Paraxial light distribution in the focal region of a lens: a comparison of several analytical solutions and a numerical result

Yang Wu* and Damien P. Kelly

Institut für Mikro-und Nanotechnologien, Technische Universität Ilmenau, Ilmenau, Germany

(Received 23 April 2014; accepted 3 July 2014)

The distribution of the complex field in the focal region of a lens is a classical optical diffraction problem. Today, it remains of significant theoretical importance for understanding the properties of imaging systems. In the paraxial regime, it is possible to find analytical solutions in the neighborhood of the focus, when a plane wave is incident on a focusing lens whose finite extent is limited by a circular aperture. For example, in Born and Wolf's treatment of this problem, two different, but mathematically equivalent analytical solutions, are presented that describe the 3D field distribution using infinite sums of U_n and V_n type Lommel functions. An alternative solution expresses the distribution in terms of Zernike polynomials, and was presented by Nijboer in 1947. More recently, Cao derived an alternative analytical solution by expanding the Fresnel kernel using a Taylor series expansion. In practical calculations, however, only a finite number of terms from these infinite series expansions is actually used to calculate the distribution in the focal region. In this manuscript, we compare and contrast each of these different solutions to a numerically calculated result, paying particular attention to how quickly each solution converges for a range of different spatial locations behind the focusing lens. We also examine the time taken to calculate each of the analytical solutions. The numerical solution is calculated in a polar coordinate system and is semi-analytic. The integration over the angle is solved analytically, while the radial coordinate is sampled with a sampling interval of $\Delta\rho$ and then numerically integrated. This produces an infinite set of replicas in the diffraction plane, that are located in circular rings centered at the optical axis and each with radii given by $2\pi m/\Delta\rho$, where m is the replica order. These circular replicas are shown to be fundamentally different from the replicas that arise in a Cartesian coordinate system.

Keywords: propagating methods; diffraction; analytical solution

1. Introduction

The complex amplitude distribution in the focal region of a converging lens remains an important problem in optics today [1]. Although the problem is well known and has been analyzed, using many different techniques [2–13], the most suitable approach depends on the nature of the problem being investigated. When an optical system is being analyzed, where the lens has a low numerical aperture, a scalar analysis is often sufficient to accurately model the behavior of the focused field. In other problems, however, for example; the design of DVD optics [14], a full vectorial treatment is required to understand and design the necessary optics [15–17].

In the past and increasingly in more recent times (due to the widespread availability of relatively inexpensive computing power), numerical calculation techniques have been used to examine the finer details of the form of the electromagnetic distribution by solving complex vectorial diffraction integrals or using grid or mesh techniques such as Finite Element Methods or the Finite Difference Time

Domain method [18]. Traditionally, however, efforts were made to find analytical solutions to scalar diffraction problems so that insight into the characteristics of the problem could be brought out. For example, in Nijboer's approach, it was shown how an analytical solution could be used to balance higher order aberrations against lower order aberrations so that the overall imaging performance of the lens approached the ideal diffraction limited case. Other approaches based on orthogonal expansions are outlined here [19,20]. Such an insight would not naturally arise from a numerical investigation. Another observation is that with increased computational power, numerical approaches often face computational limits, with apparently straight-forward diffraction calculations. Hence there are several good reasons for investigating diffraction problems using analytical techniques: (i) more insight into the diffraction process is provided, (ii) a 'correct' analytical solution serves as an excellent way of testing the predictions of numerical calculations in specific cases, and (iii) sometimes it may be desirable to use a combination of analytical

*Corresponding author. Email: yang.wu@tu-ilmenau.de

and numerical techniques to best model a given diffraction problem.

In scalar theory the Kirchhoff–Fresnel diffraction integral is often used to model the propagation of coherent monochromatic light in free space. In many cases, however, the less restrictive paraxial approximation still provides sufficiently accurate results. In this instance, the Fresnel transform can be used to model the diffraction and the propagation of light. In this manuscript, we assume that the paraxial approximation is valid and use the Fresnel transform to model the diffraction and propagation of light in free space.

When examining the diffraction of monochromatic light from a circular aperture, Lommel [1] introduced two functions to solve the integral in 1885. In 1947, Nijboer [21,22] gave another solution of the integral using Zernike polynomials. With this solution, we can calculate the 3D field distribution behind a perfect converging lens. This solution can also be extended to include the effects of aberration. In 2002, Cao [9] developed another series expansion, and used it to solve the diffraction integral.

These analytical solutions are all mathematical equivalent. However, none of them are a closed form solution and hence require an infinite sum of series terms. To calculate the result from one of these solutions, in practice, requires truncating the infinite series when a desired accuracy has been achieved. As we shall see each solution, has different convergence properties, which vary depending on the spatial location in the output plane. In this manuscript, we find some of these solutions are stable and converge quickly in some regions, however, do not perform so well in other locations. We analyze the reasons behind this and provide guidelines on how to choose the appropriate solution for a given spatial location.

In order to compare the performance of each of these analytical solutions, it was necessary to measure their accuracy against a common benchmark. We therefore also examine the numerical integration of the diffraction integral in polar coordinates. The radial variable is discretized in S uniform steps, $\Delta\rho$, over the range: $0 \leq \rho \leq 1$. By increasing S , we improve the accuracy of the numerical solution. Here, we find that this discretization process produces an infinite set of replicas in the output diffraction domain that can overlap with each other if S is not large enough. When that happens, the replicas add coherently leading to erroneous numerical results. We derive an expression describing the effect of this sampling process and find that the replicas are located in concentric circles about the origin with radii of $\lambda z / \Delta\rho$, where λ is the wavelength of the light, z the propagation distance and m is the replica order (the replica radii are given here in physical coordinates, however, later in the paper, we find it more convenient to use a normalized coordinate system). These replicas have fundamentally different properties from the replicas that arise in a Cartesian coordinate system, see for example Ref. [10]. This theoretical

Figure 1. Schematic view of the diffractive geometric.

result may have significant implications for the iterative design of optical elements in polar coordinate systems and understanding the loss of information that occurs when using unconventional sensors that measure at discrete radial locations.

We use Figure 1 to illustrate the optical system that we wish to analyze. A plane wave is incident on a perfect converging spherical lens. In the geometrical approximation, the focused light would converge to an ideal point source at the focus. In paraxial wave optics, however, diffraction introduced by the finite extent of the focusing lens aperture, causes the complex amplitude distribution in the focal plane to spread out over the plane. If one traces the intensity distribution in the focal plane, moving radially out from the focus, one observes a bright central lobe which changes to series of bright and dark rings as one moves out along the plane. We wish to examine the distribution over the entire focal volume, which requires a more complex analytical solution. We begin our analysis by defining the Fresnel transform in same manner as Ref. [23] (see Chapter 4),

$$A(x, y, z) = \frac{e^{ikz}}{i\lambda z} \iint_S A_0(x', y') e^{\frac{ik}{2z}[(x'-x)^2 + (y'-y)^2]} dx' dy', \quad (1)$$

where $k = 2\pi/\lambda$ is the wavenumber and i is the imaginary unit. We now wish to describe the operation of a perfect converging thin lens. Again from Ref. [23] in Chapter 5 we see that,

$$A_0(x', y') = e^{-i\frac{k}{2f}(x'^2 + y'^2)}, \quad (2)$$

where f is the focal length of the lens. Due to the symmetrical property of the complex field, we use a cylindrical coordinate system instead of the Cartesian coordinate system, with following relations:

$$x = r \cos \phi, \quad (3)$$

$$y = r \sin \phi, \quad (4)$$

$$x' = R \cos \theta, \quad (5)$$

$$y' = R \sin \theta. \quad (6)$$

Substituting Equations (2)–(6) into Equation (1), then using a property of the Bessel function (see Ref. [23], Chapter 2, pp 28, 2–30),

$$J_0(\alpha) = \frac{1}{2\pi} \int_0^{2\pi} e^{-i\alpha \cos(\theta-\phi)} d\theta \quad (7)$$

we arrive at the following result,

$$A(r, z) = \frac{2\pi}{i\lambda z} e^{jk\left(z + \frac{r^2}{2z}\right)} \int_0^a e^{\frac{i\pi R^2}{\lambda} \left(\frac{1}{z} - \frac{1}{f}\right)} J_0\left(\frac{2\pi Rr}{\lambda z}\right) R dR, \quad (8)$$

where J_0 is the Bessel function of the first kind. We introduce the normalized parameter ρ ,

$$0 \leq \rho = R/a \leq 1. \quad (9)$$

We also use two normalized optical coordinates u, v to make the integral concise,

$$u = \frac{2\pi a^2}{\lambda} \left(\frac{1}{f} - \frac{1}{z} \right), \quad (10)$$

$$v = \frac{2\pi ar}{\lambda z}. \quad (11)$$

Substituting Equations (9)–(11) into Equation (8), we get,

$$A(u, v) = e^{jkc} \frac{2\pi a^2 - \lambda u f}{i\lambda f} T(u, v). \quad (12)$$

where $c = f(v^2\lambda^2 + 2\pi^2 a^2)/\pi(2\pi a^2 - u\lambda f)$ and $T(u, v)$ is,

$$T(u, v) = \int_0^1 e^{-\frac{i u \rho^2}{2}} J_0(v\rho) \rho d\rho. \quad (13)$$

Using Equation (12), we can calculate the complex amplitude of the points behind the lens. In the next section, however, we first concentrate on finding a solution to the integral $T(u, v)$, using four different analytical methods. As shown in Figure 1, the diffraction field can be divided into two areas: if $|u| > |v|$, it is in the illumination area; And if $|u| < |v|$, it is in the geometrical shadow area. Note that, if $|u| = |v|$, the point is at the boundary between the illumination area and the shadow area, and that the focal point is located at the coordinates $u = 0, v = 0$. It seems reasonable to expect that the various analytical solutions have different properties in each of these regions.

We have organized the manuscript in the following manner: In Section 2, we derive four different but mathematically equivalent analytical solutions for $T(u, v)$. Then in Section 3, the relative performance of each of these solutions is compared, i.e. in relation to their numerical accuracy, speed of convergence for a range of different spatial locations behind the focusing lens and the speed of computation. In Section 4, we examine the implementations of the numerical solution in detail. Because the integral is represented discretely, it produces infinite a set of replicas near the original, the replicas are discussed and analyzed. We finish with a brief conclusion.

2. Analytical solution for $T(u, v)$

In this section, we derive four different analytical solutions for the diffraction integral, Equation (13).

Figure 2. A contour plot of $|T_{num}|^2$ in the focal region. (The colour version of this figure is included in the online version of the journal.)

I. The first Lommel solution

$T(u, v)$ is separated into real and imaginary parts [1],

$$T(u, v) = C(u, v) + iS(u, v), \quad (14)$$

where

$$C(u, v) = \int_0^1 J_0(v\rho) \cos\left(\frac{1}{2}u\rho^2\right) \rho d\rho, \quad (15)$$

$$S(u, v) = \int_0^1 J_0(v\rho) \sin\left(\frac{1}{2}u\rho^2\right) \rho d\rho. \quad (16)$$

Using integration by parts,

$$\int A(x)B'(x)dx = A(x)B(x) - \int A'(x)B(x)dx. \quad (17)$$

Where $A(x) = \cos(\frac{1}{2}u\rho^2)$ and $B'(x) = J_0(v\rho)$, and using the property of Bessel function (Ref. [24], p 18),

$$\frac{d}{dx} [x^{n+1} J_{n+1}(x)] = x^{n+1} J_n(x). \quad (18)$$

we get that,

$$C(u, v) = \frac{1}{v} \times \left[J_1(v) \cos\left(\frac{1}{2}u\right) + u \int_0^1 \rho^2 J_1(v\rho) \sin\left(\frac{1}{2}u\rho^2\right) d\rho \right]. \quad (19)$$

Using the Equation (18) and integration by parts again, we get finally,

$$C(u, v) = \frac{\cos\left(\frac{1}{2}u\right)}{u} U_1(u, v) + \frac{\sin\left(\frac{1}{2}u\right)}{u} U_2(u, v), \quad (20)$$

where $U_n(u, v)$ is the first Lommel function, as discussed in Refs. [24,25]:

$$U_n(u, v) = \sum_{s=0}^{\infty} (-1)^s \left(\frac{u}{v}\right)^{n+2s} J_{n+2s}(v). \quad (21)$$

Similar we get that,

$$S(u, v) = \frac{\sin\left(\frac{1}{2}u\right)}{u} U_1(u, v) + \frac{\cos\left(\frac{1}{2}u\right)}{u} U_2(u, v), \quad (22)$$

Substituting Equations (20)–(22) into Equation (14), we get the first of our analytical solution of Equation (13),

$$T_1(u, v) = \frac{U_1(u, v) + U_2(u, v)i}{u} e^{-\frac{iu}{2}}, \quad (23)$$

where the subscript ‘1’ indicate that this is the first solution of the integral. Note that the term $J_{n+2s}(v)$ in the first Lommel function converge to zero. In the illumination area, $|u| < |v|$, the term $\left(\frac{u}{v}\right)^{n+2s}$ converge to zero, too. It, therefore, is appropriate to use this solution in the illumination area [26,27].

II. The second Lommel solution

We now provide a solution to the Equation (13) using the second Lommel function. Integration by parts is still used. But with $A(x) = J_0(v\rho)$ and $B'(x) = \cos(\frac{1}{2}u\rho^2)$ [1,28], or using alternatively the equations from Ref. [24], Chapter 16, pp 537–542,

$$\frac{V_1(u, v) - V_0(u, v)i}{u} e^{-\frac{iu}{2}} = \int_1^{\infty} J_0(v\rho) e^{-\frac{iu\rho^2}{2}} \rho d\rho, \quad (24)$$

$$\int_0^{\infty} J_0(v\rho) e^{-\frac{iu\rho^2}{2}} \rho d\rho = -\frac{i}{u} e^{\frac{iv^2}{2u}}. \quad (25)$$

$V_n(u, v)$ is the second Lommel function,

$$V_n(u, v) = \sum_{s=0}^{\infty} (-1)^s \left(\frac{v}{u}\right)^{n+2s} J_{n+2s}(v). \quad (26)$$

Using Equations (25)–(24), the solution is,

$$T_2(u, v) = -\frac{i}{u} e^{\frac{iv^2}{2u}} - \frac{V_1(u, v) - V_0(u, v)i}{u} e^{-\frac{iu}{2}}. \quad (27)$$

Using the second Lommel function, the term $\left(\frac{v}{u}\right)^{n+2s}$ appears instead of $\left(\frac{u}{v}\right)^{n+2s}$ in Equation (21), this indicates the solution converges in the shadow area. So we propose to use this solution in the geometric shadow area [26,27].

III. Nijboer’s solution with Zernike polynomials

We now turn out attention to employing the third solution. In Equation (13), we have the term

$$e^{-\frac{iu\rho^2}{2}} = e^{-\frac{iu}{4}} e^{-\frac{iu}{4}(2\rho^2-1)}. \quad (28)$$

and from Ref. [25] and Ref. [29], we note that,

$$e^{ic(2\rho^2-1)} = \sqrt{\frac{\pi}{2c}} \sum_{n=0}^{\infty} i^n (2n+1) J_{n+\frac{1}{2}}(c) P_n(2\rho^2-1), \quad (29)$$

where P_n is the Legendre polynomial, which is related to the Zernike polynomial (Ref. [22]), as follows

$$P_n(2\rho^2-1) = R_{2n}^0(\rho). \quad (30)$$

$R_{2n}^0(\rho)$ is the Zernike polynomials,

$$R_n^m(\rho) = \sum_{k=0}^{\frac{n-m}{2}} (-1)^k \frac{(n-k)!}{k! \left(\frac{n+m}{2} - k\right)! \left(\frac{n-m}{2} - k\right)!} \rho^{n-2k}, \quad (31)$$

Substituting Equations (28)–(30) into Equation (13), and introducing the parameter $c = -u/4$,

$$T(u, v) = \sum_{n=0}^{\infty} e^{-\frac{iu}{4}} \sqrt{\frac{2\pi}{-u}} i^n (2n+1) J_{n+\frac{1}{2}}\left(-\frac{u}{4}\right) \times \int_0^1 R_{2n}^0(\rho) J_0(v\rho) \rho d\rho. \quad (32)$$

Another relationship, which is important in the diffraction theory of aberrations [22], is that

$$\int_0^1 R_{2n}^0(\rho) J_0(v\rho) \rho d\rho = (-1)^n \frac{J_{2n+1}(v)}{v}. \quad (33)$$

Substituting Equation (33) into Equation (32) gives third solution for Equation (13),

$$T_3(u, v) = \sum_{n=0}^{\infty} e^{-\frac{iu}{4}} \sqrt{\frac{2\pi}{-u}} (-i)^n (2n+1) J_{n+\frac{1}{2}}\left(-\frac{u}{4}\right) \frac{J_{2n+1}(v)}{v}. \quad (34)$$

IV. Cao’s solution

Recently, Cao defined a family of generalized Jinc functions $w_n(v)$ as follows [9],

$$w_n(v) = \frac{1}{v^{2n+2}} \int_0^v t^{2n+1} J_0(t) dt. \quad (35)$$

The zero order ($n = 0$) of this function is the traditional Jinc function. We rearrange the integral as follows, with $\rho = t/v$,

$$w_n(v) = \int_0^1 \rho^{2n+1} J_0(v\rho) d\rho. \quad (36)$$

This can also be rewritten as the form of a polynomial [9],

$$w_n(v) = \sum_{m=0}^n (-2)^m \frac{n!}{(n-m)!} \frac{J_{m+1}(v)}{v^{m+1}}. \quad (37)$$

Using the Taylor series expansion for $e^{-\frac{1}{2}iup^2}$, it can be shown that

$$T(u, v) = \sum_{n=0}^{\infty} \frac{\left(-\frac{1}{2}iu\right)^n}{n!} \int_0^1 \rho^{2n+1} J_0(v\rho) d\rho. \quad (38)$$

Figure 3. Intensity along the focal plane ($u = 0$) when five series terms are used in the calculations. (The colour version of this figure is included in the online version of the journal.)

Substituting Equations (36)–(37) into Equation (38) we get a fourth solution of the form,

$$T_4(u, v) = \sum_{n=0}^{\infty} \frac{\left(-\frac{1}{2}iu\right)^n}{n!} \sum_{m=0}^n (-2)^m \frac{n!}{(n-m)!} \frac{J_{m+1}(v)}{v^{m+1}}. \quad (39)$$

Applying these four solutions, we can calculate the intensity distribution, to within a constant multiplicative factor, at any plane behind the lens,

$$I_T(u, v) = |T(u, v)|^2. \quad (40)$$

3. A comparison of the solutions

It must be emphasized that, all of the four analytical solutions, derived in the previous section, are mathematically equivalent. They all involve summing over an infinite number of Bessel functions to solve the integral exactly. In practice, however, only a finite number of terms can be used in the calculation and so the rate of convergence of each solution is of very real practical importance. It is shown that the rate at which each method converges has strong spatial dependence and we establish a relationship between the spatial location and the convergence rate for each solution.

In order to compare the relative performance of each of the different analytical solutions, we have found it useful to compare the results to those found by directly numerically integrating the diffraction integral Equation (13). We refer to this numerical solution as T_N , and present one result for such a calculation in Figure 2. In Section 4, we detail how this calculation is implemented numerically.

We first, however, compare the analytical solutions for given numbers of series terms along different cross-sections through the focus. We then examine how many series terms are required for each solution to meet the same convergence criteria at a set of specific points, P_1 , P_2 and P_3 , see Figure 2. Various cross-sections and spatial locations are

Figure 4. Intensity distribution along the optical axis ($v = 0$) retaining five series terms. (The colour version of this figure is included in the online version of the journal.)

Figure 5. Intensity at the boundary ($|u| = |v|$) of the illumination area. Again only five series terms are used in the calculations. (The colour version of this figure is included in the online version of the journal.)

identified in Figure 2. In Section 3.3, we examine the differences between each analytical solution and the numerically calculated results and produce a set of error maps so that the significants of these errors can be visualized. In Section 3.4, we examine the time taken to calculate the results using each of the analytical solutions.

3.1. A comparison of the analytical solutions along several cross-sections through the focal region

We begin our comparison of the analytical solutions by examining their predictions along three different cross-sections, each of which passes through the focal point. Three different cross-sections are shown in Figure 2 and defined as: A. along the focal plane ($u = 0$), B. along the optical axis ($v = 0$) and C. along the boundary of the illumination area ($|u| = |v|$).

A. Along focal plane ($u = 0$)

For each solution only the first five terms of the analytical series is used for the calculation and we examine $|T(u, v)|^2$ over the range $-30 < v < 30$. We note that when $u = 0$,

Figure 6. Intensity at the boundary ($|u| = |v|$) of the illumination area with ten terms. (The colour version of this figure is included in the online version of the journal.)

one must perform a limiting operation in order to arrive at the correct result for the first Lommel function, second Lommel function and Zernike polynomial solutions. This is because of the presence of the $1/u$ term. In practice, we use $u = 0.0001$ close to $u = 0$. Figure 3 shows the log of the intensity distribution in the focal plane. The solutions, $|T_1|^2$, $|T_3|^2$ and $|T_4|^2$ converge quickly and closely agree with each other in Figure 3. However the second Lommel function, $|T_2|^2$, does not converge. As when $u \approx 0$, and $v \gg u$, the term $(v/u)^{n+2s}$ in Equation (26) requires that more terms be included so as to achieve convergence.

B. Along the optical axis ($v = 0$)

As in the previous calculation along the focal plane, we again use the first five series terms in each of the analytical solutions. The calculation range is $-30 < u < 30$. Again due to the presence of the $1/v$ term, we only calculate the result for $v = 0.0001$ instead of $v = 0$. The comparison is shown in Figure 4. In this region, $|T_2|^2$ and $|T_3|^2$ converge quickly (brown-yellow distribution). But for the other two, five terms are not sufficient to ensure that the analytical solution converges. The T_4 solution converges over the range $-6 < v < 6$ when 5 terms are used while the $|T_1|^2$ solution performs better converging over the range $-9 < v < 9$ (see the green and blue plots in Figure 4 respectively). Therefore, we conclude that it is better to use $|T_2|^2$ or $|T_3|^2$ near the optical axis, i.e. when $v = 0$.

C. Along the boundary of the illumination area ($|v| = |u|$)

In this region, we have $|v| = |u|$, and the calculation range is $-30 < v = u < 30$. The results from the four analytical solutions are presented in Figure 5. The result predicted by $|T_4|^2$ converges (green plot) for spatial locations near the focal region, however, does poorly for most other spatial locations along the cross-section. In fact each solution converges at different rates along this cross-section. To examine this feature in more detail, we calculate the intensity distributions again, however, now using ten terms, see Figure 6. It is clear that the $|T_3|^2$ distribution does not

Figure 7. Converge process at P_1 as a function of series terms. (The colour version of this figure is included in the online version of the journal.)

Figure 8. Converge process at P_2 as a function of series terms. (The colour version of this figure is included in the online version of the journal.)

change, however, the additional terms improve the convergence of the other three solutions. So we see that the $|T_3|^2$ solution converges most quickly along this boundary region while the $|T_4|^2$ solution performs the worst in this region.

3.2. Convergence rate at three specific spatial locations: P_1 , P_2 and P_3

In the previous section, we examined the performance of the analytical solutions along three different cross-sections of the focal distribution all of which pass through the focus. In this section, we wish to examine in more detail how each solution converges as a function of the number of series terms retaining. We choose three specific spatial locations as indicated in Figure 2.

P_1 . Near the focal plane $|u| = 5$, $|v| = 25$

From Figure 3 we see that at the focal plane the performance of the T_1 , T_3 and T_4 solutions are good, while the T_2 solution does not converge. A question arises as to whether we can make T_2 converge by simply including more series terms? In the Figure 7, we calculate the intensity at point P_1 ($|u| = 5$, $|v| = 25$) near the focal plane, and we see that the T_1 , T_3 and T_4 solutions converge rapidly. Increasing the

Figure 9. Converge process at P_3 as a function of series terms. (The colour version of this figure is included in the online version of the journal.)

Figure 10. Error map of solution T_1 , $\log(|T_1 - T_N|)$ is plotted. (The colour version of this figure is included in the online version of the journal.)

number of terms used with the T_2 solution does not seem to help and the solution does not converge. This lack of convergence is numerical in nature and is related to the number of significant figures that can be used to represent a rational number in a computer. ‘The floating-point machine reals follow the IEEE standard Double Format using 53 bits of machine storage (including one hidden bit) with a machine epsilon of 2^{-52} (which is approximately 2.2×10^{-16})’ [30], and hence the inaccuracy is of the order 10^{-16} . In most cases, this difference does not cause numerical instability, however here, if $v \gg u$, the second Lommel solution produces values that are larger than 10^{16} . In the T_2 series solution at P_1 , successive series terms are very large and opposite in sign. Adding large numbers of opposite sign means that the resulting small round-off errors can quickly lead to numerical stability problems. While there are means of improving the numerical accuracy, they are not pursued here and we say that second Lommel function does not provide a numerically stable solution in this region.

Figure 11. Error map of solution T_2 , $\log(|T_2 - T_N|)$ is plotted. (The colour version of this figure is included in the online version of the journal.)

Figure 12. Error map of solution T_3 , $\log(|T_3 - T_N|)$ is plotted. (The colour version of this figure is included in the online version of the journal.)

P_2 . Near the optical axis $|u| = 25$, $|v| = 5$

Now we turn to look at the point near the optical axis $|u| = 25$, $|v| = 5$. From Figure 8 we see that both the T_2 and T_3 solutions converge quickly needing only a few series terms to reach a final stable value. The T_1 and T_4 solutions converge more slowly. If we compare the manner in which the T_2 solution in Figure 7, and the T_1 and T_4 solutions in Figure 8, converge, we can see a similar trend. Although the T_1 and T_4 solutions converge for all the cases presented here, we expect from this observation that similarly there are regions where these solutions will fail due to similar numerical stability reasons. We have not found a simple means of defining these unstable regions. Hence we adopt the rule that if the solution does not stabilize as the series are increased, one should choose an alternative solution, see Section 3.3.

Figure 13. Error map of solution T_4 , $\log(|T_4 - T_N|)$ is plotted. (The colour version of this figure is included in the online version of the journal.)

Figure 14. Comparison of the computation time for the four analytical methods. (The colour version of this figure is included in the online version of the journal.)

P_3 . At the boundary of geometric shadow $|u| = 25$, $|v| = 25$ This point is located away from the optical axis and the focal plane, and we can see that more terms are required to achieve the same level of accuracy (see Figure 9). The T_1 and T_2 solutions need about 14 terms to reach the correct answer. The T_3 solution converges the fastest, needing only ten terms. The T_4 solution needs more than 30 terms to converge. We note the nature of the convergence is similar in form to T_2 in Figure 7. As we move out further along curve C, and away from the focus, this T_4 solution becomes increasingly unstable and will eventually reach a point where it does not converge due to the digital limit discussed before.

3.3. Error maps

In this section, we determine regions where the solutions converge well and where the maximum errors are to be found. We compare the four analytical solutions, each cal-

Figure 15. Error of numerical solution with increasing number of samples s . (The colour version of this figure is included in the online version of the journal.)

Figure 16. The diffractive field with replicas at the focal plane using 30 samplings. (The colour version of this figure is included in the online version of the journal.)

culated using only ten terms, with a numerical solution T_N . As we show later, the error in this numerical solution is of the order 10^{-6} , and so we can provide a definite error value for the analytical solutions. In Figures 10–13 the resulting errors can be seen and regions of convergence identified. We make the following observations:

- (1) The $T_1(u, v)$ solution converges well within the illuminated region behind the lens.
- (2) The opposite appears to hold for the $T_2(u, v)$ solution, where the largest errors are within the illuminated region. It does however converge well in the geometrical shadow.
- (3) The Nijboer solution provides the most stable solution over the ranges we have examined. The error increases as one move away from the focus of the lens.

- (4) The Cao solution seems to be have a performance that is a combination of the Lommel solutions. It is not as robust as the Nijboer and the errors appear to be symmetrical about the $u = 0$ plane.

From these results, we conclude that while all the solutions are mathematically equivalent they require a finite number of series terms to be used in practical calculations. When only a finite number of terms are used, the Nijboer solution has the most robust convergence properties.

3.4. Comparison of the calculation time

In addition to the accuracy, the total time taken to calculate a numerical result from the analytical solutions is an important parameter. Here we examine how the computation time varies for each of the four analytical solutions. We calculate 100×100 points distributed in the focal region $-20 < u < 20$, $-20 < v < 20$. The computer we used is an Intel(R) Core(TM) i7-2600K, and the computing platform is Matlab. The computation time for the four solutions are plotted as a function of the number of series terms, see Figure 14. The T_4 solution takes the most time of all, then follow the Nijboer's solution and two Lommel solutions, the time spent of the three solutions are similar. The Nijboer's solution performs very good, not only the fast computation time but also the relative stable convergence properties, and noted that the time spent can be significantly improved using look-up tables and other optimizing steps, see [15].

4. Numerical method to solve the integral

In this section, we examine in detail how the integral in Equation (13) can be calculated numerically. We represent the variable ρ with S samples over the range $0 < \rho < 1$, in steps of $\Delta\rho = 1/S$, giving us the radial spatial vector $[\rho_1, \rho_2, \rho_3, \dots, \rho_S]$, which is defined as $\rho_n = n/S$. This results in the following discrete expression,

$$T_N^S(u, v) = \Delta\rho \sum_{n=1}^S e^{-\frac{i u}{2} \rho_n^2} J_0(v \rho_n) \rho_n, \quad (41)$$

and it is used to generate the plot in Figure 2. We expect that, the more samples S we take, the more accurate our numerical solution becomes, however it takes obviously more time. We take $S = 20,000$ as the reference, and see how the error $E(s) = |T_N^s - T_N^{20,000}|$ changes with increasing number of samples in Figure 15. We can see that as S increases the error drops dramatically. In order to ensure the high accuracy of the numerical solution, we take $S = 1000$ samples, which keeps the accuracy to a level with 10^{-6} from Figure 15. This solution is used to generate Figure 2. Also this solution (which is correct to a level of 10^{-6}) is used to produce the error plots in Section 3.3. Using numerical method, we do not need to worry about the converge property at different locations, and the time spent (about 4 s using 1000

samples) is in the same level of the corresponding analytical solution T_1 , T_2 and T_3 (about 2 s using ten terms), which is a little different from the situation when solving high aperture vectorial diffraction integrals using series expansion [31]. In that case, direct integration is more quick than the analytical solutions. However the numerical approach produces an infinite set of replicas in the diffraction plane. The replicas produced by this numerical integration are shown to be fundamentally different from the replicas that arise in a Cartesian coordinate system (x, y) , which occurs in practice in digital holography and we refer the reader to the following publications for more detail [32,33]. The errors introduced by the replicas are shown to be located along an infinite set of concentric circles, each centered at $v = 0$, and with a radius $2\pi m / \Delta\rho$, $m = 1, 2, 3, \dots$

We begin our analysis of these radial replicas by rewriting Equation (41) as follows,

$$T_N(u, v) = \Delta\rho \int_0^1 e^{-\frac{i u}{2} \rho^2} J_0(v \rho) \delta_T(\rho) \rho d\rho, \quad (42)$$

where $\delta_T(\rho)$ is the Dirac delta train function,

$$\delta_T(\rho) = \sum_{m=-\infty}^{\infty} \delta(\rho - m \Delta\rho), \quad (43)$$

and $\delta(\rho)$ is a Dirac delta function [34], which we use to define the location of each sample. In Equation (42) we use the well-known Poisson formula for Dirac delta train [35],

$$\sum_{m=-\infty}^{\infty} \delta(\rho - m \Delta\rho) = \sum_{m=-\infty}^{\infty} \frac{1}{\Delta\rho} e^{i 2\pi \frac{m}{\Delta\rho} \rho}, \quad (44)$$

we get,

$$T_N(u, v) = \sum_{m=-\infty}^{\infty} \int_0^1 e^{-\frac{i u \rho^2}{2}} J_0(v \rho) e^{i 2\pi \frac{m}{\Delta\rho} \rho} \rho d\rho. \quad (45)$$

By changing integral limits and introducing $P(\rho)$, we rewrite Equation (45) as a Hankel transform of zero order,

$$\begin{aligned} T_N(u, v) &= \sum_{m=-\infty}^{\infty} \int_0^{\infty} e^{-\frac{i u \rho^2}{2}} J_0(v \rho) e^{i 2\pi \frac{m}{\Delta\rho} \rho} P(\rho) \rho d\rho \\ &= \sum_{m=-\infty}^{\infty} H_0 \left\{ \left[e^{-\frac{i u \rho^2}{2}} P(\rho) \right] e^{i 2\pi \frac{m}{\Delta\rho} \rho} \right\} (v), \end{aligned} \quad (46)$$

where $P(\rho)$ is defined as,

$$P(\rho) = \begin{cases} 1 & 0 < \rho < 1; \\ 0 & \text{otherwise.} \end{cases}$$

We note that Equation (13) can also be written as,

$$T(u, v) = H_0 \{ e^{-\frac{i u \rho^2}{2}} P(\rho) \} (v). \quad (47)$$

From [36], Chapter 17, we get the convolution property of Hankel transform,

$$H_0 \{ 2\pi f_1(x) f_2(x) \} (y) = F_1(y) * F_2(y), \quad (48)$$

and the Hankel transform of exponential function,

$$H_0\{e^{-\beta r}\}(v) = \frac{\beta}{(v^2 + \beta^2)^{\frac{3}{2}}}, \quad (49)$$

where $\text{Re}\{\beta\} > 0$. Then we get,

$$\begin{aligned} H_0\{e^{i2\pi \frac{m}{\Delta\rho} \rho}\}(v) &= \lim_{a \rightarrow 0^+} H_0\{e^{-a+i2\pi \frac{m}{\Delta\rho} \rho}\}(v) \\ &= \lim_{a \rightarrow 0^+} \frac{a - i2\pi \frac{m}{\Delta\rho}}{[v^2 + (a - i2\pi \frac{m}{\Delta\rho})^2]^{\frac{3}{2}}}, \end{aligned} \quad (50)$$

Substituting Equation (47), (48) and (50) into Equation (46),

$$\begin{aligned} T_N(u, v) &= \sum_{m=-\infty}^{\infty} H_0\{e^{-\frac{i u \rho^2}{2}} P(\rho) e^{i2\pi \frac{m}{\Delta\rho} \rho}\}(v) \\ &= \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} T(u, v) * \frac{-i2\pi \frac{m}{\Delta\rho}}{[v^2 - (2\pi \frac{m}{\Delta\rho})^2]^{\frac{3}{2}}}. \end{aligned} \quad (51)$$

Examining Equation (51) we see that, there is an infinite number of poles located at $v = 2\pi m/\Delta\rho$, $m = 0, \pm 1, \pm 2, \dots$, and the radial distance between adjacent poles is $2\pi/\Delta\rho$.

Taking $S = 30$ samples, setting $v = 0$ (i.e. we are at the focal plane), we plot the result of Equation (41) in Figure 16. In Figure 16, we can clearly see a set of concentric circles (centred at $v = 0$, radius $2\pi m/\Delta\rho$), which is agreement with the location of the poles in Equation (51). Provided that these replicas are well separated from each other, the numerical solution provides a highly accurate result. When $v \neq 0$, the replicas introduced by the sampling operation remain at these locations, $v = 2\pi m/\Delta\rho$. However, the extent of $T(u, v)$ increases due to defocus.

If we consider the point of view of power, the power is defined as $\int 2\pi |T(u, v)|^2 v dv$. We know that at the input plane, the power is $P_{in} = \pi a^2$. Because power is conserved, the power in each diffraction plane must remain constant. However, the effect of sampling acts to produce an infinite set of replicas in the diffraction plane. Using the variable P to refer to the power of the zero order replica, we find that the power in the annular region, $(2k+1)\pi/\Delta\rho \leq \rho \leq (2k+3)\pi/\Delta\rho$, is $2P$, where $k = 0, 1, 2, \dots$. This is because when $m \neq 0$ in Equation (51), negative and positive replicas overlap spatially doubling the power in these annular regions, see the negative and positive values for the index m in Equation (51). In the plot in Figure 16, $N = 30$, samples were used giving $P = 0.97P_{in}$. Because the power associated with the replicas may not necessary be well separated, it leads to numerically erroneous results. If the input field is sampled at a sufficiently high rate the power within the first replica window, i.e. when $\rho \leq \pi/\Delta\rho$, see Figure 16, P will be closed to P_{in} , and we can be assured that the calculation has been correctly implemented.

5. Conclusions

In this manuscript, we set ourselves the task of examining the complex amplitude distribution in the region behind a perfect converging spherical lens. The analysis assumes that the paraxial approximation is valid and therefore used the Fresnel transform to describe the process of diffraction. By first simplifying the resulting analytical solution the double integral (integration with respect to the polar angle) was reduced to a single integral over the R . The solution of this latter integral is more involved and four different but mathematically equivalent analytical expressions were derived in terms of an infinite series expression. We then proceeded to examine the properties of these different solutions for three different cross-sections through the focus. The convergence properties as a function of the number of terms required in each analytical solution was examined for three different spatial locations.

The solution using the first Lommel function, was found to perform well within geometrical shadow, and converges well when $|u| \ll |v|$. In geometrical illumination region more terms were needed so that the solution converged to the correct answer. The second Lommel solution T_2 performs in a contrary manner, working well within the geometrical illumination volume and particularly well near the optical axis. However, this solution failed along focus plane perpendicular to the optical axis, because of the limited machine precision in computational software program. The computation time for these two solutions is most efficient of all. For Cao's solution, we found it is similar to the solution using first Lommel function, performing well in geometrical shadow region. However in geometrical illumination it does not work as well as the first Lommel solution, requiring more terms to converge. Nijboer's solution with Zernike polynomials has the most robust solution of all the approaches: it generally requires a lower number of series terms to converge at nearly all of the locations examined here. By using this solution, it is not necessary to worry if the spatial location is in the illumination region or in the shadow region, and by the computation time it performs also very good, see Section 3.4.

A numerical solution for the diffraction integral, was examined and used to produce error maps presented in Section 3. This numerical approach take a little more time than solutions T_1 , T_2 , and T_3 , and it produces the replicas in the diffraction domain which reduce the accuracy of the calculation. These replicas are caused by the sampling process, and located in along concentric circles (radii = $2\pi k/\Delta\rho$), with an origin at the optical axis. The more samples we take, the farther each replica located, and the more computation time is required. It is possible to implement alternative numerical solutions that are based on the fast Fourier transform. This topic has been addressed by several other authors, see for example [13] and is not pursued here. Note that, we have used the Fresnel transform, under the Fresnel approximation the solution is only valid if the

numerical aperture NA is small, for high NA focusing systems $NA > 0.5$ vector diffraction theory need to be used [37].

Funding

Yang Wu is supported by Photograd (Thüringer Landesgraduiertenschule für Photovoltaik). Damien P. Kelly is now Junior-Stiftungsprofessor of Optics Design and is supported by funding from the Carl-Zeiss-Stiftung [FKZ: 21-0563-2.8/121/1].

References

- [1] Born, M.; Wolf, E. *Principles of Optics*, 6th ed.; Pergamon Press, Oxford, 1980; Chapters 8–9.
- [2] Sommargren, G.E.; Weaver, H.J. *App. Opt.* **1990**, *29*, 4646–4657.
- [3] Li, Y.; Wolf, E. *Opt. Comm.* **1981**, *39*, 211–215.
- [4] Wolf, E.; Li, Y. *Opt. Comm.* **1981**, *39*, 205–210. Focusing of vortex beams: Lommel treatment.
- [5] Sheppard, C.J.R.; Hrynevych, M. *J. Opt. Soc. Am. A* **1992**, *9*, 274–281.
- [6] Sheppard, C.J.R. *Opt. Lett.* **2000**, *25*, 1660–1662.
- [7] Sheppard, C.J.R. *Opt. Lett.* **2013**, *38*, 1074–1076.
- [8] Sheppard, C.J.R. *J. Opt. Soc. Am. A* **2014**, *31*, 644–651.
- [9] Cao, Q. *J. Opt. Soc. Am. A* **2002**, *20*, 661–667.
- [10] Kelly, D.P. *J. Opt. Soc. Am. A* **2014**, *31*, 755–765.
- [11] Mielenz, K.D. *J. Res. Nat. Inst. Stand. Technol.* **1998**, *103*, 497–509.
- [12] Janssen, A.J.E.M.; Braat, J.J.M.; Dirksen, P. *J. Mod. Opt.* **2004**, *51*, 687–703.
- [13] Stammes, J.J. *Wave in Focal Region: Propagation, Diffraction and Focusing of light, Sound and Water Waves*, 1st ed.; Taylor & Francis, New York, NY, 1986.
- [14] van Haver, S. *The Extended Nijboer-Zernike Diffraction Theory and its Applications*; Technical University Delft, Delft, 2010.
- [15] van Haver, S.; Janssen, A.J.E.M. *J. Eur. Opt. Soc. Rap. Public* **2013**, *8*, 13044-1–13044-29.
- [16] Wolf, E. *Proc. R. Soc. London A* **1959**, *253*, 349–357.
- [17] Richards, B.; Wolf, E. *Proc. R. Soc. London A* **1959**, *253*, 358–379.
- [18] Iserles, A. *A First Course in the Numerical Analysis of Differential Equations*; Cambridge University Press, Cambridge, 2008.
- [19] Siegman, A.E. *Lasers*; University Science Books, Sausalito, CA, 1986; Chapter 16.
- [20] Frieden, B.R. *Prog. Opt.* **1971**, *9*, 311–407.
- [21] Nijboer, B.R.A. *Physica X* **1943**, *8*, 679–692.
- [22] Nijboer, B.R.A. *Physica XIII* **1947**, *10*, 605–620.
- [23] Goodman, J. *Introduction to Fourier optics*, 3th ed.; McGraw Hill, New York, NY, 1996; Chapters 2, 4–5.
- [24] Watson, G.N. *A Treatise on the Theory of Bessel Functions*; Cambridge University Press, Cambridge, 1992; Chapters 2, 16.
- [25] Gray, A.; Mathews, G.B.; MacRobert, T.M. *A treatise on Bessel Functions*, 2nd ed.; Macmillan: London, 1922; Chapter 14.
- [26] Kelly, D.P.; Hennelly, B.M.; Sheridan, J.T.; Rhodes, W.T. *Opt. Comm.* **2006**, *263*, 171–179.
- [27] Kelly, D.P.; Hennelly, B.M.; Sheridan, J.T.; Rhodes, W.T. *Opt. Comm.* **2006**, *263*, 180–188.
- [28] Lommel, E. *Abh. Bayer. Akad.* **1885**, *15*, 233–328.
- [29] Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*; Dover, New York, NY, 1964; Chapter 10.
- [30] Noyes, J.L. *Numerical Computation with Mathematica: Basic Analysis and Visualization*; Wittenberg University Press, 2009; Chapter 1.
- [31] Török, P.; Hewlett, S.J.; Varga, P. *J. Mod. Opt.* **1997**, *44*, 493–503.
- [32] Kelly, D.P.; Healy, J.J.; Hennelly, B.M.; Sheridan, J.T. *J. Eur. Opt. Soc. Rap. Public* **2011**, *6*, 11034-1–11034-14.
- [33] Kelly, D.P.; Hennelly, B.M.; Pandey, N. *Opt. Eng.* **2009**, *48*, 095801-1–095801-13.
- [34] Lohmann, A.W. *Optical Information Processing*; Universitätsverlag Ilmenau, Ilmenau, 2006; Chapter 6.
- [35] Stern, A. *Signal Process.* **2006**, *86*, 1421–1425.
- [36] Poularikas, A.D., Ed. *The Handbook of Formulas and Tables for Signal Processing*; CRC Press LLC, Boca Raton, FL, 1999; Chapter 17.
- [37] Sherif, S.S.; Foreman, M.R.; Török, P. *Opt. Express.* **2008**, *16*, 3397–3407.