Super Resolution for Automated Target Recognition

Maksim Levental

December 2, 2019

Abstract

Super resolution is the process of producing high-resolution images from low-resolution images while preserving ground truth about the subject matter of the images and potentially inferring more such truth. Algorithms that successfully carry out such a process are broadly useful in all circumstances where high-resolution imagery is either difficult or impossible to obtain. In particular we look towards super resolving images collected using longwave infrared cameras since high resolution sensors for such cameras do not currently exist. We present an exposition of motivations and concepts of super resolution in general, and current techniques, with a qualitative comparison of such techniques. Finally we suggest directions for future research.

1 Introduction

Super-resolution (*SR*) is a collection of methods[1] that augment the resolving power of an imaging system. Here, and in the forthcoming, by resolving power we mean the ability of an imaging device to distinguish distinct but proximal objects in a scene. If such objects are modeled as point sources of light then the resolving power of the imaging system is defined by Rayleigh's criterion: two point sources are considered *resolved* when the first diffraction maximum[2] of one point source coincides (at most) with the first minimum of the other (see figure 1).

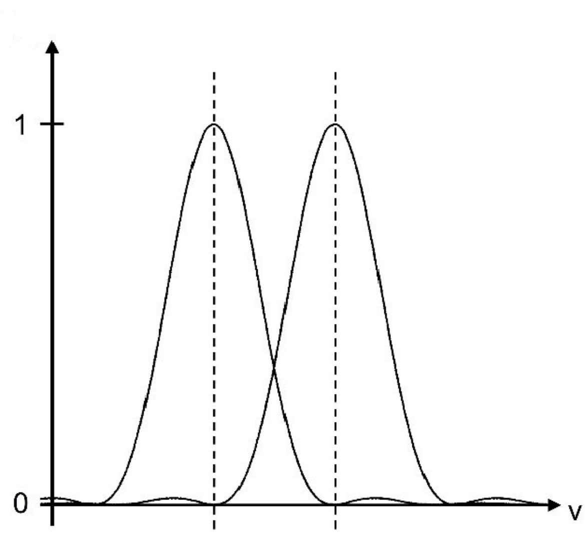


Figure 1: Rayleigh's criterion. Dashed lines indicate diffraction maxima and minima **rayleigh**.

SR techniques yield high-resolution (HR) images from one or more observed low-resolution (LR) images by restoring lost fine details and reversing deterioration produced by imperfect imaging systems. In the case where a single LR source image is used to construct the HR correspondent, the techniques are referred to as single-image-super-resolution (SISR) techniques. These techniques typically operate either by constructing a mapping from low resolution patches[3] to higher resolution patches or by estimating the HR image given the LR image. In the case when multiple LR source images are used to construct a single HR correspondent, the techniques are referred to

[1] We will often use the verb form "to super resolve" in order to denote the use of one or more such methods.

[2] The amplitude of the diffraction pattern (known as the Airy pattern) of a monochromatic point source through a circular aperture is given by

$$I(\theta) := I_0 \left[\frac{2J_1(kaR)}{kaR} \right]^2$$

where I_0 is peak intensity (at the center), $k = \frac{2\pi}{\lambda}$ is the wavenumber of the light, a is the radius of the aperture, R is the *angular resolution* (sine of the angle of observation), and J_1 is the Bessel function of the first kind of order one **goodman2005introduction**. It is maxima/minima of this function that Rayleigh's criterion concerns. See section 2.1 for a complete derivation.

[3] $m \times m$ pixel window, e.g. 3×3.

as multiple-image-super-resolution (MISR) techniques. MISR techniques rely on non-redundant and pertinent information in multiple images of the same scene (see figure 2). Note that for such information to exist there

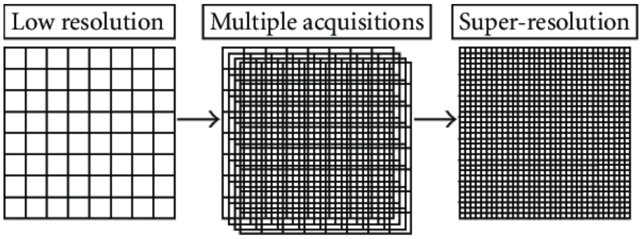


Figure 2: Multiple image super resolution. Multiple low resolution images are superposed on a higher resolution grid in order to recover non-redundant information **misr**.

should be sub-pixel[4] shifts in either the imaging system or the scene between consecutive images.

For typical imaging use-cases higher resolutions are desirable in and of themselves and as inputs to later image processing transformations that can further degrade image quality. In theory the resolving power of an imaging system is primarily determined by the number of independent sensor elements that comprise that imaging system (each of which collects a component of the ultimate image). Naturally then, a way to increase the resolution of such a system is to increase the density of such sensor elements per unit area. Unfortunately, and counter-intuitively, since the number of photons incident on each sensor decreases as the sensor shrinks, shot noise[5] thwarts that idea. Furthermore, while sensor density is primary, secondary effects limit resolution as well. For example, the point spread of a lens (distortion of a point source due to diffraction), chromatic aberrations (distortion due to differing indices of refraction for differing wavelengths of light), and motion blur all work to obscure or erase details from the image.

In domains such as satellite/aerial photography, medical imaging, and facial recognition, high-resolution reconstruction of low-resolution samples is eminently useful since ab-initio acquisition of high-resolution images is either logistically difficult or impossible. For example, in the case of satellite imagery, acquisition of high-resolution imagery is primarily hampered by optics and physics[6]. In contrast, in the case of medical imaging, where patient exposure time needs to be minimized **doi:10.1002.cmr.a.21249**, the primary challenges are logistics and access to repeat collection opportunities.

The benefits of enhancing images using SR techniques include not only more pleasing or more readily interpretable images for human consumption, but higher quality inputs for automated learning systems as well. Indeed this is our ultimate goal — not super-resolution per se but super-resolution in the service of improved object detection performance for longwave-infrared (LWIR) imagery. Towards that end, we do not consider hardware solutions for increasing the resolution of an imaging system. We instead take low resolution images as given produce of a fixed imaging system and explore techniques that allow for ex post facto recovery or inference of precise details. This necessarily constrains techniques under consideration to be algorithmic in nature and software in practice.

The rest of this survey is outlined as follows: section **??** introduces imaging systems, notation, and the mathematical framework for the proceeding sections, section **??** surveys image registration techniques (a necessary pre-processing step for MISR), section **??** surveys classical techniques (those that do not employ neural networks), section **??** surveys neural network techniques with heavy emphasis on deep learning, and finally section **??** concludes with a brief discussion of future research directions.

[4]For example when a point source wholly captured by one sensor element shifts to distributing energy equally amongst that same element and a direct adjacent.

[5]The number of photons incident on a sensor element is distributed according to a Poisson distribution (arrivals are independent and their rate is constant). For a small number of photons the variance in total arrivals (i.e., image brightness) is high.

[6]Rayleigh's criterion implies that the angular resolution R of a telescope with optical diameter $D = 2.4\text{m}$ observing visible light ($\sim$500nm) is approximately **doi:10.1080.14786447908639684**

$$R \approx 1.22\frac{\lambda}{D} = 1.22\frac{500\text{nm}}{2.4\text{m}} \approx 0.06\text{arcsec}$$

From an altitude of 250 km this corresponds to a ground sample distance of 6cm. This loss of resolving power is further exacerbated by refraction through turbulent atmosphere **Fried:66**.

2 Appendix

2.1 Rayleigh Criterion

We derive Rayleigh's criterion for the minimum angular resolution R_{min} of a λ-mono-chromatic point source through a D-diameter circular aperture

$$R_{min} = 1.22\frac{\lambda}{D} \tag{1}$$

from first principles.

2.1.1 Wave Equation in a Vacuum

Note this section primarily follows **eom2004electromagnetic**. Starting with Maxwell's equations in a vacuum (in differential form) for the electric field $\boldsymbol{E}(x, y, z, t)$ and the magnetic field $\boldsymbol{B}(x, y, z, t)$:

$$\nabla \times \boldsymbol{B} = \frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t} \tag{2}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{3}$$

Note that

$$\nabla \times (\nabla \times \boldsymbol{E}) = \frac{\partial}{\partial t}(\nabla \times \boldsymbol{B}) = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} \tag{4}$$

and with the identity

$$\nabla \times \nabla \times \boldsymbol{E} = \nabla(\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E} \tag{5}$$

we have the vector $\boldsymbol{E}$-field *vector wave equation*:

$$\nabla^2 \boldsymbol{E} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} \tag{6}$$

This decouples each components of the $\boldsymbol{E}$ field. We therefore arbitrarily choose the z component E_z of $\boldsymbol{E}$ and solve the *scalar wave equation* for $E := E_z$

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) E = 0 \tag{7}$$

We proceed by *separation of variables*:

$$E(\boldsymbol{x}, t) = U(\boldsymbol{x})T(t) \tag{8}$$

The solution for T is straightforward

$$T(t) = e^{-\mathrm{i}\omega t} \tag{9}$$

where $\omega = \beta c$ and β is separation constant. U obeys the *Helmholtz equation*

$$(\nabla^2 + \beta^2)U = 0 \tag{10}$$

To solve for U we seek a *Green's function* $G(\boldsymbol{x}, \boldsymbol{x}')$ of eqn. (10), i.e., G such that

$$\nabla^2 G + \beta^2 G = -\delta(\boldsymbol{r}) \tag{11}$$

where $-\delta(\boldsymbol{r})$ is the *Dirac delta*[7] centered at $\boldsymbol{x}'$ and $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}'$. Substituting the Fourier transform $\tilde{G}(\boldsymbol{k})$ of G

$$G(\boldsymbol{x}, \boldsymbol{x}') = \int\limits_{\mathbb{R}^3} \tilde{G} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\,\mathrm{d}\boldsymbol{k} \tag{12}$$

where $\boldsymbol{k} = (k_1, k_2, k_3)$ and the Fourier representation of $\delta(\boldsymbol{r})$

$$\delta(\boldsymbol{r}) = \int\limits_{\mathbb{R}^3} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\,\mathrm{d}\boldsymbol{k} \tag{13}$$

into eqn. (11)

$$\int\limits_{\mathbb{R}^3} \left(-k^2 + \beta^2\right)\tilde{G} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\,\mathrm{d}\boldsymbol{k} = -\int\limits_{\mathbb{R}^3} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\,\mathrm{d}\boldsymbol{k} \tag{14}$$

where $k = \boldsymbol{k} \cdot \boldsymbol{k}$. Comparing both sides of eqn. (14) we conclude that

$$\tilde{G}(\boldsymbol{k}) = \frac{1}{k^2 - \beta^2} \tag{15}$$

and hence

$$G(\boldsymbol{x}, \boldsymbol{x}') = \int\limits_{\mathbb{R}^3} \frac{e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}}{k^2 - \beta^2}\,\mathrm{d}\boldsymbol{k} \tag{16}$$

To compute this inverse Fourier transform, first rotate the $\boldsymbol{k}$ coordinate system such that the k_3-axis and $\boldsymbol{r}$ are aligned and then transform to spherical coordinates (see figure 3). That implies

$$\boldsymbol{k} \cdot \boldsymbol{r} = k|\boldsymbol{r}|\cos(\theta)$$

and the differential volume element is

$$\mathrm{d}V = k^2 \sin(\theta)\,\mathrm{d}\phi\,\mathrm{d}\theta\,\mathrm{d}k \tag{17}$$

[7] A generalized function is a function that can only be integrated against. Let

$$\delta_\epsilon(x) := \frac{1}{\epsilon\sqrt{\pi}} e^{-x^2/\epsilon^2}$$

Then for any continuous $\phi(x)$

$$\lim_{\epsilon\to 0}\left[\int\limits_{-\infty}^{\infty} \phi(x)\delta_\epsilon(x - x_0)\right] = \lim_{\epsilon\to 0}\left[\phi(x_0)\int\limits_{-\infty}^{\infty} \delta_\epsilon(x - x_0)\right] = \phi(x_0)$$

and hence $\delta(x)$ is defined as the limit of this behavior

$$\int\limits_{-\infty}^{\infty} \phi(x)\delta(x - x_0) = \phi(x_0)$$

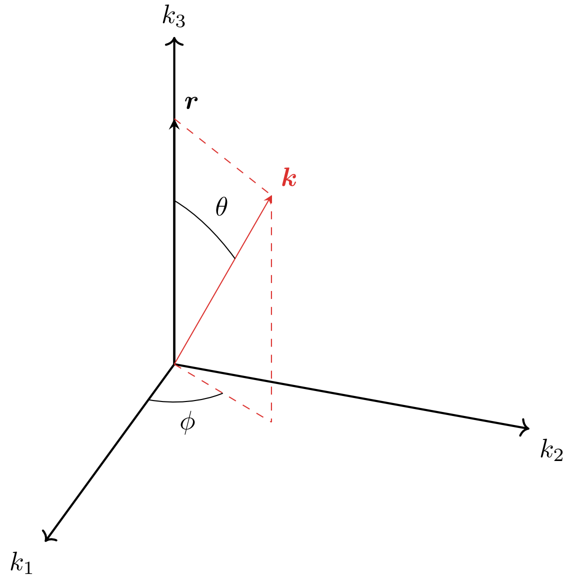


Figure 3: Rotated/aligned spherical coordinate system for computing the inverse Fourier transform of $\tilde{G}$.

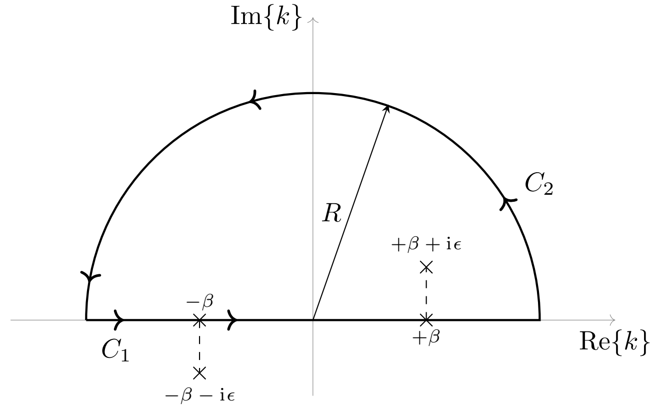


Figure 4: Contour $C = C_1 \cup C_2$ for the Helmholtz Green's function contour integral with poles $-\beta, +\beta$ and shifted poles $-\beta - \mathrm{i}\epsilon, +\beta + \mathrm{i}\epsilon$.

Then

$$G(\boldsymbol{x}, \boldsymbol{x}') = \int_0^\infty \int_0^\pi \int_0^{2\pi} \frac{e^{\mathrm{i}k|\boldsymbol{r}|\cos(\theta)}}{k^2 - \beta^2} k \sin(\theta)\, \mathrm{d}\phi\, \mathrm{d}\theta\, \mathrm{d}k = \frac{2\pi}{\mathrm{i}|\boldsymbol{r}|} \int_0^\infty \frac{e^{\mathrm{i}k|\boldsymbol{r}|} - e^{-\mathrm{i}k|\boldsymbol{r}|}}{(k-\beta)(k+\beta)} k\, \mathrm{d}k \tag{18}$$

Hence, since

$$\int_0^\infty \frac{e^{\mathrm{i}k|\boldsymbol{r}|}}{(k-\beta)(k+\beta)} k\, \mathrm{d}k = \int_{-\infty}^0 \frac{-e^{-\mathrm{i}k|\boldsymbol{r}|}}{(k-\beta)(k+\beta)} k\, \mathrm{d}k \tag{19}$$

equation (18) becomes

$$G(\boldsymbol{x}, \boldsymbol{x}') = \frac{2\pi}{\mathrm{i}|\boldsymbol{r}|} \int_{-\infty}^\infty \frac{k e^{\mathrm{i}k|\boldsymbol{r}|}}{(k-\beta)(k+\beta)} \mathrm{d}k \tag{20}$$

Note that there are there are two singularities or *poles* in the integrand in eqn. (18): the integrand goes to ∞ as $k \to +\beta$ or $k \to -\beta$. To perform the integral, despite the poles, we need to use complex integration and the residue theorem[8]: we perform a line integral (called a *contour integral*) in the k-complex plane such that its value along the real axis equals the integral in eqn. (20) and its value elsewhere is zero.

A critical requirement of the residue theorem is that the poles are wholly contained in the contour. To that end we shift the poles $-\beta, +\beta$ by a pure imaginary component $+\mathrm{i}\epsilon$ and then ultimately take the limit of the result of the perturbed integral as $\epsilon \to 0$. Consider the contour C in figure 4. It's composed of the portion C_1 along $\mathrm{Re}\{k\}$ and the semi-circle C_2 in the positive $\mathrm{Im}\{k\}$ half-plane. If we take the limit as $R \to \infty$ then the integral along C_1 agrees with the integrand in eqn. (20) and by *Jordan's lemma*[9], with

$$g(k) = \frac{1}{k^2 - \beta^2}$$

the integral along C_2 vanishes. Therefore

$$\begin{aligned} G(\boldsymbol{x}, \boldsymbol{x}') &= \lim_{\epsilon \to 0} \frac{(2\pi)^2}{|\boldsymbol{r}|} \mathrm{Res}(f, \beta + \mathrm{i}\epsilon) && (21) \\ &= \frac{(2\pi)^2}{|\boldsymbol{r}|} \lim_{\epsilon \to 0} \frac{\beta e^{\mathrm{i}\beta|\boldsymbol{r}|}}{2(\beta + \mathrm{i}\epsilon)} && (22) \\ &= A \frac{e^{\mathrm{i}\beta|\boldsymbol{r}|}}{|\boldsymbol{r}|} && (23) \\ &= A \frac{e^{\mathrm{i}kr}}{r} && (24) \end{aligned}$$

where A absorbs constant factors, $r = |\boldsymbol{r}|$, and $k \equiv \beta$

[8] Consider a contour in the complex plane that encloses poles a_j. *Cauchy's residue theorem* dictates that for a *holomorphic* function f

$$\oint_C f(z)\, dz = 2\pi\mathrm{i} \sum \mathrm{Res}(f, a_j)$$

where $\mathrm{Res}(f, a_k)$ are the *residues* at the poles a_k:

$$\mathrm{Res}(f, a_j) := \frac{1}{(n-1)!} \lim_{a \to a_j} \left[\frac{d^{(n-1)}}{da^{(n-1)}} \Big((a - a_j)^n f(a) \Big) \right]$$

where n is the order of the pole.

[9] Consider a complex-valued, continuous function f, defined on a semicircular contour

$$C_R = \{ Re^{i\theta} \mid \theta \in [0, \pi] \}$$

of positive radius R lying in the upper half-plane, centered at the origin. If the function f is of the form

$$f(z) = e^{iaz} g(z), \quad z \in C_R$$

to better align with conventional notation. Finally, recalling eqn. (8), we have the *mono-chromatic spherical wave* solution[10]:

$$E(r,t) = A\frac{e^{\mathrm{i}(kr-\omega t)}}{r} \tag{25}$$

2.1.2 *Kirchhoff-Helmholtz integral theorem*

Note this section and the proceeding ones closely follows **Ivanov2016ElementsOD**. Let P_0 be a point source emitting spherical waves and S an arbitrary spherical wave-front with radius r_0. According to the *Huygens-Fresnel principle* each point Q on S is a secondary emitter. Consider figure 5 and let S_1 be the secondary wave-front (envelope of secondary emitters). Then the field at a distant point P is the superposition of secondary waves that originate from the surface S. Kirchoff's derived Huygens-Fresnel by computing the field at point P in terms of only the field and its derivatives at points on an arbitrary surface enclosing P (see figure 6) using an *auxiliary function* U'.

First we derive a general fact about fields; if we assume two fields U, U' are continuously differentiable to second order on S (and on its interior) then we have by *Green's theorem*[11]

$$\int_v \left(U\nabla^2 U' - U'\nabla^2 U\right) dV = \\ -\int_S \left(U\frac{\partial U'}{\partial \boldsymbol{n}} - U'\frac{\partial U}{\partial \boldsymbol{n}}\right) \mathrm{d}S \tag{26}$$

where $\frac{\partial}{\partial \boldsymbol{n}}$ is the directional derivative in the direction of the **inward**[12] pointing normal $\boldsymbol{n}$. Then, since U, U' both satisfy eqn. (10), the left-hand side of eqn. (26) vanishes and we have that

$$\int_S \left(U\frac{\partial U'}{\partial \boldsymbol{n}} - U'\frac{\partial U}{\partial \boldsymbol{n}}\right) \mathrm{d}S = 0 \tag{27}$$

Suppose now that the auxiliary function U' is the Helmholtz Green's function (eqn. (24))

$$U' = \frac{e^{\mathrm{i}ks}}{s} \tag{28}$$

That is to say, to be the field at a secondary emitter on the wave-front[13]. Note that U' has a singularity at $s = 0$ (i.e. at P); since an assumption of eqn. (26) is that U' is continuous and differentiable in the entire volume v, we must exclude P from the domain of integration. We shall therefore surround P by a small sphere S' of radius ϵ, treat the integration on the left-hand side eqn. (26) as **over the volume between the exterior surface S and the interior surface** S', and then take the limit as $\epsilon \to 0$.

Hence the integration in eqn. (27)

$$\left\{\int_S + \int_{S'}\right\}\left[U\frac{\partial}{\partial \boldsymbol{n}}\frac{e^{\mathrm{i}ks}}{s} - \frac{e^{\mathrm{i}ks}}{s}\frac{\partial U}{\partial \boldsymbol{n}}\right] \mathrm{d}S = 0 \tag{29}$$

Note that surface normal $\boldsymbol{n}$ points inwardly (into the volume) at both surfaces. Therefore

$$\int_S U\frac{\partial}{\partial \boldsymbol{n}}\frac{e^{\mathrm{i}ks}}{s} - \frac{e^{\mathrm{i}ks}}{s}\frac{\partial U}{\partial \boldsymbol{n}} \mathrm{d}S = \\ -\int_{S'} \left(U\frac{e^{\mathrm{i}ks}}{s}\left(\mathrm{i}k - \frac{1}{s}\right) - \frac{e^{\mathrm{i}ks}}{s}\frac{\partial U}{\partial \boldsymbol{n}}\right) \mathrm{d}S = \\ -\int_\Omega \left(U\frac{e^{\mathrm{i}k\epsilon}}{\epsilon}\left(\mathrm{i}k - \frac{1}{\epsilon}\right) - \frac{e^{\mathrm{i}k\epsilon}}{\epsilon}\frac{\partial U}{\partial \boldsymbol{n}}\right) \epsilon^2 \,\mathrm{d}\Omega \tag{30}$$

where $\mathrm{d}\Omega$ is differential *solid angle*[14] and we have used the fact

$$\frac{\partial}{\partial \boldsymbol{n}}\frac{e^{\mathrm{i}ks}}{s} = \boldsymbol{n}\cdot\nabla\left(\frac{e^{\mathrm{i}ks}}{s}\right) = \\ \boldsymbol{n}\cdot\hat{\boldsymbol{s}}\left(\frac{\partial}{\partial s}\frac{e^{\mathrm{i}ks}}{s}\right) = \frac{\partial}{\partial s}\frac{e^{\mathrm{i}ks}}{s} \tag{31}$$

since $\boldsymbol{n}\cdot\hat{\boldsymbol{s}} = 1$ (since S' is a sphere). Taking the limit of the right-hand side of eqn. (30) we see that terms proportional to ϵ don't contribute (first and third) and

with a positive parameter a, then Jordan's lemma states the following upper bound for the contour integral:

$$\left|\int_{C_R} f(z)\,dz\right| \leq \frac{\pi}{a}M_R \quad \text{where} \quad M_R := \max_{\theta\in[0,\pi]}\left|g\left(Re^{i\theta}\right)\right|$$

[10]So named due to the spherical symmetry of the solution, i.e., dependence on only r and a single frequency ω

[11] Itself a special case of the *divergence theorem*, Green's theorem states that for ψ, φ both twice continuously differentiable on $U \subset \mathbb{R}^3$

$$\int_U \left(\psi\,\nabla^2\varphi - \varphi\,\nabla^2\psi\right) dV = \oint_{\partial U}\left(\psi\frac{\partial\varphi}{\partial \mathbf{n}} - \varphi\frac{\partial\psi}{\partial \mathbf{n}}\right) \mathrm{d}S$$

where $\frac{\partial}{\partial \boldsymbol{n}}$ is the directional derivative of in the direction of the outward pointing normal $\boldsymbol{n}$ to the surface element $\mathrm{d}S$, for example

$$\frac{\partial\varphi}{\partial \boldsymbol{n}} = \boldsymbol{n}\cdot\nabla\varphi$$

[12]We use inward rather than outward in order to handle a forthcoming nuance.

[13]s being the distance from P to the emitter as in figure 5

[14]The solid angle element is the angular component of the differential surface element

$$d\Omega = \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$$

For example, for a surface integral over a sphere of radius r (in spherical coordinates)

$$\int_{dA} = \int_\Omega r^2 d\Omega = r^2 \int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi} \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$$

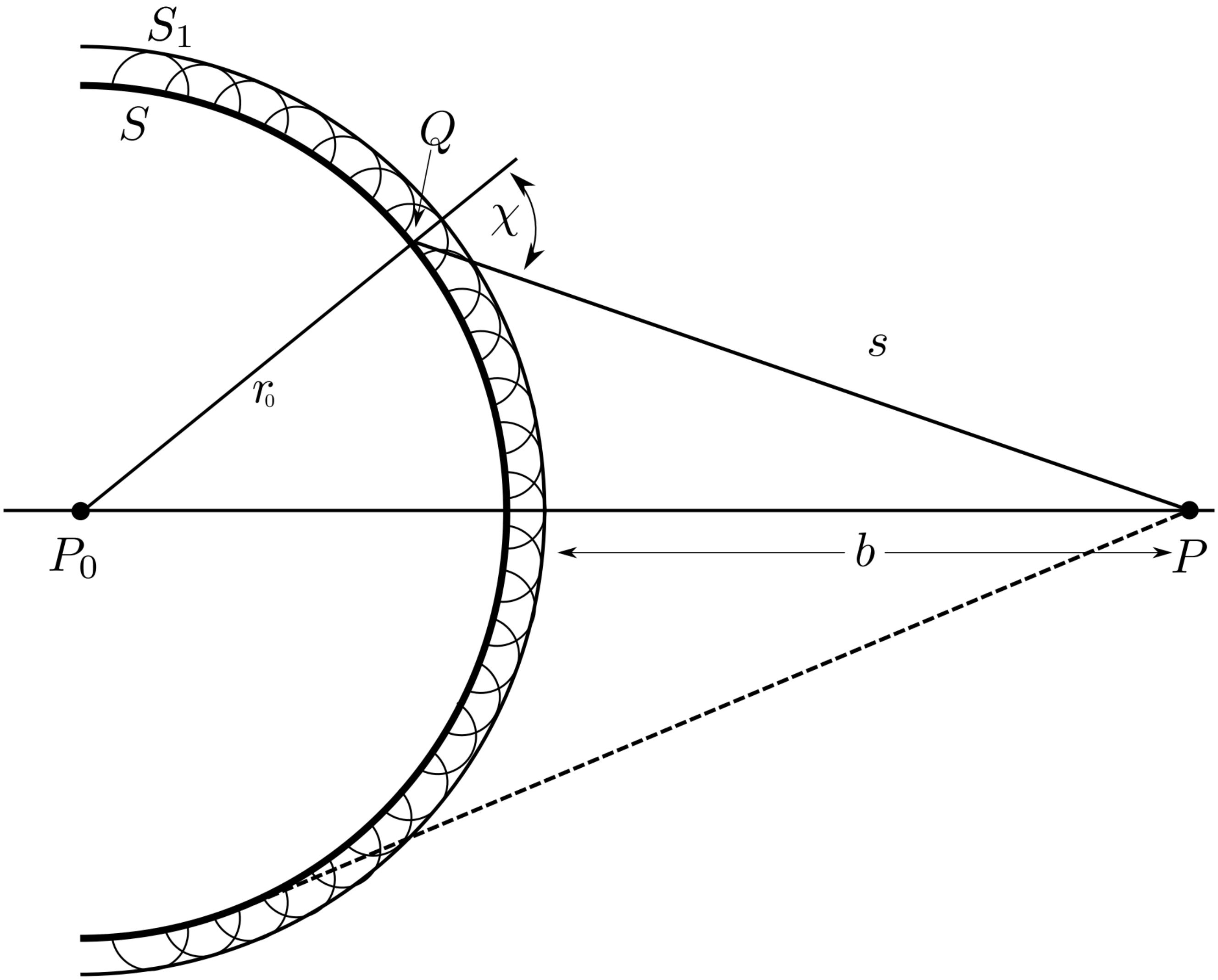


Figure 5: Illustration of the Huygens-Fresnel principle. P_0 and P are the light source and the receiver, respectively, r_0 is the radius of the wave front with surface S, S_1 is the wave front at a later instant and χ is the inclination angle at some point Q **Ivanov2016ElementsOD**.

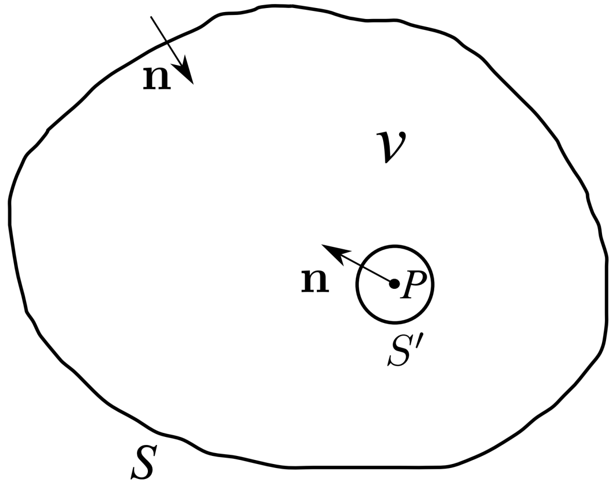


Figure 6: Schematic for derivation of Kirchhoff-Helmholtz integral **Ivanov2016ElementsOD**.

therefore

$$\int_S \left(U \frac{\partial}{\partial \boldsymbol{n}} \frac{e^{\mathrm{i}ks}}{s} - \frac{e^{\mathrm{i}ks}}{s} \frac{\partial U}{\partial \boldsymbol{n}} \right) \mathrm{d}S = \lim_{\epsilon \to 0} \int_\Omega U e^{\mathrm{i}k\epsilon} \, \mathrm{d}\Omega \quad (32)$$

$$= 4\pi U(P) \quad (33)$$

since $\lim_{\epsilon \to 0} e^{\mathrm{i}k\epsilon} = 1$ and the integration over $\mathrm{d}\Omega$ doesn't depend on ϵ. Therefore the field U at point P is determined by the *Kirchhoff-Helmholtz integral theorem*

$$U(P) = \frac{1}{4\pi} \int_S \left(U \frac{\partial}{\partial \boldsymbol{n}} \frac{e^{\mathrm{i}ks}}{s} - \frac{e^{\mathrm{i}ks}}{s} \frac{\partial U}{\partial \boldsymbol{n}} \right) \mathrm{d}S \quad (34)$$

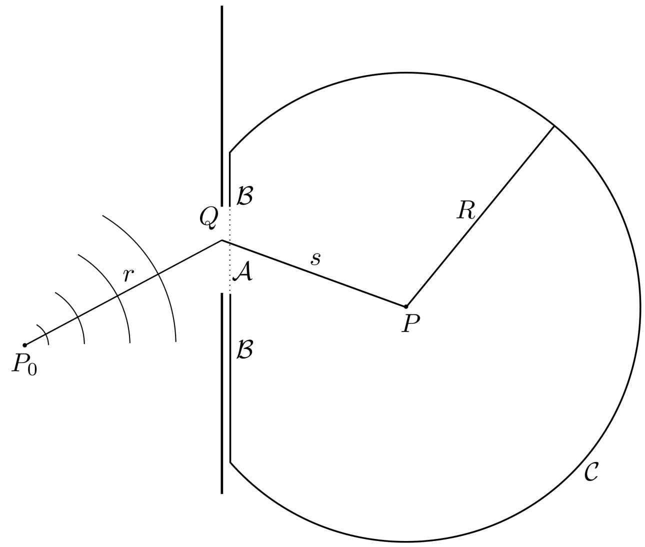
Note that the Kirchhoff-Helmholtz integral theorem determines $U(P)$ as a function of the values of U and $\partial U / \partial \boldsymbol{n}$ on the surface S alone.

2.1.3 Fresnel-Kirchhoff diffraction

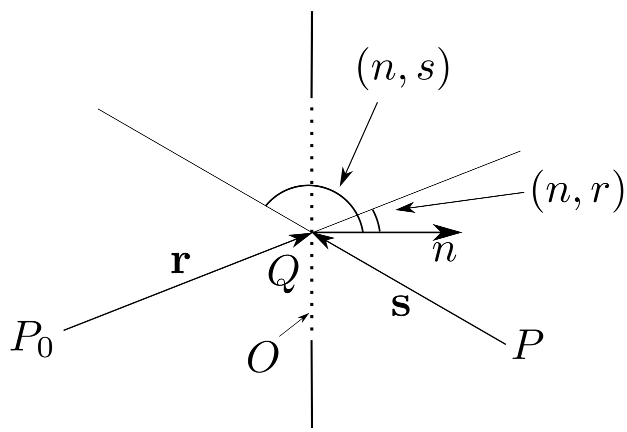
We now apply eqn. (34) to the typical case of waves propagating through an aperture in a screen (see figure 7a). We take as our surface $S = \mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$ where $\mathcal{A}$ is the aperture itself, $\mathcal{B}$ is the unilluminated side of the screen, and $\mathcal{C}$ is a portion of a large sphere centered at P. The Kirchhoff-Helmholtz integral dictates that

$$U(P) = \frac{1}{4\pi} \left\{ \int_{\mathcal{A}} + \int_{\mathcal{B}} + \int_{\mathcal{C}} \right\} \left[U \frac{\partial}{\partial \boldsymbol{n}} \frac{e^{\mathrm{i}ks}}{s} - \frac{e^{\mathrm{i}ks}}{s} \frac{\partial U}{\partial \boldsymbol{n}} \right] \mathrm{d}S \quad (35)$$

Note that a fortiori the field is ~ 0 on $\mathcal{B}$ because it is unilluminated. Similarly we'd hope the field on $\mathcal{C}$ is approximately zero as well, since we are free to choose R as large as we want. Unfortunately, this isn't exactly the case; while it is true that both $U, \partial U / \partial \boldsymbol{n} \to 0$ as $R \to \infty$ it's also the case that the area of $S \to \infty$



(a) The surface S consists of portions $\mathcal{A}, \mathcal{B}, \mathcal{C}$ **Ivanov2016ElementsOD**.



(b) Definition of angles and directions. $\mathbf{r}$ and $\mathbf{s}$ point in the direction of increasing r and s, respectively, and $\mathbf{n}$ is normal to the screen. (n, s) and (n, r) indicate the angles between the normal $\boldsymbol{n}$ and $\boldsymbol{s}, \boldsymbol{r}$ respectively **Ivanov2016ElementsOD**.

Figure 7: Illustrating the derivation of the Fresnel-Kirchhoff diffraction formula of propagation through an aperture.

and so we don't necessarily have that the integral on $\mathcal{C}$ vanishes. To assure this we need to further assume *Sommerfeld's radiation condition*:

$$\lim_{R\to\infty} R\left(U(R) - \frac{\partial U}{\partial \boldsymbol{n}}\right) = 0 \tag{36}$$

which assumes we are only dealing with *outgoing* waves. With this assumption included the integral on $\mathcal{C}$ indeed vanishes.

This leaves the field on $\mathcal{A}$. Kirchoff assumed that the field on $\mathcal{A}$ is the same as if the screen were absent, i.e., the field is determined by a spherical wave emanating from P_0 (see eqn. (24)). Therefore we have

$$U(P) = \frac{1}{4\pi}\int\limits_S \left(U\frac{\partial}{\partial \boldsymbol{n}}\frac{e^{iks}}{s} - \frac{e^{iks}}{s}\frac{\partial U}{\partial \boldsymbol{n}}\right)\mathrm{d}S \tag{37}$$

$$= \frac{A}{4\pi}\int\limits_{\mathcal{A}} \left(\frac{e^{ikr}}{r}\frac{\partial}{\partial \boldsymbol{n}}\frac{e^{iks}}{s} - \frac{e^{iks}}{s}\left(\frac{\partial}{\partial \boldsymbol{n}}\frac{e^{ikr}}{r}\right)\right)\mathrm{d}S \tag{38}$$

Assuming the wavelength of the field is much smaller than the distances involved ($\lambda \ll r, s$ both and therefore $k \gg 1/r, 1/s$ both), we can approximate

$$ik - \frac{1}{r} \approx ik - \frac{1}{s} \approx ik$$

Then the *Fresnel-Kirchhoff diffraction formula* reads

$$U(P) = \frac{1}{4\pi}\int\limits_S \left(U\frac{\partial}{\partial \boldsymbol{n}}\frac{e^{iks}}{s} - \frac{e^{iks}}{s}\frac{\partial U}{\partial \boldsymbol{n}}\right)\mathrm{d}S \tag{39}$$

$$= \frac{iA}{2\lambda}\int\limits_{\mathcal{A}} \left(\frac{e^{ik(r+s)}}{rs}\left(\cos(n,s) - \cos(n,r)\right)\right)\mathrm{d}S \tag{40}$$

where (n, s) is the angle between $\boldsymbol{n}$ and $\boldsymbol{s}$ and similarly (n, r) (see figure 7b).

2.1.4 Fraunhofer Diffraction

We now make further assumptions in order to be able to evaluate eqn. (40) analytically. Assume that the dimensions of the opening are large compared to λ but are small compared to r, s. That is to say, $\lambda \ll a \ll r, s$ where a is the characteristic size of the aperture (radius or edge length).

The assumption $a \ll r, s$ implies $\cos(n, s) - \cos(n, r)$ varies very little over the aperture (see figure 8) and hence

$$\cos(n,s) - \cos(n,r) \approx \cos\big(n,s'\big) - \cos\big(n,r'\big) \tag{41}$$

and similarly $1/rs \approx 1/r's'$. Hence

$$U(P) \approx \frac{iA}{2\lambda}\frac{\big(\cos\big(n,s'\big) - \cos\big(n,r'\big)\big)}{r's'}\int\limits_{\mathcal{A}} e^{ik(r+s)}\,\mathrm{d}S \tag{42}$$

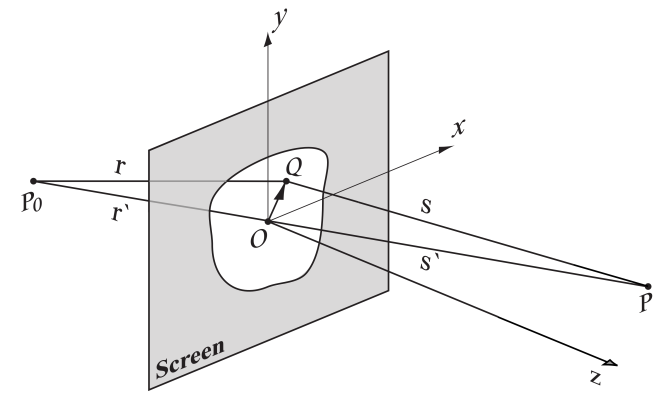


Figure 8: Diffraction at an aperture in a plane screen. We introduce Cartesian coordinates with origin O, where the xy-plane and the aperture plane coincide and positive z direction points into the half-space containing P. An arbitrary point Q within the aperture is specified by the coordinates ξ, η **Ivanov2016ElementsOD**.

The integral in eqn. (42) is still only numerically computable and so we make further assumptions in order to approximate $e^{ik(r+s)}$. First, we introduce a Cartesian coordinate system with origin O in the aperture, with the xy-axes in the plane of the aperture. We then choose the positive z direction to point into the half-space that contains the point P (see figure 8). Let $P_0 = (x_0, y_0, z_0)$, $P = (x, y, z)$, and an arbitrary point $Q = (\xi, \eta)$ in the aperture plane. Then

$$r^2 = (x_0 - \xi)^2 + (y_0 - \eta)^2 + z_0^2 \tag{43}$$

$$s^2 = (x - \xi)^2 + (y - \eta)^2 + z^2 \tag{44}$$

$$(r')^2 = x_0^2 + y_0^2 + z_0^2 \tag{45}$$

$$(s')^2 = x^2 + y^2 + z^2 \tag{46}$$

and hence

$$r^2 = (r')^2 - 2(x_0\xi + y_0\eta) + \xi^2 + \eta^2 \tag{47}$$

$$s^2 = (s')^2 - 2(x\xi + y\eta) + \xi^2 + \eta^2 \tag{48}$$

Taking into account the assumption $a \ll r, s$, we can neglect the distance $\xi^2 + \eta^2$ of Q from the origin:

$$r^2 = (r')^2 - 2(x_0\xi + y_0\eta) \tag{49}$$

$$s^2 = (s')^2 - 2(x\xi + y\eta) \tag{50}$$

Now expanding expanding r, s using the power series for $\sqrt{1+x}$ and neglecting higher order terms owing to $a^2/\lambda \ll r, s$

$$r \approx r' - \frac{x_0\xi + y_0\eta}{r'} \tag{51}$$

$$s \approx s' - \frac{x\xi + y\eta}{s'} \tag{52}$$

Finally substituting r, s into eqn. (42) the *Fraunhofer's diffraction integral* is defined

$$U(p,q) := U(P) \approx C\int\limits_{\mathcal{A}} \exp\big\{ik(p\xi + q\eta)\big\}\,\mathrm{d}\xi\,\mathrm{d}\eta \tag{53}$$

where

$$C = \frac{\mathrm{i}A}{2\lambda}\frac{\cos(n, s') - \cos(n, r')}{r's'}e^{\mathrm{i}k(r'+s')} \tag{54}$$

and

$$p = \frac{x_0}{r'} + \frac{x}{s'} \quad q = \frac{y_0}{r'} + \frac{y}{s'} \tag{55}$$

are the coordinates p, q of a point P in the diffraction pattern.

Notice that eqn. (53) can be rewritten as a Fourier transform of the characteristic function[15] $M(\xi, \eta)$ of the aperture:

$$U(p, q) \coloneqq U(P) \approx C\int M(\xi, \eta)\exp\left\{\mathrm{i}k(p\xi + q\eta)\right\}\mathrm{d}\xi\,\mathrm{d}\eta \tag{56}$$

Thus the diffraction pattern is the spatial Fourier transform of the aperture (called the *optical transfer function*).

2.1.5 *Diffraction through a Circular Aperture*

We now investigate Fraunhofer diffraction through a circular aperture. Using polar coordinates (ρ, θ) to parameterize the aperture coordinates (ξ, η) and (R, ψ) to parameterize the diffraction pattern point P coordinates (p, q)

$$\rho\cos(\theta) = \xi \quad \rho\sin(\theta) = \eta \tag{57}$$

$$R\cos(\psi) = p \quad R\sin(\psi) = q \tag{58}$$

Note that R, called the *angular resolution*, represents maximum resolvable distance in the diffraction pattern. We express the argument of the exponential in the integrand of eqn. (53)

$$\begin{aligned}
-\mathrm{i}k(p\xi + q\eta) &= -\mathrm{i}k\rho(p\cos(\theta) + q\sin(\theta)) \\
&= -\mathrm{i}k\sqrt{p^2+q^2}\left(\frac{p\cos(\theta)}{\sqrt{p^2+q^2}} + \frac{q\sin(\theta)}{\sqrt{p^2+q^2}}\right) \\
&= -\mathrm{i}k\rho R\left(\cos(\psi)\cos(\theta) + \sin(\psi)\sin(\theta)\right) \\
&= -\mathrm{i}k\rho R\cos(\theta - \psi)
\end{aligned} \tag{59}$$

The integral in eqn. (53) then becomes

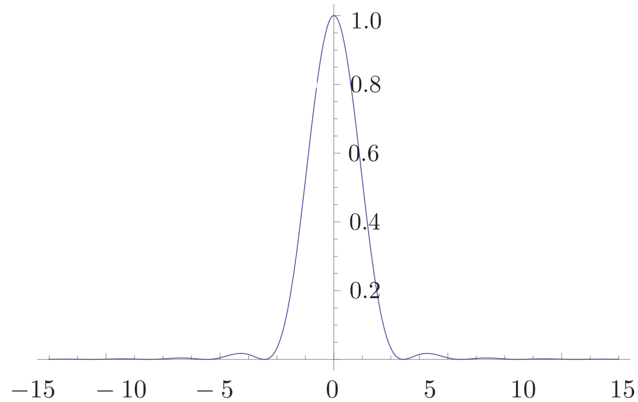
$$U(R, \psi) \coloneqq C\int_0^a\int_0^{2\pi} e^{-\mathrm{i}k\rho R\cos(\theta-\psi)}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta \tag{60}$$

where a is the radius of the aperture. Using *Bessel functions of the first kind* J_0, J_1 we have

$$U(R) \coloneqq 2\pi C\int_0^a J_0(k\rho R)\rho\,\mathrm{d}\rho \tag{61}$$

$$= \pi C a^2\frac{2J_1(kaR)}{kaR} \tag{62}$$

[15] $M(\xi, \eta) = 1$ in $\mathcal{A}$ and 0 otherwise



(a) Modulation transfer function $y = \left[\frac{2J_1(x)}{x}\right]^2$ for a circular aperture **Ivanov2016ElementsOD**.

(b) Simulated Fraunhofer diffraction from a circular aperture **Ivanov2016ElementsOD**.

Figure 9: Fraunhofer diffraction through a circular aperture.

with intensity (see figure 9a)

$$I(R) \propto \left|U(R)\right|^2 = I_0 \left[\frac{2J_1(kaR)}{kaR}\right]^2 \tag{63}$$

where I_0 is peak intensity (which happens to be the intensity at the central peak).

Note that the central maximum of eqn. (63) comprises 98% of the transmitted power. Therefore the effect of a circular aperture is to disperse a point source into a large circular shape (see figure 9b). Note also that the first zero appears at $kaR \approx 1.22\pi$ and so we recover Rayleigh's criterion as

$$R_{min} \approx 1.22\frac{\pi}{ka} = 1.22\frac{\lambda\pi}{2\pi a} = 1.22\frac{\lambda}{D} \tag{64}$$

where $D = 2a$ the diameter of the aperture.

Index