1 Introduction to Digital Holography

1.1 Basic Concept of Holography

When light impinges from one medium to another, part of it will be transmitted while the rest will bounce back to the environment or be absorbed into the medium. Figure 1.1(a) shows a ray of monochromatic light (i.e., single frequency or wavelength) L falling on a planar interface between two different homogeneous media M_1 and M_2.The beam is inclined at an angle θ_1 with respect to the normal of the interface. Assuming that M_2 is partially transmissive, and L is a plane wave, the direction of part of the light transmitting through M_2 will be changed from θ_1 to θ_2. This process is known as *refraction*, and the relation between the pair of angles is governed by Snell's law.

If the interface between the two media is smooth, the light that is not refracted will be reflected back as specular (mirror-like) reflection, in which case the angle of incident θ_1 will be identical to the angle of reflection θ_2. For an interface with a rough surface composed of microscopic irregularities, the incident beam will be scattered, resulting in diffusive reflection, as shown in Figure 1.1(b). *Scattering* is a generic term that refers to the change in the propagation path of an electromagnetic wave as it is intercepted by some form of non-uniformities in a medium. Optical scattering results when a beam of light hits one or more tiny particles.

The light impinging on a physical object undergoes one or more of the above-mentioned modes of propagation. Light emitted from the illuminated object surface conveys information on the shape and color of the object. As the emitted light waves fall on the retina of human eyes, an image of the object will be perceived by the brain.

Photography and optical holography are methods for capturing the emitted light waves from an object scene onto a photographic film. When the photographic film carrying the recorded optical signals is presented to an observer, the light waves of the object scene will be partially or fully reproduced to the eyes of an observer. At present, there are two major approaches in capturing the light waves of an illuminated object: photography and holography. Photography is the art of capturing the part of the optical waves corresponding to a two-dimensional (2-D) projected view of an object, while holography captures the entire optical waves. A photograph only presents the intensity or color of a scene; it is a planar image and does not contain any disparity or depth information. With holography, all the light waves emitted by the object are captured,

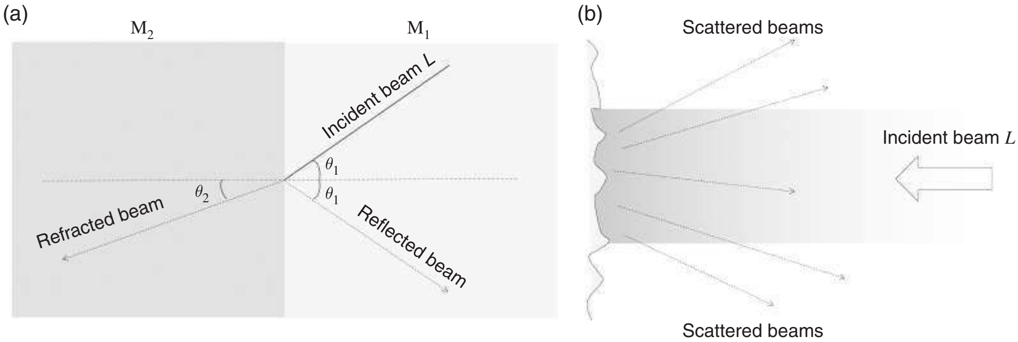

Figure 1.1 (a) Refraction and reflection of a light beam on a smooth surface. (b) Scattering of a light beam on a rough surface.

and hence are capable of reproducing a three-dimensional (3-D) view of the object. These two image-recording techniques are explained as follows.

Light, or optical signal, is a kind of electromagnetic wave in the form of a time-varying sinusoidal wave oscillating with a certain amplitude and frequency. When a person observes a scene, optical waves of different intensities and frequencies will be received by different groups of sensory cells in the retina, creating the impression of color and brightness. The instantaneous amplitude $v(t)$ of a single beam of monochromatic light L is generally expressed as:

$$v(t) = A\cos(\omega t + \theta_L). \tag{1.1}$$

In Eq. (1.1), t represents the time variable, while the terms ω and θ_L denote the frequency and phase of the light beam, respectively. For a monochromatic light wave of constant frequency, the frequency term is sometimes left out, and the representation of the wave can be simplified as $v = A\exp(i\theta_L)$, where i denotes the imaginary unit. In the physical world, an object reflects an infinite number of light beams from its surface, and the collective excitation of these optical signals on the retina leads to the formation of an image of the object to an observer.

Next, we shall explore how the optical waves of an object can be captured. Imagine the light wave of an object, which is sometimes referred to as the object wave, is intercepted by a transparent medium that is capable of recording the amplitude and phase of every optical beam that passes through it, as shown in Figure 1.2(a). For the time being, let's ignore the details of the mechanism behind the recording process. As this medium is illuminated by a coherent plane optical wave (i.e., the amplitude and phase at every part of the wavefront is identical), the light beam that emerges from every point on it will be modulated (in amplitude and phase), by an amount that is identical to that which has been recorded. As illustrated in Figure 1.2, the observer will see the same set of optical waves from either the object or the recording media. Theoretically, the optical waves received by the retina in both cases should be identical, hence casting an impression of the same image.

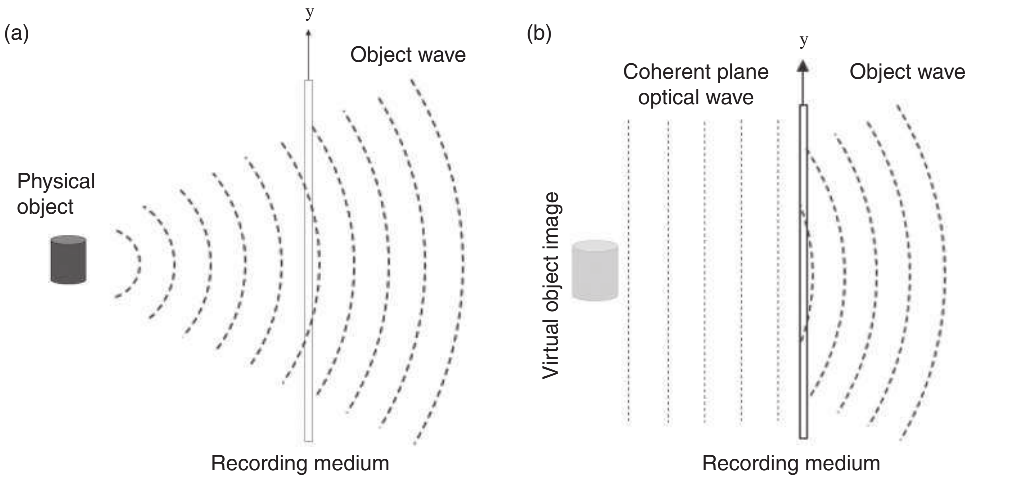

Figure 1.2 (a) Capturing object waves of an object on a recording medium. (b) Reconstructing a virtual object image from the recording medium with a coherent plane optical wave.

The magical recording media is known as a hologram, a word that finds its origin from the Greek words ὅλος (whole) and γραφή (drawing). Literally, the word hologram carries the meaning of something that is capable of recording the complete information of a visible image. Theoretically, there is little difference between looking at the virtual image of a hologram and the actual object. Hence, from a hologram, an observer should be able to see not only the color and intensity of the object, but also all the depth cues and disparity information. Depth cues are the visual information in an image that instigates the perception of distances between the observer and different parts of the object. For example, the lens of our eye will change its focus when objects at different distances are observed. Another example of depth cues is when the viewpoint of an observer is panned (shifted along a certain path): an object that is closer will appear to translate more than one that is farther away, which again creates the sensation of difference in depth between the objects. Disparity, sometimes referred as “binocular disparity,” is the deviation in the location of an object as observed by the left and the right eyes. Through the disparity, our brain is able to interpret the depth of different objects in the scene.

The art of capturing a hologram of an object is commonly referred to as optical holography. The method was invented in late 1940 by Dennis Gabor, who was awarded the Nobel Prize in Physics in 1971 for his contribution to the invention and development of the holographic method. Works of Gabor in optical holography were reported in [1], and later by Bragg in [2]. There have been numerous advancements in optical holography since, and a comprehensive description on the developments can be found in [3,4]. For completion, a brief outline of holography is given here.

The concept of holography can be illustrated from a more analytical point of view. Figure 1.3 shows the side view of an object scene containing a single object point. The intensity of the object point is white, represented with a value of unity. A recording medium is positioned at the hologram plane to record the light wave of the object scene. As the white object point is illuminated with a coherent plane wave (i.e.,

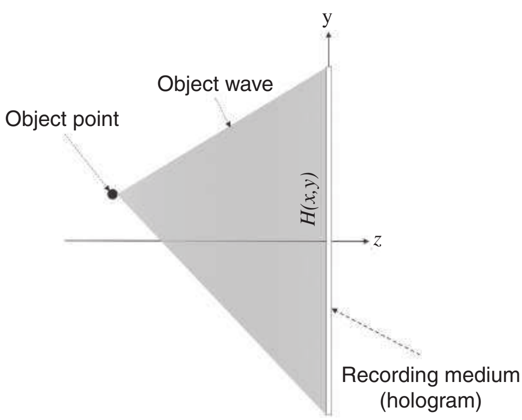

Figure 1.3 Object wave projected by an object point of unit intensity on the recording medium.

a wavefront that is homogeneous in magnitude and phase), it will behave as an isotropic point light source that scatters an object wave $O_0(x,y)$ onto the hologram. Extending this concept, for a single object point with intensity A, the object wave $O(x,y)$ that is projected on the hologram is given by

$$O(x,y) = A|O_0(x,y)\ |\exp[i\theta_0(x,y)]\exp[i\phi_0]. \tag{1.2}$$

In Eq. (1.2), the magnitude and phase of the object wave of a point source are denoted by $|O_0(x,y)\ |$ and $\theta_0(x,y)$, respectively. The term ϕ_0 is the phase of the light illuminating the object point, which will be added to the phase component of the object wave.

The above principles of hologram formation can be easily applied to a scene with multiple object points. An arbitrary object surface can be considered as comprising many isotropic point sources, each emitting its own object wave when illuminated. The optical signal recorded on the hologram is the superposition of the optical wave emitted from each object point. The mixing of these waves is a process that is commonly known as optical interference:

$$O(x,y) = \sum_{p=0}^{P-1} A_p|O_p(x,y)|\exp[i\theta_p(x,y)]\exp[i\phi_p], \tag{1.3}$$

where P is the total number of object points, and subscript p is the index of the object points. The object wave recorded on the hologram is referred as the interference pattern. For a coherent plane wave propagating along the axial direction (i.e., normal to the hologram plane), the phase ϕ_p of the illuminating wave is constant for all the object points, which can be ignored or included as a constant phase term in the computation of the object wave. If the illumination light beam is incoherent, the value of ϕ_p is random for different object points, and the object wave can no longer represent the light field of the source object.

For the time being, assume there are some means that enables the full object wave to be recorded on a medium, resulting in a hologram (denoted by the function $H(x,y)$ in

Figure 1.3). The signals on the hologram are complex-valued (with magnitude and phase components), and in the form of high-frequency fringe patterns. Different names have been adopted in various literatures to describe hologram signals, such as – but not limited to – fringe patterns, hologram fringes, diffractive waves, or holographic signals. These terms will be used throughout this book.

As shown in Figure 1.2(b), a coherent plane wave passing through a hologram is modulated, both in amplitude and phase to give the same object wave $O(x, y)$. From the reconstructed object wave, the observer will be able to see a realistic 3-D image of the object from the hologram. The process of recovering a 3-D image from the hologram is known as "reconstruction." The use of a coherent optical plane wave is mandatory as it will not change the amplitude and phase of the object waves – that is, A_p and $\theta_p(x, y)$ of each object point in both the recording and the reconstruction processes.

1.2 Optical Recording in Practice

A conceptual optical setup for recording the hologram of an object scene is shown in Figure 1.4. The actual recording system is slightly more complicated. A thin laser beam is converted into a plane wave with the use of a beam expander (BX), and illuminates the specimen after passing through the semi-transparent beam splitter (BS). A laser beam is used because it is coherent, so the phase distribution of its wavefront is homogeneous. As explained in Section 1.1, it is necessary that the phase ϕ_p of the illumination beam on every point on the object surface is constant. The beam expander magnifies the cross-sectional area of the thin laser beam so that the illumination light will be wide enough to cover the entire object of interest. The object wave emitted by the object of interest is partially reflected by the beam splitter to the recording media. If both the amplitude and phase of the object wave can be captured,

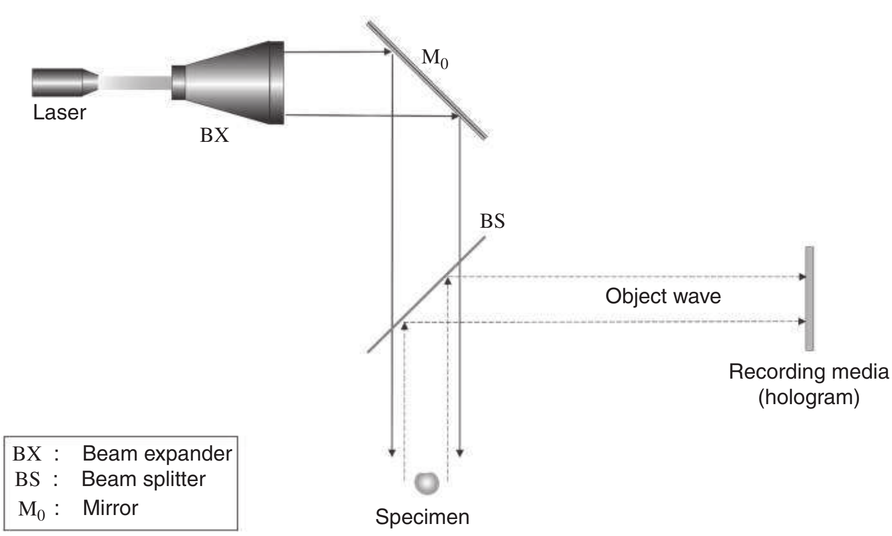

Figure 1.4 Conceptual optical setup in hologram recording.

the recording medium, will become the hologram of the specimen. Unfortunately, this ideal scenario is difficult, if not impossible, to achieve in practice.

The reason is that optical recording materials developed to date, commonly known as photographic films, are usually implemented with a layer of gelatin emulsion coated onto a plastic substrate. The emulsion comprises microscopic grains of silver halide crystal, each of which will be partially converted into silver upon exposure to light. Upon exposing the film to an optical image, the chemical property of each silver halide crystal will change proportionally to the intensity of light falling on it, forming an invisible or latent image. Subsequently, the film can be chemically developed into a visible image. Since the silver halide crystal is only responsive to the intensity of light, a photographic film is only capable of recording the intensity, but not the phase, of a light field. Instead of the full holographic information, the signal recorded on the photographic film is only constituted by the magnitude component of the object wave, as given by

$$O_I(x,y) = |O(x,y)|^2. \tag{1.4}$$

Apparently, the signal recorded by the photographic film in Eq. (1.4) is different from the object wave, as it does not carry any information on the phase component. Due to the absence of the phase terms, an image of the specimen cannot be reconstructed from the signal $O_I(x,y)$.

1.3 Photography

The classical art of photography can be interpreted as a special case of optical holography, although it was invented at a much earlier date when the art of holography was unknown to the world. As seen in Eq. (1.4), it is not possible to record the phase component of the object wave with a photographic film. Photography is based on capturing a small part of the object waves that carries the optical image of a specific viewpoint of the object scene. The concept is to place a small aperture between the object scene and the photographic film (a.k.a. a photograph), so that there is a one-to-one mapping between the light wave emitted from each object point and a single point on the recording plane. In another words, the light waves of different object points will not interfere with one another as they fall onto the recording media. Such an implementation is known as a "pinhole camera." Figure 1.5(a) shows the imaging of a pair of point light sources, A and B, with a pinhole camera. The tiny pinhole only allows a light wave of a unique orientation for each object point to pass through, and impinges on an image plane at which a photographic film is placed. Restricted by the pinhole, each location on a photograph can only receive the object wave of a single object point. Mathematically, the magnitude of the object wave received by a photograph from a single object point at location (x_p, y_p) is expressed as

$$O_P(x,y) = \begin{cases} |A_p|^2 & (x,y) = T(x_p, y_p) \\ 0 & \text{otherwise} \end{cases}, \tag{1.5}$$

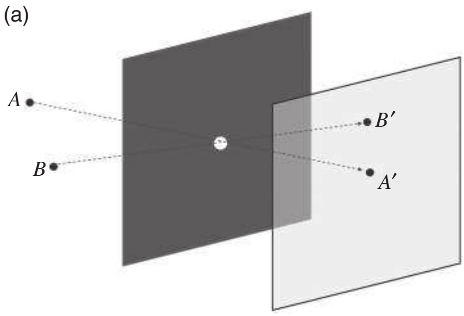

Figure 1.5 (a) A pinhole camera imaging a pair of points A and B. (b) Imaging points A and B with a pinhole at a different location.

where $T(x_p, y_p)$ denotes a one-to-one transformation on the coordinates (x_p, y_p). From Eq. (1.5), the function $O_P(x, y)$ is simply a transformed image of the object scene, with the object point recorded at the position $T(x_p, y_p)$. On the image plane, the image of an object point is flipped horizontally and vertically, and relocated to a new position according to the location of the pinhole. Note that the object wave contains only the magnitude component (i.e., the intensity) of the light wave of a unique point source on the photograph. As a result, the phase component of the light wave is not involved in the recording process, and the object can be illuminated with incoherent light (e.g., ambient lighting from the sun and fluorescent lamps), whereby the phase of the illumination beam on each object point can be totally random and uncorrelated with each other.

The photograph can only capture an image of the object scene at a particular viewpoint that is constrained by the position of the pinhole. With a single view of the object scene, the recorded image cannot convey any depth cue or disparity information. To capture another viewpoint of the object scene, the position of the pinhole has to be changed. Figure 1.5(b) shows the imaging of the pair of object points, A and B, with the pinhole shifted to a different location. From these two figures, the images of the two object points are relocated to different positions on the recording plane.

1.4 Recording Setup in Optical Holography

Although holographic technology can be traced back to 1940, hologram recording was not practiced until the development of the laser in the 1960s, which is capable of generating a coherent light wave. As mentioned before, coherent illumination is mandatory in holographic recording. The first hologram was demonstrated in 1962 by the Soviet scientist Yuri Denisyuk [5], as well as by Emmett Leith and Juris Upatnieks of the University of Michigan in the USA [6]. Different to photography, optical holography records a 3-D – instead of 2-D – image of the object scene onto a photographic film by preserving both the amplitude and phase components of the

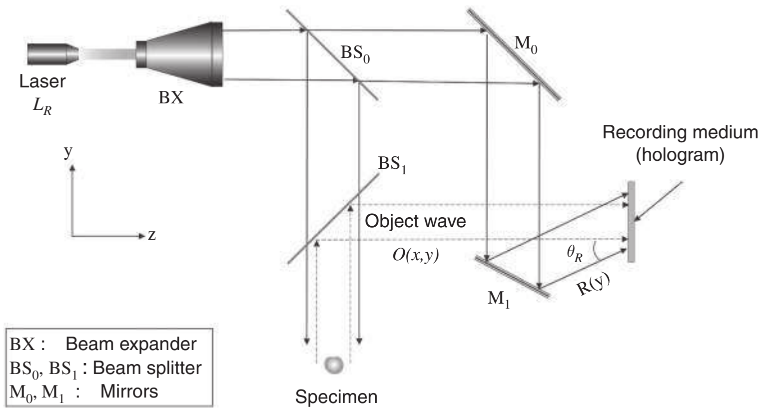

Figure 1.6 Practical optical setup for recording an off-axis hologram.

object wave. Intuitively, this may sound a bit contradictory, as a photographic film is only responsive to the intensity of the light wave. In holographic recording, this problem is overcome by converting the complex-valued holographic signal into intensity information prior to recording on a photographic film. The conversion is achieved by adding the complex-valued object wave with an inclined reference plane wave, which is also derived from the coherent beam that illuminates the object. The result of adding the pair of waves is an interference pattern (which is also commonly referred to as a beat pattern), encapsulating both the magnitude and phase components of the object waves. A typical optical setup for recording an off-axis hologram is given in Figure 1.6.

Similar to Figure 1.4, the source of illumination is a monochromatic laser beam L_R which is expanded into a plane wave via a beam expander. The expanded beam is divided into two separate paths after passing through a beam splitter. The first beam is reflected by the beam splitter BS_0, propagates through the beam splitter BS_1, and illuminates the specimen. The object wave $O(x, y)$ from the specimen is reflected by BS_1, and impinges on the hologram. The second part of the expanded beam is reflected to the hologram via mirrors M_0 and M_1, projecting the beam onto the hologram at an angle of incidence θ_R along the vertical direction, resulting in an inclined reference plane wave

$$R(y) = \exp i2\pi y\lambda^{-1} \sin(\theta_R), \tag{1.6}$$

where λ is the wavelength of the laser beam. Note that $R(y)$ is a pure phase function. If $\theta_R = 0$, the reference wave is simply the plane wave that is used to illuminate the specimen. When the object wave and the reference wave meet on the surface of the photographic film, they combine to form an interference pattern. The magnitude of the interference pattern is then recorded on the photographic film, resulting in a hologram with the intensity given by

$$\begin{aligned} H(x,y) &= |O(x,y) + R(y)|^2 \\ &= [|O(x,y)|^2 + |R(y)|^2] + O(x,y)R^*(y) + O^*(x,y)R(y) \\ &= C_1 + C_2 + C_3, \end{aligned} \tag{1.7}$$

where $C_1 = [|O(x,y)|^2 + |R(y)|^2]$, $C_2 = O(x,y)R^*(y)$, and $C_3 = O^*(x,y)R(y)$.

From Eq. (1.7), the hologram $H(x,y)$ is a real signal that can be recorded onto the photographic film. Subsequently, the object wave can be optically reconstructed from the hologram by illuminating it with a plane wave, as shown in Figure 1.7. The light wave emitted by the hologram is contributed by three different components, each corresponding to a term in Eq. (1.7).

The first term C_1 is the sum of the intensities of the object and the reference beams. Due to the absence of the phase component, this part of the light wave does not carry useful information on the 3-D image of the specimen, and appears as a patch of highlighted region. The second term C_2 is identical to the object wave apart from the constant phase term $R^*(y)$, while the last term C_3 is the conjugate of the object wave multiplied with the reference wave $R(y)$. Being different from C_1, the complex object wave of the specimen are encapsulated in both C_2 and C_3.

From the optical wave of the illuminated hologram, three images can be observed. The first one is a focused, virtual image of the object scene that is reconstructed from the component C_2. This component of the optical wave contains the object wave $O(x,y)$ in the hologram $H(x,y)$, with the addition of the phase term $R^*(y)$. The virtual image can be viewed from an oblique angle, as shown in Figure 1.7. The second image is a blurred, defocused twin image corresponding to C_3, which carries $O^*(x,y)$, the conjugate of the object wave, and the phase term $R(y)$. Similar to the virtual image, the twin image can be observed from another oblique angle. The third image, usually referred to as the zero-order beam, is constituted from the remaining term C_1, and can be observed from the axial direction. Apart from the virtual image, the other pair of images are unwanted content as they do not form a correct reconstruction of the object scene.

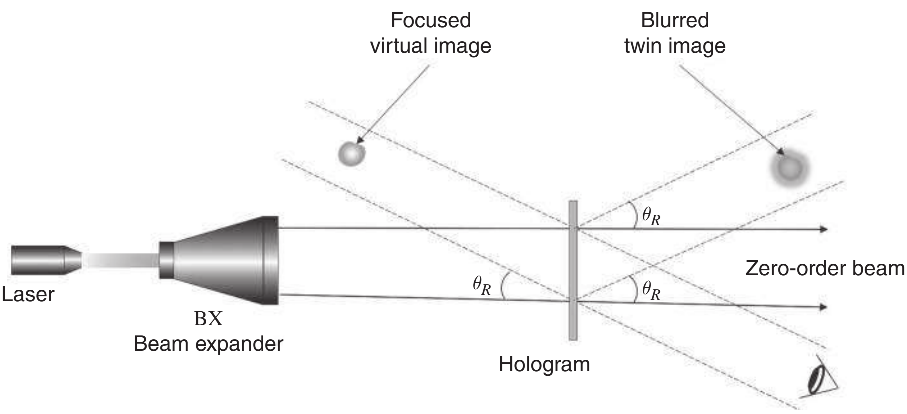

Figure 1.7 Optical reconstruction of an off-axis hologram.

From Figure 1.7, it can be seen that mixing of the object wave $O(x,y)$ with the inclined reference wave $R(y)$ has imposed an angular separation θ_R between each consecutive pair of images that are reconstructed from the hologram. It can be inferred that if θ_R is small, the virtual image could be partially overlapping with, and hence contaminated by, the twin image and the zero-order beam. As such, the angular separation θ_R must be large enough that the virtual image can be separated from the unwanted content, and optically reconstructed as a standalone visible image. Even if this is achieved, there is still a good chance that all three components are visible to the observer, although they are not spatially overlapping with one another. This is one of the major disadvantages of an off-axis hologram.

Theoretically, the virtual image can be fully separated from the unwanted content by increasing the value of θ_R. In practice, due to the limited resolution of the photographic film, a ceiling on the angle θ_R is imposed. As explained previously, a photographic film comprises individual microscopic grains of photosensitive materials, each of which constitute a pixel on the film. The resolution of the photographic film, therefore, is limited by the size of the pixel. Assuming that all the pixels are square in shape and have identical size of $\delta_d \times \delta_d$, an image recorded on it is effectively discretized with a 2-D sampling lattice with a sampling interval δ_d along the horizontal and vertical directions. According to sampling theory, this will impose an upper limit to the spatial frequency of the recorded signal.

To determine the maximum angle of incidence, consider a simple case where only the reference wave is present. For a given θ_R, the spatial frequency of the hologram fringe pattern is the minimum, as compared to the case when the object wave is present. Figure 1.8 shows the side view of a reference wave which is inclined at an angle θ_R along the vertical axis. As the signal on the hologram is discretized into pixels when it is recorded on the photographic film, the vertical axis is denoted by the variable n, where n is an integer representing the row indices of the pixels. The physical location of the kth row along the vertical axis is given by $y = k\delta_d$. A pair of light beams of the reference wave is shown in Figure 1.8, impinging on the hologram at $n = 0$ and $n = k$. The relative phase difference between these two beams is given by

$$d_p = 2\pi\lambda^{-1}d = 2\pi\lambda^{-1}k\delta_d \sin(\theta_R) = \omega k, \tag{1.8}$$

where $\omega = 2\pi\delta_d\lambda^{-1} \sin(\theta_R)$ rad/s. From Eq. (1.8) it can be visualized that the fringe pattern on the hologram is a sinusoidal wave with frequency ω. As the maximum frequency that can be represented in the discrete signal is π rad/s,

$$2\pi\delta_d\lambda^{-1} \sin(\theta_R) \leq \pi \Rightarrow \quad \sin(\theta_R) \leq 0.5\lambda\delta_d^{-1}. \tag{1.9}$$

For small value of θ_R, this expression can be approximated as

$$\theta_R \leq 0.5\lambda\delta_d^{-1}. \tag{1.10}$$