1

Making fringes

1.1 The need for angular resolution

Progress in astronomy is dependent on the development of new instrumentation that can provide data which are better in some way than the data which were available before. One improvement that has consistently led to astronomical discoveries is that of seeing finer detail in objects. In astronomy, the majority of the objects under study cannot easily be brought closer for inspection, so the typical *angular* scales subtended by objects, which depend on the ratio of the typical sizes of the objects to their typical distances from the Earth, are a more useful indicator of how easily they can be seen than their linear sizes alone. The angular separation of two features in a scene which can be just be distinguished from one another is called the *angular resolution* and the smaller this scale is, the more detail can be seen.

The impact of increased angular resolution can be appreciated from comparing the important angular scales of objects of interest with the angular resolution of the instrumentation available at different times in history. Prior to the invention of the telescope, the human eye was the premier 'instrument' in astronomy, with an angular resolution of about 1 arcminute (about 300 microradians). With the notable exceptions of the Sun, Moon and comets, most objects visible in the night sky are 'star-like': they have angular sizes smaller than 1 arcminute and so appear as point-like objects. The first telescopes improved the angular resolution of the naked eye by factors of three to six: Galileo's telescopes are thought to have had angular resolutions of about 10–20 arcseconds (Greco *et al.*, 1993; Strano, 2009) and it became possible to see that planets appear as discs or crescents and have their own moons. Subsequent improvements to telescopes have culminated in telescopes like the Hubble Space Telescope (HST) which have typical angular resolutions of around 50 milliarcseconds (about 250 nanoradians) – better by a factor of more than a thousand than the naked eye.

The increased angular resolution offered by the HST and other high-angular-resolution telescopes has transformed the study of astrophysics. Astronomers have seen many phenomena that were undreamed of even a few decades earlier: young stars surrounded by discs of material left over from their formation, hugely complex filamentary structure in the 'planetary nebulae' surrounding stars at the end of their lives and bright 'cusps' of stellar emission at the centres of galaxies indicating the presence of black holes.

Nevertheless, the angular resolution available with a conventional optical telescope is still inadequate to resolve many important astrophysical phenomena. Amongst the most obvious examples are the following:

Stars – The photospheres of the nearest stars (except for the Sun) are a few milliarcseconds across.

Planet formation – A planet in an Earth-like orbit forming around a star in the nearest star-forming region (around 150 parsecs away) will be about 6 milliarcseconds from its parent star.

Black-hole accretion – The standard model for active galactic nuclei consists of an accreting black hole surrounded by a broad-line region which reprocesses the radiaton emerging from the accretion disc, and a torus of dusty material which can block direct radiation from the accretion disk. The dust tori in the nearest active galactic nuclei have angular diameters of a few milliarcseconds and the broad-line regions are predicted to have sub-milliarcsecond angular radii. The accretion disks themselves are thought to have *micro* arcsecond-scale diameters.

Undeniably, then, there is scope for observing new phenomena if angular resolutions much greater than those available with current telescopes could be achieved.

1.2 The resolution of a single telescope

If a telescope is built so that all optical imperfections are overcome and the distorting effects of the Earth's atmosphere are removed (for example by placing the telescope in space), then the angular resolution of the telescope will be limited by diffraction. This can be understood by considering the observation of a point source of light using such an idealised telescope.

The telescope can be modelled as a perfect lens projecting an image onto a detector as shown in Figure 1.1. The finite size of the telescope is modelled as a circular aperture of diameter d placed in front of the lens. When observing

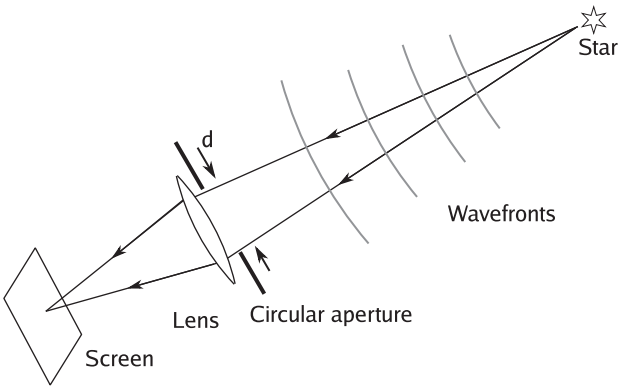

Figure 1.1 A telescope focussing the light from a point source of light (e. g. a star) onto a screen. The telescope is represented as a perfect lens with a circular aperture of diameter d in front of it.

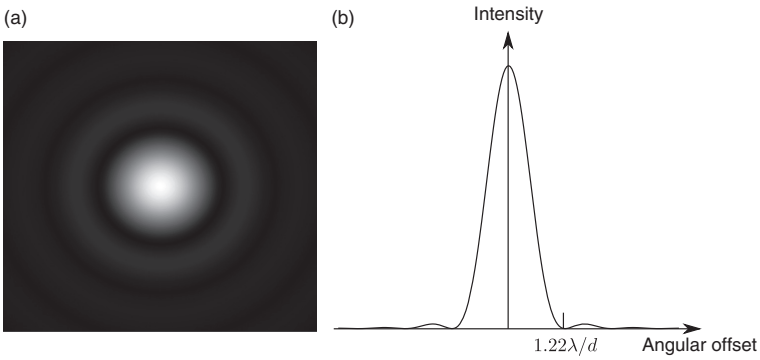

Figure 1.2 The diffraction-limited intensity pattern (known as the 'Airy disc') seen on the screen in the focal plane of the telescope in Figure 1.1 (a) and a cut through the intensity pattern (b).

a point-like object (which will be referred to as a 'star', since most stars are close enough to point-like for these purposes), this arrangement corresponds to a Fraunhofer diffraction experiment. What is seen on the screen is not an infinitely sharp point of light but rather the diffraction pattern of the circular aperture, known as an *Airy disc*, which consists of a central spot surrounded by circular rings as shown in Figure 1.2. This diffraction pattern is known as the *point-spread function* (PSF) of the telescope. Diffraction therefore introduces a finite amount of 'blurring' to the image of the point-like source, even though the lens is modelled as being free from any defects.

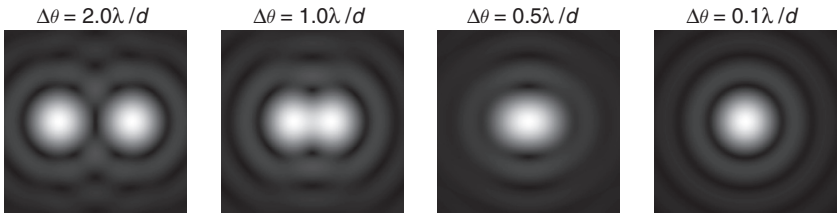

Figure 1.3 Patterns seen in the focal plane of a telescope when pairs of stars of different separations $\Delta\theta$ are observed through a telescope.

The effect of the blurring on angular resolution can be quantified by considering what happens if there is a second star close to the first one as shown in Figure 1.3. The light from the second star will produce a second identically shaped diffraction pattern on the screen, offset by an angular distance $\Delta\theta$ where $\Delta\theta$ is the angular separation of the two stars. Since the phase of the light waves from one star varies randomly and independently of the phase of the light waves from the other star, there is no interference between the two (the justification for this lack of interference is discussed further in Section 1.4.3), and so what is seen on the screen is simply the sum of the intensity patterns that would be seen with either star alone.

If the two stars are brought closer and closer to one another as shown in Figure 1.3, then at some point it becomes impossible to tell whether there is one star or two. At this point the pair of stars is said to be 'unresolved' and the separation at which this occurs is the angular resolution of the telescope. While the exact separation at which the stars become indistinguishable depends on a number of factors such as their relative brightnesses, the *Rayleigh criterion* defines the stars as being 'just resolved' when the peak of the blur pattern produced by one star overlaps with the first null of the blur pattern produced by the second. As shown in Figure 1.2, the angular distance from the peak to the first null of the Airy disc is given by $1.22\lambda/d$, where λ is the wavelength of the light being observed. Thus, the required overlap occurs when

$$\Delta\theta = 1.22\lambda/d. \tag{1.1}$$

By the Rayleigh criterion, Equation (1.1) gives value of the angular resolution of any sufficiently well-corrected telescope. As an example, we can consider the HST, which has a 2.4-m-diameter primary mirror. When observing at a visible wavelength of 500 nm, the diffraction spot from this telescope will have an angular radius of 52 milliarcseconds and so two stars closer together than this cannot be reliably distinguished.

The angular resolution can be improved by building larger telescopes. However, to achieve 1 milliarcsecond resolution at a wavelength of 500 nm would require a telescope 126 m in diameter. Even the largest optical telescopes being proposed at present will have aperture diameters of less than 40 metres, and they come with billion-dollar price tags. The cost of a telescope scales as the square or cube of the aperture diameter so building telescopes more than three times as large seems unlikely in the medium term.

What is required is a method of gaining large factors of improvement in angular resolution which do not require unfeasibly large telescopes. The only known method with this property is long-baseline interferometry. Interferometry uses the interference of light from two or more small telescopes separated by a large distance B to get images with an angular resolution of order λ/B without the mechanical and optical complexities inherent in constructing a single large telescope of diameter B. The next sections serve to show the principles of this method.

1.3 A long-baseline interferometer

An astronomical interferometer collects the light originating from a single region of the sky at two or more locations and brings the collected beams of light together to form an interference pattern. Figure 1.4 shows perhaps the simplest interferometer which could be implemented in practice. Most real interferometers include magnifying and/or demagnifying optics to make the construction of the interferometer easier and cheaper, but the design shown has the advantage that it achieves all the essential functions of an interferometer using only plane (flat) mirrors and so the optical functions of all the elements are readily understandable.

The example interferometer collects starlight using a pair of siderostats, flat mirrors which can be tilted appropriately to reflect the light from a chosen region of sky into a fixed direction. The diameter of the siderostat mirrors can be modest, perhaps only 5 cm if only bright objects are to be observed. In an interferometer used for studying faint objects, the siderostats would typically be replaced by individual telescopes acting as light collectors, each perhaps several metres in diameter.

The distance between the two collectors is typically much larger than the size of any feasible individual collector, perhaps hundreds of metres. The orientation in space of the collector separation is also important: the *baseline vector* $\boldsymbol{B}_{pq}$ between the light-collecting elements p and q of an interferometer is defined as $\boldsymbol{B}_{pq} = \boldsymbol{x}_p - \boldsymbol{x}_q$, where the elements are situated at locations $\boldsymbol{x}_p$

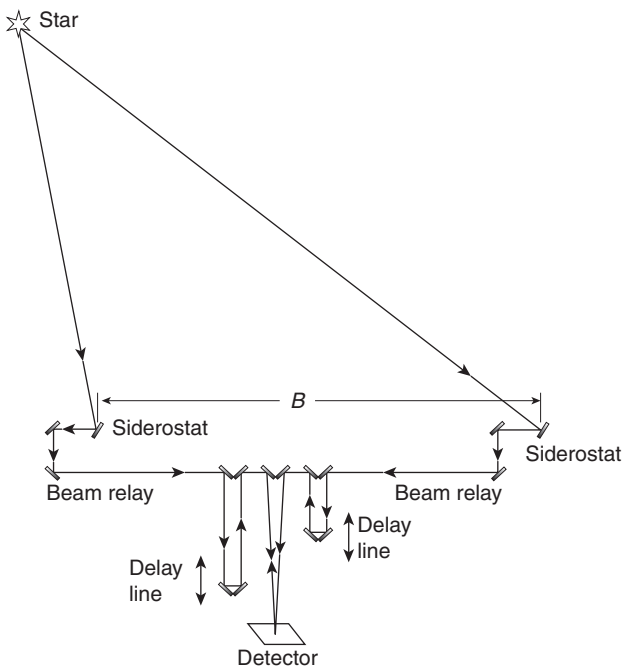

Figure 1.4 A simple long-baseline interferometer constructed out of plane (flat) mirrors. Some of the mirrors have been included to maintain certain symmetries of the optical path – these symmetries are explained in Section 4.4.2.

and $\boldsymbol{x}_q$. In much of the following the pq suffix is dropped, but it will be used later when considering multi-baseline interferometers. The length of the baseline vector is called the 'baseline length' or just the 'baseline' and is shown as B in Figure 1.4.

The collected starlight is brought to a central point using reflections off a series of 'beam-relay' mirrors. Included in the beam path are a pair of mirrors, which can be moved backwards or forwards, acting like an 'optical trombone' to vary the distance the light travels before reaching the central combination point. These 'path compensators' or 'delay lines' serve to control the relative delays between the light beams coming from different collectors: as will be discussed in Section 1.7, in practice it is necessary for the times taken for the light to travel from the object to the point of interference via the two collectors to be matched with one another to in order to see interference.

The light beams are combined in a 'beam combiner' to produce interference fringes. There are many arrangements which can be used to do this: the arrangement shown here uses a so-called 'pupil-plane' arrangement where the

two beams are simply allowed to overlap on a screen. The intensity pattern on the screen can be observed visually but it is more usual to replace the screen by an electronic detector in order to obtain more quantiative information and to observe fainter objects. The detector converts the intensity at each location on its face into an electronic signal, which is digitised, analysed and displayed on a computer.

Interference between the two beams results in a sinusoidal intensity pattern. In the next section it will be demonstrated that this 'fringe pattern' contains information about the size and shape of the object being observed.

1.4 The interferometric measurement equation

The properties of the fringe pattern seen when observing complex objects is derived in the following analysis by considering first the fringe pattern formed when observing a point source, and then how the characteristics of the fringe pattern change when the object consists of two closely spaced point sources. Finally, the properties of the fringe pattern formed when observing an arbitrary object will then be derived by considering it as a collection of closely spaced point sources.

1.4.1 The fringe pattern from a point source

The form of the fringe pattern is derived here using a model of the interferometer which is shown schematically in Figure 1.5. In this model, light arrives from a source of light that is the object of interest. The source is assumed to be a point-like 'star', which is sufficiently distant that the light can be accurately represented as a plane wave, in other words there are plane surfaces known as 'wavefronts' over which the instantaneous electromanetic field E_0 is the same at any given moment in time.

Light propagates from this wavefront along two parallel rays and arrives at the two collectors. The rays then travel via the interferometer optics to the beam combination point, at which point the light waves are superposed and then converted into an intensity $i(x)$. These rays are subject to a series of delays consisting of three different components:

1. An 'external' or 'geometric' delay τ_{ext} due to the light travel time from the wavefront to the collector.
2. An 'internal' delay τ_{int} due to the light travel time along the beam-relay and delay-line beam paths.

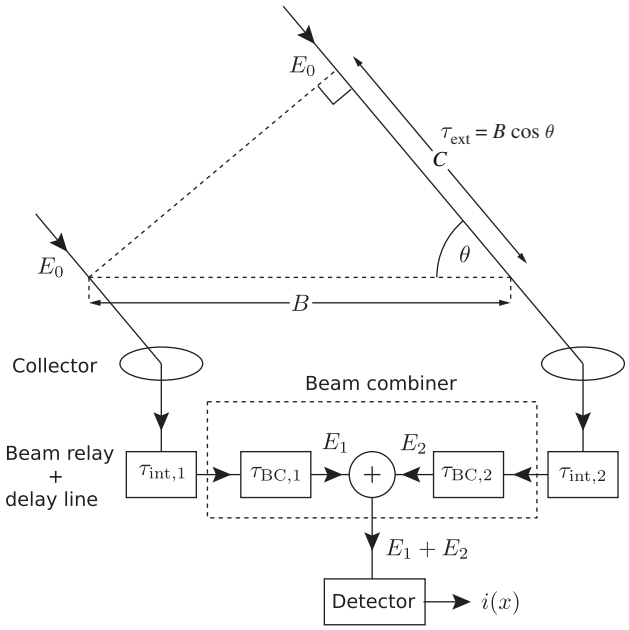

Figure 1.5 A simplified model of fringe formation in an interferometer.

3. A beam-combiner delay $\tau_{BC}(x)$, which is dependent on the location x on the detector on which the beam lands, as shown in Figure 1.6. It will be seen in the following analysis that an important function of any beam-combiner design is to allow the sampling of the interference patterns at multiple locations in 'delay space' and in this case these locations are dependent on the detector coordinate x.

This model neglects the effects of light losses in the interferometer. It also neglects other effects such as optical imperfections along the beam path because it assumes that all rays passing through one collector and arriving at the entrance of the beam combiner experience the same delay. The benefits of using this model are that the analysis is simpler and, perhaps more importantly, the results can be readily applied to interferometers of different designs. For example, the model can be straightforwardly applied to an interferometer which uses temporal coding of the fringes (see Section 4.7) instead of spatial fringes by replacing the detector coordinate x with a time coordinate t.

The analysis starts by considering the electromagnetic field incident on the interferometer. The light is assumed to be perfectly monochromatic so the light wave consists of an electromagnetic wave oscillating at frequency ν and the

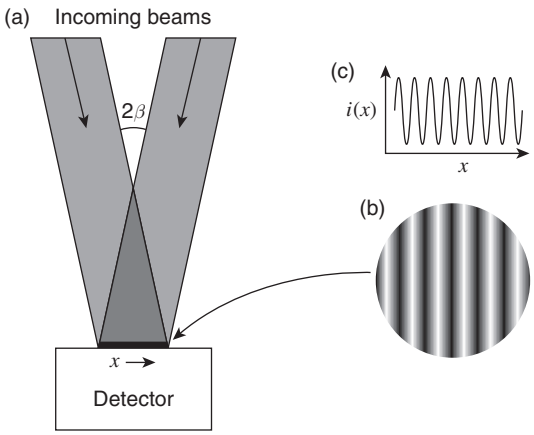

Figure 1.6 (a) Beams of starlight arriving at angles of $\pm\beta$ on a detector. (b) The fringe pattern on the detector. (c) A one-dimensional cut through the fringe pattern.

therefore instantaneous electric field at any point on the initial wavefront is given by

$$E_0 = \frac{2}{\epsilon_0}\mathrm{Re}\left[\Psi_0 e^{-2\pi i\nu t}\right], \tag{1.2}$$

where ϵ_0 is the electric permittivity of free space and Ψ_0 is a 'complex wave amplitude' given by

$$\Psi_0 = |\Psi_0|\, e^{i\phi_0}, \tag{1.3}$$

where ϕ_0 is the phase of the wave. The electric field is not represented as a vector in this 'scalar wave' analysis. It is assumed that the properties of the system being analysed are the same for any polarisation, and so the vector properties of the electromagnetic field are ignored.

At optical frequencies, the oscillations of the wave are typically not directly observable: optical detectors effectively measure the accumulated light energy received over an 'exposure time' or 'integration time', which can be anywhere from several picoseconds to many minutes, whereas the oscillation period is a few femtoseconds. What is observable is the mean intensity of the wave (i. e. the mean energy crossing unit area per unit time, also known as the 'flux' from the object) given by

$$F_0 = \left\langle \epsilon_0 E_0^2 \right\rangle, \tag{1.4}$$

where $\langle\rangle$ represents averaging over the integration time of the detector. Substituting Equation (1.2) into Equation (1.4) and using the relationship

$$\mathrm{Re}\,\{X\} = \tfrac{1}{2}(X + X^*), \tag{1.5}$$

where X is any complex number and X^* denotes the complex conjugate of X, gives

$$F_0 = |\Psi_0|^2\,, \tag{1.6}$$

after dropping terms which average to zero over the exposure time. The simplicity of this expression explains the seemingly arbitrary factor of $2/\epsilon_0$ in Equation (1.2), which serves to define Ψ_0.

In an interferometer, the incident intensity F_0 is not measured directly. Instead, the wave is incident on the two collectors and travels via the beam-relay optics to the beam combiner where the beams are combined and arrive at a given location on the detector surface denoted by a coordinate x. At this location the electric field is given by the superposition of the two fields $E_1(x)$ and $E_2(x)$ that would have been observed from each collector alone. The mean intensity of the light received by the detector at a given location x is therefore given by

$$\begin{aligned} i(x) &= \epsilon_0 \left\langle (E_1(x) + E_2(x))^2 \right\rangle \\ &= \left\langle \left(\mathrm{Re}\left[\Psi_1(x) e^{-2\pi i \nu t} + \Psi_2(x) e^{-2\pi i \nu t} \right] \right)^2 \right\rangle, \end{aligned} \tag{1.7}$$

where $\Psi_1(x)$ and $\Psi_2(x)$ are defined analagously to Ψ_0.

Expanding and dropping terms which average to zero over the integration time of the detector gives

$$i(x) = |\Psi_1(x)|^2 + |\Psi_2(x)|^2 + 2\mathrm{Re}\left[\Psi_1(x)\Psi_2^*(x)\right]. \tag{1.8}$$

The intensity is therefore the sum of the intensities which would be observed on either beam alone, plus a cross term which depends on a product of the two wave amplitudes. This 'interference term' can be positive or negative, corresponding to constructive and destructive interference respectively.

Assuming that the interferometer optics introduce no losses in the light intensity, then the waves arriving at the detector are simply time-delayed versions of the incident wave $E_0(t)$, given by

$$E_1(t) = E_0(t - \tau_{\mathrm{ext},1} - \tau_{\mathrm{int},1} - \tau_{\mathrm{BC},1}(x)) \tag{1.9}$$

and

$$E_2(t) = E_0(t - \tau_{\mathrm{ext},2} - \tau_{\mathrm{int},2} - \tau_{\mathrm{BC},2}(x)). \tag{1.10}$$