Chapter 1

Introduction to Light and Optical Systems

Optical systems, which provide much convenience to our lives and industries, manipulate light to satisfy the particular requirements of end users. For end users and optical engineers, understanding optical systems is of fundamental importance for using, designing, or manufacturing optical systems. In order to better understand optical systems, it is essential to be familiar with the behavior and properties of light. In history, the behavior and properties of light have been gradually discovered in the process of explaining some optical phenomena, and validated by many experiments. However, the nature of light is still difficult for many people to understand, especially for beginner students of optics. In this chapter, we give readers a very brief description of light and optical systems. The question "What is light?" is firstly addressed. The generation of light and the three basic theories of light—geometrical optics, wave optics, and quantum optics—are then briefly discussed, after which, the concept of statistical optics is mentioned. Finally, an overview of optical systems is presented, and some examples are also provided to give readers a sense of what is entailed in such systems.

1.1 What Is Light?

Light is familiar, but also mysterious to human beings. We are accustomed to experiencing sunlight and manmade light every day. However, the nature of light is extremely foreign to us. What is light exactly? For several centuries, people have been devoting much effort to answering this question. At first, people thought that light consists of massive corpuscles that obey Newtonian mechanics, which can easily explain the phenomena of rectilinear propagation, reflection, and refraction of light. However, this theory cannot explain the phenomena of diffraction, interference, and polarization of light. For these phenomena, the wave nature of light, which was proposed later on, is more pertinent. Finally, human beings found that within a certain wavelength region, light is nothing but electromagnetic waves obeying Maxwellian electromagnetic theory. When people found the photoelectric effect, i.e., that many metals emit electrons when light shines on them, neither of the two theories above could explain this phenomenon any longer. Einstein proposed that a beam of light is not a wave, but rather a collection of photons; thus, the quantum theory of light emerged. Currently, people think that light has the attributes of a wave–particle duality.

1.1.1 Light as electromagnetic waves

On the basis of previous research on electricity and magnetism, Maxwell summarized a set of equations, i.e., Maxwell's equations, in rigorous mathematical form. This set of equations predicted the existence of electromagnetic waves propagating in vacuum with the speed of $3 \times 10^8 \mathrm{m/s}$, the same as the speed of light. This led Maxwell to surmise that light is a kind of electromagnetic wave. Gradually, light as electromagnetic waves was verified and accepted. The wave attribute of light can be observed in many experiments, such as the single-slit diffraction and Young's double-slit interference experiments. A detailed discussion of the wave characteristics of light will be presented in Chapter 3.

However, although light is a kind of electromagnetic wave with all of the characteristics of waves, it is different from mechanical waves, such as sound and water waves. For example, light can propagate in both vacuum and media, while mechanical waves can propagate only in media.

1.1.2 Light as particles: photons

The theory that deals with light as a kind of electromagnetic wave, namely, wave optics, successfully explains the optical phenomena of diffraction, interference, and polarization of light. However, wave optics is unable to explain phenomena involving the interaction of light and matter, such as the photoelectric effect.

The photoelectric effect was first observed by Hertz in 1887. As shown in Fig. 1.1, when blue light is incident on the surface of caesium, electrons are emitted from that surface. These emitted electrons are also called photoelectrons.

According to classical electromagnetic radiation theory, the photoelectric effect is due to the transfer of energy from light incident on caesium to electrons in it. From this perspective, two predictions can be easily obtained: (1) the intensity of the incoming light completely determines whether photoelectrons are released or not as well as the speed of the photoelectron; (2) the brighter the incoming light the more photoelectrons are emitted, and the brighter the incoming light the faster the photoelectrons are emitted. However, experimental results are contradictory to these predictions in two points: (1) blue light can always release electrons from caesium, no matter

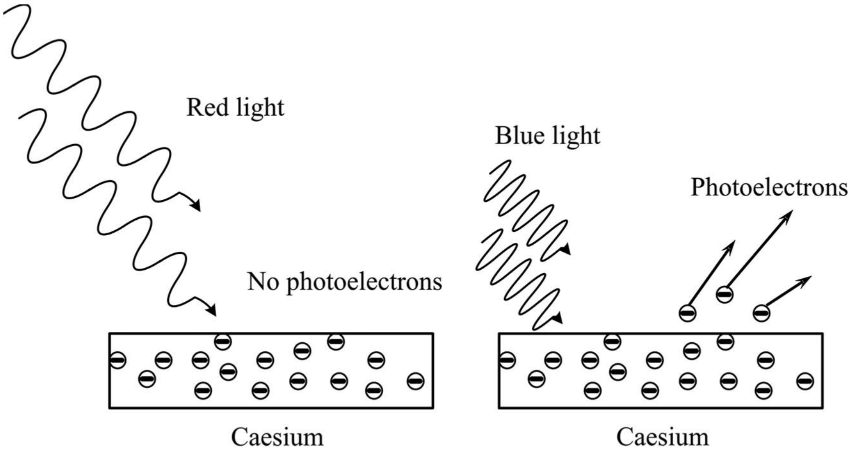

Figure 1.1 Diagram of the photoelectric effect.

how dim it is, and not even one photoelectron is emitted from caesium by exposing it to red light, regardless of how bright the red light is (assuming that nonlinear optical effects do not occur); (b) the speed of an emitted photoelectron is completely determined by the frequency of the light that releases the photoelectron. These contradictions between experimental results and the predictions from the electromagnetic theory indicate that a new theory should be invoked to explain this effect.

In 1905, Einstein explained the photoelectric effect using the concept of photons. In Einstein's theory, light is a collection of photons. Each photon has a fixed amount of energy, which only depends on the frequency of light. When photons fall on the surface of a metal, the energy of a photon is split into two parts if light of this frequency can release electrons in this metal. One part is used for releasing an electron from the metal, and the other part transforms into the kinetic energy of the released electron. According to the law of conservation of energy, the process of energy conversion in the photoelectric effect can be expressed as

$$h\nu = \frac{1}{2}m\upsilon^2 + w, \tag{1.1}$$

where h is Planck's constant, ν is the frequency of light, m is the mass of the electron, υ is the speed of the photoelectron, and w is the work function of the metal, which represents the smallest energy needed for electrons to escape from the constraints of the metal. According to Eq. (1.1), both the release of the photoelectron from the metal by light and the speed of the photoelectron are determined by the frequency of the incoming light, irrespective of the intensity of light incident on the metal.

The concept of light as photons successfully explains the photoelectric effect and is confirmed by more and more experimental results. This reveals that the particle nature is another aspect of light.

1.1.3 Wave–particle duality of light

In the quantum theory of light, the wave nature and the particle nature of light are unified by invoking the wave–particle duality of photons. For example, the connection between the wave and particle natures of light is furnished by Planck's constant h.[1] In view of the particle nature of light, a photon has a packet of energy E,

$$E = h\nu = \frac{h}{T}, \tag{1.2}$$

and a momentum p,

$$p = \frac{h}{\lambda}. \tag{1.3}$$

Obviously, both the energy and the momentum of a photon are expressed by the wave attribute of light in terms of T (the period of the light wave) or λ (the wavelength of the light wave), respectively. By rearranging Eqs. (1.2) and (1.3), Planck's constant can be expressed as

$$h = ET = p\lambda. \tag{1.4}$$

Note that E and p express the attribute of a particle of light, while T and λ characterize the attribute of a wave of light. This means that Planck's constant can be expressed as the product of the particle and wave attributes of light, i.e., ET or $p\lambda$. If light exhibits more wave nature, the particle nature of light will manifest less, and vice versa. Furthermore, it should be pointed out that in the quantum theory of light, the wave nature of photons involves a probability wave, not a classical wave.

In practical issues, when light interacts with matter—such as in the generation of light and the detection of light with photoelectric detectors, which will be discussed in Chapter 4—it displays the particle characteristics. On the other hand, light exhibits more wave characteristics in situations involving diffraction of light and the interaction between light waves, such as the interference of light.

1.2 How Do Light Sources Produce Light?

The generation of light from light sources can be roughly explained by the classical electromagnetic radiation theory or, more precisely, by the quantum theory of light. These two explanations are illustrated in the following subsections.

1.2.1 Explanation by electromagnetic wave theory

Matter is made of atoms, and each atom is composed of a nucleus and some electrons around the nucleus. When an amount of energy is continuously injected into matter, the temperature of the matter will gradually increase, and electron movement around nuclei will be gradually accelerated. However, according to the law of conservation of energy, the temperature of matter and speed of electrons cannot increase infinitely. The injected energy must be consumed in some way, and radiation is one of the ways it can be consumed. Moreover, in classical electromagnetic radiation theory, all accelerated particles with charges will radiate energy, so the accelerated electrons will radiate energy in the form of electromagnetic waves, i.e., light.

According to blackbody radiation theory, a material in thermal equilibrium that absorbs more energy than another same material (leading to a higher temperature) also radiates more light, and the radiated light has more light constituents with shorter wavelengths. Conversely, a material that absorbs less energy than another same material (leading to lower temperature) also radiates less light, and the radiated light has more light constituents with longer wavelengths. For example, when the temperature of heated steel is low, the steel appears to be red due to the radiation of the steel containing more light with longer wavelengths. When the temperature of the steel is high enough, the steel appears to be white because the radiation of the steel contains more light with shorter wavelengths. The sun, having a high temperature due to the thermonuclear reactions occurring in it, radiates electromagnetic waves across most of the electromagnetic spectrum, including γ rays, x rays, ultraviolet, visible light, infrared, and even radio waves. Visible light, defined as light with wavelengths that are visible to normal human eyes, falls in the range of the electromagnetic spectrum between ultraviolet (UV) and infrared (IR). It has wavelengths of about 380 nm to 740 nm.

1.2.2 Explanation by quantum theory

In this section, we explain the generation of light using quantum theory. We first introduce the basic concept of energy levels in matter. In quantum theory, all matter particles, e.g., electrons, also have wave properties (matter waves) and possess the wave–particle duality. The relationship between energy and wavelength or frequency for a matter wave is the same as that for light described in Section 1.1.3.

In order to understand the formation of energy levels in matter, we will suppose that a particle, e.g., an electron, is trapped inside a potential well, shown as a one-dimensional (1D) well in Fig. 1.2(a). In this case, the form of the electron wave is very much like a sound wave from a guitar string. We can

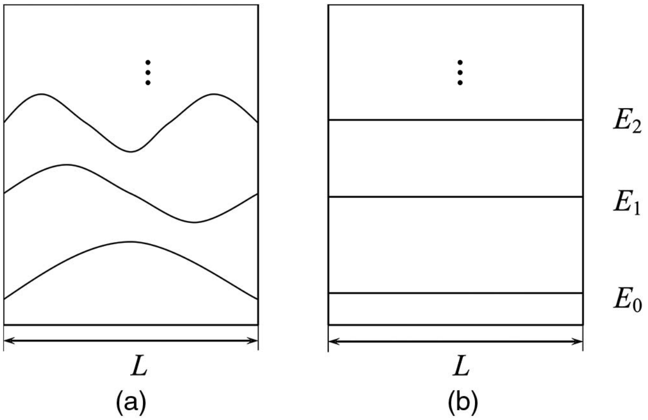

Figure 1.2 Diagram of (a) electron waves in a 1D potential well and (b) the corresponding energy levels.

understand the formation of energy levels in matter using an analogy to a bounded sound wave.

Let us assume that electron waves are constrained by two parallel barriers (a potential well) separated by a distance L. Only waves that can form standing waves after interfering with their corresponding reflected waves from the two parallel barriers can exist in the potential well. For standing waves to exist, the length of the round trip between the two parallel barriers must be an integral multiple of their corresponding wavelengths. As shown in Fig. 1.2(a), wavelengths of electron waves should be $2L$, $2L/2$, $2L/3\ldots$, $2L/m,\ldots$, where $m = 1, 2, 3, \ldots$. So the constraint of the potential well can be seen as a wavelength selector that only retains waves with discrete wavelengths equal to $2L/m$. In other words, only some discrete frequencies exist for electron waves in a bounded region. Because the energy of an electron wave is proportional to its frequency (see Section 1.1.3),[2] values of energies of existing waves are discrete, as shown in Fig. 1.2(b).

The lowest energy level E_0 shown in Fig. 1.2(b) is called the ground state, at which electrons are the most stable. The other energy levels above the ground state are called excited states, where electrons have limited lifetimes and can transit to the ground state or to other low energy levels at any time.

Generally, according to Boltzmann distribution, most of the electrons in matter are in the state with the lowest energy, i.e., the ground state. Once matter is heated or injected with an amount of energy, electrons in the ground state will absorb a portion of the energy and transit to excited states. An example of this is the transition in the two-level system shown in Fig. 1.3(a). Because electrons in excited states have higher energy and limited lifetime, they can quickly transit from excited states to the ground state or to other low energy levels. During this transition process, a portion of energy, in the form of light, will be radiated. As shown in Fig. 1.3(b), the energy $h\nu$ of the radiated

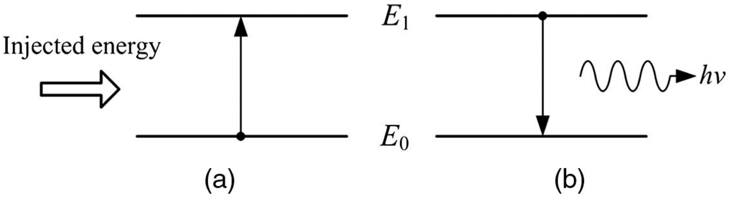

Figure 1.3 Diagram of transitions between E_0 and E_1 and in a two-level system.

light equals the energy gap between the two energy levels. The larger the energy gap the larger the energy of light, and the shorter the wavelength of light.

Furthermore, energy levels and energy gaps between those energy levels for different types of matter are completely different due to the different configurations (or different constraints to electrons) of atoms. This is the reason that each type of matter has its own characteristic absorptions and emissions. For example, wavelengths of light emitted from any type of sodium lamp are all at 589.0 or 589.6 nm. Note that if transitions between states with many different energy gaps occur simultaneously, light in multiple wavelengths or even white light can be produced.

In fact, the above just explained the generation of light from incoherent sources. Light generated by coherent sources, e.g., a laser, will be briefly introduced in Chapter 4.

1.3 Theories of Light: An Overview

Theories of light can currently be divided into three main branches: geometrical optics, wave optics, and quantum optics. In the following subsections, brief introductions to these three branches are presented.

1.3.1 Geometrical optics

Geometrical optics, which deals with light as rays that travel in straight lines in a homogeneous medium, formulates optical laws in the language of geometry. In geometrical optics, the laws of reflection and refraction can provide very good explanations for many optical phenomena, such as specular reflection, the shallower appearance for the depth of water, the dispersion of light, etc. Furthermore, with the help of these two laws and the rectilinear propagation of light rays in homogeneous media, the path of a light ray can be traced throughout an optical system, revealing the main characteristics of that system. In addition, by introducing the diffraction phenomenon in terms of the laws of reflection and refraction, the diffraction of light by edges, corners, or vertices of boundary surfaces can be predicted using geometrical optics.[3] The diffraction ray can also be traced based on geometrical optics.[4] The details of diffraction raytracing will be presented in Section 2.5.2. This geometrical ray-trace methodology for the diffraction ray is commonly used in many optical design software packages.

In order to simplify the calculation of paths of rays and give a quick evaluation of an optical system, the input rays to an optical system are limited in the paraxial region, and this is called paraxial optics. If paraxial optics is extended to the whole space of the optical system, it is called Gaussian optics. In Gaussian optics, optical imaging systems can produce perfect images, and characteristics of an optical imaging system, including the object–image relationship, the magnifications, and the field of view, can be easily calculated in the framework of this theory.

However, as perfect imaging systems do not exist in practice due to the nonlinearity of the law of refraction and dispersion, images of optical imaging systems are blurred. The differences between the blurred and the corresponding perfect image are known as geometrical aberrations. Although aberrations of an optical imaging system are impossible to completely remove from the entire field of view, the performance of the optical imaging system can be greatly improved by correcting most of the aberrations. Hence, correcting geometrical aberrations is one of the most important steps in optical system design.

Usually, geometrical optics can give reasonable explanations for most optical phenomena. However, because geometrical optics is determined from the approximation of wave optics as the wavelength of light approaches zero ($\lambda \rightarrow 0$), the wave nature of light is neglected. Therefore, it is impossible to explain the physical reasons for the diffraction and interference of light using geometrical optics.

1.3.2 Wave optics

Wave optics, which deals with light as waves, studies optical phenomena involved in the wave nature of light, such as diffraction, interference, and polarization.

As light is a form of electromagnetic wave, all wave characteristics of light can be deduced from Maxwell's equations. For example, starting with Maxwell's equations, both the wave equations and the Rayleigh–Sommerfeld diffraction formula for light can be determined. The wave equations of light clearly reveal the wave nature of light; the Rayleigh–Sommerfeld diffraction formula describes an optical field at a reference plane as a superposition of spherical waves. In addition, it is highly convenient to describe the propagation of light in free space with the Rayleigh–Sommerfeld diffraction formula.

According to the Rayleigh–Sommerfeld diffraction formula, the propagation path of light is not along a straight line due to the diffraction of light. In order to simplify the Rayleigh–Sommerfeld diffraction formula, the Fresnel and Fraunhofer approximations are obtained under the near- and far-field approximations, respectively. Similar to the interference of water waves, a light wave can interfere with another light wave, in what we call interference of light waves.

The study of wave optics using the Fourier transform is called Fourier optics. In Fourier optics, a wave is regarded as the superposition of a set of plane waves with the same wavelength, and the direction of the propagation of each plane wave stands for one spatial frequency of the Fourier spectrum of the wave. The Fourier spectrum, which expresses the wave in the spatial frequency domain, is also known as the angular spectrum. In view of the angular spectrum expansion, propagation of light can be seen as the propagation of the angular spectrum. The angular spectrum expansion can describe the propagation of light with more accuracy than the Fresnel and Fraunhofer approximations, but it is not more accurate than the Rayleigh–Sommerfeld diffraction formula.

Wavefront aberrations are the deviations of actual wavefront from the ideal wavefront. Note that directions of rays corresponding to the local wavefront are normal. For the sake of convenience, wavefront aberrations are usually expanded in a series of orthonormal polynomials, e.g., the Zernike polynomials.

In wave optics, the resolution of an optical imaging system is ultimately limited by the diffraction of light, which is different from the case of geometrical optics. Generally, the resolution of an optical imaging system is determined using the following: the resolution limits of the optical imaging system; the Sparrow criterion and Rayleigh criterion; and a measure of image quality, the Strehl ratio.

Wave optics not only explains optical phenomena such as the diffraction and interference of light, but is also commonly used in optical system designs and optical metrology. However, as wave optics ignores the particle aspect of light, it cannot be used in scenarios involving in the interaction between light and matter, such as the photoelectric effect.

1.3.3 Quantum optics

Optical phenomena concerning the interaction between light and matter, such as characteristic emission, the photoelectric effect, etc., is in the realm of quantum optics. The concept of the photon, proposed by Einstein in 1905, is fundamental to quantum optics, and the interaction between light and matter can be considered as interactions between photons and atoms of matter.

For example, the invention of the laser is the most famous application of quantum optics. This new type of optical source provided an important experimental tool for the development of modern optics. Furthermore, commonly used photoelectric detectors, such as charge-coupled devices (CCDs), photodiodes, and photomultiplier tubes, which will be introduced in Chapter 4, are all successful applications of quantum optics.

Quantum optics describes the wave–particle duality of light well. So far, it is the most accurate theory of optics.

Here, we also briefly mention the concept of statistical optics, or the theory of optical coherence. This theory involves the study of the properties of random light fluctuations in terms of statistics. Randomness of light fluctuations is caused by unpredictable fluctuations of light sources, e.g., a hot object, or by a medium, e.g., the atmosphere through which light propagates. Furthermore, the interaction between light and matter is a random or stochastic process that is demonstrated by quantum theory. As a consequence, any detection of light will be accompanied by random fluctuations. For dealing with these situations, a statistical approach must be invoked. This branch of optics is called statistical optics.

In this book, the theories behind geometrical optics and wave optics are predominantly explained, while some of the concepts behind quantum and statistical optics are occasionally depicted as well.

1.4 Overview of Optical Systems

1.4.1 What are optical systems?

An optical system is usually composed of a number of individual optical elements, such as lenses, mirrors, gratings, detectors, etc. However, as the goal of an optical system is to achieve certain functions by manipulating light, an optical system is *not* simply a combination of optical elements. The optical system must be carefully designed to constrain the propagation of light in it. From this point of view, an optical system either (1) processes light to produce an image for viewing or to collect energy for detection or (2) analyzes light to determine a characteristic of the light or to reveal properties of the surroundings that are interacting with light.

In the next subsection, some main types of optical systems are taken as examples to give readers a general idea of what optical systems are. These types of optical systems are classified based on their main functions. Moreover, as the functions of optical systems are achieved by manipulating light, it should be borne in mind that all optical systems must transfer optical energy.

1.4.2 Main types of optical systems

1.4.2.1 Optical imaging systems

Optical imaging systems are one of the most important and widely used optical systems. Generally, optical imaging systems map an object of interest in object space into a corresponding image in image space. These systems allow objects to be seen more clearly either by improving the resolution of an optical system (e.g., an adaptive optics system for a telescope) or by magnifying the image of an object.

By correcting aberrations of a particular optical system via another optical system, the blurred image of an object of interest becomes clear. For example, the eyeglass for correcting vision is an optical imaging system generally used in our lives. With the help of eyeglasses, wearers can see the world more clearly. An adaptive optics system, widely used to improve the image quality of the astronomical telescope, is another optical imaging system that pursues high-resolution images of astronomical targets by correcting aberrations introduced by atmospheric turbulence. Different from eyeglasses, adaptive optics systems need to actively correct dynamic aberrations caused by atmospheric turbulence at a very high frequency. The details of adaptive optics systems will be discussed in Chapter 7.

By magnifying images of objects using optical imaging systems, the targets of interest, too far or too small to be seen, can be observed. Telescopes and microscopes belong to this kind of optical imaging systems. For example, by magnifying the angle subtended to observers, telescopes can display more details of objects at a long distance compared to those visible with the naked eye. As shown in Fig. 1.4, the original angle θ subtended to observers is magnified to become a larger angle θ′. The principle of a microscope for magnifying the image of an object will be presented in Chapter 5.

1.4.2.2 Optical systems for energy collection

Energy-collecting systems are designed to collect optical energy, regardless of their imaging functions. Solar energy collectors are the most widely used energy-collecting systems. Solar energy is renewable and clean. However, as the solar energy density on the surface of the earth is relatively low, it is difficult to use solar energy directly. By using solar energy collectors, the solar energy distributed on a large area can be concentrated on a small area to

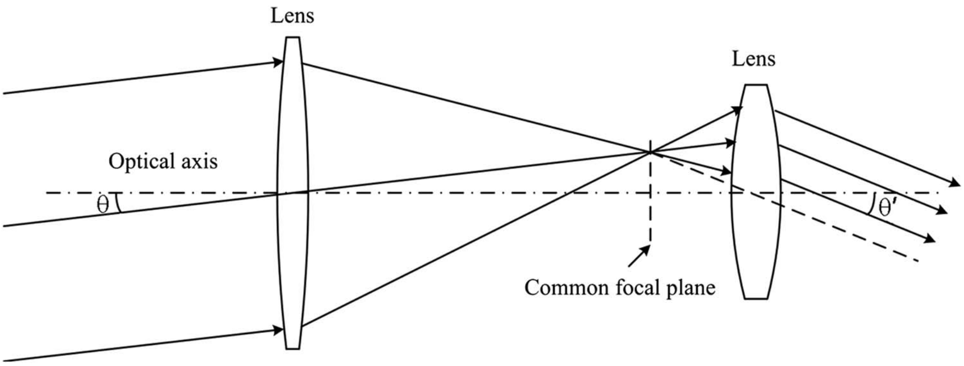

Figure 1.4 Diagram of the magnification of the subtended angle by a telescope.

increase the energy density, and the gathered energy can be used to heat or provide power.

An astronomical telescope can also be considered as a type of energy collector, e.g., for astronomical spectroscopy, which aims to gather as much light as possible to detect the spectrum of the dim objects in space. Since the intensity of light collected by the telescope increases with the square of the aperture size of the telescope, telescopes with large apertures can detect dimmer objects in space compared to telescopes with smaller apertures. This is one important reason that apertures of astronomical telescopes are being designed to be increasingly larger.

Another example of energy-collecting systems is the optical probe shown in Fig. 1.5. This optical probe is designed to measure the lifetime of fluorescence or phosphorescence emitted from a moving target. The purpose of this system is not imaging but collecting light from a moving fluorescent source in a relative wide field of view without losing intensity caused by the conventional probe with a small field of view.[5] To achieve this, an optical conjugate relationship (to be explained in Chapter 2) exists between the first surface of the first lens of the probe and the first surface of the fiber bundle, as shown in Fig. 1.5. This conjugate relationship makes the system very useful for measuring the intensity of light emitted from a moving object.

The normalized energy transfer of the optical probe as a function of the incident angle of light (the normalized energy transfer function) is shown in Fig. 1.6. When the incident angle changes in the range of –6 deg to + 6 deg, the energy-transfer function of the optical probe is close to 1.0 and substantially flat for practical applications. This property of the probe ensures that the measurement of the lifetime of fluorescence from a moving target is almost unaffected by a small field of view in the conventional probe. It should also be noticed that for this system the acceptance angle of the fiber bundle determines the field of view of the system.

The illumination system is a very important type of optical system that is used in almost every corner of our world. In our daily lives, both street lamps on the sides of roads and fluorescent lights in buildings and homes are illumination systems that give us a bright world. In addition to the normally used lights, illumination systems can also play an important role in industry and academic research. For example, by applying a structured illumination system in a microscope, observers can distinguish objects in great detail, or even achieve subdiffraction-limited resolution.[6]

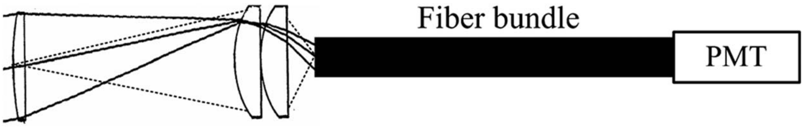

Figure 1.5 Diagram of an optical probe with a photomultiplier tube (PMT) as a detector.

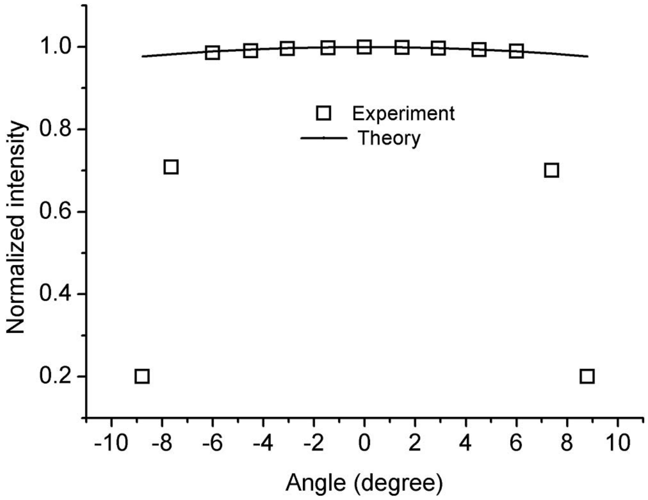

Figure 1.6 Normalized energy-transfer function of the optical probe.

References

1. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, Fourth Edition, McGraw-Hill Education, New York (1976).
2. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Volume III, Commemorative Issue Edition, Addison-Wesley, Boston (1989).
3. J. B. Keller, "Geometrical theory of diffraction," *Journal of the Optical Society of America* **52**(2), 116–130 (1962).
4. Y. G. Soskind, *Field Guide to Diffractive Optics*, SPIE Press, Bellingham, Washington (2011) [doi: 10.1117/3.895041].
5. J. Feist and S. Zhang, "Optical Probe and Apparatus," UK Patent 2484482 (2012).
6. L. Schermelleh, P. M. Carlton, S. Haase, L. Shao, L. Winoto, P. Kner, B. Burke, M. C. Cardoso, D. A. Agard, and M. G. Gustafsson, "Subdiffraction multicolor imaging of the nuclear periphery with 3D structured illumination microscopy," *Science* **320**(5881), 1332–1336 (2008).