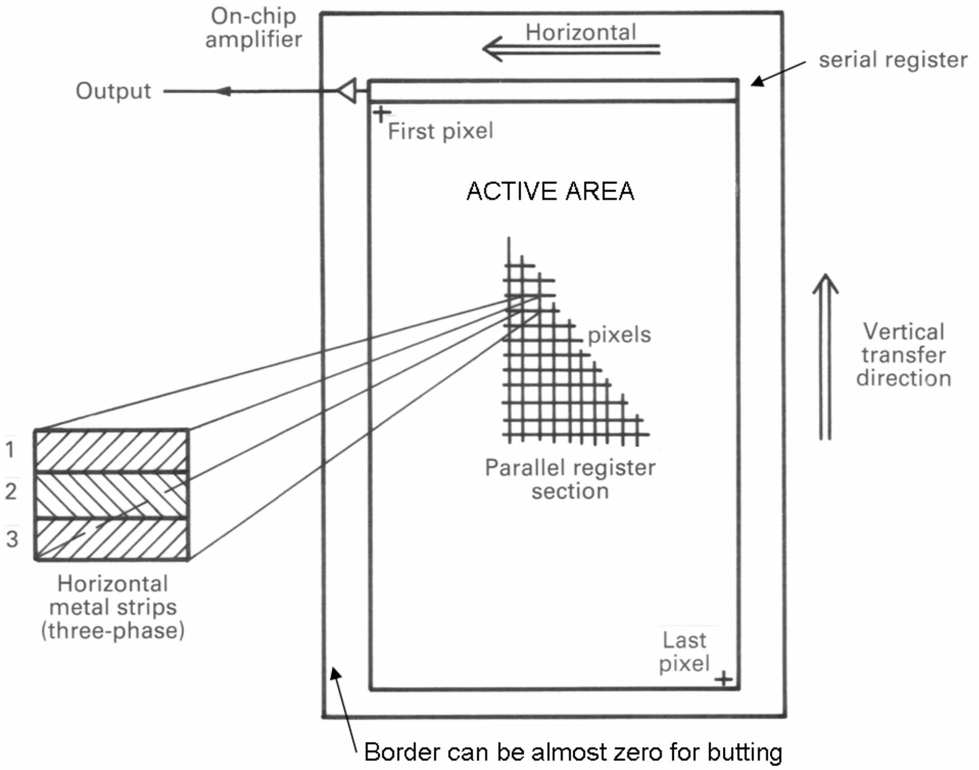


Fig. 7.3 The general layout of a CCD as a grid of pixels on a slab of silicon is shown.

If the material is p-type (the usual case) then a positive voltage on the metal gate will repel the holes which are in the majority and sweep out a region depleted of charge carriers. These conditions are illustrated in Fig. 7.4 When a photon is absorbed in this region it produces an electron-hole pair, but the hole is driven out of the depletion region and the electron is attracted towards the positively charged electrode. The MOS capacitor is the combination of two parallel plate capacitors namely, the oxide capacitor and the silicon depletion region capacitor, and therefore the capacitance is proportional to the area of the plates (electrodes) and inversely proportional to their separation. As the voltage on the plate can be controlled, then the depletion width can be increased or decreased, and so the capacity to store charge can also be controlled. The depletion region shown in Fig. 7.4 is an electrostatic "potential well" or "bucket" into which many photo-generated charges can be collected. Typically, the number of electrons stored is just $Q = CV/e$, where e is the charge on the electron (1.6×10^{-19} C), V is the effective voltage and the capacitance C is given by the "parallel-plate" formula $C = A\kappa\varepsilon_0/d$ in which A is the area of the pixel or gate electrode, d is the thickness of the region, κ is the dielectric constant of the SiO_2 insulator (~3.9) and ε_0 is the permittivity of free space ($8.85 \text{ x } 10^{-12}$ farad/m). As the voltage on the electrode increases, the "depth" of the well increases; other ways are needed to create side-walls to the well. Eventually, at a certain "threshold" voltage, even the minority charge carriers due to impurities, electrons for a p-type semiconductor, will be drawn to the electrode.

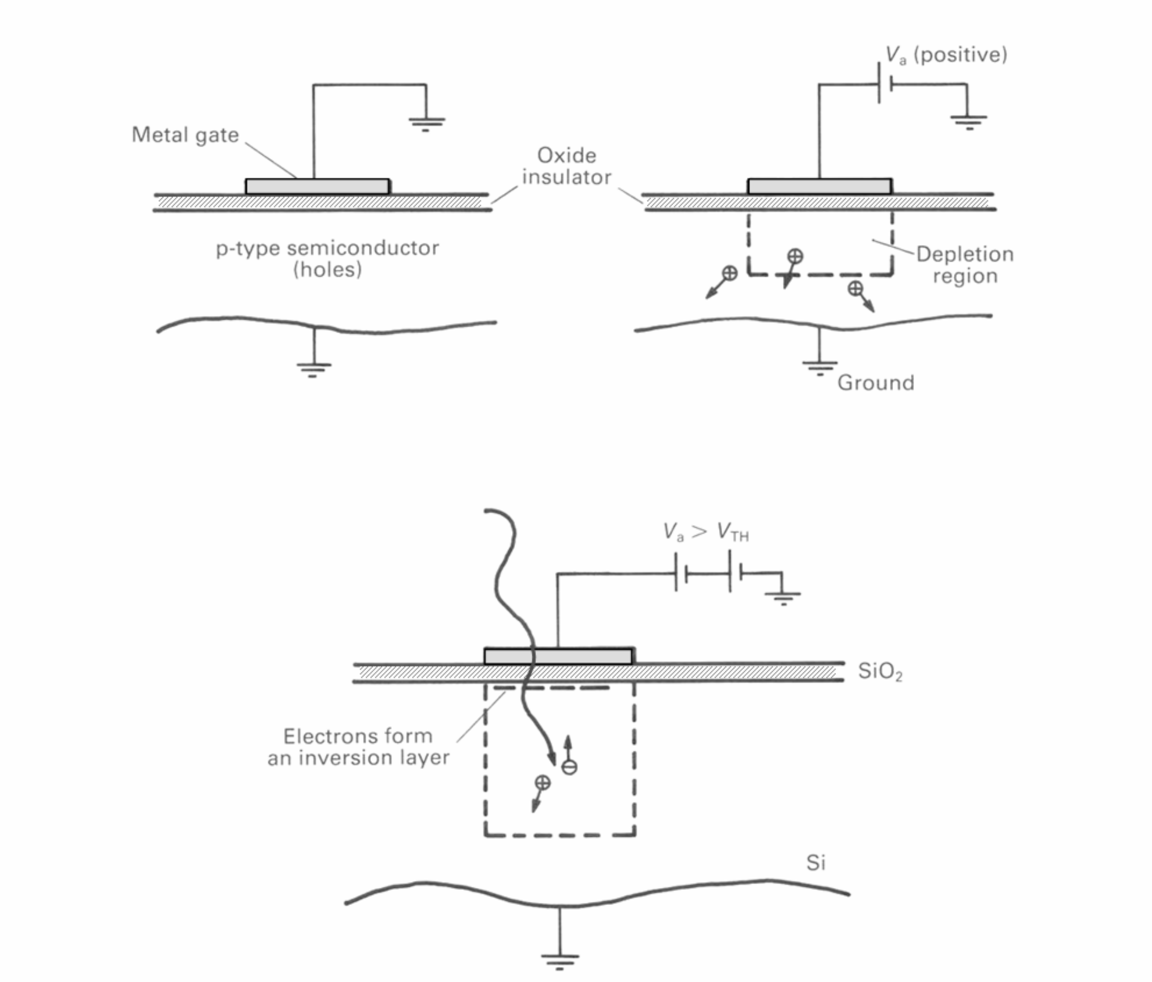


Fig. 7.4 The development of a single metal-oxide-semiconductor (MOS) storage well, the basic element in a CCD, is shown for different applied gate voltages.

Following Janesick (2001) we can provide a more quantitative description. A p-MOS capacitor consists typically of p-type (boron doped) silicon (Si), a thermally grown layer of silicon dioxide (SiO_2) about 100 nm thick to act as a dielectric insulator and a conductive (metallic) gate usually made of deposited polysilicon (silicon with randomly oriented crystal grains). In the p-type material holes are the majority carriers. If a negative voltage is applied to the gate while the silicon substrate is at ground potential, then a highly-conductive layer of holes will accumulate in a few nanoseconds at the Si-SiO_2 interface. This is called accumulation mode. If d is the thickness of the oxide insulator and $\varepsilon_{ox} = \kappa\varepsilon_0$ is the permittivity of SiO_2 (3.45×10^{-11} F/m), then the capacitance per unit area (F/m^2) of the oxide is just $C_{ox} = \varepsilon_{ox}/d$. Assuming $d = 100$ nm then $C_{ox} = 3.45 \times 10^{-4}$ F/m^2. Sometimes there are two insulating layers, one of SiO_2 and the other of silicon nitride ($\varepsilon_{nit} = 6.63 \times 10^{-11}$ F/m) and so these capacitors must add in series so that $C_T = C_{ox}C_{nit} /(C_{ox}+C_{nit})$. Now, if a positive voltage is applied to the gate, holes are driven away from the surface leaving behind negatively charged boron ions and thus creating a depletion region devoid of mobile charge carriers. The number of holes driven away in depletion mode equals the number of positive charges on the gate electrode, thus, $Q_i = e\ N_A x_d$ where Q_i is the ionized acceptor charge concentration (C/m^2) beneath the depleted gate and x_d is the depth (m) of the depletion region. N_A is the concentration of boron (acceptors) in atoms/m^3 and e is the numerical value of the charge on the electron. As the depletion region is non-conductive it acts like an insulator with a capacitance per unit area of $C_{dep} = \varepsilon_{Si} / x_d$ where $\varepsilon_{Si} = 1.04 \times 10^{-10}$ F/m; the dielectric constant of silicon is ~11.7. Thus the net gate capacitance in depletion mode is the series combination of C_{ox} and C_{dep} and because C_{dep} is the smaller capacitance then it dominates the series combination.

Gate voltage is constant throughout its thickness because it is a conductor. There is a voltage drop across the oxide later and then the voltage in the depleted p-type silicon will depend on the charge distribution, but it must eventually drop to the ground potential of the substrate, i.e. $V = E = 0$ at $x = x_d$. The variation of voltage (V) with depth (x) is determined by Poisson's equation:

$$\frac{d^2V}{dx^2} = -\frac{\rho}{\varepsilon_{Si}} \tag{7.1}$$

where ρ is the charge density. The origin (x =0) is taken as the Si-SiO$_2$ interface. In principle ρ is given by $e[p + n + N_A + N_D]$ where p is the number density of free holes, n the number density of free electrons, N_A the number density of localized fixed ionized acceptors (atoms/m^3) and N_D the number density of fixed ionized donors (atoms/m^3). Most free carriers in the depletion region are swept away by the electric field and thus for a p-channel device we expect $\rho = -eN_A$, where the sign is negative due to the absence of holes. Therefore $d^2V/dx^2 = eN_A/\varepsilon_{Si}$. Integrating Eq. 7.1 with this expression for ρ and applying the condition that $dV/dx = 0$ at $x = x_d$ gives

$$\frac{dV}{dx} \equiv -E_x = \frac{eN_A}{\varepsilon_{Si}}\left(x - x_d\right) \tag{7.2}$$

Integrating again using the fact that V = 0 when $x = x_d$ gives

$$V = \frac{eN_A}{2\varepsilon_{Si}}\left(x - x_d\right)^2 \tag{7.3}$$

Equation 7.3 implies that the most positive voltage relative to the substrate occurs at the Si-SiO$_2$ interface where $x = 0$ and that the surface voltage V_S is

$$V_S = \frac{eN_A}{2\varepsilon_{Si}}\left(x_d\right)^2 \tag{7.4}$$

and the electric field at the surface is

$$E_S = \frac{eN_A}{\varepsilon_{Si}} x_d \tag{7.5}$$

8

Practical operation of CCDs

Modern CCDs are fairly predictable in their operation and characteristics, but there are still many subtleties to successful operation, especially for devices customized for astronomy. Issues include maximizing the ratio of signal to noise, obtaining stability and repeatability in performance, and finding suitable methods of control. These topics are treated by recounting some of the developments which uncovered problems and led to today’s solutions. There is no intention here to provide a "constructors manual", only to alert the potential user to a host of practical issues. Many of the same practical issues apply for other detectors and other wavelength regimes.

8.1 CLOCK VOLTAGES AND BASIC ELECTRICAL FACTORS

CCD manufacturers provide a data sheet which gives the electrical pin connection diagram for the device (Fig. 8.1), the names and symbols for each pin, the voltages or range of voltages to be applied to each pin and the timing diagram, that is, a diagram showing the time sequence of the CCD drive signals and the relationship between them. Terminology varies, but certain basic functions are common to all. Voltages applied to CCDs are of two types. Fixed voltages, referred to as “dc bias” levels, which remain unchanged after switch-on, and pulsed or “clock” voltages which can be switched back and forth between two voltage levels known as the high and low levels. As described in Chapter 7, clock voltages are applied in a precise order and time sequence to charge-couple the electrons from one storage-well to the next. Although the type of mounting package used (its size, number of pins and their names) differs from device to device, certain functions are required in all CCDs, and in particular in the frame transfer image sensors used widely in astronomy, namely:

1. **Serial (horizontal) register clocks**. One pin for each phase or electrode used to define the pixels in the horizontal register is required, i.e. three for a 3-phase CCD. In many larger CCDs two completely separate serial output registers are provided on opposite sides of the active area. Sometimes each serial register has outputs at both ends giving four outputs in total.

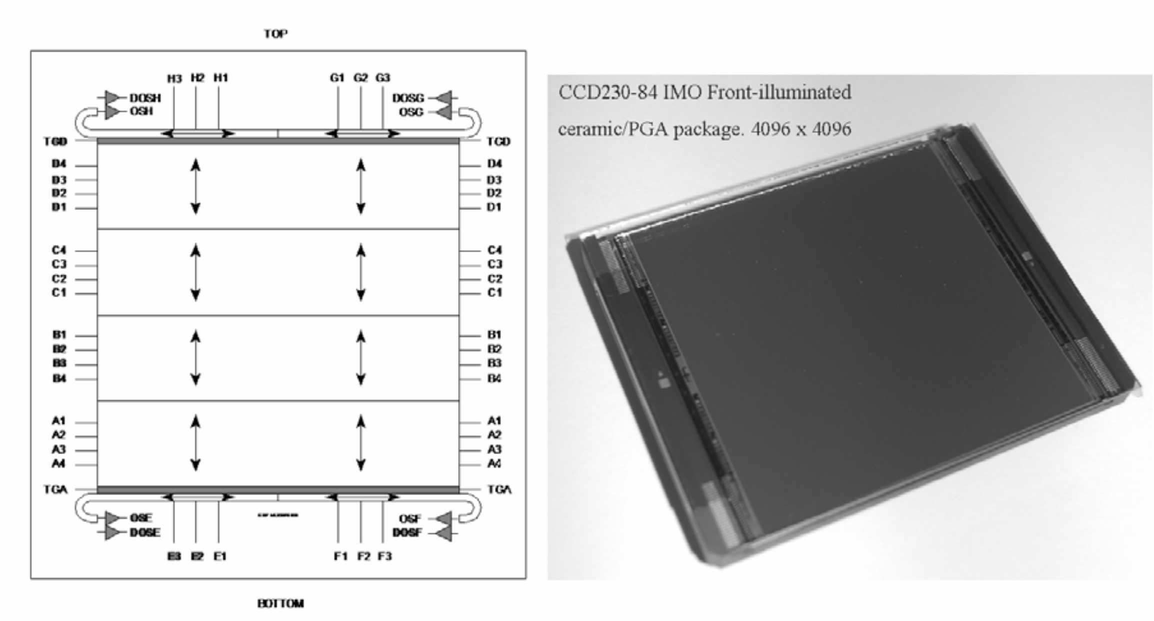


Fig. 8.1 The device schematic for a 4K x 4K frame transfer CCD to illustrate typical pin assignments and terminology. Each manufacturer will provide information of this kind on a data sheet. Credit: Paul Jorden and e2v technologies.

2. **Parallel (vertical) register clocks**. Again one voltage line is needed for each phase. In frame transfer CCDs the two-dimensional vertical or parallel register is split into two identical sections — the "image" section and the "store" section — which can be controlled separately and therefore two sets of pins are needed.

3. **Reset transistor clock**. A single, periodically recurring voltage pulse is required to reset the CCD output amplifier or more accurately, the output charge collecting capacitor during the readout process. Each of these clock voltages will have a specified high and low level and, since the voltage difference or "swing" — usually in the range 5 to 10 volts — can have important consequences on performance, it is often arranged to select the levels on demand from the CCD controller. The total range of voltage required to operate a CCD is generally less than 20 volts, except for Electron Multiplied CCDs that require one clock at about 40 V.

The most important dc bias voltages are as follows:

1. **Substrate voltage** (V_{sub}). This is the reference for all other voltages. This voltage is usually, but not always, kept at ground or zero volts.

2. **Reset drain voltage** (V_{RD}). This voltage is applied to the "drain" terminal of the on-chip "reset" field-effect-transistor (FET) at the CCD output to establish the level to which the output node (capacitor) must return after each charge packet is read out.

3. **Output drain voltage** (V_{OD}). This voltage applied to the drain terminal of the on-chip output amplifier determines the operating point of that transistor.

4. **Output gate voltage** (V_{OG}). The output gate is essentially an extra (last) electrode in the serial output register.

Nomenclature varies slightly from one manufacturer to another, but details are always given on the data sheet.

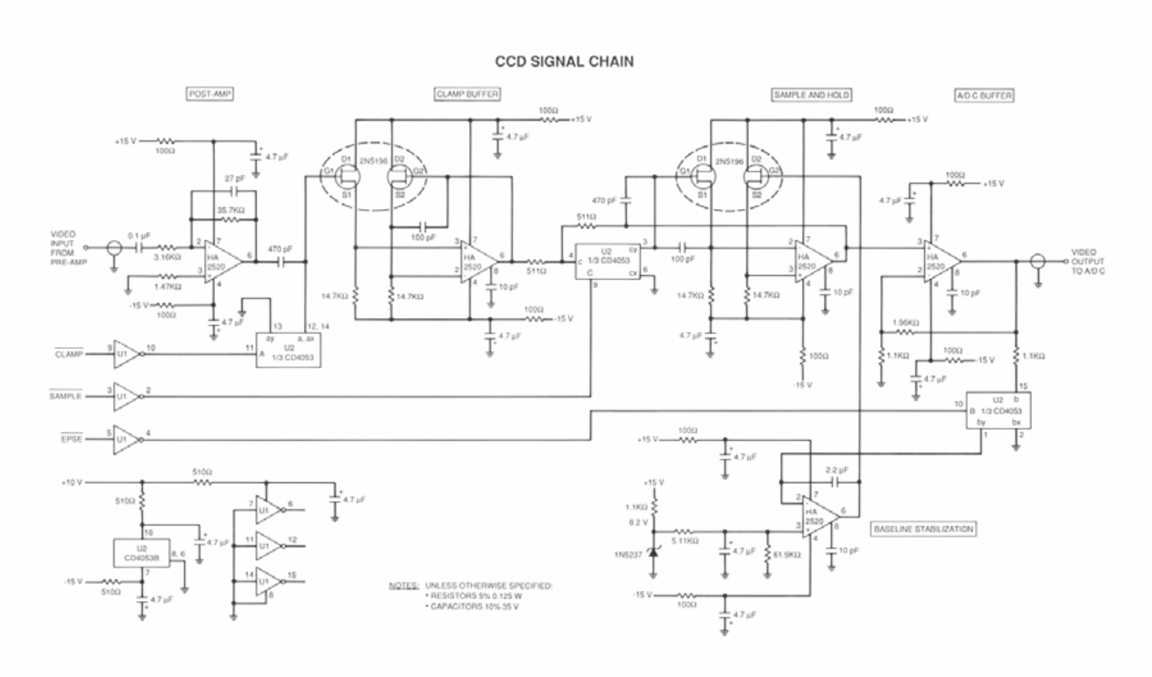
Pulsed voltage signals, corresponding to individual charge packets, emerge from the output transistor (OS)—also called the video output—which is connected to ground via another transistor external to the CCD to provide a constant current load. The output transistor source current (I_{OS}) is important for most CCDs. Some CCDs have additional options. For example, the vertical register can be clocked up or down to independent serial registers. One serial register may terminate with an on-chip amplifier designed for low-noise slow-scan operation whereas the other register might have an output amplifier optimized for TV video rates. Some chips have a separately clocked gate known as a "summing well" which has the storage capacity of two serial pixels. Almost all CCDs have electrical input connections and test points; the manufacturer will specify whether these pins should be fixed at a high or low voltage. Remember also that voltages are relative to the substrate voltage which may not always be zero volts. Normally, a CCD manufacturer will provide an initial set of operating voltages for a particular chip, but will not necessarily optimize these voltages for cooled slow-scan astronomical work. Builders of commercial astronomical CCD camera systems can provide this service. For some CCDs, small changes of one-tenth of a volt to clock swings or dc bias values can often yield substantial improvements in low light level behavior.

Safe Handling: CCDs, including thinned devices, can withstand substantial illumination overloads without permanent damage. Mechanically, CCDs are also quite robust unless heavily thinned. However, they are integrated circuits of the CMOS type and their tiny gate connections can be short-circuited by static electricity discharges. Precautions against static must be taken when handling CCDs. For example, CCDs should be stored in electrically conducting containers, earthing-straps should be tied to the wrist during handling operations, no nylon clothing worn, and the work performed in a clean, ionized (electrically conducting) airflow. Finally, some kind of protection on the drive signals is essential if power supplies are used which have a rating even slightly above the maximum recommended voltages for the CCD. Usually this is achieved by the use of Zener diodes on the drive outputs. With these precautions, a CCD camera can last a long time.

8.1.1 The Analog Signal Chain

The analog signal chain includes the preamplifier, post-amplifier, noise removal circuits and analog-to-digital converter (ADC). In addition, low noise dc power supplies are required to provide the bias voltages and a "level shifter" circuit is usually needed to convert simple 0-5 V TTL clock pulses to the levels required for the chip. Figure 8.2 shows a typical signal chain after the preamplifier.

The heart of the electronic system is a signal processing unit designed to sample and filter noise, but proper grounding, good power supply de-coupling and optimum control timing are also very important factors in achieving a low-noise system. Undoubtedly, "ground-loops" are the most common cause of noisy CCD systems. A ground-loop forms when two interconnecting parts of an electronic system are separately connected to ground via small but different impedance paths. As a result, a voltage difference can exist between "grounds", and currents can flow. Ground-loops between the telescope and camera body and driver electronics, within the driver electronics, or between the driver electronics and the computer system can cause interference with the readout electronics.



NOTES: UNLESS OTHERWISE SPECIFIED:
• RESISTORS 5% 0.125 W
• CAPACITORS 10% 35 V

Fig. 8.2 Part of the analog signal chain is shown. This is a practical correlated double sampling circuit used at JPL. Credit: Jim Janesick.

The observable effect is a clearly visible pattern of diagonal stripes in the image; these patterns are synchronized with the mains frequency. The solution to this problem is to have only a single ground point in the system, to which all the zero reference points and shields are connected; because many wires may radiate from such a single ground point it is often called a "star ground" (see Fig. 8.3). All connections to the star ground should be made as short as possible (< 1 m for frequencies up to 10 MHz), with the lowest-resistance electrical wires available. Wherever possible, circuit boards should use copper ground planes. Many designers carefully isolate the entire instrument from the telescope structure, even if the telescope is known to have an excellent earth ground, which means that extreme care is required when attaching any other piece of electrically-powered apparatus to the instrument in case a ground-loop is formed.

Similarly, electrical noise from motors, light dimmers, or computer parts can be "picked up" by capacitive coupling, inductive coupling or radiative coupling if inadequately shielded wires or components are used in the CCD system. Remember, the CCD system is capable of detecting signals of only a few millionths of 1 volt! Co-axial cables which have a surrounding braided copper shield are quite effective (90%), and "twisted-pair" wire is often used in preference to a single wire for carrying critical signals to the CCD. Signals transmitted over long distances, such as from telescope to control room, usually use optical fibers if possible. Most astronomical CCD systems have a low-noise preamplifier inside or very close to the cryostat to boost the CCD signal and most systems convert from the weak analog output to healthy digital signals before transmitting the data over long cables. The physical environment at cold, high-altitude mountain-top observatories can cause "drifts" in the operating points of many electronic components; this effect can be overcome by careful component selection and packaging.

8.5 NOISE SOURCES

A CCD is by its very nature a digital imaging device. The clocking procedure described earlier delivers a stream of charge "packets" from pixels in the image area all the way to the output amplifier; the charge (*Q*) in each packet is proportional to the amount of light in that part of the original image scene. As each charge packet arrives at the output field- effect transistor it causes a change in voltage to occur (of amount $V = Q/C$, where C is the capacitance at the output node); the smaller the node capacitance, the larger the voltage change for the same size of charge packet. For the earliest CCDs the output capacitance was fairly high, e.g. $C \approx 0.6$ picofarads (pf) which yields about 0.25 microvolts (µV) per electron (in the charge packet), whereas for modern CCDs the node capacitance is < 0.1 pf which gives a healthy > 1.6 µV per electron. Much larger values are possible; some Kodak CCDs using an extremely small output MOSFET give 15 µV/electron, yet the overall noise under slow-scan conditions is greater than 10 electrons because other noise sources become larger as the MOSFET gets smaller.

It is desirable that the noise performance of a CCD camera system be limited only by the output transistor of the CCD and not by any other part of the electronic system. To achieve this goal, one must understand the noise sources associated with the CCD and take steps to get them to an irreducible minimum; this minimum is the ultimate “readout noise” (R), usually quoted as the root mean square noise in electrons. There are several potential sources of unwanted electronic noise. These include,

- background charge associated with fat zero offsets
- transfer loss fluctuations
- reset or "kTC" noise
- MOSFET noise
- fast interface state noise

When a preflash is used to introduce a fat zero charge to aid transfer efficiency or eliminate charge skimming, the consequence is a noise equal to the square-root of the total number of charges in a pixel.

During charge transfer a fraction of the charges are left behind. However, this fraction is not constant but may fluctuate and so an additional noise component is added to the signal noise. This "transfer noise" is given by

$$\sigma_{tr} = \sqrt{2\varepsilon n\, N_0} \qquad (8.4)$$

where ε = 1 - CTE is the fraction of charges not transferred, n is the number of transfer and N_0 is the original charge. The factor of two occurs because the Poisson-distributed noise happens twice, once for trapping and once for release. This effect can be of order 70 electrons for surface channel CCDs but is typically ten times smaller or better for buried channel CCDs and astronomical light levels. For very large CCDs this effect implies that exceptionally good charge transfer efficiency must be achieved.

Noise associated with the re-charging of the output node is given by $\sqrt{(kTC)}/e$ or about $284\sqrt{(C)}$ at a CCD temperature of 150 K, where the capacitance, C, is in picofarads. This effect is called “reset noise” and it is the dominant source in most cases.

Other noise sources associated with the output MOSFET, such as one-over-f noise (i.e. noise that varies roughly as 1/frequency), can generally be made quite small by good manufacture, typically a few electrons. Traps that absorb and release charges on very short time scales, thereby causing a fluctuation in the charge in any pixel, are called "fast interface states". In this case the noise is given by

$$\sigma_{SS} = \sqrt{2kTn\, N_{SS}\, A} \tag{8.5}$$

where k is Boltzmann's constant, T is the absolute temperature, n is the total number of transfers (not the number of pixels), N_{SS} is the surface density of traps and A is the surface area. This effect is very serious for surface channel CCDs but is normally quite small (of order 5 electrons or less) for good buried channel devices. It is a remarkable tribute to the foresight of Boyle and Smith, and the skills of every scientist and engineer who has worked hard on CCD technology, to realize that devices are now routinely made with a readout noise of less than 5 electrons (see Fig. 8.11 a,b). The lowest noises that have been consistently measured (without electron multiplication) are in the range 2 - 3 electrons.

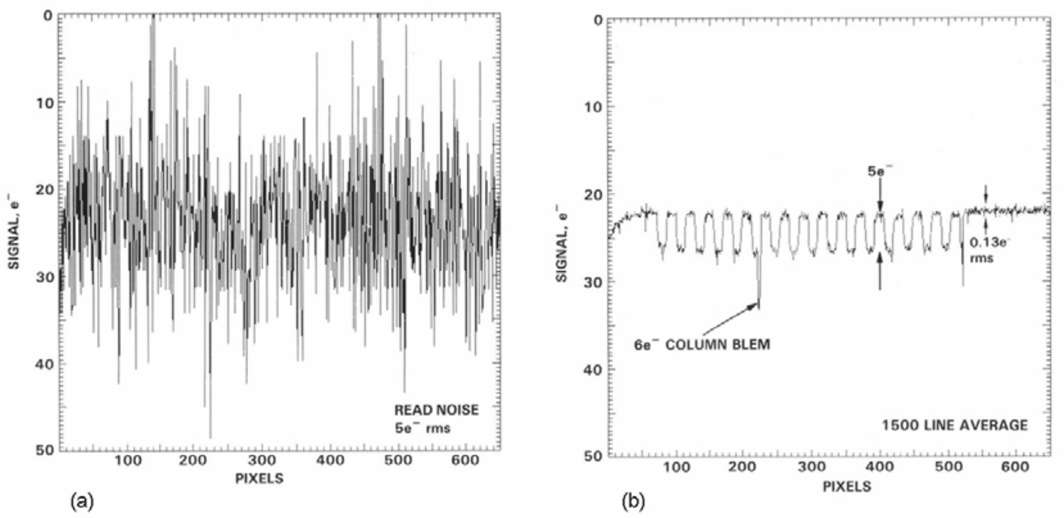


Fig. 8.11 (a) A single raw trace of the signal from a Loral CCD at the 5 electron level. Embedded in the noisy trace is an unseen $5e^-$ peak-to-peak square-wave pattern. (b) After 1500 lines have been averaged the random noise is only $0.13e^-$ as seen in the overscan region. A $6e^-$ column blemish has also emerged. Credit: Jim Janesick.

8.6 SIGNAL PROCESSING AND DIGITIZATION

Here, we consider CCD signal processing in more detail, and in particular explain the crucially important technique known as “correlated double sampling” or CDS. As each charge packet arrives at the output node it produces a voltage change which must first be amplified and then digitized by an analog-to-digital (A/D) converter. This process is not instantaneous; it requires a finite amount of time and hence the term “slow-scan”**. A** high degree of accuracy is required, such as can be achieved with a 16 bit A/D; a 16 bit A/D divides a specified voltage range, typically 10 volts, into 65,536 (2^{16}) parts and therefore each voltage interval is 152.5 μV in size. The A/D circuit matches up the actual voltage to the nearest number on the scale of 0 - 65,535.

To measure the voltage of each charge packet we need a "reference" voltage. We could use ground, but it is important to reset the output capacitance back to some nominal value on each readout cycle otherwise we would be forming the difference between one charge packet and the previous one while drifting away (in voltage) from the ideal operating point of the MOSFET.

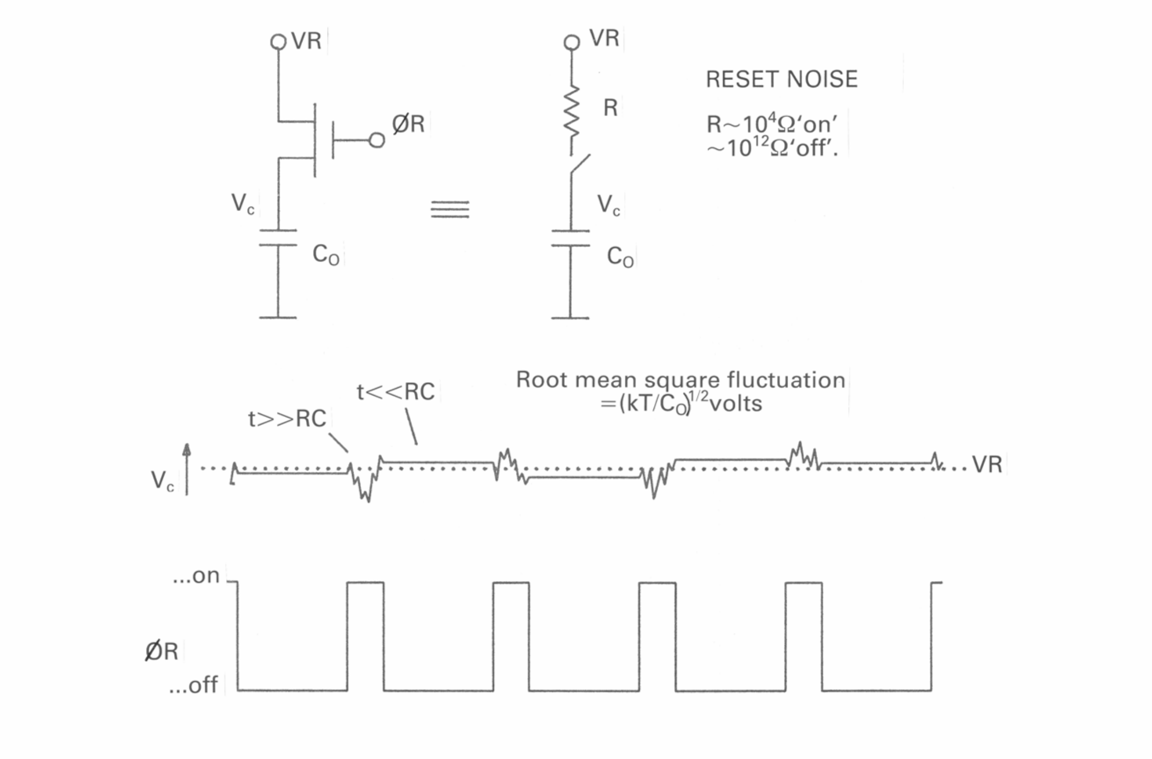


Fig. 8.12 An equivalent "switch" circuit to explain the operation of the reset transistor. The resistance, and therefore the RC-time constant, is very different between off and on states.

There is another way. The output capacitor can be recharged to a fixed voltage by briefly pulsing the gate of another transistor, called the reset transistor (see Fig. 8.12), to briefly turn that transistor on (like closing a switch) so that current can flow from a power supply to charge-up the node to the supply level. When the reset pulse disappears the reset transistor is turned off (like opening a switch) and the output becomes isolated to await the next charge-dump from the horizontal register. As a capacitor is charged to a certain voltage level (V_{RESET} or V_{RD}) it does so exponentially, rising steeply at first and then leveling off to approach its final value as shown in Fig. 8.13. Again, due to random thermal agitation of electrons in the material, there is "noise" or uncertainty on the mean value and so the final voltage can lie anywhere within a small range, given by

$$Reset\ Noise = \sqrt{\frac{kT}{C}}\ volts\ \ or\ \ \frac{\sqrt{kTC}}{e}\ electrons \tag{8.6}$$

In this expression, k is Boltzmann's constant (1.38 x 10^{-23} Joules/K), e is the charge on the electron, T is the absolute temperature of the output node in degrees Kelvin (K) and C is the node capacitance. If C is expressed in picofarads (i.e. 10^{-12} Coulombs/volt) then this noise uncertainty (called the "reset noise" or, from the formula, kTC (kay-tee-cee) noise) is simply 400√(C) electrons at room temperature and about 250√(C) at 120 K (-153 °C); for a typical modern device this would yield < 80 electrons noise which greatly exceeds the readout noise of the MOSFET alone, and so some means must be found to remove it.

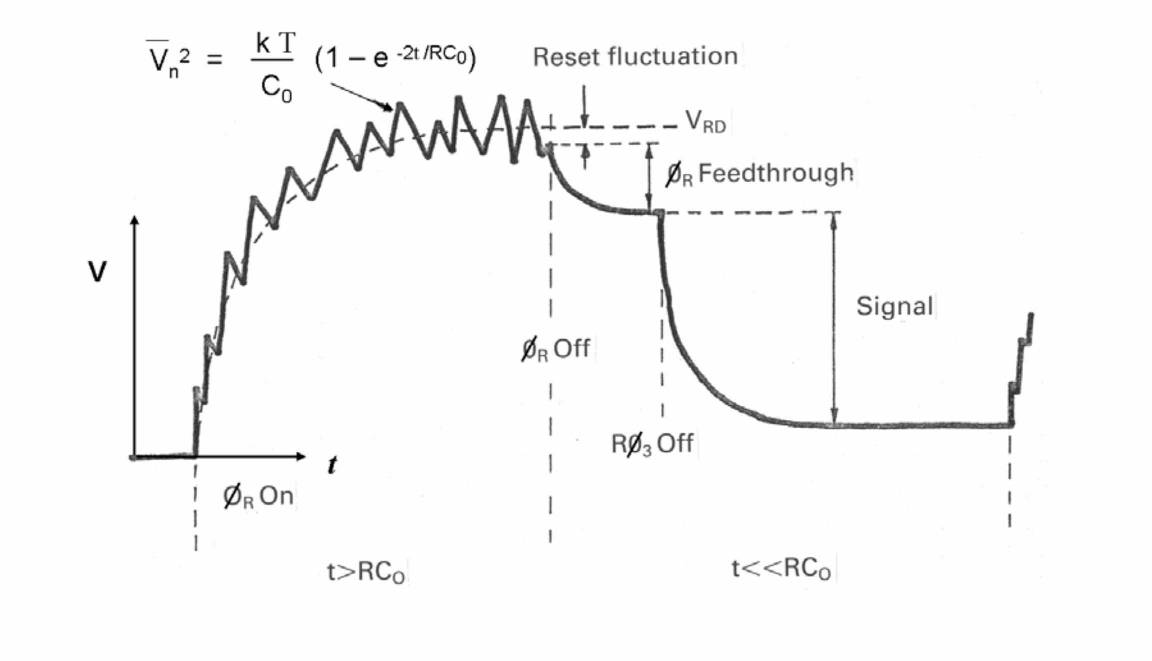


Fig. 8.13 The charging profile of the output of a CCD when reset. The curve is "noisy" but when the reset pulse disappears the last value of the signal becomes frozen.

8.6.1 Correlated double sampling

Fortunately, removal of reset noise is quite straightforward due to the fact that whatever the final reset voltage actually is, and it must be in the range (V_{RESET} - √(kT/C)) to (V_{RESET} + √(kT/C)), it will get "frozen" at that value because the leakage of current through the switched-off reset transistor is exceedingly slow (its "RC time constant" is seconds compared to the microseconds between arrivals of discrete CCD charge packets) and hence, if this uncertain reset level is sampled by the A/D just prior to a charge packet being dumped at the output, and then again after the charge has been added, it will have *exactly* the same value in each sample. Forming the difference of these two signals therefore automatically eliminates this voltage level without ever knowing exactly what it was. This technique is known as correlated double sampling (or CDS).

9

Characterization and calibration of array instruments

All electronic imaging devices require calibration in order to be used for quantitative work in photometry and spectroscopy. It is important to understand how the properties of the detector can be measured and how the behavior of the detector affects photometric and spectroscopic analyses. This chapter describes important steps in these calibrations in terms of CCDs, but the same considerations apply to other array detectors. Signal-to-noise expressions for array instruments are also developed.

9.1 FROM PHOTONS TO MICROVOLTS

The observed quantity in an experiment is the stream of photons, but the detected quantity is a small voltage (V_o) which is amplified and digitized. If N_p photons are absorbed in the integration time (t), then $\eta G N_P$ electrons will be detected. Here η (< 1) is the quantum efficiency and G ~ 1 is called the photoconductive gain and allows for intrinsic amplification within the detector in some cases (G=1 for a CCD). Multiplying by the charge on the electron (e) gives the total number of coulombs of charge detected, and the resulting voltage at the output pin of the array detector is

$$V_o = \frac{A_{SF}\,\eta G\, N_p\, e}{C} \tag{9.1}$$

In this expression, C is the capacitance of the output node of the detector (CCD or infrared array) and A_{SF} is the amplification or "gain" of the output amplifier which is usually a source follower (typically A_{SF} ~ 0.8); the suffix SF stands for "source follower". In practical terms, we need only know the combined quantity $A_{SF}\eta G/C$, but it is desirable to know these quantities individually too. Therefore, the first step is usually to determine the quantum efficiency or QE.

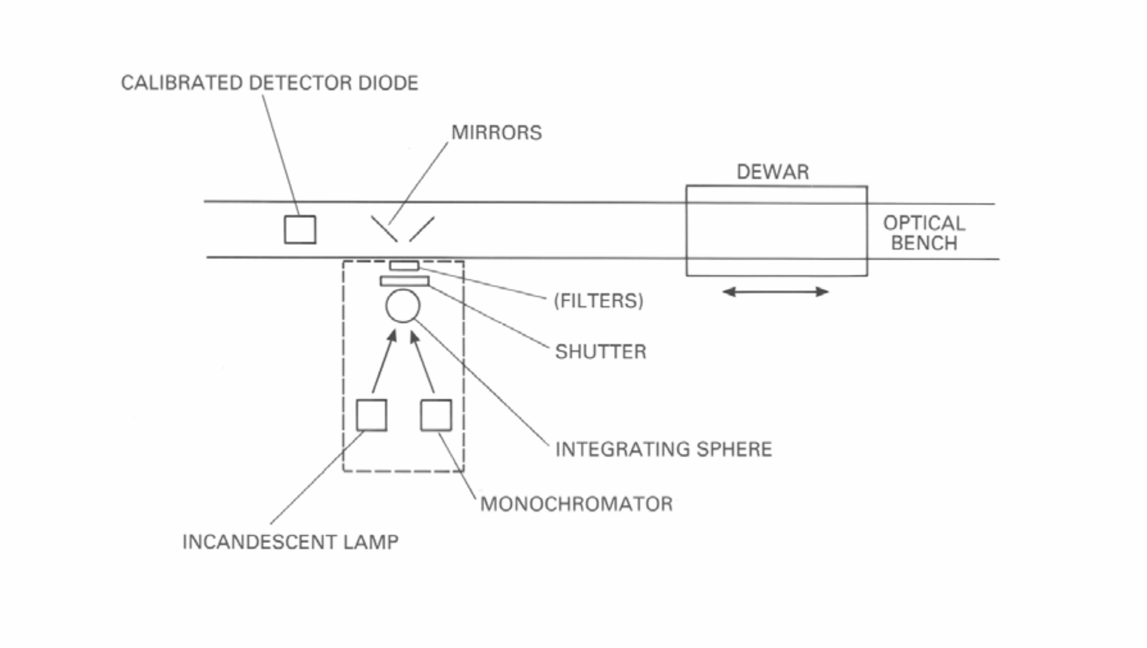


Fig. 9.1 A possible laboratory arrangement for calibration and characterization of CCDs.

9.1.1 Quantum efficiency and DQE

In principle, quantum efficiency can be determined in the laboratory with a stable and well-designed calibration system constructed to properly illuminate the detector through a known spectral passband with the minimum of other optics in the beam. One example of an experimental set-up is shown in Fig. 9.1 and a practical realization is illustrated in Fig. 9.2 which shows the UCO/Lick Observatory automated QE measurement system. Either an incandescent lamp or a grating spectrometer can be used as a source of illumination. After passing through a device called an "integrating sphere" which randomizes the light rays and produces a uniformly illuminated source, the light passes through a shutter and a filter holder. At longer wavelengths a stable blackbody source (commercially available) can be used and the integrating sphere is not needed. The light is then split by mirrors; part is directed towards the detector cryostat and part toward a calibrated photodiode. Exposures are taken at the desired wavelengths and recorded along with the signal from the calibrated photodiode. Mounting the camera unit on an optical bench allows it to be moved closer or farther from the light source in a controlled manner. This enables the experimenter to use the inverse square law for light as a way of changing the illumination on the detector. For some wavelengths and passbands it is also possible to use non-wavelength-dependent attenuating filters called "neutral density" filters since their attenuation can be determined fairly accurately with the calibrated photodiode. With this set-up it is easy to obtain the relative QE as a function of wavelength, but to convert this to absolute quantum efficiency at any given wavelength requires a precise calibration of the illumination. The exact transmission or "profile" of the filter passband at each wavelength and an accurate determination of the solid angle on the source subtended by a pixel are also required. It is usually easier to obtain the solid angle with a well-defined geometry controlled by baffles rather than optics because adding refractive or reflective elements to the setup just introduces other unknown quantities into the experiment. Filter profiles are measured in commercial spectrophotometers.

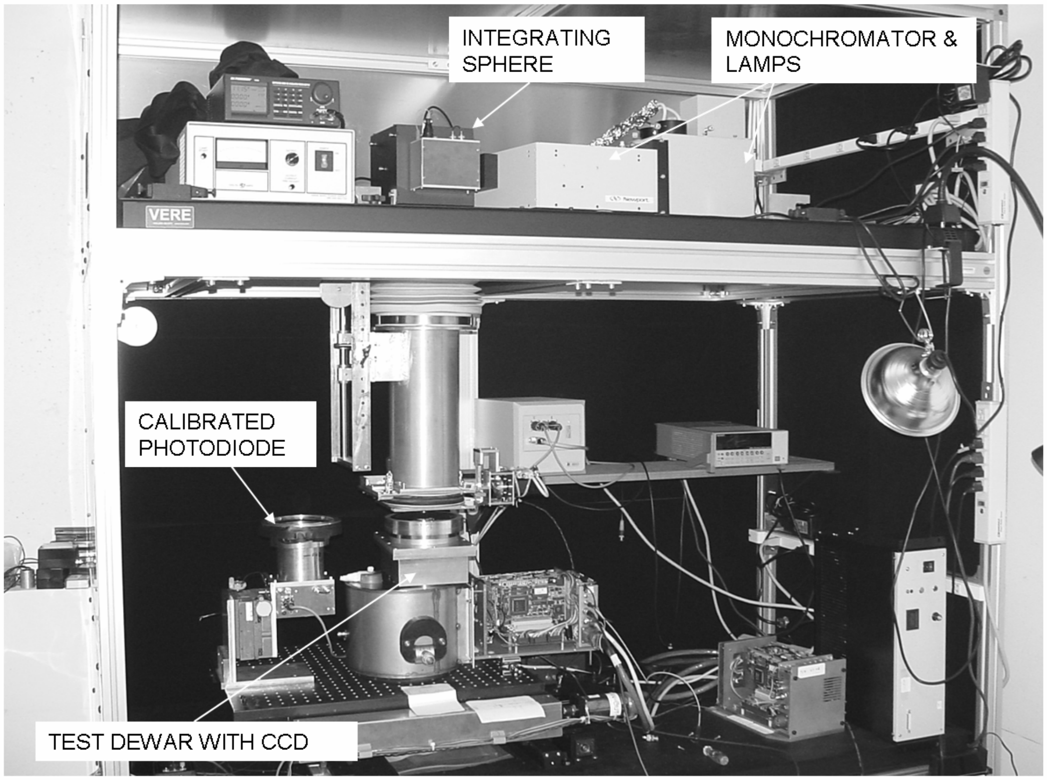


Fig. 9.2 The UCO/Lick Observatory automated QE measurement facility. Credit: Richard Stover and UCO.

Note that in the case of an infrared filter the scan needs to be done at the operating temperature (e.g. 77 K) because the passband broadens and shifts to shorter wavelengths as temperature decreases. At infrared wavelengths it is easier to be sure of the illumination level by using a blackbody source at a known temperature because the energy spectrum is given by the Planck function $B_\lambda(T)$ which is determined only by the absolute temperature (T). Good laboratory setups can yield both the relative quantum efficiency as a function of wavelength and the absolute QE. The quantum efficiency of a deep-depletion CCD measured with the UCO/Lick Observatory equipment by Richard Stover is shown in Fig. 9.3. The reflectance (R) is also measured and there is good agreement between 1-R and QE (η) except at the short and long ends of the wavelength range. At the shortest wavelengths the absorbed photons create electron-hole pairs too far from the depletion region in these thick devices, and for the longest wavelengths the absorption lengths are too long and no electron-hole pairs are created.

Electrical measurements can be used to determine A_{SF} independently. For A_{SF}, the simplest approach is to change the output drain voltage and observe the change in the output source voltage; the ratio will yield A_{SF}. To measure C a controlled charge Q can be injected and the voltage V measured, then $C = Q/V$. Alternatively, one can expose the detector to a substantial light level to yield a large output signal in which the dominant noise is shot noise. If N is the total number of charges collected then the measured voltage is $V = eN/C$, and the noise is $\sigma_V = e\sqrt{N}/C$. By squaring the noise and forming the ratio we get

$$C = \frac{eV}{\sigma_V^2} \tag{9.2}$$

thus allowing C to be determined from the *mean* signal V and the *variance* σ_V^2 of the measured voltage noise on the signal.

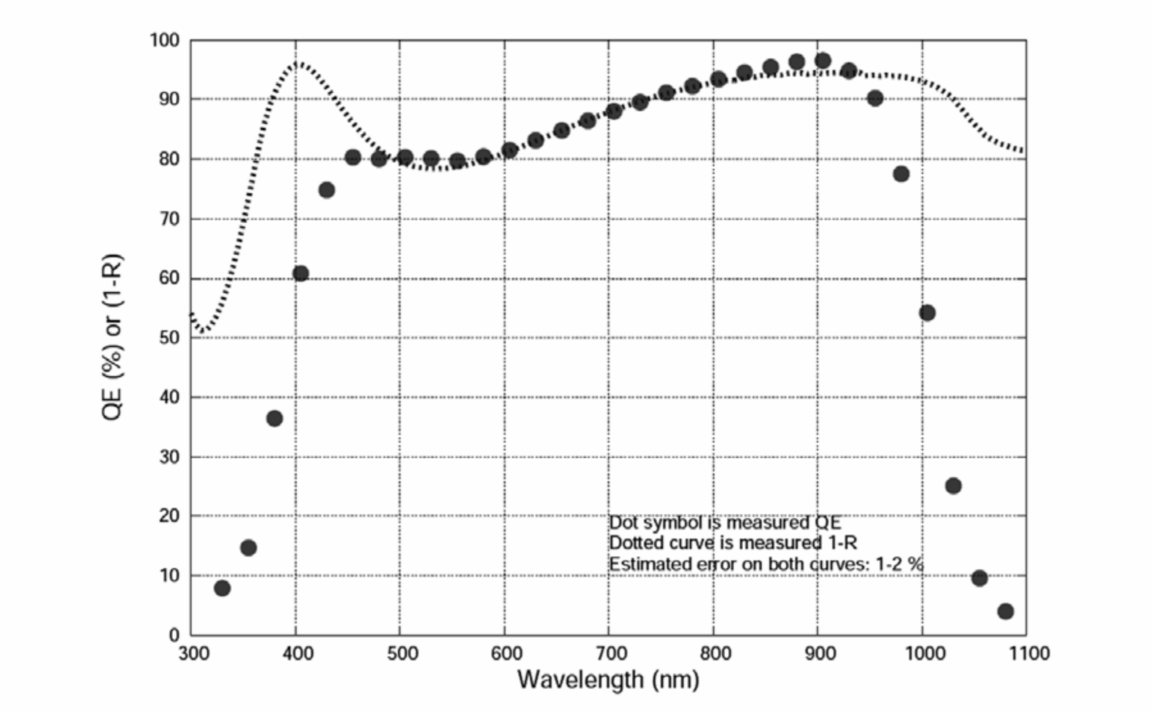


Fig. 9.3 Curves of the measured QE and reflectance of a deep-depletion CCD using the UCO/Lick automated system. Credit: Richard Stover.

By observing the signal from a star of known brightness and energy distribution, one can take advantage of the fact that the solid angle (on the sky) of a telescope is very well-defined. Unfortunately, the product (τ) of all the unknown optical transmissions is now included and so the derived quantity (assuming $G = 1$ and A_{SF} and C known from electrical measurements), is $\tau\eta$. While this is all that is needed for calibration, it is still very helpful to know where light is being lost so that improvements can be made.

When discussing systems which exhibit readout noise, as opposed to systems with pure photon counting detectors (PCDs), it is useful to introduce the concept of Detective Quantum Efficiency or DQE. The DQE is defined as the quantum efficiency of an idealized imaging system with no readout noise but which produces the same signal-to-noise ratio as the actual CCD system in question.

$$\overline{X} = \frac{(X_0 + 4\,X_{1/2} + X_1)}{6} + O(e) \qquad (9.22)$$

where $X_{\frac{1}{2}}$ is the airmass at the mid-point of the integration and O(e) is a small error of about 1 part in 10,000. Color correction, is still tricky because even two stars at the same airmass will undergo different extinctions if they are not identical in color.

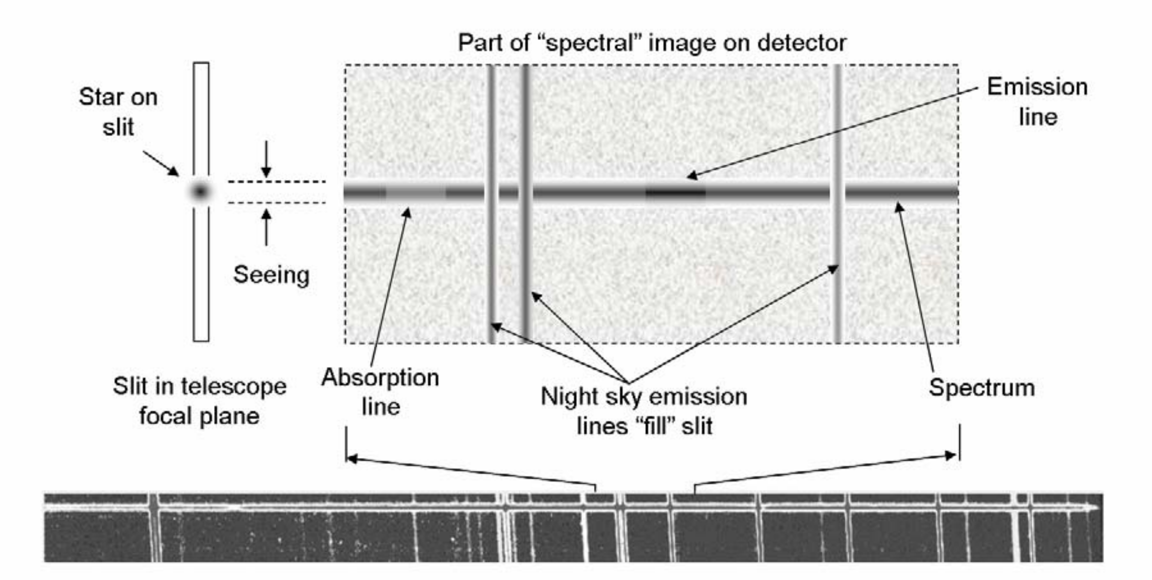

Part of "spectral" image on detector
Star on slit
Emission line
Seeing
Slit in telescope focal plane
Absorption line
Night sky emission lines "fill" slit
Spectrum

Fig. 9.14 An illustration of the typical appearance of a spectrum on an array detector, including the presence of night sky lines. The width is determined by the seeing.

9.7 SPECTROSCOPY

Spectroscopic applications put the most stringent requirements on detectors and drive the quest for the lowest possible noise performance. A majority of large astronomical spectrographs employ diffraction gratings to disperse light, although prisms with transmission gratings applied to them are also frequently used. Figure 9.14 shows the "image" of a spectrum on the detector. A cross-dispersed echelle spectrum that fills the detector is shown in Fig. 9.15. Spectroscopic calibrations proceed in much the same way as with imaging. A flat-field is required to remove optical interference effects caused by the near-monochromatic light and variations in the thickness of thinned backside illuminated CCDs. Bias frames must proceed the derivation of the flat-field as before, but now, because of the much weaker signals and longer integrations, dark current may be more serious. Many dark frames may need to be averaged and subtracted from the object frame. The flat-fielded spectra must be "sky-subtracted". To do this it is normal to collapse or sum together all rows of the spectrum containing source flux and all rows containing sky spectra. After allowing for the fact that there will probably be more rows of sky spectrum than source spectrum, the pair of flat-fielded, summed spectra are subtracted. Next, the relationship between pixel number and wavelength is determined using arc lamp (emission-line) spectra containing numerous lines with accurately known wavelengths. In the near-infrared it is often convenient to use the numerous, sharp OH night-sky lines for wavelength calibration as they are already “built-in” to the observed spectrum. Atmospheric extinction corrections are applied to the intensity and correction to absolute flux levels is accomplished by forming the ratio of the observed spectrum to that of a flux standard. Stars of type A0 V that are not too distant and therefore not reddened by interstellar dust are used. Fainter objects, such as white dwarfs which have almost featureless spectra, are also utilized as flux standards.

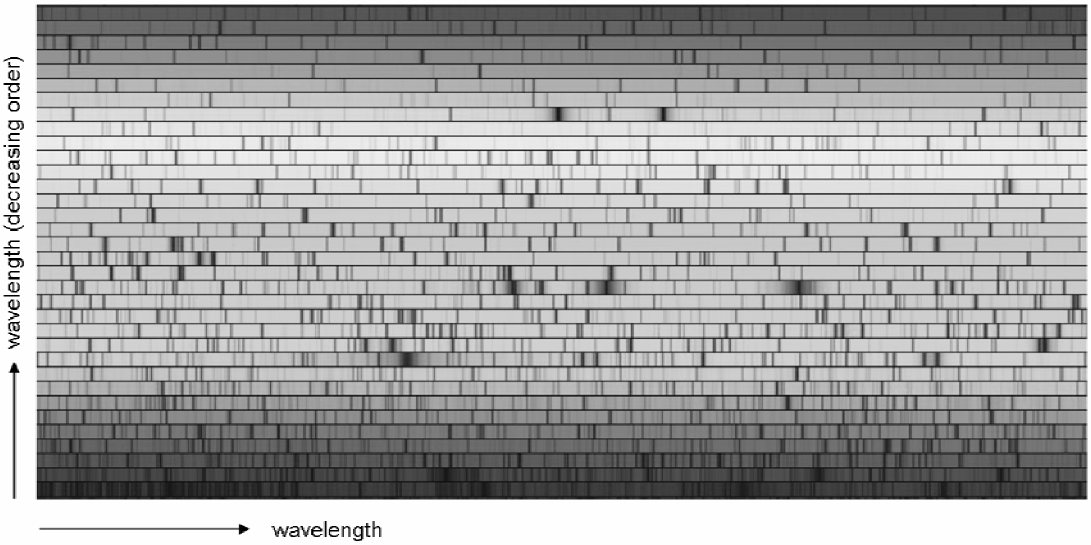


Fig. 9.15 A cross-dispersed echelle spectrometer fills the detector array with many spectral segments.

A summary of the key steps are as follows:

(1) Identify the direction of dispersion (is increasing wavelength the same as increasing pixel numbers?)

(2) Interpolate over dead pixels or columns to prevent these extreme deviant values from ruining the subsequent steps, but keep a "map" of these locations so as not to forget that there was no real data in those pixels.

(3) Sum up and normalize to unity the flat fields. Flat fields are usually taken with the spectrograph slit wide open if that is an option. In this way, the orders overlap considerably giving a uniform illumination on the CCD when viewing a quartz lamp illuminating a white screen on the inside of the dome.

(4) Divide the observed stellar spectra by the flat field to remove pixel-to-pixel sensitivity variations.

(5) Some software packages (e.g. IRAF) require that you define the positions of the orders across the CCD. This may require observations of a brighter star if the program star is too faint. The program (e.g. "aptrace") will then know where to find spectra.

(6) Extract the rectangular subsets of CCD pixels corresponding to the stellar spectrum. Do the same for the arc lamp using the normal slit width and quartz-lamp exposures. In IRAF this is done with a program called "apsum".

(7) Divide the flat-fielded stellar spectra by the "white-light" spectrum obtained with the normal slit and quartz lamp. This white light spectrum must itself be flat-fielded

10

Image processing and analysis

Computers are used in the control of telescopes and instruments, for acquisition of digital data from electronic detectors, for image display, analysis, numerical simulations and more. In this chapter we present some important terminology and explain the standard astronomical data file format known as FITS. Astronomical software is a vast subject beyond the scope of a single chapter. Our goal here is to introduce some well-known software packages but focus more on basic concepts including image visualization, high- and low-pass filtering, false color and image restoration by deconvolution.

10.1 COMPUTERS

The growth rate of computer technology is enormous. Keeping up to date is best done by watching for reviews and surveys in both trade and popular magazines. Our emphasis here is on those aspects of computer technology likely to be of most interest to astronomers, such as acquisition of digital data from a detector, image analysis software and data storage facilities. The next two sections introduce some terminology.

10.1.1 Data acquisition and data transfer

The "data rate" is the number of digitized pixel values transferred per second to the host computer. For an array detector, this rate is largely determined by the electronic configuration, although the actual application and device physics may also be factors. Recall that the integrating correlated double sampling (CDS) method used with CCDs requires an interval of time to digitize a pixel value, typically about 20 to 100 μs, corresponding to pixel readout rates of 50 kHz and 10 kHz respectively. For a 20 μs pixel time and digitization to 16 bits (65,536 voltage levels, or 0 to 65,535), the maximum output data rate is 800,000 bits/second. For detector systems running at 5MHz however, the rate would be 100 times larger. Note that this level of digitization is mandatory if one is to capitalize on the large dynamic range of CCDs. So this data rate is 100 kilobytes/s. Serial data rates are usually specified in bits per second (bps) and parallel rates in Mbytes/s. Driven by diverse applications ranging from audio reproduction to military radar, analog-to-digital converters (ADCs) are getting faster each year. When selecting an ADC it is important to look carefully at performance details. A key parameter is the signal-to-noise ratio in decibels (db) – more is better. Today's low noise CCDs can reveal limitations in ADCs. For example, as the digitization process is discrete then signals will be rounded up or down by 0.5 DN, leading to digitization noise. Also, if the conversion "accuracy" or signal-to-noise ratio is equivalent to several bits (e.g. 2 bits = 4 DN) then at some ADC step the error will be ±4*g* electrons, where *g* is the number of electrons per DN, but which step has this error will be unknown. For a 16-bit ADC with a 1 bit error (>90 db signal-to-noise ratio) then the effect is only 0.003%. Transfer of data from the ADC to the computer can be accomplished in two ways. Either by a high-speed serial link which is a single line along which signals are sent in a sequential pattern, or by a multi-way cable called a parallel link which sends all 16 bits at once.

One of the most commonly encountered serial standards is RS-232-C in which a "logic 1" level is represented by a voltage in the range -5 to -15 V, and "logic 0" by a voltage between +5 and +15 V; up to 25 lines are specified, although seldom used. In the RS-423-C standard, only two wires are used and the logic voltages are 0 and 5 V. The RS-422-C is similar but the lines are "balanced" to achieve much higher data rates. In personal computers, RS-232 has been superseded by USB (Universal Serial Bus) which is faster, has lower voltages and smaller connectors, and can also handle data transfer. The current version (USB 2.0) runs at 480 Mbit/s.

A rather common method of data transfer today is the use of a high-speed, single-cable serial link called Ethernet which is also used as a general means of inter-linking several computers in a Local Area Network or LAN. In the original Ethernet approach (due originally to Xerox, DEC and Intel) messages and data are broadcast along a coaxial cable from the sender station to the receiver station. To send a message, the sending station senses the cable to see if it is free. If it is free the station transmits but it also continues to sense in case some other station on the net transmitted at the same time. If this were to happen a "collision" would have occurred and the transmission would be unsuccessful, so both stations stop sending, wait a random time interval and then try again. This is called "carrier sense multiple access with collision detection" or CSMA-CD. Originally, coaxial cable was used but it was replaced by thin-wire Ethernet based on unshielded twisted pairs, with a system of linked hubs ultimately replacing the CSMA-CD method for a full duplex system in which devices at both ends can receive and transmit at the same time. Rates improve steadily, but 100 Mbits/s to 1Gbits/s is common at present.

Of extreme importance has been the introduction of optical fibers. Optical transmission, which is insensitive to electrical interference, has rapidly emerged as the best high-speed, high-capacity, low-error communications link for data transmission. In an optical fiber, the electrical pulse which would have been sent down a copper wire is converted to a pulse of light by a transmitter at one end and then back into an electrical pulse by a receiver at the other end. Typically, the optical fiber itself has a very small diameter of only 50 - 75 microns.

For parallel data transfers, many computers handle this flow by using Direct Memory Access (DMA). Parallel data highways within a computer system are called "buses" and there are different buses for data transfer to memory, for control and for addressing. The number of lines in the address bus determines the number of memory locations that can be specified. A 32-bit bus gives over 4 billion locations. Two commonly encountered systems are SCSI (pronounced SKUH-ZEE) and PCI. In general, digital data from an astronomical detector are saved initially on a magnetic disk drive (the "hard drive") and later transferred to another medium such as an optical disk (CD or DVD) for longer term storage; this process is called "archiving".

For compatibility between two devices, several parameters must match besides the rate, in particular, the voltage levels, the timing, the format, the "code" by which numbers and characters are digitized, and the rules or "protocols" by which data transfers are acknowledged—called "hand-shaking"—and by which computer operations are tracked, a process called "housekeeping". One instrument control philosophy/environment known as EPICS, developed at the Los Alamos labs in the USA, is being used at several large astronomical centers. EPICS is based on a transport layer called Channel Access, a message system protocol over Ethernet hardware, which connects Unix host computers to smaller systems running VxWorks (Wind River Systems, Alameda, CA), typically in a VME crate. VME is a bus system used in mid-size computers such as the early Sun Microsystems machines. Many different CPUs are available to run in a VME bus system including Motorola MC68000 and Sun Sparc families. VME is capable of very high speeds, extremely compact hardware and receives widespread commercial support. The VME bus is designed for input-output paths which are 32 bits wide.

Another type of data-acquisition system uses the IEEE-488 standard which is based on the Hewlett-Packard (Palo Alto, CA) general-purpose interface bus (GPIB). Input-output data paths are 8 bits wide (parallel) and so transfer of a 16-bit word consists of two serial bytes. Each unit on the bus is independent and connections to the bus are by means of a standard 24-pin connector which allows other connectors to be added on to it to create a "stack" of several devices on the same bus. Instruments connected to the bus are designated talkers, listeners, and controllers. Again, a computer program is required to control the modules on the bus, but the support for this system is very widespread and many commercially-available pieces of lab apparatus (such as oscilloscopes, temperature monitors, voltage-frequency converters) come with an IEEE-488 interface.

Many astronomical instruments employ powerful Digital Signal Processors (DSPs) for detector and instrument control. DSP chips are made by several manufacturers (e.g. Motorola and Texas Instruments) and are widely used in equipment of all kinds, from cellular phones to automobiles. Several observatory-built CCD and infrared camera systems employ DSP chips as the "intelligence" in a controller/sequencer.

10.1.2 Data file formats

In a CCD system the “photon image” on the detector is converted to a “charge image” in the pixels. The (x,y) location of the charge image accurately mimics the location of the arriving photons and the amount of charge is (usually) linearly proportional to the number of photons at that location in the image. During the readout process each charge packet is converted to a voltage and then digitized with an analog-to-digital converter (ADC) in order to associate a number with each pixel location. If the ADC is a 16 bit device then the numbers representing the intensity in the image will range from 0-65,535. Those numbers are stored in a two-dimensional array or table in a computer such that the correct intensity is associated with the appropriate pixel (x,y). This is the “digital image”. It may be stored on a disk or held in memory, but the CCD image is now an array of numbers arranged just like the columns and rows of the detector itself. Usually, the image on disk is a master copy that is protected from change. Copies can be manipulated by operations on the individual digital pixel values.

There are numerous formats for image files such as bitmap (BMP), GIF, TIFF, JPEG, and description of these can be found by a simple web search. However, the recognized standard format among professional astronomers is the Flexible Image Transport System (FITS) developed by Don Wells, Eric Greisen and Ron Harten (1981). Sometimes instruments save data in a "native" form consistent with the software environment at the observatory. Nevertheless, files are always converted to FITS for transport away from the observatory. FITS is used throughout astronomy. FITS files consist of three parts: a "header", the image "data" in binary form, and a "tailer". Unlike JPEG and GIF images, FITS images cannot be viewed in web browsers. Special display software is required.

A FITS “header” comprises an integer multiple of 36 lines of 80 bytes (the 80 bytes is a relic of 80-character punched cards) giving 2,880 bytes, or 5,760 bytes for 72 lines and so on. If less than 36 lines are used then the remainder must be filled out with the ASCII (American Standard Code for Information Interchange) character for a blank space (hexadecimal value of 20). Each line, also called a "card image", begins with a "keyword" in bytes 1 through 8, which identifies the information type for that line. The construction of the keywords is very specific. Each word must be left-justified and consist of only eight valid ASCII characters with no blank spaces except at the end, to pad out the keyword to eight characters if necessary. Longer keywords such as TELESCOPE are contracted to TELESCOP to remain within the 8-character limit. Only uppercase letters, the digits 0 through 9, periods and hyphens are all that is allowed. Bytes 9 and 10 may contain an equal sign and a space if the keyword has an associated numerical or text value. Numerical values are always right-justified between bytes 11 through 30, whereas text strings begin with a single quote at byte 11 and must end (with a single quote) by byte 80. An optional "comment" can be added after the value if separated by a space followed by a slash (/). When a keyword has no associated value, then bytes 9 through 80 can contain any ASCII text characters. The following order of keywords is required: SIMPLE, BITPIX, NAXIS, NAXIS1, NAXIS2, ... NAXISn, and END.

Table 10.1 FITS keywords and their meanings.

SIMPLE	has the value in byte 30 of either T (true) of F (false): simply a statement of whether or not the file conforms to the FITS standard.
BITPIX	an integer describing the number of bits in the data values. Options are 8, 16, 32 for 8-bit, 16-bit and 32-bit unsigned integers. Floating-point data can be represented e.g. -32 and -64 for 32-bit and 64-bit respectively.
NAXIS	the dimension of the data array. If value is zero, no data follows. Value of 1 for 1-d data such as intensity values in a spectrum. For image data, NAXIS = 2 (e.g. rows and columns of CCD) and NAXIS = 3 would be used for a data cube of spatial coordinates versus velocity. The maximum value of NAXIS is 999.

NAXIS1, NAXIS2, NAXISn	Each specify number of elements along that axis, with convention that NAXIS1 is the axis whose index changes most rapidly and NAXISn is the axis whose index changes the slowest. For example, in a CCD image the number of columns would go in NAXIS1 and the number of rows in NAXIS2.

Completing the header is the keyword END which is located in bytes 1-3 and the remaining fields (to 80) are filled with ASCII blanks. Several optional keywords may be inserted after NAXISn and before END. For instance, BSCALE and BZERO relate the array values and the true values through the relation:

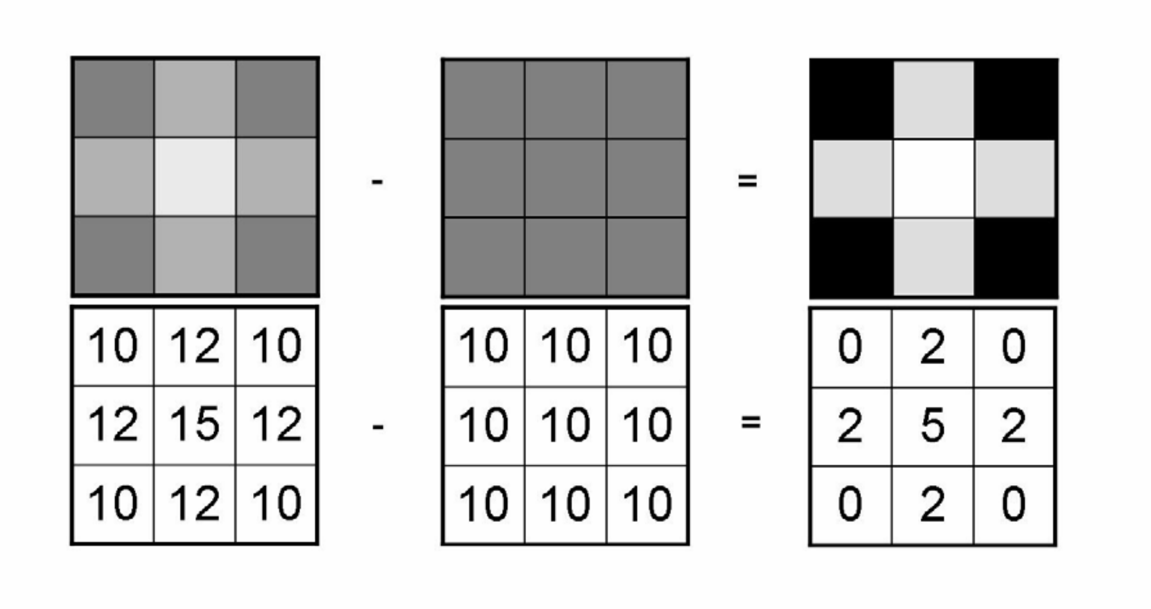
True value = BSCALE x array value + BZERO

and can be used to convert signed 16-bit array values (-32,768 to +32,767) into unsigned 16-bit pixel values (0 to 65,535) by setting BZERO to 32,768.0 and BSCALE to 1.0. Standard self-explanatory additional keywords with associated character strings are OBJECT, TELESCOP, INSTRUME, and OBSERVER. DATE-OBS and DATE have character string values and are intended to record the date on which the observations were obtained and the date on which the header was written respectively. The usual format for the date is dd/mm/yy and Universal Time is preferred. The keyword ORIGIN is used with a character string to identify the institution creating the FITS file. COMMENT and HISTORY are two keywords which do not have associated values and any valid ASCII text can be inserted in bytes 9 through 80. Any number of COMMENT or HISTORY lines is allowed, consistent with the 36 line header. If more than 36 lines are required then the keyword EXTEND should be inserted before line 36 and unused lines up to 72 will need to be padded with the ASCII blank character.

Immediately following the header (at byte 2,881) begins the “data” in a continuous sequence according to the NAXIS parameters already declared. According to the FITS standard, 8-bit integer data must be represented by unsigned binary integers contained in one byte and 16-bit data values must be stored as signed binary integers in two bytes with the most significant byte first. This convention is not followed by many computers (especially PCs) or programs and consequently, "byte-swapping" may be needed to import FITS data into another system. Also, although 16-bit digitization is standard, many imaging systems use 12, 14 or 15 bits and some use 24 bits. The BZERO keyword can be helpful in offsetting the zero point of the stored values and the 24 bit numbers would require to be handled using the 32-bit convention of four bytes with the most significant byte first. Finally, the “tailer” of a FITS file is ASCII null (00) characters used to pad out the final 2,880-byte record.

10.2 DATA REDUCTION AND ANALYSIS SYSTEMS

For most astronomers, image processing means simply, "data reduction", and is largely associated with the simplest aspects of visualization and mathematical manipulation of the grid of numbers stored in the computer to represent the intensity $I(x,y)$ (photon counts) at each (x,y) pixel location on the array detector during the given exposure time. Raw electronic imaging data may be uneven in appearance until it is correctly calibrated.



$$\begin{bmatrix} 10 & 12 & 10 \\ 12 & 15 & 12 \\ 10 & 12 & 10 \end{bmatrix} - \begin{bmatrix} 10 & 10 & 10 \\ 10 & 10 & 10 \\ 10 & 10 & 10 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 0 \\ 2 & 5 & 2 \\ 0 & 2 & 0 \end{bmatrix}$$

Fig. 10.1 Illustration of a simple operation on digital image data to subtract one frame from another.

Usually the two-point correction of dark/bias subtraction and flat-field division is adequate. Most processing steps must be carried out pixel-by-pixel. For example, we will need to remove the pattern of dark current which was accumulated during the CCD integration. To do this we subtract a data array containing the "dark" frame from the data array containing the "raw" frame using a vector arithmetic algorithm which moves from pixel to pixel calculating the difference and entering the answer in a new data array. The process is illustrated by the simple graphic in Fig. 10.1. Next, the dark-subtracted frame can be divided by another digital image representing the (normalized) flat-field. Again using a pixel-by-pixel division, the number in the first data array is divided by the corresponding pixel value in the second data array, and the quotient is entered into that pixel location in the output data array. Of course, simple scalar arithmetic is also possible. For example, each and every pixel value can be multiplied (or divided) by a constant number and the product is entered into that pixel location in the output array. One example might be division by the exposure time so that the numbers now represent counts per second. Software is also required for a number of other steps required to “reduce” the data to a form suitable for further analysis.

There are several major suites of computer programs which have been developed specifically to support the reduction of astronomical data, and especially for CCD-type imaging and spectroscopic systems. Most of these programs have been "packaged" within an environment which allows the users to select the appropriate task and even set up a sequence of tasks to be performed without writing and compiling computer code from scratch. The most well-known packages are:

- **AIPS**
- **STARLINK**
- **IRAF**
- **MIDAS**
- **STSDAS**
- **IDL**

Initially, this display is a linear mapping of true intensity to values in the display range, which might be 0 - 255 for instance to give 256 "levels". The conversion is stored in a LUT (or Look-Up Table). If the weakest intensity is set to correspond to 0 and the brightest signal assigned 255, then all intermediate signals are binned into the intermediate levels as shown in Fig. 10.3a. It is generally advantageous if the display software has a cursor which can be moved over the displayed image to "read back" to the screen the (*x*,*y*) pixel coordinates and the true intensity at that point in the image (not the scaled value from 0-255). If the dynamic range in the image is large, $I_{max} >> I_{min}$, then the resulting linear mapping does not have good contrast. One way to bring up the faint end is to significantly reduce the intensity level assigned to 255 (white). For instance, if we set the white level to 10% of the peak signal, then the remaining range of signals is mapped into 255 resulting in a display in which all the brighter objects are white, but all the fainter signals are now visible. The display is said to have been "stretched" and it is clear that the "transfer function" is steeper. In fact, the 0 (black level) can also be moved and need not correspond to the weakest signal. Thus, any "window" of signal levels can be stretched from 0 to 255 display levels (see Figure 10.3b).

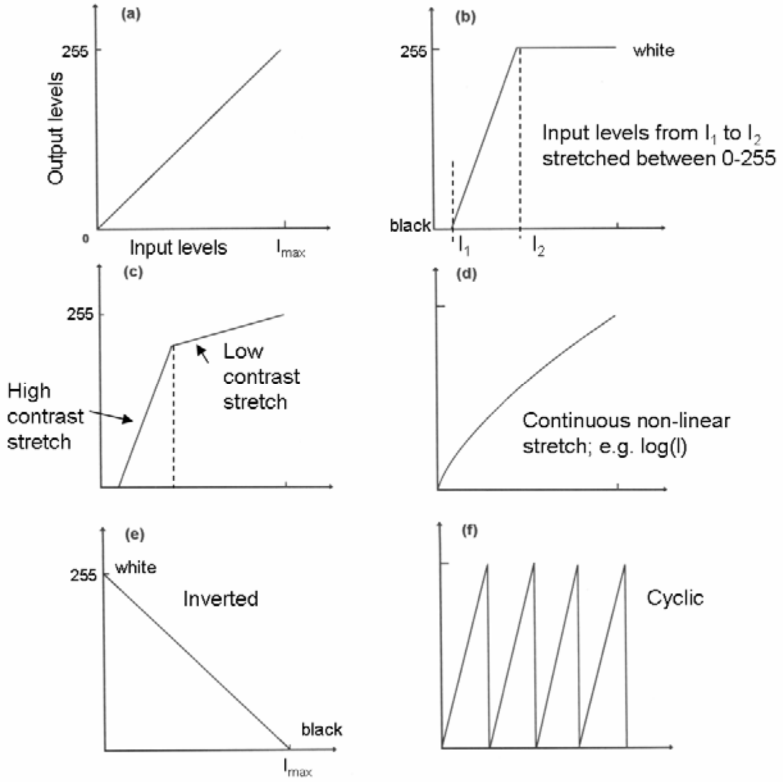


Fig. 10.3 Examples of six Look-Up Tables (LUTs) or display-stretching transformations are shown. (a) linear; (b) linear between two intensities; (c) 2-step linear; (d) logarithmic; (e) inverse and (f) saw-tooth or wrap-around.

A variation on this linear stretching approach is to add a point in between the black and white levels and use a different steepness of transfer function in each part. For instance, a steep transfer function could be applied to display data with signal levels of 0 - 10% of the peak value and these could be mapped into 0 - 200 levels, with the remaining signal levels (10% - 100% of peak) being mapped to the 201 - 255 levels as in Figure 10.3c. Such a plot gives a good stretch to the faint end without grossly "over-exposing" the bright end. Because the transfer function is now non-linear, although composed of two straight line segments, we might as well consider *any* non-linear mapping of signal to display levels. Enhancing the contrast of faint objects near the sky brightness level can be done with a non-linear transformation such as a "logarithmic curve" (Fig. 10.3d) which rises steeply at first to increase the contrast of faint objects, but levels off more slowly to compress the bright end of the map. Repetitive, linear (saw-tooth) ramps can also be used to "wrap-around" all the grey levels several times (Fig. 10.3f). Finally, one of the most powerful non-linear distortions is a transformation which "equalizes" the histogram of signal values versus the number of pixels with that signal. This condition is illustrated in Fig. 10.4. Histogram equalization is very good at bringing out faint objects near the background level.

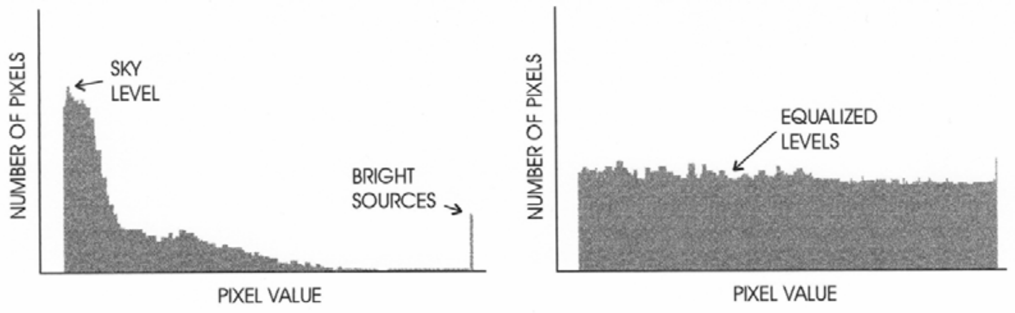


Fig. 10.4 (Left): Histogram of the distribution of signal values in the image. In a sparse field most of the pixels record the sky value. (Right): A display transformation that results in equalization of the histogram brings up faint objects.

To see the effect of changing the stretch consider the left hand image in Figure 10.5 which shows the result of stretching the input levels linearly from 0-255. Compare this with the result on the right where “white” is set at 25 instead of 255. This figure was made using FITS Liberator. Notice that a histogram of the distribution of signal values is provided.

In the above discussion we have assumed that the 256 display levels represent shades of "grey", but it is also possible to match each interval to a particular shade of color. When the distribution of brightness in an image is represented by arbitrary colors it is known as a “false-color” representation. False-color display is really a simple form of image enhancement. Compare Figure 10.6 in shades of grey with its false-color version in the Color Plates section.

Even in visible light, astronomical images are usually obtained with a combination of optical filters and detector sensitivities that don’t remotely match that of the human eye. Therefore, the term “true color” must be used very cautiously. Many spectacular color images are available to the general public and they are often represented implicitly as natural color, but if you could really see these objects with your eyes they would simply not appear as colorful as shown.

11

Electronic imaging at infrared wavelengths

Because of the intrinsic band-gap of silicon, CCDs do not respond beyond 1.1 µm. To cover the huge infrared range out to at least 150 µm requires different materials and techniques. In this chapter we describe the remarkable development and impact of infrared "array" detectors and the enormous explosion of infrared facilities. In contrast to UV, X-ray and gamma-ray astronomy, there are extensive opportunities to make infrared observations from ground-based sites, as well as from the stratosphere and from space.

11.1 INTRODUCTION

So great was the impact of the CCD that it is frequently said to have "revolutionized" optical astronomy. That same pronouncement would seem like an understatement for the advent of the "infrared array" about a decade later. Infrared observations are extremely important in astrophysics for many reasons. For example, because of the Hubble expansion of the universe, the visible light from distant galaxies is stretched, effectively moving the observed spectrum into the infrared for the most distant objects[2]. Equally important, infrared wavelengths are much more penetrating than visible light, and can therefore reveal the processes at work in star forming regions which are typically enshrouded in clouds of gas and dust. Similarly, infrared observations can allow us to "see" all the way to the center of the Milky Way and reveal the nature of the central mass of the galaxy. Cold interstellar material emits no visible light but it does emit in the far infrared, which provides a means to study the dust itself. Lastly, energy transitions in molecules that involve quantized rotation and vibration states result in the emission of low-energy infrared photons, and thus infrared

[2] The wavelength shift is measured by the scale factor (1 + z) where z is called the "redshift"; that is, $\lambda = \lambda_o(1 + z)$. For example, for a galaxy at redshift z = 2.5, when the universe was only 1/(1+2.5) = 0.29 its present size, the familiar red H-alpha line from hot hydrogen gas at 656.3 nm is found at the infrared wavelength of 2297 nm (≈2.30 µm).

spectroscopy is a powerful diagnostic tool to probe the chemistry of the interstellar medium and the coolest stars.

For many years the infrared part of the spectrum was considered to be the region just beyond the red limit of sensitivity of the human eye, at a wavelength of about 720 nm (or 0.72 µm). With the advent of CCDs, "optical" astronomy extended its territorial claims to about 1.1 µm, the cut-off wavelength for detection of light imposed by the fundamental band-gap of silicon (recall that $\lambda_c = 1.24/E_G$; for $E_G = 1.13$ eV, $\lambda_c = 1.1$ µm). So where is the "real" optical-IR boundary for ground-based astronomy? A reasonable response is that it occurs at 2.2-2.4 µm because, as shown in Chapter 2, at these wavelengths there is a marked and fundamental change in the nature of the "background" light entering the telescope/detector system. Consequently, there is a practical change in observing methods and instrument design. For wavelengths shorter than ~2.2 µm the background light comes mainly from OH emission in the Earth's upper atmosphere, whereas at longer wavelengths the dominant source of background radiation is the thermal (heat) emission from the atmosphere and telescope optical components. The domain of infrared astronomy is typically subdivided as follows. Near-infrared (NIR) is now taken to be the interval from about 0.9-5.5 µm, although the term short wave infrared (SWIR) is used specifically for 0.9-2.5 µm and the *thermal* near-infrared refers to the part from 2.5-5.5 µm. Current NIR detectors already overlap with CCDs for wavelengths less than 1.1 µm and new devices will perform down to ~0.5 µm. Large format IR arrays are available and thus NIR merges smoothly with the classical optical regime. Mid-infrared (MIR) extends from ~5-30 µm and far-infrared (FIR) stretches from ~30 to ~200 µm. Observations at these longer wavelengths are more challenging from the ground hence the interest in observations from the stratosphere. Wavelengths longer than about 200 µm (or 0.2 mm) are now referred to as the sub-millimeter, and although sub-millimeter astronomy is closely allied with infrared wavelengths in terms of the objects and regions of space which are studied, some of its techniques are more akin to those of radio astronomy. To appreciate the remarkable transformations in IR astronomy in recent years and the impact of technology, it is worthwhile to consider briefly the historical development.

11.1.1 Early history of infrared astronomy

Infrared astronomy had an early albeit somewhat accidental origin when, in 1800 in a series of papers, Sir William Herschel (1738-1822) discoverer of Uranus, noted that a thermometer placed just beyond the reddest end of a spectrum of sunlight not only increased its temperature compared with two other thermometers set off to the side, but also showed a greater heating than any other location within the spectrum. Herschel called these unseen radiations "calorific rays" and proved that they were refracted and reflected just like ordinary light. Herschel's discovery occurred about 65 years before Maxwell showed that light was only one form of electromagnetic radiation. The prolific Sir William also made another observational discovery which he called "holes in the sky" by which he meant irregularly-shaped dark regions where the dense distribution of Milky Way stars, so visible in the eyepiece of his large telescope, seemed simply to vanish. These dark blobs, absent of stars, would be cataloged by pioneering American astrophotographer E. E. Barnard (1857-1923) almost a century later, and recognized as dense clouds of gas and dust. But it would not be until the Swiss-born US astronomer Robert J. Trumpler (1886-1956) proposed in 1930 that the general interstellar medium was filled with interstellar “dust” which affected distance measurements by dimming and reddening blue light much more than red light, that a motivation for infrared studies would slowly emerge. Herschel had no way of knowing that the “calorific rays” he had discovered would one day provide the means to penetrate and explore his “holes in the sky.”

11.1.2 The beginning of modern infrared astronomy

Despite some additional development of infrared radiometers, and a variety of infrared observations of stars and solar system bodies up through the 1920s (see Martin Harwit’s chapter in *The Century of Space Science*, 2001 and Harwitt 1999 for more details), the major breakthroughs did not come until after World War II. Rapid developments in infrared detector technology were stimulated not by any commercial market but by military requirements. Lead sulfide (also known as galena; PbS) is a semiconductor with a fundamental (direct) band gap of 0.41 eV at room temperature dropping to 0.286 eV (4.3 μm) at 4.2 K. PbS was used in the classical photoconductor mode at 77 K (Chapter 5) with a fixed voltage across the detector.

In the early 1960s two physicists Gerry Neugebauer and Bob Leighton at the California Institute of Technology (Caltech) began a "two-micron sky survey" (TMSS) with an angular resolution of 4 minutes of arc. The pair constructed their own survey telescope, a 1.57-m (62-in) f/1 parabolic dish, by machining the primary mirror from aluminum metal on a lathe. Then, to improve the surface finish, they poured on a layer of epoxy and spun the paraboloid about a vertical axis until the material set; this technique is similar to the principle used today in the spinning furnace to shape molten glass into a deep parabolic curve. After applying a reflective coating the telescope produced images of about 2 arc minutes, which was sufficient considering the survey had to cover about 30,000 square degrees of the sky above a declination of -30°. Their detector was eight separate PbS photoconductors used in a pair-wise fashion to alternate between detectors in order to remove background radiation. In 1965 Neugebauer, Martz and Leighton announced the discovery of incredibly bright, “first magnitude” infrared sources with extremely faint optical counterparts. This was not expected. Then, based on observations made in 1966 by graduate student Eric Becklin, he and Neugebauer (1968) announced the infrared detection of the objects at the center of the Milky Way, not seen in visible light due to 30 magnitudes of extinction. Another Becklin and Neugebauer find lay in the heart of the well-known Orion Nebula where they discovered a very bright yet optically invisible young star (now named the BN-object).

Many of the new sources showed a trend of increasing brightness at longer wavelengths where PbS was not sensitive. A crucial step forward to exploring longer wavelengths was the invention of the liquid helium-cooled gallium-doped germanium (Ge:Ga) bolometer in 1961 by Frank Low, another physicist who was at that time working for the Texas Instruments Corporation. Frank later moved to the National Radio Astronomy Observatory and then to the University of Arizona in Tucson, where he not only established a formidable infrared program, but also set up a company, called IR Labs, to provide cryogenic detector systems to other researchers. Using a cryogenic instrument at the telescope was a challenging prospect in those days. However, because of its wavelength-independent response, the gallium-doped germanium detector opened up much longer wavelengths to astronomers. Frank built detectors for 10 and 21 µm. He also developed a telescope and bolometer system that could be mounted on a NASA Lear Jet and flown above most of the terrestrial water vapor, thus enabling him to observe at 70 µm. Infrared observations from balloons and rockets also began in the mid-60s. In 1967 Doug Kleinmann and Frank Low reported observations at 22 µm that led to the discovery of an extended infrared-glowing cloud near the BN object in Orion, now known as the Kleinmann-Low nebula, and by 1970 it was realized that some distant galaxies emitted far more infrared radiation than all other wavelengths combined (Kleinmann and Low 1970).

Still driven by military requirements for heat-seeking devices, the lead sulfide cell was replaced by a more sensitive photodiode made from indium antimonide (InSb). Don Hall (now at the University of Hawaii) played a key part in the introduction of InSb to astronomy (Hall *et al*. 1975) and has continued to push the frontiers of infrared detector developments ever since. Stimulated by the discovery rate, there was a push for telescopes that were "optimized" for infrared work. In 1970 the Mount Lemon Infrared Observatory was established in the Catalina Mountains near Tucson, Arizona, while UK astronomers built a 1.5-m infrared “flux collector” on Tenerife in the Canary Islands in 1971. Harry Hyland pioneered the study of the southern hemisphere skies at Mount Stromlo using a series of detectors (Hyland 1971). By 1978 infrared observations with a single-element detector were being made on the Anglo-Australian Telescope by David Allen (1946-1994) and colleagues, and Allen had produced the first book on the “new astronomy”. The need for better far-infrared observations led to the development of the Kuiper Airborne Observatory (KAO), a modified C-141A jet transport aircraft with a 91.5 cm (36-in) Cassegrain telescope, capable of operating at altitudes of up to 14 km (45,000 ft). The KAO made many significant discoveries including the first detection of faint rings around Herschel’s planet – Uranus – in 1977. The KAO operated out of the NASA Ames Research Center at Moffett Field, California from 1974-1995.

By 1979, a new generation of 3-4 meter class telescopes dedicated to infrared astronomy had come into operation including, the United Kingdom 3.8-m Infrared Telescope (UKIRT) and the NASA 3-m Infrared Telescope Facility (IRTF) whose first director was Eric Becklin. Both of these telescopes were located on the 4.2 km (14,000 ft) summit of Mauna Kea, Hawaii which was rapidly becoming recognized as an exceptional site. Other telescopes optimized for infrared astronomy soon followed, and some "optical" telescopes began to provide optional configurations for infrared work, including the 5-m Hale telescope. At wavelengths longer than 2.4 µm, moonlight is almost undetectable and so infrared became the "bright time" or full Moon option. The period up to the launch of IRAS has been reviewed by Low, Rieke and Gehrz (2007).

11.1.3 The launch of IRAS

The Anglo-American-Dutch Infrared Astronomical Satellite (IRAS) mission gave astronomers their first deep all-sky survey in the infrared. With the launch of IRAS on January 25, 1983, infrared astronomy took a quantum leap. The IRAS mission mapped the entire sky at wavelengths of 12, 25, 60 and 100 µm, produced a point source catalog of over 245,000 sources (more than 100 times the number known previously), and made numerous unexpected discoveries, including a dust shell around the "standard star" α Lyrae (Vega) and about 75,000 galaxies believed to be in a “starburst” state. IRAS had a lifetime of only 10 months in operation until the on-board supply of 475 liters of superfluid helium coolant, which held the detectors at 1.8 K and everything else at about 10 K, was finally exhausted. The 60-cm telescope and its detectors then warmed-up and lost their sensitivity. IRAS was so successful that follow-up missions involving "observatory class" cryogenic satellites were planned by both ESA and NASA. The European project, called ISO (Infrared Space Observatory), was launched successfully in late 1995 and operated until 1998, while the American project, initially called SIRTF (Space Infrared Telescope Facility) was delayed and then finally launched in 2003, at which time it was renamed the Spitzer Space Telescope. In the interim period an infrared instrument (NICMOS) was placed into service on the Hubble Space Telescope.

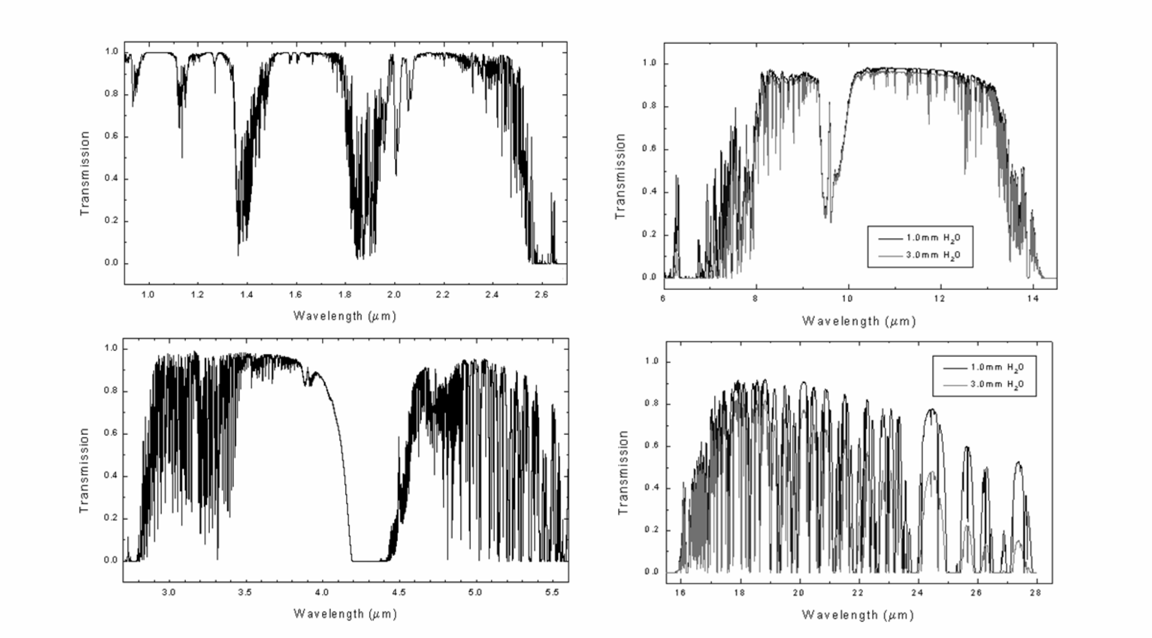


Fig. 11.1 Details of the near-infrared transmission profile of the atmosphere above Mauna Kea (14,000 ft) for a typical water vapor level. Plots created using the ATRAN code developed by Steve Lord. Credit: Gemini Observatory web site.

11.2 INFRARED WAVEBANDS

11.2.1 Atmospheric windows

Water vapor (H_2O) and carbon dioxide (CO_2) block out a lot of infrared radiation from space. Figure 11.1 shows a more detailed transmission spectrum of the Earth's atmosphere in the near- and mid- infrared than given in Chapter 2. These plots are derived from models using the ATRAN software developed by Steve Lord (1992) and are available on the Gemini Observatory web site. Water vapor absorption is sensitive to altitude and occurs in certain wavelength intervals, between which the atmosphere is remarkably transparent. These atmospheric "windows" of transparency allow astronomers to define photometric bands. The standard windows are listed by central wavelength and the full width at half maximum intensity (FWHM) in Table 11.1.

Table 11.1 Infrared windows in the Earth's atmosphere

CENTER WAVELENGTH (μm)	DESIGNATION OF THE BAND	WIDTH (FWHM) (μm)
1.25	J	0.3
1.65	H	0.35
2.2	K	0.4
3.5	L	1.0
4.8	M	0.6
10.6	N	5.0
21	Q	11.0

There is also a relatively poor window (designated X) from about 30-35 μm which is accessible from dry high-altitude sites or from Antarctica. There are variations of some bands (such as K_{short}; 2.0-2.3 μm, K′; 1.95-2.30 μm and L′; 3.5-4.1 μm) which have been developed to improve performance at given sites. Interference filters can be manufactured to match these windows. Several different filter sets are in use and therefore care must be taken when comparing photometric observations with those of others.

11.2.2 The high background problem

There are two major sources of unwanted background photons. One component is OH emission lines and the other is the blackbody thermal emission from the telescope which, even at an ambient temperature close to 0° C (273 K), emits prodigiously in the infrared. Thermal emission from any warm optics in the beam can be predicted from two quantities: the absolute temperature T (K) which determines the spectrum of the radiation from the Planck function $B_\lambda(T)$, and the emissivity $\varepsilon\ (\lambda)$ of each component which determines the fraction of blackbody radiation added to the beam. Objects which "appear" black to our visual senses may not be black at longer wavelengths, i.e. they may reflect some infrared light. To estimate the emissivity (ε) of telescope mirrors (due to absorption) we can apply Kirchhoff's Law and take one *minus* the measured spectral reflectivity. For example, if the reflectivity is measured to be 96% then the emissivity is 4% and is additive for the train of warm optics. There will be an additional component of emissivity from dust on the mirror surface. To emphasize how bright the night sky is at infrared wavelengths we can compare the brightness in magnitudes of one square arcsecond in the blue (λ=0.43 μm) $m \approx 24$ (no moonlight), with that at 2.2 μm in the near IR, $m \approx 13.5$, and also at 10 μm where the sky and telescope combined are brighter than $m \approx 0.0$ (depending on emissivity and temperature)! The most effective way of eliminating telescope background is to cool the entire telescope. On Mauna Kea at 14,000 ft above sea-level the temperature hangs around 1 °C. Temperatures in Antarctica near the South Pole are lower still, ranging from -13.6 °C to

-82.8 °C and hence have stimulated the development of Antarctic astronomy. In the stratosphere the average temperature is about -50 °C and the residual level of water vapor is extremely low. Consequently, flying a telescope on an airplane to this altitude can be a very effective alternative to space missions.

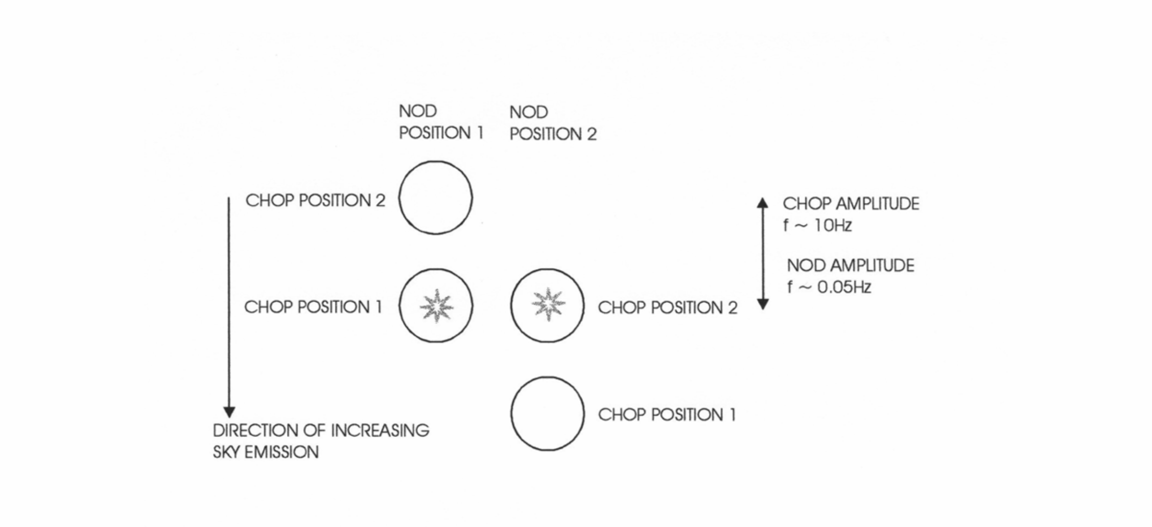


Fig. 11.2 Chopping and nodding remove background flux and gradients at infrared wavelengths.

11.2.3 Chopping

Early infrared astronomers found a solution to the problem of a bright sky, it is a technique called "chopping". The infrared beam is rapidly switched between the source position on the sky and a nearby reference position, by the use of an oscillating or "wobbling" secondary mirror in the telescope itself. Typical wobbling secondary mirrors on IR-optimized 3-4 meter telescopes such as the IRTF and UKIRT are small, 0.24-0.31 m in diameter, and have a slow f/ratio (~f/35). On the larger 8-10 meter telescopes these secondary mirrors are bigger and more massive (e.g. 1.0 m for Gemini). Moreover, on alt-az telescopes the chop direction must be variable. Chopping typically takes place at a frequency of ~10-20 Hz. In a photometer, this method involves isolating the astronomical object in a small aperture and first measuring the total brightness of "object plus sky" included in the aperture. Chopping changes the location of the image in the focal plane quickly so as to record the signal from a nearby piece of sky containing no objects in view. By forming the difference, the sky signal is eliminated provided it has remained constant. In addition, it is usually necessary to move the entire telescope every minute or so to enable the sky on the "other side" of the object to be measured and thereby eliminate any systematic trend or gradient; this step is called "nodding" and the amount of the nod is usually the same as the "throw" of the chop for symmetry (Fig. 11.2). The difference between the pair of chopped signals for nod position 1 is given by

$$C_1(x) = S + B_{tel,1} - B_{tel,2} + \left(\frac{d}{dx} B_{sky}\right)\Delta x \qquad (11.1)$$

where B_{tel} and B_{sky} are the telescope and sky backgrounds at the two chop positions separated by Δx. These terms are usually always much larger than the source flux, S. For the second nod position the signs are reversed and the difference signal is

$$C_2(x) = S - B_{tel,1} + B_{tel,2} - \left(\frac{d}{dx}B_{sky}\right)\Delta x \tag{11.2}$$

and adding these two results gives the required source signal

$$S = \frac{1}{2}\left(C_1(x) + C_2(x)\right) \tag{11.3}$$

Chopping and nodding are generally required at wavelengths longer than about 3.5 μm, and nodding alone is required for good background subtraction at shorter wavelengths too.

Another reason for using a secondary mirror with a slow f/ratio is to significantly reduce the background on a given detector pixel by stretching the plate scale. For example, going from an f/9 secondary to an f/36 gives a smaller scale in arc seconds per mm by a factor of 4, and reduces the flux falling on each square mm by a factor of $4^2 = 16$. An infrared secondary mirror is "undersized" to permit chopping and therefore it is over-filled by the beam from the primary mirror. This means that the primary mirror no longer defines the entrance pupil of the system; it is now defined by the size of the secondary mirror. In general, the secondary mirror is not surrounded by a black baffle tube in the normal way, because it is important to ensure that any subsequent image of the secondary formed inside the instrument is surrounded by sky, which produces a lower background than a warm, black baffle. Often, the secondary will be gold-coated for best infrared performance since gold is more reflective that aluminum in the IR. In addition, there will be either a small deflecting mirror or a hole at the center of the secondary with access to the sky. Such precautions eliminate thermal photons from the central Cassegrain hole in the primary mirror. Telescopes built this way are said to be infrared optimized.

11.3 INFRARED ARRAY DETECTORS

11.3.1 The infrared "array" revolution — déjà vue

Lacking the long pre-CCD history of photographic imaging enjoyed by optical astronomy, it is easy to appreciate the staggering boost to infrared astronomy that occurred when the first true array detectors were introduced (e.g. McLean 1988, 1995). Reviews of infrared detectors and materials are given by Paul Richards and Craig McCreight in *Physics Today*, February 2005 and by George Rieke (2007) in the *Annual Reviews*.

Many forms of infrared array devices with closely-packed pixels were constructed during the period 1974-1984 by several different companies, due mainly to the extreme importance of the infrared for military applications. Both CCD and CID (charge-injection device) readouts were used and different materials were evaluated. For example, Koch *et al.* (1981) described work on InSb monolithic charge-coupled infrared arrays at Santa Barbara Research Center; Kosonocky *et al.* (1981) reported on the development of a 256-element

PtSi Schottky-barrier IR CCD line sensor at the RCA Labs; Baker *et al*. (1981) described the work at Mullard in the UK to make a 32x32 HgCdTe photovoltaic array hybridized to silicon circuitry; and Rode *et al*. (1981) reported on Rockwell's hybrid arrays fabricated in HgCdTe or InAsSb and multiplexed to a Si CCD via direct injection. Hybrid array development was reviewed by D. H. Alexander of Hughes Aircraft Co. (1980). However, despite a rich technical literature, even by 1982 when I carried out a detailed survey at the suggestion of Malcolm Longair, Director of the Royal Observatory Edinburgh (ROE) and the UK Infrared Telescope, pixel formats for available devices that had made it outside the classified arena to astronomers were very small (32 x 32 or less), very few devices were actually for sale and none had the performance needed for low-background astronomical applications. Prospects seemed bleak.

An early champion of astronomical infrared array devices was Craig McCreight of the NASA Ames Research Center who led in-house tests and coordinated a major NASA-funded program involving a number of other groups (e.g. McCreight 1981). As early as 1979, John (Eric) Arens and co-workers at the Goddard Space Flight Center tested a 32 × 32 pixel bismuth-doped silicon CID array made by Aeroject ElectroSystems at a wavelength of 10 μm (Arens *et al*. 1981, 1983). Astronomical observations were published in Arens *et al*. (1984) and in the thesis of Richard Tresch-Fienberg. Elsewhere, a 32 × 64 platinum silicide (PtSi) Schottky Barrier array was evaluated by the Kitt Peak National Observatory (Dereniak *et al*. 1984) and at the NASA Infrared Telescope Facility (Rich Capps) in collaboration with the US Airforce.

A lot of military funding had gone into the development of mercury-cadmium-telluride (HgCdTe) devices for mid-IR work, but much of that work was classified. Long linear photodiode arrays using individual, switched MOSFET multiplexers were tested and described in the technical literature; one of the best of these arrays was a 32 element linear array of indium antimonide (InSb) developed by Jim Wimmers, Dave Smith and Kurt Niblack at Cincinnati Electronics Corporation which was used successfully by astronomers in near infrared spectrographs (Niblack 1985). Each of the early devices always had some drawback for astronomy, such as poor quantum efficiency or high readout noise (~1,000 electrons) and of course, just not enough pixels! The most hopeful sign came from tests of a 32×32 array of InSb detectors reported in 1983 by Judith Pipher and Bill Forrest of the University of Rochester (USA) at a NASA Ames detector workshop organized by Craig McCreight. This device was a "reject" loaned to them by Alan Hoffman, a former colleague who was now employed by Santa Barbara Research Center (SBRC) in California (now Raytheon Vision Systems). Astronomical results from this camera were published in Forrest *et al*. (1985).

Having visited Judy and Bill in Rochester and made them aware of UKIRT's interest in this kind of array from an, as yet, undisclosed source, they passed on the information to Alan Hoffman. Alan and I had already met, but I was not aware that he was the source of Judy and Bill's detector. We made contact again and Alan approached his senior management with the prospect that a major observatory was interested in helping to develop a device optimized for astronomy. Wisely, SBRC also approached the Kitt Peak National Observatory (now the National Optical Astronomy Observatories - NOAO) and found interest there too. From 1982-1984, I worked closely with senior SBRC engineers (Alan Hoffman and Jim West) and marketing personnel (Dick Brodie and Carol Oania) to define the technical specification and costs of a new infrared array device suitable for astronomy (McLean and Wade 1984). Our goal was to image using sub-arcsecond pixels just like CCDs. Smaller beams on the sky meant much lower background levels per pixel, an unusual condition compared to the high-background strategic applications. Unfortunately, the original readout design proposed by SBRC had to be abandoned as too complex and subject to amplifier glow. Luckily, an alternative readout scheme was suggested and already tested by Al Fowler at NOAO, but with platinum silicide (PtSi) rather than InSb as the detector material. This was the basic source-follower per detector. We all agreed that the alternative device was going to be noisier than desired but should still work, and fabrication began on the well-known 58 × 62 InSb array in 1984. Meanwhile, Judy Pipher, Bill Forrest, Giovanni Fazio (Harvard) and others were working towards an instrument definition for NASA's Space Infrared Telescope Facility and were also negotiating for detector development work at SBRC; this instrument would become the Infrared Array Camera (IRAC) on Spitzer.

At about the same time, contracts were being developed for second-generation Hubble Space Telescope instrumentation. Several groups of US astronomers (in Arizona, Chicago and Hawaii) had obtained access to new arrays made from mercury-cadmium-telluride (HgCdTe or MCT) from Jon Rode of the Rockwell International Science Center in Thousand Oaks, California. By a circuitous route I had already met Jon in 1982 when carrying out the detector survey for the Royal Observatory and UKIRT. Having learned that the UK had a program of HgCdTe array development at Mullard Ltd., I visited Ian Baker there, but he suggested that if our interests were only 1-5 µm then I should try Rockwell in California. At our meeting in 1982 however, Jon and I concluded that the development of a low-background, near-infrared MCT array would be too expensive for ground-based astronomy at that time. In addition, this material was still largely classified. By the mid-80s however, things were beginning to change. Stimulus for the astronomy MCT array development was generated by funding for a proposed new instrument for the Hubble Space Telescope. The instrument was called NICMOS and the principal investigator was Rodger Thompson at the University of Arizona. Marcia Rieke and Rodger Thompson at University of Arizona, and Mark Herald at University of Chicago were among the first to successfully demonstrate the new Rockwell arrays at the telescope. Subsequently, Jon Rode moved up in the company at Rockwell and his role was taken over by Kadri Vural. Both Jon and Kadri have been strong supporters of the astronomy detector programs ever since and over the years Kadri has played a very significant role in encouraging and promoting the development of (MCT) arrays for wide use. This development would prove to be very important, because these array devices can be customized to the shortest IR wavelengths (~2.5 µm) and could be run at 77 K using liquid nitrogen and operated by existing CCD controllers.

Several European sources of infrared array technology, for space applications in particular, were also recognized, and French astronomers were already using an InSb array of 32x32 pixels with the charge-injection principle (Sibille *et al.* 1982). Preparations for both ISO and SIRTF (Spitzer) had stimulated work on longer wavelength devices in both Europe and America, again much of the NASA development being supported by Craig McCreight's program. In addition, a very important development that had occurred in 1979 would soon make its way to astronomy arrays. Mike Petroff and Dutch Stapelbroek, at what was then another division of Rockwell (now DRS Technologies), perfected a new way to construct extrinsic silicon photoconductors that significantly improved their performance. The method was called blocked-impurity-band (BIB) and this array technology would be developed later for Spitzer and ground-based applications. In addition, work was under way at the University of Arizona and elsewhere to develop doped germanium detectors into a viable form of array device. Thus excitement was high!

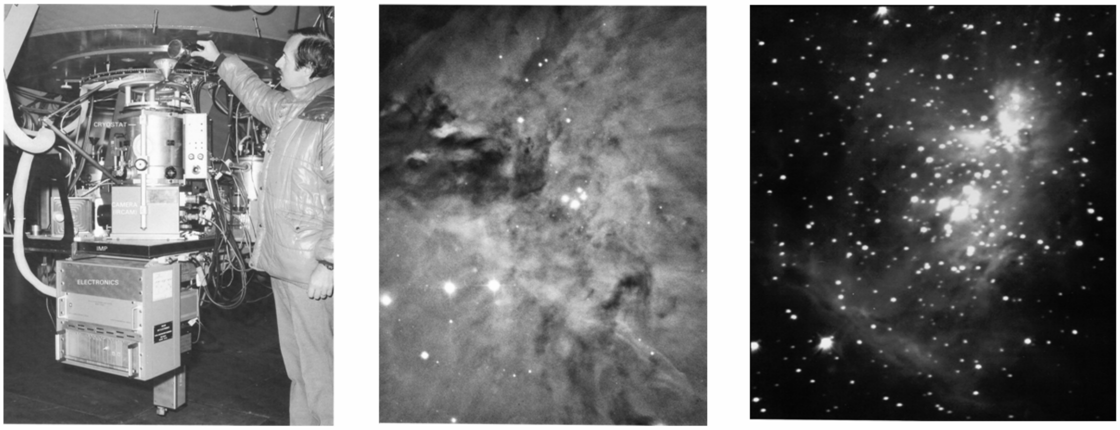


Fig.11.3 (a) The author with IRCAM (1986), the first common-user camera system on the UKIRT 3.8-m infrared telescope to employ the 58x62 InSb arrays from SBRC (Raytheon). (b) A visible light image of the Trapezium region of the Orion Nebula. (c) An infrared image of the same region obtained with IRCAM at a wavelength of 2.2 microns. The bright source above the Trapezium is the Becklin-Neugebauer (BN) object.

The first project to develop a true, user-friendly, facility-class infrared array camera based on the new 58 × 62 InSb array began at the Royal Observatory Edinburgh (ROE) in June 1984. I was the principal investigator for that development and two years later in September 1986 we delivered IRCAM to the 3.8 m UKIRT in Hawaii (Fig. 11.3). I was fortunate to have the enthusiastic support of Eric Becklin who was at ROE on sabbatical leave from the University of Hawaii. Al Fowler received his device at NOAO at about the same time, and so he and I were frequently in touch across continents, trying to compare results as we went along. Indeed, the development was not trouble free. For example, the batch of InSb used to make the astronomy devices suffered a loss of quantum efficiency when cooled to operating temperature (~30 K). This was a trying time for SBRC manager Dave Randall and scientists like Alan Hoffman and Geoff Orias. Fortunately, InSb material with completely different doping had been developed simultaneously for the SIRTF (Spitzer) project, and so the ground-based program was able to acquire some of that material which performed very well. So little was known about how these “astronomy arrays” would work that Al Fowler and me, as well as Alan Hoffman's team at SBRC, wanted to be very cautious until we had "first light" on the telescope. In my case that event occurred at 8 am on the morning of October 23 1986, in broad daylight! Together with my long-time colleague Colin Aspin—who had written much of the software for the camera—we obtained the first infrared image of a cosmic source with IRCAM, it was the Orion nebula (far right in Fig. 11.3). Also present at the telescope were our colleagues Gillian Wright (now PI of the MIRI instrument for JWST) and UKIRT operator Dolores Walther, who were both more than a little surprised by the quiet manner in which Colin and I accepted this momentous event. Of course, we had taken thousands of test images and knew that the camera would work, and certainly on something as bright as the BN object. We became much more jubilant in the months ahead as more and more challenging targets were observed and it became clear that the flood gates had opened (McLean 1985, 1987). Two graduate students working on that project with me were Mark McCaughrean (1988) and John Rayner (1988), and they were responsible for producing remarkable images of OMC-1 and OMC-2 with IRCAM. Although the SBRC detector had an array of only 3,596 pixels that was 3,595 more pixels than we had before!

By March 1987, the first astronomical results from several of the new infrared arrays had begun to appear. A key moment in infrared astronomy was a "workshop" on infrared array detectors in Hilo, Hawaii in March 1987. The meeting was organized by Eric Becklin and Gareth Wynn-Williams of the University of Hawaii, Honolulu, with local support from David Beattie and me on behalf of UKIRT. Don Hall gave the summary at the end of the meeting. The first true images from the new arrays were very encouraging and we all realized that infrared astronomy had changed. To those of us who had straddled the apparent divide between optical and infrared astronomy, it was like history repeating itself. For me, the euphoria was similar to the Harvard-Smithsonian meeting on optical CCDs in 1981.

Eric Becklin and I joined forces in 1989 and moved to the University of California, Los Angeles to build infrared instruments for the new Keck 10-m telescopes. Six years later, in July 1993, we hosted a meeting at UCLA entitled "Infrared Astronomy with Arrays: the Next Generation". By then, everyone was already using 256 × 256 detectors for near-infrared work and 128 × 128 devices at mid-IR wavelengths, and plans for 1024 × 1024 arrays were announced at that conference (McLean 1994). Short-wavelength mer-cad-tel was emerging as a powerful means for optical telescopes to extend their capability to 2.5 μm, while the longer wavelength arrays of InSb and extrinsic silicon (Si:As) were proving better than anyone had dared hope. Moreover, it was clear from the papers presented by 300 participants from all over the world that the new detectors had been quickly assimilated into the subject and that a wide range of new astrophysics was being produced. At the time of writing (2008), near-infrared arrays (1-5 μm) with formats of 2048 × 2048 pixels and mid-infrared arrays (5-30 μm) with 1024 x 1024 pixels are standard. Gallium-doped germanium detectors for 70 and 160 μm have been made into arrays of 32x32 pixels for space applications. Moreover, just like CCDs before them, these devices are being built into larger mosaics for both cameras and spectrometers, or becoming the heart of more complex instruments such as diffraction-limited cameras and integral field spectrometers.

12

Electronic imaging at ultraviolet, X-ray and gamma-ray wavelengths

Astronomy in the ultraviolet, X-ray and gamma-ray parts of the spectrum can only be carried out from above the Earth's atmosphere using satellites or rockets. In the UV and X-ray regimes we again come across the CCD, but other important electronic imaging technologies are also used for reasons that will be explained. High energy photons with extremely small (sub-atomic) wavelengths require different kinds of detectors and even the telescope design must change. Within the scope of this text, we can only illustrate some of the most important innovations and advances that have made imaging possible in these exciting fields.

12.1 INTRODUCTION

In terms of photon wavelengths, the ultraviolet region spans roughly 300 nm to ~10 nm, X-rays from ~10 nm to ~0.01 nm and gamma-rays the regime below that down to and below nuclear dimensions (one millionth of 1 nm). This also represents an enormous range in photon energies, more than a factor of a billion. As described in Chapter 2, photons with wavelengths shorter than 300 nm (3000 Å) are not transmitted by the Earth’s atmosphere. Consequently, observations in the ultraviolet (UV), X-ray and gamma-ray regimes must be carried out from space. If normal optical telescopes with parabolic mirrors can be used then the diffraction-limited formula for angular resolution with a circular aperture ($\theta \sim\lambda/D$) suggests that ultraviolet observations should have better resolution than visible light images for a given telescope aperture because of the smaller wavelength. However, the surface quality of the mirror becomes very important as the wavelength is decreased. Recall that the Strehl Ratio is related to the rms amplitude of the surface roughness by $S = \exp[-(4\pi\sigma/\lambda)^2]$ and a $\sigma = \lambda/20$ surface will scatter 33% of the light out of the diffraction spot. At 500 nm this corresponds to a 25 nm rms surface smoothness, but at 100 nm in the UV it would imply a surface smooth to 5 nm. The Hubble Space Telescope is a conventional Ritchey-Chrétien telescope possessing one of the smoothest primary mirrors ever polished with a surface roughness of ~2-3 nm which allows it to perform into the UV. Clearly, for even smaller wavelengths, achieving diffraction-limited performance from reflecting telescopes becomes challenging. But there is another problem too. In general, highly polished metallic surfaces have a higher reflectance than dielectric materials like glass. At normal incidence, silver and aluminum reflect over 90% of all visible light, which is why metallic coatings are applied to glass telescope mirrors. The amount reflected increases to 100% at grazing angles of incidence. Metallic reflection is a function of wavelength. For example, silver (Ag) has a strong minimum reflectance (<10%) around 320 nm. The optical properties of a dielectric are specified by its refractive index (*n*), but for a metal we need another property, the absorption index ($\kappa_0 = a\lambda/4\pi$) to measure the attenuation of photons due to interactions with the high density of free electrons in metals. In this expression *a* is the normal absorption coefficient (e.g. units of μm^{-1}) used in the standard exponential law of absorption $I = I_0 e^{-ax}$ and thus the intensity drops to $1/e^{4\pi\kappa_0}$ in going the distance $x = \lambda$ into the medium. The complex refractive index is described $n' = n - i\ \kappa_0$, where $i = \sqrt{(-1)}$ and the ratio of reflected to incident intensities is given by

$$r = [(n-1)^2 + \kappa_0^2] / [(n+1)^2 + \kappa_0^2] \qquad (12.1)$$

which in the absence of absorption ($\kappa_0 = a = 0$) reduces to the well-known relationship for dielectrics. In practice n' can be expressed as $1 - c - id$ at very short wavelengths where *c* and *d* are values that depend on the wavelength and the material. Extreme UV photons (λ < 90 nm) and X-rays (λ < 10 nm) are either completely absorbed (large *a*) or pass right through the mirror material (small *a*, low cross-section for interactions) at normal incidence. Fortunately, these photons will reflect off the surface of certain materials if grazing incidence angles are used due to the fact that the refractive index of metals at short wavelengths is *less* than 1 ($c > 1$ in expression for n'), which creates a situation similar to light going from a denser medium (glass, water) into air and leads to a critical angle for total internal reflection. In this case, the incident photons suffer total *external* reflection at some critical angle. An approximate empirical formula for the critical angle expressed in minutes of arc is

$$\theta_{crit} = 2.2\,(\sqrt{\rho})/E \qquad (12.2)$$

where ρ is the density of the mirror material in kg/m^3 and $E = hc/\lambda = 1.24$ keV/λ (nm) is the photon energy expressed in kilo-electron volts (keV); 1eV = 1.602×10^{-19} J. For energies of a few keV the critical angle is of order 1° for materials like gold and iridium.

12.1 Grazing incidence telescopes

Employing grazing incidence implies nearly parallel orientation of the reflecting surfaces, which leads to very long focal length systems. The simplest focusing system is the parabolic dish used extensively from the near-ultraviolet to the radio to eliminate spherical aberration, but a single parabola experiences coma for off-axis rays. Abbe's sine condition requires that to be free of coma we need $r = h/\sin\theta$ where θ is the angle of incidence and h is the height of the ray from the optical axis and r is the radius of a circle around the focal point. If the radius of curvature is very large, as it is for a typical optical telescope, then coma is minimized. But x-ray telescopes require grazing incidence and so cannot have large radii of curvature; the parabola must be highly curved. This problem was solved in 1952 by the German scientist Hans Wolter (1911-1978) while working on the development of an X-ray microscope. He used a coaxial hyperboloid as a secondary mirror which was properly confocal with the initial paraboloid, just as is done at optical wavelengths, but in this case both mirrors are used at grazing incidence (Figure 12.1).

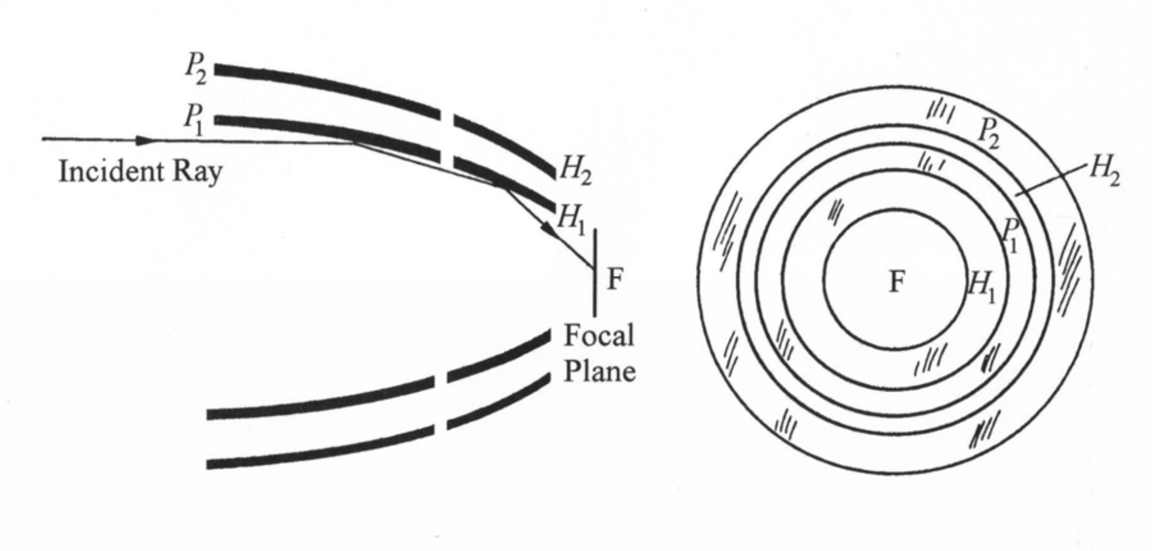


Figure 12.1 The basic concept of a Wolter 1 grazing incidence X-ray telescope.

This arrangement very nearly meets the Abbe sine condition. The focal point is found from the relation

$$Y_0 = Z_0 \tan(4\theta) \quad (12.3)$$

where Y_0 is the distance from the optical axis to the point where the two conic surfaces intersect, and Z_0 is the distance along the optical axis from the focal point to the intersection plane of the two surfaces. The angle θ is the sum of the two grazing angles on each mirror. In practice, the curvatures and angles are much less than shown and the focal point is far behind the mirrors (Figure 12.2). Wolter described three different imaging configurations, now known as Wolter Types I, II, and III, but the design most commonly used by X-ray astronomers is the Type I because it has the simplest mechanical configuration and offers the possibility of nesting several reflecting surfaces inside one another, thereby increasing the useful collecting area. Wolter Type II uses a convex secondary hyperboloid and has a longer focal length but narrower field of view. In the Type III system the initial paraboloid is convex and the secondary is a concave ellipsoid. Almost all past space-borne X-ray and extreme UV missions have used stacked Wolter-I telescopes. By stacked we mean that the paraboloids and hyperboloids are replaced with conical off-axis sections of the solids of revolution and many such sections are nested inside one another to increase the effective collecting area (Fig. 12.2).

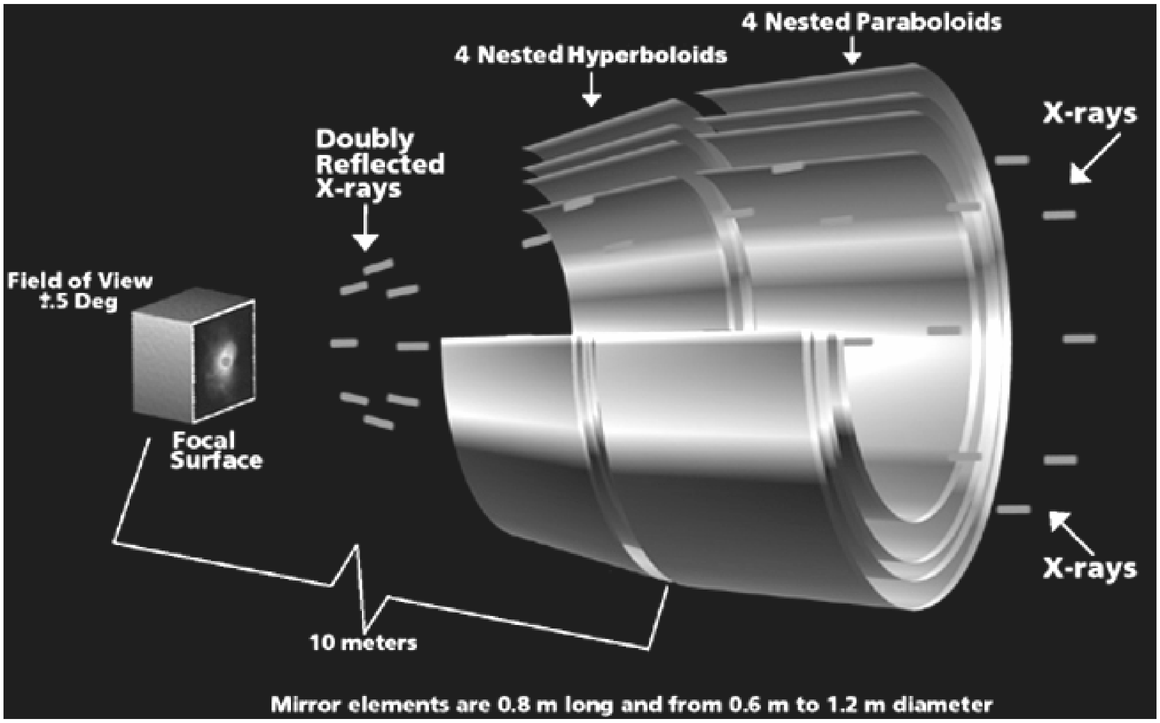


Figure 12.2 The arrangement of grazing incidence optics in the Chandra X-ray telescope. Credit: NASA/CXC

Example: For a grazing incidence angle of 0.86 degrees and an aperture height of 0.6 m the focal length is approximately 10 m. Thus the plate scale is 206265/10000 =20.63 arcsec/mm. If we want to match this telescope directly to a CCD with 24 µm pixels, then 20.63x0.024 yields a pixel resolution of about 0.5 arcsec. With conical sections about 0.8 m long formed as four concentric stacks, this set up describes the Chandra X-ray Telescope.

NASA's Chandra X-ray Observatory has 4 stacked Zerodur reflectors with an outer mirror diameter of 120 cm and a geometrical sensitive area of 1145 cm^2, corresponding to an aperture filling factor of only 0.1. In ESA's XMM-Newton observatory, the resolution is 5 arcsec using 3 modules of 58 stacked reflectors with an outer foil diameter of only 70 cm. Each module has a geometrical area of 1750 cm^2 and an aperture filling factor of 0.45. The multi-foil approach gives more collecting area but the difference in resolution between Chandra and XMM is due to the difficulties in fabricating thin X-ray mirror shells with high optical performance.

Detectors of UV and X-rays are many and varied. In the UV, a major problem is the requirement for the rejection of long-wavelength photons, in other words visible light. There are essentially two classes of detectors that cover a wide range of photon energies: photo-emissive devices and solid state devices. The latter category includes the silicon CCD. Among the photo-emissive devices are other kinds of panoramic detectors using photocathodes with large work functions like CsI and various methods of achieving pixel definition. Together with the change from classical to Wolter telescopes for the extreme UV and X-ray wavelengths, imaging is therefore possible in the UV and X-ray bands, but what about even higher energy photons?

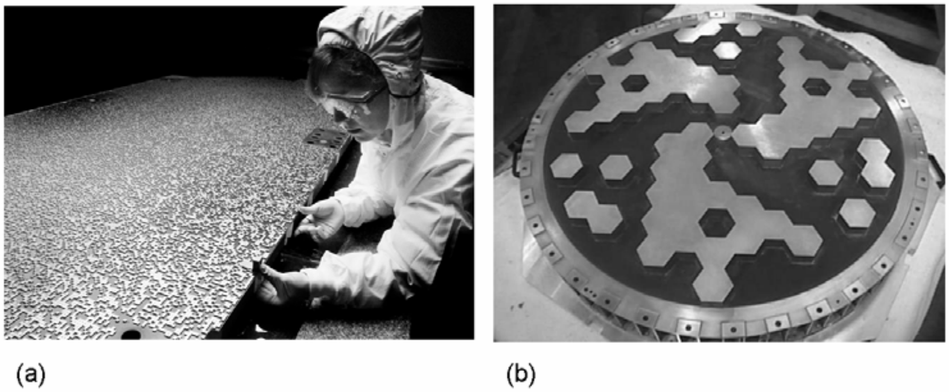


Figure 12.3 (a) The random coded-mask telescope for the SWIFT gamma-ray satellite and (b) the cyclic coded mask for the INTEGRAL gamma-ray satellite. Credit: SWIFT and INTEGRAL Teams.

12.1.2 Coded Mask Telescopes

High energy X-ray and gamma-ray photons with sub-atomic-sized wavelengths cannot be focused easily because they are too penetrating, but some degree of directionality can be imparted using a “collimator” to restrict the angle of acceptance. The simplest kind of collimator is a honeycomb of long, closely packed tubes. The angular field of view is determined by the width divided by half the length of the tube ($w/2L$). Of course, the walls have to be of dense enough material to stop the high energy photon from reaching the detector by passing through other tube walls. A variation of this approach is the “lobster eye” collimator which is a honeycomb collimator curved into a spherical shape to help direct rays to a common location. An alternative approach used on recent missions including ESA’s INTEGRAL satellite and NASA’s SWIFT spacecraft is called the “coded mask” telescope. Figure 12.3 shows the coded masks used on SWIFT and INTEGRAL.

A generic coded mask telescope is sketched in Figure 12.4. Basically, the mask contains both transparent and opaque patches and each detector pixel records the sum of the signals from a different combination of incident directions. The point source function of a coded mask telescope is not just a slightly blurred image at one location, like an Airy function for an optical telescope, but is in fact spread over the entire detector plane. However, all is not lost. The principle of the coded mask is as follows. Photons from a certain direction in the sky project the mask pattern (shadow) onto the detector; this projection has the same coding as the mask pattern, but is shifted relative to the central position by a distance that corresponds uniquely to the direction of the photons.

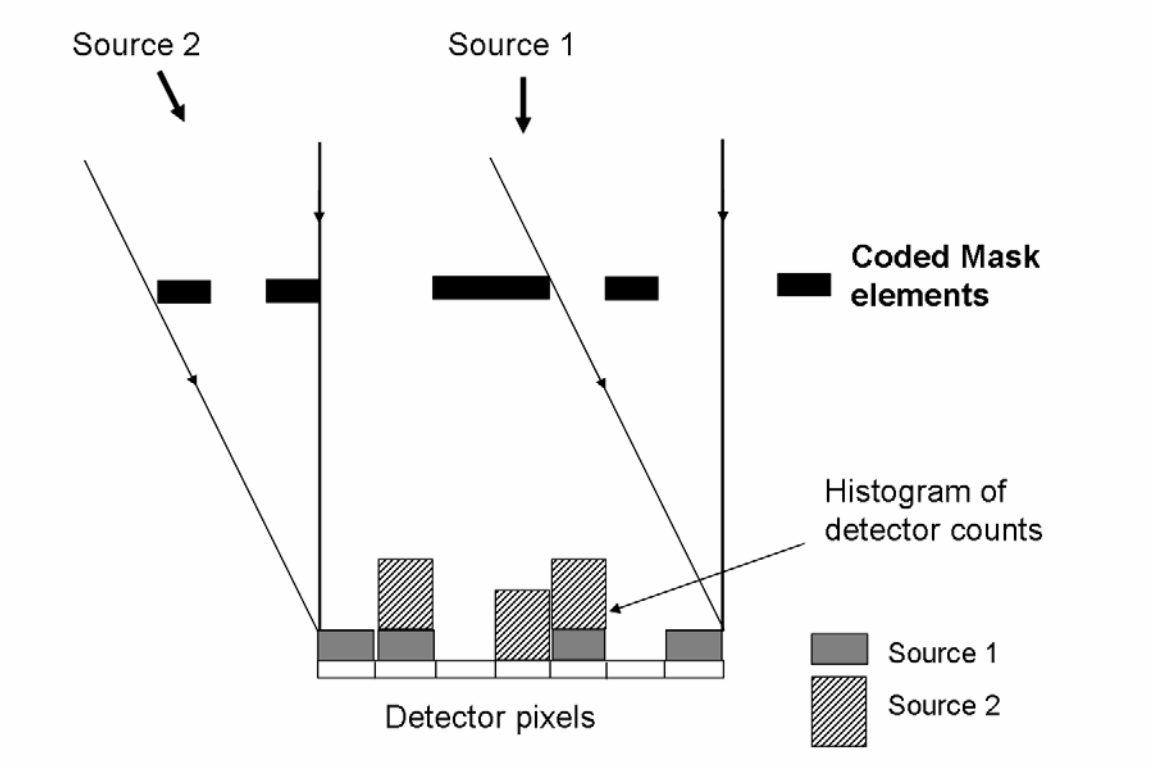


Figure 12.4 An illustration of the basis of the coded mask technique.

A two-dimensional detector, which needs to be well-matched to the mask elements, accumulates the summation signals from a number of shifted mask patterns. Each shift encodes the position, and the signal strength encodes the intensity of the sky at that position. Clearly, each part of the detector array may receive photons from any position within the observed sky. After a certain illumination period, the accumulated detector image can be decoded to an image of the sky by determining the strength of every possible shifted mask pattern using an autocorrelation algorithm. Proper performance of a coded-mask camera requires that every sky position is encoded on the detector in a unique way. Stated in terms of the autocorrelation function (ACF) of the mask pattern, this means that the ACF should consist of a single peak and flat side-lobes (a delta function), which therefore puts constraints on the type of mask pattern and on the way its (displaced) projections are detected. An important difference compared to direct-imaging systems is the fact that Poisson (√N) noise from any source in the observed sky is, in principle, induced at any other position in the reconstructed image. Thus, the imaging quality of the camera is determined by the type of mask pattern, the spatial response of the detector and the reconstruction method.

Two types of mask patterns were proposed initially: a pattern of Fresnel zones (Mertz & Young 1961) and the random pinhole pattern (Dicke 1968, Ables 1968). A camera with a Fresnel zone plate has not yet been applied to extra-solar X-ray and gamma-ray astronomy, but the concept of the random pinhole pattern has. The random pinhole pattern is an extension of the pinhole camera. A pinhole camera has ideal imaging properties but delivers a poor signal-to-noise ratio because the sensitive area is severely restricted by the size of the pinhole. Sensitivity can be increased by enlarging the pinhole, but at the expense of angular resolution. However, the open area can be increased while still preserving angular resolution by using many duplicate pinholes at random. The random character of the pinholes is necessary in order to meet the condition that the auto-correlation function be as close to a delta function as possible. Non-random patterns also exist, based on "cyclic difference sets" or "uniformly redundant arrays", which approach this ideal. The angular resolution limit in seconds of arc is the quadratic sum ($\theta^2 = \theta_m^2 + \theta_d^2$) of the mask element size term $\theta_m = 206\ (m/L)$ and the detector resolution element $\theta_d = 206\ (d/L)$ where the mask to detector separation L is given in meters and the detector pixel size and mask element size are expressed in mm; the angular resolution is degraded by $\cos^2\varphi$ for off-axis sources. The image reconstruction algorithm must collect photons from all over the detector, but noise events will also be collected from all over the detector, so all the noise contributes to each pixel. On SWIFT the coded-mask telescope has a mask size of 2400x1200 mm with 52,000 lead (Pb) elements each 5x5x2 mm in a random pattern, and the separation length results in a resolution of about 17 minutes of arc. INTEGRAL uses a cyclic difference mask 1200 mm in diameter with 72 large tungsten (W) elements each 60x60x50 mm in size. Representing the mask with an array M of 1 (open) and 0 (opaque) elements, the detector array D will be given by the convolution of the sky image S by M, plus an un-modulated background array term B thus: D = S * M + B. We need to find a special array M for which there exists a correlation inverse G such that M * G = δ-function. In which case we have that

$$S' = D * G = S * M * G + B * G = S * \delta + B * G = S + B * G \qquad (12.4)$$

S′ differs from the real sky image S only by the term B*G, which for a flat array B, is a constant term that can be measured and removed (Goldwurm et al. 2001). The implication of this discussion about imaging telescopes even for high energy photons is that there are corresponding "array" detectors, and there are. We will now pursue the detector technology by reviewing each wavelength regime consecutively.

12.2 ULTRAVIOLET DETECTORS AND INSTRUMENTS

For practical reasons associated with the technologies used, the UV is often subdivided into four regions; the Extreme UV (EUV) from ~5-90 nm (~50 to 900 Å), the Far UV (FUV) from 90-120 nm (~900 to 1200 Å), the UV from 120-200 (1200 to 2000 Å) and the Near UV (NUV) from 200-300 nm (2000 to 3000 Å). Unlike the well-defined infrared windows however, there are no sharp boundaries and sometimes three divisions are used instead, namely the near-, mid- and far-UV. Departures from visible light techniques become greater as the wavelength becomes shorter. For example, not many materials have good transmission in the UV (magnesium fluoride MgF_2 is one that does transmit down to ~180 nm), any contamination that settles on the optics will be opaque in the UV, it is hard to make UV filters that don't leak slightly at longer wavelengths and, as we have seen, as the EUV is approached the design of telescopes and other optical systems must change to use grazing incidence angles.

While several ultraviolet satellites, in particular Copernicus, TD-1 and ANS laid the early groundwork, it was the launch of the highly successful IUE (International Ultraviolet Explorer) satellite in 1978 that really opened up this vast field to all astronomers. Originally suggested by Bob Wilson as early as 1964, this remarkable NASA/ESA/UK satellite was one of the longest running space operations (1978-1996). Located in geosynchronous orbit 36,000 km (22,700 miles) from Earth, IUE carried a telescope with a diameter of 45 cm (18 in.) and was equipped with both high- and low-dispersion ultraviolet spectrographs covering the wavelength interval from 1,250 - 3,200 Å. With the launch of the Roentgen Satellite (ROSAT) in 1990 and the Extreme Ultraviolet Explorer (EUVE) in 1992, ultraviolet astronomy pushed its boundaries into the 100 - 1000 Å region. The EUVE satellite (1992-2001) supplied images in four wavelength regions across the whole EUV band and carried three EUV spectrometers. Its all-sky survey found 801 objects including the first extra-galactic detection at these wavelengths (Bowyer and Malina 1994). The majority of the sources discovered by EUVE lay within a few hundred light years of the Sun, and included such hot, young luminous stars as Eta Canis Minoris, white dwarf stars and cataclysmic variable stars like SS Cygni. Interstellar hydrogen atoms absorb EUV radiation so efficiently that if the density around the solar system was about 100 atoms per cubic centimeter then there would be enough absorption to limit our view to within 10 light years of the Sun. The fact that EUV stars were discovered at all implies that the density is much lower. The solar neighborhood lies in a low-density bubble, but EUVE's electronic imaging systems revealed that there are "tunnels" through the neutral gas in some directions.

Several UV experiments also utilized the Space Shuttle, such as WUPPE (the Wisconsin Ultraviolet Photo-Polarimeter Experiment) and ORFEUS (Orbiting and Retrievable Far and Extreme Ultraviolet Spectrometer), and of course, the Hubble Space Telescope (HST) is itself a superb ultraviolet collector. Past UV-sensitive instruments on HST include FOC and GHRS and current instruments include STIS and ACS (see Chapter 2 for more on HST). After Service Mission 4 (SM4) the no longer needed COSTAR corrector will be replaced by the UV spectrometer COS (Cosmic Origins Spectrograph).

The Far Ultraviolet Spectroscopic Explorer (FUSE), which operated from 1999-2007 was another successful mission. FUSE observed from about 90-120 nm and employed several unique design features. Instead of a single mirror four separate mirror segments (off-axis parabolas) were used, two of which were coated with silicon carbide to enhance reflectivity in the FUV and the other two were coated with lithium fluoride over aluminum which performs better at longer UV wavelengths. Light from the four optical channels was dispersed by four spherical, aberration-corrected holographic diffraction gratings on a large Rowland circle with a resolving power of $\lambda/\Delta\lambda$ = 24,000-30,000. Imaging devices using photocathodes as described below were used.

Historically, the emphasis of UV missions has been spectroscopy of point sources rather than imaging or studies of extended sources such as galaxies. Therefore, until 2003, there were only shallow all-sky surveys in the UV. GALEX (the Galaxy Evolution Explorer

Mission) launched in April 2003 is an all-sky survey down to AB magnitudes ~21 using a 50-cm (20-inch) telescope imaging in two broad UV bands centered at 150 nm and 230 nm. Understanding how galaxies were formed is the primary mission of GALEX, led by the California Institute of Technology (Martin *et al*. 2005), but its state-of-the-art UV cameras have provided some remarkable images of all kinds of astronomical sources.

13

Electronic imaging at sub-mm and radio wavelengths

Photon detectors can operate into the far-infrared, but a point is reached where no suitable shallow-doped materials exist and a transition is required to thermal detectors for the sub-mm and coherent detectors for the radio. Even so, the creation of two-dimensional arrays of detectors is still possible in principle. Recent technology developments in the sub-mm and mm bands have led to cameras with moderately high pixel densities. Aperture synthesis methods at longer wavelengths allow high-resolution mapping using multiple telescopes. Microwave images of the entire sky by the COBE and WMAP satellites have revealed the cosmic background in great detail. In this chapter we review the devices and techniques for creating images at these wavelengths.

13.1 INTRODUCTION TO RADIO ASTRONOMY

In 1932, after about four years of work studying background "static" or noise in ship-to-shore communications at a radio wavelength of 15 meters, a young radio engineer named Karl Jansky working at the Bell Telephone Laboratories in Holmdel, New Jersey (USA)—the same laboratories from which would later come the invention of the transistor, the CCD and the discovery of the cosmic microwave background—realized that a certain kind of radio noise developed a peak approximately once every 23 hours 56 minutes. The signal seemed strongest when the constellation of Sagittarius was high in the sky. As the center of the Milky Way galaxy lies in the direction of Sagittarius, Jansky correctly concluded that he was detecting radio waves from outer space. Unfortunately, Jansky's work went unnoticed by professional astronomers, but not by an engineer in Illinois, named Grote Reber. During the period 1936 - 1944 Reber completed a map of the radio emission from the Milky Way using a "backyard" concave dish 9.1 m (about 30 ft) in diameter and "tuned" to a wavelength of 1.87 m. The antennas used by Jansky and Reber are on display at the National Radio Astronomy Observatory (NRAO) in Green Bank, West Virginia (USA). Of course, the development of radar during World War II (1939 - 1945) stimulated the technology and very soon afterwards "radio observatories" began to appear all over the world.

Reber's radio dish may seem quite large by comparison with "backyard" optical telescopes—Jansky used a large rotating assembly of linear antennas (aerials) rather than a dish—but despite its physical size, the angular resolution of this and many radio telescopes is worse than the human eye. Angular resolution for a radio telescope is usually controlled by the wave phenomenon of diffraction for which the Rayleigh criterion gives:

$$\theta = 1.22\frac{\lambda}{D}\,radians \approx 70^\circ\frac{\lambda}{D} \qquad (13.1)$$

Example:
For a radio wavelength of $\lambda = 1$ m and a telescope diameter $D = 10$ m, the angular resolution is only 7° on the sky. At a wavelength of 1 mm the resolution of a 10-m telescope improves by a factor of 1000 to 25″, if the surface of the dish is smooth enough.

Radio astronomy began with equipment for detecting electromagnetic waves with wavelengths of about 1 meter. Large macroscopic wavelengths (from ~0.3 mm to ~30 m) enable groups of charged particles to produce coherent emission with fixed phase relationships between waves, which accounts for the extraordinary brightness of pulsars at radio wavelengths. And because radio wavelengths are much larger than interstellar dust grains, scattering is negligible and so the radio sky is dark both day and night, and the neutral interstellar medium is transparent. Stimulated by the huge military and civilian demand for communications, radio receiving equipment improved and observations were extended to the centimeter band and most recently to millimeter and sub-millimeter wavelengths. The sub-millimeter regime lagged behind mainstream radio astronomy largely because of lack of technology development outside of astronomy. For example, the communications industry abandoned millimeter waveguides in favor of fiber optics in the seventies. All that has changed now, and sub-mm and mm astronomy is one of the fastest growing areas in astronomy. For example, one of the largest astronomical projects to date is the international development of the Atacama Large Millimeter Array (ALMA). The immense importance of this (sub-mm and mm) waveband lies in the fact that numerous molecules have strong emission lines in this region of the spectrum, making it ideal for mapping the cold molecular gas clouds from which new stars are born.

The "window" for ground-based radio observations is quite large. For wavelengths less than 2 cm, atmospheric water vapor begins to attenuate radio signals and high altitude sites are obligatory. The warm (~300 K) attenuating atmosphere also emits radio noise that degrades sensitivity. For example, emission by water vapor above Green Bank precludes summer observations at this national facility for wavelengths below 3 cm. On the other hand, wavelengths longer than 10-20 m suffer absorption and scattering in the Earth's ionosphere. As the radio waveband is used extensively for a wide range of communication purposes, it has become necessary to regulate that certain wavebands are allocated purely for radio astronomy; a list can be obtained from the Federal Communications Commissions (FCC) or from the National Radio Astronomy Observatories (NRAO). For example, the band from 1400-1427 MHz includes the 21 cm line of hydrogen, a critical diagnostic for mapping the distribution of this otherwise invisible gas.

It is customary in radio astronomy to use frequency (ν) rather than wavelength (λ), but the two are of course easily interchanged using the relationship

$$\nu\lambda = c \equiv 2.9979 \times 10^{8}\ m/s \tag{13.2}$$

where c is the speed of light. The meter waveband corresponds to frequencies lower than 300 MegaHertz (MHz), the band from 1-10 cm corresponds to frequencies from 30-3 GigaHertz (GHz), and the millimeter and sub-millimeter waveband corresponds to frequencies above 300 GHz. The far infrared wavelength of 100 µm (0.1 mm) corresponds to 300 TeraHertz (THz). The range from 30-300 MHz is also called VHF (Very High Frequency) and the range from 300 MHz to 3 GHz is called UHF (Ultra High Frequency). Most classical radio astronomy occurs in the 1-30 GHz range and this is also called the microwave region. There are named bands within this region including the L-band at 1.5 GHz (20 cm), the S-band at 3 GHz (10 cm), the X-band at 10 GHz (3 cm), the Ku-band at 15 GHz (2 cm) and the K-band at 30 GHz (1 cm).

Example:
Using the rule of thumb that 1 mm is equivalent to 300 GHz (as opposed to 299.79 GHz), what is the frequency of the 6 cm band used by the Very Large Array (Socorro, New Mexico)? As 6 cm = 60 mm, it is 60 times larger than 1 mm and therefore the frequency will be 60 times smaller: ν(GHz) = 300 GHz/λ (mm) = 300/60 = 5 GHz.

The power received at a unit surface element per unit frequency (or wavelength) interval is called the Flux Density S_ν (or S_λ) and is usually measured in W m^{-2} Hz^{-1} (or W m^{-3}); an alternative terminology is spectral irradiance. Radio astronomers also use a flux unit called the jansky; 1 Jy = 1 Flux Unit = 10^{-26} W m^{-2} Hz^{-1}. In the 1940s, a sensitivity of a few jansky was considered good. Nowadays, signal strengths are measured in millionths of a jansky (µJy) in some cases. For an extended source the brightness (b_ν) is the flux density per unit solid angle measured in jansky per steradian. The total power (P) collected by the antenna (ignoring efficiency factors and polarization state) is given by

$$P = b_\nu A\Omega\Delta\,\nu \tag{13.3}$$

where A is the collecting area, Ω is the solid angle on the source and $\Delta\nu$ is the frequency bandwidth. For a radio telescope, the $A\Omega$ product or étendue is given by $A\Omega = \lambda^2$ because of diffraction.

13.2 RADIO TELESCOPES

13.2.1 Antennas

The simplest form of antenna (also called an aerial) is the classical half-wave dipole which consists of two conducting rods (usually copper) each one-quarter of the desired wavelength and separated by a small gap with coaxial cable going from the closer ends to a receiver. Free electrons in the conductors are set into motion by the incoming electromagnetic wave, generating an alternating electric current at the same frequency as the incoming wave. Although it may seem rather pointless to use a dipole antenna, because it is sensitive to only a rather narrow range of frequencies, some specific frequencies are of great interest such as the 1427 MHz (21 cm) line of hydrogen which has enabled astronomers to map the structure of the Milky Way. More important is the fact that the half-wave dipole has a broad point source function, which radio astronomers call the "antenna response" function or pattern. The strongest response is for a plane polarized wave with a direction of propagation perpendicular (broadside) to the conductors because the electric field in the wave is parallel to the rods in that case. Let's take this to be the optical axis. There is no response to waves coming along the axis of the rod (which makes the electric field of the wave perpendicular to the rod and so no current can flow along the rod). The response function is represented by a polar diagram which plots the response as a function of angle on the sky, and there is a reciprocity theorem that says that the transmitting and receiving pattern should be the same. It can be shown that the power pattern of a dipole depends on the field angle θ as P ~ $\sin^2 \theta$. Here the ~ sign means proportional to in this case. Therefore, the antenna gain is given by $G = G_0 \sin^2 \theta$ and for a half-wave dipole $G_0 = 3/2$ thus:

$$G(\theta, \varphi) = (3/2) \sin^2 \theta \qquad (13.7)$$

In Equation 13.7, φ is the azimuthal angle about the optical axis and in the plane perpendicular to the optical axis. In general, an antenna having a peak gain G_{max} must beam most of its power into a solid angle $\Omega \sim 4\pi/G_{max}$. The Beam Width at the Half-Power (BWHP) points represents the equivalent term to full width at half maximum (FWHM) for the Airy diffraction pattern or the seeing disk at visible wavelengths. Note that the Airy function for an optical telescope is usually represented by an (x, y) plot, but it could also be drawn as an (r, θ) plot, in which case it would be a polar diagram with side lobes just like the standard radio antenna pattern; the side lobes are tiny and very close to the main beam because λ/D is so small at optical frequencies.

In general, a dipole has rather poor directional sensitivity. By combining the signals from a stacked array of dipoles spaced at half wavelength intervals the directionality is much improved. However, as this combination behaves like the slits of a diffraction grating, interference causes additional secondary peaks or side lobes to appear in the polar pattern. Side lobes are the equivalent of the secondary maxima in the Airy diffraction pattern and they imply that the antenna will detect radiation at high angles of incidence to the main beam. The angular width of the main beam between the first minima (nulls) for n end-to-end elements is given by $\sin a = 1/n$ whereas for a broadside array of elements it is $\sin a = 2/n$. Even more directionality can be obtained by adding extra rods which are coplanar with the dipole (usually several in front and one behind acting as a pure reflector) but *not* electrically connected to the dipole or the receiver, and hence are said to be "parasitic". This is called a parasitic antenna or Yagi-Uda antenna (after its inventors Shintaro Uda and Hidetsugu Yagi of Tohoku Imperial University, Japan in 1926). The reflector element concentrates the power into the forward direction.

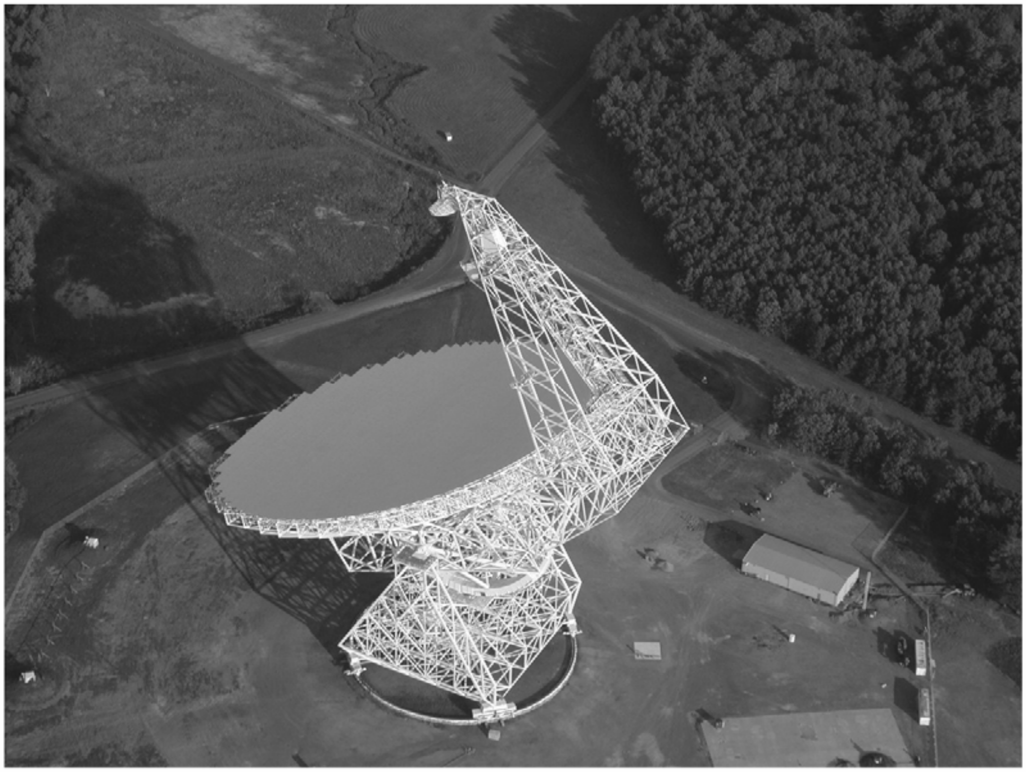


Fig. 13.1 The 100-m Green Bank Telescope (GBT) of the US National Radio Astronomy Observatories (NRAO) in West Virginia. Credit: NRAO/AUI.

The effective collecting area (A_e) of an antenna is defined in terms of the collected power ($P_1 = P_v \Delta\nu$) compared to the flux density S_1 of that part of the radio wave whose polarization coincides with the antenna. Thus $P_1 = A_e S_1 \Delta\nu$ and for random polarization $S_1 = (1/2)\ S$ and therefore the power extracted is $P = (1/2)\ A_e\ S\ \Delta\nu$. It can be shown that the effective collecting area of all lossless antennas is given by:

$$A_e = \lambda^2/4\pi \qquad (13.8)$$

This remarkable result applies to the simple dipole antenna or to a large parabolic dish antenna. Effective collecting area and gain are related by

$$A_e\,(\theta, \varphi) = \lambda^2\, G(\theta, \varphi)\ /4\pi \qquad (13.9)$$

An obvious disadvantage of the directional antennas just described is the fact that the wavelength is at once fixed by the choice of the length of dipole. A much more versatile approach is to use a large parabolic dish of metal panels as the reflector. The dish will collect more energy and bring the radio waves to a focus, effectively narrowing the power pattern so that $\theta_{\mathrm{HPBW}} \sim \lambda/D$. In practice, $1.2\lambda/D$ is a good rule of thumb. A dish can increase the maximum gain and effective area, which can in fact approach its normal geometric area of $\pi D^2/4$.

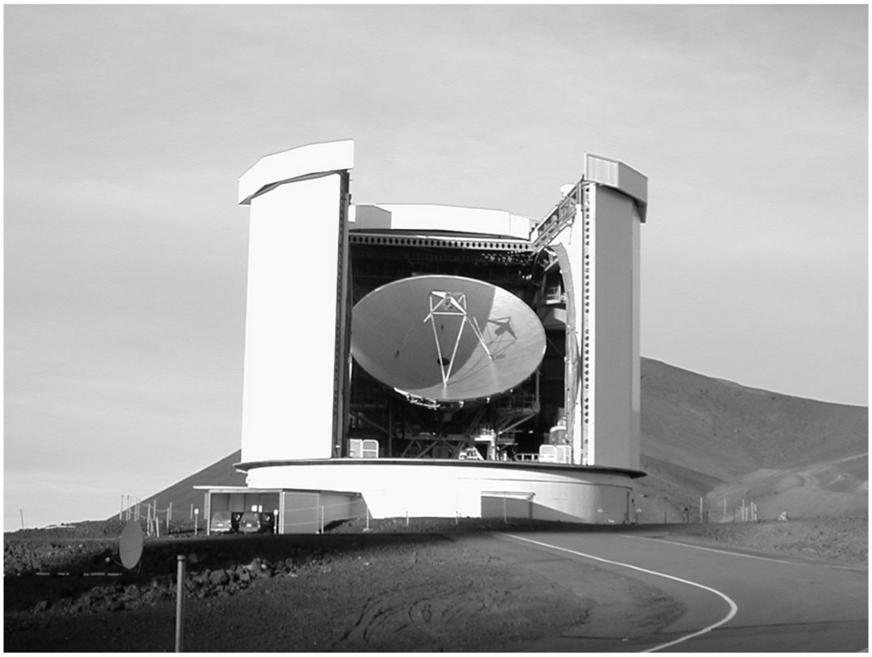


Fig. 13.2 The 15-m James Clerk Maxwell sub-millimeter telescope (JCMT). Credit: JCMT.

As long as the diameter of the dish is large compared to the wavelength then the rules of geometric optics still apply and the parabolic shape is best. Selectable dipole feeds or better, a collecting horn and waveguide that channels the wave to a probe at the focus can be used. Prime focus can be used, but so too can the Cassegrain and Gregorian foci by using convex and concave secondary reflectors respectively. Most large radio telescopes must have a very fast focal ratio $f/D \sim 0.4$ to ensure that the support structure for secondary or sub-reflector is not unreasonably large. As mentioned previously, to obtain an angular resolution on the sky of ~0.7° requires that the ratio of the diameter of the dish to the observing wavelength must be at least 100:1. To ensure diffraction-limited performance the surface must be smooth to $\sim\lambda/20$ but, as the wavelength is now much larger, the actual physical deviations from a smooth parabolic surface can be correspondingly bigger, ranging from 50 μm rms at 1 mm to 5 mm rms at 10 cm. Therefore, the world's largest steerable parabolic dishes are very large. For example, the largest single-dish radio telescope in the world is the partially-steerable RATAN-600 built in 1977 in present-day Russia which has a 576 m diameter circle of rectangular radio reflectors. The Effelsberg Radiotelescope located 40 km south of Bonn in Germany is 100 m in diameter. It has a high quality polished surface of metal plates and typically works at wavelengths around 6 cm (4996 MHz). At shorter wavelengths the surface accuracy degrades the collecting efficiency. The Robert C. Byrd Green Bank

Telescope (GBT) in West Virginia (USA), which became operational in 2000, has a surface area of 100 x 110 m. Its 2,004 panels are made from aluminum with a surface accuracy of better than 76 µm rms and 2,209 actuators adjust the panel to correct for distortions due to gravity as the telescope moves. The dish of the GBT is an off-axis parabola which means that instruments at the prime focus do not obscure the beam (Fig. 13.1). The IRAM Millimeter Radiotelescope at Pico Veleta in Spain is a carbon fiber structure 30 m in diameter with panels machined to an average surface precision of 100 µm, which permits diffraction-limited performance at a wavelength of about 2 mm. Significantly better surface accuracy (30 µm) is obtained by the 15-m James Clerk Maxwell Telescope (Fig. 14.2) on Mauna Kea, Hawaii and better still (15 µm) with the 10-m dish and panels of the Heinrich Hertz Sub-Millimeter Telescope Observatory on Mount Graham, Arizona—a joint project between the University of Arizona and the Max Planck Institute for Radio Astronomy in Bonn. Made from carbon-fiber reinforced plastic that is 20 times less sensitive to thermal change than polished metal panels, this dish permits the telescope to work at sub-millimeter wavelengths as short as 0.350 mm. The largest non-steerable radio telescope is the 300-m dish at Arecibo, Puerto Rico and the first very large steerable dish was the 76.3 m (250 ft) Lovell Telescope at Jodrell Bank, England built in 1957 and famous for detecting Sputnik 1.

13.2.1 Receivers

Typically, a wave received at the focus of a radio telescope enters a “feed” which may be a direct dipole or a "corrugated horn" which matches a waveguide and feeds a resonant cavity, which in turn defines a frequency interval or bandpass near the frequency of the wave. The horn suppresses side bands and couples the radiation onto the much smaller detection system. At the base of the horn is a dipole or a pair of crossed dipoles (to detect both polarizations). Enormous amplification of the detected signal is needed (10^{14}-10^{19}), which therefore implies a cascade of amplifiers, but the initial one, the “pre-amplifier” is critical. An example of a horn is shown in Fig. 13.3 and the typical arrangement of components (for cm wavelengths) is shown in Fig. 13. 4.

The simplest radiometer measures the average total power received over a well-defined radio frequency bandwidth $\Delta\nu$ and over a time interval τ. Just as for optical and infrared wavelengths, the weak astronomical source is measured against a background of many other radio signals such as the cosmic microwave background, the atmosphere and the noise in the receiver itself. Power is usually expressed as a temperature and the total system power is T_{sys}. The total noise in the measurement is given by the practical form of the radiometer equation:

$$\sigma_{\mathrm{T}} = \mathrm{T}_{\mathrm{sys}}[\ (1/\Delta\nu_{\mathrm{RF}}\ \tau) + (\Delta G/G)^2\ ]^{1/2} \qquad (13.10)$$

where ΔG represents possible fluctuations in gain. If those fluctuations are negligible then the equation simplifies to its ideal form ($\mathrm{T}_{\mathrm{sys}}/\sqrt{(\Delta\nu_{\mathrm{RF}}\tau)}$). In a manner similar to chopping in the infrared, one way to minimize fluctuations in receiver gain and atmospheric emission is to perform differential measurements by switching rapidly between two adjacent feeds as first suggested by Robert Dicke (1916-1997) in the 1940s. The main drawback of Dicke switching is that the measured noise is doubled to 2 $T_{sys}/\sqrt{(\Delta\nu_{RF}\tau)}$ as a result of the difference measurement.

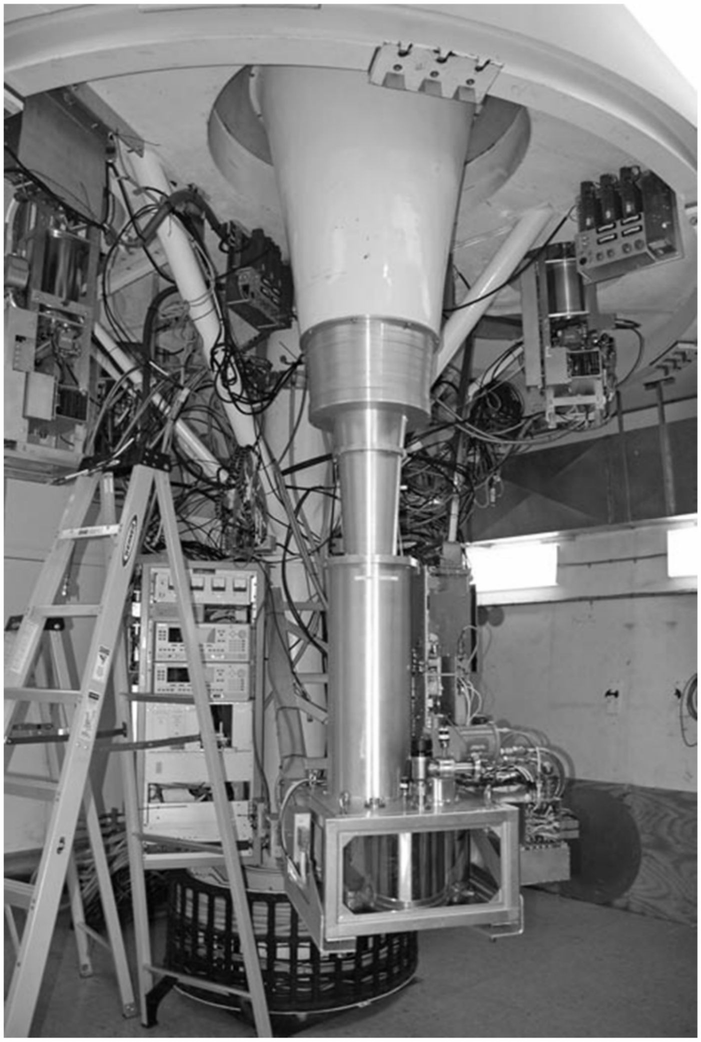


Fig. 13.3 A feed horn at the Gregorian focus of the GBT. Credit: NRAO/AUI.

Nearly all practical radiometers are more complex. As is well-known from elementary physics, if two signals of different but similar frequencies are added they produce a signal at the "beat" frequency, which is the difference between the two original frequencies and is therefore a much lower frequency. While the resulting "mixed" signal contains frequencies only from the original two signals, its amplitude is modulated at the difference, or beat frequency. Heterodyne receivers measure this amplitude. Effectively, the mixer device multiplies the RF signal by a sine wave of frequency ν_{LO} generated by a Local Oscillator (LO). To illustrate the effect consider the product of two sine waves where *t* is time:

$$2 \sin(2\pi\, \nu_{LO} t) \text{ x } \sin(2\pi\, \nu_{RF} t) = \cos[2\pi\,(\nu_{LO} - \nu_{RF})t] - \cos[2\pi\,(\nu_{LO} + \nu_{RF})t] \qquad (13.11)$$

The difference frequency ν_{LO} - ν_{RF} is called the intermediate frequency (IF). At microwave frequencies the mixer comes first because low-noise amplifiers are difficult to design and the LO signal is fed directly into the horn (waveguide) along with the antenna signal.

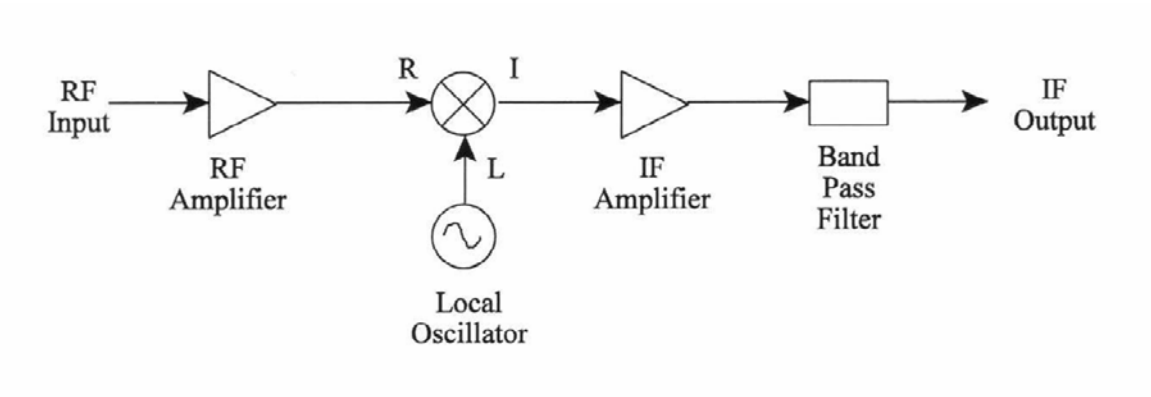


Fig. 13.4 Basic layout of a heterodyne radio detection system showing antenna, mixer, local oscillator, IF amplifier and detector.

Otherwise, as shown in Fig. 13.4, the mixing is done after the signal becomes an electrical current. The advantage of heterodyne receivers lies in the "down conversion" of the frequency from the large (GHz or MHz) radio frequencies to the much lower intermediate frequency (IF) range (kHz) where conventional electronics can be used. In addition, control over the RF range being covered depends only on tuning the oscillator and therefore back-end devices following the un-tuned IF amplifier can operate over fixed frequency ranges. An ambiguity exists in the sign of the difference signal. It is not possible to tell whether the true frequency was larger or smaller than the local oscillator frequency. Because the IF signal can arise from a combination of two possible inputs it is called a "double sideband" or DSB signal. This is a serious problem for observations of spectral lines at radio frequencies and therefore it is usually desirable to operate in a "single sideband" or SSB configuration if possible by using a narrow bandpass rejection filter in front of the receiver or by tuning the mixer.

Again, even if there are no sources in the beam, a non-zero current will be measured. This signal arises from various causes such as residual thermal radio emission from the atmosphere, telescope or waveguide; thermal radio emission from the ground detected in the side lobe pattern of the antenna; thermal noise in the detector itself. The noise is equivalent to a small amount of power and can be given an effective noise temperature as mentioned earlier which represents the background power against which the signal is to be detected. If the receiver operates in the double sideband (DSB) mode then the bandwidth for the noise measurement is generally $\Delta\nu = 2\Delta\nu_{IF}$. In the best case the limit is imposed by the detector itself and cannot be lower than the physical temperature of the device. Noise temperatures in the range 10 - 200 K are typical, with 50 K being typical at wavelengths of 6 to 21 cm. Figure 13.5 shows a 4-beam cryogenic (20 K) receiver and feed horns for the GBT Q-band receiver (40-52 GHz).

14

Future developments

In this concluding chapter we briefly review some new technologies, discuss the current trends in astronomical detectors and instrumentation and summarize plans for some future new facilities. Despite all the advances since the invention of the CCD, many important questions about the universe remain to be answered and the experiments needed to study these questions will spur the development of new technology. Larger telescopes, better detectors and more efficient instruments are a few of the driving factors.

14.1 SCIENTIFIC CHALLENGES

Among the many intriguing puzzles in astronomy still unsolved are several over-arching topics that are typically repeated in decadal reviews and proposals for future new facilities. This list of key topics includes the following:

- The discovery of nearby Earth-like worlds, the statistics of planetary systems and the evidence for biological activity elsewhere
- The detection and tracking of near-Earth asteroids
- A deeper understanding of the origin and formation of stars and planetary systems
- The Black Hole at the center of the Milky Way and tests of General Relativity
- The origin and evolution of the supermassive Black Holes in quasars
- The origin of cosmic gamma-ray bursts
- The nature and distribution of dark matter in the cosmos
- The nature and distribution of dark energy in the cosmos
- Detection of the first starlight in the early universe; formation of the first galaxies

To be sure, some of these problems can be tackled with lengthy programs using existing facilities or upgraded ones, but others will require new and sometimes radically different approaches. Electronic imaging across the electromagnetic spectrum, with even larger telescopes and with finer detail than today, is likely to be at the heart of most of these new developments.

14.2 NEW GROUND-BASED TELESCOPES

It is now about four centuries since Galileo Galilei turned his tiny telescope to the sky and made his first sketch of the Moon's surface. As shown at the beginning and throughout this book, the rate of progress in telescope development has been steady until recently when the pace seems to have picked up due to rapid advances in technology. In 1998 when the previous edition of this book appeared the only optical telescopes over 6 m in diameter were the twin 10-m segmented-mirror telescopes of the W. M. Keck Observatory operated by the California Association for Research in Astronomy. A decade later, in 2008, the number of telescopes in the Very Large Telescope (VLT) category had grown to 15 and plans to develop "Extremely" Large Telescopes (ELTs) with diameters in the range 20-40 m had surfaced. There has even been one proposal for a 100-m telescope known as OWL (for Over-Whelmingly Large telescope)! In addition, because large telescopes tend to be used for relatively narrow fields of view, several radically new ideas for large-aperture but wide-field telescopes for surveys have been proposed, including the Large Synoptic Survey Telescope (LSST) and Pan-STARRS. All of these new facilities would serve ground-based visible and infrared astronomy, but new ground-based facilities for the radio regime also got under way during this period. ALMA, the Atacama Large Millimeter Array, will dramatically change the field of sub-mm and mm astronomy, and radio astronomy at cm wavelengths will be boosted by projects like the Extended VLA (E-VLA) and the Square Kilometer Array (SKA). Below is a brief summary of each project as of 2008. More information and status reports can be found at the web sites provided.

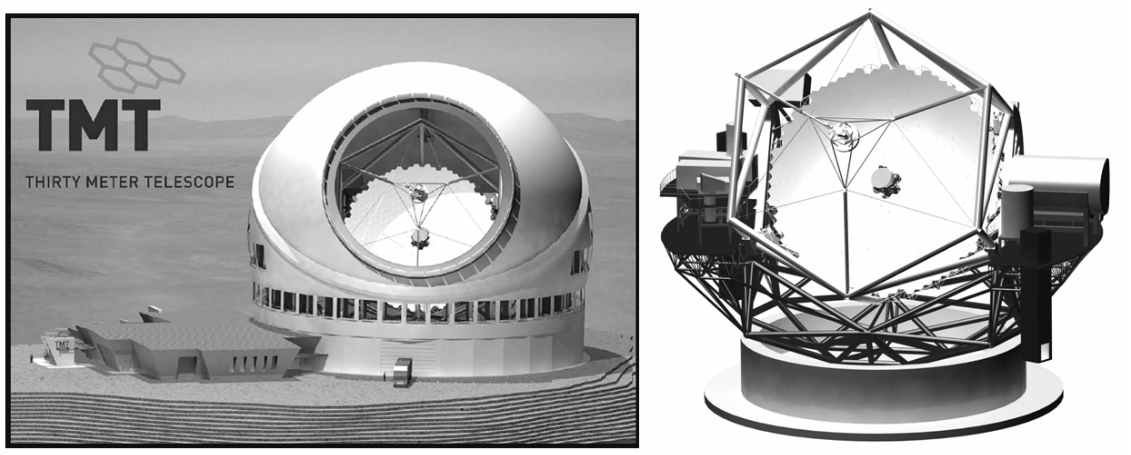


Figure 14.1 Left: Artist's concept of the Thirty Meter Telescope (TMT). Note the unusual dome and the size scale as judged by the human figures and vehicles. Right: Close-up view of the telescope and structure. Credit: TMT Corporation.

GMT: The Giant Magellan Telescope

The GMT is a multi-mirror telescope that employs today's largest stiff monolith mirrors as segments. Six off-axis 8.4 m segments surround a central on-axis segment, forming a single optical surface with a collecting area equivalent to a filled aperture 21.4 m in diameter, and the resolving power of a 24.5-m (80 ft) primary mirror. The focal length of this primary mirror combination is 18 m and the focal ratio is *f*/0.7. Each mirror will be made using the same honeycomb borosilicate mirrors that have already been deployed successfully on the Magellan telescopes in Chile, and the Multiple Mirror Telescope (MMT) and Large Binocular Telescope (LBT) in Arizona. The secondary mirror is composed of seven thin adaptive shells, with each segment mapping to a single primary mirror segment. By making the secondary an adaptive optics component the telescope should get diffraction-limited performance over modest fields of view and with the addition of ground-layer adaptive optics at the focus, the corrected field will be 10-20 minutes of arc. The final focal ratio at the straight Gregorian focus will be *f*/8.4 with an image scale of 1.0 "/mm.

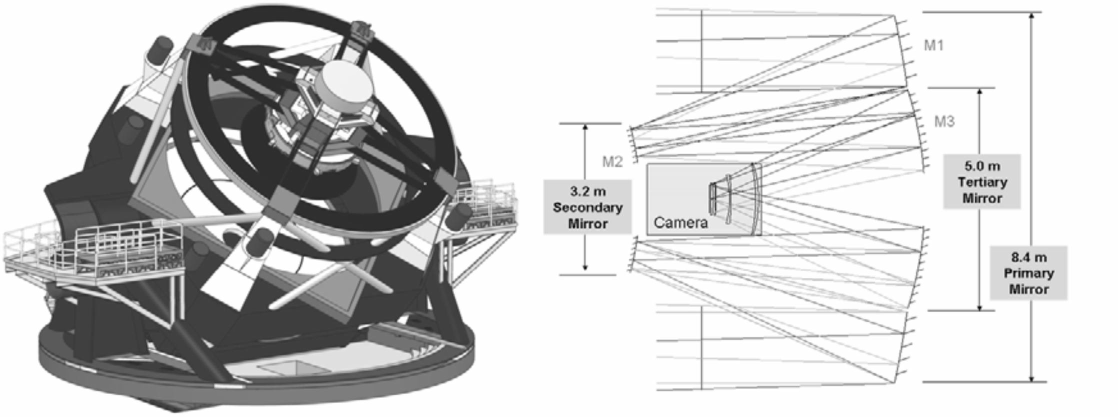


Figure 14.2 Mechanical design and light-path for the Large Synoptic Survey Telescope (LSST). Credit: Tony Tyson.

TMT: The Thirty Meter Telescope

Based on lessons learned from the Keck telescopes, the TMT (Fig. 14.1) is a wide-field, Ritchey-Chrétien telescope with a 30 m (98 ft) diameter *f*/1 hyperboloidal primary mirror composed of 492 hexagonal segments, a fully active 3.1-m secondary mirror and an articulated tertiary mirror. The optical beam of the telescope feeds a suite of adaptive optics (AO) systems and science instruments mounted on very large Nasmyth platforms surrounding the telescope azimuth structure. These platforms will be large enough to support at least eight different AO/instrument combinations covering a broad range of spatial and spectral resolution. The final focal ratio at the Nasmyth foci is *f*/15 with an image scale of 0.46"/mm. At the time of writing TMT is in a very advanced design phase and has received partial funding from the Gordon and Betty Moore Foundation to begin the construction phase following site selection.

E-ELT: The European Extremely Large Telescope

This telescope is a 5-mirror concept based on a 42 m (140 ft) *f*/1 segmented primary mirror composed of 906 segments each 1.45 m wide, and a secondary mirror of diameter 6 m. A tertiary mirror 4.2 m in diameter will relay the light to the adaptive optics system which is composed of two large mirrors, a 2.5 m mirror supported by 5000 or more actuators able to distort its shape a thousand times per second, and one 2.7 m mirror that gives accurate image stabilization.

14.7 CONCLUSION

The years since the invention of the CCD in 1970 have been a remarkable period of growth and development for astronomy. This time-period matches my own career quite well as I graduated with my first degree in 1971, and it is sobering to look back at all that has transpired in astronomy since then. New technologies have led to new discoveries and the promise of even greater discoveries has stimulated the drive for even better instrumentation. As expected, the access to space has meant that gamma-ray, X-ray, and ultraviolet astronomy have all blossomed, and CCD-like imaging is now available even at the shortest wavelengths. Infrared astronomy underwent a tremendous surge with the advent of InSb, HgCdTe and Si:As array detectors in the mid-eighties, and yet another boost with the introduction of adaptive optics techniques for the elimination of atmospheric turbulence. New telescopes and new detectors for far-infrared to millimeter wavelengths have finally opened up that part of the spectrum to CCD-like imaging too. Conventional ground-based optical astronomy, far from becoming entrenched, has continued to expand with new telescopes and better CCDs. Even the very method of building optical reflecting telescopes underwent a radical change during this period. Adaptive optics and laser guide stars have enabled the suite of very large telescopes now in operation to achieve their diffraction-limited performance, at least in the near-infrared. Of course, the amount of data being obtained, studied and archived is enormous, but Moore's Law has enticed computer manufacturers to keep pace, and hopefully concepts like the Virtual Observatory will be fully realized.

Throughout this book I have tried to show that the underlying reason for these advances has been a willingness of astronomers to grasp the very latest technologies and push them to their limits. From Galileo Galilei's eyes (1609) to Gigapixel CCD cameras (2009), from sketches of the uneven surface of the Moon to the detection of the cosmic microwave background and ripples in the fabric of spacetime, we have come a long way. Our understanding of the Universe has grown in leaps and bounds, hand-in-hand with developments in technology, and the cycle continues. I hope it will always be so, and I think Galileo would be pleased.