1

Light and Colour

- **What is colour?**
- **Why do hot objects become red or white hot?**
- **How do e-books produce 'printed' words?**

1.1 Colour and Light

Colour is defined as the subjective appearance of light as detected by the eye. It is necessary, therefore, to look initially at how light is regarded. In fact, light has been a puzzle from earliest times and remains so today. In elementary optics, light can usefully be considered to consist of light *rays*. These can be thought of as extremely fine beams that travel in straight lines from the light source and thence, ultimately, to the eye. The majority of optical instruments can be constructed within the framework of this idea. However, the ray concept breaks down when the behaviour of light is critically tested, and the performance of optical instruments, as distinct from their construction, cannot be explained in terms of light rays. Moreover, colour is not conveniently defined in this way. For this, more complex ideas are needed.

The first testable theory of the nature of light was put forward by Newton (in 1704) in his book *Optics*, in which it was suggested that light was composed of small particles or 'corpuscles'. This idea was supported on philosophical grounds by Descartes. Huygens, a contemporary, thought that light was wavelike, a point of view also supported by Hooke. Young provided strong evidence for the wave theory of light by demonstrating the interference of light beams (1803). Shortly afterwards, Fresnell and Arago explained the polarisation of light in terms of transverse light waves. However, none of these explanations was able to refute the particle hypothesis completely. Nevertheless, the wave versus particle theories differed in one fundamental aspect that could be tested. When light enters water it is refracted (Chapter 2). In terms of corpuscles, this implied a speeding up of the light in water relative to air. The wave theory demanded that the light should move more slowly in water than in air. The experiments were complicated by the enormous speed of light, which was known to be about

Colour and the Optical Properties of Materials: An Exploration of the Relationship Between Light, the Optical Properties of Materials and Colour, Second Edition
Richard J. D. Tilley
 ISBN: 978-0-470-74696-7

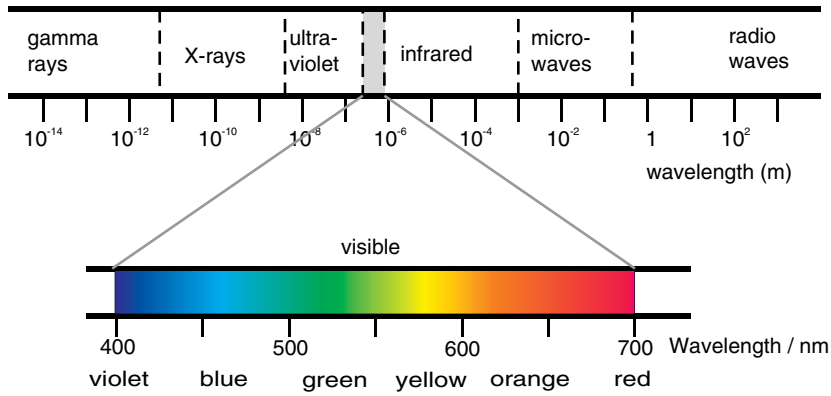


Figure 1.1 ***The electromagnetic spectrum. Historically, different regions have been given different names. The boundaries between each region are not sharply defined but grade into one another. The visible spectrum occupies only a small part of the total spectrum***

$3 \times 10^8\ \mathrm{m\,s^{-1}}$, and it was not until April 1850 that Foucault first proved that light moved slower in water than in air, and seemingly killed the corpuscular theory then and there. Confirmation of the result by Fizeau a few months later removed all doubt.

Over the years the wave theory became entrenched and was strengthened by the theoretical work of physicists such as Fresnel, who first explained interference and diffraction (Chapter 6) using wave theory. Polarisation (Chapter 4) is similarly explained on the assumption that light is a wave. The wave theory of light undoubtedly reached its peak when Maxwell developed his theory of electromagnetic radiation and showed that light was only a small part of an *electromagnetic spectrum*. Light was then imagined as an electromagnetic wave (Figure 1.1). Maxwell's theory was confirmed experimentally by Hertz, whose experiments led directly to radio.

The problem for the wave theory was that waves had to exist in something, and the 'something' was hard to pin down. It became called the *luminiferous aether* and had the remarkable properties of pervading all space, being of very small (or even zero) density and having extremely high rigidity. Attempts to measure the velocity of the Earth relative to the luminiferous aether, the so-called aether drift, by Michelson and Morley, before the end of the nineteenth century, proved negative. The difficulty was removed by Einstein's theory of relativity, and for a time it appeared that a theory of light as electromagnetic waves would finally explain all optical phenomena.

This proved a false hope, and the corpuscular theory of light was revived early in the twentieth century, principally by Einstein. Since 1895, it had been observed that when ultraviolet light was used to illuminate the surfaces of certain metals, negative particles, later identified as electrons, were emitted. The details of the experimental results were completely at odds with the wave theory. The electrons, called *photoelectrons*, were only observed if the frequency of the radiation exceeded a certain minimum value, which varied from one material to another. The kinetic energy of the photoelectrons was linearly proportional to the frequency of the illumination. The number of photoelectrons emitted increased as the intensity[1] of the light increased, but their energy remained constant for any particular light source. Very dim illumination still produced small numbers of photoelectrons with the appropriate energy.

[1] The imprecise expression 'intensity' has largely been replaced in the optical literature by well-defined terms such as irradiance (Appendix 1.1). The term intensity is retained here (in a qualitative way to designate the amount of light) because of the historical context.

The explanation of this 'photoelectric effect' by Einstein in 1905 was based upon the idea that light behaved as small particles, now called *photons*. Each photon delivered the same amount of energy. If this was sufficiently large, then the electron could be ejected from the surface. The energy of each photon, E, was proportional to the frequency of the illumination, so that the photoelectron could be ejected when the frequency passed a certain threshold, but not before that point was reached. Thereafter, increasing the frequency of the illumination allowed the excess energy to be displayed as an increase in kinetic energy. The kinetic energy of the photoelectrons ejected from a metal under this hail of photons could then be written as:

$$\tfrac{1}{2}mv^2 = E - \phi$$

where ϕ is known as the *work function* of the metal and is simply the energy required to liberate the electron from the metal surface. The intensity of the light simply indicated the number of photons arriving at the surface, so that the number of photoelectrons emitted is a function of irradiance, but the energy of these electrons is a function of the frequency of the radiation. Einstein thus rescued the wave theory from the dilemma of the luminiferous aether and then seemingly wrecked the self-same theory via his explanation of the photoelectric effect.

At present, all experiments show that light and its interaction with matter (i.e. atoms) is best described in terms of photons. At its simplest level, the statistical behaviour of a large number of photons is then represented very well by an electromagnetic wave. That is to say, photons are the components of a light beam, whilst waves are a mathematical description of a beam of light.

In this book, explanations are given in terms of the simplest approach that is in accord with the observations. For large-scale phenomena, such as the operation of a magnifying glass, it is adequate to use the idea of a ray of light. When objects having dimensions of the order of hundreds of nanometres are encountered it is necessary to consider light to be a wave. Atomic processes require a photon approach. It needs to be stressed that these are not different fundamentally. All are contained within the theory of optics available today, generally described as quantum optics or quantum electrodynamics.

No matter how it is described, light has no colour as such. Light simply leaves the generating source, possibly interacts with matter in the course of passage and then enters the eye. Colour, or more accurately the perception of colour, is the result of an eye–brain combination that serves to discriminate between light of different wavelengths or energies. In the following chapters, the production of light and its interaction with matter is discussed from the point of view of colour – that of the original light source, and how this is modified by interaction with matter to generate new colours.

1.2 Colour and Energy

Colour is generated by interactions of light and matter – atoms and molecules, or, more strictly, the electrons associated with these. If light is considered as an electromagnetic wave, then the energy density of the wave, which is the energy per unit volume of the space through which the light wave travels, is given by:

$$E = \varepsilon_0(\mathcal{E}_0)^2$$

where ε_0 is the vacuum permittivity and $\mathcal{E}_0$ is the amplitude of the electric component of the wave. Classical optics, the interaction of light with a transparent solid, in the main, is concerned with scattering of the light. This leads to the phenomena of reflection, refraction and so on. In these processes, colour is produced by interaction between various light waves, and energy exchange considerations hardly matter. These aspects of colour formation are covered in Chapters 2–6.

When light is absorbed by or emitted from a material, say a gemstone such as ruby, energy changes are paramount. In this case, light is best regarded as a stream of photons; the energy of each photon being defined as:

$$E = h\nu = \frac{hc}{\lambda} \tag{1.1}$$

where ν is the frequency of the equivalent light wave, λ is the wavelength of the equivalent light wave, h is Planck's constant and c is the velocity of light in vacuum.

The absorption of light by isolated atoms or molecules involves a change in energy of the electrons surrounding the atomic nuclei. These occupy a series of atomic or molecular orbitals, each of which can be assigned a precise energy. The energies of the orbitals form a sort of ladder (with variable rung spacing) from low to high, each separated from the next by an energy gap. Electrons are fed into the orbitals from lowest to highest energy until all of the electrons have been allocated, leaving the extra outer electron orbitals empty. The total energy of all the electrons in the atom at modest temperatures is represented by an *energy level* called the *ground state*. The absorption of light will cause an electron to move from the low-energy ground state E_0 to an empty orbital at a higher energy. The new energy situation is represented by an energy level at energy E_1 (Figure 1.2a). (These energy levels and how they are enumerated will be described in detail in later chapters.) The relationship between the energy change ΔE and the frequency ν or the wavelength λ of the light absorbed is:

$$E_1 - E_0 = \Delta E = h\nu = \frac{hc}{\lambda} \tag{1.2}$$

where h is the Planck constant and c is the speed of light. When energy is lost from an isolated atom it moves from the excited state back to the ground state. The simplest case is when the species passes directly from E_1 to E_0 (Figure 1.2a), with an energy output given by:

$$E_1 - E_0 = \Delta E = h\nu = \frac{hc}{\lambda}$$

identical to that of the absorbed radiation. However, the release of energy often takes place by more complex mechanisms that will be explored in later chapters.

In both cases, if the frequency associated with the energy change ΔE lies in the band that is registered by the eye, then colour is perceived.

When atoms unite to form a solid (or a liquid) the precise energies of the orbitals are broadened out into continuous bands of energy. The main energy landscape in a solid is the band structure – which is the geometrical form of the energy bands throughout the matrix. In a solid, electrons are allocated to the energy bands, from the lowest energy up, until all have been allocated. The energy bands of highest energy are then empty, similar to the orbitals with highest energy in an atom. In the simplest depictions, the highest filled energy band (the *conduction band*) is separated from the lowest empty energy band (the *valence band*) by a constant band gap (Figure 1.2b). In real structures, the band architecture is more complex. Light absorption, emission and colour generation in a solid cannot be discussed without consideration of the role of the band structure.

In this case, the energy difference ΔE which corresponds to colour registration might correspond to the promotion of an electron from a full conduction band to an empty valence band. However, impurities and defects can introduce further energy levels into the energy gap between the conduction and valence bands. In these cases, energy transitions between these levels or between them and the energy bands of the solid may then be of an appropriate energy to act as important sources of colour (Figure 1.2b). Examples of these instances are presented in Chapters 7–10.

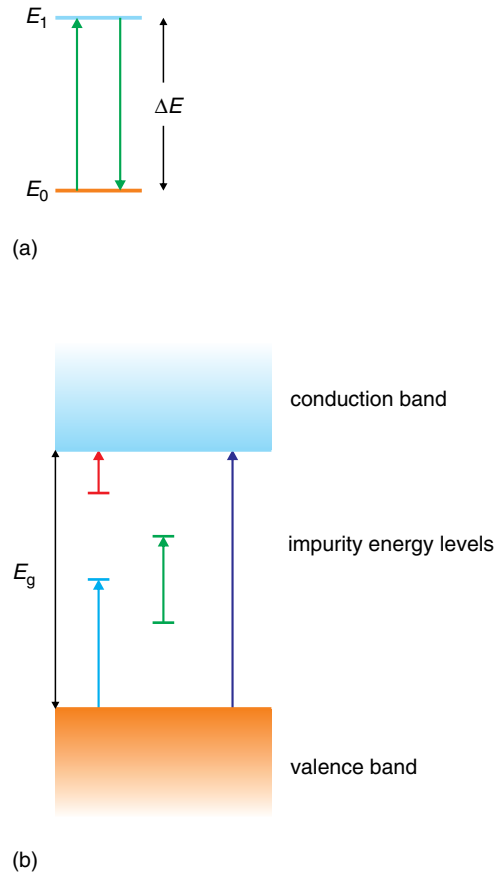


Figure 1.2 ***Energy transitions leading to colour production, shown as arrows: (a) transitions between energy levels in isolated atoms or molecules; (b) transitions between impurity levels and energy bands in a solid. Note that each single energy level shown may actually be composed of several closely spaced energy levels in real systems. E_g is the magnitude of the energy gap between the valence and conduction band***

1.3 Light Waves

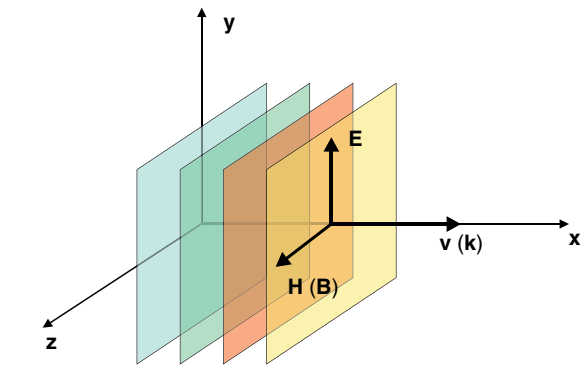
In terms of the wave theory, light waves comprise a small segment of the electromagnetic spectrum (Figure 1.1). Any part of the electromagnetic spectrum is regarded as a wave of wavelength λ with an electrical and magnetic component, each described by a vector and moving with a velocity, the 'speed of light', in a vacuum. The electric field vector[2] **E** is perpendicular to the magnetic vector, described in terms of the magnetic induction **B** or the magnetic field **H**, and they are in phase, so that a peak in the electric field component coincides with a peak in the magnetic field component. Moreover, these vectors lie in a plane perpendicular to the direction in which the wave is moving, described by the *velocity vector* **v**, or the *propagation vector* or *wave vector* **k**. Thus, **E** and **B** both lie perpendicular to the *direction of propagation*, so that light is regarded as a *transverse electromagnetic* (TEM) wave. The wave is a *progressive wave*, a *travelling wave* or a *propagating wave*, all

[2] Vectors are given in bold throughout this book.

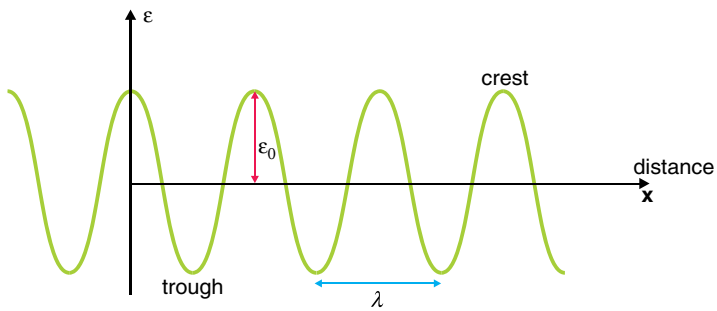
terms being used more or less interchangeably. The electric field vector, the magnetic field vector and the velocity vector can be represented by the three (right-handed) Cartesian axes (Figure 1.3a).

As far as the topics in this book are concerned, the electric field **E** can usually be considered in isolation. In this case, a one-dimensional continuous electromagnetic wave moving in the $+x$ direction can be conveniently depicted by the equation:

$$\mathcal{E}_y = \mathcal{E}_0 \cos[(2\pi/\lambda)(x - vt)] \tag{1.3}$$



(a)



(b)

Figure 1.3 ***Light waves. (a) Light can be thought of as a TEM wave. The electric*** **(E)** ***and magnetic (*** **H** ***or*** **B***) vectors lie perpendicular to each other and to the vector representing the direction of travel of the wave (*** **v** ***or*** **k***). The shaded planes represent the positions of peaks in the electric and magnetic fields. (b) Part of a light wave travelling along x. The curve represents the magnitude $\mathcal{E}$ of the electric field vector as a function of position. The distance between the crests or troughs is the wavelength λ. Any point on the wave moves with a speed v.*** **If the electric field vector remains in the plane of the paper, as drawn, the light is linearly polarised. If the orientation of the electric field with respect to the plane of the page varies at random so that the curve continually adopts differing angles with the plane of the paper, the light is unpolarised**

Here, $\mathcal{E}_y$ is the magnitude of the electric field vector at position x and time t and v is the wave speed (or velocity[3]). The term $\mathcal{E}_0$ is the *amplitude* of the wave (the maximum value that the electric field vector takes) and is a constant. The speed v at which any point on the wave, say a peak or a trough, travels is called the *phase speed* or *phase velocity*. The velocity of an electromagnetic wave in vacuum, denoted by the speed of light c, is an important physical constant.

Taking t as fixed gives a snapshot of the wave at a single instant (Figure 1.3b). The spatial period of the wave, which is the distance over which the wave subsequently repeats itself, is called the *wavelength* λ. The peaks in the wave are referred to as *crests* and the valleys as *troughs*. The term in square brackets, $[(2\pi/\lambda)(x-vt)]$, i.e. the argument of the cosine function, is called the *phase* of the wave, represented by ϕ. The phase of the wave is usually quoted in radians, in the form $(m\pi/n)$, i.e. $3\pi/4$. Clearly, the phase of the wave varies along its length and changes by 2π in one wavelength. The phase of a light wave cannot be determined. However, the phase difference between corresponding points on two different waves, say two equivalent crests, can be measured with considerable precision.

Taking x as fixed will show that the magnitude of the electric field vector $\mathcal{E}_y$ will oscillate up and down between values of $\pm\mathcal{E}_0$. The temporal period of the wave τ, which is the time over which the wave subsequently repeats itself, is more usually encountered as the reciprocal $1/\tau$ and is equal to the temporal frequency ν, which is the number of waves that pass a point per second.

The speed of the wave v is related to the frequency ν by:

$$v = \lambda\nu$$

(or in a vacuum by $c = \lambda\nu$).

A beam of light is said to be *monochromatic* when it is comprised of a very narrow range of wavelengths and it is *coherent* when all of the waves which make up the beam are completely *in phase*; that is, the crests and troughs of all the waves are in step. The way in which the electric field vector is constrained describes the *polarisation* of the wave. If the electric field vector remains in one plane, then the light is said to be *linearly* (or *plane*) *polarised*. In general, the polarisation of the light wave must be considered when describing optical phenomena.

Normal light, such as that from the sun, say, is not emitted in a continuous stream, but in short bursts lasting about 10^{-8} s. Within each burst all of the light waves are in phase and linearly polarised. However, both the phase and polarisation change from burst to burst in a random fashion, so that the phase and polarisation of each burst are unrelated to those in the preceding burst. This means that the phase and the polarisation of a light wave fluctuate continuously and at random within a fraction of a second. Normal light is thus described as being *incoherent* and *unpolarised*. Because of this, the interaction of daylight with objects can be interpreted (at least as a good approximation) without considering polarisation. Light from lasers (Section 1.9) is, by and large, coherent and polarised, and these aspects cannot usually be ignored.

1.4 Interference

One of the advantages of the wave description of light is that the interactions between two beams are easily explained. If two light waves occupy the same region of space at the same time then they add together, or *interfere*, to form a product wave. This idea, called the *principle of superposition*, was stated by Young some

[3] Strictly speaking we are discussing wave speed, which is a scalar quantity. Velocity is a vector quantity. However, it makes things simpler to brush over this distinction in the present case.

two centuries ago, in 1802. If two identical waves are exactly in step then they will add to produce a resultant wave with twice the amplitude (Figure 1.4a–c) by the process of *constructive* interference. If the two waves are out of step, then the resultant amplitude will be less, due to *destructive* interference. If the waves are sufficiently out of step that the crests of one correspond with the troughs of the other, then the resulting amplitude will be zero (Figure 1.4d–f).

Interference can occur between light waves with different relative frequencies, amplitudes and phases. However, for this to be *observed* the phase difference between the beams must remain constant. That is, the waves must be coherent. Many of the difficulties inherent in observing interference effects using normal light stem from the incoherent nature of the wave trains used, and efforts must be made to ensure that the incoherence does not destroy any visible interference patterns that may be generated. The use of laser light makes the observation of interference much simpler.

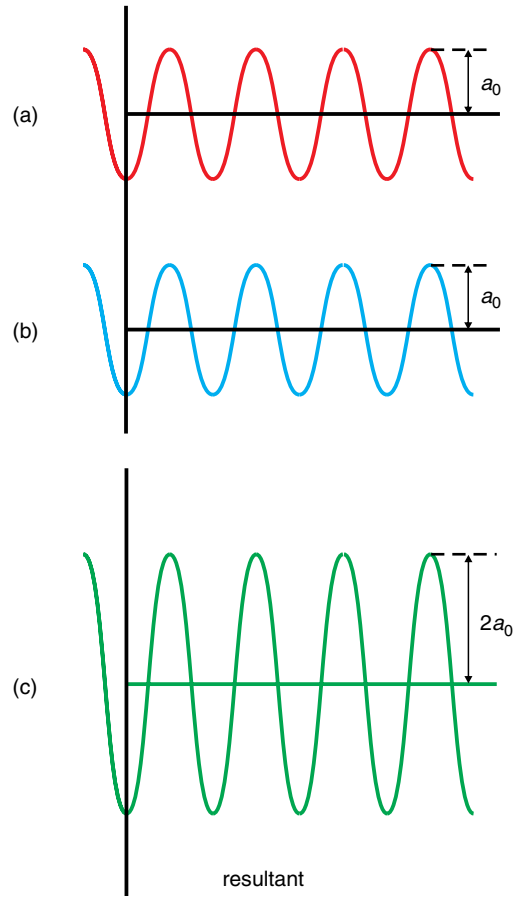


Figure 1.4 ***Interference of light waves. (a)–(c) The addition of two waves in phase, (a), (b), will produce a wave of twice the amplitude of the original wave, (c). (d)–(f) The addition of two waves out of phase by λ/2, (d), (e), will produce a wave with zero amplitude, (f)***

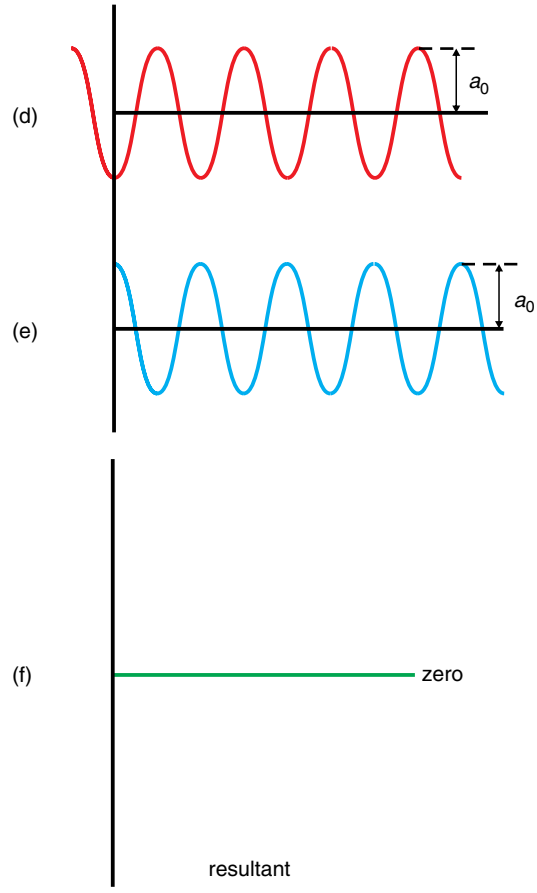


Figure 1.4 ***(Continued)***

The effects of interference can be assessed analytically using algebraic methods. An intuitive feeling for the phenomenon is best gained by adding waves represented by formulae such as Equation 1.3 using a computer and displaying the results graphically.

1.5 Light Waves and Colour

Our eyes can detect only a small part of the whole electromagnetic spectrum, called the *visible spectrum* (Figure 1.1). The amount of light that the eye records in any situation, which can *loosely* be called the brightness or intensity of the light, is not the amplitude of the wave but is the *irradiance I*, which is proportional to the square of the amplitude:

$$I = K(\mathcal{E}_0)^2$$

Table 1.1 ***The visible spectrum***

Colour	λ/nm	$10^{-14}\nu$/Hz	$10^{-15}\omega$/rad s^{-1}	$10^{19} \times$ Energy/J	Energy/eV
Infrared	750	4.00	2.51	2.65	1.65
Deep red	700	4.28	2.69	2.84	1.77
Orange–red	650	4.61	2.90	3.06	1.91
Orange	600	5.00	3.14	3.31	2.07
Yellow	580	5.17	3.25	3.43	2.14
Yellow–green	550	5.45	3.42	3.61	2.25
Green	525	5.71	3.59	3.78	2.36
Blue–green	500	6.00	3.77	3.98	2.48
Blue	450	6.66	4.19	4.42	2.75
Violet	400	7.50	4.71	4.97	3.10
Ultraviolet	350	8.57	5.38	5.68	3.54

where the value of the constant of proportionality K depends upon the properties of the medium containing the wave. (See Appendix A1.1 for information on units.)

The extent of the visible spectrum is defined in terms of the wavelength or frequency of the light waves involved. *Perception* of the different wavelengths is called *colour*. The precise measurement of colour involves a determination of the energy present at each wavelength in the light using a spectrometer.

The wavelength range that an eye can perceive varies from individual to individual. In general, it is assumed that the shortest wavelength of light that an average person can detect corresponds to the colour violet, with a wavelength near to 400 nm. Similarly, the longest wavelength of light registered by an average observer corresponds to the colour red, with a wavelength close to 700 nm. Between these two limits the other wavelengths of the spectrum are associated with the colour sequence from red to orange, green, blue, indigo and finally to violet (Figure 1.1 and Table 1.1). The divisions between these colours are, of course, artificial, and each colour blends into its neighbours. (Note that these colours are simply approximate labels for the wavelength. The perceived colour of an object is a function of a number of factors (Section 1.10).) It is known that the sensitivity of the eyes of animals is different than those of humans. Many insects, for example, can detect wavelengths shorter than humans but do not see so far into the red.

Radiation with wavelengths shorter than violet falls in the *ultraviolet* region of the spectrum. Ultraviolet A (UVA) is closest to the violet region and is taken to have a wavelength range of 400–320 nm. This radiation is largely responsible for suntan. Ultraviolet B (UVB), with an approximate wavelength range of 320–280 nm, is more damaging and causes sunburn. Ultraviolet radiation with shorter wavelengths is called the *far ultraviolet*, (280–200 nm) and *vacuum ultraviolet* (below 200 nm). UVB and shorter wavelengths are able to damage biological cells severely, and excessive exposure leads to the occurrence of skin diseases. Radiation with wavelengths longer than red is referred to as *infrared* radiation. Although not visible, the longer wavelengths of infrared radiation, called *thermal infrared*, are detectable as the feeling of warmth on the skin.

1.6 Black-Body Radiation and Incandescence

There are many ways in which light can be generated, but the action normally tales place at an atomic level. Individual atoms (or molecules) lose energy, which is given out as radiation. These processes generally need to be discussed in terms of photons rather than waves. In this section, just one example is given, the generation of light by a hot body. This was the first light-generating process to be understood at a fundamental level, and led directly to the photon concept as well as to an appreciation of our idea of the make-up of white light.

Incandescence is the emission of light by a hot body. The sun and tungsten-lamp filaments provide commonplace examples, and both are regarded as producing (more or less) white light. The light characterising the upper part of a candle flame also arises from incandescence. In this case, small particles of carbon are heated to high temperatures in the flame and emit light which is perceived as yellow in colour. When light from an incandescent object is spread out according to wavelength by a prism (Chapter 2) the result is a continuous fan of colours following the sequence listed in Table 1.1 and called a *continuous spectrum*. The radiation emitted is both incoherent and unpolarised.

Incandescence comes about in the following way. At absolute zero all atoms and molecules making up the solid are in the lowest possible energy state. As the temperature increases they absorb energy and are promoted to higher energies and, at the same time, atoms and molecules which have already absorbed energy lose some and they fall back to lower energies. (The energy levels involved in this process will be described in more detail in later chapters.) The radiation emitted in this way effectively extends over a continuous range of energies. For a solid a little above room temperature all the wavelengths of the emitted energy lie in the infrared; although the radiation is invisible, it is detectable as a sensation of warmth. At a temperature of about 700 °C the shortest wavelengths emitted creep into the red end of the visible spectrum. The colour of the emitter is seen as red and the object is said to become *red hot*. At higher temperatures the wavelengths of the radiation given out extend increasingly into the visible region and the colour observed changes from red to orange and thence to yellow, as in the example of a candle flame, mentioned above. When the temperature of the emitting object reaches about 2500 °C all visible wavelengths are present and the body is said to be *white hot*. The sun provides a perfect example, and the 'colour' white as applied to light is a combination of energies or wavelengths that spans the visible spectrum with the same composition as that of the radiation from the sun.

These qualitative colour changes can be understood in terms of the radiation emitted by a *black body*. A black body is an idealized object which absorbs and emits all wavelengths perfectly. A reasonable approximation to a source of black-body radiation would be a small pinhole in the wall of a hot furnace. If the irradiance of the radiation issuing from the pinhole is measured as a function of wavelength, a characteristic curve is obtained called a black-body spectrum (Figure 1.5). The shape of the curve is dependent only upon the temperature of the body, and the maximum in the curve moves to shorter wavelengths as the temperature of the black body increases. The curve also mirrors the energy distribution inside the black body when in thermal equilibrium.

The explanation of the form that this curve takes played a significant role in the physics of the twentieth century. Despite many attempts, the form of the black-body spectrum could not be explained by the classical wave theory of electromagnetic radiation. The successful theoretical description of this curve by Planck in 1901, now known as the *Planck law of black-body radiation* or *Planck's radiation law*, signalled the start of the quantum theory. The equations describing the *spectral radiance* of all the radiation components within a black body at equilibrium at temperature T in the frequency range ν to $\nu + \delta\nu$ or the wavelength range λ to $\lambda + \delta\lambda$ are:

$$L_\nu = \frac{2h\nu^3}{c^2[\exp(h\nu/k_\mathrm{B}T)-1]} \qquad \text{units: W m}^{-2}\,\text{sr}^{-1}\,\text{Hz}^{-1} \tag{1.4a}$$

or

$$L_\lambda = \frac{2hc^2}{\lambda^5[\exp(hc/\lambda k_\mathrm{B}T)-1]} \qquad \text{units: W m}^{-3}\,\text{sr}^{-1} \tag{1.4b}$$

In these equations, h is a constant that is now called Planck's constant, c is the speed of light, λ is the wavelength, k_B is Boltzmann's constant and T (K) the temperature of the body. These equations are often seen in

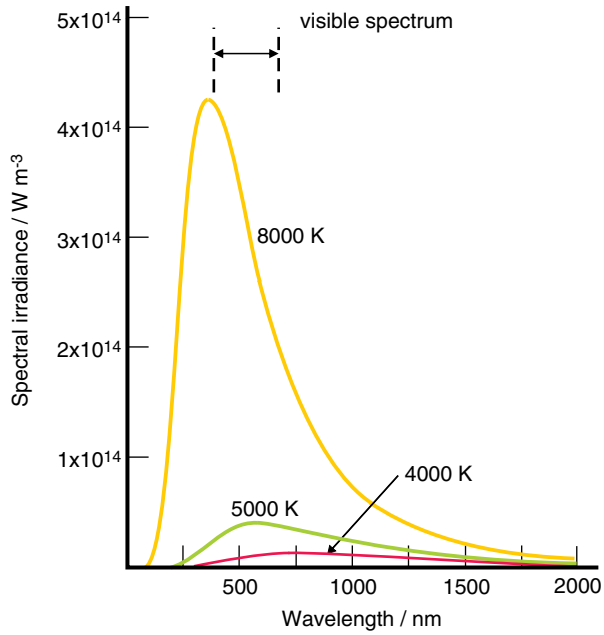


Figure 1.5 ***The radiation emitted from a black body as a function of wavelength. As the temperature of the body is increased, the maximum of the curve both increases and moves towards shorter wavelengths (higher energy). The spectrum emitted by the sun is similar to that for a black body at 6000 K and that from a red-hot object is similar to the curve for a black body at 1000 K***

the form describing the *spectral irradiance* (if the energy falls upon a surface) or the *spectral exitance* (if the energy is observed after leaving a black body via a pinhole not large enough to disturb the thermal equilibrium within), I_ν in the frequency range ν to $\nu + \delta\nu$ or I_λ in the wavelength range λ to $\lambda + \delta\lambda$ as a function of the wavelength λ for a black body at a temperature T:

$$I_\nu = \frac{2\pi h\nu^3}{c^2[\exp(h\nu/k_B T)-1]} \qquad \text{units: W m}^{-2}\,\text{Hz}^{-1} \tag{1.5a}$$

or

$$I_\lambda = \frac{2\pi hc^2}{\lambda^5[\exp(hc/\lambda k_B T)-1]} \qquad \text{units: W m}^{-3} \tag{1.5b}$$

or as the corresponding spectral energy density u_ν in the frequency range ν to $\nu + \delta\nu$ or u_λ in the range λ to $\lambda + \delta\lambda$ as a function of the wavelength λ for a black body at a temperature T:

$$u_\nu = \frac{8\pi h\nu^3}{c^3[\exp(h\nu/k_B T)-1]} \qquad \text{units: J m}^{-3}\,\text{Hz}^{-1} \tag{1.6a}$$

or

$$u_\lambda = \frac{8\pi hc}{\lambda^5[\exp(hc/\lambda k_B T)-1]} \qquad \text{units: J m}^{-4} \tag{1.6b}$$

The revolutionary concept that Planck employed in the derivation of these equations to successfully reproduce the black-body curve was that the energy absorbed or given out by the atoms and molecules (the 'oscillators' in Planck's time) in the black body could not take any value from a continuous spread of energies, but had to be delivered only in packets or *quanta* q_0, $2q_0$, $3q_0$ and so on. The relationship between the energy of a quantum E and the frequency of the radiation ν was given by what has since become one of the most famous equations of science:

$$E = h\nu \tag{1.1}$$

The constant h, Planck's constant, is one of the important fundamental physical constants.

More recently, in the mid-twentieth century, it was realized that the cosmos was filled with some sort of background electromagnetic radiation. The peak of the radiation lies in the microwave part of the electromagnetic spectrum. Naturally, it is invisible to optical instruments and was first mapped using radio telescopes and latterly by satellites. The spectrum of this radiation fits that of a black body; and indeed, this radiation, called the cosmic microwave background radiation, is possibly the most accurately measured black-body radiation curve available. It is interpreted as lending strong support to the 'Big Bang' theory of the origin of the universe.

1.7 The Colour of Incandescent Objects

From the point of view of the colour of incandescent objects, one of the most important attributes of the emission curve is the variation in the position of the maximum as the temperature of the black body increases. (This was derived before the Planck radiation law and represents the final success of classical electromagnetic theory.) The relationship, known as the *Wien displacement law*, is:

$$\lambda_{\max} T = \text{constant}$$

where T (K) is the temperature of the body and the constant has a value of 0.002 898 m K. It can be derived from Equations 1.5a and 1.5b by differentiating with respect to λ and setting the result equal to zero. The colour of an incandescent object is then controlled by the maximum of the black body curve (or an approximation to it), as mentioned below. The second factor of importance is the spread of the spectrum. A cool body will be perceived as initially showing a colour when the peak of the curve is close enough to the visible range that some radiant energy creeps into the low-energy (red) end of the spectrum. As the temperature of the incandescent object increases, the peak moves to higher energies, following the displacement law, and the spread moves further across the visible spectrum, resulting in the colour sequence of dull red, red hot to white hot to blue–white.

The colour of an incandescent object is described by its *colour temperature* if the spectrum resembles that of a black body closely. Most solids behave like black bodies over some range of temperature and wavelength, and stars are a close approximation over the whole of the wavelength range. If the match is approximate, the term used is *correlated colour temperature* and this expression is used for light sources that are not incandescent,

Table 1.2 ***Colour temperature of incandescent sources***

Light source	Correlated colour temperature/K
Mean noon sunlight	5 400
Electronic flash	~7 000
Blue flash bulb	~6 000
Tungsten-filament photographic lamps	~3 400
Tubular triphosphor fluorescent lamp, 36 W	3 000
Household tungsten-filament light bulb, 100 W	2 850
Standard candle	1 930

Table 1.3 ***Effective star temperatures***

Star colour and example	Effective temperature/K
Blue–white, Bellatrix	25 000
White, Sirius	11 000
Yellow–white, Sirius–Solar	7 500
Solar, the Sun	6 000
Orange–yellow, Arcturus	4 200
Orange, Antares	3 000
Deep orange–red, μ-Cephei	2 600

such as fluorescent lighting (Table 1.2). Colour photographs taken on film designed to be used in daylight (colour temperature of about 5 400 K) will show incorrect tones when used to photograph objects illuminated with tungsten lights (colour temperature of about 3 400 K) or fluorescent lights (colour temperature of about 3 000 K) unless correcting filters are used.

The most important incandescent object for us is the sun, which is the ultimate source of energy on Earth. The solar spectrum has a form quite similar to a black-body curve corresponding to a solar temperature of about 5 780 °C (about 6 000 K), which has a maximum near 480 nm. The form of the spectrum when it reaches the surface of the Earth is a function of a number of variables, including the elevation of the sun, the amount of scattering material in the atmosphere and so on. Light is perceived as white if it has a make-up like that of the solar spectrum from an overhead sun on a clear day. Stars which are cooler than the sun give a redder colour, whilst those which are hotter are perceived as whiter. The *effective temperature* of a star is the temperature calculated as if it were a black body radiating with the same energy over the same wavelength ranges (Table 1.3). The effective temperature is generally a good approximation to the surface temperature of a star. The hottest visible stars are the Bellatrix type, with blue–white colour and an effective temperature of approximately 25 000 K, whilst the reddest naked-eye star is μ-Cephei, the Garnet Star, with a temperature of approximately 2 600 K.

1.8 Photons

The quantization of radiation proposed by Planck in the derivation of the radiation law was not seized upon instantly. After a lapse of some years it was exploited by Einstein in his explanation of the photoelectric effect

in 1905 (Section 1.1). He proposed that the quantization of radiation contained in Planck's formula for black-body absorption and emission of energy, i.e.:

$$E = h\nu$$

where ν was the frequency of the radiation and h is Planck's constant, could be applied to the radiation itself, not just to the energy exchange with atoms or molecules. That is to say, light was to be regarded not as a wave but as a hail of bullet-like objects (which are now called *photons*), each of which had an energy $h\nu$. Each photon delivered the same amount of energy. If this was sufficiently large then the electron could be ejected from the surface. The energy of each photon was proportional to the frequency of the illumination, so that when the frequency passed a certain threshold, the photoelectron could be ejected, but not before that point was reached. Thereafter, increasing the frequency of the illumination allowed the excess energy to be displayed as an increase in kinetic energy. The kinetic energy of the photoelectrons ejected from a metal under this hail of photons could then be written as:

$$\tfrac{1}{2}mv^2 = h\upsilon - \phi$$

where ϕ is known as the *work function* of the metal and is the energy required to liberate the electron from the metal surface. The irradiance of the light indicated the number of photons arriving at the surface, so that the number of photoelectrons emitted is a function of irradiance, but the energy of these electrons is a function of the frequency of the radiation.

A description of light in terms of photons is mandatory when dealing with events at an atomic scale. The energy E of a photon is given by Equation 1.1:

$$E = h\nu = \frac{hc}{\lambda} \tag{1.1}$$

where ν is the frequency of the equivalent light wave, λ is the wavelength of the equivalent light wave, h is Planck's constant and c is the velocity of light in vacuum.

This conjunction of the particle and wave descriptions, called wave–particle duality, is evident in the fact that ν is the frequency and λ is the wavelength of the wave-like properties associated with the photon. In fact, all particles exhibit wave-like properties. The momentum p of a particle (such as an electron, say), is given by:

$$p = \frac{(E^2 - m^2c^4)^{1/2}}{c}$$

where E is the energy, m the particle mass and c the speed of light. For a photon, $m = 0$, so that:

$$p = \frac{E}{c}$$

The wavelength of a particle is:

$$\lambda = \frac{h}{p} = \frac{hc}{(E^2 - m^2c^4)^{1/2}}$$

For a photon, $m = 0$, so that:

$$\lambda = \frac{hc}{E}$$

The velocity of a particle is:

$$v = \frac{pc^2}{E} = c\left[1 - \left(\frac{m^2c^4}{E^2}\right)\right]^{1/2}$$

For a photon, $m = 0$, so that:

$$v = c$$

(For particles such as electrons, m is not zero.)

For many purposes the wave and particle aspects of light can be used interchangeably, as dictated by experiment. The wave aspect of light expresses the fact that the photons do not obey deterministic laws of motion, but laws of probability. The waves associated with light photons are a way of describing these probabilities.

1.9 Lamps and Lasers

1.9.1 Lamps

Until the end of the nineteenth century artificial illumination was via incandescence – either firelight, candles, oil lamps or gas light. At the end of this period, new light sources began to be invented in parallel with the generation and availability of electricity. In 1897 Nernst invented the 'glower'. This lamp consisted of a bar of electrically conducting ceramic made from a mixture of lanthanide oxides that became incandescent under the action of an electric current. Although Nernst glowers were widely used and were more efficient than the competing incandescent carbon-filament electric light bulbs developed by Edison, they fell into disuse following the successful introduction of tungsten-filament lamps after the invention of the Coolidge process for the production of ductile tungsten wires for the fabrication of lamp filaments. Throughout the twentieth century, tungsten-filament lamps dominated the lighting market.

Although incandescence was the most widespread source of artificial light, other lighting was well known. Neon signs (Chapter 7) and various forms of luminescence (Chapter 9) were used in specialist light-generating ways, such as, in the case of neon signs, for advertising. These latter mechanisms relied directly upon atomic transitions in a way that was obscured in the complex incandescence reactions. In addition, instead of generating a continuous 'white light' spectrum, these new light sources tended to give out coloured light, the wavelengths produced depending upon the actual atoms emitting the photons.

All of these light sources, however, were similar to each other in one way – the light emitted was incoherent and usually unpolarised. Towards the middle of the twentieth century, advances in communications technologies reinforced the utility of using light directly to carry signals. This necessitated the use of coherent radiation. Initially the push came from radio, as radio waves are normally emitted as a coherent wave train, not as incoherent waves. The wavelength of the waves used for carrying signals continually decreased via long waves, medium waves and short waves. At the same time, the engineering skills required to encode greater and greater information on these waves increased to an amazing extent, making television and stereo broadcasting a norm. Unfortunately, the production of coherent radiation seemed to be stuck somewhere in the microwave region. The idea of using lower wavelengths, though, especially optical wavelengths, was enormously attractive, and a

great deal of effort was invested into breaking into this wavelength range. Success came in the 1960s, with the invention of the *laser*. Lasers are a completely new sort of lamp compared with those already described.

The word *laser* is an acronym for the expression *Light Amplification by Stimulated Emission of Radiation*. The first laser to be made was the ruby laser, and the first laser light emitted was on 15 May 1960. Since then a vast number of lasers have been produced, including solid-state lasers, gas lasers, semiconductor diode lasers and dye lasers. From an exotic beginning lasers have become ubiquitous in modern life, being used as pointers, at check-outs in supermarkets, in surveying and measurement, in micromachining, microsurgery and so on. Here, the general principles of laser action will be outlined. Examples which illustrate particular facets of laser light generation will be discussed throughout the text.

1.9.2 Emission and absorption of radiation

When a photon of energy $h\nu$ is *absorbed* by an atom or molecule it passes from the normally occupied lower energy state, often called the *ground state*, to an upper or *excited state*, as described above. The transition will take place if the frequency of the photon ν, is given exactly by:

$$E_1 - E_0 = \Delta E = h\nu = \frac{hc}{\lambda} \tag{1.2}$$

where E_0 is the energy of the ground state, E_1 is the energy of the excited state and h is Planck's constant. If the atom is in the excited state E_1 and makes a transition to the ground state E_0, energy will be emitted with the same frequency, given by the same equation.

In this description the actual emission mechanism is ignored. In 1917 Einstein suggested that there should be *two* possible types of emission process (Figure 1.6):

1. An atom in an excited state can randomly change to the ground state, by a process called *spontaneous emission*.
2. A photon having an energy equal to the energy difference between the two levels (i.e. $E_1 - E_0$) can interact with the atom in the excited state, causing it to fall to the lower state and emit a photon at the same time, a process called *stimulated emission*.

The light emission from 'ordinary', i.e. non-laser, sources is the result of spontaneous emission. Lasers are concerned with stimulated emission. In spontaneous emission, the light photons all have the same frequency but possess random phases and polarisation so that the light is *incoherent*. In stimulated emission the photon produced has the same frequency, phase and polarisation, as the one which caused the emission so that the light is *coherent*. It is these important features of stimulated emission on which the special properties of laser light depend.

1.9.3 Energy-level populations

Under conditions of *thermal* equilibrium the relative populations of a series of energy levels will be given by the Boltzmann law, which for two energy levels can be written as:

$$\frac{N_1}{N_0} = \exp\left[\frac{-(E_1 - E_0)}{k_B T}\right]$$

where k_B is Boltzmann's constant, T is the absolute temperature, E_1 and E_0 are the energies of the excited state and the ground state respectively and N_1 and N_0 are the numbers of atoms (the *populations*) in each of these energy levels. For ordinary atoms, in a gas, liquid or solid at ordinary temperatures, the fraction N_1/N_0 will be

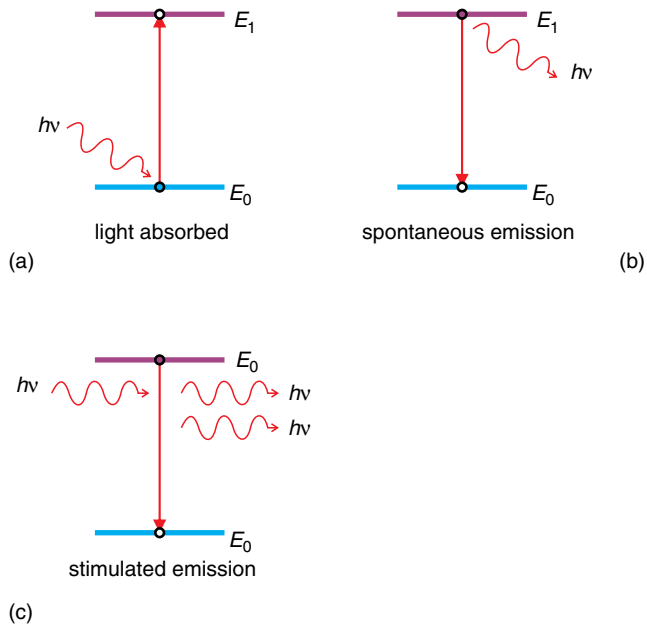


Figure 1.6 ***Light absorption and emission. (a) Light absorption occurs when a photon excites an atom (or molecule) from the ground state E_0 to an excited state E_1. (b) During spontaneous emission, the atoms lose energy and release photons at random. (c) During stimulated emission, an atom in an excited state is triggered to lose energy by interaction with a photon of energy ($E_1 - E_0$)***

negligible for energy levels which are sufficiently separated to give rise to visible light. Atoms can be assumed to be in the ground state as far as visible light emission is concerned.

When a photon of the appropriate energy interacts with an atom in the ground state it will be absorbed and shortly afterwards re-released by spontaneous emission (Figure 1.7a). This will be repeated at each atom in the ground state. There will be no amplification and we may well see a net absorption of energy. To obtain laser amplification one needs to ensure that stimulated emission is the dominant process occurring. This means that there are more atoms in the excited state of energy E_1 than in the ground state E_0. In this instance, a photon interacting with an excited atom can cause energy to be released by stimulated emission and two photons emerge. If most atoms are in the excited state then amplification may occur (Figure 1.7b). The situation in which more atoms are in the excited state than in the ground state is called a *population inversion*.

From the Boltzmann equation it is obvious that an increase in temperature cannot achieve this objective. Even an infinite temperature will only result in equal numbers of atoms in E_0 and E_1. To obtain a population inversion, therefore, a *nonequilibrium state* must be achieved. The crux of laser action is how to create such a nonequilibrium situation in a material and then exploit it to produce the desired amplification. Examples of practical ways in which this is achieved are given later in the text (i.e. see Chapters 7 and 10).

1.9.4 Rates of absorption and emission

In the previous section it was implicitly implied that the rate of spontaneous emission was fast. This aspect must be looked at in more detail to obtain a better understanding of laser action. When equilibrium between absorption and emission holds, the rate of depopulation of an upper level ($-\mathrm{d}N_1/\mathrm{d}t$) by spontaneous emission

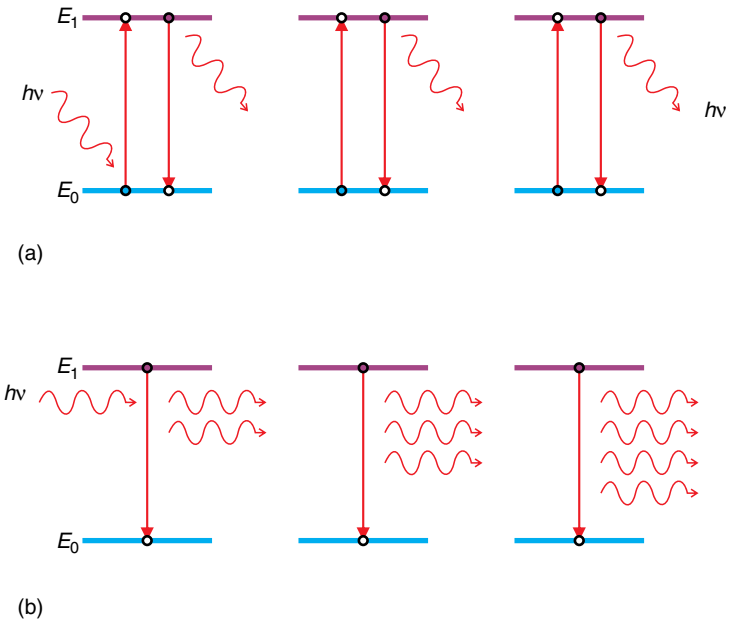


Figure 1.7 ***Amplification. (a) When most atoms are in the ground state the absorption of a photon and the subsequent spontaneous re-emission will not lead to amplification. (b) When most atoms are in the excited state, stimulated emission can lead to amplification***

will be given by a first-order rate law:

$$-\frac{\mathrm{d}N_1}{\mathrm{d}t} = A_{10}N_1$$

where the negative sign denotes that the number N_1 of atoms in the upper state E_1 (per cubic metre, say) is decreasing with time. The rate is proportional to the number of atoms N_1 in the state. The rate constant, denoted here as A_{10}, is called the *Einstein coefficient for spontaneous emission*, where the suffix '10' means that we are considering a transition from the excited state E_1 to the ground state E_0. The number of downward transitions due to spontaneous emission, per second, will be given by:

$$A_{10}N_1$$

Similar rate laws can be written for the cases of stimulated emission and for absorption, but in this case the rates are proportional to the numbers of atoms in the relevant state and, in addition, the number of photons present. The reactions can be taken to be first order with respect to both of these quantities.

The rate at which atoms in state E_0 are excited to state E_1 is then given by:

$$-\frac{\mathrm{d}N_0}{\mathrm{d}t} = B_{01}\rho(\nu_{01})N_0$$

where N_0 is the number of atoms in state E_0 (per cubic metre, say), $\rho(\nu_{01})$ is the radiation density responsible for absorption, which is the number of quanta per cubic metre incident per second at the correct excitation frequency ν_{01}, and B_{01} is the *Einstein coefficient for absorption of radiation*. Similarly, the rate of depopulation

of state E_1 by stimulated emission is given by:

$$-\frac{\mathrm{d}N_1}{\mathrm{d}t} = B_{10}\rho(\nu_{10})N_1$$

where N_1 is the number of atoms in state E_1 (per cubic metre), $\rho(\nu_{10})$ is the radiation density responsible for depopulation, which is the number of quanta per cubic metre incident per second at the correct frequency ν_{10}, and B_{10} is the Einstein coefficient for stimulated emission of radiation. Now, the correct frequency for excitation will be the same as that for depopulation, so that $\nu_{10} = \nu_{01}$, which we can simply write as ν, and the radiation density will be the same in each case, so that we can write:

$$\rho(\nu_{10}) = \rho(\nu_{01}) = \rho(\nu)$$

The number of stimulated downward transitions per second will be given by:

$$N_1 B_{10}\rho(\nu)$$

while the total number of upward transitions in the same time will be given by:

$$N_0 B_{01}\rho(\nu)$$

At equilibrium, the total number of transitions in each direction must be equal; hence:

$$N_0 B_{01}\rho(\nu) = N_1 A_{10} + N_1 B_{10}\rho(\nu)$$

so

$$\rho(\nu) = \frac{N_1 A_{10}}{N_0 B_{01} - N_1 B_{10}}$$

In addition, at equilibrium the Boltzmann distribution applies; thus:

$$\frac{N_1}{N_0} = \exp\left(\frac{-h\nu}{k_\mathrm{B}T}\right)$$

and by making this substitution we have:

$$\rho(\nu) = \frac{A_{10}}{\exp(h\nu/k_\mathrm{B}T)B_{01} - B_{10}}$$

This expression represents the radiation density at frequency ν. At thermal equilibrium, this should be identical to Planck's equation, Equation 1.6a:

$$\rho(\nu) = \frac{8\pi h\nu^3}{c^3[\exp(h\nu/k_\mathrm{B}T) - 1]}$$

which leads to the conclusion that:

$$B_{01} = B_{10} = B$$

and:

$$\frac{A_{10}}{B} = \frac{8\pi h\nu^3}{c^3}$$

The *ratio* of the rate of spontaneous emission to stimulated emission under conditions of thermal equilibrium is given by:

$$R = \frac{A_{10}}{\rho(\nu)B} = \exp\left(\frac{h\nu}{k_B T}\right) - 1$$

This is an extremely interesting result. At 300 K, at visible wavelengths, $R \gg 1$. This shows that, for light, stimulated emission will be negligible compared with spontaneous emission and reinforces the idea that it will be impossible to make a laser under equilibrium conditions. On the other hand, if the wavelength increases beyond the infrared into the microwave and radio-wave regions of the electromagnetic spectrum, R becomes *much less* than unity and *all* emission will be stimulated. Hence, radio waves and microwaves arise almost entirely from stimulated emission and are always coherent. This is one of the main reasons that communications in the early part of the twentieth century used radio waves.

Perhaps because of this equation, and the towering reputation of Einstein, it seems that for the first part of the twentieth century it was felt that lasers were not feasible. In the middle of the century, scientists started to explore stimulated emission at microwave frequencies, developing the *maser*. This soon led to the first lasers, the ruby laser and then the He–Ne gas laser, produced in 1960 with these early devices often being called *optical masers*. Once the way to overcome the production of laser light was understood, laser development became prolific. Later sections show how the equilibrium problem has been bypassed and how the difficulty of achieving stimulated emission at optical wavelengths has been overcome.

1.9.5 Cavity modes

Supposing that a population inversion is obtained between energy levels that would give rise to visible light, it is still necessary to design the equipment so that amplification of the signal takes place. The losses from the laser must be less than the total emission for amplification to be achieved. Losses in oscillating systems are often defined in terms of a quality factor Q, a term borrowed from radio technology. In effect, a high value of Q is needed to ensure amplification.

One of the most important of these design features is the shape of the *cavity* that the laser medium occupies. Suppose that this is simply a crystal rod. The population is an unstable state and after a short time some spontaneous emission will occur from E_1. Naturally, these photons will rapidly leave the crystal rod; and although in so doing a few other atoms might lose energy via stimulated emission, no amplification will occur. It is necessary to prevent the photons from leaving the crystal in order to increase the chances of stimulated emission occurring. The simplest way to achieve this is to coat the ends of the crystal rod with a highly reflecting mirror. In this case the photons are reflected to and fro, causing stimulated emission from the other populated E_1 levels. Once started, the stimulated emission rapidly depopulates these levels in an avalanche. In order to permit some light to emerge, one of the mirrors is not perfect and allows a small

amount of light to pass. There will then be a burst light emerging from the cavity which is not only coherent but also shows amplification. Thus, the simplest cavity geometry is simply cylindrical with one end fitted with a completely reflecting mirror and the other with an almost perfect mirror, appropriate to the wavelength of the light generated by the stimulated emission.

There are several consequences of this simple geometry which are easiest to explain if the light trapped in the cavity is regarded as a wave. Taking the cavity as a rod with reflecting end faces, it is clear that initially all photons will be emitted at random, but only those that are emitted more or less parallel to the long axis of the cavity will bounce to an fro and so cause the stimulated emission avalanche. In terms of wave optics, the photons form a series of standing waves in the cavity, which is described as *resonance*. The standing waves form only if there is a node at each reflecting surface. The allowed waves are called *longitudinal cavity modes* and are given by the condition that a complete number of half wavelengths must fit into the length l of the cavity, i.e.:

$$m_c = \frac{l}{\lambda/2} = \frac{2l}{\lambda}$$

where m_c is an integer, l is the cavity length and λ is the wavelength of the mode. The frequency of a mode is given by:

$$\nu_m = \frac{m_c v}{2l}$$

where v is the velocity of the light waves in the cavity, given by $\nu_m \lambda$, and ν_m is the frequency of the mode m_c. The separation of the modes is given by:

$$\nu_m - \nu_{m-1} = \Delta\nu = \frac{v}{2l}$$

The velocity of light in the cavity is given by:

$$v = \frac{c}{n}$$

where c is the velocity of light in a vacuum and n is the refractive index of the cavity medium (Chapter 2), so that:

$$\nu_m - \nu_{m-1} = \Delta\nu = \frac{c}{2ln}$$

How does this work out in practice? The emission from the upper to the lower energy level has been written as a single energy with a negligible width. In the case of real materials, atoms and molecules are in continuous motion, vibration in solids, translation in gases, and the sharp energy levels idealized in Figures 1.6 and 1.7 give rise to a spread of energies (or of frequencies or wavelengths) called the transition bandwidth (Figure 1.8a). Only that part of this output that fulfils the longitudinal mode criterion will be allowed to grow. The output from the cavity will then be composed of a set of modes (Figure 1.8b). These modes will depend upon the shape of the initial emission pulse and the overall power of the excitation process.

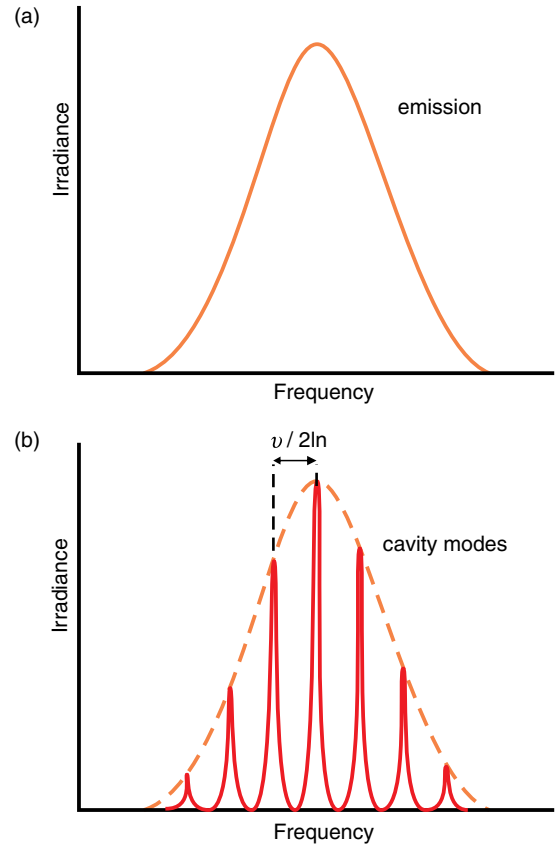


Figure 1.8 ***(a) The emission from an excited state E_1 to the ground state E_0 is not sharp, but consists of a range of frequencies dependent upon temperature and other factors. (b) In a laser cavity, only certain frequencies, the cavity modes, are allowed to propagate***

By extension, it is apparent that, in general, there will be transverse modes as well as longitudinal modes in the laser emission. These must be taken into account when the optics of the laser beam are considered. Laser cavity design is, therefore, of considerable importance in practice.

1.10 Vision

As stated earlier, light has no colour as such. Light radiation leaves the source, possibly interacts with matter in the course of passage and then enters the eye. Light is perceived by the eye–brain combination, and colour is a description of this perception. The colour that the observer is conscious of is thus a combination of many factors, including the energy spread of the source light, the addition or subtraction of energy during any interactions with other materials and the sensitivity of the eye. For example, the blue sky contains all the colours of the spectrum, as can be demonstrated by passing this light through a prism (Chapter 2). Blue is the colour attributed to the sky when all the factors mentioned above are taken into account.

The physiological response of the eye–brain combination arises when light waves fall upon the light-sensitive *retina*, which makes up the inner surface of the eye. In 1876 Boll reported that the red–purple pigment found in this part of the eyes of animals bleached in the presence of light to a colourless form. The change was found to be reversible, and in the dark the purple colour was regenerated. This important photochromic reaction is the source of vision. The compound involved became known as 'visual purple' and is now called *rhodopsin*.

Vision in humans and other animals involves a complex set of reactions which take place in two types of photoreceptor cells located in the retina of the eye: *rods* and *cones*. There are about 10^8 rods and 4×10^6 cones in an eye. In humans the rod cells, of about 0.002 mm diameter, are about four times as sensitive as cones and are responsible for vision at low light intensities. Although they detect light all across the visible, the peak sensitivity is at ~500 nm. The light not absorbed, red and blue/violet, gives rise to the purple colour of the membrane. The rod cells are not sensitive to colour and give rise to a monochrome image. Moreover, they saturate in high light levels, making them unresponsive under these conditions. The cone cells, approximately 0.006 mm diameter, are sensitive to bright light and form the daylight colour detection system. They exist in three varieties with peak sensitivities in three different regions of the visible: L cones, most sensitive to red, λ(peak) ~560 nm, M cones, most sensitive to green, λ(peak) ~530 nm, and S cones, most sensitive to blue λ(peak) ~420 nm (Figure 1.9a). The human eye is optimally sensitive to green light and is noticeably less sensitive to red and especially blue light (Figure 1.9b). The sensitivity of the eye to colour depends not only upon the amount of light, but also upon which area of the retina is being stimulated. The most sensitive region, called the *fovea*, is almost directly behind the lens of the eye and predominantly contains cone cells. The maximum sensitivity of a normal eye to bright white light focused on the fovea, which is the sum of the contributions of the cone and rod cells, is for a wavelength close to 555 nm (Figure 1.9c). Colour blindness results from a fault or deficiency in one or more varieties of the cone cells or in the way in which these cells communicate with the brain.[4]

Human vision is said to be trichromatic. There is considerable variation across the human population in the sensitivity ranges of the cone cells, giving rise to a variation in colour vision. Trichromaticity is common among primates, but most nonprimate animals can only detect two colours and are referred to as dichromats. However, some birds, fish and reptiles have four different cone cell receptors and can detect ultraviolet light with λ(peak) as low as 360–380 nm in addition to three 'normal' colours.

When light photons impinge on both rod and cone cells they are absorbed by stacks of photoreceptor molecules which are bleached in the process. This sends a nerve impulse to the brain. The system is remarkably sensitive and there is considerable evidence to suggest that in the rod cells just one photon is enough to stimulate the nerve. The light-absorbing pigments consist of a protein, an *opsin*, bound to a light-absorbing molecule, *retinal*. The receptor in the rod cells is called rhodopsin, while those in the cone cells are called *cone opsins*. The opsin part of the receptor, consisting of 364 amino acid residues in humans, is arranged in the form of seven helices, which penetrate the cell wall and enclose the retinal, which is bound to the amino acid lysine 296 (Figure 1.10). The opsin proteins differ from one cone cell to another and from rhodopsin in the rods, and it is these differences that confer the differing sensitivities to the receptors. However, the differences are rather small. For example, the amino acid sequences in the green (M) and red (L) cone receptors in humans differ in only three of the amino acid residues in 364.

[4] The existence of colour blindness itself was first recorded as such by John Dalton, who realized that his own perception of colours was different than the majority of his friends (but the same as his brother's), and for many years the condition was known as *Daltonism*. A more recent study of his careful observations suggests that he was unable to distinguish the colour red. It is of interest to learn that Dalton himself felt that he possessed some visual advantages over his friends because of the nature of the abnormal sensitivity of his eyesight. He did not find that he was at a disadvantage at all. (Also see Figure 1.15.)

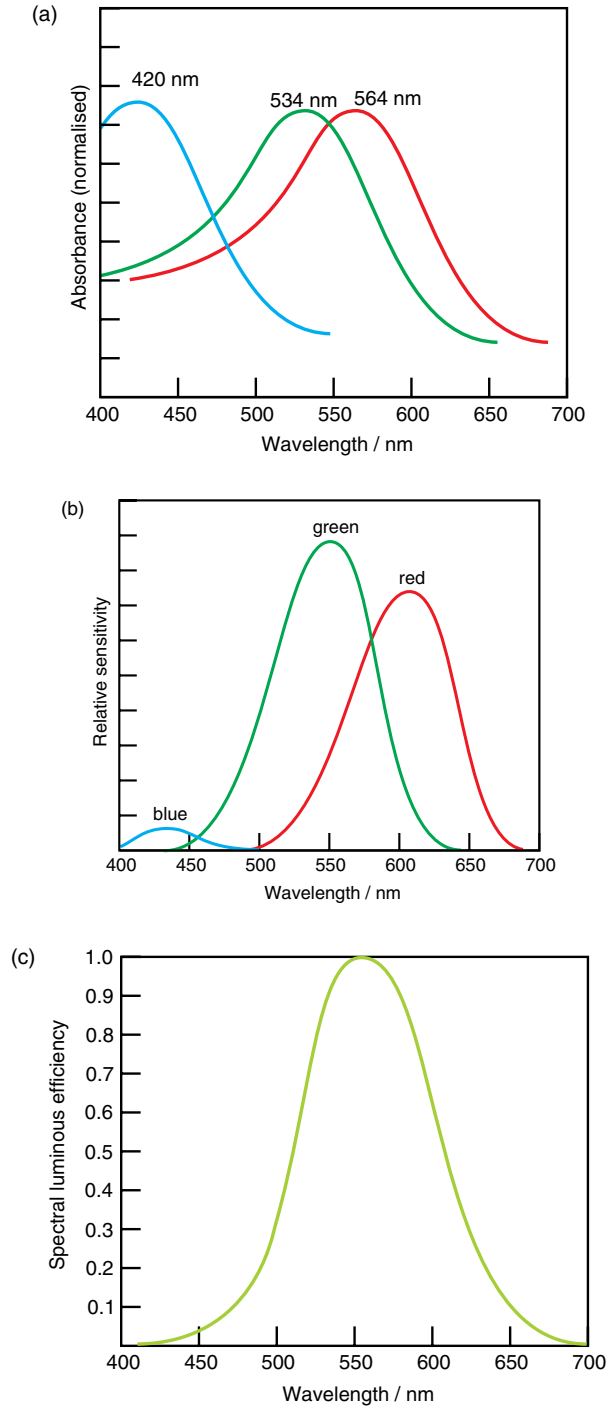


Figure 1.9 ***The sensitivity of the eye to light, schematic. (a) Sensitivity of the cone cells in a normal eye to light as a function of the wavelength. (b) Visual sensitivity of a normal eye to red, green and blue light. (c) Overall visual sensitivity of a normal eye to light; the photopic spectral luminous efficiency function. The maximum sensitivity is for a wavelength close to 555 nm***

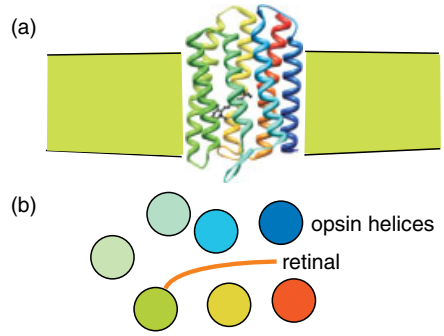


Figure 1.10 ***(a) The schematic structure of an opsin protein in the cell wall of a photorectetor. (b) The opsin protein molecule is in the form of seven helices arranged to enclose a retinal molecule. [(a) is adapted from http://en.wikipedia.org/wiki/Rhodopsin]***

In humans, retinal is derived from the compound β-carotene (Section 8.5), an orange pigment found in carrots. This is transformed into vitamin A in the liver, which then forms retinal. The visual pigments in animals then consist of retinal plus opsin. (There are two forms of vitamin A: A_1, which gives retinal$_1$ (11-*cis*-retinal, the aldehyde of vitamin A_1), and A_2, which gives retinal$_2$ (3-dehydro-retinal). Retinal$_1$ is used by all mammals and birds and will just be referred to as retinal in what follows.)

The framework of the processes triggering vision is well established. It is described here with respect to rod cells, which have been studied in most detail. The *chromophore* (light absorbing part) of rhodopsin is the *cis*-form of the molecule *retinal*, 11-*cis*-retinal (Figure 1.11a). This *cis*-retinal molecule is bound to the opsin via the amino acid lysine, to form rhodopsin (Figure 1.11b). The *cis*-retinal by itself is not coloured and has an absorption maximum between 370 and 380 nm. However, when joined to the opsin the absorption maximum moves to about 500 nm. Molecules which can cause the deepening of the colour of a chromophore are called *bathochromes* and the resultant movement of the absorption maximum is referred to as a *bathochromic shift*. The bathochromic shift comes about because of the particular conformation of the *cis*-retinal molecule in conjunction with the protein. The bonding and slight differences in the various forms of the opsin molecules produce different bathochromic shifts and, hence, make the cones sensitive to the different wavelengths of red, green and blue light.

The molecular mechanism leading to the nerve impulse hinges on the fact that retinal can exist in two isomeric forms, the *cis*-form already described and a *trans*-form, called *all-trans*-retinal. Under the influence of a photon the *cis*-retinal molecule changes to *all-trans*-retinal rhodopsin (Figure 1.11c). Absorption of light by rhodopsin drives the molecule through several intermediates to the bleached state, which can consist of a number of different molecules (metarhodopsin I, metarhodopsin II and so on), depending upon the conditions experienced. Thereafter the reaction reverses, again passing through a number of intermediates, so that the *trans*-retinal readopts the *cis*-conformation and reforms rhodopsin (Figure 1.11d). Another photon can trigger the cycle again. Each cycle takes only a fraction of a second and can repeat indefinitely in normal light conditions so as to send a stream of nerve impulses to the brain. These impulses end when the light is extinguished and all molecules revert to rhodopsin.

It is worth commenting on the enormous complexity of vision. The description of the cycle occurring in rod cells and presumed to occur in the cone cells described above is only true at moderate light intensities. At lower light intensities the *trans*-retinal molecule in rod cells is released completely from the opsin. Two processes then operate, dependent upon the weakness of the light signal. At the 'higher' of these lower intensities the *trans*-retinal is transformed back to the *cis*-conformation by the action of enzymes in the eye itself, whereupon

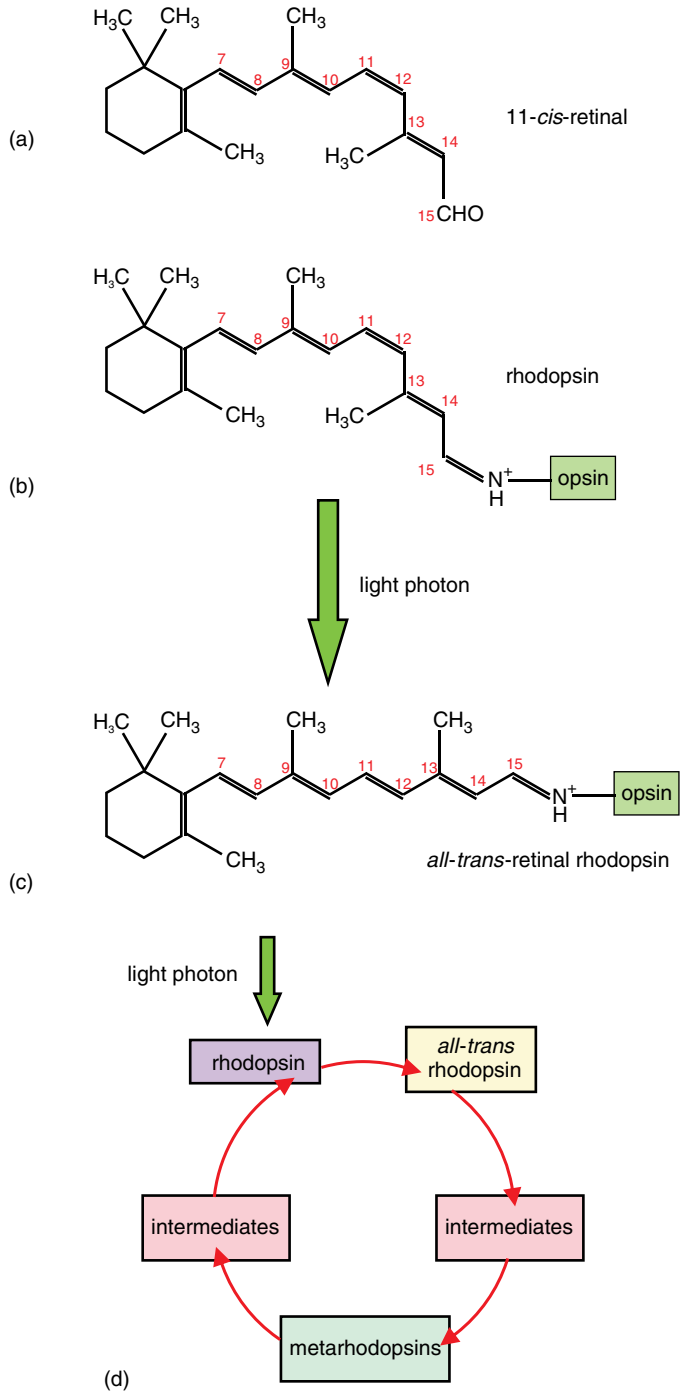


Figure 1.11 ***The structures of (a) 11-cis-retinal, (b) rhodopsin and (c) all-trans-retinal rhodopsin, produced by the action of light on (b); (d) cycle of chemical changes producing vision. In normal illumination this process is repeated many times a second. Each cycle results in the transmission of a signal along the optic nerve to the brain***

the molecule is reattached to the opsin. At the lowest light intensities, the *trans*-molecules actually leave the eye completely, enter the bloodstream and are reprocessed to the *cis*-form in the liver, an occurrence which contributes to the length of time that it takes to become fully 'dark adapted'.

Rhodopsin has another role to play in the broader picture of life. It has been found that some purple halobacteria, bacteria which inhabit very salty environments, are coloured purple by a version of rhodopsin called bacteriorhodopsin. This consists of 247 amino acid residues, arranged in seven helices, with the photoactive retinal attached to lysine 216. It is, however, not used for vision, but in an analogous fashion to chlorophyll in plants. Absorption of light by chlorophyll initiates a chain of electron transfer reactions which eventually provide the energy for plant growth. In the purple halobacteria, the rhodopsin converts sunlight into energy for the metabolism of the bacterium. In essence it appears that the *cis–trans* change acts as a proton pump, and the resulting electrochemical potential created initiates the energy building steps.

1.11 Colour Perception

Recognition of colour is a function not only of the physical make-up of the light falling on the eye and physiological factors, but also of psychological biases. The 'colour' of an object in this sense is changed by factors such as surface roughness or texture. Subsurface scattering, which returns some incident light on a body to the exterior, is of importance in the appearance of skin, cosmetics and paint. Because of this interplay it is possible to distinguish a hard red plastic surface from a red velvet surface even though in terms of physics the colours of both may be identical, originating in the same dye or pigment. It is clear that when describing the appearance of an object in colour terms it is necessary to consider specular (mirror-like) reflection, diffuse (non-mirror-like) reflection and subsurface scattering, as well as the make-up of the light which is reflected or scattered. Moreover, human eyes vary in colour-interpreting ability. It appears that an average person can distinguish more than a million different colours. All of these aspects are implied when the colour of an object or a light source is mentioned in a colloquial way. Because of this, colour is difficult to quantify.

Despite the complexity inherent in the concept of colour and its perception, it has been found that all colours can be precisely specified by three parameters. Colours can then conveniently be represented by points in a three-dimensional coordinate system. There are many diagrammatic ways of representing the three attributes, and these are called *colour spaces*. The way in which the coordinates of any colour in the colour space are derived is called a *colour model*. There are many colour models, of which only three will be described briefly in this book. (More information can be found in Section 1.17.)

One widely used colour model takes as initial parameters the three attributes *hue*, *saturation* and *brightness* to give the *HSB model*. These characteristics are generally taken to be:

1. *Hue*, which corresponds to the wavelength or frequency of the radiation. The hue is given a colour name such as red or yellow.
2. *Saturation* or *chroma*, which corresponds to the amount of white light mixed in with the hue and allows pale 'washed out' colours to be described.
3. *Brightness, lightness, luminance*, or *value*, which describes the intensity of the colour, the number of photons reaching the eye.

This model is also given the acronyms *HVC* (hue, chroma, value), *HSL* (hue, saturation, luminance), *HIS* (hue, intensity, saturation) and *HCL* (hue, chroma, luminance). One way of building a colour space in terms of this colour model is to arrange the hue around the periphery of a disc with the degree of saturation of the colour represented by the distance from the centre of the disc along the radius. Brightness is defined by an axis perpendicular to the centre of the disc (Figure 1.12a). This arrangement has been quantified in constructions such as the *Munsell colour cylinder* or *Munsell colour solid* (Figure 1.12b).

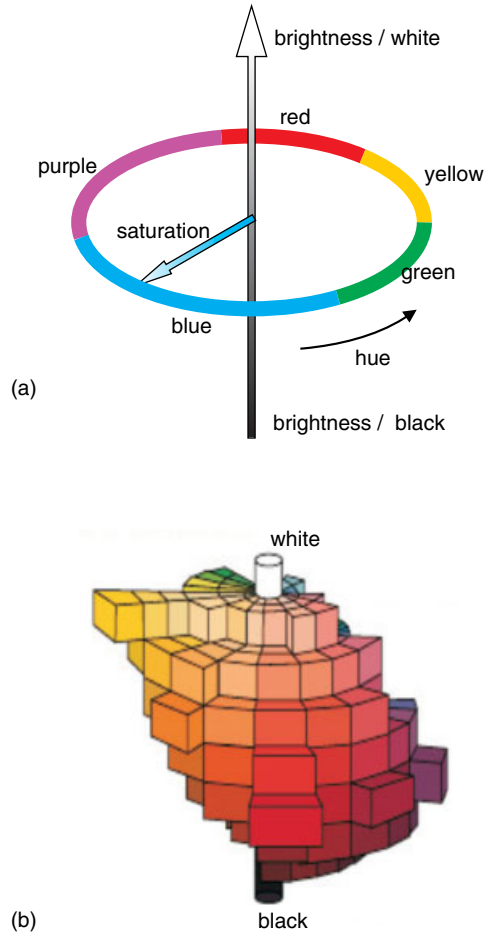


Figure 1.12 ***The representation of colours on a cylindrical colour space in the HSB colour model. (a) The hue is given by a point on the circumference of a planar disc, the saturation by the distance along the radius from the centre of the disc and the lightness by the vertical axis of the system. (b) The solid representation of the colours forms a colour cylinder, the best known of these being the Munsell colour cylinder [adapted from the Epson Online Printer Guide]***

1.12 Additive Coloration

Additive colour mixing occurs when two or more beams of differently coloured light combine (i.e. overlap on a perfectly white surface, or arrive at the eye simultaneously).

Colours on television screens are produced by additive coloration, as the screen is composed of small dots of three different phosphors each of which shines with one of three primary colours when activated. Additive coloration is also used in the painting technique known as pointillism. In this method of painting, the image is built up by placing small dots of relatively saturated colour onto the canvas, making sure that they do not overlap. When viewed from a distance of a few metres such pictures appear bright and dynamic.

The colour patterns on the wings of many butterflies and moths are produced in a similar way. The wings are tiled with a fine mosaic of scales, each of which reflects only one colour. The colour perceived by the eye is an

additive colour arising from the numerous closely spaced scales. The range of colours which can be produced by rather a few basic pigments is remarkable. For example, some perceived purples arise from mixtures of black, white and red scales, while some greens arise from mixtures of yellow and black scales.

It has been found that the majority of additive colours can be produced by mixing just three *additive primary colours*, red, green and blue. (Strictly speaking, any fairly monochromatic light near to these colours will suffice). Moreover, mixing equal quantities of these three primary colour lights will produce white light. There are a number of ways of quantifying the amounts of each primary colour light present, which can represented by the values, r of the *red* component, g of the *green* component and b of the *blue* component; thus:

$$\text{colour} = r + g + b$$

Use of these three additive primaries is called the *RGB colour model*.

A simple colour space can be constructed by using Cartesian axes to represent the amount of the three primary colours, red, green and blue, while the diagonal represents the transformation from black to white (Figure 1.13a). Sections through this colour space allow one to represent colours by a planar figure. Such representations are called *chromaticity diagrams*. A simple example is given by taking the triangular sheet running diagonally through the cube normal to the black–white diagonal and cutting the corners of the cube that represent pure red, green and blue. This produces a *colour triangle* (Figure 1.13b). Other colours can be

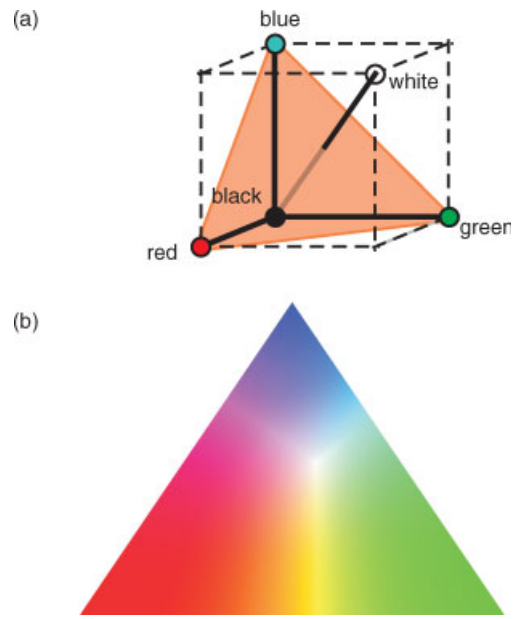


Figure 1.13 ***Colour spaces and chromaticity diagrams. (a) RGB colours represented by Cartesian axes, with black to white along the body diagonal. (b) A colour triangle, a section of (a) taken normal to the body diagonal passing through red, green and blue corners of the cube. A combination of the three primary colours at the vertices of the triangle will yield grey, but is shown white here. Other colours within the triangle (the gamut) can be represented by a point in the plane of the triangular system***

specified by coordinates in the plane of the colour triangle. The location given by the coordinates corresponds to the amounts *r*, *g* and *b* making up the colour. The coordinates which specify the case when the three primary colours are mixed in equal amounts will correspond to a shade of grey, but is usually represented by the colour white. The range of available colours that can be obtained by mixing lights corresponding to the three vertices is the *gamut* of colours available. Chromaticity diagrams generally represent hue and saturation, but not lightness (i.e. the grey tone), which must still be added as a third axis perpendicular to the chromaticity diagram if this information has to be displayed.

The study of light mixing has been quantified by the Commission Internationale de l'Eclairage (CIE), which has, on a number of occasions, refined the rather simple colour triangle concept so as to allow colour perceptions to be more accurately characterised. A colour is specified by a pair of *x*- and *y*-coordinates, which are derived from the *r*, *g* and *b* values noted above by the application of a standardized set of equations. In this representation, the triangular shape has been distorted into an outline something like a parabola, depending upon the way in which the *x*- and *y*-axes are plotted. A commonly encountered form of the CIE chromaticity diagram is that first proposed in 1931 (Figure 1.14a). The spectral colours are arranged around the outer edge of the shape and colours not seen in the spectrum, the purples and browns, are found to lie between the red and violet ends of the curve. The colours are *fully saturated* along the outer edge of the curve and become less and less saturated as the centre of the diagram is approached. Standard daylight white is represented by a point close to the coordinates $x = y = 0.33$, shown as W in Figures 1.14a, b.

If a straight line is drawn through the point W and extended to the boundaries of the curve, the pair of colours reached, when mixed, will give white light. For example (see Figure 1.14b), a line connects the colours red, of wavelength 700 nm and blue–green, of wavelength 492 nm and passes through the point W. The proportions of the end colours red and blue–green light needed to produce white light is given by the *lever rule* (Figure 1.14c):

$$\text{amount of red light} = r/(r + c)$$

$$\text{amount of blue-green light} = c/(r + c)$$

Measurement shows that mixing red of wavelength 700 nm and blue–green light of wavelength 492 nm in the proportions 39% red to 61% blue will produce white light. The colours at the ends of a line through the point W are called a *complementary pair* of colours. If one of these colours is subtracted from white light then the colour remaining is called the *complementary colour* to the first.

As with the colour triangle, all planar chromaticity diagrams represent hue and saturation, but not the exact value of lightness, which must still be added as a third axis perpendicular to the chromaticity diagram if this information has to be displayed. In general terms, therefore, the white region on the chromaticity diagram should be represented by grey, with white and black being extremes on the vertical axis perpendicular to the plane of the figure.

The accurate rendition of additive coloration is of prime importance in displays, such as television screens and computer monitors. Additive coloration and the interconvertion between various colour models is most easily explored using a computer which has photography or drawing editing software installed. On most of these packages, seven or eight or so different colour models are available, including RGB and at least one CIE model. The coordinates of any colour are given and comparisons between several systems are rapidly made. The instructions and help facilities give full information upon these options and how they affect colour rendition.

The confusion that colour blindness can cause is easily understood in terms of a chromaticity diagram. For example, Dalton had a lack of red receptors (Footnote 4). The CIE 1931 chromaticity diagram can be used to illustrate this. Any colour formed by mixing red with another colour, C, around the periphery of the curve will not be differentiated from any other colour along the line joining red to C. These lines show the loci of *colour confusion* (Figure 1.15). Other types of colour blindness will lead to other loci of colour confusion.

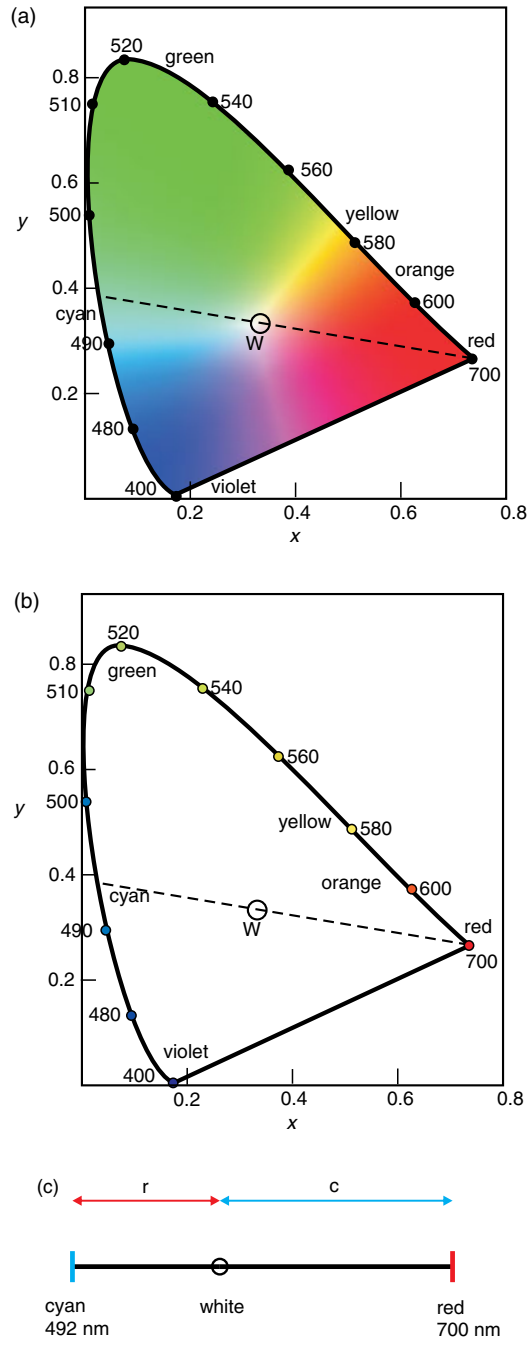


Figure 1.14 *The CIE 1931 chromaticity diagram. (a) The colours of the spectrum are arranged around a curved line and nonspectral colours fall on the line joining violet (400 nm) and red (700 nm). The figures marked around the outer edge of the curve denote the wavelength of the colour. Points within the area of the diagram represent colours formed by the additive mixing of light and can be specified by the appropriate x- and y-values. The point W represents white light. (b) A straight line through W links two complementary colours on the periphery of the diagram, in this example red and cyan. (c) The lever rule gives the proportions of complementary colours which are needed to create white light. In this example, the amount of red light is given by* $r/(r + c)$ *and the amount of cyan light by* $c/(r + c)$

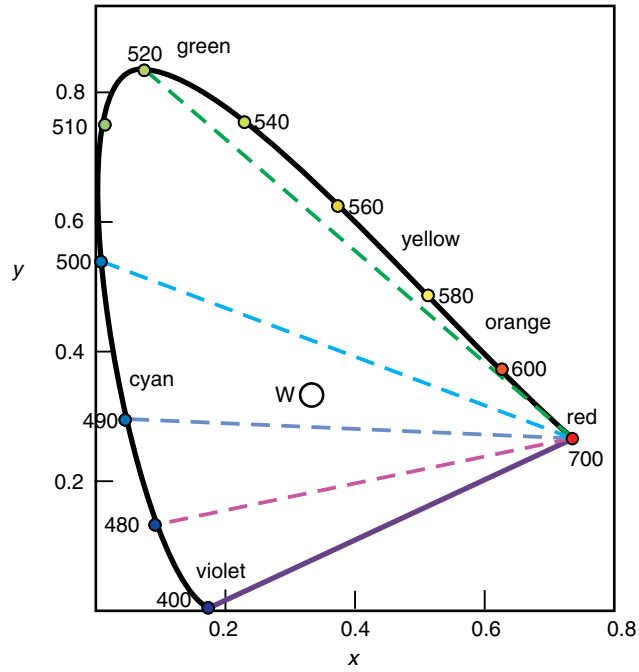


Figure 1.15 ***The dashed lines represent the loci of colour confusion for a person with red-defective vision plotted on the CIE 1931 chromaticity diagram. Because of a fault in the red perception, all colours on each line appear similar to the colour at the low wavelength extremity***

1.13 The Interaction of Light with a Material

Colour is inherent in the light that leaves an emitting source; but most often before it reaches the eye it interacts with matter of many types: gases, liquids and solids. The colour observed is thus a function of both the source radiation and the interactions that have occurred.

The way that light interacts with a material can be described in terms of scattering or absorption. To a first approximation, scattering is well treated by assuming that the light behaves as an electromagnetic wave, while absorption is best treated in terms of photons. If the energy of the scattered wave/photon is the same as that of the incident wave/photon then the scattering is called *elastic* scattering, and otherwise *inelastic* scattering.

For historical reasons, the term scattering itself, especially elastic scattering, is usually reserved for the interaction of light with randomly distributed small particles. Elastic scattering from a surface is normally called *reflection*, and elastic scattering into a transparent solid is called *refraction*. Scattering from ordered collections of small particles, or from small detail on larger objects, is called *diffraction*. For the purposes of this book these terms are retained as they stand, although all are simply different aspects of scattering. All of these processes are wavelength dependent, and so can result in the production of coloured light from white light.

Inelastic scattering arises when energy is transferred from the light photons to an absorption centre. Absorption is generally the term reserved for use when some or almost all of the incident radiation is taken up by the material and inelastic scattering when the changes are rather small. During absorption the energy is used to excite the component atoms or molecules that constitute the absorption centres into higher energy levels. Often,

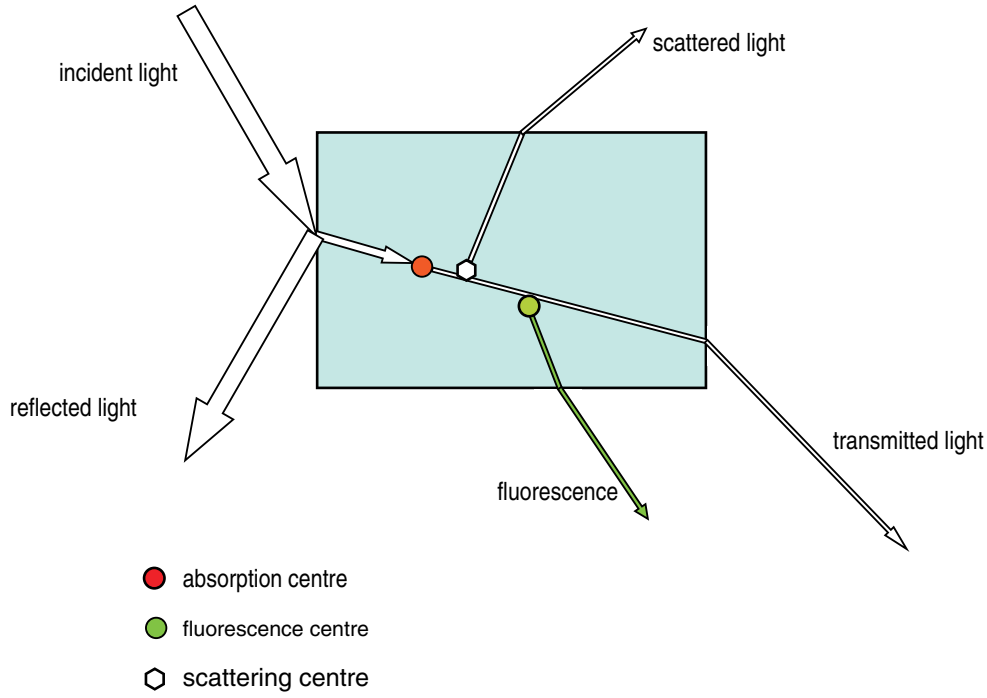


Figure 1.16 ***The interaction of light with a transparent material. The light can be reflected, absorbed or scattered. Some absorption centres are able to re-emit light as fluorescence or luminescence. All of the processes labelled are wavelength dependent and can lead to colour production***

the absorbed energy is manifested as a rise in temperature of the body. On occasion, some of this energy might be re-emitted as light, giving rise to *fluorescence* and related features. A material that does not absorb significantly is said to be *transparent*. Absorption may be minimal and transparency maximal for high-quality optical components over the visible spectrum, but no material is transparent over all wavelength ranges. Silicon, for example, appears 'metallic' over the visible spectrum but is transparent to infrared wavelengths. Absorption is wavelength dependent and an important source of colour production. It is often difficult experimentally to separate the relative roles that absorption and scattering play in the interaction of light with a material.

As a beam of light passes through a material it gradually loses intensity, a process generally called *attenuation* (formerly *extinction*). Attenuation is due to the interaction of light with a material in two basic ways: scattering or absorption (Figure 1.16). When attenuation takes place in a homogeneous solid the amount of light transmitted by a semitransparent plate of thickness x is given by:

$$I_x = I_o \exp(-\alpha_e x) \tag{1.7}$$

where I_x is the irradiance leaving the plate,[5] I_o is the incident irradiance and α_e (m^{-1}) is the (*Napierian*) *linear attenuation coefficient* (formerly *extinction coefficient*). Equation 1.7 is known as *Lambert's law* or *Beer's law*, although it was first clearly set out by Bouguer and should, by rights, be called Bouguer's law. The

[5] The symbol I is used for irradiance instead of E to avoid confusion with the use of E for energy throughout this book. See also Appendix 1.1.

attenuation length is defined as $1/\alpha_e$. The amount of light removed from the beam is thus:

$$I_{rem} = I_o - I_x = I_o - I_o \exp(-\alpha_e x) = I_o[1 - \exp(-\alpha_e x)]$$

If the attenuation of the beam is solely due to absorption, then the attenuation coefficient is replaced by the (*Napierian*) *linear absorption coefficient* α_a. Similarly, if the attenuation is solely due to scattering, then the attenuation coefficient is replaced by the (*Napierian*) *linear scattering coefficient* α_s. For nonhomogeneous solids these coefficients may vary with direction. Note that the degree of attenuation will vary significantly across the spectrum and the attenuation coefficient is not a constant.

It is sometimes convenient, as when discussing the absorption of X-rays, to define a *mass absorption coefficient* μ, which describes the decrease in transmitted irradiance through a homogeneous material of density ρ and thickness x:

$$I_x = I_o \exp(-\mu\rho x)$$

In this case:

$$\mu = \frac{\alpha_e}{\rho}$$

where μ has units $m^2\,kg^{-1}$ (in older literature $cm^2\,g^{-1}$).

Attenuation is often associated with the presence of chemical or physical 'centres', which may be atoms, molecules or larger particles, distributed throughout the bulk of a material. In the case of the mass absorption coefficient described above these are the totality of the atoms that make up the material itself. In this case, if the atoms in the material are supposed to absorb radiation independently of each other, then the mass absorption coefficient of the phase is simply related to the weight fraction of each atom species present. Thus, the mass absorption coefficient of a material M with a formula $A_xB_yC_z$ is:

$$\mu_M = (\text{wt fraction A}) \times \mu_A + (\text{wt fraction B}) \times \mu_B + (\text{wt fraction C}) \times \mu_C$$

The weight fraction of each species is given by:

$$\text{wt fraction A} = \frac{\text{mass of A present}}{\text{total mass}} = \frac{x(m_A)}{x(m_A) + y(m_B) + z(m_C)}$$

and so on, where m_A is the molar mass of species A, m_B is the molar mass of species B and m_C is the molar mass of species C.

More often, extinction is associated with a dilute concentration of centres distributed throughout the bulk phase. In this case, the degree of extinction is often taken to be a function of the concentration of these centres. This is taken into account in the *Beer–Lambert* or *Beer–Lambert–Bouguer law*:

$$\log\left(\frac{I_x}{I_o}\right) = -\varepsilon c x$$

where I_x is the irradiance after passage through a length of sample x, I_o is the incident irradiance and c is the *molar concentration* (mol L^{-1}, i.e. mol dm^{-3}) of the active centres or species. The quantity ε is called the *molar (decadic) attenuation coefficient* and has units[6] of $m^2\,mol^{-1}$. The attenuation coefficient has units of area and can, therefore, be regarded as an attenuation *cross-section*. In practical terms the units employed are often $L\,mol^{-1}\,m^{-1}$ (i.e. $dm^3\,mol^{-1}\,m^{-1}$). Writing 1 L as 0.001 m^3, the molar attenuation coefficient can be expressed as 0.001 $m^2\,mol^{-1}$ or 1 $m^2\,mmol^{-1}$.

The dimensionless product $A = \varepsilon cx$ is called the *absorbance* (sometimes the *optical density*) and the ratio I_x/I_o is the *transmittance* or *transmissivity* T. Thus, we can write:

$$\log T = -A$$

The Beer–Lambert law finds use in the measurement of concentrations. For example, the clarity or otherwise of polluted air is often measured by comparing the irradiance of light at a certain time with the irradiance on a fine day.

These interactions with a material can be expressed thus:

$$I_o = I_r + I_s + I_a + I_t$$

where I_o is the incident irradiance, I_r is the amount reflected, I_s is the amount scattered, I_a is the amount absorbed and I_t is the amount transmitted, or as:

$$1 = R + S + A + T$$

where R is the *fraction* of light reflected, S is the *fraction* of light scattered, A is the *fraction* of light absorbed and T is the *fraction* of light transmitted and the quantities measured are the appropriate irradiance values. In good-quality optical materials the amount of light scattered and absorbed is small and it is often adequate to write:

$$I_o = I_r + I_t$$

or

$$1 = R + T$$

In a pure liquid the Beer–Lambert law is often written in the form:

$$\log\left(\frac{I_x}{I_o}\right) = -ax$$

where $a\,(m^{-1}) = \varepsilon c$ is the *molar (decadic) attenuation (or absorption) coefficient*. The absorption will be due to molecular or atomic processes taking place in the pure medium.

[6] Chemists frequently use the term molarity for concentration in mol L^{-1}, given the symbol M. Thus, ε is given the units $M^{-1}\,m^{-1}$, or more often $M^{-1}\,cm^{-1}$. To convert values of ε in $M^{-1}\,cm^{-1}$ to $M^{-1}\,m^{-1}$, multiply the value by 100.

Figure 1.17 ***Mediaeval stained glass window in Gloucester Cathedral, viewed from inside the building. [Reproduced with permission from Gloucester Cathedral www.gloucestercathedral.org.uk]***

1.14 Subtractive Coloration

Absorption has been used for many centuries to produce colour. For example, the colour of stained glass and the colours seen in ordinary colour filters are examples of colour production in this way (Figure 1.17). The colours perceived by the eye in which absorption and selective reflection or transmission are important are said to be due to *subtractive colour mixing*. For example, the photosensitive pigments in green leaves preferentially absorb red and blue light and reflect more of the green component of the incident white light. Similarly, colour filters absorb some wavelengths strongly and transmit the remainder. Figure 1.18a shows the fraction of light transmitted as a function of wavelength for a commercial glass colour filter. The range of the visible spectrum is indicated above the transmittance curve. The filter absorbs red light strongly and transmits violet and blue–green light (Figure 1.18b). If the filter is held up to the light it will look blue–green. When it is viewed in reflected light it appears dark, as red light is absorbed and blue–green passes through the film. This is the reason why stained glass windows in medieval churches look impressive when viewed inside the building, with light transmitted through the glass, yet often look dull when viewed from outside the building in reflected light.

By analogy with additive coloration, one would expect to be able to combine three *subtractive primary colours* to produce the whole range of subtractive colours. These subtractive primary colours are: *cyan*, which absorbs red and transmits blue and green; *magenta*, which absorbs green and transmits blue and red; and *yellow*, which absorbs blue and transmits green and red. If the three subtractive primaries are mixed in equal amounts we obtain *black*, as one primary will absorb red, one will absorb green and one will absorb blue, thus removing the whole of the visible spectrum. Colour construction using these three subtractive primary colours is described as employing the *CMY model*, where the letters simply represent the initial letters of the colorants.

If the wavelength range of light absorbed is rather small, then the colour remaining is called the *complementary* colour to that absorbed (Table 1.4). It is seen that the additive and subtractive primary colours are complementary colours.

Colour printers use cyan, yellow and magenta dyes to produce the coloured images. These dyes are deposited upon white paper and absorb the appropriate subtractive primary colour. White light reflected from the dyes is depleted in these colours and yields the appropriate toned image by subtractive coloration. Although the

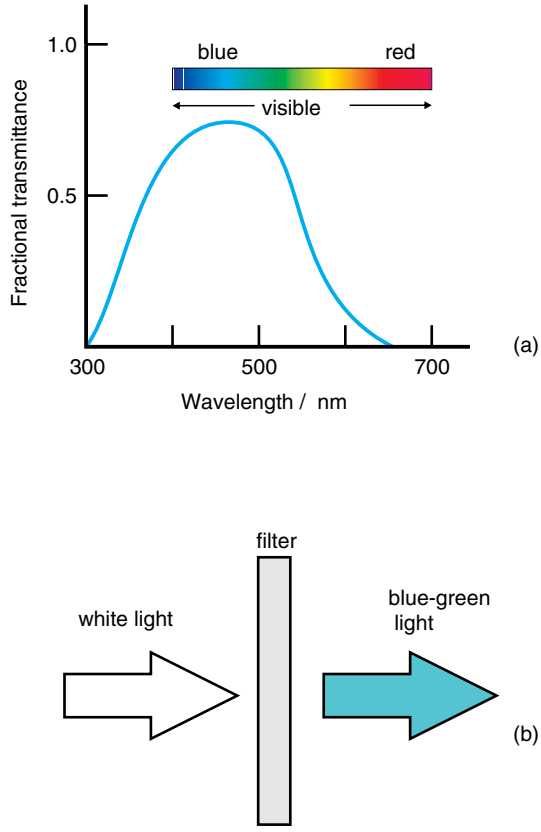


Figure 1.18 ***(a) The fractional transmittance of a commercial blue colour filter. About three-quarters of the blue light incident on the filter is transmitted, but most red light is absorbed. (b) When the filter is viewed in transmitted white light it will appear blue–green***

Table 1.4 ***Complementary colours***

Wavelength/nm	Colour absorbed	Complementary colour
400–435	Violet	Yellow–green
435–480	Blue[a]	Yellow[b]
480–490	Blue–green	Orange
490–500	Green–blue	Red
500–560	Green[a]	Magenta[b]
560–580	Yellow–green	Violet
580–595	Yellow	Blue
595–605	Orange	Blue–green
605–700	Red[a]	Cyan[b]

[a] Additive primary colours.
[b] Subtractive primary colours.

overlap of cyan, yellow and magenta produces black, this tone is often not dark enough for many representations. Printers, therefore, often add black to the trio. This system of colour production is known as the *CYMK model* of colour formation, where the letter K stands for the black component. Although these four colours are satisfactory for many colour printing applications, more hues, intermediate between the CYMK set, are used to obtain more accurate colour rendition, in, for example, high-quality art reproductions.

1.15 Electronic 'Paper'

Paper is an extremely convenient way of displaying information using subtractive coloration, but once a page is printed it is permanent. Electronic paper, with the advantages of a printed page, but the flexibility of electronic erase and rewrite has been pursued for over 30 years. As of 2000, e-book readers, which are rigid units displaying one paper-like page at a time, have been increasingly available.

There are two aspects to electronic paper. In the first, electronic 'ink' must be developed that will retain the display indefinitely but is erasable at will. At least for black-and-white displays this has been accomplished. The second is the production of a flexible page that can support the electronic circuitry needed to drive the display. In this section the characteristics of the 'ink' are the main focus of attention, as this is the aspect that impinges upon the topic of colour.

The first electronic paper, using the *Gyricon* process, consisted of small polyethylene spheres of approximately 90 μm diameter, coloured white on one hemisphere and black on the other. The white part held a positive charge and the black portion a negative charge, due to additives to the polymers used. These spheres were embedded in a transparent silicone film and the sheets were immersed in clear oil. This penetrated the sheets and coated the beads, so that they were effectively encapsulated in a bubble of oil. The application of a negative charge to an electrode on the surface will attract the positively charged white side facing one side of the 'page'. In this way pixels of the display could be made black or white at will (Figure 1.19a). Rearranging the applied voltage allows the image to be erased and rewritten.

The *e-ink* process is rather similar but uses the movement of charged particles in an electric field, the process of *electrophoresis*. Once again, small polymer capsules containing submicrometre particles of white titanium dioxide, TiO_2, holding a negative charge due to appropriate surfactants, and black particles holding a positive charge are central to the system. The microspheres also contain a nonviscous liquid and are embedded in a clear plastic film. A charge applied to surface electrodes will attract white or black particles depending upon the polarity of the electrodes. Reversal of the charge on the electrodes reverses the particles that are attracted and the area will swap colour (Figure 1.19b). Erasure and rewriting is carried out as before.

Naturally, the use of polymer spheres to contain the black and white particles is not mandatory, and any cell structure could be used. The device also becomes simpler if the black particles are replaced by a dark-coloured fluid. The white particles are then the only active species present. When attracted to a surface the appropriate pixel looks white and when not attracted the dark fluid is seen.

The colour of the pixels is due to absorption and scattering. Titanium dioxide is a well-known white scatterer (Section 5.7) and the dark colour is simply absorption of the incident light by the dye present. The system can be made into a colour display by putting red-, green- and blue-coloured filters in front of the electrodes (Figure 1.19c). A white pixel will now become a coloured subpixel corresponding to one of the colours.

The electrodes used to control the display can be a simple passive array of vertical strips on one face of the device and horizontal strips on the reverse face. Application of charges to appropriate columns and rows ensures that pixels can be made black or white as required. The active matrix method of control, as used in flat-panel television, in which a transistor controls each pixel, is also widely used. An advantage of these displays is that once the page has been created, no further electrical input is needed until the page is rewritten. Of course,

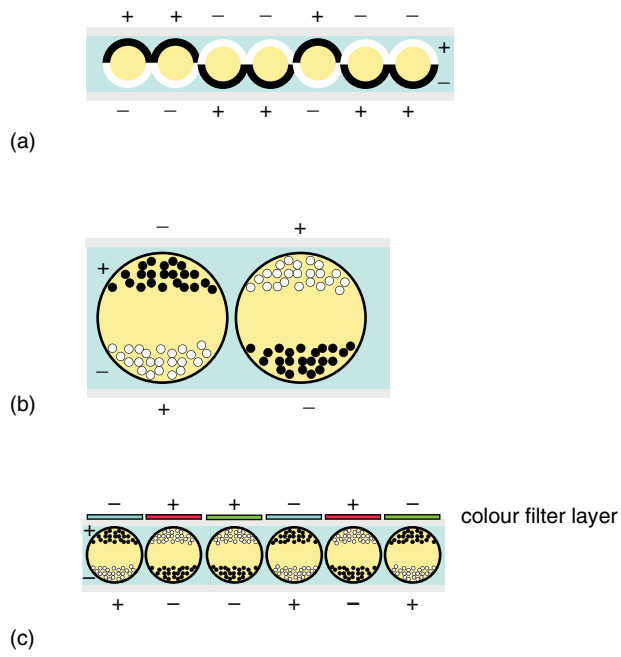


Figure 1.19 ***Electronic paper displays: (a) the rotating sphere Gyracon system; (b) the electrophoretic e-ink system; (c) coloured filters allow for a full colour display to be achieved***

the requirements are less demanding for a rigid e-book than for a portable and flexible sheet-like page, which has still to achieve widespread commercialization.

1.16 Appearance and Transparency

Scattering and absorption give rise to the world of colour around us (Figure 1.20). Even small changes in the relative amounts of each wavelength band present in a light beam will contribute significantly to colour and appearance. A striking example of this is the blue sky. Blue sky is so coloured because of light scattering (see Chapter 5). However, blue sky contains all of the wavelengths of the spectrum – something easily proved by passing the light through a prism. The sky appears blue because the balance in the various colours has been tipped slightly in favour of the blue end of the visible spectrum.

The appearance of an object will depend on a number of factors, especially on roughness and surface texture. These will alter the reflectivity of the surface considerably. If the surface is smooth then the reflection is said to be *specular*, while if the surface is rough then the reflection is *diffuse* (Figure 1.21a). The diffuse reflection component increases with surface roughness at the expense of the specular component, so that a finely ground powder shows only diffuse reflection. The *gloss* of a surface is a measure of the relative amounts of diffuse to specular reflections. Glossy surfaces have a large specular component. As well as diffuse reflection, subsurface scattering is of considerable importance in modifying the appearance of a surface. This is particularly so when the surface is composed of layers with different optical properties, such as skin. Controlling these forms of scattering and reflection are of great importance to the cosmetics industry, and imitating them is vital to both artists and personnel involved in the representation of skin tones in computer-generated images.

Figure 1.20 ***A moorland scene displaying colours due to scattering (blue sky), reflection (the blue stream) and absorption (the green–browns of grass and soil)***

Closely related to this is the property of transparency or invisibility in animals. Invisibility confers obvious advantages to both predator and prey in the living world. It is not surprising, therefore, that many marine animals almost achieve this object. To attain invisibility, the interactions of light with a material described above must be bypassed. That is to say, reflection and refraction at surfaces and scattering and absorption from internal centres need to be suppressed.

Reflection and refraction at the surface of the animal's body can be substantially reduced by making the refractive indices on both sides of the boundary the same (Chapters 2 and 3). For many marine animals, including numerous species of zooplankton, jelly fish and similar creatures, the inner body fluid is essentially watery, and reflection and refraction are virtually eliminated. This alone serves to make the animals virtually invisible.[7] However, any inhomogeneities in the tissues and membranes will act so as to scatter light and render the animal visible to a greater or lesser degree. Air pockets are particularly problematic. For instance, small bubbles of air in water are easily visible in ordinary light and shine like silver spheres. Pigments, which colour by absorption, cannot be totally avoided. The photoreceptors of the eye are pigments which absorb visible radiation (Section 1.10). Similarly, the prey of the animal, once consumed, will be visible in the gut, unless this matches the surroundings in both transparency and refractive index. Thus, although many marine animals can be extremely difficult to spot, and may be termed invisible for all practical purposes, some traces will remain visible.

Solids and liquids cannot be manipulated so that their refractive index matches that of air, but can be made with a refractive index that matches that of a liquid. A transparent solid immersed in a liquid of the same refractive index will be invisible. For many solids, internal surfaces are a major cause of loss of transparency.

[7] This is the basis of the famous story *The Invisible Man* by H. G. Wells, published in 1897.

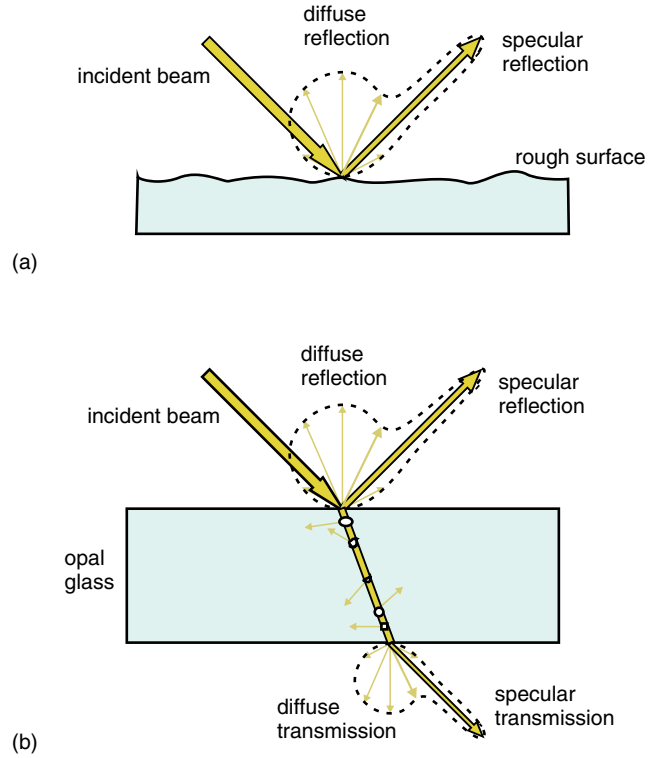


Figure 1.21 ***(a) Reflection of light from a rough surface consists of two components, diffuse reflection and specular reflection. The ratio of diffuse reflection to specular reflection increases as the surface roughness increases. The ratio is an indication of surface gloss. (b) The passage of light through a translucent material containing many scattering centres gives rise to both surface reflection and transmitted light with diffuse and specular components***

Glass, the best known of transparent solids, is, in effect, in the liquid state, and no internal boundaries occur. However, it is relatively simple to cause a glass to crystallize and the ensuing tiny crystallites act as scattering centres. The resulting scattering, which can contain diffuse and specular components, renders the material nontransparent although the solid transmits a certain amount of the incident light. Such materials are termed *translucent*. The light emerging from a translucent material will also contain a diffuse and specular component (Figure 1.21b). Translucency is a desirable property of fine porcelain, which consists of crystallites of mullite ($\sim Al_6Si_2O_{13}$) dispersed throughout a glassy matrix. More opaque glasses, such as *opal* glasses, are deliberately made with large numbers of scattering centres present. The resultant scattering renders the material white because the scattering affects all wavelengths of the incident light equally.

Similarly, most plastics as fabricated are noncrystalline and have no internal boundaries, rendering them transparent. If these contain impurities, inhomogeneities or polymer crystallites they become translucent and take on a slightly milky appearance.

Non-glassy solids are mainly composed of polycrystalline aggregates or ‘grains’. The grain boundaries between each crystallite scatter light, and any impurity phases that exist in the matrix or the grain boundary regions enhance this effect, so that polycrystalline solids are invariably opaque. However, it is of considerable benefit to make such materials transparent. This can be achieved by careful processing that achieves a high

density, so that internal pores and bubbles of gas are eliminated, and produces a solid composed of small, evenly sized crystallites with no impurity grain-boundary phases present. In this way, transparent refractory ceramics such as alumina (Al_2O_3), aluminium oxynitride ($\sim Al_{23}O_{27}N_5$) and SiAlONs (materials occurring in the SiO_2–Al_2O_3–Si_4N_3–AlN system) have been produced. These and similar materials have uses as lamp housings and windows which need to be stable in air to temperatures of 2000 °C or more. In addition, these are hard and durable ceramics and are favoured for applications such as specialist optical windows and domes where resistance to abrasion and erosion are important selection criteria.

Appendix A1.1 Definitions, Units and Conversion Factors

A1.1.1 Constants, conversion factors and energy

Constants

The important constants for light are:

velocity of light in vacuum c	$2.99792 \times 10^{8}\,\mathrm{m\,s^{-1}}$
Planck constant h	$6.62608 \times 10^{-34}\,\mathrm{J\,s}$
Boltzmann constant k_B	$1.38066 \times 10^{-23}\,\mathrm{J\,K^{-1}}$

Conversion Factors

$E\,(\mathrm{J}) = E\,(\mathrm{eV}) \times 1.60219 \times 10^{-19}$
$E\,(\mathrm{J}) = E\,(\mathrm{cm^{-1}}) \times 1.98645 \times 10^{-23}$
$E\,(\mathrm{eV}) = E\,(\mathrm{cm^{-1}}) \times 1.23987 \times 10^{-4}$
$\lambda\,(\mathrm{\AA}) = \lambda\,(\mathrm{nm}) \times 10$
$\lambda\,(\mathrm{nm}) = \lambda\,(\mu\mathrm{m}) \times 1000$
$\lambda\,(\mathrm{nm}) = 1239.9/E\,(\mathrm{eV})$
$\lambda\,(\mathrm{nm}) = 198\,645 \times 10^{-21}/E\,(\mathrm{J})$
$\lambda\,(\mathrm{nm}) = 10^{7}/\bar{\nu}\,(\mathrm{cm^{-1}})$

Energy

The SI unit of energy is the joule (J). A wide variety of energy units are used in the literature connected with light apart from the joule. A common nonstandard unit of energy in atomic work is the electron volt (eV). Spectroscopy often uses energy values given in $\mathrm{cm^{-1}}$. These are not energy values at all really, but E/hc values. To convert 'energy' in $\mathrm{cm^{-1}}$ to joules, multiply the value in $\mathrm{cm^{-1}}$ by h (J s) and c ($\mathrm{cm\,s^{-1}}$); see Conversion Factors above.

A1.1.2 Waves

Waves

The wave equation, Equation 1.2, is a one-dimensional continuous harmonic wave that represents the electric field vector **E**:

$$\mathcal{E} = \mathcal{E}_0 \cos[(2\pi/\lambda)(x - vt)] \qquad \text{(A1.1)}$$

$\mathcal{E}$ is the magnitude of the electric field vector at position x and time t, $\mathcal{E}_0$ is the *amplitude* of the wave, λ is the wavelength of the wave, $[(2\pi/\lambda)(x-vt)]$ is the *phase* of the wave (radians), v is the speed at which any point on the wave, say a peak or a trough, travels in the positive x direction, and is called the *phase speed* or *phase velocity*. The velocity of an electromagnetic wave in vacuum (the speed of light) has the symbol c.

The relationships described below allow the equation to be written in other equivalent forms. Those most frequently met are:

1. a standing (non-travelling) wave:

$$\mathcal{E} = \mathcal{E}_0 \cos[(2\pi/\lambda)x]$$

2. a wave travelling in the negative x direction:

$$\mathcal{E} = \mathcal{E}_0 \cos[(2\pi/\lambda)(x + vt)]$$

3. a wave travelling in the positive x direction, where ω is the angular frequency:

$$\mathcal{E} = \mathcal{E}_0 \cos[(2\pi x/\lambda) - \omega t]$$

4. a wave travelling in the negative x direction, where ω is the angular frequency:

$$\mathcal{E} = \mathcal{E}_0 \cos[(2\pi x/\lambda) + \omega t]$$

Frequency

The *temporal frequency* ν of light (the number of waves that pass a point per second) has units of cycles per second, hertz (Hz) or s^{-1}. It is usually just called the frequency. The reciprocal of the temporal frequency, $1/\nu$, is the *temporal period* τ, which is the amount of time for a complete wave oscillation to pass a stationary observer at a fixed value of x.

The *angular (temporal) frequency* ω of a wave is given by:

$$\omega = \frac{2\pi}{\tau} = 2\pi\nu \qquad \text{units: rad } s^{-1}$$

Using the relationship $c = \lambda\nu$ gives $\omega = 2\pi c/\lambda$.

Wavelength

The wavelength λ. is the *spatial period* of the wave – the distance over which the wave subsequently repeats itself. In wavelength designations concerning light, nanometre (nm) is the preferred unit, but a commonly used nonstandard unit, especially in X-ray diffraction, is the ångström (Å), 10^{-10} m. To convert between units, see Conversion factors above.

Wavelength and Energy

Planck's law ($E = h\nu = hc/\lambda = h\omega/2\pi = \hbar\omega$) relates energy to wavelength. To convert between units, see Conversion factors above.

Wavenumber

The wavenumber is the reciprocal of the *spatial period* of the wave (the number of waves per unit length) and so is the reciprocal of the wavelength, $1/\lambda$. The wavenumber is given the symbol σ when the light traverses a transparent medium or $\bar{\nu}$ when in a vacuum. In spectroscopy it is often given units of cm^{-1}.

$$\sigma(\bar{\nu}) = \frac{1}{\lambda}$$

Using Planck's law, in a vacuum:

$$E = h\nu = \frac{hc}{\lambda} = hc\bar{\nu}$$

$$\frac{E}{hc} = \bar{\nu}$$

Similar equations can be written for light in a substance, by replacing c by the velocity v in the medium and replacing $\bar{\nu}$ with σ.

Spectroscopy often uses wavenumbers (units: cm^{-1}). To convert these to wavelength (units: nm):

$$\lambda\,(\mathrm{nm}) = \frac{10^7}{\bar{\nu}\,(\mathrm{cm}^{-1})}$$

In physics, the magnitude of the propagation vector or wave vector **k** is called the propagation number or wavenumber, given by $k = 2\pi/\lambda$. (By analogy with temporal frequency and temporal angular frequency, it might be better to call $k = 2\pi/\lambda$ the angular spatial frequency to avoid confusion.) Additionally, physics also uses $k = 1/\lambda$ for the wavenumber, omitting the factor 2π. To avoid confusion, the wave vector will be written as $2\pi/\lambda$ or $1/\lambda$ rather than k.

A1.1.3 SI units associated with radiation and light

There are two parallel sets of units in use for the measurement of radiation and light. *Photometric* units measure the *perception* of a light as it appears to the eye of an average observer. Radiometric units measure the *amount* of electromagnetic radiation, including light, in terms of absolute quantities, without any reference to the eye. The difference can be understood by considering the light output of four small light-emitting diodes (LEDs), one infrared, one deep red, with an emission at 670 nm, one green with an emission at 555 nm and one blue with an emission at 490 nm. These may all emit exactly the same absolute power (measured in radiometric units, say 5 mW), but the green light will appear 'brighter' than the other two visible LEDs because the eye is more sensitive to green than red or blue. Calculation shows that the visible outputs will be: green, approximately 3.4 lm; blue, approximately 0.75 lm; red, approximately 0.1 lm; infrared, 0 lm. The green LED will appear about 31 times brighter than the red LED, and the blue LED about seven times brighter than the red LED. The infrared-emitting LED will be invisible to the eye and will not register at all in terms of photometric units, although it still emits the same amount of power as the visible ones.

Clearly, photometric units are of importance in the design of displays and lighting, whereas radiometric units are of more importance when comparing the energy requirements of the same structures. Although the two sets of units are analogous, as set out in Table A1.1, because they measure different aspects of light, they cannot be trivially interchanged in this regime.

Table A1.1 ***Units used in radiometry and photometry***

Radiometry			Photometry		
Name, symbol	Comments	Units	Name, symbol	Comments	Units
Radiant power, radiant flux, Φ, P	Rate of flow of energy emitted by a source	W	Luminous power, luminous flux, F, Φ_v	Rate of flow of luminous energy emitted by a source	lumen (lm) (cd sr)
Radiant intensity $I = d\Phi/d\Omega$	The power of an emitting source per unit solid angle	$W\,sr^{-1}$	Luminous intensity, I_v	Light emitted from a source per unit solid angle; SI base unit	candela (cd) $= lm\,sr^{-1}$
Radiance, $L = d^2\Phi/(dA\,d\Omega)$	Radiant power per unit area per unit solid angle. Radiant intensity of a radiating source per unit surface area.	$W\,m^{-2}\,sr^{-1}$	Luminance, L_v	A measure of 'brightness'; luminous intensity of a light-emitting source per unit area of source; may vary over the source surface	cd m^{-2} (nit!)
Irradiance E (I) $= d\Phi/dA$	Radiant power incident upon a unit area of a surface.	$W\,m^{-2}$	Illuminance, E_v	A measure of illumination; the luminous flux falling on a surface per unit area	lux ($lm\,m^{-2}$)
(Radiant) Exitance $M = d\Phi/dA$	Radiant power emitted by a surface per unit area	$W\,m^{-2}$	Luminous exitance, M_v	Luminous flux emitted from a surface	lux ($lm\,m^{-2}$)

Flux is the amount of something flowing through a specified surface per unit time.
Luminous flux or *luminous power F*, unit lumen (lm): 1 lm is the amount of luminous flux passing in 1 s through a unit solid angle emitted by a point source of 1 cd. The total luminous flux of such a point source is 4π lumens.
Luminous intensity I_v, unit candela (cd): 1 cd is the photometric measurement of luminous intensity in a given direction of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and that has a radiant intensity in that direction of (1/683) watts per steradian. One square metre of a black body at 2042 K emits 600 000 cd.
Radiance L, units $W\,m^{-2}\,sr^{-1}$: the *radiance* is the incoming radiation collected from a small angle of surroundings (measured in steradians) as if, for example, the detector is at the bottom of a tube. The units of radiance are energy per unit area per unit solid angle, $W\,m^{-2}\,sr^{-1}$. The radiance is direction sensitive – the value recorded depends upon the direction in which the tube is pointing.
Irradiance E or *I*, unit $W\,m^{-2}$: the radiometric term *irradiance* is the total energy that a detector 'sees' from a hemisphere of surroundings. The preferred symbol for irradiance is *E*, but because of the use of *E* for energy, and of $\mathcal{E}$ for the amplitude of an electromagnetic wave, it is less confusing here to use the symbol *I*.
Illuminance E_v, unit lux (lx): this is the photometric analogue of irradiance, being the total luminous flux incident upon unit area of a surface, with units of lux $= lm\,m^{-2}$. The photometric term illuminance has replaced the term brightness.
Radiant exitance or *radiant emittance M*, unit $W\,m^{-2}$: the amount of electromagnetic radiation leaving a surface is described by the radiometric term (*radiant*) *exitance*. The exitance is the opposite of the irradiance, as it measures the total energy emitted by a surface into a hemisphere of the surroundings. The exitance has the same units as irradiance.
Luminous exitance or *luminous emittance* M_v, unit lux (lx): this is the photometric analogue of the radiometric radiant exitance.
Spectral units. These give the distribution of the quantity under discussion with respect to the wavelength or frequency of the radiation. For example, the *spectral irradiance* takes the form irradiance per unit wavelength, written E_λ, or irradiance per unit frequency E_ν. The units of spectral quantities must contain the units of wavelength or frequency as appropriate. Thus, the units of spectral irradiance are $W\,m^{-2}\,m^{-1} = W\,m^{-3}$.

Further Reading

The following five books contain a vast amount of material of relevance to the whole of this book. The two books by Bohren present material in a nonmathematical format and will repay repeated reading.
E. Hecht, Optics, 4th edition, Addison-Wesley, San Francisco, 2002.
K. Nassau, *The Physics and Chemistry of Color*, 2nd edition, Wiley-Interscience, New York, 2001, Chapters 1 and 2 and Appendix A.
C. F. Bohren, *Clouds in a Glass of Beer*, Dover, New York, 2001 (originally published by John Wiley and Sons, Inc., New York, 1987).
C. F. Bohren, *What Light Through Yonder Window Breaks*? Dover, New York, 2006 (originally published by John Wiley and Sons, Inc., New York, 1991).
B. E. E. Saleh, M. C. Teich, *Fundamentals of Photonics*, John Wiley and Sons, Inc., New York, 1991.

Colour, from the point of view of artist's pigments, is the subject of
V. Findlay, *Colour; Travels through the Paintbox*, Folio, London, 2009.
Goethe's *Theory of Colour*, written 1808, published 1810, gives an interesting historical view of colour. It is included in
D. Miller, (ed. and trans.), *Goethe: Scientific Studies*, Suhrkamp, New York, 1988 (Goethe Edition Vol. XII).

A clear discussion of radiometric and photometric units is given by J. M. Palmer (2003) to be found at http://www.optics.arizona.edu/Palmer/rpfaq/rpfaq.htm. Also see
C. F. Bohren, E. C. Clothiaux, *Fundamentals of Atmospheric Radiation*, Wiley–VCH, Weinheim, 2006, Chapter 4.

The electromagnetic theory of radiation is clearly set out in
M. Kotlarchyk, Electromagnetic radiation and interactions with matter, in *Encyclopedia of Imaging Science and Technology*, J. P. Hornak (ed.), Wiley-Interscience, 2002.
N. Braithwaite (ed.), Electromagnetism, Book 3, *Electromagnetic Waves*, The Open University, Milton Keynes, 2006.

Light described in terms of quantum electrodynamics is explained lucidly and nonmathematically in
R. P. Feynman, *QED: The Strange Theory of Light and Matter*, Princeton University Press, Princeton, 1985.

The fascinating history of the theories of light is given by
G. N. Cantor, *Optics after Newton*, Manchester University Press, Manchester, 1983.

A comparison between the wave and particle explanation of the photoelectron effect and profound discussions of the relationship between particle and wave theories of atomic physics are given by
D. Bohm, Quantum Theory, Prentice-Hall, Englewood Cliffs, NJ, 1951.

A detailed discussion of the solar spectrum and related topics is given by
C. F. Bohren, E. E. Clothiaux, *Fundamentals of Atmospheric Radiation*, Wiley–VCH, Weinheim, 2006, Chapter 1.
D. K. Lynch, W. Livingston, *Color in Light and Nature*, Cambridge University Press, Cambridge, 1995, Chapters 2 and 7.

The (coincidental) relationship between the sensitivity of the human eye and the solar spectrum is discussed by
B. H. Stoffer, D. K. Lynch, *Am. J. Phys.* **67**, 946–958 (1999).

For discussions of the molecular basis of vision, see
D. M. Hunt, L. S. Carvalho, J. A. Cowing, W. L. Davies, *Phil. Trans. R. Soc. Lond. Ser*. B **364**, 2941–2945 (2009).

For a discussion of bacteriorhodopsin, see
J. Whitford, *Proteins, Structure and Function*, John Wiley and Sons, Ltd, Chichester, 2005, pp. 114–119.

The evolution of primate colour vision is detailed by
G. H. Jacobs, J. Nathans, *Sci. Am.* **300** (April), 40–47 (2009).

The complexity of the retina and much information about vision in specialist circumstances is to be found by consulting
S. Temple, N. S. Hart, N. J. Marshall, S. Collin, *Proc. R. Soc. Lond. Ser.* B **277**, 2607–2615 (2010).

Many aspects of vision and the interpretation of visual images, including optical illusions, are detailed in the series of articles by
J. C. Russ, Seeing the scientific image, *Proc. R. Microscop. Soc.* **39** (2004); Part 1: 97–114; Part 2: 179–193; Part 3: 267–281.

The complexities of analysing colour and descriptions of the construction and use of chromaticity diagrams are detailed in the following sources. A large number of articles concerning colour, colour theory, colour systems and colour spaces will be found on Wikipedia (http://en.wikipedia.org/wiki/). Up-to-date details of colour and colour reproduction will be found in the Instructions and Help functions for computer drawing and photograph editing software, many of which are available: typically as in manuals for Nikon Coolscan, Coreldraw, Photoshop and so on. Other sources are
http://www.efg2.com (this site has programs for the display and representation of chromaticity diagrams, colour mixing and many other topics of relevance to the material in this chapter).
R. McDonald, *Colour Physics for Industry*, 2nd edition, Society of Dyers and Colourists, Bradford, UK, 1997.
F. Grum, C. J. Bartleson, *Colour Measurement*, Academic Press, New York, 1980.
R. Jackson, L. MacDonald, K. Freeman, *Computer Generated Colour*, John Wiley and Sons, Ltd, Chichester, 1994.

Other interesting sources on colour and colour perception are as follows:
Animal colour patterns
J. A. Endler, J. Linn. Soc. **41**, 315–352 (1990).

Reviews of transparency in biological tissues
S. Johnsen, E. A. Widder, *J. Theor. Biol.* **199**, 181–198 (1999).
S. Johnsen, *Sci. Am.* **282** (February), 62–71 (2000).

Texture and computer modelling of surfaces
J. Dorsey, P. Hanrahan, *Sci. Am.* **282** (February), 46–53 (2000) and references cited therein.

There are a number of demonstrations of relevance to this chapter, including diffuse versus specular reflection, available at http://demonstrations.wolfram.com/index.html.