Université de Neuchâtel
Institut de Microtechnique

Far-field beam shaping elements for deep UV lithography

Thèse

Présentée à la Faculté des Sciences
pour obtenir le grade de docteur ès sciences
par

Olivier Ripoll

Neuchâtel, janvier 2003

UFO Dissertetion Band 428

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Ein Titeldatensatz für diese Publikation ist bei
Der Deutschen Bibliothek erhältlich.

Dissertation der Universität Neuchâtel
Datum der mündlichen Prüfung: 04.01.2003
Referenten: Prof. Dr. R. Dändliker
Prof. Dr. H. P. Herzig
Prof. Dr. M. Kuittinen
Dr. V. Kettunen
Dr. M. Maul

UFO Atelier für Gestaltung & Verlag GbR · D-78476 Allensbach
Internet: www.ufo-verlag.de

ISBN 3-935511-29-9

IMPRIMATUR POUR LA THESE

Far-field beam shaping elements for deep UV lithography

de M. Olivier RIPOLL

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel, sur le rapport des membres du jury

MM. R. Dändliker (directeur de thèse), H.-P. Herzig, M. Kuittinen (Joensuu Finland), V. Kettunen (Zürich) et M. Maul (Oberkochen D)

autorise l'impression de la présente thèse.

Neuchâtel, le 28 mai 2003

Le doyen:

François Zwahlen

Abstract

For the efficient fabrication of micro-electronic circuits by photo-lithography, the beam of light used to illuminate the mask has to be appropriately structured. This shaping is realised by a set of optical elements. During the design of the beam-shaping element, constraints due to the nature of the machinery, the characteristics of the light and the fabrication technology of the optical element have to be taken into account.

In this work, we tackle the task of designing and analysing such elements. Simulation of optical structures whose dimensions are large compared to the wavelength and the minimum feature size has required the development of a completely new representation of the geometry of the elements. Instead of the two-dimensional sampling usually used, we have introduced a hybrid digital-analytical Fourier transform which drastically reduces the memory requirements.

Various design methods have been compared, and improvements were proposed to several candidates. The sensitivity to fabrication errors have been characterised for the elements resulting of these design techniques. Conclusions were drawn about which design to use depending on the tolerances available.

Contents

Introduction 1

For the manufacturing of micro-electronics circuits such as microprocessors, one needs to produce micro-structures in a substrate (usually silicon). Typically, the substrate is covered with some photo-sensitive material, and the pattern to be etched is written into this material by high energy illumination. This step changes the property of the covering layer at the places where it has received energy, and chemical or physical processes can then be used to transfer the pattern into the substrate. The projection of the pattern onto the substrate is schematically illustrated in Fig. 1.1. The pattern is drawn on a mask (the reticle R) that is imaged onto the target T. In the plane of the pupil P, the source is imaged by lenses L_1 and L_2. Additionally, as shown in Fig. 1.2(a), the pupil of the imaging lens L_3 acts as low-pass filter for the spatial frequencies of the mask pattern. The pattern is resolved if at least two diffraction orders are transmitted by the pupil. Figure 1.2(b) illustrates this situation for a pattern composed of vertical lines. The zero order image of the source is centered in the pupil. Since the source is spatially incoherent, only those points of the source which generate at least two images in the pupil contribute to the pattern. An example of such a point is point A, whose zero order (A_0) and +1 order (A_{+1}) images are within the pupil P. Point B, on the other side of the source provides also two image points inside the pupil, B_0 and B_{-1}. The dark-gray areas indicate all points of the source that contribute to the formation of the image of the mask. The light-gray area does not contribute to the imaging of the pattern, as it consists of points like C, which has only one image point in the pupil (C_0). The light in this area produces a background noise which lowers the contrast. From the top row, one can see that the important contribution of the orders ± 1 (the parts in the pupil) comes from the peripheral area of the source. Thus, to improve the contrast of small features, one should use sources with off-axis light emission. The bottom row of Fig. 1.2(b) shows the image of such a source. The amount of light contributing

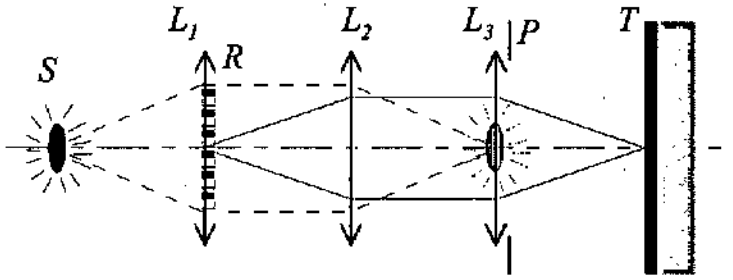


Figure 1.1: Schematic view of the projection lithography principle for a 1:1 magnification. The source S is imaged onto the plane of the pupil P, the reticle R is imaged onto the target wafer T.

to the background illumination is considerably reduced. To adapt the source to the pattern to be transferred, there are in lithographic systems several types of geometry for the illumination source: disc (often called conventional illumination), ring (called annular), quadrupole (such as the quadrupole shaped annular – QUASAR^{TM} [1]), and dipole. These geometries are available with different values of the partial coherence factor σ, defined as the ratio between the size of the source image and the size of the pupil [2]. Depending on the geometry and spatial frequencies of the mask pattern, one has to choose the most suited source distribution. For example, the conventional illumination is more suited for isolated features, while dipole illumination is more suited for horizontal or vertical dense line patterns [1, 2].

Creating such a light distribution, has to be performed by *beam shaping* from a somehow non structured primary light source with the highest possible efficiency. Beam shaping should be done with the smallest number of elements to avoid absorption which gets critical for shorter wavelengths. Additionally to the loss of power, absorption generates heat that would cause the destruction of transmissive optical elements. Thus, the optical elements have to be made the thinest possible. For deep ultra-violet illumination, the available laser sources are multi-mode excimer lasers, such as argon fluoride (ArF) or krypton fluoride (KrF). The structure of the beam presents a speckled appearance which changes orders of magnitude faster than the exposure time, resulting in a time-averaging of the beam that can be described as an incoherent Gaussian beam [3–5]. Moreover, the beam-shaping element should be spatially invariant, i.e., it should not depend on the input wavefront and the spatial intensity distribution. This property is realised by tiling many small individual beam-shaping elements. Indeed, if the single cell size is small compared to the beam size, the wavefront incident on the cell structure can

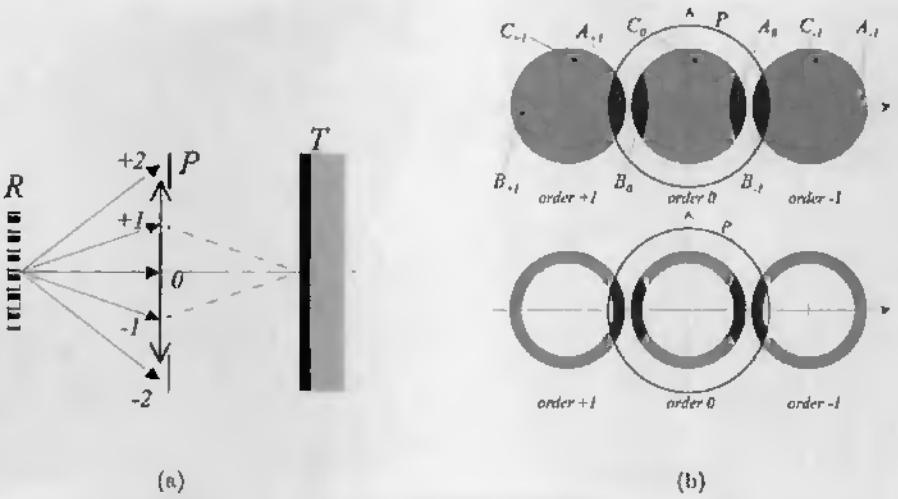


Figure 1.2: The mask generates diffraction orders that are filtered by the pupil (a). The limit of resolution is reached when no first order is transmitted. The amount of energy contributing to the pattern is represented by the points of the source which have at least two images inside the pupil, shown in dark-gray (b). An annular illumination (bottom) allows a better contrast that a conventional circular one (top).

be locally approximated by a plane wave. The final resulting intensity will thus be the incoherent sum of the intensities from each element, which have similar profiles but are slightly translated in the output plane, as illustrated in Fig. 1.3. With respect to the spatial intensity distribution, as shown in Fig. 1.4, the distribution resulting from the whole beam-shaping element can be described as the incoherent addition of the intensities resulting from all single cells. This results in an averaging of the light distribution and makes it independent of the input beam distribution.

The present work tackles the use of diffractive and refractive optical elements to structure the light distribution in lithographic systems. Simulating their effect is important in order to understand their limits, to be able to analyse eventual deficiencies, and to propose possible improvements. In addition, we will compare different types of optical elements whose characteristics will be presented.

This work has been realised in collaboration with Colibrys SA, formerly CSEM microsystems division, in Neuchâtel, Switzerland.

In chapter 2, we will describe the various theories used to model the propagation and interaction of light, namely rigorous diffraction theory, scalar optics

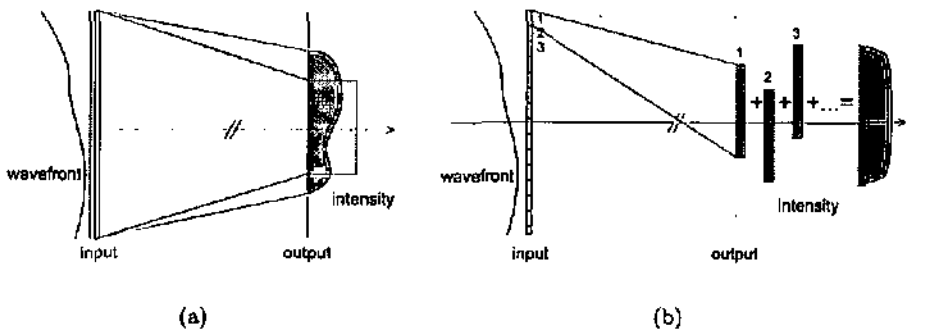


Figure 1.3: Sensitivity to spatial variations of the wavefront. (a) The output of an element designed for a plane wave (dashed line) is seriously distorted if the incident wave is not plane (solid line). (b) An element composed of small beam-shaping cells is less sensitive to distorted wavefronts, because the total output is the incoherent sum of the distributions resulting from each cell illuminated locally by an oblique plane wave.

and ray-tracing. The first one provides a better description of the interaction of light with the material structures used to modify it. The two others are based on simplifications but still provide surprisingly useful predictions. Because the first is based on a precise knowledge of both the incoming light and the optical elements, the physical effects – intuitive understandings of critical parameters – can be hidden among the number of variables. Moreover, the actual implementations of this theory on computers will be limited to very small optical elements and not suitable for real world applications, where the accuracy of the model is balanced by materialistic issues like speed, computational power or data storage capacities. The simplified theories, while restricted in their field of applications and not perfectly accurate, have provided useful criteria and working tools for optical designer for many decades. We will thus present how the various models may be used for the design, simulation and tolerancing of an optical element.

Chapter 3 will focus on the simulation step, and introduces a new technique which greatly reduces the requirements for data storage, allowing desktop computers to be used in the field of physical optics. The classical implementation of this model is based on a sampled description of the optical element, similar to the way computers represent images, thus close to the insides of the computer. However, the optical element is not itself best described by this decomposition. This usually leads to some sub-optimal representation, which may introduce artifacts (moiré effects, or other "numerical contamination").

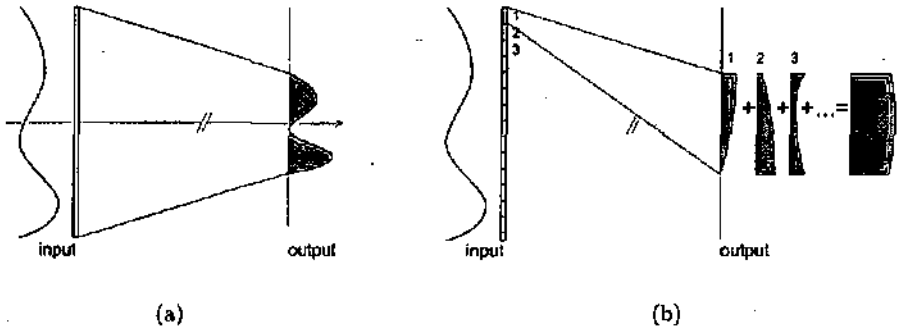


Figure 1.4: Comparison of the sensitivity to spatial variations of the intensity. For an single element (a) the output intensity depends on the input intensity while for a tiling of small sub-cells (b), the output light distribution is an average of the distributions resulting of every sub-cell.

For a set of optical elements – general enough to describe most of the DOEs for illumination – we introduce a new computation scheme. It is based on a representation that is less computer oriented, and leans more towards the geometry of the element itself. The main advantage is a tremendous cut in the data storage requirements, with no loss in accuracy. Also, we shall expose cases where the sensitivity to artifacts is reduced, and new possibilities in the simulation are offered to the designer. While, for the purpose of our applications, we restrict ourselves to a subset of the available optical structures, we will show that the principle of the technique can be extended to a larger set of situations.

For a better understanding of the constraints, which we have to take into account in our designs and in our tolerance studies, we will then discuss the fabrication methods of beam-shaping elements. Various techniques will be presented for the fabrication of micro-optical elements on an industrial scale. Both diffractive and refractive diffusers were studied within this thesis. We will outline the similarities and the differences between both families. This section should provide the basic information for the comprehension of the design techniques which will be covered in chapter 5. There, we shall see that beam shaping can be realised by two different classes of solutions: re-mapping type and grating type elements.

Re-mapping type elements are based on the correspondence between two distributions of light – input and output – and intends to find a point to point relation between them. Its origins lie in the ray-tracing history, and thus do not take into account wave optics properties, such as diffraction. Most of

the beam shaping optics is still designed with these techniques. We will first present specific solutions that can be found analytically, but are restricted to cases of special symmetry. A more general technique, mesh based inverse ray-tracing, will be discussed afterwards, fulfilling some other goals. We will tackle issues of both techniques, and provide improvements for the mesh based ray-tracing originally presented by Dresel *et al.* [6]. Finally, we will underline the drawbacks of such designs that do not account for wave optics properties.

On the other hand, grating type elements rely on the Fourier transforming properties of the propagation of light. The resulting beam shaping elements do not rely any more on point to point correspondence, but rather on a position to frequency relationship between the input and output light distributions. Thus, not only wave optics is taken into account, but the elements will not have the limitations of the re-mapping type ones. On the other hand, it may introduce its own defects. Iterative Fourier transform algorithm will be specifically discussed in this section.

Refinements and improvements of optical elements will be the subject of chapter 6. Effects of various fabrication errors will be characterised, and some solutions aiming at reducing the sensitivity to such errors will be presented: pre-compensation of line width errors or three level elements that are less sensitive to etching depth variations. Alignment errors result in different behaviours depending on the nature of the design family, as presented in chapter 5. This difference will be studied. The second part of this chapter will be dedicated to design-related defects, like oscillations due to the rim of the elementary cell or to the quantification of the profile. We will again propose improvements that may correct these defects.

Interaction and propagation of light

2

In this chapter, we will present and compare the different theories used for the modelling of optical elements, from the most rigorous to the fastest and lightest. All of them present domains where they are more suitable than the others, given that accuracy of the predicted result is not the only criterion for design and optimisation. Indeed, the capacity to apply a model to a wide set of optical elements and the ease of computation have to be taken into account. Understanding the limits of each is thus of great importance, and allows for a set of improved implementations of these models, aimed at specific situations.

2.1 Rigorous theory

2.1.1 Electromagnetic theory

Electromagnetic theory is based on the Maxwell equations (2.1-2.4). They describe the relations between the electric field $\mathbf{E}(\mathbf{r})$, the magnetic field $\mathbf{H}(\mathbf{r})$, the electric displacement $\mathbf{D}(\mathbf{r})$, the magnetic induction $\mathbf{B}(\mathbf{r})$, the electric current density $\mathbf{j}(\mathbf{r})$ and the electric charge density $\rho(\mathbf{r})$. The dot stands for the derivative with respect to time.

$$\nabla \times \mathbf{E} = -\dot{\mathbf{B}}, \tag{2.1}$$

$$\nabla \times \mathbf{H} = \mathbf{j} + \dot{\mathbf{D}}, \tag{2.2}$$

$$\nabla \cdot \mathbf{D} = \rho, \tag{2.3}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{2.4}$$

The interaction with a material medium is expressed through the constitutive equations

$$\mathbf{D} = \varepsilon_0 \varepsilon \mathbf{E}, \qquad \text{with} \quad \varepsilon_0 = \frac{1}{36\pi} 10^{-9} \mathrm{AsV^{-1}m^{-1}}, \tag{2.5}$$

$$\mathbf{B} = \mu_0 \mu \mathbf{H}, \qquad \text{with} \quad \mu_0 = 4\pi \cdot 10^{-7} \mathrm{VsA^{-1}m^{-1}}, \tag{2.6}$$

$$\mathbf{j} = \sigma \mathbf{E} \tag{2.7}$$

ε is the dielectric permittivity, μ is the permeability and σ is the electric conductivity. Note that $\varepsilon(\mathbf{r})$ and $\mu(\mathbf{r})$ are scalar for isotropic materials, and tensorial for anisotropic materials. Herein we will only discuss non-magnetic and isotropic materials (i.e. $\mu = 1$, and ε is scalar).

From Maxwell equations can be derived the boundary conditions describing the behaviour of the electromagnetic field at the interface between two media (1 and 2). $\mathbf{n}_{12}$ represents the unity vector normal to the surface separating the two media. ρ_s and $\mathbf{j}_s$ are the surface charge and current densities, respectively.

$$\mathbf{n}_{12} \cdot (\mathbf{B}_2 - \mathbf{B}_1) = 0, \tag{2.8}$$

$$\mathbf{n}_{12} \cdot (\mathbf{D}_2 - \mathbf{D}_1) = \rho_s, \tag{2.9}$$

$$\mathbf{n}_{12} \times (\mathbf{E}_2 - \mathbf{E}_1) = 0, \tag{2.10}$$

$$\mathbf{n}_{12} \times (\mathbf{H}_2 - \mathbf{H}_1) = \mathbf{j}_s. \tag{2.11}$$

For dielectric materials, both the charge and current densities are zero. This leads to the conservation of both the electric and magnetic fields at the interface. For metallic materials however, this is no longer the case.

Maxwell equations can be combined in order to decouple both components of the field. For an inhomogeneous, but isotropic, dielectric and non-magnetic medium, the differential equations (2.12) and (2.13), also known as the wave equations, are derived, as

$$\nabla^2 \mathbf{E}(\mathbf{r}) - \varepsilon_0 \varepsilon(\mathbf{r}) \mu_0 \mu \ddot{\mathbf{E}}(\mathbf{r}) = 0 \tag{2.12}$$

$$\nabla^2 \mathbf{H}(\mathbf{r}) - \varepsilon_0 \varepsilon(\mathbf{r}) \mu_0 \mu \ddot{\mathbf{H}}(\mathbf{r}) = (\nabla \times \mathbf{H}(\mathbf{r})) \times \frac{\nabla \varepsilon(\mathbf{r})}{\varepsilon(\mathbf{r})} \tag{2.13}$$

The absence of a term on the right side of Eq. (2.12) is due to the constant value of μ. For a magnetically inhomogeneous material with $\mu(\mathbf{r})$, a term similar to the right hand side of Eq. (2.13) would be present (with $\mu(\mathbf{r})$ replacing $\varepsilon(\mathbf{r})$, and $\mathbf{E}(\mathbf{r})$ instead of $\mathbf{H}(\mathbf{r})$).

The time dependent electromagnetic fields can be decomposed on the time harmonic function basis. Any vector is then a sum of many monochromatic

functions. The problem can then be reduced to a harmonic field. Electromagnetic vectors can then be written as the product of a time independent vector with a harmonic function

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0(\mathbf{r})\exp(\mathrm{i}\omega t), \quad (2.14)$$

$$\mathbf{H}(\mathbf{r},t) = \mathbf{H}_0(\mathbf{r})\exp(\mathrm{i}\omega t). \quad (2.15)$$

Introducing this decomposition into Eqs. (2.12) and (2.13), we obtain

$$\nabla^2\mathbf{E}_0(\mathbf{r}) + \varepsilon_0\varepsilon(\mathbf{r})\mu_0\mu\omega^2\mathbf{E}_0(\mathbf{r}) = 0 \quad (2.16)$$

$$\nabla^2\mathbf{H}_0(\mathbf{r}) + \varepsilon_0\varepsilon(\mathbf{r})\mu_0\mu\omega^2\mathbf{H}_0(\mathbf{r}) = (\nabla\times\mathbf{H}_0(\mathbf{r}))\times\frac{\nabla\varepsilon(\mathbf{r})}{\varepsilon(\mathbf{r})} \quad (2.17)$$

2.1.2 Numerical methods

There exist several mathematical methods to solve Maxwell equations. The integral methods are of common use in the antenna and tomographic domains. They are well suited to compute the fields of aperiodic structures [7]. Nevertheless, in diffractive optics the solution is often performed through differential methods. They are based on the mathematical treatment of the grating diffraction, that resorts on the assumption of periodicity on the diffractive structure. The structure is localised on a plane. Among the various implementations of the grating diffraction, of particular interest are the rigorous coupled wave algorithm (RCWA) and the modal methods. While the RCWA expands the field on a harmonic basis in the whole space, the modal method uses locally the basis that is made of modes of the propagation. The rigorous modal method is thus limited to a small set of geometries, where the modes are known (e.g. binary, multilevel, saw-tooth). In the general case, the modes are themselves expanded on a harmonic basis, hence the name Fourier modal method (FMM), which leads to a computer implementation equivalent to the one of the RCWA. In this thesis, whenever we had to rely on a rigorous diffraction algorithm, FMM programs were used. Other differential methods worth of interest are the C-method [8,9], and Morf method [10,11], which is a variation of the modal method where the modes are expressed with polynomials instead of harmonics. Both are suited for continuous profiles, but are harder to implement and not as general as the FMM.

Finally, it is possible to solve Maxwell equations locally with a finite elements method. This technique usually allows to study periodic and aperiodic solutions, depending on the conditions at the borders of the structure.

2.1.3 Limitation for implementations of the rigorous theories

The main issue with all of the rigorous methods is the amount of time and data required. Indeed, for the FMM, the profile has to be sliced in Q superposed lamellar layers and the infinite sums have to be truncated so that a sufficient number N of diffraction orders is taken into account. The size of the matrices is evolving as $\mathcal{O}\left(N^2Q^2\right)$. The time needed to solve this eigenvalue problem is hence typically $\mathcal{O}\left(N^3Q^3\right)$ for such arrays [12]. As a consequence, the rigorous theory implementations are unsuitable for elements whose dimensions are large compared to the optical wavelength. Unfortunately, we are interested in elements whose sizes are typically a few ten thousand wavelengths in two directions.

For elements whose profiles cannot be described in a plane (often referred to es a three-dimensional situation), the problem of data size and computation time is even more restrictive, and only structures whose dimensions are of a few wavelengths can be simulated. Therefore, rigorous methods can be used for optimisation routines [13, 14], but they are not practical because of the time requirements.

The fact that large grating-like structures can be simulated is due to their periodicity. The period itself is much smaller. While adequate for gratings, this bypothesis is not appropriated for more complicated structures, like computer generated holograms (CGH) or diffractive lenses, that are used in beam shaping applications. Hence, the FMM is not suitable for our large optical structures.

The limitations of rigorous methods prohibit their use for the optical elements we will study for deep UV lithography. However, they are useful to study local behaviour, when the approximate theories fail to describe the interaction of the light and the structured medium with enough details.

2.2 Scalar diffraction theory

Scalar optics, strictly speaking, is the approximation of electromagnetic optics by scalar fields, i.e. fields that are not vectorial. The polarisation is thus not considered in this model, and one should reasonably expect a bad accuracy for structures with rapid spatial variations and deep features. But for many situations, scalar optics has proved to be effective enough. Moreover, some of its implementations rely on Fourier transform which can be implemented as a fast algorithm on computers.

2.2.1 Propagation of a scalar field

Under the scalar approximation, the Eqs. (2.16) and (2.17) can be unified in a single scalar equation, called Helmholtz equation

$$\left(\nabla^2 + n^2(\mathbf{r})\, k_0^2\right) U(\mathbf{r}) = 0, \tag{2.18}$$

where

$$k_0 = \frac{k}{n} = \frac{\omega}{c}. \tag{2.19}$$

The refraction index and is defined as

$$n(\mathbf{r}) = \sqrt{\varepsilon(\mathbf{r})\,\mu}, \tag{2.20}$$

and

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} \tag{2.21}$$

is the speed of light in the vacuum. Using the Green theorem

$$\iiint_V \left(G\nabla^2 U - U\nabla^2 G\right) \mathrm{d}v = \oiint_S \left(G\frac{\partial U}{\partial n} - U\frac{\partial G}{\partial n}\right) \mathrm{d}s \tag{2.22}$$

with the three-dimensional free space Green's function,

$$G(\mathbf{r}_1) = \frac{\exp(ikr_1)}{r_1} \tag{2.23}$$

a spherical wave, the integral theorem of Helmholtz and Kirchhoff

$$U(\mathbf{r}_0) = \frac{1}{4\pi} \iint_S \left\{ \frac{\partial U}{\partial n}\left[\frac{\exp(ikr_{01})}{r_{01}}\right] - U\frac{\partial}{\partial n}\left[\frac{\exp(ikr_{01})}{r_{01}}\right] \right\} \mathrm{d}s \tag{2.24}$$

is derived. As illustrated by Fig. 2.1, $\mathbf{r}_{01} = \mathbf{r}_1 - \mathbf{r}_0$ is the vector between a point $P_1(\mathbf{r}_1)$ on the surface and the point $P_0(\mathbf{r}_0)$ where the field is observed. So far, no further assumptions have been made, with the exception of the ones used for the Helmholtz equation, namely the isotropy of the medium. This implies that we can apply the following equations to both components of the electromagnetic field independently in such a medium.

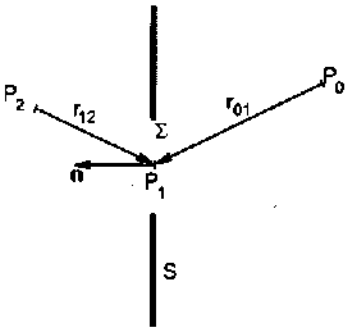


Figure 2.1: Geometry used for the theorem of Helmholtz and Kirchhoff, in the case of diffraction at a planac interface.

2.2.2 Diffraction at a planar interface

Fresnel-Kirchhoff and Rayleigh-Sommerfeld formulas

Originally derived for diffracting apertures in black screens, the Fresnel-Kirchhoff and Rayleigh-Sommerfeld formulas are used for much more general apertures. They are also used for optical elements, where the notion of aperture cannot be defined in the same manner as for a screen. However, they are well suited to describe the propagation of light diffracted by a structured medium. An extensive derivation and discussion of these formulas can be found in chapter 3 of Ref. [15].

The Fresnel-Kirchhoff formula is based on the assumption that both the field and its first derivative are zero on the plane outside the aperture. Within, they are assumed to be equal to the values in the absence of the screen. The correctness of this assumption can be discussed physically and mathematically: it is inconsistent but can be reformulated consistently [16]. Anyway, it leads to the formula

$$U(\mathrm{r}_0) = \iint_{\Sigma} U'(\mathrm{r}_1)\,\frac{\exp(\mathrm{i}kr_{01})}{r_{01}}\,\mathrm{d}s, \tag{2.25}$$

where Σ is the surface of the aperture with

$$U'(\mathrm{r}_1) = \frac{1}{\mathrm{i}\lambda}\,\frac{\exp(\mathrm{i}kr_{21})}{r_{21}}\,\frac{[\cos(\mathrm{n},\mathrm{r}_{01}) - \cos(\mathrm{n},\mathrm{r}_{21})]}{2}. \tag{2.26}$$

The mathematical inconsistency can be avoided when using different Green's functions. This leads to the two different Rayleigh-Sommerfeld formulas

$$U_I(\mathrm{r}_0) = \frac{1}{\mathrm{i}\lambda}\iint_{\Sigma} U(\mathrm{r}_1)\,\frac{\exp(\mathrm{i}kr_{01})}{r_{01}}\cos(\mathrm{n},\mathrm{r}_{01})\,\mathrm{d}s, \tag{2.27}$$

$$U_{II}(\mathrm{r}_0) = \frac{1}{2\pi}\iint_{\Sigma} \frac{\partial U(\mathrm{r}_1)}{\partial n}\frac{\exp(\mathrm{i}kr_{01})}{r_{01}}\mathrm{d}s. \tag{2.28}$$

The first formula is based on the condition that the field is zero on the surface outside the aperture, while the second formula is based on the condition that its first derivative is zero on the surface outside the aperture. One could think the boundary conditions used to derived Eqs. (2.27) and (2.28) may be interpreted in order to choose which solution to use for a specific problem. Indeed, the first solution is modelling a "hard" screen and the second solution is closer to a "soft" screen (acoustic description of the screens properties). However, one should avoid such assumptions, as the exact electromagnetic computation can contradict this intuition [17, 18].

For both Fresnel-Kirchhoff and Rayleigh-Sommerfeld formulas, additional hypotheses have been made:

1. the size of the aperture Σ is large compared with the wavelength λ.
2. the observation point is "far enough" from the aperture, i.e. $|\mathrm{r}_{01}| \gg \lambda$.

Note that the Kirchhoff formula is actually the arithmetic mean of the two Rayleigh-Sommerfeld formulas. This implies that if one has no reason to prefer one or the other of the Rayleigh-Sommerfeld formulas for a diffraction problem, Kirchhoff is also a good approach [16, 17, 19]. The difference between these three formulas have been studied extensively, and the conclusions do not indicate any superiority of one on the others [20–22].

Another useful diffraction formula is the angular spectrum approach, as presented in chapter 3 of Ref. [15]. It expresses the propagation as a transfer function, and is thus easy to implement. However, it can be proven that it is equivalent to the first Rayleigh-Sommerfeld formula [23, 24]. This formulation, based on the Fourier transform

$$U(\mathrm{r}_0) = \mathcal{F}^{-1}\left\{\mathcal{F}\{U(\mathrm{r}_1),(u,v)\}\exp\left(2\mathrm{i}\pi z_{01}\sqrt{\frac{1}{\lambda^2}-u^2-v^2}\right),(x_0,y_0)\right\}, \tag{2.29}$$

is extremely well suited for numerical computation. The exponential term is usually referred to as the transfer function of free space, and is based on the Weyl expansion of a spherical wave into plane waves [25] (the Weyl expansion of a spherical wave may also be replaced by the Whittaker expansion [26] or by the approximated circular plane wave [27]). The Fourier transform is defined herein as

$$\mathcal{F}\{f(x),u\} = \widetilde{f}(u) = \int_{-\infty}^{+\infty} f(x)\exp(-2\mathrm{i}\pi ux)\,\mathrm{d}x. \tag{2.30}$$

for a one-dimensional space and can be written similarly for multiple dimensions.

Fresnel and Fraunhofer approximations

The previous formulas can be simplified further for paraxial optics. Firstly, if the angles involved in Eqs. (2.25) and (2.27) are small, the value of the cosine can be assumed to be equal to one. As a result, the two equations are identical. As we mentioned before, Eq. (2.25) describing Fresnel-Kirchhoff diffraction is equal to the arithmetic mean of the two Rayleigh-Sommerfeld formulas (2.27), implying that (2.28) is equal to the two others.

Additionally, we will make use of the two following approximations:

1. $r_{01} \simeq z_{01}$, in the numerator.
2. $r_{01} \simeq z_{01}\left(1+\frac{1}{2}r_{01}^2\right)$, in the complex exponential.

Hence, we get that Fresnel-Kirchhoff and Rayleigh-Sommerfeld equations can be simplified to

$$U(x',y',z) = \frac{\mathrm{e}^{\mathrm{i}kz}}{\mathrm{i}\lambda z}\iint_{-\infty}^{+\infty} U(x,y,0)\exp\left[\mathrm{i}\frac{k}{2z}\left((x-x')^2+(y-y')^2\right)\right]\mathrm{d}x\mathrm{d}y, \tag{2.31}$$

that is a convolution of the input field with the Fresnel phase term. Equation (2.31) can be reformulated as

$$U(x',y',z) = \frac{\mathrm{e}^{\mathrm{i}kz}}{\mathrm{i}\lambda z}\mathrm{e}^{\mathrm{i}\frac{k}{2z}\left(x'^2+y'^2\right)}\iint_{-\infty}^{+\infty} U(x,y,0)\,\mathrm{e}^{\mathrm{i}\frac{k}{2z}\left(x^2+y^2\right)}\mathrm{e}^{-2\mathrm{i}\pi\left(\frac{x'}{\lambda z}x+\frac{y'}{\lambda z}y\right)}\mathrm{d}x\mathrm{d}y \tag{2.32}$$

which is recognised to be a Fourier transform. Both formulas can thus be implemented easily on a computer, and one can use either one or the other depending on the conditions of the problem. The Fourier transform is usually faster, but the resolution in the output plane is limited, and numerical problems, such as aliasing, have to be avoided with care.

Finally, for long distances of propagation z, namely $z > 2D^2/\lambda$, with D being the characteristic dimension of the aperture, the quadratic phase inside the integral of (2.32) can be neglected, which leads to the Fraunhofer diffraction formula

$$U(x',y',z) = \frac{\mathrm{e}^{\mathrm{i}kz}}{\mathrm{i}\lambda z}\mathrm{e}^{\mathrm{i}\frac{k}{2z}\left(x'^2+y'^2\right)}\iint_{-\infty}^{+\infty} U(x,y,0)\,\mathrm{e}^{-2\mathrm{i}\pi\left(\frac{x'}{\lambda z}x+\frac{y'}{\lambda z}y\right)}\mathrm{d}x\mathrm{d}y. \tag{2.33}$$

If the distance z tends towards infinity, we can express the output field in a space where distances x', y' and z are replaced by the tangents $\alpha = x'/z$ and $\beta = y'/z$, yielding

$$U_{\text{out}}(\alpha,\beta) \propto \iint_{-\infty}^{+\infty} U_{\text{in}}(x,y) \exp\left(-2i\pi\left(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y\right)\right) \mathrm{d}x\mathrm{d}y \qquad (2.34)$$

Note that, to avoid definition problems, the constant and the multiplicative phase terms are disregarded. This formula expresses the output field as the Fourier transform of the input light distribution. While the list of approximations leading to Eq. (2.34) is long, its domain of validity covers a wide set of applications, and is thus widely used in design and simulation.

Fresnel diffraction is commonly referred to as *finite distance* or *near-field* diffraction as opposed to *far-field* diffraction described by Fraunhofer approximation. In this work, we will use the first term. Indeed, nowadays, *near-field* should preferably be reserved to the area of optics dealing with small distances in the order of the optical wavelength, which was excluded by the assumptions needed for the Fresnel-Kirchhoff and Rayleigh-Sommerfeld formulas.

2.2.3 Interaction with a structured interface

Rayleigh-Sommerfeld and Fresnel-Kirchhoff were originally intended to be used for holes in infinitely thin, black or reflecting screens. However, they provide us with a tool which is also suitable for dielectric structures.

Transmittance

The boundary conditions for the black screen can be expressed as a binary transmittance applied to the input field – the field without the diffracting plane. Hence, the integrals in Eqs. (2.25-2.28) that are limited to the slit area can be extended to the whole plane S, using a transmittance

$$T_\Sigma = \begin{cases} 1, & \text{for } \mathrm{P}_1 \in \Sigma \\ 0, & \text{for } \mathrm{P}_1 \in S \setminus \Sigma \end{cases} \qquad (2.35)$$

For Fresnel-Kirchhoff, this would be expressed as

$$U(\mathrm{r}_0) = \frac{1}{\mathrm{i}\lambda} \iint_S T_\Sigma(\mathrm{r}_1) \cdot U'(\mathrm{r}_1) \frac{\exp(\mathrm{i}kr_{01})}{r_{01}} ds \qquad (2.36)$$

This reformulation introduces the concept of transmittances for thin apertures. This transmittances can be interpreted as the interaction of the plane with the input light. Hence, it is straightforward to extend this concept to non binary or complex transmittances

$$U_{\text{out}}(\mathbf{r}) = T(\mathbf{r}) \cdot U_{\text{in}}(\mathbf{r}), \tag{2.37}$$

where U_{out} is the scalar field just after the diffracting area, and U_{in} is the incident scalar field.

Thin element approximation (TEA)

For very thin elements, the transmittance can be presented as a legitimate approximation of the interaction between light and the material medium. This function was however defined for an infinitely thin plane. For a real optical element with three-dimensional spatial extension as seen in Fig. 2.2(a), the concept can be used as an approximation if the phase difference along the longitudinal dimension is of the order of the wavelength. This approximation is called thin element approximation and is illustrated in Fig. 2.2(b). While there are no justification for such a simplification, it happens to give reasonably good results, and tremendously simplifies the computations. It starts to break down when the element is not thin compared to both the wavelength and its features (large deflection angles). This is also the case when the interaction cannot be seen as a simple multiplicative transmittance (the Venetian blind). Bragg diffraction is another situation where TEA fails to predict the correct behaviour of the light.

One may wonder if the interaction inside the structured medium could be simulated without TEA in the scalar approximation. When scalar fields are exact models, like in acoustics or for water waves, there exists indeed interest to avoid TEA and to apply the same rigorous methods as presented in section 2.1. However, the computation problem does not lie in the vectorial nature of the electromagnetic field, but more in the modelling of the interaction in the volume of the structure, which requires the inversion of large matrices. Moreover, the scalar approximation of the light field will be deficient where the coupling between both components is non-negligible, i.e. in deep and narrow structures. This coincides with the area where TEA is not valid any more. Consequently, resorting on a rigorous-like scalar method would not bring great speed improvements nor it would be much more accurate than TEA.

Figure 2.2: Approximation of a real lens (a) by the thin element approximation (b) and the beam propagation method (c).

Beam propagation method (BPM)

If the interaction area is thick, but no small features are present, one can improve the interaction model by slicing the element in small layers where the TEA is still valid. Such layers are then modelled by a thin transmittance, and the field is propagated between such slices. This algorithm is known as the beam propagation method (BPM). Nevertheless, the thickness of the element should not be too large, since the required time for calculation is proportional to the number of layers. The principle of the BPM is illustrated in Fig. 2.2(c). A refractive lens is sliced in a stack of thin transmittances separated by a distance equal to the slice width. The gray levels represent the phase of the thin transmittance used to model the lens.

The BPM is known to be useful for lenses with a small numerical aperture ($\pm 10°$). There exists a variation of the BPM called the wave propagation method (WPM), which extends the area of validity to non paraxial cases, at the expense of more computational effort [28]. The WPM inserts some corrections derived from Snell's law in Eq. (2.29).

2.2.4 Improvements to scalar optics

To compensate for the weaknesses of the scalar theory and the TEA, various improvements have been proposed in the literature. Two of them are of particular interest: The extended scalar theory and the step transition method.

Extended scalar theory

Considering the TEA starts to be inaccurate for non paraxial angles. Therefore, Swanson has proposed a simple correction to the transmittance for gratings [29,30]. This correction allows for an improved computation of diffraction efficiencies up to 20°. It is based on a multiplicative factor, the duty cycle

$$DC = 1 - \frac{d\lambda}{\Lambda^2 \sqrt{1 - \left(\frac{\lambda}{\Lambda}\right)^2}}, \tag{2.38}$$

which is a function of the wavelength λ, the grating period Λ and the grating depth d.

This factor represents the portion of one period that is subject to the geometrical shadow of its neighbours. The energy is considered lost in this shadow. If the grating is not deep, or if the wavelength is small compared to the grating period, the duty cycle tends towards one, and the extended scalar theory is equivalent to the paraxial case. This treatment is based on some ray considerations. The correction for the 1^{st} order is expressed as

$$\eta_{1,\text{corr}} = DC^2 \cdot \eta_1. \tag{2.39}$$

Similar considerations based on ray-tracing have brought a correction to the depth of the grating [31, 32]. The depth of a grating that maximises the diffraction efficiency is found to be

$$d_{\text{opt}} = \frac{\lambda}{n\cos(\theta_i) - \sqrt{1 - \left[\frac{\lambda}{\Lambda} + n\sin(\theta_i)\right]^2}}, \tag{2.40}$$

where θ_i is the incidence angle. For normally incident illumination, Eq. (2.40) simplifies to

$$d_{\text{opt}} = \frac{\lambda}{n - \sqrt{1 - \left(\frac{\lambda}{\Lambda}\right)^2}}. \tag{2.41}$$

Similar formulas have been derived for multi-level gratings and are presented in Ref. [31].

Step transition method

While the previous corrections are based on rather empirical considerations, it is also possible to correct the TEA results more rigorously. Kettunen *et al.* have proposed such a method [33, 34]. Firstly, a collection of amplitude and phase functions generated by several step transitions is created. Then, for every step transition, the corresponding corrective function is added locally to the amplitude and phase resulting of the TEA. This method is accurate as long as the neighbouring transitions are a few wavelengths apart from each other.

As the rigorous computation is performed once for all, the speed of this method is comparable to the TEA computation. However, it can take care of the difference between TE and TM polarisation.

2.2.5 Geometrical optics

When the wavelength is small enough in comparison with the dimensions present in the system, it is possible to rewrite the propagation laws in a simple form. This is the field of optics using the notion of rays of light, also known as geometrical optics. A complete discussion of this domain may be found in chapter 3 of Ref. [35]. We will succinctly recall the main equations leading to the concept of ray, refraction and to Snell's law.

Assuming that the light field can be written as

$$U(\mathbf{r},t) = A(\mathbf{r}) \exp\left(\mathrm{i}\left(\omega t - k_0 \mathcal{S}(\mathbf{r})\right)\right), \tag{2.42}$$

the Helmholtz equation (2.18) leads to

$$\mathrm{e}^{-\mathrm{i}k_0\mathcal{S}}\left[\nabla^2 A - \mathrm{i}k_0\left(2\nabla A\cdot\nabla\mathcal{S} + A\nabla^2\mathcal{S}\right) - k_0^2 A\left((\nabla\mathcal{S})^2 - n^2\right)\right] = 0 \tag{2.43}$$

with

$$k_0 = \frac{\omega}{c}. \tag{2.44}$$

If we assume that

$$k_0 A \gg |\nabla A| \tag{2.45}$$

and

$$k_0 \gg |\nabla n|, \tag{2.46}$$

which means that the relative variations of A and of the refractive index n are small at the wavelength scale (weakly perturbative medium). Equation (2.43) leads to

$$(\nabla\mathcal{S})^2 = n^2, \tag{2.47}$$

known as the Eikonal equation. Note that Eq. (2.42) is general enough to describe both spherical waves ($\mathcal{S} = r$) and plane waves ($\mathcal{S} = \frac{\mathbf{k}}{k}\cdot\mathbf{r}$). Since the Eikonal is related to the phase by Eq. (2.42), its iso-surfaces represent the wavefronts. The optical rays, perpendicular to the wavefronts, are thus the three-dimensional linear trajectories $\mathbf{r}(s)$ defined by

$$n\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}s} = \nabla\mathcal{S}. \tag{2.48}$$

Further manipulation of Eq. (2.48) leads to

$$\frac{\mathrm{d}}{\mathrm{d}s}\left[n\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}s}\right] = \nabla n, \tag{2.49}$$

a differential equation implying that in a homogeneous medium ($\nabla n = 0$) the rays are straight lines. Variational analysis states that solving Eq. (2.49) is equivalent to finding the curves along which the optical path length

$$\mathrm{OPL} = \int_M^N n(\mathbf{r}(s))\,\mathrm{d}s \tag{2.50}$$

between two fixed points M and N in space is stationary, which is known as the principle of Fermat. From Eq. (2.50), one can deduce Snell's law

$$n_1 \sin(i_1) = n_2 \sin(i_2) \tag{2.51}$$

describing the behaviour of light rays at an interface between two homogeneous media, of refractive indices n_1 and n_2, respectively, when i_1 and i_2 are the angles with respect to the surface normal.

While geometrical optics does not include diffraction phenomena, it is nonetheless interesting for the design of beam shaping optical elements. Indeed, it is very fast and can be applied using commercial software. These programs perform ray-tracing, a technique consisting in simulating beams of light by a large number of individual rays and propagating them through the system following Eq. (2.51). There exist two types of ray-tracing, *sequential* and *non-sequential* [36]. Sequential ray-tracing models the interaction of light as a sequence (an ordered set) of interactions between the rays and the optical elements. It thus cannot model partial reflection, multiple interactions of a ray with a structure. This is similar to the approximation used in the TEA or in its derivatives like the BPM or the WPM. The speed of all these methods is actually due to the hypothesis that interaction occurs in a sequential and predetermined order. Non-sequential ray-tracing, on the other hand, although using the very same equations, can model the division of light rays, partial reflection, scattering, multiple interactions of rays with structures. Similarly to rigorous methods or scalar optics (strictly speaking), non-sequential ray-tracing requires huge computation times, sometimes many days, compared to the few seconds typically needed in sequential ray-tracing.

It appears that the frontier lines that are usually placed between ray-tracing, scalar optics and rigorous theories do not describe completely the relative accuracy of the methods. It would be useful to draw another distinction concerning the accuracy of the interaction with extended media, where non-sequential ray-tracing, rigorous theories and scalar optics (strictly speaking) would be separated from sequential ray-tracing, TEA and its cousins.

Nevertheless, considering the beam shaping in the far-field, under paraxial approximation, the elements are not extended more than one wavelength in depth and the application of the second set of methods is completely justified.

Simulation of large optical elements

3

We have seen in chapter 2 that design and simulation for large optical elements should preferably be based on ray-tracing or scalar optics, but not on rigorous theory, because of time and amount of data. However, using scalar optics and thin element approximation (TEA) does not yet guarantee that the memory requirements will not overrun the capabilities of the computer. In fact, most of the optical beam shaping elements presented in this work are too large for a straightforward application of these methods.

Consequently, we have developed a computational technique that tremendously reduces the memory requirements [37, 38]. The technique, which we shall call the *shape technique*, is based on a geometric description of the phase structures of a diffractive optical element (DOE). It makes it possible to simulate large elements (large meaning here "whose dimensions are of thousands or tens of thousands of wavelength in both directions"). This technique was originally developed for Fraunhofer diffraction of multilevel optical elements. While not as versatile as two-dimensional sampling, the geometric description can be generalised to linearly blazed gratings. Because of the reduced memory requirements, it can be used for both periodic and aperiodic structures.

3.1 Description of an optical element

Be it for simulation, design or fabrication, the element structure has to be defined in terms of its geometric and optical properties. We may distinguish three uses for the description of an optical element, each with its own constraints. We shall illustrate this distinction with the example of a multi-level Fresnel zone plate (FZP).

First, we have to be able to design the element, that is to describe it with respect to a goal. Practically, the angles defining the light distribution, the wavelength and the size of the individual FZP lead to a radial phase profile, which in turn can be expressed as a set of radii and heights of concentric rings.

Second, we have to simulate this optical element. Therefore, it must be described in such a way that we can apply one of the propagation equations (2.25) to (2.34). For our example, this may be performed through a sampled representation of the FZP. But it could also be done with Bessel functions or Hankel transform.

Finally, we have to represent the optical element for the fabrication. Mask generation usually makes use of polygons [39], and we therefore would need to convert the rings of our FZP into an approximated set of polygonal primitives, or into pixels.

We would like to emphasize that these three steps are independent, even if they may use a common description of the structure.

Defining an optical element is equivalent to defining the interaction properties all over its surface. Under the approximation of the TEA, this is equivalent to the knowledge of the profile and the absorption. For phase only elements, only the height profile is required.

There are classically two alternatives for the profile definition. Digital descriptions, well adapted to the computer architecture, and analytical ones, less versatile but with certain advantages. Two-dimensional sampling and polygonal decomposition are examples of the first kind. Polynomial expansion and geometric decomposition illustrate the second possibility. Further on, we shall coin the term geometric decomposition for any description of an element based on domains whose borders are defined following their geometry. The value of the profile in each such domain is described by means of an analytical function. A good illustration is the description of the multi-level Fresnel zone plate as a set of rings. Each ring is defined by its internal and external radii. The profile is constant and defined by the height value. This example illustrates the small number of data needed to store the profile.

3.1.1 Tessellated description

The principle is explained in Fig. 3.1. On the left we have represented a continuous function. The sampling points, localised on a two-dimensional grid are represented by the arrows whose heights are proportional to the function value. This description does not define the element all over the interaction plane, but only at a discrete number of points. To achieve our goal, we then define a pixel representation of the profile function. Every sample point is now

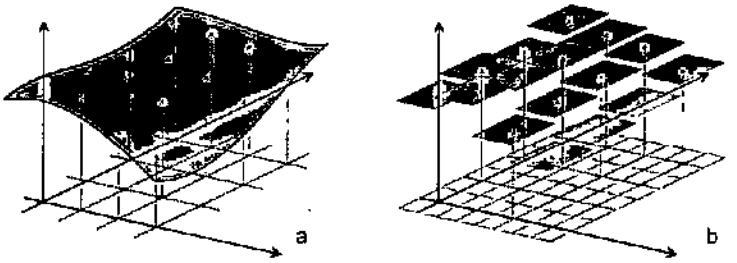


Figure 3.1: A profile is sampled on a two-dimensional grid (a). It is then tessellated corresponding to this grid (b).

the centre of a square. The process of representing a continuous function by an approximated mosaic is called *tessellation*. By analogy with computer science and image processing where the pictures are composed of pixels, we shall use indifferently pixelation and tessellation.

Two-dimensional pixelation of the optical element profile (or equivalently, of the phase) is widely used, because of the advantages it brings for computation. Firstly, its similarity with computer images implies that all the filters and operators of image processing can be applied to it (convolution, edge detection, interpolation, etc.). This also brings practical bonus for the fabrication of the element. Indeed, the pixel representation is well suited to the mask generation process. Electron beam machines used e.g. for writing lithography masks typically have preferred directions, because of the X-Y translation stages.

Finally, this discretisation corresponds also to the commonly used measuring devices. One can measure the input beam with a pixelated sensor and easily use the result on the pixelated transmittance, implementing Eq. (2.37).

On the other hand, this tessellation of the phase has drawbacks. Replacing a sample (the value at a defined position) by an extended pixel is not exact. The sampling grid might not be fine enough to catch all the details of the profile function, but may instead add high frequency variations to it. The fabrication errors may concentrate in some directions due to the pixel edges and can consequently generate artifacts. Even in the absence of fabrication errors, the pixels are responsible for a sinc envelope, lowering the efficiency of the element [40]. Finally, the sampling frequency may beat with local frequencies of the structure, resulting in a moiré phenomenon [41–45]. Precautions should be taken, such as accounting for the Nyquist criterion, stating that the sample frequency should be at least twice the highest frequency of the structure.

However, the main restriction of the two-dimensional sampling is the amount of data required to store the structure of real elements. Indeed, if N is the number of samples in one direction, the data evolves as $\mathcal{O}\left(N^2\right)$. Even if the

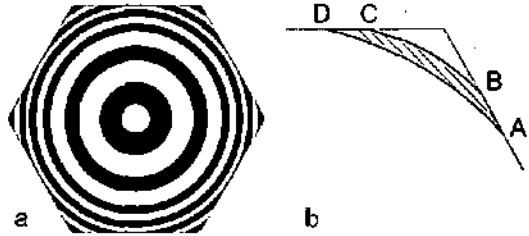


Figure 3.2: An hexagonal Fresnel zone plate (a) may be decomposed into concentric rings where it is symmetrical. Where the symmetry does not hold, it still can be defined by zones whose contour is a sequence of primitives (b). Here, two segments of lines AB and CD and two sectors of circles BC and DA.

profile variations are slow compared to the pixel size, the pixel representation still has to store as many points, and cannot take advantage of the areas of contiguous pixels of same value. In other words, the pixel grid is not adaptive, but constant all over the element.

3.1.2 Geometric description

The other widely used description of an optical element is geometric: Fresnel zone plates are made of concentric rings. If the profile of each ring is constant, it is obvious that the data required to describe such an element is very limited: the radii of the rings and their transmittance values (phase and amplitude). If the phase is not constant, the transmittance may be stored with a few more parameters. For instance, a linearly blazed profile would require the equation of the blaze, defined by two parameters.

Describing an element via primitives can he done if we have a curve database. For example, a Fresnel aperture modulated diffuser (AMD) consists of sectors of rings truncated by the cell shape. It is thus straightforward to describe such a DOE with arcs of circles and straight line segments, as illustrated by Fig. 3.2. For a more general DOE structure, the cubic Bézier curves might be a useful set of primitives (see appendix A). Moreover, this primitive decomposition is very well adapted to the mask generation process.

The main advantages of this kind of decomposition compared with the pixel representation are the small number of data used to described the DOE and the capacity to represent structures with convoluted geometries with great fidelity.

3.1.3 Other representations

Other representations worth mentioning here are the polygonal and polynomial descriptions of a function. The first approximates the profile by polygons of constant transmittance. Consequently, it is well suited for multi-level diffractive structures. However, the main application is for mask generation. Mask data is usually stored in a polygon format, e.g. Caltech intermediate format (CIF) or GDS II. These polygons are projected on a grid whose pitch can be as small as 10 nm [39, 46]. These formats are intermediate, which means they are transformed into a final format that depends on the machine used for mask generation [47].

Alternatively, the polynomial description expands the profile function on a basis of polynomials [6]. It is therefore much more suited for continuous profiles. However, the described profile cannot be very complicated. The polynomials basis might be of the form $\sum_{k=0}^{N} a_k x^k \cdot y^{N-k}$ or different, like Zernicke polynomials. They might also be replaced by a more general function basis, such as splines. Expanding a profile on a basis of functions is however time consuming. This restricts drastically the usefulness of this description in practice.

3.2 Implementation of scalar optics

Now that we have defined the function of our structure, we want to compute the light distribution away from the element. Consequently, we need to apply Fresnel or Fraunhofer formulas. We shall now present the classical alternatives used for this purpose. We shall demonstrate that none of them fulfils our requirements, and that we need a new implementation of the scalar optics kernel.

3.2.1 Sampled implementation

Similarly to the two-dimensional pixelation of the structure used to describe an optical element, a two-dimensional sampling can be used to calculate the propagation. The main interest of the sampled description lies in its versatility. Indeed, the sampling process gets rid of the non separability of the problem. The phase function $f(x, y)$ is transformed into a discrete and finite array. Many mathematical operators and transformations can be easily applied in a separable way to such arrays. Another bonus is the ease of implementation. Since Eqs. (2.32) and (2.34) are Fourier transforms, the two-dimensional

discrete Fourier transform (DFT) can be used. In practice, the computation is based on algorithms like the fast Fourier transform (FFT).

Furthermore, the sampled description is naturally adapted to simulate the pixelated optical elements. This similarity between both the structure of the element and the simulation greatly simplifies the relations between design, analysis and fabrication. It is straightforward to implement a feedback loop from the simulation to the design if every pixel is related to a sampling point. This allows the use of effective iterative algorithms like iterative Fourier transform (IFTA) [48,49].

Finally, it should be noted that every description can be expressed as a set of samples. Thus, this implementation can in principle be used with any of the element description. For instance, from the radii defining a Fresnel lens, it is straightforward to define a two-dimensional sampling for the phase of the lens.

Nevertheless, the main drawback of this simulation method is the data storage required for large elements. It is at least as large as for the pixelated element design, and often larger because one usually has to apply zero-padding in the analysis. Zero-padding consists of simulating the surrounding of the function by a large quantity of zeros in order to work around aliasing artifacts [12]. In a one-dimensional simulation, this doubles the number of data. In two-dimensional simulations, the amount of data is multiplied by four, three quarters of which are just zeros.

3.2.2 Analytical implementation

For rotationally symmetrical structures, the simulation of propagation can be performed either using Bessel functions, or fast Hankel transform [50]. Unfortunately, not all the optical elements exhibit rotation symmetry. Aperture modulated diffusers (AMDs) are based on Fresnel zone plates, but they have to be tiled in order to exhibit a fill factor close to the one [51]. Only rectangular, triangular and hexagonal cells allow the tiling of a plane. Thus AMDs are most often based on square or hexagonal elementary cells. Similarly, one-dimensional elements – cylindrical lenses for example – may also be simulated analytically.

Other structures do not have any symmetry of rotation, as for instance the ones obtained from iterative algorithms (IFTA, simulated annealing, etc.) or some phase zone plates (PZP) obtained from other design techniques, to be presented in chapter 5. Thus, analytical solutions are not suitable any more. Neither can we rely on them if the input light distribution does not share the symmetry of the element.

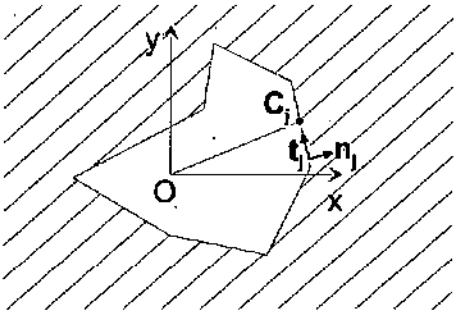


Figure 3.3: Definition of the parameters used in the formula of Fraunhofer diffraction by a polygonal aperture.

3.2.3 Polygonal implementation

There exists a third implementation of the Fraunhofer equation. It is based on the Fourier transform of a triangular aperture [52, 53]. Since a polygon can be decomposed in a set of triangles, authors have proposed a formula implementing Fraunhofer diffraction for polygonal apertures in a screen. This formula has been extended for Fresnel diffraction in order to be able to study the artifacts due to polygonisation of the Fresnel zone plates [54]. The formula is based on the Abbe transform

$$U(\mathbf{u}) = \frac{\mathrm{i}}{Sku^2} \oint_\ell \exp(-\mathrm{i}k\mathbf{u}\cdot\mathbf{x})\,\mathbf{u}\cdot\mathbf{n}\mathrm{d}\ell, \tag{3.1}$$

where ℓ is a closed contour around the aperture of area S. The vector $\mathbf{n}$ is the normal to ℓ, oriented towards the outside of the aperture, $\mathbf{x} = (x, y)$ is moving on the curve ℓ and $\mathbf{u} = (u, v)$ is the vector of the output plane coordinates. As shown in Fig. 3.3, a polygon is defined by the positions $\mathbf{x}_j$ of the centre points C_j of the N sides of length L_j. With $\mathbf{n}_j$ and $\mathbf{t}_j$ the normal and the tangent vectors of the side, the field at $\mathbf{u}$ can be then written as

$$U(\mathbf{u}) = \frac{\mathrm{i}}{Sku^2} \cdot \sum_{j=1}^{N} \mathbf{u}\cdot\mathbf{n}_j L_j \exp(-\mathrm{i}k\mathbf{u}\cdot\mathbf{x}_j)\,\mathrm{sinc}\left(k\mathbf{u}\cdot\mathbf{t}_j\frac{L_j}{2\pi}\right) \tag{3.2}$$

3.2.4 The need for a new formulation

While two-dimensional sampling is very practical and can describe almost any optical structure, the data requirements are still to big for large two-dimensional elements. Analytical solutions and symmetrical solutions are the only ones to allow the analysis of large planar structures. However, they can

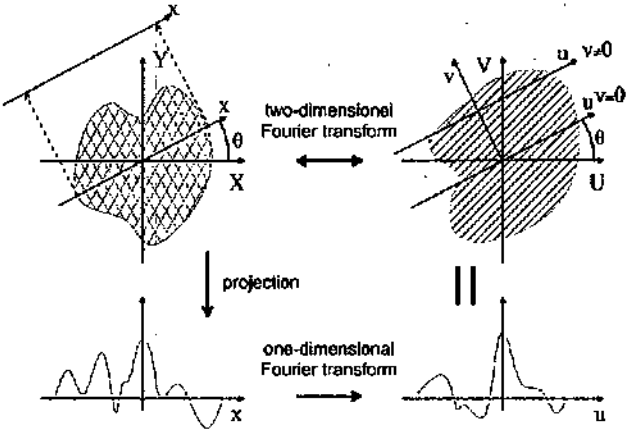


Figure 3.4: Principle of the central slice theorem. The line at angle θ and at $v = 0$ of the two-dimensional Fourier transform is equal to the one-dimensional Fourier transform of the projection at angle θ.

only be applied to a restricted subset of the optical elements, excluding most of the ones needed in beam shaping. Thus, there is a need to fill the gap between both representations. The description of the element should be geometric to reduce the amount of data used. On the other hand, we have to propose an implementation of the Fresnel and Fraunhofer formulas that can make use of such geometric properties.

3.3 Hybrid two-dimensional Fourier transform

We shall demonstrate a hybrid two-dimensional Fourier transform. This procedure is composed of a continuous Fourier transform in one direction, followed by a discrete Fourier transform in the perpendicular direction. It is consequently less sensitive to aliasing. We shall first apply it to the Fraunhofer diffraction formula, then extend it to Fresnel diffraction.

3.3.1 The Radon transform and the central slice theorem

In tomography, the Radon transform

$$\mathcal{R}\{f\} = F(x) = \int_{-\infty}^{+\infty} f(x, y)\,\mathrm{d}y, \tag{3.3}$$

is used to rebuild the object function $f(x,y)$ from radial lines of its Fourier transform [55]. This is related to the central slice theorem, illustrated by Fig. 3.4. The Radon transform of Eq. (3.3) is the mathematical expression of the projection parallel to the y axis of $f(x,y)$ onto the x axis, as represented by the dashed arrows. With the notation of Fig. 3.4, we see that the whole Fourier space can be addressed with radial scans, by varying the angle θ.

From this theorem, we see that it is possible to avoid the two-dimensional Fourier transform which requires the knowledge of the transmittance at every point of the two-dimensional sampling grid. We just have to be able to compute the Radon transform of the phase of the optical element and its Fourier transform

$$\mathcal{F}\{F(x),u\} = \int_{-\infty}^{+\infty} F(x)\exp(-2i\pi ux)\,\mathrm{d}x = \tilde{f}(u,v=0). \tag{3.4}$$

The expression *central slice* means that the line of the Fourier transform that is computed passes by the origin of the Fourier plane $(U,V)=(0,0)$, which is equivalent to state $v=0$.

3.3.2 Application to Fraunhofer diffraction

We shall now define an hybrid Fourier transform based on the principle of the central slice theorem. From now on, for Fraunhofer diffraction, we will use the notation of Fig. 3.5. Two conjugate (input-output) systems of axes are introduced, the global ones (O,X,Y) and (O',U,V), and the rotated ones (O,x,y) and (O',u,v), the rotation angle between both being θ. The optical element is described with respect to the global axes, while the computation is performed in the rotated coordinate system. To differentiate the output plane coordinates in the Fraunhofer diffraction (spatial frequencies) from those in the Fresnel diffraction (positions), we shall call them (O,X',Y') and (O',x',y').

For the sake of clarity, we will first limit ourselves to an optical element whose transmittance is a set of M zones of complex amplitude C_m, and which is illuminated by a uniform, normally incident plane wave $(T(x,y)=1)$. Some generalisations will be introduced later in the chapter. Under the considered simplifications, the amplitude of the field right after the element can be written as

$$U(x,y,0) = \sum_{m=1}^{M} \delta_m(x,y)\cdot C_m, \tag{3.5}$$

following Eq. (2.37), where $\delta_m(x,y)$ is one for (x,y) inside zone number m and zero elsewhere.

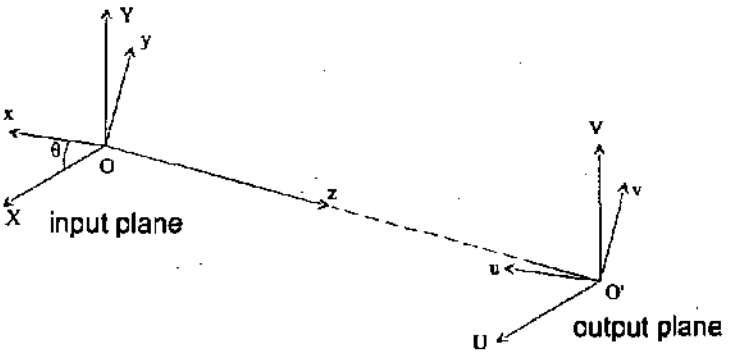


Figure 3.5: Definition of the input plane and output plane for the Fraunhofer diffraction geometry.

In addition, we will assume the uniqueness of the lower and upper boundaries $a_m(x)$ and $b_m(x)$ as illustrated by Fig. 3.6(a). The projection at angle θ of a such a function is expressed as:

$$F(x) = \sum_{m=1}^{M} \int_{a_m(x)}^{b_m(x)} C_m \mathrm{d}y = \sum_{m=1}^{M} C_m \left[b_m(x) - a_m(x)\right]. \tag{3.6}$$

If the element contains shapes whose contours are not unique, we have to take care that we do not integrate over areas which are outside the contour, as in the case of the dash-dotted line in Fig. 3.6(b). A straightforward workaround is to decompose every such shape into simple contours, as illustrated in the figure. Thus, Eq. (3.6) remains valid.

For the Fraunhofer diffraction, Eq. (2.34) can be rewritten as a Fourier transform

$$\tilde{f}(u,v) = \iint_{-\infty}^{+\infty} f(x,y) \exp\left[-2i\pi\left(ux + vy\right)\right] \mathrm{d}x\mathrm{d}y. \tag{3.7}$$

Using the central slice theorem, we get a one-dimensional Fourier transform. Additionally, if $f(x,y)$ is expanded as a set of zones of constant value C_m, this equation is simply a sum of one-dimensional Fourier transforms

$$\tilde{f}(u, v=0) = \mathcal{F}\left\{\sum_{m=1}^{M} C_m \left[b_m(x) - a_m(x)\right], u\right\} \tag{3.8}$$

We shall now extend Eq. (3.8) to axes of u that are not bound to be passing by $(U,V) = (0,0)$. Since computation of the borders $a_m(x)$ and $b_m(x)$ has

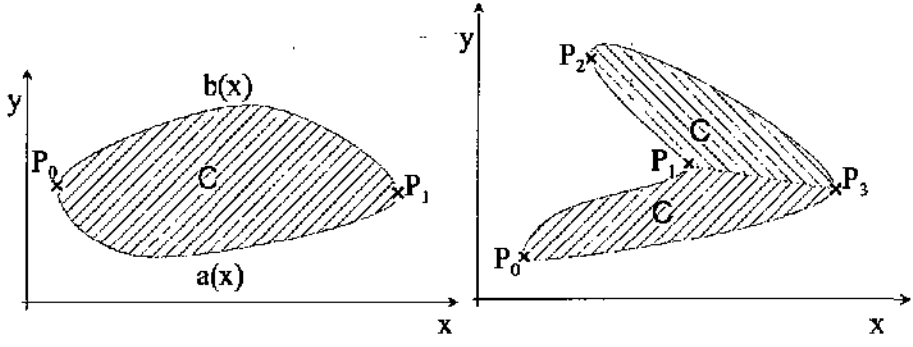


Figure 3.6: (a) A simple shape with unique upper and lower contour lines $a(x)$ and $b(x)$. A complex contour (b) can be decomposed into simple contours, here $P_0P_1P_3P_0$ and $P_1P_2P_3P_1$.

to be repeated for every angle of projection θ, it is preferable, if possible, to scan along the v axis of Fig. 3.4 instead. The Radon transform is a Fourier transform applied only in one dimension to a two dimensional distribution.

For a fixed v, and if $f(x,y)$ can be expressed as a set of constant-phase shapes, Eq. (3.7) can be rewritten as:

$$\tilde{f}(u,v) = \int_{-\infty}^{+\infty} \left[\sum_{m=1}^{M} C_m \int_{a_m(x)}^{b_m(x)} \exp(-2i\pi vy)\,dy \right] \exp(-2i\pi ux)\,dx, \quad (3.9)$$

which is a one-dimensional Fourier transform. The term between the brackets is identical to Eq. (3.6) for $v = 0$ and can be computed analytically for other values of v. Thus we obtain in the general case

$$\tilde{f}(u,v) = \int_{-\infty}^{+\infty} f_v(x) \exp(-2i\pi ux)\,dx, \quad (3.10)$$

with

$$f_v(x) = \begin{cases} \sum_{m=1}^{M} \frac{C_m}{2i\pi v} \left[\exp(-2i\pi v a_m(x)) - \exp(-2i\pi v b_m(x)) \right] & v \neq 0 \\ \sum_{m=1}^{M} C_m \left[b_m(x) - a_m(x) \right] & v = 0 \end{cases} \quad (3.11)$$

We have expressed one line of the two-dimensional Fourier transform of a function as a one dimensional Fourier transform. It is worth noting that that the sums of Eq. (3.11) are easily evaluated. However, to obtain this expression, we have had to make various simplifications, namely

1. The element is illuminated by a plane wave.
2. This plane wave falls on the element at normal incidence.
3. The elements can be decomposed in a set of contours.
4. The profile is constant inside such contours.

We shall now see that this approach can be extended to Fresnel propagation, obliquely incident plane waves and linearly blazed profiles.

3.3.3 Obliquely incident plane waves

If the illuminating plane wave is falling on the element at oblique incidence, Eq. (3.9) is changed to

$$\tilde{f}(u,v) = \int_{-\infty}^{+\infty} \left[\sum_{m=1}^{M} C_m \int_{a_m(x)}^{b_m(x)} \mathrm{e}^{-\mathrm{i}(2\pi v + k_y)y} \mathrm{d}y \right] \mathrm{e}^{-\mathrm{i}(2\pi u + k_x)x} \mathrm{d}x, \qquad (3.12)$$

where the wave vector of the incident light is

$$\mathbf{k} = \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix}, \qquad (3.13)$$

and its components are related by

$$\|\mathbf{k}\| = \sqrt{k_x^2 + k_y^2 + k_z^2} = k = \frac{2\pi}{\lambda}. \qquad (3.14)$$

Then, Eqs. (3.10) and (3.11) are reformulated as

$$\tilde{f}(u,v) = \int_{-\infty}^{+\infty} f_v(x) \exp\left[-2\mathrm{i}\pi \left(u + \frac{k_x}{2\pi} \right) x \right] \mathrm{d}x, \qquad (3.15)$$

and $f_v(x)$ becomes

$$f_v(x) = \begin{cases} \sum_{m=1}^{M} \frac{C_m}{i(2\pi v + k_y)} \left[e^{-i(2\pi v + k_y) a_m(x)} - e^{-i(2\pi v + k_y) b_m(x)} \right] & v \neq -\frac{k_y}{2\pi} \\ \sum_{m=1}^{M} C_m \left[b_m(x) - a_m(x) \right] & v = -\frac{k_y}{2\pi} \end{cases} \tag{3.16}$$

In terms of implementation, we now have several possibilities to compute the far-field resulting from an obliquely incident plane wave:

1. Computing the result of a normally incident plane wave and translating the output. This is correct in the paraxial approximation.

2. Computing from Eqs. (3.15) and (3.16).

3. Computing from Eq. (3.16) the result of the Fourier transform in y, and then using Eq. (3.10) with a translation in x. Or the opposite way, using Eq. (3.11) with (3.15) and a translation in y.

To illustrate the effectiveness of Eqs. (3.15) and (3.16), we compared the theoretical and measured light distribution produced by two hexagonal flat-top aperture modulated diffusers (AMDs). The two diffractive optical elements (DOEs) were binary Fresnel zone plates (FZPs) of different phase offset. In Fig. 3.7, the magnification of the central part of the output plane distributions are shown, when both lenses are illuminated by a plane wave from a He-Ne laser at 633 nm. The simulations not only predict the radii of the light and dark rings, but also their relative intensity. Moreover, the far-field of the DOE and of gratings of known frequencies were projected on a screen were measurements were performed, the far-field of the gratings being used

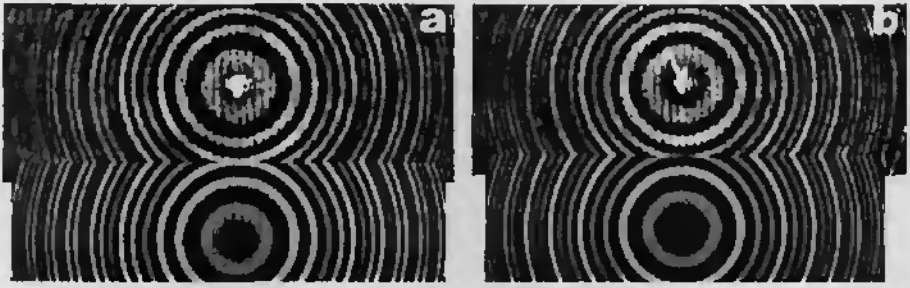


Figure 3.7: Magnification of the centre of the light distribution generated by two Fresnel zone plates of different phase offsets ($0.18 \times 2\pi$ (a) and $0.25 \times 2\pi$ (b)). Comparison between simulated (bottom) and measured (top) far-field intensities.

as angular references. The full width at half maximum (FWHM) of the two light distributions generated by the DOEs where measured and found to be predicted with an error of less than 1%, that is better than the measurement accuracy in this experiment. The FWHM was about 6° and the outer diameter of the FZP was 2 mm. The central peak is due to fabrication errors (etching depth mainly) and the loss of contrast of the outer rings can be attributed to the sensor pixel size.

3.3.4 Fresnel propagation

Equation (2.32) for the Fresnel propagation can also be rewritten as a one-dimensional Fourier transform

$$\begin{aligned} U(x',y',z) &= \frac{e^{ikz}}{i\lambda z}\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty}U(x,y,0)\,e^{i\frac{k}{2z}(y-y')^2}dy\right]e^{i\frac{k}{2z}(x-x')^2}dx \\ &= \frac{e^{ikz}}{i\lambda z}e^{i\frac{k}{2z}x'^2}\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty}U(x,y,0)\,e^{i\frac{k}{2z}(y-y')^2}dy\right]e^{i\frac{k}{2z}x^2}e^{-2i\pi\frac{x'}{\lambda z}x}dx \end{aligned} \tag{3.17}$$

For a multilevel element illuminated by a normally incident plane wave, this equation takes the form

$$U(x',y',z) = \frac{e^{ikz}}{i\lambda z}e^{i\frac{k}{2z}x'^2}\mathcal{F}\left\{\left(\sum_{m=1}^{M}C_m\int_{a_m(x)}^{b_m(x)}e^{i\frac{k}{2z}(y-y')^2}dy\right)e^{i\frac{k}{2z}x^2},\frac{x'}{\lambda z}\right\} \tag{3.18}$$

As previously, we have to be able to compute analytically the integral inside the sum to achieve a separation of both dimensions in the propagation kernel. This integral is of the form of the Fresnel integral $C(z)+iS(z)$, where $z = x + iy$). It can be solved using of the complex error function $w(z)$, presented in appendix B. The Fresnel integral of Eq. (3.18) can be computed from

$$\int_0^z \exp\left(t^2\right)dt = i\frac{\sqrt{\pi}}{2}\left[1 - w(z)\cdot\exp\left(z^2\right)\right]. \tag{3.19}$$

if we introduce the change of variables

$$\begin{aligned} t &= e^{i\frac{\pi}{4}}\sqrt{\frac{\pi}{\lambda z}}(y-y') \\ dt &= e^{i\frac{\pi}{4}}\sqrt{\frac{\pi}{\lambda z}}dy \end{aligned} \tag{3.20}$$

The final result is

$$\begin{aligned}\int_{a_m(x)}^{b_m(x)} \mathrm{e}^{\mathrm{i}\frac{k}{2z}(y-y')^2}\,\mathrm{d}y &= \mathrm{e}^{-\mathrm{i}\frac{\pi}{4}}\sqrt{\frac{\lambda z}{\pi}}\int_A^B \exp\left(t^2\right)\mathrm{d}t \\ &= \mathrm{e}^{\mathrm{i}\frac{\pi}{4}}\frac{\sqrt{\lambda z}}{2}\cdot\left[w\left(A\right)\cdot\mathrm{e}^{A^2} - w\left(B\right)\cdot\mathrm{e}^{B^2}\right] \quad (3.21)\end{aligned}$$

where A and B are defined as

$$\begin{aligned}A &= \mathrm{e}^{\mathrm{i}\frac{\pi}{4}}\sqrt{\tfrac{\pi}{\lambda z}}\,(a-y') \\ B &= \mathrm{e}^{\mathrm{i}\frac{\pi}{4}}\sqrt{\tfrac{\pi}{\lambda z}}\,(b-y')\end{aligned} \quad (3.22)$$

Obliquely incident plane waves

Equations (3.18), (3.21) and (3.22) may be generalised to obliquely incident plane waves as previously. The Fresnel propagation becomes then

$$\begin{aligned}U\left(x',y',z\right) &= \tfrac{\mathrm{e}^{\mathrm{i}\left(kz+k_y y'\right)}}{\mathrm{i}\lambda z}\exp\left[\mathrm{i}\tfrac{k}{2z}\left(x'^2-\tfrac{k_y^2}{k^2}z^2\right)\right] \\ &\quad\times\mathcal{F}\left\{\left[\sum_{m=1}^{M}C_m\int_{a_m(x)}^{b_m(x)}\mathrm{e}^{\mathrm{i}\frac{k}{2z}\left[y-\left(y'-\frac{k_y}{k}z\right)\right]^2}\mathrm{d}y\right]\mathrm{e}^{\mathrm{i}\left(\frac{k}{2z}x^2+k_x x\right)},\tfrac{x'}{\lambda z}\right\},\end{aligned} \quad (3.23)$$

and for the Fresnel integral corresponding to one contour analytical computation one gets

$$\int_{a_m(x)}^{b_m(x)} U\left(x,y,0\right)\mathrm{e}^{\mathrm{i}\frac{k}{2z}\left[y-\left(y'-\frac{k_y}{k}z\right)\right]^2}\mathrm{d}y = \mathrm{e}^{\mathrm{i}\frac{\pi}{4}}\frac{\sqrt{\lambda z}}{2}\cdot\left[w\left(A\right)\cdot\mathrm{e}^{A^2} - w\left(B\right)\cdot\mathrm{e}^{B^2}\right], \quad (3.24)$$

with

$$\begin{aligned}A &= \mathrm{e}^{\mathrm{i}\frac{\pi}{4}}\sqrt{\tfrac{\pi}{\lambda z}}\left[a-\left(y'-\tfrac{k_y}{k}z\right)\right] \\ B &= \mathrm{e}^{\mathrm{i}\frac{\pi}{4}}\sqrt{\tfrac{\pi}{\lambda z}}\left[b-\left(y'-\tfrac{k_y}{k}z\right)\right]\end{aligned} \quad (3.25)$$

Application: single hexagonal lens

We applied the previous formulas to the computation of the focal pattern of a hexagonal lens illuminated with a normally incident plane wave at 633 nm. The profile of the unwrapped lens is parabolic, encoded as a two-level phase

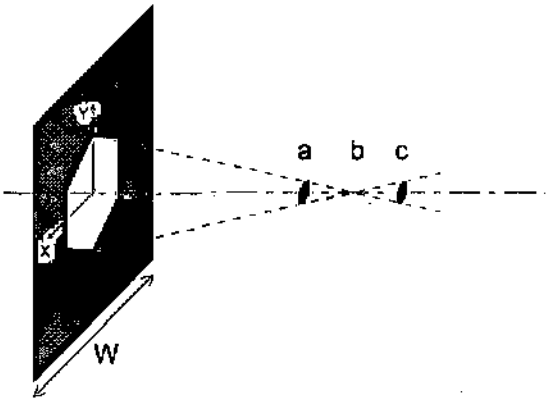


Figure 3.8: The setup for the Fresnel propagation simulation of an hexagonal lens. The zero-padding W is 128 mm. Three planes chosen for the simulation are (a) just before the focus, in focus (b) and after the focus (c).

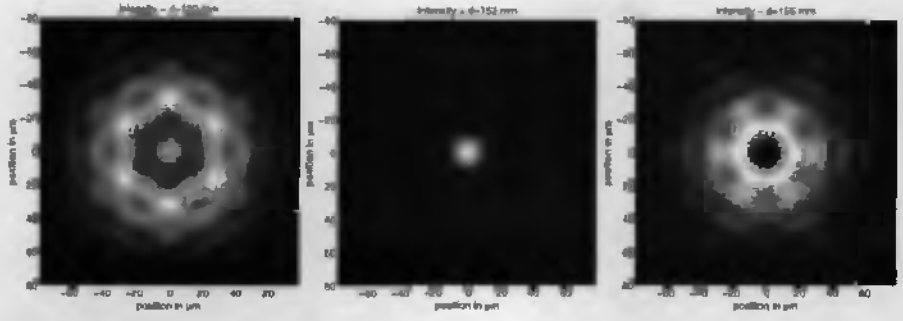
lens. The outer diameter of the hexagon is Ø=8 mm. The lens has an angular aperture θ_{NA} of $\pm 1.5°$. This corresponds to a focal length of

$$f = \frac{\text{Ø}}{2\tan(\theta_{NA})}, \tag{3.26}$$

which in the present case yields $f = 152.7$ mm. The simulation was performed with 32×1024 points sampling an input window of width $W = 128$ mm. This wide zero-padding was used to obtain a high resolution in the output planes. The setup is presented in Fig. 3.8.

Images of the light distribution in the focal area are presented in Fig. 3.9. Three transversal planes were chosen to display the light distribution. One was placed before the first order focus, one near the focus, and the last one after the focus. The effect of the non rotationally symmetrical aperture is clearly visible. In the focus, the light distribution is the Fourier transform of the aperture. The formulas of Fresnel diffraction also allow to see the other diffraction orders. Since the lens profile is binary, no even orders are present, but odd orders can clearly be observed, for instance order three at $f/3$.

Figure 3.10 shows the intensity of the light distribution on transversal lines at different positions along the optical axis. The scans are radial, and at angles $\theta = 0°$ and $\theta = 30°$. The focus is between 152.5 mm and 153 mm as predicted from the designed far-field numerical aperture. It can be seen from Fig. 3.10, that near the focus the influence of the shape of the aperture is reduced, but still present. Scans at 150 mm, 152 mm and 155 mm, respectively, correspond to the left, centre and right images of Fig. 3.9. The difference between the two



(a) (b) (c)

Figure 3.9: Images of the intensities in the planes at 150 mm (a), 152 mm (b) and 155 mm (c).

characteristic directions of the hexagon is more visible out of focus, especially at 150.5 mm, 151 mm and 155 mm. The depth of focus deduced from these scans is about 2.5 mm.

3.3.5 Linearly blazed profile

We have seen that we can apply the same principle based the central slice theorem, namely the hybrid transform, to both the Fourier and Fresnel diffraction kernels. Both formulas, although first established for normally incident plane waves, can be easily extended to obliquely incident plane waves. We will now tackle the last restriction we set at the beginning of this section, the multi-level nature of the profile.

While we cannot apply the separation of the kernel in two directions for any profile, we still can imagine how the other most often used profile, the linearly blazed, could be expressed using the same principle. Instead of being defined by a unique parameters (height or phase value), the blazed profile is of the form

$$h(x,y) = c(x) \cdot y + d(x), \tag{3.27}$$

which changes Eq. (3.5) into

$$U(x,y,0) = \sum_m C_m(x)\, \delta_m(x,y) \cdot \exp\left(\mathrm{i}k\,(n-1)\, c_m(x) \cdot y\right), \tag{3.28}$$

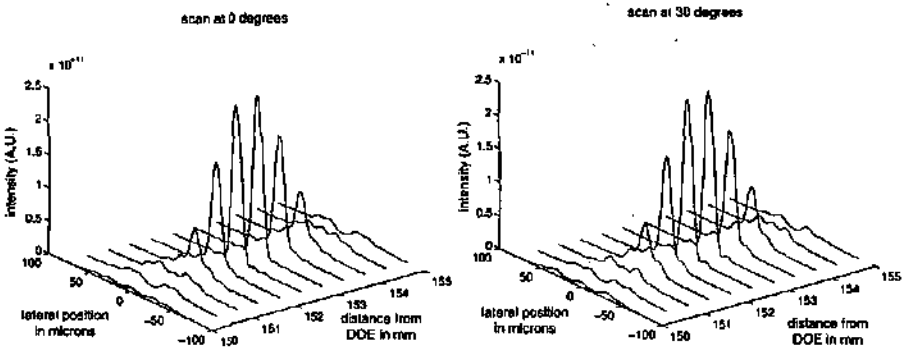


Figure 3.10: Light distribution at different values of the position on the axis around the focus. Scans in the hexagons inner diameter (a) and outer diameter (b) directions.

where the term $d_m(x)$ has been merged with the constant phase term into the function $G_m(x)$. Using Eq. (3.28), we get for the Fraunhofer diffraction at oblique incidence

$$\tilde{f}(u,v) = \int_{-\infty}^{+\infty} \left\{ \sum_{m=1}^{M} C_m(x) \int_{a_m(x)}^{b_m(x)} \mathrm{e}^{-\mathrm{i}[2\pi v + k_y - (n-1)kc_m(x)]y} \mathrm{d}y \right\} \mathrm{e}^{-\mathrm{i}(2\pi u + k_x)x} \mathrm{d}x. \tag{3.29}$$

This equation is solved similarly to the multilevel version, with

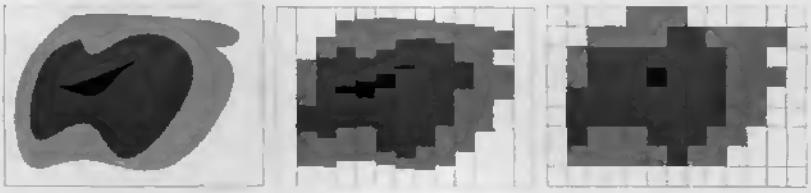
$$K_m = k_y - (n-1)\, kc_m(x), \tag{3.30}$$

for the sake of simplifying the equation. The final result is

$$\tilde{f}(u,v) = \int_{-\infty}^{+\infty} f_v(x) \exp\left[-2\mathrm{i}\pi \left(u + \frac{k_x}{2\pi}\right) x\right] \mathrm{d}x, \tag{3.31}$$

and $f_v(x)$ is defined as

$$f_v(x) = \begin{cases} \sum_{m=1}^{M} \frac{C_m(x)}{\mathrm{i}(2\pi v + K_m)} \left[\mathrm{e}^{-\mathrm{i}(2\pi v + K_m)a_m(x)} - \mathrm{e}^{-\mathrm{i}(2\pi v + K_m)b_m(x)}\right] & v \neq -\frac{K_m}{2\pi} \\ \sum_{m=1}^{M} C_m(x) \cdot [b_m(x) - a_m(x)] & v = -\frac{K_m}{2\pi} \end{cases} \tag{3.32}$$



(a) (b) (c)

Figure 3.11: Comparison of two-dimensional and one-dimensional sampling. The DOE structure (a) is more accurately described by one-dimensional sampling of its borders (b) than by two-dimensional sampling of its values (c).

3.4 Advantages of the hybrid Fourier transform

We shall now review the most interesting advantages offered by this hybrid Fourier transform.

3.4.1 Accuracy of the boundary definition

As we can see from Eq. (3.11), the computation is based on the knowledge of $a_m(x)$ and $b_m(x)$, the lower and upper boundaries of the zones contours. These functions are sampled in the x direction. Contrary to the two-dimensional sampling, the resolution in y direction is not tied to the sampling. As illustrated in Fig. 3.11, we see that a direct consequence is the ability to follow the contour more accurately than by two-dimensional sampling. Finer details are not lost, as illustrated by the small dark areas.

3.4.2 Data requirements

Another direct advantage of the use of Eqs. (3.10) and (3.11) is the fact that the sampling is only one-dimensional. For two-dimensional sampling, the data requirements are evolving as $\mathcal{O}(N_x \times N_y)$, where N_x and N_y are the numbers of sample points in the x and y directions. Most often, the numbers of samples are chosen so that $N_x = N_y = N$. The data evolves then in $\mathcal{O}(N^2)$. For a contour-based one-dimensional sampling, the data requirements evolve in $\mathcal{O}(M \times N)$ instead, where M is the number of shapes. As M is constant for a given optical element, we can increase the resolution in the DOE plane without a dramatic increase of the memory requirements. As an example, the case of two different eight-level hexagonal flat-top beam-shaping elements

for 633 nm has been studied. Both DOEs are 8 mm wide and zero-padding is used, resulting in a total window of 16 mm. The first one generates a light distribution of angular extension $\pm 1°$, the second one of $\pm 5°$.

If sampled with $N \doteq 32 \times 1024$ points in each direction, both require 16 GiB to be stored in Matlab. With the same resolution, the contour-based description requires only 30 MiB for the first ($M = 110$ rings) and 150 MiB for the second ($M = 550$ rings).

Additionally, to perform a Fourier transform, Matlab needs twice the input data memory (three times if the output does not overwrite the input). Since the hybrid transform is sequential, this ratio is applied to the a one-dimensional vector. The use of the proposed contour-based computation represents then a total reduction of the amount of data by a factor 220 for the $\pm 5°$ DOE and more than 1000 for the $\pm 1°$ DOE.

This feature of the hybrid Fourier transform is illustrated in Fig. 3.11. The number of rectangular blocks in Fig. 3.11(b) is smaller than the number of square pixels in Fig. 3.11(c). Increasing the DOE resolution by a factor of two would split vertically every block in two rectangles and divide every square pixel in four.

To increase the speed of computation, it is preferable to store the curves $a_m(x)$ and $b_m(x)$. While this is not mandatory, in practice the gain is appreciable. For the above cited examples this approach has been used. Nevertheless, if similar contours are present in the optical structure, there is a possibility to store them only once. Tiling of the input plane with many analogous structure is done at no additional cost. Sharing contours does not imply sharing all of the zone properties, like complex transmittance C_m. This allows to compute arrangements of lenses like positive-negative random phase plates [56–58], or to simulate variations of the incident beam amplitude on a wide scale with a single DOE cell stored in memory. There are in practice two possibilities to reuse the contours for distinct zones:

1. The difference in position is represented by a multiplicative phase factor, taking advantage of the Fourier transform translation property [12]. This method is however not applicable to the Fresnel propagation, as the complex error function is not linear.

2. The difference in position is taken into account directly in the contour functions. This method does not require more computation time than the first one, and can be applied to both Fraunhofer and Fresnel propagation kernels. This demands that the sampled contours can be superimposed, as illustrated in Fig. 3.12. Although the contours are similar, the gray shape cannot share its boundaries with the two white ones, because gray

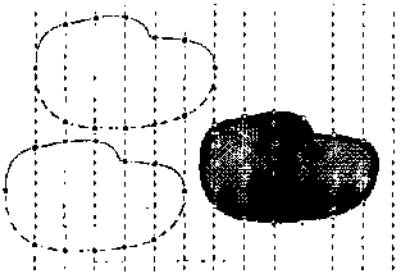


Figure 3.12: Similar elements are translated versions of the contours. This can be expressed as multiplicative phase factors in y and a curve offset in x.

and white dots cannot be superimposed. However, if the sampling step is small enough, this drawback tends to be of minor influence.

3.4.3 Rotational degree of freedom

Another benefit arising from the proposed technique is the possibility to study the element under different angles θ. The element description is done in the (O, X, Y) axes and the computation in the (O, x, y) axes, as seen in Fig. 3.5. For any θ, the contours $a_m(x)$ and $b_m(x)$ have to be recomputed. Consequently, it is usually preferable to choose θ fixed and to vary v to compute more than one line. However, recomputing the contours for a different angle is not different from recomputing the two-dimensional array in the sampled description. The freedom in the choice of the angle θ is just an additional possibility not available with the two-dimensional sampling.

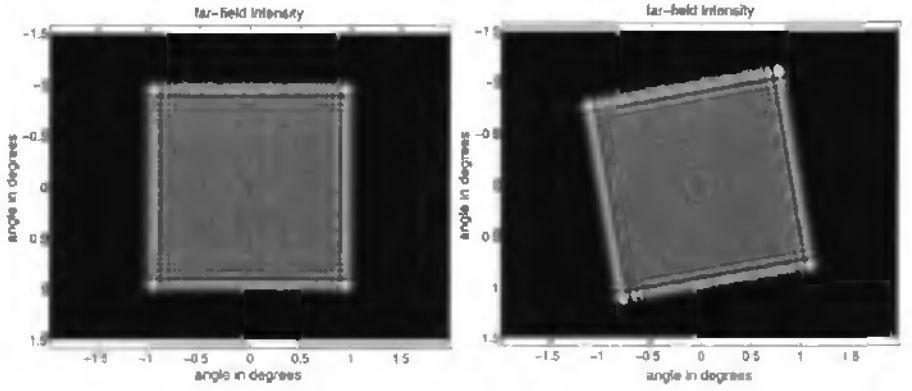
This rotation capacity is illustrated in Fig. 3.13. Figure 3.13(a) represents the far-field pattern of a beam shaping element. The generated far-field is a $\pm 1°$ square flat-top. Figure 3.13(b) is the same pattern rotated by 10 degrees.

The interests of this freedom in the computation are multiple:

1. Characteristic dimensions or profiles in the far-field can be computed exactly in every direction, without using interpolation techniques.

2. Possible artifacts due to the sampling in x direction can be monitored, as the hybrid Fourier transform is continuous, thus artifact-free, in the y direction. Since the artifacts for different rotations are not identical, numerical artifacts can be distinguished from real phenomena.

3. Together with the choice of the output line coordinate v, this rotation freedom allows to limit the computation to areas of special interest in of the output plane.



(a) (b)

Figure 3.13: Far-field intensity of a beam-shaping element (AMD), calculated with $\theta = 0$ (a) and $\theta = 10°$ (b).

This third point is related to the fact that the proposed technique computes only one line of the Fourier transform. Two-dimensional sampling allows to know completely the output plane after one step. While this is a strength, there are many cases where we are not interested in the whole output plane. Often, only some special directions are of interest. In these cases, computing only the desired lines is rather an advantage than a limitation. When an extended area of the far-field is required, we can still compute sequentially enough lines to retrieve the far-field "image".

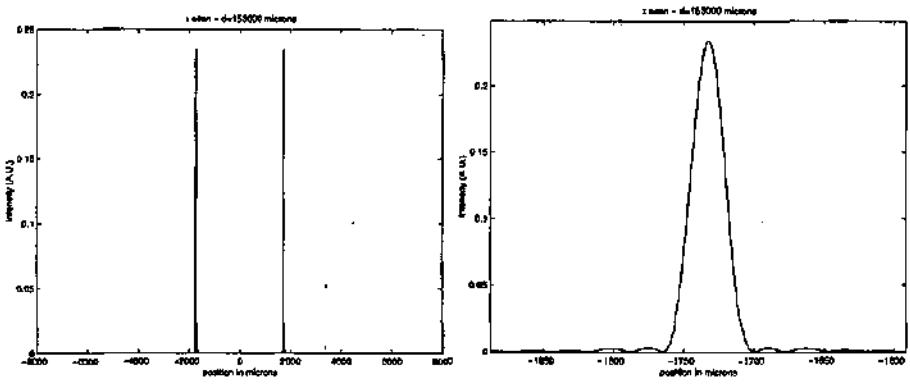
3.4.4 Freedom of the output plane resolution

The resolution in the output plane (i.e. far-field or finite distance) is directly related to the dimension of the signal window used in the input plane. The discrete Fourier transform does not only sample the input, but also the output. It follows from Fourier analysis that the pixel size δu in the far-field is inversely proportional to the extension W of the function in the input plane, namely

$$\delta u = \frac{\lambda}{W} \tag{3.33}$$

for the Fraunhofer diffraction and

$$\delta x' = \frac{\lambda z}{W} \tag{3.34}$$



(a) (b)

Figure 3.14: Scan in the focal plane of two adjacent hexagonal Fresnel zone plates (a). The high resolution due to the large zero-padding becomes evident if the focal area is magnified (b).

for the Fresnel diffraction. Increasing the resolution in the far-field requires to increase the size of the input window (zero-padding) while the resolution remains constant. Since this window is outside the diffracting structure, the contribution to Eq. (3.18) is zero. As a consequence, zero-padding appears only in the one-dimensional DFT. From this observation, we conclude that increasing the resolution of the far-field is virtually at no cost. This was illustrated by the hexagonal lens we presented as an example for the Fresnel diffraction, where we used a complete window of $W = 12.8\,\mathrm{cm}$ sampled with $N = 32 \times 1024$ points.

In the other direction, where the Fourier transform is computed analytically, the notion of signal window has no real sense. The value of W is infinite. Thus, there are no such pixel sizes. The lines may be computed as close as one wishes.

An additional consequence of the possibility to use wide zero-padding is the capability to suppress unwanted foci alias in finite distance diffraction. Indeed, the DFT implies a periodicity of the input. This would result in an infinity of similar foci. This artifact is called aliasing. After a certain distance, the light from the other lenses will overlap with the light of the real one. The only method to reduce this defect is to move those parasitic lenses far away from the original lens. Hence the use of a window of zero values, simulating a black screen.

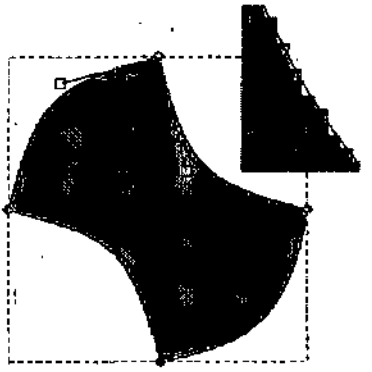


Figure 3.15: A pixelated beam-splitting element cell is fitted by four cubic Bézier curves. The resolution is increased and the resulting contour is smoothed, as shown by the magnification (top right). (Design of the element by V. Kettunen)

To illustrate both the resolution increase and the aliasing suppression, we replaced the hexagonal lens of section 3.3.4 by two hexagonal lenses of outer diameter 4 mm (inner diameter 3.464 mm). Fig. 3.14 shows both the intensity plot in the focal plane on a wide scale (16 mm) and a magnification of one of the two foci. Not only there are no aliases of the two lenses, but also, the resolution is of a few micrometers.

3.4.5 Refinement of pixelated elements

Another advantage of the geometric description is that smooth contours do not exhibit the artifact caused by the pixel extension [59]. As it will be shown in chapter 6, a sinc envelope distorts the light distribution in pixelated structures. The classical procedure to avoid this artifact is to reduce the pixel size [60], however this strategy is limited by the amount of data. On the other hand, the geometric description is free of the presence of pixels, and requires much smaller data. Thus, it follows that one can use smooth contours to improve pixelated structures.

To demonstrate the principle of refinement of existing structures by smooth curves, we used a beam-splitting DOE that, when periodically tiled, generates a 3×3 orders pattern with central order missing. Four cubic Bézier curves (see appendix A) are used to approximate this tessellated element, as shown in Fig. 3.15. For symmetry reasons, the whole contour can be reduced to three independent points (one end point and two control points), representing six scalar parameters. The intensity and amplitude generated by the elementary cell when illuminated by a uniform plane wave are presented in Fig. 3.16.

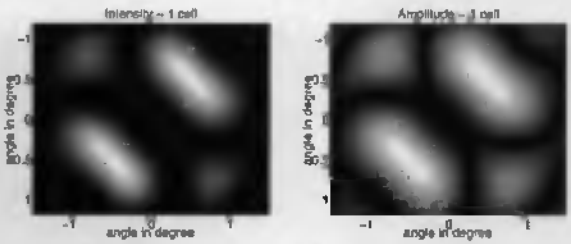


Figure 3.16: Intensity distribution in the far-field of a single cell of the beam-splitting element presented in Fig. 3.15.

One can imagine using this refinement on a computer generated hologram (CGH) after it has been designed by IFTA. Also, as the number of parameters is reduced, an additional optimisation stage could then be applied to the curves, similarly to the two-stage design used by Blair for trapezoidal CGH [61]. Finally, one can also imagine to modify the iterative Fourier transform algorithm so that it can benefit from this additional freedom.

3.4.6 Arrays of optical elements

The small data requirements allow to study larger structures, like arrays of elements. These arrays are not restricted to be rectangular, but can have any geometry. Periodic structures can be simulated by large arrays of similar elements. As an example, the individual cell of the computer generated hologram (CGH) presented in Figs. 3.15 and 3.16 can be tiled as arrays of 2×2, 3×3 and 5×5 cells, as illustrated in Fig. 3.17. The construction of the spot pattern can be seen as cells are added, and the orders are progressively decoupled, as seen in Figs. 3.18 and 3.19. The coupling between adjacent orders is more visible for the modulus of the amplitude than for the intensity.

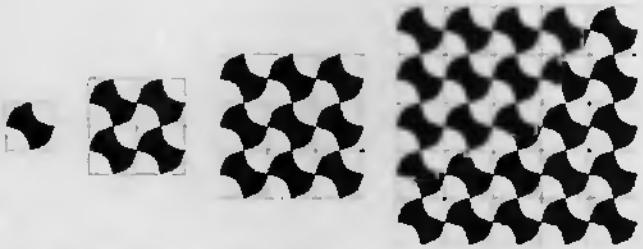
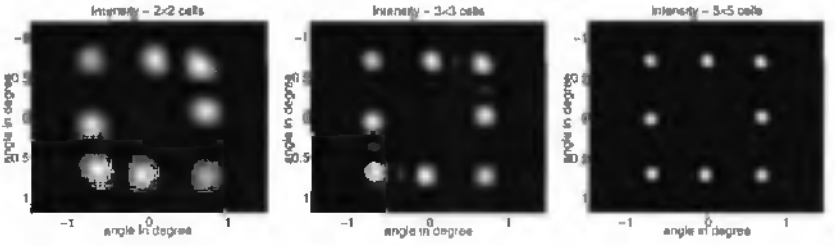
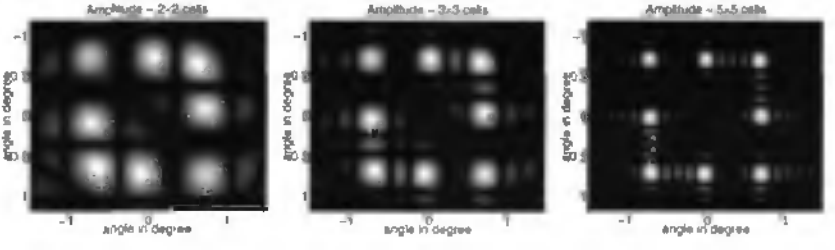


Figure 3.17: The tiling of the individual cells in 1, 2×2, 3×3 and 5×5 arrays.



(a) (b) (c)

Figure 3.18: Intensity distribution in the far-field of a beam-splitting element. The optical element is tiled in 2×2 cells (a), 3×3 cells (b) and 5×5 cells (c).

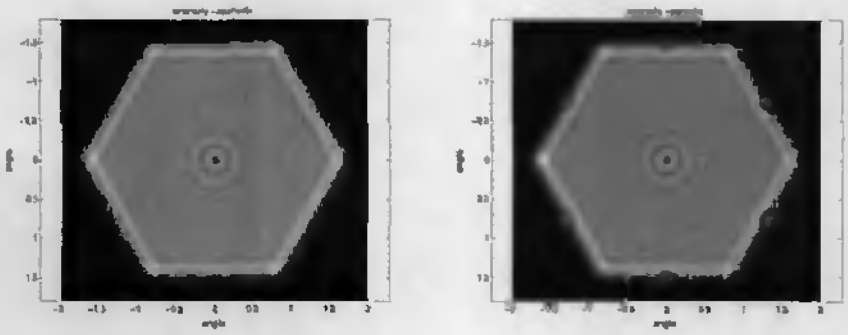


(a) (b) (c)

Figure 3.19: Modulus of the amplitude in the far-field corresponding to Fig. 3.18.

The element of Fig. 3.15 is designed to be effective when tiled as an array. On the other hand, aperture modulated diffusers (AMDs) are designed as individual elements. The consequences of the tiling can be studied. The first expected effect is the presence of a sampling in the output space, as the input becomes quasi-periodic. This *comb* pattern has the symmetry of the array. Another consequence of the arrangement of cells is the appearance of interferences between the neighbours, resulting in oscillations in the far-field distribution. This effect is visible in Fig. 3.20, where we compare the far-field resulting of an individual hexagonal PZP and of an array of 7 such hexagons (1 at the centre surrounded by 6 similar cells). The resulting light distribution was smoothed to suppress the comb effect. Oscillations appear at the middle of the sides of the hexagonal light distribution.

Measurements, as shown in 3.21, have confirmed the presence of such oscillations. It is worth noting that the far-field of the rectangle DOE exhibits



(a) (b)

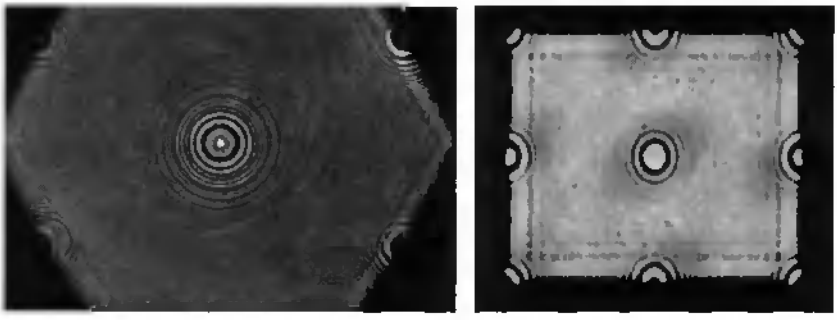
Figure 3.20: Far-field light distribution generated by a single hexagonal cell (a) and a periodic tiling of hexagonal cells (b).

additional oscillations in the corners, contrary to the hexagon. We explain this difference by the fact that the rectangular array has a periodicity of the dimension of one cell in both directions of the sides and the corners, while the hexagonal array has only a periodicity in the direction of the sides.

3.5 Conclusions

We have presented a technique suited to compute the Fresnel and Fraunhofer diffraction formulas for elements whose structures are described vectorially and not pixelated. This technique presents the advantages of the vectorial description, namely small data requirements and high accuracy at any resolution. In addition, some degrees of freedom appear that allow to avoid classical artifacts (aliasing, pixel extension). Finally, the geometric description is well suited to be converted into data for the mask shop, whatever fabrication process chosen (based on polygons, circles, Bézier curves, etc.).

The trade-off of this technique is that the structures have to be decomposed in geometric domains, where the diffraction kernel can be analytically computed along one direction. Luckily, multi-level and linearly blazed structures, which are of common use in optical elements fabrication, fulfil these constraints.



(a) (b)

Figure 3.21: Oscillations on the border of the far-field distribution have been observed in experiments for a hexagonal array illuminated by a plane wave from an He-Ne laser (a) and a rectangular array of Fresnel zone plates (b). The pixels of the sensor act as a low pass filter on the image, smoothing the sampling of the output.

Fabrication technologies 4

In this chapter, we shall introduce concepts about fabrication of refractive and diffractive optical elements. Fabrication technology shall be described, and related errors tackled. An extensive comparison of these technologies may be found in Refs. [62–64].

Micro-optical elements require a very precise control of the geometry on their surface. The high precision technology involved comes mainly from the micro-electronics circuit fabrication. Micro-machining also allows such tolerances. In addition to precision, for most industrial applications the capability to fabricate at reasonable cost a large quantity of elements, in the shortest possible time, is a necessity. This last goal is in contradiction with fabrication oriented towards precision. Thus the general fabrication processes are usually divided into two steps.

Firstly, the precision of the structure is obtained via slow processes to generate a template of the element. This may be either the masks used in photolithography or a master element. The second step is the reproduction of the previously created objects. The low-cost and high-quantity constraints are only applied to this stage. Both steps are sources of fabrication errors.

We shall now review the most commonly used technologies for optical element fabrication, laying the emphasis on their respective errors.

4.1 Mask-making and photolithography

One of the possibilities to fabricate optical elements is photolithography, which transfers an original pattern from a mask into a material as a depth profile.

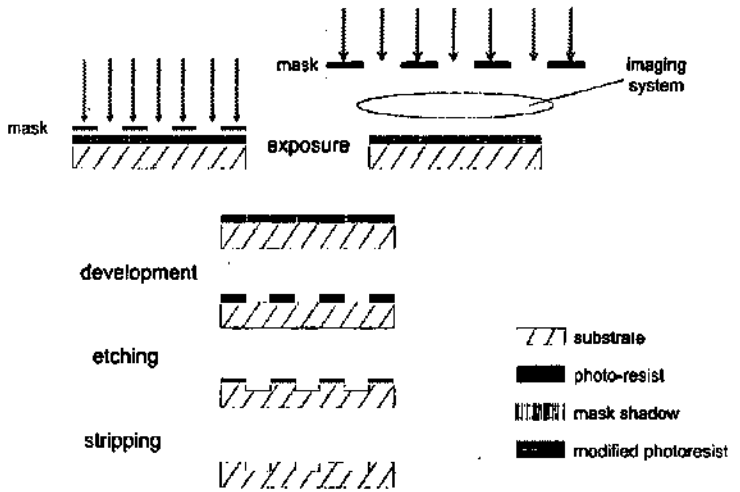
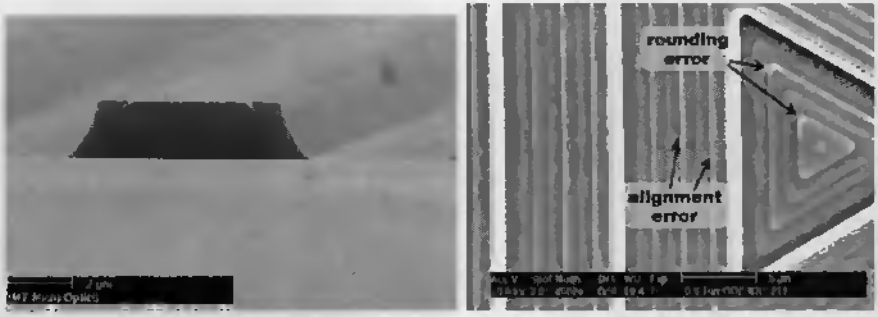


Figure 4.1: Steps of microlithographic fabrication of optical elements. The photoresist is illuminated and developed. The unwanted material is removed, then the profile can be transferred into the substrate.

4.1.1 Photolithography

The optical elements designed and fabricated during this work are mainly based on photolithography, a technology used for very large scale integration (VLSI) manufacturing of microelectronics. The principle of this technique is illustrated in Fig. 4.1. A substrate is covered with a photosensitive material (photoresist). The image of the pattern is then projected onto the layer. Similarly to photography, the result is developed and fixed. Structures in the photoresist are hardened. Depending on the nature of the photoresist, the illuminated areas will be removed (*positive* resist - the one used in our cases) or conserved (*negative* resist). As a result, the desired pattern is now converted into a variation of height, as a binary profile of photoresist.

The exposure of the resist is detailed in the top of Fig. 4.1. Different techniques may be employed to structure the illumination on the photo-sensitive material. The most straightforward one is to place a mask containing the desired pattern in the proximity or in contact of the resist. The last variant permits finer features, but also damages the mask, and is thus avoided (Süss announced recently improvements in this area with their MPT^{TM} technology, though). These two techniques are mainly limited by the diffraction of the light on the features of the mask, that are usually close to the wavelength. The resulting error is a distortion of the profile. This distortion is visible at the transition between the two levels, as can be seen in Fig. 4.2. This error shall



(a) (b)

Figure 4.2: Common fabrication errors are the slope error (a), here on photo-resist before transfer into substrate, and the rounding of the angular features (b). For multilevel profiles, alignment errors also appear, as seen in (b) for a 8-level AMD.

be categorised in line-width error and slope error. Slope error in Fig. 4.2(a) is around 60°, for a feature size of 3 μm and a wavelength between 320 nm and 400 nm. Angular features are rounded due to diffraction by the mask, as seen from Fig. 4.2(b).

These defects are less pronounced for the projection of the pattern by steppers, permitting scaling between the mask and its image on the photoresist layer. This allows to reduce the minimum feature size, or at equal feature size, reduces the influence of the diffraction. However, steppers cannot illuminate a complete wafer in one step, and large areas have to be divided into subfields. This may result in unwanted gaps or overlaps at the borders of the fields. However, this stitching error is only present at these borders, thus is less important than widely spread errors, like profile errors.

After development and exposure, the photoresist profile is transferred into the wafer substrate, most often by reactive ion etching (RIE) or ion beam milling (IBM). In the case of RIE, depending on the composition of the ionised gas, the speed of the etching is different for the photoresist and for the substrate. If the substrate is etching faster than the photoresist, the profile can be reproduced scaled in depth, allowing for a first correction of the slope error. While RIE is mainly an anisotropic process, there exist an isotropic reaction. This reaction can be additionally used to attack the sidewall laterally and compensate for the profile errors. After stripping the residual photo-resist, the optical element is completely transfered into the substrate. The main error

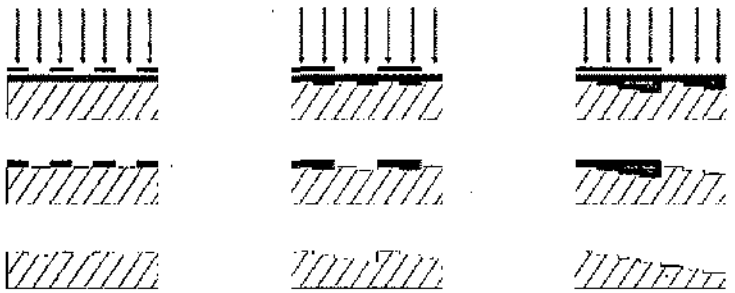


Figure 4.3: For multilevel profiles, several masks are used in successive lithographic steps. Each of the lithographic steps introduces an alignment error, as seen in Fig. 4.2(b).

produced by this step is the etching depth error. This error can be typically reduced to 5% or less.

For elements that need more than two levels of height, after a careful realignment of the substrate, the process is repeated, as show in Fig. 4.3. By successive writing-etching steps, a multilevel profile is realised. The alignment, a critical issue, is achieved using corresponding marks, usually crosses, on the mask and the substrate. Gruber *et al.* [65] proposed an improvement, where both sides of the wafer are used. Alignment errors, illustrated in Fig. 4.2(b), can be reduced by a factor two with this solution, but some areas of the back side of the wafer have to be patterned with lenses, perturbing locally the function of the optical element.

4.1.2 Mask patterning

The previously described step is the high-quantity low-cost process. It only copies the pattern carried by the mask into the optical material. The precise DOE description is mainly contained in the mask. Commonly e-beam lithography is employed to pattern chromium masks. This technology uses a beam of electrons to structure the resist deposited on the chromium. By processes similar to lithography, the pattern is transfered into the chromium. The advantage of electrons is their small wavelength, allowing to write very small features (a few tens of nm). The pattern is realised locally by the electron beam, while an X-Y stage scans the whole element. This technology has nevertheless a drawback. The final pattern is not smooth, because the movement of the stage is Cartesian. Curves and circles are approximated by polygons, trapezes or rectangles. This results in edge roughness. While it is possible to design optical elements based on the trapezoidal nature of the pattern [61,66–68], it is also possible to approximate the smooth curves by a large

number of lines [47]. On the other hand, some electron-beam writing machines (like Raith ELPHY plus and LION-LV1) provide a feature called path control or continuous path control that drives the X-Y stage on the whole mask along curves. The precision is obtained by a control loop that drives the electron beam to correct for the X-Y stage errors [69,70]. The width of the lines can be controlled by varying the size of the spot. This is achieved by defocusing the electron beam. This smooth control of the beam as been used to fabricate very accurate axicons [71]. The proximity effects due to the interaction of electrons with the resist are reduced by using on low energy electrons.

Laser beam writing may also be used to pattern chromium masks. The technique is similar to electron beam writing, but the optical wavelength limits the resolution. As in electron beam writing, X-Y based pattern generators, causing edge roughness of the curves, can be replaced by polar coordinate laser plotters [72, 73].

For further improvement and flexibility, the binary masks can be replaced by gray-tone masks. No multiple alignment and exposure is needed any more, because all the information of the profile can be stored on one mask and transferred in one step. Unfortunately, gray-tone mask technology is not completely controlled yet, and the profiles may be distorted. It will doubtlessly be a very effective technology as soon as these problems are solved.

4.2 Direct writing by e-beams and lasers

In the past decade, direct writing of optical components has been developed with the high quality demanded by optics. Instead of using the previously described e-beam or laser writer to structure a mask, they are directly used to structure the photoresist on the substrate with the desired profile. This requires a precise control of the laser or electron flux and a development process adapted to remove photoresist proportionally to the exposure dose [69, 74]. For the laser case, due to the limited spot size, profiles with high spatial frequencies look blurred. However, corrections for the limited spot size can be partly incorporated in the design algorithm. Additionally, the rounding of the sharp edges affects the the overall phase distribution balance, responsible of the zero order. This phenomenon can be compensated by scaling the depth of the structure [32]. Moreover, to achieve small features, the spot size needs to be reduced, which implies a smaller depth of focus. Consequently, fine features can only be realised for thin elements.

An interesting bonus of the direct writing processes is the ability to generate continuous height profiles, whereas mask photolithography, with the exception

of gray-tone masks, can only produce multilevel profiles. The efficiency is increased, and scattering can be reduced.

The main weakness of direct writing is the time needed to structure a single element, and it is thus not realistic to write high-quantity low-cost elements this way. The solution is to use one element as a model for others, that are produced by replication technologies, such as injection moulding, UV-casting or embossing [63]. These techniques are effective at very low cost, but they introduce slight distortions in the profile depth and other defects which can cause scattering. Also, replication into hard materials, such as quartz and calcium fluoride, is not possible. A possible technique to structure such materials could then be imprint lithography [75], where a layer of resist is deposited on top of the material, embossed and then etched as in section 4.1.1.

4.3 Resist melting technology

For the fabrication of refractive micro-lenses, the technique of resist melting has been developed [76]. Cylinders of photoresist realised by photolithography are heated until melting, as shown in Fig. 4.4(a). The surface tension forces form a spherical or cylindrical lens that can then be transferred by etching into the substrate [77, 78]. Additionally, if the selectivity of the etching is chosen with care, the profile of the etched element can be different from the original photoresist lens. This allows to fabricate aspherical lenses or to scale the vertical dimension of the lens [79]. Resist melting requires only one lithographic step, and is thus insensitive to errors such as mask mis-alignment or slope error. Moreover, refractive elements are highly efficient which makes them interesting candidates for beam-shaping.

Despite these many qualities, the variety of available profiles and the fill factor of arrays restricts this technique to micro-lenses or to one-dimensional beam-shaping elements, like the one of Fig. 4.4(b). As slope error tends to reduce the free space between two adjacent lenses, there is an upper limit for the fill factor. This limit tends to decrease with the thickness of the photoresist layer. A workaround consists of fabricating a thin pedestal with a high fill factor, on which a thick layer of resist with a lower fill factor is deposited. After melting, the refractive optical element (ROE) has the fill factor of the thin layer, but the height of the thick layer. The cylindrical lens of Fig. 4.4(b) was manufactured by this technique.

Refractive elements manufactured by resist-melting process are mainly sensitive to profile errors. Etching depth error results in a change of focal length or equivalently, in the far-field, in a change of the aperture angle [80, 81].

(a) (b)

Figure 4.4: The melting resist process (a): deposition and development (top), melting to achieve a nearly spherical (or cylindrical) photoresist lens (middle), final transfer into the substrate (bottom). Example of a cylindrical lens with a width of 300 μm and a height of 30 μm (b).

4.4 Conclusions

We have presented succinctly some technologies used in fabrication of refractive and diffractive elements for beam shaping. The constraints of low-cost and high quality have led to two-step techniques, based on the replication of a precise template, be it a mask or a master element.

We aimed at presenting the main constraints resulting from the processes, in terms of features and design freedom. In chapter 5, we shall present different design methods suited to create elements to be realised by these fabrication technologies. In chapter 6, we will finally review some consequences of the errors mentioned here, and try to present possible corrections schemes.

Design of optical beam-shaping elements 5

In the present chapter, we will tackle the design of micro-optical beam-shaping elements that will be fabricated using the techniques described in chapter 4. Mainly two techniques are available for the design of diffractive and refractive optical beam-shaping elements: The first one, re-mapping, is based on geometrical optics, while the second one, more evolved, is based on wave optics.

5.1 Re-mapping type elements

We shall firstly present the re-mapping type elements and some analytical and numerical methods used for their design.

5.1.1 Principle

The principle of re-mapping is illustrated in Fig. 5.1. This design is based on a map transformation between two light distributions. The physical implementation of the point to point transformation is performed via the law of grating diffraction or Snell's law of refraction.

The theory of map transformation was mathematically described by Bryngdahl [82]. He computed filters that realised coordinate transformations, such as converting a rectangular line pattern into a circular line pattern. Frère, Leseberg, Jaroszewicz and others later proposed methods to focus an input light distribution into an arbitrary line segment [83–86]. By combination of such elements, three-dimensional patterns were produced [87–90]. The emphasis was put on the shape of the pattern and not on the intensity distribution.

In beam shaping, the output intensity distribution is a requirement. Using power conservation, intensity redistribution was then taken into account.

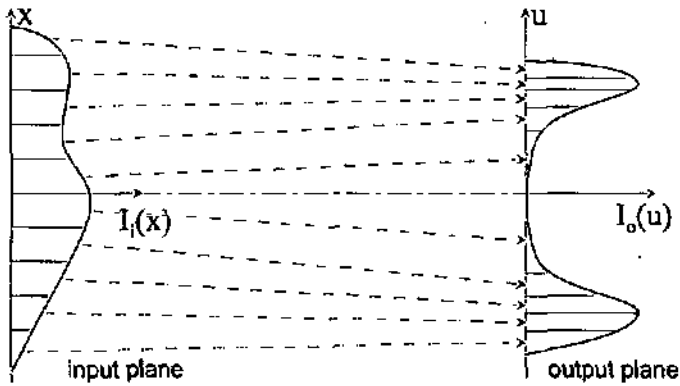


Figure 5.1: The principle of re-mapping: The input is transformed into the output by means of the laws of geometrical optics. The energy is redistributed in order to produce the desired profile.

One of the first beam-shaping device using map-transform techniques was introduced by Han [91]. It converts a Gaussian beam into a uniform light distribution by means of two consecutive elements. Later, axicons [92–94] and beam-shaping elements for lasers [51, 95–98] were designed using a single element. All these solutions are either rotationally symmetrical, separable in Cartesian coordinates or one-dimensional.

We shall first present the one-dimensional analytical solutions for common beam shaping problems and their extension to separable elements. The problem of two-dimensional problems with symmetry of rotation will be tackled afterwards. We will propose, to our knowledge, new formulas for flat-top and Gaussian distributions, and outline some specificities of the rotationally symmetrical case in terms of degrees of freedom. Finally, we will describe mesh based inverse ray-tracing as proposed by Dresel [6] and improved by Hermerschmidt [99–101]. This method allows beam shaping of non-symmetric or non-separable input and output. An improvement, resulting in smoother meshes, will be proposed.

We will restrict our demonstrations to paraxial far-field optics, but all these design techniques are equally applicable to finite distances.

5.1.2 Map transform and energy redistribution

Based on the ray equation (2.48), the relation between the phase $\varphi(x, y)$ of the field emerging from the optical element and the local direction of the rays

are given by

$$\frac{\partial \varphi(x,y)}{\partial x} = \frac{2\pi}{\lambda}\sin\alpha \tag{5.1}$$

$$\frac{\partial \varphi(x,y)}{\partial y} = \frac{2\pi}{\lambda}\sin\beta \tag{5.2}$$

where α and β are the angles between the optical axis Oz and the x and y components of the ray vector, respectively. Using the convention presented in Fig. 3.5, we can relate the angles α and β to the output plane coordinates. In the far-field, where the coordinates are the spatial frequencies $\mathbf{u}(u,v)$,

$$\sin\alpha = \lambda u \tag{5.3}$$

$$\sin\beta = \lambda v \tag{5.4}$$

and at a finite distance z, where the output coordinates are the positions $\mathbf{x}'(x',y')$,

$$\sin\alpha = \frac{x-x'}{z} \tag{5.5}$$

$$\sin\beta = \frac{y-y'}{z}. \tag{5.6}$$

It is worth noting that the same conclusions can be drawn from the application of the stationary phase method to the wave optics equations, such as the Fraunhofer equation (2.33). Additionally, for a two-dimensional design problem, the existence of an exact solution $\varphi(x,y)$ is conditioned to

$$\frac{\partial^2 \varphi}{\partial x \partial y} = \frac{\partial^2 \varphi}{\partial y \partial x}. \tag{5.7}$$

Now that the map-transformation relations are known, we shall introduce the light distributions $I_i(x,y)$ and $I_o(u,v)$ in the input and in the output planes and use the energy (or power) conservation law

$$\mathcal{E} = \iint_{-\infty}^{+\infty} I_i(x,y)\,\mathrm{d}x\mathrm{d}y = \iint_{-\infty}^{+\infty} I_o(u,v)\,\mathrm{d}u\mathrm{d}v \tag{5.8}$$

to determine the phase function $\varphi(x,y)$. Note that the light distributions I_i and I_o are densities per unity surface of their respective space. I_i is expressed by spatial position (x,y), while I_o is given as a function of the spatial frequencies (u,v).

Finally, we shall search for a continuous solution with a continuous derivative. We will see later that diffraction at the borders of the cells results in oscillations in the far-field. These cells are defined by the area where the phase (or the profile) and its derivative (or the profiles' slope) are continuous. Breaking the slope or the phase results in strong oscillations in the far-field distribution. Moreover, a monotonic profile, i.e. a derivative without change of sign, is also preferable, as this condition allows to reduce the number of solutions to a finite number. Also, one can expect that this last condition will result in a smoothly varying profile, avoiding high frequencies, which is definitely an advantage in terms of fabrication tolerances. This assumption allows to write Eq. (5.8) in a differential form

$$I_i(x,y)\,\mathrm{d}x\mathrm{d}y = I_o(u,v)\,\mathrm{d}u\mathrm{d}v. \tag{5.9}$$

When Eq. (5.9) fails to provide a solution to the problem, one should use Eq. (5.8) instead.

5.1.3 One-dimensional analytical solutions

One-dimensional analytical solutions are of interest for one-dimensional beam shaping, but also for separable beam shaping functions. The map-transformation of Eqs. (5.1) to (5.4) are simplified to

$$\frac{\mathrm{d}\varphi(x)}{\mathrm{d}x} = 2\pi u, \tag{5.10}$$

and the differential energy conservation (5.9) to

$$I_i(x)\,\mathrm{d}x = I_o(u)\,\mathrm{d}u. \tag{5.11}$$

Uniform to flat-top beam-shaping element

The simple case of the flat-top beam shaping element in one dimension is known to have a parabolic phase [51]. However, we shall demonstrate this property for a general case. We assume that the incident light distribution is a uniform plane wave $I_i(x) = C_1$, and the element is delimited by x_1 and x_2. The desired far-field is flat, which means $I_o(u) = C_2$ in the angular range from α_1 to α_2. These angles are related to the spatial frequencies u_1 and u_2 by Eq. (5.3). Differentiating Eq. (5.10) with respect to x, we get

$$\frac{\mathrm{d}^2\varphi}{\mathrm{d}x^2} = 2\pi\frac{\mathrm{d}u}{\mathrm{d}x}. \tag{5.12}$$

where $\mathrm{d}u/\mathrm{d}x$ can be deduced from Eq. (5.11), leading to

$$\frac{\mathrm{d}^2\varphi}{\mathrm{d}x^2} = 2\pi\frac{C_1}{C_2}, \tag{5.13}$$

which defines a parabola

$$\varphi(x) = \pi\frac{C_1}{C_2}x^2 + bx + c. \tag{5.14}$$

$a = \pi\frac{C_1}{C_2}$ is homogeneous to the inverse of a squared distance, and c is a constant named *phase offset* [97]. In practise, the value of c is free and the values of a and b are determined with the help of Eq. (5.10). If we decide arbitrarily that the light at the frequency u_i is coming from the position x_i, which is mathematically expressed as

$$\left.\frac{\mathrm{d}\varphi}{\mathrm{d}x}\right|_{x_i} = 2\pi u_i, \tag{5.15}$$

then we obtain

$$\begin{aligned} a &= \pi\frac{u_1-u_2}{x_1-x_2} \\ b &= 2\pi\frac{u_2x_1-u_1x_2}{x_1-x_2} \end{aligned} \tag{5.16}$$

If we had chosen the opposite situation, where x_1 is mapped onto u_2 and vice-versa, we would have obtained

$$\begin{aligned} a &= -\pi\frac{u_1-u_2}{x_1-x_2} \\ b &= 2\pi\frac{u_1x_1-u_2x_2}{x_1-x_2} \end{aligned} \tag{5.17}$$

While the phase offset c has no influence at this stage, but may have when the encoding technique is taken into account, we see from Eqs. (5.16) that a relates to the overall energy conservation, and b is an asymmetry term. Indeed, if the situation is symmetrical, $x_1 = -x_2$ and $u_1 = -u_2$, which implies that $b = 0$. This term may be used to displace the distribution off-axis or to compensate for an oblique incidence.

Gaussian to flat-top beam-shaping element

Another classical situation is the transformation of a Gaussian beam into a uniform flat-top distribution [91, 95, 96, 98]. We have $I_i(x) = C_1 \exp\left(-\frac{x^2}{\sigma^2}\right)$

and $I_o(u) = C_2$. As in the previous case, by differentiating Eq. (5.10) and using the differential energy relation (5.11), we obtain

$$\frac{\mathrm{d}^2\varphi}{\mathrm{d}x^2}\mathrm{d}x = 2\pi\frac{C_1}{C_2}\exp\left(-\frac{x^2}{\sigma^2}\right), \tag{5.18}$$

that in turn, integrated between 0 and x, leads to

$$\frac{\mathrm{d}\varphi}{\mathrm{d}x} = 2\pi\frac{C_1}{C_2}\int_0^x \exp\left(-\frac{x^2}{\sigma^2}\right)\mathrm{d}x = \pi^{\frac{3}{2}}\frac{C_1}{C_2}\sigma\mathrm{erf}\left(\frac{x}{\sigma}\right). \tag{5.19}$$

which is similar to Eq. (11) of Ref. [95]. The integral of the error function is given by [102]

$$\int \mathrm{erf}(x)\,\mathrm{d}x = x\cdot\mathrm{erf}(x) + \frac{\exp\left(-x^2\right)}{\sqrt{\pi}}. \tag{5.20}$$

The phase of the optical structure is then defined by

$$\varphi(x) = \sqrt{2\pi}\frac{C_1}{C_2}\left[\sqrt{\pi}\sigma^2\exp\left(-\frac{x^2}{\sigma^2}\right) + \pi\sigma x\cdot\mathrm{erf}\left(\frac{x}{\sigma}\right)\right] + b, \tag{5.21}$$

which in turn, is similar to Eq. (14) of Ref. [98] or Eq. (18) of Ref. [96]. Here again, the phase offset b is a free parameter.

Two-dimensional beam shaping separable in Cartesian coordinates

One-dimensional solutions can be used to find solutions for situations which are separable in Cartesian coordinates. In this case, the input and the desired light distributions can be written as

$$\begin{aligned} I_i(x,y) &= i_x(x)\cdot i_y(y) \\ I_o(u,v) &= o_u(u)\cdot o_v(v) \end{aligned} \tag{5.22}$$

Separable distributions are generally encountered with Gaussian beams, or super-Gaussian beams, whose intensities are defined by

$$I_{SG}(x,y) = \exp\left[-\left(\frac{x^2}{\sigma_x^2}\right)^m\right]\cdot\exp\left[-\left(\frac{y^2}{\sigma_y^2}\right)^n\right]. \tag{5.23}$$

An analogous equation can be written in the output plane. As the rectangular flat-top is also separable, beam shaping from a Gaussian [95] or super-Gaussian [98] beam to a rectangular flat-top distribution is easily obtained. Indeed, both directions are solved separately, and the result is the product of the two solutions.

5.1.4 Rotation symmetrical analytical solutions

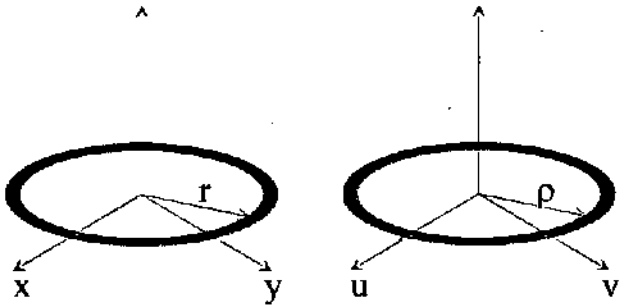


Figure 5.2: Rotationally symmetrical beam shaping. $r = \sqrt{x^2 + y^2}$ in the input plane and $\rho = \sqrt{u^2 + v^2}$ in the output plane.

To complement separable and one-dimensional formulas, we shall study the beam shaping in two dimensions of distributions with a symmetry of rotation (also called isotropic). The polar coordinates for the input and output are shown in Fig. 5.2, and the map transform is then

$$\frac{\mathrm{d}\varphi(r)}{\mathrm{d}r} = 2\pi\rho, \tag{5.24}$$

while the energy conservation can be written in the differential form

$$I_i(r) \cdot 2\pi r \cdot \mathrm{d}r = I_o(\rho) \cdot 2\pi\rho \cdot \mathrm{d}\rho. \tag{5.25}$$

We shall tackle now two most important beam-shaping element designs, the uniform to flat-top and the uniform to Gaussian distribution.

Uniform to flat-top beam shaping

The input and output light distributions are $I_i(r) = C_1$ and $I_o(\rho) = C_2$. Deriving Eq. (5.24), we get

$$\mathrm{d}\rho = \frac{1}{2\pi}\frac{\mathrm{d}^2\varphi}{\mathrm{d}r^2}\mathrm{d}r. \tag{5.26}$$

Introduced in Eq. (5.25) this leads to

$$C_1 r = \frac{C_2}{4\pi^2} \cdot \frac{\mathrm{d}\varphi}{\mathrm{d}r} \cdot \frac{\mathrm{d}^2\varphi}{\mathrm{d}r^2} = \frac{C_2}{4\pi^2} \cdot \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}r}\left[\left(\frac{\mathrm{d}\varphi}{\mathrm{d}r}\right)^2\right]. \tag{5.27}$$

By integration, we find the equation describing the square of the derivative of the phase

$$\left(\frac{\mathrm{d}\varphi}{\mathrm{d}r}\right)^2 = ar^2 + b, \tag{5.28}$$

with $a = 4\pi^2 \frac{C_1}{C_2}$ and a constant of integration b, which is equivalent to

$$\frac{\mathrm{d}\varphi}{\mathrm{d}r} = \pm\sqrt{ar^2 + b}. \tag{5.29}$$

The two solutions with opposite signs for the square root correspond to a positive and a negative lens. We will only study one of them, the other being straightforward to deduce. While a is related to the energy value, thus the size of the cell and the aperture angle, the meaning b is obviously found as

$$b = \left(\left.\frac{\mathrm{d}\varphi}{\mathrm{d}r}\right|_{r=0}\right)^2. \tag{5.30}$$

Firstly, if $b = 0$, the slope at the centre of the lens is zero, we get

$$\varphi(r) = \frac{\sqrt{a}}{2}r^2 + c, \tag{5.31}$$

which is the equation of a parabola. The main difference with respect to the parabola of Eq. (5.14) is the absence of the linear term. We noticed that this term was due to asymmetry around $x = 0$. In the rotationally symmetrical situation, this asymmetry cannot exist. The constant c is the phase offset, and does not influence the shape of the lens. Finally, a is determined by the external radius of the lens R and the desired far-field angle α, with Eqs. (5.3) and (5.24), as

$$a = \left(\frac{2\pi \sin\alpha}{\lambda R}\right)^2 - \frac{b}{R^2}, \tag{5.32}$$

and with $b = 0$

$$, \quad a = \left(\frac{2\pi \sin\alpha}{\lambda R}\right)^2. \tag{5.33}$$

If $b \neq 0$ but $a = 0$, the solution is of the form

$$\varphi(r) = \sqrt{b}r + c, \tag{5.34}$$

which is the equation of a cone. The far-field pattern is a thin ring, whose angle is defined by the value of b. This element is known as an axicon [93].

Unfortunately, the light distribution on the optical axis is not constant, but grows linearly as the surface of the rings, thus the energy in the rings, grows from the centre to the edge of the lens. One should then prefer the *generalised axicon* or *logarithmic axicon* that compensate for this drawback [71,92,94,103].

For more general, non degenerated situations where $a \neq 0$ and $b \neq 0$, we shall define two distinct radii R_1 and R_2, and two angles α_1 and α_2 (which may be equal) so that the frequencies at R_i are associated to the angles α_i. The values of coefficients a and b are then obtained with the help of Eqs. (5.24) and (5.28) as

$$\begin{aligned} a &= \left(\tfrac{2\pi}{\lambda}\right)^2 \tfrac{\sin^2\alpha_1 - \sin^2\alpha_2}{R_1^2 - R_2^2} \\ b &= \left(\tfrac{2\pi}{\lambda}\right)^2 \sin^2\alpha_i - aR_i^2 \end{aligned} \tag{5.35}$$

If $b = 0$, then Eq. (5.33) holds, with both sets of radii and angles. The solution of Eq. (5.29) depends on the signs of the coefficients inside the square root.

For $a > 0$ and $b > 0$, the resulting phase is

$$\varphi(r) = \frac{b}{2\sqrt{a}}\left[\sqrt{\frac{a}{b}}\cdot r\cdot\sqrt{1+\frac{a}{b}\cdot r^2} + \operatorname{asinh}\left(\sqrt{\frac{a}{b}}\cdot r\right)\right] + c, \tag{5.36}$$

while for $a > 0$ and $b < 0$, the profile is given by

$$\varphi(r) = \frac{|b|}{2\sqrt{a}}\left[\sqrt{\frac{a}{|b|}}\cdot r\cdot\sqrt{-1+\frac{a}{|b|}\cdot r^2} - \operatorname{acosh}\left(\sqrt{\frac{a}{|b|}}\cdot r\right)\right] + c. \tag{5.37}$$

Finally, if $a < 0$ and $b > 0$, the phase is

$$\varphi(r) = \frac{b}{2\sqrt{|a|}}\left[\sqrt{\frac{|a|}{b}}\cdot r\cdot\sqrt{1-\frac{|a|}{b}\cdot r^2} + \operatorname{asin}\left(\sqrt{\frac{|a|}{b}}\cdot r\right)\right] + c. \tag{5.38}$$

The case $a < 0$ and $b < 0$ is rejected, because it is in contradiction with Eqs. (5.35). Indeed, they would imply that

$$0 \leqslant \left(\frac{2\pi}{\lambda}\right)^2 \sin^2\alpha_i < aR_i^2 < 0. \tag{5.39}$$

The signification of the three possible solutions is explained in Fig. 5.3. The two first ones are extensions of the classical parabola to cases where either the far-field starts at $\alpha > 0$ (ring beam-shaping element), or where the lens itself starts at $R > 0$ (ring lens). The last one is an inverted beam shaping, that can be used for ring or disc beam-shaping element, proposing an alternative

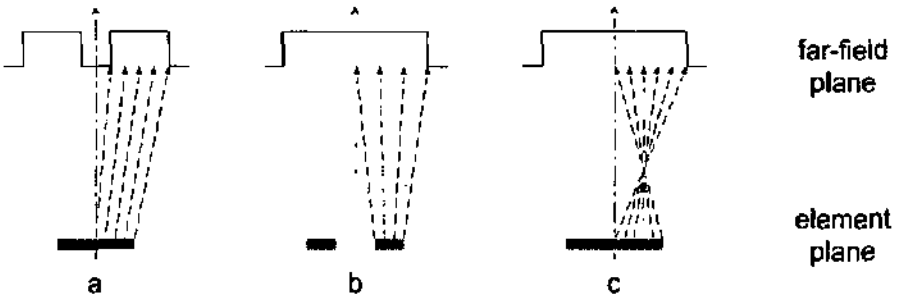


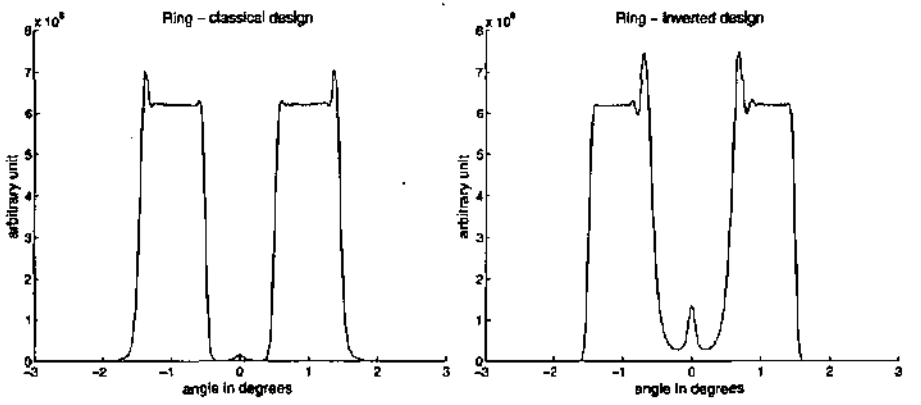
Figure 5.3: Illustration of the three general flat-top beam shaping solutions: (a) with Eq. (5.36), (b) with Eq. (5.37) and (c) with Eq. (5.38).

to Eq. (5.36). All three equations might be reformulated equivalently with logarithms instead of trigonometric and exponential functions. Eq. (5.37) is then similar, but not equivalent, to Eq. (5) of Ref. [104], where a set of two DOEs performing annular-to-circular transformation is presented.

Examples of the classical ring of Eq. (5.36) and the inverted ring of Eq. (5.38) are shown in Fig. 5.4. The DOE was designed as a 16-level lens of 2 mm, shaping the beam into a ring between 0.5° and 1.5°. In practise, the input illumination is an incoherent Gaussian beam of wavelength 248 nm and of divergence 1 mrad. To simulate the effect of such an illumination, the intensity resulting from a plane wave illumination is convolved by the Gaussian beam intensity in the far-field. The high frequency oscillations, such as those caused by the quantisation of the phase, are thus filtered out. The main difference between Fig. 5.4(a) and 5.4(b) lies in the location of the diffraction oscillations. They are located at the angles associated to the frequencies present at the border of the lens. Additionally, the noise in the centre area is much higher for the inverted design. The reason for this noise and the possibilities to reduce it by appropriate design will be discussed in chapter 6.

Uniform to Gaussian beam-shaping

We will study here the analogue to section 5.1.3, and restrict ourselves to cases where the lens will be defined from $r = 0$ to $r = R$. Similarly, the far-field will be a Gaussian profile defined between $\alpha_i = 0$ and α_o. Gaussian rings will thus not be tackled in this section. The light distributions are $I_i(r) = C_1$ and $I_o(\rho) = C_2 \exp\left(-\frac{\rho^2}{\sigma^2}\right)$. Hence, with Eq. (5.24) we can write the energy



(a) (b)

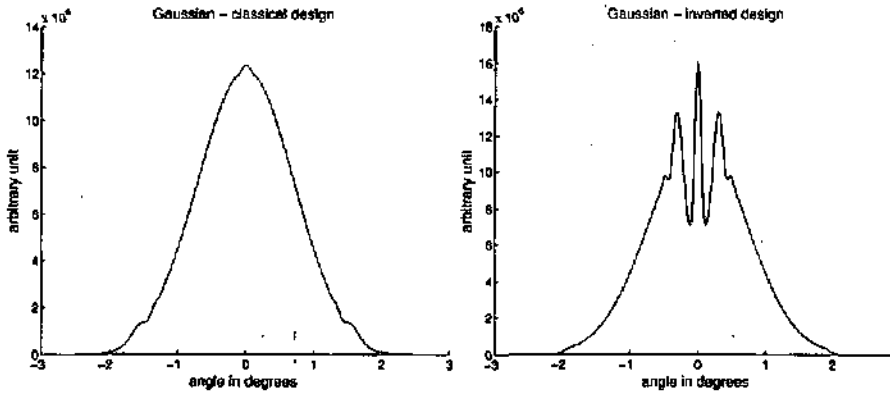
Figure 5.4: Comparison of the far-field of a ring beam-shaping element designed by Eq. (5.36) (a) and Eq. (5.38) (b).

conservation as

$$\begin{aligned} C_1 r &= \frac{C_2}{4\pi^2} \exp\left[-\frac{\left(\frac{\mathrm{d}\varphi}{\mathrm{d}r}\right)^2}{4\pi^2\sigma^2}\right] \cdot \frac{\mathrm{d}\varphi}{\mathrm{d}r} \cdot \frac{\mathrm{d}^2\varphi}{\mathrm{d}r^2} \\ &= \frac{C_2}{4\pi^2} \exp\left[-\frac{\left(\frac{\mathrm{d}\varphi}{\mathrm{d}r}\right)^2}{4\pi^2\sigma^2}\right] \cdot \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}r}\left[\left(\frac{\mathrm{d}\varphi}{\mathrm{d}r}\right)^2\right] \\ &= -\frac{C_2}{2}\sigma^2 \frac{\mathrm{d}}{\mathrm{d}r}\left\{\exp\left[-\frac{\left(\frac{\mathrm{d}\varphi}{\mathrm{d}r}\right)^2}{4\pi^2\sigma^2}\right]\right\}. \end{aligned} \tag{5.40}$$

This equation, contrary to the similar one-dimensional case, is straightforwardly integrated into

$$\left(\frac{\mathrm{d}\varphi}{\mathrm{d}r}\right)^2 = -4\pi^2\sigma^2 \ln\left(ar^2 + b\right), \tag{5.41}$$



(a) (b)

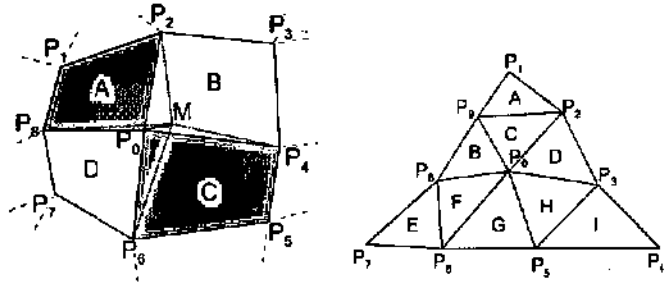
Figure 5.5: The two solutions to generate a Gaussian beam from a uniform illumination. The classical branch (a) and the inverted one (b).

where the coefficients a and b are

$$\begin{aligned} a &= \frac{1}{R^2}\left\{\exp\left[-\frac{1}{\sigma^2}\left(\frac{\mathrm{d}\varphi}{\mathrm{d}r}\Big|_R\right)^2\right] - \exp\left[-\frac{1}{\sigma^2}\left(\frac{\mathrm{d}\varphi}{\mathrm{d}r}\Big|_0\right)^2\right]\right\} \\ b &= \exp\left[-\frac{1}{\sigma^2}\left(\frac{\mathrm{d}\varphi}{\mathrm{d}r}\Big|_R\right)^2\right] \end{aligned} \quad . \tag{5.42}$$

Equation (5.41) defines both the positive and the negative lenses, as previously. Depending on the values of the derivative of the phase at $r = 0$ and $r = R$, two solutions for a exist for Eq. (5.42), as illustrated by Fig. 5.5. The classical solution maps the centre of the lens to the centre of the far-field and the edge of the lens to the border of the Gaussian, and the inverted solution maps the centre of the lens to the border of the far-field and the edge of the lens to the centre of the far-field.

The two profiles of Fig. 5.5 are the results of elements designed to generate a Gaussian distribution of width $\sigma = 1°$. Moreover, the Gaussian profile is truncated at 2°. As for the ring beam-shaping element, the illumination is an incoherent Gaussian beam of divergence 1 mrad and wavelength 248 nm. The oscillations due to diffraction have a stronger influence in the centre of Fig. 5.5(b) than at the border of Fig. 5.5(a). Indeed, their intensity is spread over a wider area in the case of Fig. 5.5(a) (the edge of the Gaussian) than for Fig. 5.5(b) (the centre of the Gaussian).



(a) (b)

Figure 5.6: Mesh adaptation. A mesh is defined by nodes P_i. The node P_0 is moved towards its ideal position M that equalises the areas of the patches. Two topologies are available: rectangular with eight neighbouring nodes (a) and triangular, with nine neighbours (b).

5.1.5 General solutions with mesh adaptation

All the solutions presented in the previous sections were making use of some symmetry. Unfortunately, only a few geometries allow for an analytical solution of Eqs. (5.3), (5.4) and (5.8). In most cases, if a solution exists, it cannot be found analytically. And even more frequently, one has to accept approximated solutions. To derive such approximations, Dresel at al. have proposed an algorithm based on the meshing of both the DOE and the target [6]. The meshes are then adapted in both planes to have patches of equal areas, and the gradient of the phase is retrieved from the coordinates of the nodes of the two meshes. A final step intends to find a phase that matches this gradient. We will describe this algorithm in detail, and see that ways to improve it can be proposed.

Mesh adaptation algorithm

The principle of the mesh adaptation is illustrated in Fig. 5.6. Every mesh node, except the border ones, can be described as the centre P_o of a set defined by the neighbouring nodes P_i, $i \neq 0$. These nodes define rectangular or triangular patches, depending on the topology. An ideal central point M would equalise the area of every patch. For every node of the mesh, such a point M is computed, then the nodes are translated towards these points. The rate at which the nodes are moved towards the ideal position is controlled with a parameter in the range between zero and one. If this parameter is set to zero,

the node P_0 is not moved. If it is set to one, the node P_0 is moved directly onto M. Our experience shows that for values of this parameter which are greater than 0.7, the algorithm tends to diverge.

To compute the position of M, Dresel *et al.* proposed to specify for the rectangular case

$$\begin{aligned} \mathcal{A}(A) + \mathcal{A}(B) &= \mathcal{A}(C) + \mathcal{A}(D) \\ \mathcal{A}(A) + \mathcal{A}(D) &= \mathcal{A}(B) + \mathcal{A}(C) \end{aligned}, \tag{5.43}$$

where the operator $\mathcal{A}$ stands for the area of the patch. This set of equations is equivalent to

$$\begin{aligned} \mathcal{A}(A) &= \mathcal{A}(C) \\ \mathcal{A}(B) &= \mathcal{A}(D) \end{aligned}, \tag{5.44}$$

equalising the areas of the opposite patches, as illustrated by the gray tints in Fig. 5.6(a). This leads to the equations

$$\begin{aligned} & x_M (y_2 + y_4 - y_6 - y_8) - y_M (x_2 + x_4 - x_6 - x_8) \\ & = x_1 (y_2 - y_8) + x_5 (y_4 - y_6) - y_1 (x_2 - x_8) - y_5 (y_4 - y_6), \end{aligned} \tag{5.45}$$

$$\begin{aligned} & x_M (y_2 - y_4 - y_6 + y_8) - y_M (x_2 - x_4 - x_6 + x_8) \\ & = x_3 (y_2 - y_4) + x_7 (y_8 - y_6) - y_3 (x_2 - x_4) - y_7 (x_8 - x_6), \end{aligned} \tag{5.46}$$

For the triangular topology, the areas to equalise are

$$\begin{aligned} \mathcal{A}(D) + \mathcal{A}(H) + \mathcal{A}(I) &= \mathcal{A}(B) + \mathcal{A}(E) + \mathcal{A}(F) \\ \mathcal{A}(A) + \mathcal{A}(B) + \mathcal{A}(C) &= \mathcal{A}(G) + \mathcal{A}(H) + \mathcal{A}(I) \end{aligned}, \tag{5.47}$$

which leads to

$$\begin{aligned} & x_M (-y_2 + y_5 + y_6 - y_9) - y_M (-x_2 + x_5 + x_6 - x_9) \\ & \quad = x_7 (y_6 - y_8) - y_7 (x_6 - x_8) + x_4 (y_5 - y_3) - y_4 (x_5 - x_3) \\ & \qquad + x_9 y_8 - x_8 y_9 + x_2 y_3 - x_3 y_2 \end{aligned} \tag{5.48}$$

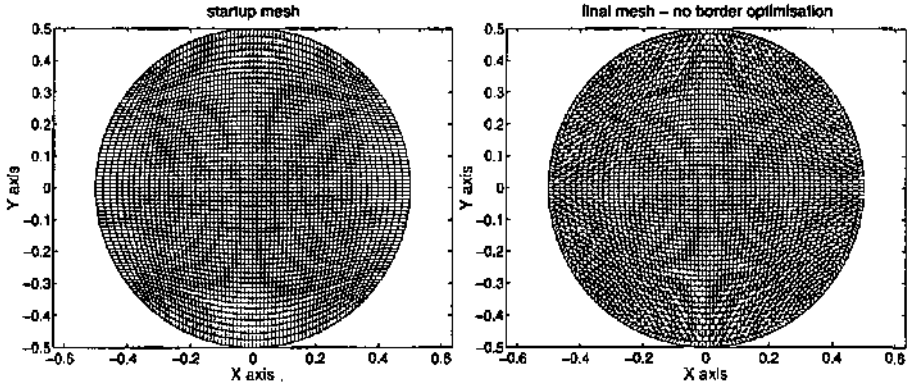


Figure 5.7: The startup mesh (a) in a rectangular topology is optimised to an adapted mesh (b) for a circular geometry.

and

$$\begin{aligned} x_M\left(-y_2-y_3+y_6+y_8\right)-y_M\left(-x_2-x_3+x_6+x_8\right) \\ =x_1\left(y_9-y_2\right)-y_1\left(x_9-x_2\right)+x_4\left(y_5-y_3\right)-y_4\left(x_5-x_3\right) \\ +x_9y_8-x_8y_9+x_5y_6-x_6y_5. \end{aligned} \tag{5.49}$$

An example of the process of mesh adaptation is shown in Fig. 5.7. The algorithm starts with the meshing of a disc in a rectangular topology composed of 65×65 nodes. After 1000 iterations, the mesh patches are equalised. A zoom on the optimised mesh will be presented later in Fig. 5.10(a).

The mesh adaptation gives the two components of the gradient of the phase. As noticed by Dresel, numerical integration is generally not possible, thus one needs another way to retrieve an approximate phase function from its two-dimensional gradient. The solution proposed is the use of a basis of polynomials of degree D. The binomial $x^k y^l$ is replaced by a monomial $\Omega_m = x^k y^l$, with $m = k\frac{2D-k+3}{2} + l + 1$. The approximated phase function can then be written as

$$\varphi\left(x,y\right) = \sum_{k=0}^{D}\sum_{l=0}^{D} a_{kl}x^k y^l = \sum_{m=2}^{\frac{(D+1)(D+2)}{2}} a_m \Omega_m, \tag{5.50}$$

with $a_m = a_{kl}$. There exists a bijection between the m and the (k,l) set. This approximated two-dimensional phase function can then be derived in x and y, and the partial derivatives fitted to the gradient in the least-squares way. The problem is finally reduced to a system of equations which can be solved by a matrix inversion. This is most commonly performed by Gaussian elimination, triangular lower-upper decomposition (LU) or singular-value decomposition (SVD) [12]. Although the SVD method is said to be most stable, we found that all three methods reach already their limits for degrees of polynomials around $D = 6$ in the case of the square-to-disc map-transformation. However, despite the warnings returned when using *Matlab* to perform the inversion, the results were still acceptable, while higher degree polynomials did not bring any improvement. This numerical limitation is due to the precision of the range available when adding two floating point values.

The previously described algorithm is based on equalising the areas of all the patches composing the mesh. It is thus implicitly assuming that both the input and output light distributions are uniform. While this is the case for our applications, it may not always satisfy one's needs. This led Hermerschmidt *et al.* to propose a different algorithm that fills this gap [99]. They successfully applied it to generate various shapes from a Gaussian beam [100]. The difference mainly lies in the mesh adaptation steps. The nodes, depending on their position, are weighted correspondingly to the intensity of the light. This value is obtained either analytically from their position (Gaussian beam) or by interpolation from a measurement (real beam). Whenever they are moved, this intensity has to be recomputed. From the values of the four surrounding nodes, the energy contained in the patch is deduced. The algorithm tries to adjust the position of the central node P_0 so that the four patches have equal energy. However, the authors find that the energy is locally uniform, but not on the whole mesh because of the local nature of the procedure. It is thus necessary to add a second variation of Dresel's algorithm. The mesh is started with a small number of nodes, and extra nodes are inserted by interpolation in order to refine the mesh. Between two node insertions, the mesh adaptation procedure is performed in order to obtain a reasonably stable state.

Extracting contours from the mesh gradients

A diffractive optical element (DOE) performing the desired mapping can be realised using the phase $\varphi(x,y)$ calculated in Eq. (5.50). For this purpose, we have to compute the loci of $\varphi(x,y) = \ell$, where ℓ are the values of the discrete phase level. We introduce again the piecewise cubic Bézier curves, already mentioned in chapter 3 and described in appendix A. These parametric

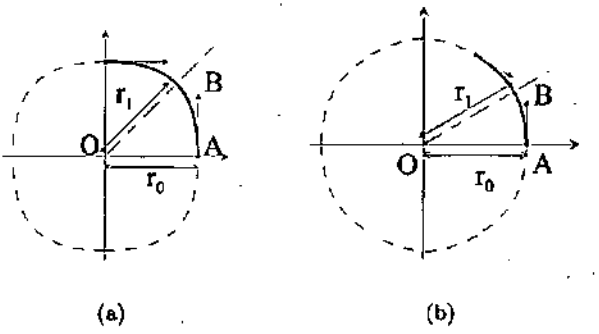


Figure 5.8: Decomposition of a 360° nearly circular ring (dashed line) into four (a) or six (b) cubic Bézier curves (thick plain line), for square and hexagonal cells geometries respectively. We impose the continuity of the tangents (arrows) between adjacent curves, which reduces the number of free parameters.

curves are defined by the knowledge of the start and end points and the tangent vectors of the curve at these two points.

As the geometry depends on every specific problem, we shall now restrict our study to the problem of shaping a uniform square beam into a uniform disc. In this case, we have observed from the polynomial fit that the contours are nearly circular. We choose thus to use one cubic Bézier curve to describe one sector of 90°. In Fig. 5.8, the geometry for square-to-disc and hexagon-to-disc beam-shaping elements is presented with the corresponding cubic Bézier curves.

To fit the contour with the Bézier curve, we use the following procedure. We impose that the Bézier curve and the ideal contour are superimposed on the 0° and 45° axes (30° in the triangular topology). To calculate the position of these points, we call their radial position on these axes r_0 and r_1, respectively. The coordinates of the points are thus

$$\begin{aligned} x &= r_0, \quad y = 0 \quad \text{at } 0° \\ x &= \tfrac{r_1}{\sqrt{2}}, \quad y = x \quad \text{at } 45° \end{aligned} \tag{5.51}$$

For a polynomial phase function of degree $D = 4$, with odd degrees polynomials absent because of the central-symmetry of the DOE, the contours are found by solving

$$\varphi(x, y) = a_{00} + a_{20}\left(x^2 + y^2\right) + a_{22}x^2y^2 + a_{40}\left(x^4 + y^4\right) = \ell, \tag{5.52}$$

where ℓ are the values of the discrete phase level. Using Eqs. (5.51), the values of r_0 and r_1 are determined from the coefficients a_{kl} as

$$\begin{aligned} r_0(\ell) &= \sqrt{\frac{-a_{20}+\sqrt{a_{20}^2-4a_{40}(a_{00}-\ell)}}{2a_{40}}} \\ r_1(\ell) &= \sqrt{2}\sqrt{\frac{-a_{20}+\sqrt{a_{20}^2-(a_{22}+2a_{40})(a_{00}-\ell)}}{a_{22}+2a_{40}}} \end{aligned} . \tag{5.53}$$

We now want to find the cubic Bézier curve passing through these two points. Taking into consideration the symmetry properties and the continuity of the tangents at the connection between two curves, only two parameters are left free, the radial position of point A and the size AB of the tangent vector. OA is easily identified to be r_0. We then adjust the value of AB so that the middle point of the Bézier curve is superimposed to the point of the contour curve on the 45° axis (30° for the triangular topology). The Bézier curve can thus be defined by $OA = r_0$ and the ratio

$$\frac{AB}{OA} = \frac{4}{3}\left(\frac{r_1}{r_0}\sqrt{2}-1\right). \tag{5.54}$$

For the triangular topology presented in Fig. 5.8(b), we would similarly get

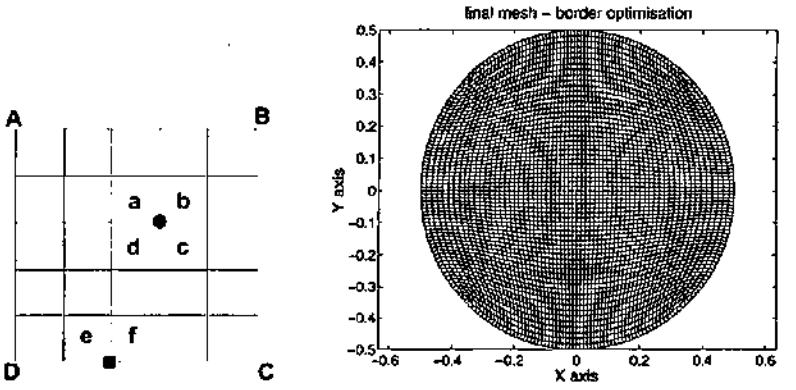
$$\frac{AB}{OA} = \frac{4}{3}\left(2\frac{r_1}{r_0}-\sqrt{3}\right). \tag{5.55}$$

For the general case of phase design by mesh adaptation, the modus operandi can be summarised as follows: We first study the problem with the original algorithm. In the resulting image, then we look for symmetries which can be exploited for the solution. Finally, we use a set of smooth curves to fit the ideal contour.

Improvement of the mesh smoothness

Figure 5.6 shows that the mesh adaptation is done for every node based on the areas of the adjacent patches, hence of the neighbouring nodes. Consequently, it is not possible to use the same equations when these neighbours are absent. As seen in Fig. 5.9(a), this is the case for the border of the mesh (black square). In the original algorithm, it is proposed to place the border nodes uniformly on the border and let them not move during the optimisation.

We refine this algorithm by adding the possibility for the border nodes to move. For the normal nodes (dots), we go on and equalise $\mathcal{A}(a)+\mathcal{A}(c)$ and $\mathcal{A}(b)+\mathcal{A}(d)$ as previously. For the border nodes, we move them in order

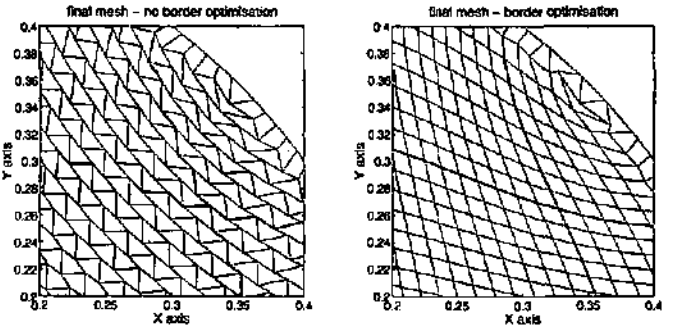


(a) (b)

Figure 5.9: The principle of the mesh adaptation with border optimisation (a), and the resulting mesh from the modified algorithm (b), corresponding to the startup mesh of Fig. 5.7(a).

to balance $\mathcal{A}(e)$ and $\mathcal{A}(f)$. The node is also constrained to be located on a defined curve (usually a circle or a line segment). This modification introduces more freedom for the optimisation. Only the four corner points A, B, C and D are fixed. The resulting mesh is much smoother, as can be seen when comparing the final meshes of Figs. 5.7(b) and 5.9(b). A zoom into one corner of the meshes is shown in Fig. 5.10. The patches are now less distorted from their square shape and the horizontal and vertical lines of the mesh are now smoothed with respect to the zigzag lines obtained with the basic algorithm.

A measurement of the convergence improvement can be obtained from the areas of the patches. Of special interest are the standard deviation and the peak-to-valley value of this set. Indeed, as the whole surface is constrained by the border of the mesh, the mean value of the patch areas is always constant, and therefore not significant. A comparison of the results for the original and the improved algorithms is given in Fig. 5.11, showing clearly the advantage of the modified algorithm. We have also observed that for very coarse meshes the convergence is faster at the beginning with this algorithm, but will reach a point where the influence of the corner patches and its neighbours prevent further improvement. The original algorithm offers a better convergence in the long run. However, there are two arguments in favor of the modified algorithm: Firstly, for increasing numbers of iteration, the original algorithm



(a) (b)

Figure 5.10: Zoom into the corner of the meshes of Figs. 5.7(b) and 5.9(b). (a) mesh without and (b) with optimisation of border nodes.

tends to diverge. Second, it is preferable to refine the mesh than to add more iterations.

To characterise the smoothness, we introduced two measures. The first one, which we call *aspect ratio*, can be used for any topology. It is defined as the ratio between the area and the circumference of a patch, normalised by the equivalent ratio for the "ideal" shape in the topology. This "ideal" shape is the one with the maximum ratio of area and circumference. This is a square or a rectangle in a rectangular topology, and an equilateral triangle in a triangular topology. For smooth meshes, this parameter should be close to one. Figure 5.11(c) shows the comparison of this parameter for the two algorithms. The second parameter, which we call *line smoothness*, measures the behaviour of the lines of the mesh. It is defined as the difference between the direction of one segment of the line and the mean direction of the previous and the following segments. This parameter quantifies the zigzag appearance observed in Fig. 5.10(a). For an ideal line, this parameter should be close to zero. Figures. 5.11(c) and 5.11(d) show how the border optimisation smoothes the mesh in comparison with the original algorithm. It should be emphasised that smoothness of the mesh is not only cosmetic. Fitting either the whole phase or specific lines of the mesh by polynomials is easier when the derivatives are smooth functions.

Our experience is that the presence of the distorted patches at the borders and in the corners makes it difficult to obtain the best possible element. Their influence can be minimised efficiently by increasing the number of nodes, so that the area covered by the problematic patches becomes negligible compared

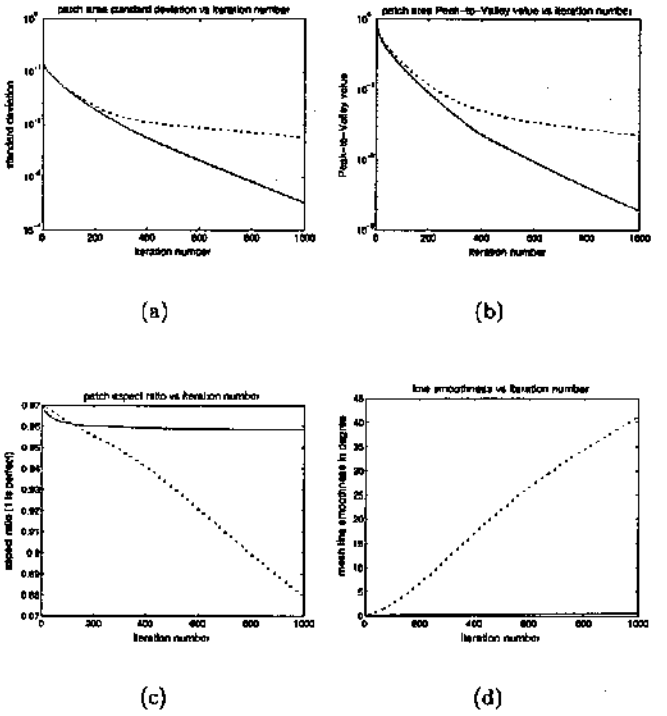


Figure 5.11: Standard deviation (a), peak-to-valley value (b) of the areas of the patches, normalised aspect ratio of the patches (c) and line smoothness (d) obtained with the original algorithm (dashed line) and compared with the modified algorithm with border optimisation (solid line).

to the whole mesh area. This is illustrated in Fig. 5.12. The left column shows the resulting light distribution from DOEs designed with a coarse mesh of only 41×41 nodes, while the right column results from a finer mesh of 401×401 nodes. Nevertheless, the square-to-disc problem is not representative, since the disc has no sharp angles. We have observed that the node distribution at the border is critical for shapes where corners are more pronounced, like a triangle or a "pie slice". The problem of transforming an hexagon into a disc can be treated using such shapes in a triangular topology. We would like to stress that the influence of corners and borders can be important for general beam-shaping problems.

As shown by the comparison of top and bottom rows of Fig. 5.12, the simple model we used for 360°, shown in Fig. 5.8(a), is not precise enough to equal

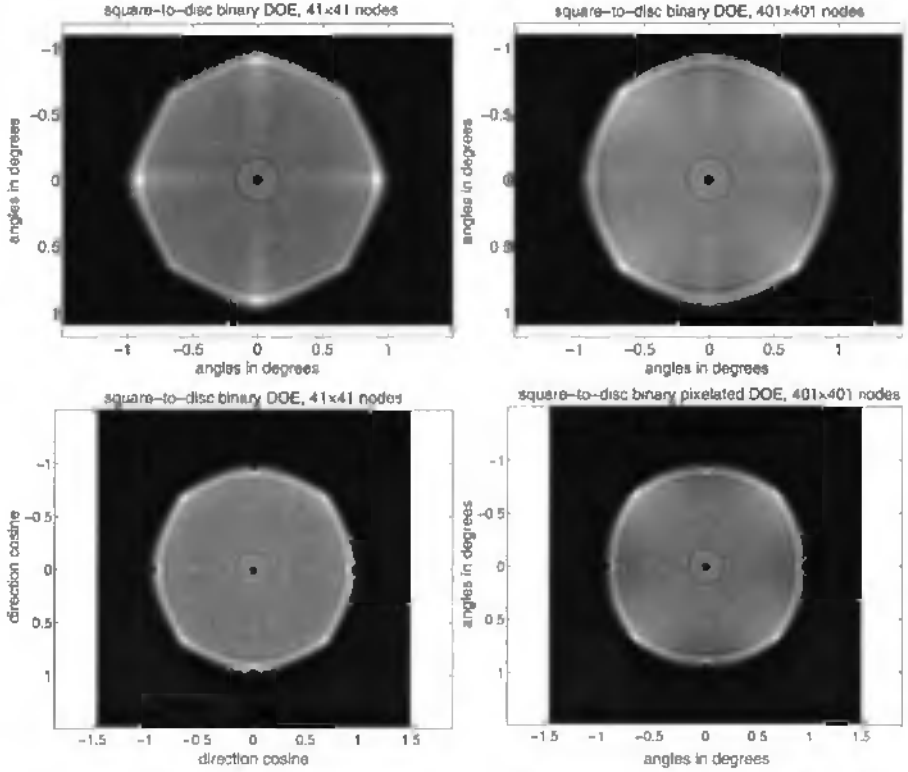


Figure 5.12: Far-field light distribution for a square-to-disc beam-shaping element. Top row are Bézier based DOEs, bottom row are pixelated DOEs. Left column are with 41×41 nodes, right column with 401×401 nodes.

the polynomial fit, however, it gives already an acceptable result. A further improvement could be proposed, based on two Bézier curves for 90° instead of one only.

5.1.6 Advantages and limitations of re-mapping elements

The fundamental characteristic of the re-mapping design is that it is based on geometrical optics. Consequently, it provides a straightforward and intuitive relation between the element and the resulting light distribution. This property allows rapid design and understanding of the sources of errors. Unfortunately, geometrical optics does not take diffraction effects into account, and therefore undesired oscillations will appear in the real light distribution. Their number

is given by the Fresnel number

$$N_F = \frac{a^2}{\lambda f}, \tag{5.56}$$

where a is the aperture of the element and f is the focal length of the element. For a far-field beam shaping element, the pertinent parameter is the maximum angle of illumination $2\alpha \cong a/f$, and the Fresnel number can be rewritten as

$$N_F = \frac{2\alpha a}{\lambda}. \tag{5.57}$$

The larger the number of oscillations, the smaller they are, thus the easier it is to get rid of them by filtering (an incoherent beam with a small divergence acts like a low pass filter, so does any measuring device composed of extended pixels). We will study in more details these oscillations in chapter 6.

The second drawback of these type of element is not related to the design method, but to the principle of map transform as presented in Fig. 5.1. The point to point relation defines a bijection between input and output. This implies that any local perturbation in the input distribution will be propagated to the output. This makes these elements sensitive to fabrication errors or fluctuations of the input beam, as we shall see in chapter 6.

5.2 Grating-type elements

Mainly due to the ease of computation of the Fourier transform with modern computers, the far-field and finite-distance diffraction for paraxial approximation can be evaluated for moderate size optical elements in reasonable times. This allows designs which take diffraction effects into account that are not predicted by geometrical optics. As the Fourier transform is the basis of this design concept, every output point is a function of all input points. This represent a fundamental difference from the previous design techniques. The optical element can be seen as a sum of gratings, corresponding to the output spots of light, thus the name grating-type, or diffuser-type, elements.

Due to the constraints imposed for most of the design problems, it is not possible to inverse only the propagation to retrieve the optical element structure. The problem must be solved by iterative techniques, which can be classified in two main families: direct and indirect approaches [105]. Direct approaches intend to minimise a merit function defined for the output by changes in the input distribution. The computed DOE thus always satisfies the constraints. Simulated annealing [106], direct binary search [107] and genetic algorithms [108] are popular algorithms for that purpose. On the other

hand, the indirect approaches alternate between the input and output planes, accepting a violation of the constraints on one side when they enforce those constraints in the other side. They are based on the error reduction algorithm, also known as the iterative Fourier transform algorithm (IFTA) or Gerchberg-Saxton algorithm [48, 109, 110]. Fienup has demonstrated how this algorithm succeeds to reduce both errors at every iteration [111].

5.2.1 Direct approaches

We shall now briefly tackle various techniques inspired by life and material sciences, laying the emphasis on simulated annealing.

Simulated annealing and quenching

Simulated annealing is based on an analogy with material science and statistical physics. If a metal (or another material) is cooled sufficiently slowly, its atoms will move in order to be in a state of minimal energy (as a crystalline configuration). As the notion of temperature is defined only at equilibrium, the cooling process must be slow enough to allow the system to reach a state of equilibrium after each step of cooling during the annealing.

To reach the equilibrium after each step of cooling, the following algorithm, first introduced by Metropolis, is applied. Changes are introduced in the configuration of the system, and they are accepted if they decrease the energy. If they result in an increase ΔE of the energy, they are accepted with a probability of

$$P(\Delta E) = \exp\left(-\frac{\Delta E}{k_B T}\right), \tag{5.58}$$

which simulates the statistical nature of the behaviour of the atoms. After a sufficient number of changes, the system reaches a stable state and the temperature can be changed to a lower value. The process of successive cooling and Metropolis procedures will be stopped when the configuration is frozen or when the element satisfies the constraints. In the following, we will consider that the Boltzmann constant k_B in Eq. (5.58) is equal to one.

For the operation of simulated annealing we need a temperature T, an energy E, given by the merit function of our process, and an annealing schedule, defining how fast T is decreased between two Metropolis procedures.

There exists several variants of simulated annealing. Depending on the system complexity, it is possible to rely on fast annealing or adaptive simulated annealing [112] (or very fast simulated re-annealing). However, all of the applications for computer generated hologram (CCH) designs are adapted from the basic simulated annealing [12].

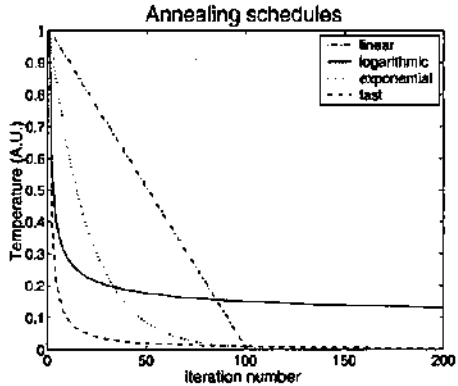


Figure 5.13: Comparison of the cooling of various annealing schedules. Any schedule that cools the temperature faster than the logarithmic one does not guarantee the convergence of the algorithm towards the global optimum.

The simulated annealing algorithm is composed of two nested iterative routines (Metropolis procedure inside a temperature cooling process), therefore it is important to understand how both stages can be influenced by the implementation.

Most of the implementations applied in optics use an exponential annealing schedule

$$T(k) = T_0 c^k = T_0 \exp(k \ln c), \tag{5.59}$$

where k is the iteration number in the annealing algorithm and $c < 1$ [107, 113–116]. Recent examples use a linear annealing schedule [117]. However, it can be proved that the basic simulated annealing is guaranteed to converge to the global optimum only if the cooling process in not faster than

$$T(k) = \frac{T_0}{\ln(1+k)}. \tag{5.60}$$

Figure 5.13 illustrates various annealing schedules, and compares their speed with the much slower logarithmic annealing. These implementations are also called simulated quenching [118, 119]. Simulated quenching is much faster than simulated annealing and provides optima that are usually sufficient to satisfy the constraints. A further improvement proposed by Gillet is to re-heat the system when a stagnation is reached [120, 121].

The other parameter that can influence the convergence is the Metropolis procedure, trying to stabilise the system in an equilibrium state where the temperature is defined. Some algorithms perform a fixed number of iterations in each Metropolis stage [113, 114, 117], while others prefer to test the presence

of an equilibrium [116]. Gillet asserts that a good compromise between speed and stability is to perform twice more iterations than the number of degrees of freedom [121].

Other direct approaches

Among the various direct approaches, one can mention the evolutionary algorithms based on life science analogies, such as genetic algorithms [108] and evolutionary strategies [122]. A simplified version of simulated annealing, the direct binary search [107], can be used for binary-phase or binary-amplitude DOE design. Finally, Chen [123] has proposed to design holograms by a non-linear least-squares algorithm (NLS). Unfortunately, every iteration of this last method is $\mathcal{O}\left(N^2\right)$ times slower than an IFTA iteration, where N is the number of pixels in one direction of the CGH which makes it unsuitable in practice for most design situations.

Advantages and weaknesses

The main strength of direct methods is that the implementation requires only forward propagation. The output data does not have to be used as parameter and as data for backward propagation. Especially, while IFTA variants all have in common the pixelated two-dimensional sampling characteristic, the variety of parameters to optimise is wider with direct methods. Examples of CGH with irregular pixels [66] or trapezoidal pixels have been optimised with simulated annealing [67, 68, 120]. There is an example of trapezoidal pixels CCH designed by IFTA [61], however, the optimisation of the pixels shapes is only done at the end, and using a parametric optimisation. The refinement of pixelated kinoforms presented in section 3.4.5 would be a good candidate for a direct optimisation.

On the other hand, the principal drawback of direct search approaches is the time needed to reach an acceptable solution, especially when the number of parameters is growing. Evolutionary strategies are particularly slow, even for small designs (array size 64×64), as shown by Birch [122]. They may only be used as a final possible refinement stage if the constraints are not satisfied by a first design stage. Moreover, the evolutionary algorithms have important memory requirements, as they need to store many CGHs at the same time. They are thus not well suited for the design of large elements.

5.2.2 Iterative Fourier transform algorithms

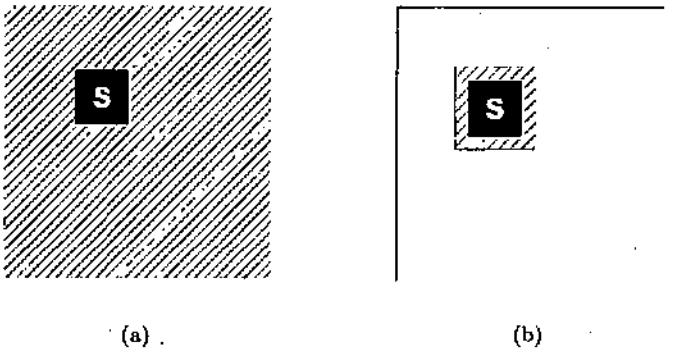
The popularity of iterative Fourier transform (IFT) algorithms is due to their low requirements of time and memory as compared to the direct approach. Since the original error-reduction algorithm [109], many improvements have been proposed. The problems that arises from the original IFTA are the slow convergence and the poor performance in terms of efficiency and uniformity. For the case of multilevel elements, the stagnation of the solution in a local optimum has been the main limitation of these algorithms.

To solve these problems, many variants have been proposed, which we can categorise in two families. Firstly, most of the techniques are only different from the original algorithm by the application of operators specific to the constraints. Smooth projection is popular to quantise progressively amplitude or phase [124,125]. Filtering of high frequencies may be used to suppress small features [126,127].

Secondly, based on the work of Fienup, IFTA using a driving function have emerged [110,111]. The result of an iteration is used to determine the changes to be applied to the original function in order to improve the DOE. Fienup's input-output algorithm, Prongué's over-compensation [128] or Johansson's up-scaling [129] are examples of such methods. The latter two are characterised by the fact that during the signal constraint step, the orders are changed to an amplitude value that is not only a function of the goal, but also a function of the amplitude at the previous iteration.

As outlined by Wyrowski [105,125], the degrees of freedom are important to identify and to use. He pointed out that for beam shaping and beam splitting, the phase of the signal is subject to few constraints (*phase freedom*), as is the field outside the area of the signal window shown in Fig. 5.14(a) (*amplitude freedom*). Additionally, some latitude is available in the strength of the signal relative to the noise. Although this is not strictly speaking a new degree of freedom, it is often called *scale freedom*. Scale freedom can be seen either like an enhancement of the signal or an attenuation of the noise. Most of the IFT algorithms do not change the value of the orders outside the signal window. However, due to the conservation of energy, up or down-scaling the signal orders reduces the noise orders. Scale freedom is thus just the possibility to reduce the noise level with respect to the signal level.

The principle of the IFTA is illustrated in Fig. 5.15. At iteration k, the complex amplitude in the signal plane, called A_k is back-propagated to the DOE space, leading to the complex amplitude in the object space a'_k. Then the object constraints are applied to produce a new amplitude a_k, which is propagated to the signal plane leading to a complex amplitude A'_k. The goal to achieve is A_{goal}, thus the signal space constraints are now enforced, to



(a) (b)

Figure 5.14: The signal window (in black) is the place of the image field where the signal amplitudes are constrained. The noise (dashed) is the energy outside the signal orders if no orders are constrained to be zero (a). If some orders are constrained to be zero (dashed frame), then the noise is the total intensity of such orders (b).

produce A_{k+1}, which is the start of a new iteration. If the constraints for the DOE are enforced strictly, the error in the DOE domain is not relevant; only the error in the signal is of interest. The two propagation steps and the DOE space step may be grouped, symbolised by the dashed box in Fig. 5.15, as the input-output algorithm.

Efficiency and uniformity

We have studied and compared various methods for DOE designs based on IFTA. The goal was to design a circular flat-top with a periodic phase element. The profiles were binary, eight-level or continuous. The start phase of the signal was either resulting from a quadratic lens or random (but identical for all designs). The two main criteria used to measure the performance of the algorithms were the diffraction efficiency defined as the ratio of the amount of light in the desired orders and the total amount of light, and the uniformity error

$$U = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}, \tag{5.61}$$

inside the disc with $I_{\min}$ and $I_{\max}$ being the energies of the weakest and the strongest orders of diffractions. At first, the amplitude freedom was applied everywhere outside the disc, but we also considered the capacity to impose a dark window, i.e. an area where the orders have the least possible energy. The noise was then characterised by the ratio of the energy in this dark window with respect to the whole energy, as illustrated by Fig. 5.14(b).

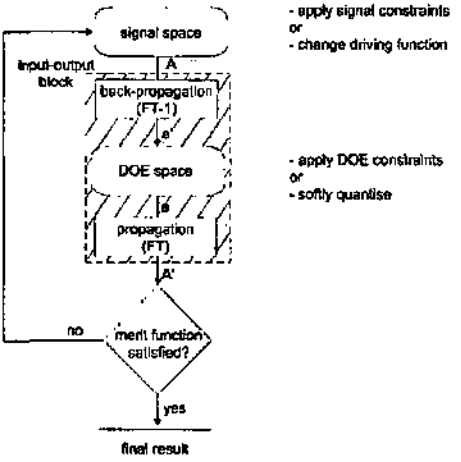


Figure 5.15: Principle of the IFTA. The steps of back-propagation, application of DOE constraints and forward propagation can be grouped as a unique stage (dashes), for the input-output algorithm. When the specifications are met, the algorithm is left.

We compared the following variants of the IFTA:

1. The original error-reduction algorithm, with the possibility of noise attenuation described above.

2. The Fienup input-output algorithm, given by Eqs. (9) and (10) of Ref. [110]. Fienup called this variant output-output in a later article [111]. The signal orders are changed from iteration k to $k+1$ according to

$$A_{k+1} = A'_k + \beta \Delta A_{\text{driving},k}$$

with

$$\Delta A_{\text{driving},k} = |A_{\text{goal}}| \{2 \exp[\mathrm{i} \cdot \arg(A'_k)] - \exp[\mathrm{i} \cdot \arg(A_k)]\} - A_k.$$

β is a free parameter, usually chosen close to one. The input-output family is specially designed to give high efficiency.

3. The over-compensation proposed by Prongué *et al.* [128] and modified by Schilling [77]. The amplitudes A'_k in the signal window of the output plane are modified for A_{k+1} according to

$$|A_{k+1}| = |A_k| \frac{\langle |A'_k| \rangle}{|A'_k|}, \tag{5.62}$$

where $\langle|A'_k|\rangle$ is the average of the $|A'_k|$. Thus, the low energy orders are set to higher values and the high orders are decreased. This technique requires to take care to the conservation of the energy, by re-normalising the whole energy at every step. Also, one has to avoid the cases where the denominator tends towards zero and implies divergence of the compensated value. It has to be noted that the amplitudes in Eq. (5.62) can be replaced by intensities, without changing much the convergence properties of the algorithm [77]. This algorithm is restricted to binary signal distributions, but can be easily extended to general distributions.

4. The algorithm proposed by Arrizón *et al.* [130]. It is characterised by the possibility to reduce the desired energy within the disc. Whenever the uniformity is not improving enough during an iteration, the efficiency goal is slightly decreased, allowing the uniformity error to be lowered again. Although this algorithm was originally used for parageometric starting solutions (quadratic phase DOE in our case), we used it successfully with random start phases. Here again, we used the possibility to enhance the signal in comparison to the noise as a possible design parameter. The algorithm is stopped when the goal for uniformity error is reached or when stagnation appears. In practise, we used a goal of 1% or 0.5% for the design.

5. The up-scaling algorithm proposed by Johansson *et al.* [129] (the term up-scaling was coined by Schäfer [131]). This algorithm is based on a concept similar to over-compensation. Two real thresholds are defined, a lower $A_{\min}$ and a higher $A_{\max}$, around the desired real signal value A_{goal}. The signal orders are changed according to

$$|A_{k+1}| = \begin{cases} A_{\max} & \text{if } |A'_k| \leqslant A_{\min} \\ A_{\min} & \text{if } |A'_k| \geqslant A_{\max} \\ 2A_{\text{goal}} - |A'_k| & \text{otherwise} \end{cases} . \tag{5.63}$$

The noise orders are not modified in the original algorithm. We applied here again the possibility to up-scale the signal with respect to the noise.

6. For multi-level elements, we also considered the error-reduction algorithm with the soft-quantisation proposed by Wyrowski [125]. The algorithm consists of three stages. First, a phase synthesis is performed where no noise is allowed and the element is continuous (use of phase freedom only). Then the amplitude freedom is introduced outside the signal

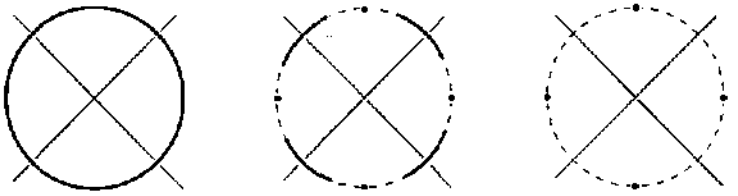


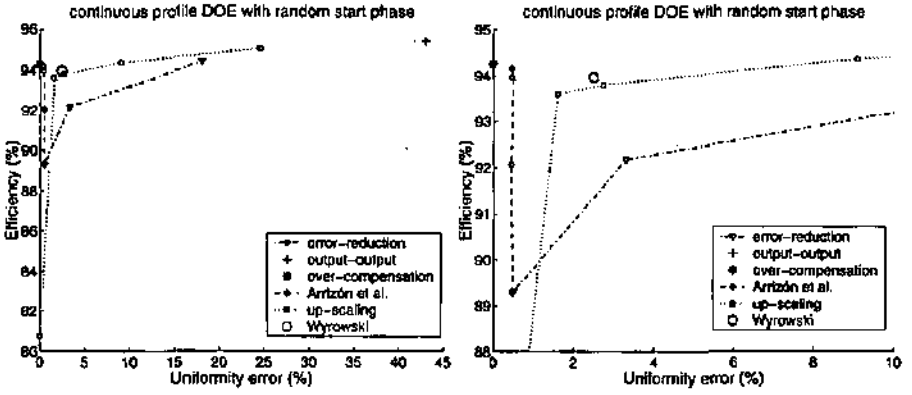
Figure 5.16: Soft quantisation on four levels of phase in the complex plane. The phases are spread at the start over the entire angular interval of 2π (a), and they are progressively projected onto the four phase levels (b) until the quantisation process is ended (c).

window. Finally, the soft quantization is performed. Soft-quantisation consists, as shown in Fig. 5.16, to quantise progressively the phases in the DOE space. There exists variants of this algorithm, but we will concentrate on this widespread three stages version [132]. For the study of continuous elements, this method was used, although the soft-quantisation that characterises it was not applicable.

The results of such a comparison are presented in the next figures. We first studied the situation of a random start phase. In this case, the resulting beam shaping element is usually classified as a diffuser. The same start phase was used for all the algorithms. We aimed at designing a circular flat-top distribution with a CGH of 128×128 pixels illuminated by a uniform plane wave.

For a continuous element, as illustrated in Fig. 5.17, all the methods can achieve good uniformities, but the over-compensation and Arrizón's algorithm are clearly giving the best efficiency and uniformity compromises. The up-scaling can be seen to be very versatile, depending on the up-scaling parameter. Fienup's output-output gives the highest efficiency, but the uniformity is far from being optimum. Finally, Wyrowski's technique returns good uniformity and efficiency, but is harder to tune in order to choose the compromise. From these observations, Arrizón's technique and the up-scaling appear as very powerful design techniques for continuous phase profiles. Wyrowski's three stages method and the over-compensation also show interesting performances.

The results for elements with discrete phase levels are presented in Figs. 5.18 and 5.19. Figure 5.18 shows the result of a design of an eight-level CGH. Two techniques appear most useful. Wyrowski's algorithm is best because it can now make full use of its soft-quantisation stage. Moreover, it shows a good ca-

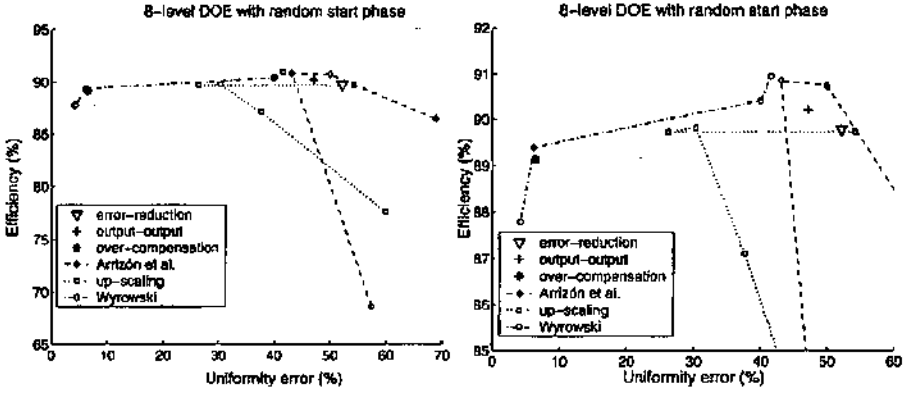


(a) (b)

Figure 5.17: Comparison of efficiency and uniformity error of a continuous profile beam-shaping element calculated with different IFTA algorithms and a random start phase; (b) is a zoom-in on the details of (a).

pability to be tuned in order to modify the efficiency-uniformity compromise. A close second is the over-compensation, whose only drawback is the fixed efficiency-uniformity balance. Both up-scaling and Arrizón's algorithm need to be tuned in order to find an optimum input parameter value. For smaller or larger values of this parameter, either the uniformity or the efficiency will degrade without improvement of the other. Error-reduction and output-output are unable to provide acceptable uniformities and are not tunable. Figure 5.19 illustrates the binary DOE designs. There again, over-compensation offers a good solution. Wyrowski's algorithm starts to fail, but is still tunable. At the expense of efficiency, it can somehow offer good uniformity. As previously, output-output is very efficient but poorly uniform. One can see that the uniformity offered by the algorithm is almost independent of the number of quantisation levels. As for the eight-level CGH, up-scaling and Arrizón's algorithm have an optimum point, and the input parameter does not provide a useful way to choose the efficiency-uniformity compromise.

The role of the start phase distribution To illustrate the influence of the start phase, we performed the same designs with a start phase distribution corresponding to the one generated by a quadratic-phase lens whose angles would be close to the desired distribution. The efficiency and the uniformity



(a) (b)

Figure 5.18: Comparison of efficiency and uniformity error of a disc beam-shaping element with eight phase levels calculated with different IFTA algorithms and a random start phase; (b) is a zoom-in on the details of (a).

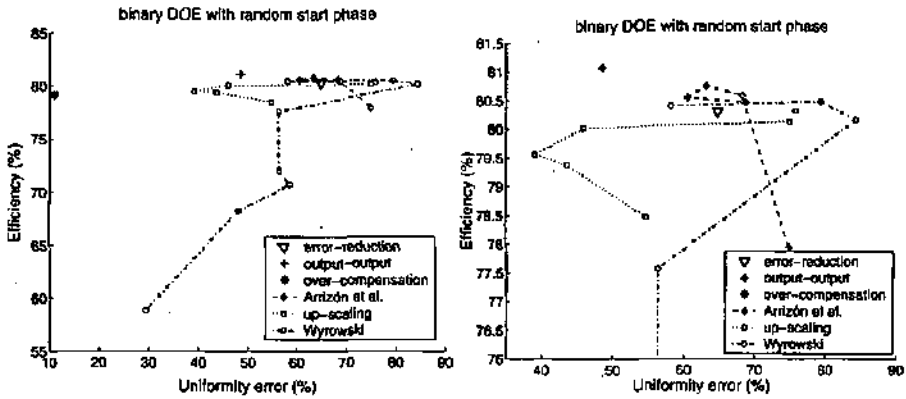
are significantly better in that case. The noise is now not randomly located, but appears to be more systematic.

General observations and conclusions

To conclude, we list some peculiarities of the tested algorithms.

Convergence of IFTA for multi-level DOE We have observed that in order to use the error-reduction or the up-scaling algorithms for multilevel elements, we have to attenuate the noise (or equivalently to enhance the signal) at every step. Otherwise, it was not possible to come to a solution.

Over-compensation Although it performs very well in most situations, over-compensation cannot be tuned as most of the other algorithms. It rapidly approaches its optimum results and starts oscillating between similar configurations. Adding more iterations will not ensure the appearance of a better solution, except from a statistical perspective. It is thus important to compare every iteration with the best of all the previous ones with respect to the merit function. As most of the other techniques, over-compensation is not very efficient at imposing zero intensity areas. Nevertheless, it is very efficient when the others fail, especially for binary element design.



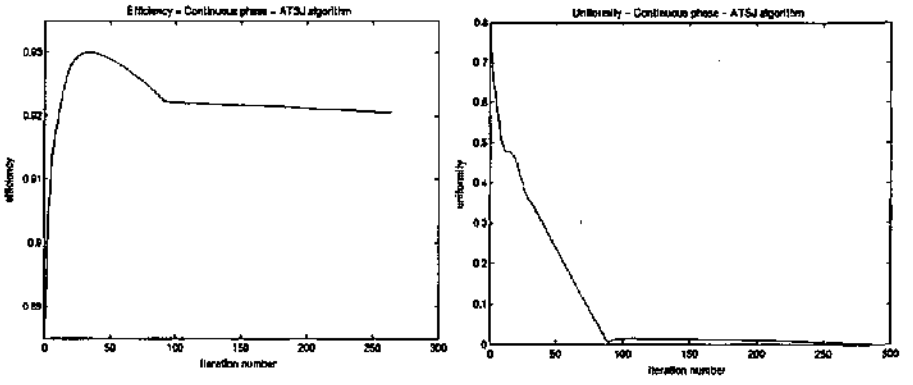
(a) (b)

Figure 5.19: Comparison of efficiency and uniformity error of a disc beam-shaping element with two phase levels calculated with different IFTA algorithms and a random start phase; (b) is a zoom-in on the details of (a).

Arrizón's algorithm Arrizón's algorithm was designed to get a small uniformity error and is efficient at this task for continuous profile. The large number of parameters makes it versatile, with a possibility to balance efficiency and uniformity. One can also easily set a uniformity goal, where the algorithm will stop. As seen from Fig. 5.20, and in the original article, efficiency reaches a maximum (around $k \simeq 30$ in our example) and starts then to be traded for more uniformity. After some iterations ($k \simeq 80$), both uniformity error and efficiency do not evolve much more, as a stable state has been reached. For these particular examples, the goal chosen for the error was 0.5%.

It must be noted that there exists a possibility to over-tense the algorithm. This allows to obtain a higher efficiency as seen in Fig 5.21, by enforcing a lower noise outside of the signal window. The convergence behaviour is different, with an additional initial stage where no convergence is reached as the constraints are too strict. As they are relaxed, the algorithm reaches a point where it behaves as previously described, that is an increase of efficiency until a maximum ($k \simeq 1000$) followed by a decrease of efficiency and uniformity error and a final stagnation stage ($k \simeq 1050$).

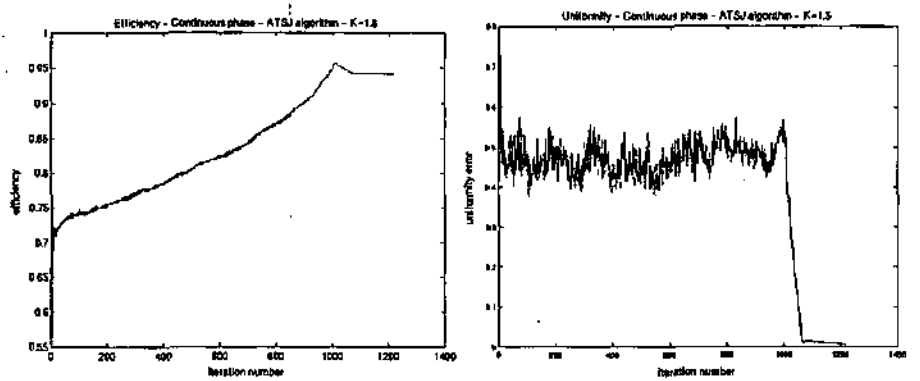
Wyrowski synthesis and soft quantization algorithm Wyrowski three steps algorithm performs extremely well with multilevel structures and can be tuned in order to provide efficiency and uniformity. It is mainly characterised



(a) (b)

Figure 5.20: Efficiency (a) and uniformity error (b) of the original Arrizón's algorithm. When uniformity error does not decrease enough, some efficiency is sacrificed. The algorithm reaches 92.05% efficiency with a uniformity error of 0.45%.

by the ability to enforce the zero intensity window, as opposed to all the others. However it is not very well adapted for binary phase elements. We noticed that, in this case, it tends to generate isolated pixels. Nevertheless, good efficiencies can be achieved using noise attenuation, which results in the reduction of the uniformity, as can be seen in the top part of the curves in Fig. 5.19(a). It has been proposed to overcome this drawback by a special filtering of the DOE phase [126].



(a) (b)

Figure 5.21: Efficiency (a) and uniformity error (b) of the over-tensed Arrizón's algorithm. The decrease of efficiency is compensated by the tensed-up goal, resulting in an efficiency of 94.15% for a uniformity error of 0.48%.

Compensation of design and fabrication errors 6

We have seen that designs based on geometrical optics do not take into account the influence of diffraction and that fabrication can cause differences between the desired and the measured intensity distribution. We shall review here the main types of errors observed in DOEs and, when possible, we shall present some solutions to decrease their influence. We have noticed that beam-shaping elements can be subdivided into different categories with respect the design method and to the fabrication errors. In chapter 5, we have put the emphasis on the distinction between the design methods: re-mapping elements based on geometrical optics and grating-type elements based on wave optics. The first are sensible to diffraction oscillations. From the perspective of fabrication, the refractive and diffractive nature of the optical structure must also be taken into account. Additionally, grating-type elements are themselves divided into diffusers and para-geometric elements. Diffusers are optical elements whose structure is close to random, while para-geometric elements are similar to re-mapping elements. Para-geometric elements are usually grating-type elements whose phase is built from a geometrical start phase, commonly quadratic, as presented at the end of chapter 5.

6.1 Preliminary remarks

Before studying the effect of specific errors, we shall first tackle the differences between beam-shaping and focusing for diffractive optical elements. Then we shall emphasize the distinction between refractive and diffractive elements, and between grating type and re-mapping type elements. Finally, we shall study the limitations of the model used to simulate the influence of the errors.

6.1.1 Parasitic orders in beam-shaping and focusing

First of all we would like to put the emphasis on the difference between far-field beam shaping and focusing with respect to the influence of the errors. For the second case, the energy of the desired diffraction order is concentrated at the focal point, resulting in a high density of energy for the considered order. The signal to noise ratio is thus considerably enhanced, as the energy of the parasitic orders is dispersed over a large surface. For far-field beam shaping, the situation is different, especially with respect to the zero order. Indeed, only the orders ± 1 (one or both) are usually used as signal orders, while the orders $\pm n$, $n \neq 1$ are parasitic scaled duplicates of the signal. Their angular extension is n times larger than the first orders in both directions, resulting in an energy density evolving as

$$\frac{\eta_n}{\eta_{\text{signal}}} \cdot \frac{1}{n^2} \tag{6.1}$$

where η_n is the efficiency of the order n. We see that the contribution of the parasitic higher orders to the desired illumination is usually small. However, the case of the zero order is clearly problematic. The energy is concentrated within a single direction and causes a high peak at the origin even for a small total energy. For a one-dimensional far-field beam-shaping, the situation is different ; the n^2 in Eq. (6.1) changes to n. The zero frequency peak is less present, but higher orders do not decrease as fast as for two-dimensional beam shaping.

In practice, the influence of the zero order is reduced by two phenomena. Firstly, the limited spatial extension and the spatial coherence of the illumination beam widens the peak. Secondly, in our application, the angular divergence of the input beam causes a smoothing of the resulting light distribution. The second effect is more important than the first one. However, the influence of the zero order is still important. As an example, we compare the energy contained in a uniform distribution of an angle of 25 mrad and in a Gaussian spot of a width of $\sigma = 1$ mrad, simulating the signal order and the zero order of a typical flat-top beam-shaping element, respectively. For the one-dimensional case, the Gaussian spot reaches the desired uniform level already with 4.6% of the total signal. For the two-dimensional case, already 0.2% of the total signal energy is sufficient to bring the spot at the same level.

6.1.2 Localised error for re-mapping and grating DOEs

An important distinction between grating and re-mapping type elements exists regarding their tolerance to a localised error $\epsilon(\mathbf{x})$. The effect of such an error

is the multiplication of the input intensity in Eq. (5.8) by the local grating efficiency $\eta(\mathbf{x}) = \mathrm{f}(\epsilon(\mathbf{x}))$. We have seen in chapter 5 that the elements obtained by re-mapping design are a subset of all possible solutions. This subset is characterised by the fact that the energy conservation of Eq. (5.8) is given by the differential formulation

$$I_i(\mathbf{x})\,\mathrm{d}S_{\mathbf{x}}(\mathbf{x}) = I_o(\mathbf{u})\,\mathrm{d}S_{\mathbf{u}}(\mathbf{u}). \tag{6.2}$$

where $\mathrm{d}S_{\mathbf{x}}$ and $\mathrm{d}S_{\mathbf{u}}$ are the elements of surface in the input and in the output planes, respectively. This simplification is possible because there exists a map-transform $g : \mathbf{x} \rightarrow \mathbf{u}$ between the coordinates of the input and output planes given by Eqs. (5.1) to (5.4). This map-transform is only dependent on the local frequency of the optical element, not on the grating profile inside a period which influences essentially the intensity repartition among the grating orders. Consequently, the map-transform is not affected by the errors. From this observation, we can draw a first conclusion on the effect of a local error on an optical element. The diffusers will spread the effect of the error over the entire far-field light distribution. On the other hand, the elements which exhibit a map-transform relation, i.e. re-mapping and para-geometric grating-type elements, will transfer a localised error to the far-field distribution. This result can be expressed as

$$I_{o,\mathrm{error}}(\mathbf{u})\,\mathrm{d}S_{\mathbf{u}}(\mathbf{u}) = \eta(\mathbf{x})\,I_i(\mathbf{x})\,\mathrm{d}S_{\mathbf{x}}(\mathbf{x}) = \eta(\mathbf{x})\,I_o(\mathbf{u}(\mathbf{x}))\,\mathrm{d}S_{\mathbf{u}}(\mathbf{u}(\mathbf{x}))\,. \tag{6.3}$$

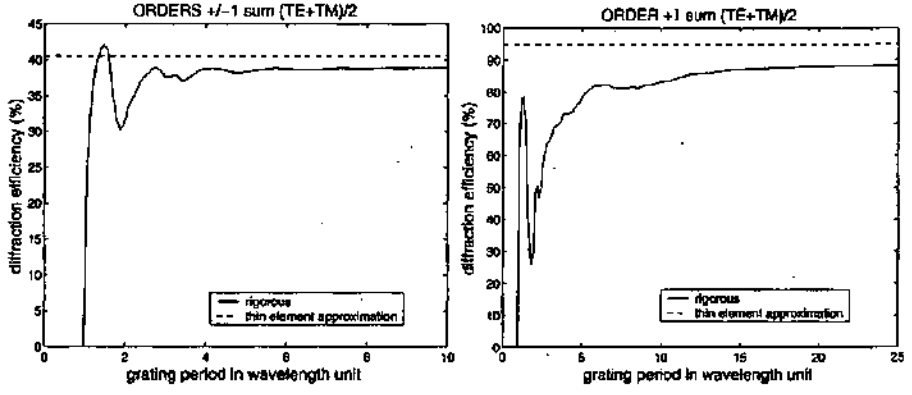
The difference between para-geometric and diffusers grating type elements has been reviewed earlier by Aagedal or Senthilkumaran [133, 134].

6.1.3 Differences for diffractive and refractive types of elements

The previous property is not shared by refractive elements. Indeed, an error for a refractive element cannot be modelled by the a change of efficiency. It will instead influence the profile, thus the local grating frequency. This will in turn modify the map-transform relation, leading to a global geometrical distortion instead of a local intensity modification. This difference has been worked out by Ehbets [80] and Hessler [81]. The latter has elegantly outlined how the position and the intensities of the focal spots are changed by fabrication errors.

6.1.4 Error simulation method

Another point to consider in the study of the errors is the methods used to simulate their influence. Of course, rigorous methods are the only ones to take



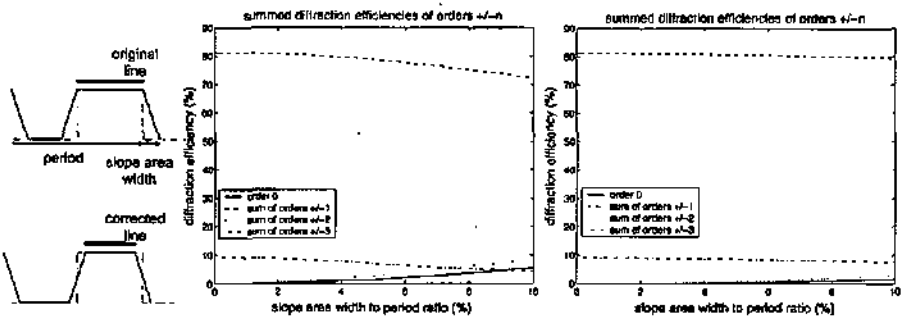
(a) (b)

Figure 6.1: Comparison of the efficiency of the 1st order of a binary (a) and an eight-level (b) grating for rigorous calculation and paraxial scalar optics. The rigorous curve is the average of both polarisations. The constant line is the value predicted by scalar theory (40.53% for a binary and 94.96% for an eight-level grating).

into account all the peculiarities of the errors, but their use is limited for the reasons explained in chapter 2. However, they shall be used to validate or complement the use of scalar theory and thin element approximations.

In order to judge the accuracy of the thin element approximation (TEA) in our domain of application, one can compare the diffraction efficiency of a grating calculated by rigorous methods (Fourier modal method in our case) and the behaviour predicted by TEA, i.e. a constant value. As an example, we consider a grating in calcium fluoride (refraction index $n = 1.559$ at 157 nm). Figure 6.1 shows the corresponding results for comparison. The paraxial scalar approximation for a binary grating is valid as soon as the grating period is about four times larger than the wavelength. Though ten wavelengths are required for an eight-level grating, the behaviour is already stable when the period is five wavelengths. This validates its use for the study of most fabrication errors, especially the ones which do not create small features.

When the nature of the error cannot be fully explained through a one-dimensional grating, we will also give a brief description of the two-dimensional behaviour, with respect to symmetries and to the differences between the optical elements of re-mapping type or grating-type. However, as we are mostly interested in a qualitative study of the errors, the one-dimensional simulations should provide enough information.



(a) (b) (c)

Figure 6.2: Illustration of the slope correction principle (a). The original illumination resulting in a narrowed groove (top) is reduced to balance both grating levels (bottom). Comparison of the efficiencies of the of the original (b) and the corrected (c) gratings for different slope area to period ratios.

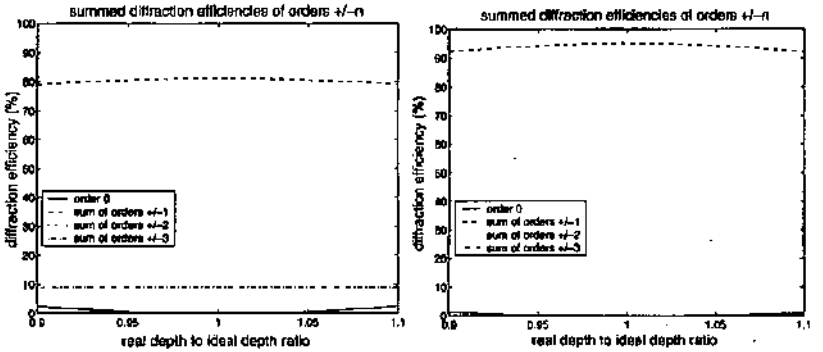
6.2 Fabrication errors

Fabrication errors have extensively been studied for focusing lenses and grating elements [135, 136]. However, for far-field beam shaping the aspect of fabrication errors has not received much interest until now. We shall present here the characteristic effects of some of the errors presented in chapter 4.

6.2.1 Grating profile errors

As we have seen in chapter 4, the lithographic process may create a line-width error and a profile slope error, both of them can be partly corrected during the etching step. Additionally, one can also reduce their influence in the design stage if they are reproducible. Under this condition, one can introduce a precompensation of the error, by making the transparent apertures of the mask wider in order to achieve the originally desired width at half height instead of the top of the fabricated structure [136]. The effect of such a correction is demonstrated in Fig. 6.2. While, for a slope area of 10% of the period, the original element signal orders have decreased from 81% to 72% and the zero order has reached 5.6%, the corrected element signal stays nearly stable at 79.5%, and the zero order is limited to 1.6% of the total signal energy.

For line-width and slope errors which are constant over the entire DOE, one can expect that the effect is higher for smaller grating periods. Thus the resulting pattern shares the symmetries of the original DOE. For a re-



(a) (b)

Figure 6.3: Effect of etching depth error for a binary grating (a) and an eight-level grating (b).

mapping element based on a Fresnel zone plate, the error exhibits a symmetry of rotation.

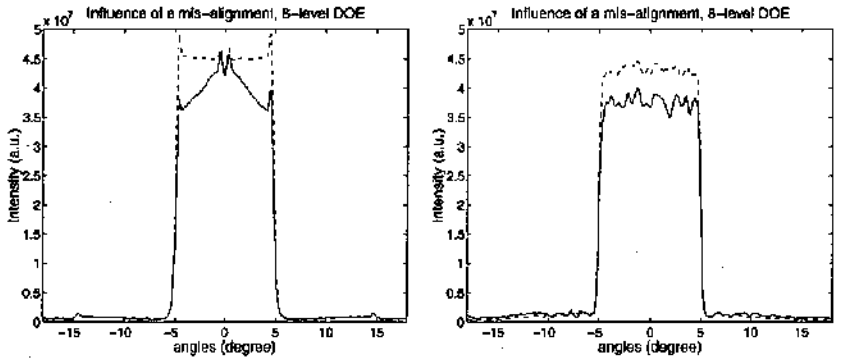
6.2.2 Etching depth errors

Refractive elements

Contrarily to line-width, slope or alignment errors, etching depth errors affect both refractive (ROE) and diffractive (DOE) elements. Due to the large sensitivity of the refractive elements to the profile height, beam-shaping with such elements requires a high degree of control during the reactive ion etching (RIE) stage. Besides the profile height, also the selectivity of the process and its changes must be mastered during the whole process. As previously stated, etching depth errors will result in a wrong angular dimension of the far-field light distribution. Errors in the selectivity control will result in a distortion of the phase profile, thus in an inaccuracy of the intensity profile.

Diffractive elements

Etching depth errors are typically around 5% for most DOEs. The resulting effect is illustrated in Fig. 6.3. For the binary grating of Fig. 6.3(a) and for an etching depth error of 10%, the main effect is the transfer of energy from the signal orders (decrease from 81% to 79%) to the zero order (increase from 0% to 2.5%). However, for an error of 5%, the zero order has only raised to 0.6%.



(a) (b)

Figure 6.4: Simulation of one-dimensional eight level re-mapping type (a) and grating type (b) elements with (solid line) of without (dashed line) misalignment of the intermediate mask.

For an eight-level grating, the situation is more complicated depending on the way the three successive errors are composed. For the case that the errors are systematically in the same direction, and proportional to the ideal depth of the etching for all three processes, the resulting efficiencies are plotted in Fig. 6.3(b). For an error of 10%, the loss of efficiency in the signal orders is still small (decrease from 95% to 92%) and the energy transferred to the zero order only 1%. For an error of 5%, the zero order contains only 0.3% of the energy. This shows that for relatively easily achievable tolerances, the etching depth is not the most serious issue.

An improvement has been proposed by Kettunen to lower the sensitivity of diffractive structures to etching depth errors [137]. The principle is to use a three-level optical element. However the phase is designed to be binary, and it requires one additional mask when compared with the original process. This will in turn introduce other errors, like mis-alignment of the masks, which makes this improvement preferable for direct-writing technologies.

6.2.3 Alignment error

Misalignment of the masks during the exposure step in the lithography process is known to be the main source of errors for multilevel lenses in focusing applications [138–142]. We have studied the effect of this error for both re-mapping type and grating type DOEs [143], as illustrated in Fig. 6.4. From section 6.1,

in the first case, we should expect a decrease of the efficiency which is higher for the parts of the phase zone plate (PZP) where the grating period is smaller. Moreover, the geometrical behaviour of the alignment errors is different from the profile errors. Indeed, misalignment is possible in two dimensions, but it exists a direction of maximum and a direction of zero error for each mask. Thus the effect should be a deformation of a flat-top distribution into a roof-like distribution, whose ridge is perpendicular to the direction of maximum misalignment, as shown in Fig. 6.5. The combination of the misalignment of several masks is difficult to predict, as the directions of the ridges are random for each process.

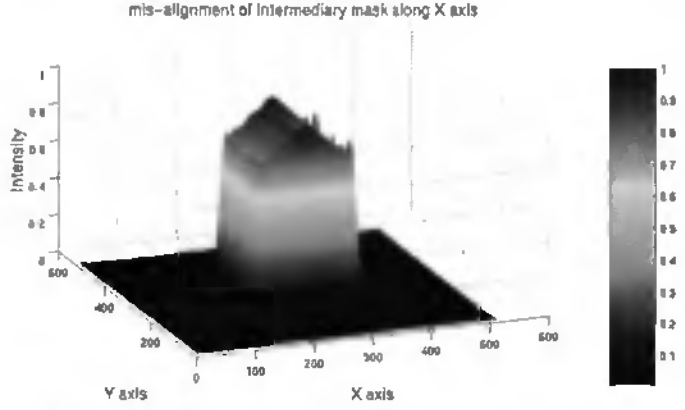


Figure 6.5: Effect of a misalignment of the intermediary mask of a two-dimensional Fresnel zone plate.

6.3 Influence of the diffraction

This error is only present for DOEs that have been designed in the geometrical optics approximation, i.e. with the design methods presented in section 5.1.

6.3.1 Error characteristics

The diffraction of light at the edge of the elementary cell of the DOE results in oscillations located all over the far-field light distribution. The number of oscillations is given by the Fresnel number

$$N_F = \frac{2\alpha a}{\lambda}, \tag{6.4}$$

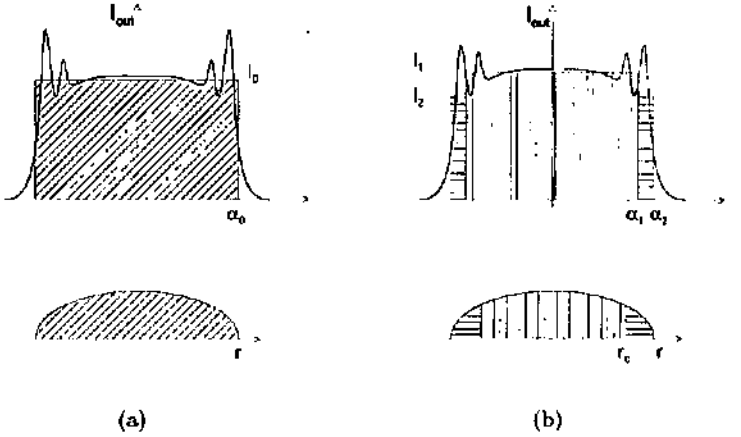


Figure 6.6: Reduction of the diffraction oscillations. The flat-top geometrical optics design (dashed) results in strong oscillations (a), the energy is redistributed from the sides of the signal to the central part (b). At the bottom: the re-mapping element.

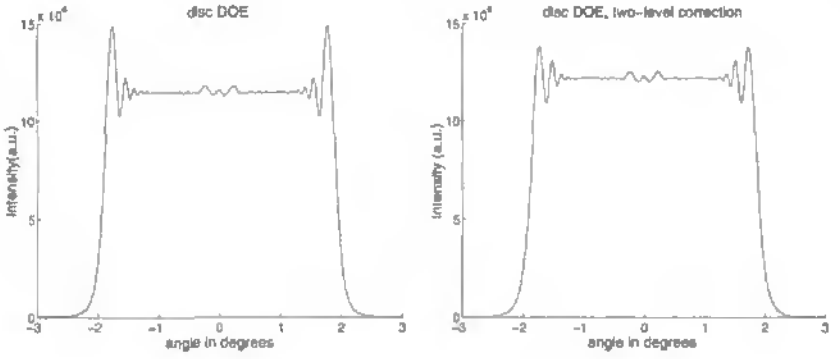
where α is the far-field angle and a the cell dimension. Kopp has proposed an analysis of the oscillations for a quadratic phase in a separable geometry [58]. The light intensity has extrema located at

$$u_j = \frac{\alpha_j}{\lambda} = \frac{j}{a}, \quad j \in \mathbb{Z}. \tag{6.5}$$

For more general cases, as seen from previous examples in chapters 3 and 5, the strongest oscillations are always located at the angles corresponding to the edges of the cell. The exact positions of the extrema cannot be determined analytically.

6.3.2 Correction schemes

The Gibbs phenomenon is usually corrected by apodisation [94], but this solution involves absorption of light, that is not easy to fabricate neither acceptable for thermal reasons in our application. Locally controlled fabrication errors such as line width and etching depth can be used to generate a local loss of efficiency. This allows to simulate an apodisation without absorption in the material [71,103]. However, both solutions are lossy and it is usually preferable to redistribute the light in the signal and to avoid increasing the noise. We propose here a simple corrective scheme that reduces the importance of Gibbs oscillations while maintaining the energy in the signal orders. We will restrict



(a) (b)

Figure 6.7: Illustration of the reduction of the side oscillations of a flat-top beam-shaping element (a) by a design based on two intensity levels (b).

the scope of our demonstration to rotationally symmetrical two-dimensional problems.

The iden behind this correction scheme is explained in Fig. 6.6. The geometrical optics predict a binary light distribution and the diffraction at the edge of the cell introduces strong over-shots at the border, as in Fig. 6.7(a). In order to reduce these over-shots, we intend to design an element which generates a far-field distribution in two different levels, as illustrated in Fig. 6.7(b). To calculate such a phase function, we divide the optical element into two concentric areas (vertical and horizontal dashes). The radius of the contact r_C between the two sub-elements in chosen so that the energy conservation implies that the central far-field area (vertical dashes) will have a higher intensity level than the border area. We then use Eqs. (5.36) and (5.37) to compute the profile of the two sub-elements, and choose the phase offsets in order to ensure the continuity of the phase at the connection. As a consequence, some energy is transferred from the side peaks to the central part, increasing the intensity level in the far-field. An example of this correction scheme is shown in Fig. 6.7.

One can easily imagine to extend this process to more than two intensity levels, aiming to reduce the Gibbs phenomenon even further. However, it has to be noted that this correction is difficult to apply if the DOE is not rotationally symmetrical or one-dimensional. This limitation makes the use of locally controlled errors more suitable for two-dimensional structures with

no symmetries, such as hexagonal or square arrays of lenses used in aperture modulated diffusers (AMDs).

6.4 Influence of the encoding

For multilevel structures, the decomposition of the phase onto a discrete number of values is responsible for many phenomena in the resulting light distribution. It is thus necessary to review the role of the various parameters influencing the encoding in order to choose the most suited scheme for our application.

6.4.1 Quantisation of the phase

The multilevel approximation of a continuous phase is responsible for strong perturbations in the far-field profile, as illustrated by Fig. 3.7. These oscillations are different from the ones caused by diffraction. Their geometry is given by the encoding, not by the aperture. They are stronger when the ideal phase is not well approximated by the multilevel profile, for instance in the areas of slow phase variations. For a Fresnel lens, they are circular and mostly visible at the centre. Diffraction related oscillations, on the other hand, are following the shape of the cells and are located at the edges. This difference is clearly visible in Fig. 3.21 for both the rectangle and hexagonal cells.

The oscillations due to quantisation are not influenced by the diameter of the lens, as seen in Fig. 6.8. The lines of the image represent the radial profile of the far-field light distribution. When the lens size is increased, for instance by opening progressively a diaphragm in the lens plane, the angular aperture increases, but the position of the maxima and minima of the intensity are preserved. The diffraction oscillations are located at the edge of the far-field and do not perturb the quantisation oscillations either.

6.4.2 The phase offset

As seen in chapter 5, in both cases, diffractive and refractive, a constant phase offset does not alter the functionality of the elements. However, this degree of freedom becomes interesting to use when the profile is approximated by a multilevel structure. The phase offset does change the positions of the steps (the transition points), as seen in Fig. 6.9.

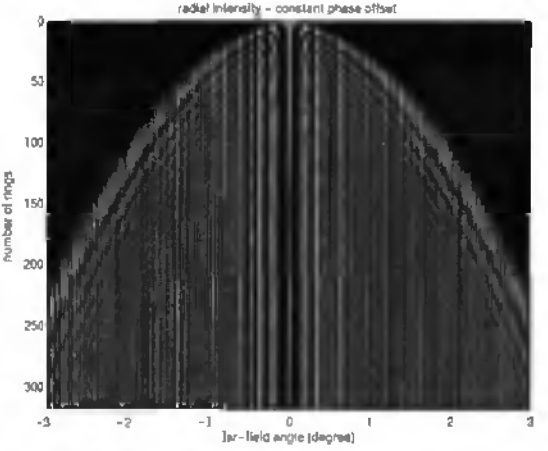
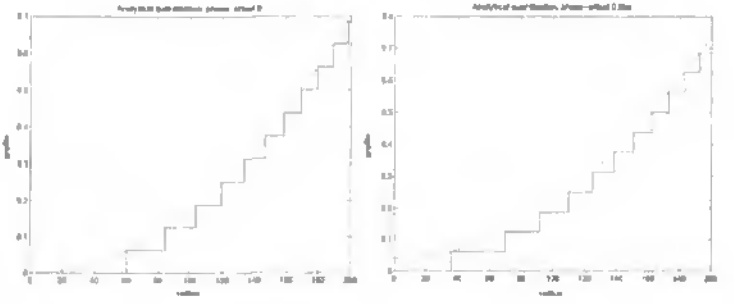


Figure 6.8: Influence of the lens size, expressed in number of rings, on the quantisation oscillations for a binary lens.

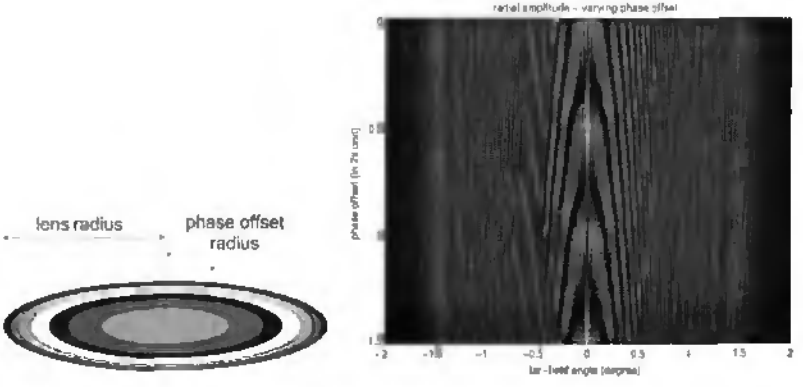


(a) (b)

Figure 6.9: The phase offset defines a translation of the profile with respect to the quantisation phase level. Zero phase offset (a) and $0.04 \times 2\pi$ phase offset (b).

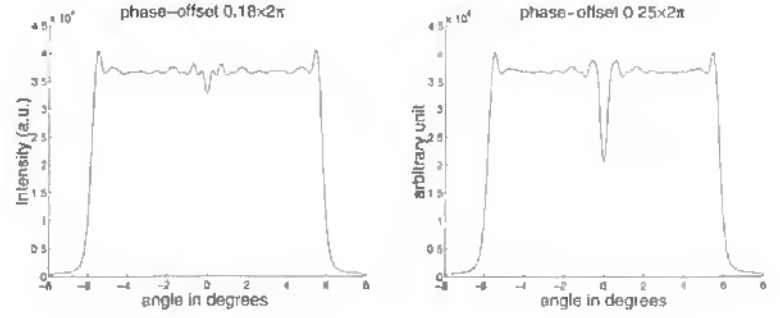
The phase offset influences two features of the light distribution. We outline these two properties in Fig. 6.10(b). The lines of the image present the radial light distribution due to a binary circular Fresnel zone plate. The phase offset is increased from 0 to 3π. The main observation is that the angular positions of the minima and maxima of light are influenced by the phase offset, in particular at the small angles. The phase offset in thus an efficient parameter to control the presence of a hole or a bump of light at the centre of the far-field [57, 97, 144]. In Fig. 6.10(b), the period of this pattern is π. Indeed, for

a binary lens, such a phase offset corresponds to suppressing one ring. The new lens has the same transition points than the original lens, but the inverse phase. Similarly, the period is $\pi/2N$ for a N-level lens. Examples of use of this degree of freedom are shown in Fig. 6.11 or in Fig. 3.7.



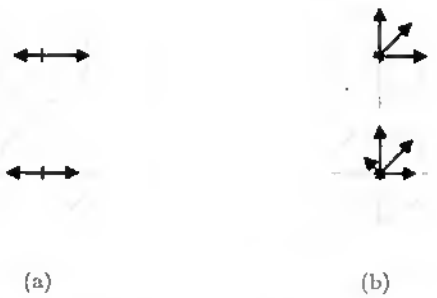
(a) (b)

Figure 6.10: The phase offset is changing the radii of the Fresnel zones, introducing an additional degree of freedom (a). Influence of the phase offset on the modulus of the amplitude of the far-field light distribution for a binary lens (b).



(a) (b)

Figure 6.11: Simulated light distributions for an incoherent divergent beam for two different hexagonal flat-top beam-shaping DOE. with phase offsets $0.18 \times 2\pi$ (a) and $0.25 \times 2\pi$ (b).



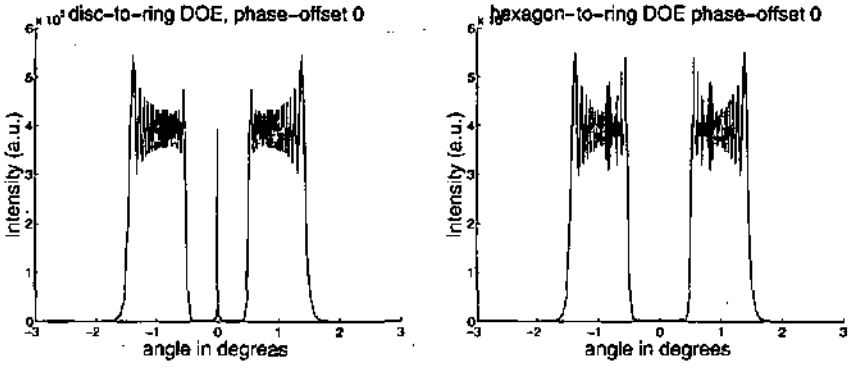
(a) (b)

Figure 6.12: For a binary lens, the phase offset can cancel the zero order value by balancing the phasors of opposite angle (a). For more phase levels, the levels that are increased do in general not compensate the mean value of the phase of the lens (b). Top row is without phase offset, bottom row with an arbitrary phase offset.

Secondly, one can also see that the zero order has a period of 2π. Its value is zero when there is a balance between the number of rings of zero and π phase, which equals to suppressing two rings for a binary lens and N for a N-level lens, etc. The period of the zero order pattern is thus always 2π. [97]

It should be noted, however, that the capability to control the zero order peak is related to the number of rings in the Fresnel zone plate. For a quadratic phase, as all the rings have the same area, they contribute equally to the peak. For a N-level DOE, the contribution of a whole period, composed of N rings, is zero. The phase offset can only be used to modify the contributions of the first and last rings of the lens. This implies that, as shown in Fig. 6.12(a) it is, for a binary lens, always possible to balance the two extreme phasors so that the zero order gets zero. However, if the lens contains more levels, as illustrated in Fig. 6.12(b), the phase offset is not enough to suppress the zero order. The number of rings as to be carefully chosen so that it is a multiple of the number of levels.

Nevertheless, when the cell is not rotationally symmetrical, the central Fourier coefficient tends to be reduced. Indeed, as the corners of the cells contain pieces of rings with all phase levels, one can intuitively understand that they statistically compensate each other, and that the last sectors are so small that they will not generate an important peak. The effect is illustrated in Fig. 6.13 by the comparison of two different ring beam-shaping elements. The disc-to-ring beam-shaping element cannot be adjusted enough to suppress the zero-frequency coefficient, while the hexagon-to-ring DOE gives already a small peak without special effort.



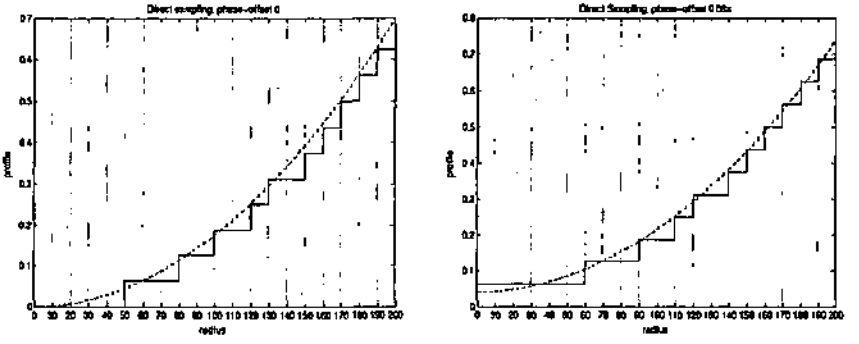
(a) (b)

Figure 6.13: Comparison of the central peak in Fourier space for a disc-to-ring (a) and hexagon-to-ring (b) elements. The DOEs are 8-level lenses, 2 mm wide, designed to generate a ring between 0.5° and 1.5°.

6.4.3 Lens encoding

Until now, we have presented and analysed the results of the so called analytical quantisation (AQ). However, this is not the only available quantisation scheme. Direct sampling (DS) was proposed by Kuittinen as an optimum version of the radially symmetrical iterative discrete on-axis encoding (RSIDO) [145, 146]. Direct sampling can work around the limitations due to the minimum feature size and gives higher efficiencies for lenses with large numerical apertures. Figure 6.14(a) illustrates the principle of direct sampling. A fixed pitch defines a set of sampling points. The phase value of the continuous profile between such points is approximated by the closest phase level available. Figure 6.14(b) shows how the phase offset modifies the encoded profile in the case of the direct sampling.

However, when the rings are sufficiently wide, the fixed pitch of the DS results in a loss of freedom in the choice of the ring radii. To outline this drawback, we have simulated an eight-level disc-to-ring beam-shaping element spreading the light between 0.5° and 1.5°. The DOE is supposed to be illuminated by an incoherent beam with 1 mrad divergence, which smoothes the rapid oscillations in the far-field. This element, encoded with analytical quantisation, requires a minimum feature size (mfs) of 1.16 μm and produces a far-field presented in Fig. 6.15(a). When encoded with direct sampling of pitch 1 μm, the far-field distribution is less uniform and the noise inside the

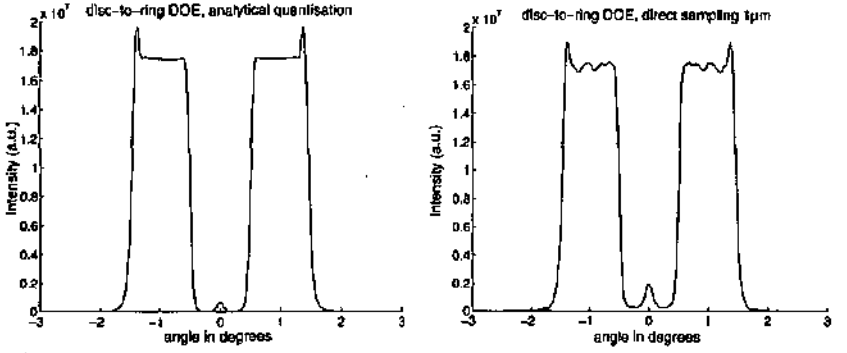


(a) (b)

Figure 6.14: Direct sampling for the same lens as in Fig. 6.9 with phase offset zero in (a) and $0.04 \times 2\pi$ in (b).

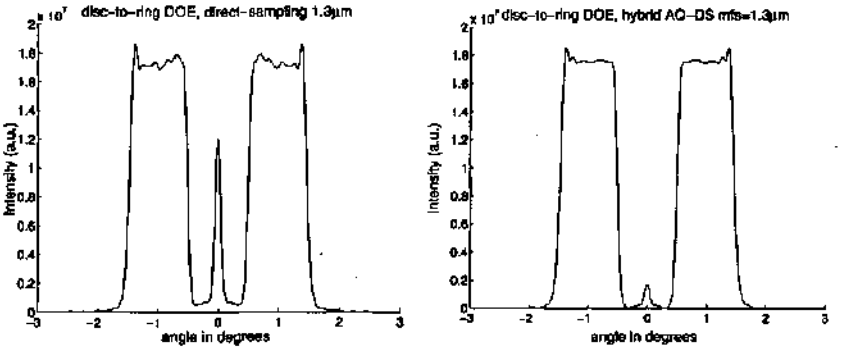
ring is increased, as seen in Fig. 6.15(b). The direct sampling already shows its limits although the pitch is smaller than the minimum feature size for the analytical quantisation. Therefore, direct sampling should be used only when analytical quantisation requires feature sizes smaller than offered by the fabrication process. Even in this case, it is preferable to use the direct sampling only for those parts of the element where it is strictly mandatory. Encoding the low frequencies areas using analytical quantisation results in a appreciable gain of uniformity and a reduced noise in the centre of the ring, as illustrated by Fig. 6.16. We assumed a minimum feature size of $1.3\,\mu$m, which is too large for AQ in the high frequency part of the element. The noise produced by a purely direct-sampled solution is considerably higher than for an hybrid encoding.

This weakness of the direct-sampling encoding in terms of noise and uniformity compared to analytical quantisation can be understood by the constraint of equal pitch. This constraint ensures that no feature will be smaller than a minimum threshold, but goes far beyond. Analytical quantisation, on the other side, cannot enforce the feature size constraint, but allows for more freedom in the value of the radii.



(a) (b)

Figure 6.15: Comparison of a disc-to-ring beam-shaping element encoded with analytical quantisation (a) of mfs = 1.16 μm and with direct sampling of pitch 1 μm (b).



(a) (b)

Figure 6.16: Comparison of a disc-to-ring beam-shaping element encoded with Direct Sampling of pitch 1.3 μm (a), and hybrid element with mfs = 1.3 μm (b).

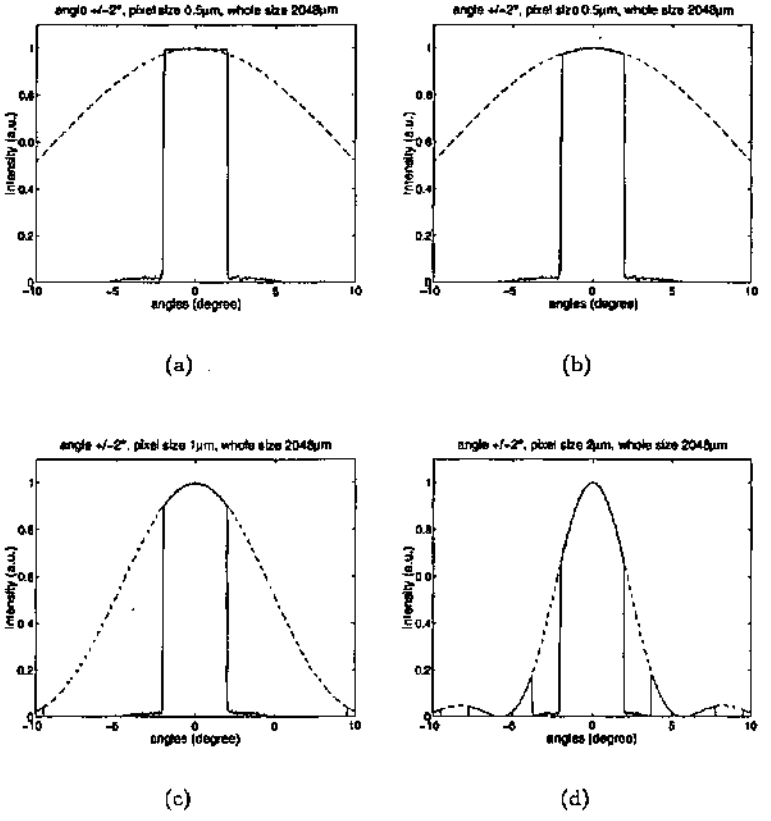
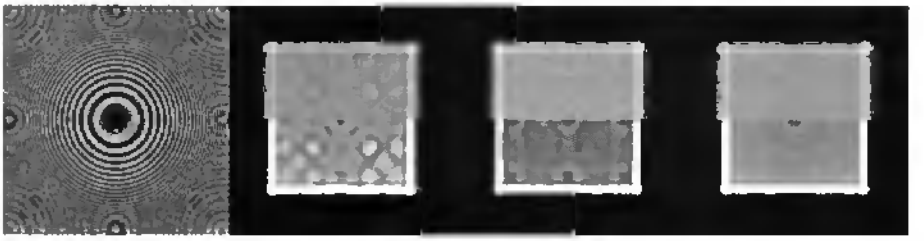


Figure 6.17: Effect of tessellation on the far-field light distribution of a one-dimensional CGH (solid line). No effect is predicted if the resolution of the simulation is 1 point per pixel (a), but for a higher resolution (b), the signal is degraded by an envelope that corresponds to the sinc square due to the diffraction of one pixel (dashed line). When the pixel size is increased, the distortion gets stronger (c) and (d).

6.4.4 Spatial quantisation

In chapter 3, we have mentioned that the use of pixels (tessellation) to fabricate the structure of the optical elements introduces artifacts. The first one is the presence of a sinc square curve that distorts the desired far-field, as illustrated in Fig. 6.17. The element is a one-dimensional CCH, designed by IFTA to generate a flat-top distribution of angle $\pm 2°$ for a wavelength of 200 nm. If the pixel size is not taken into account, i.e. if the simulation is performed with a



(a) (b) (c) (d)

Figure 6.18: Effect of the tessellation for a 256 × 256 pixels 8-level lens (a) and the resulting far-field (b). Far-fields of similar lenses with 512 × 512 (c) and 1024 × 1024 (d) pixels. The top half of the images (b)-(d) is the intensity, the bottom half is a contrast enhanced version of the intensity.

resolution of one sample point per pixel, the light distribution corresponds to the desired goal, as shown in Fig. 6.17(a). However, when the resolution of the simulation is increased to sixteen points per pixel, the profile is distorted by an envelope, as shown in Fig. 6.17(b). As a consequence, the efficiency is reduced. The envelope corresponds to the theoretical sinc square curve due to the diffraction by one pixel, drawn as dashed lines in Fig. 6.17. Indeed, the pixel realises a convolution of the sampled representation of the optical element by a rectangular function. In Fourier space, this corresponds to the multiplication of the ideal distribution by the Fourier transform of the pixel. As this error is known from the size of the pixel, it is possible to take the envelope directly into account in the IFT algorithm [40]. However, this will correct the distortion, but not the overall loss of efficiency. For re-mapping elements, the effect of the pixel cannot be easily taken into account in the design formulas.

Moreover, a second artifact is observed for re-mapping elements, due to a moiré effect between the pixel grid and the local grating. This results in faint parasitic lenses structures, as seen in the corners of Fig. 6.18(a). These structures perturb the beam-shaping effect, as shown in Fig. 6.18(b). To reduce their influence, it is necessary to use smaller pixels, as presented in Figs.6.18(c) and 6.18(d). The effect of such moiré artifacts in the neighbourhood of the focal plane has been studied by Carcolé *et al.* [42–45]. Carcolé observed secondary off-axis foci. Such tessellated structures can be used for the design of spot array generators [147].

These two drawbacks of the tessellated structures suggest the use of smooth or polygonal descriptions. Kallioniemi *et al.* have studied the effect of the ap-

proximation of rings by polygons and shown which defects can be expected depending on the resolution of the polygons [54]. We presented in chapter 3 a possible method to work around the tessellation by refining the two-dimensional sampling by smooth curves. Finally, it is also possible to use smaller pixels to reduce the pixelation influence. It is known that when the pixel size is smaller than half of wavelength, the light does not "see" the pixels any more [59].

6.5 Beam size and spatial invariance

By design, remapping-type and para-geometric elements are fundamentally spatially variant and so is the input beam. To achieve spatial invariance of the beam-shaping device with respect to the input beam, beam-shaping elements are built from small individual optical elements tiled into arrays. The far-field distribution is the sum of all the light distributions generated from the individual cells, thus realising the average of all cells. To evaluate the number of cells required, we studied the repartition of light created by a Gaussian beam over a hexagonal array. The light distributions from each hexagon are then summed. The result is shown in Fig. 6.19. The illumination is homogeneously spread (in a 10% range) over the hexagon geometry as soon as the width of the Gaussian beam is greater than 1.3 times the outer diameter of the hexagonal cell.

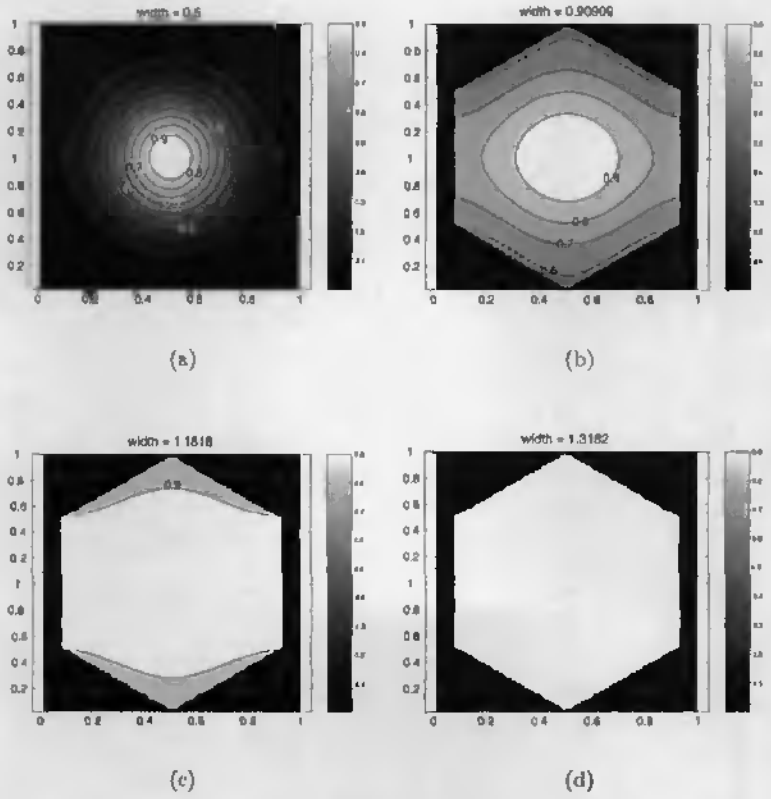


Figure 6.19: Uniformity of the light distribution from a hexagonal array illuminated by a Gaussian beam whose width (at 1/e) is 0.5 (a), 0.9 (b), 1.2 (c) and 1.3 (d) times the outer diameter of the hexagon.

Conclusion 7

The present work has tackled the design and characterisation of far-field beam-shaping elements used for the patterning of illumination in lithographic systems. This work adds up to an already extensive literature on the subject [48,51,97,144,148]. However, new concepts and ideas have been presented.

Firstly, the capability to simulate large structures with low memory requirements has been introduced. This has allowed to study real two-dimensional beam-shaping elements whose dimensions are commonly thousands of times the wavelength of the light. Based on this new capacity, the influence of diffraction has be studied.

Secondly, the design of re-mapping type elements has been revised. New formulas have been proposed for specific light distributions, and an improvement has been demonstrated for the iterative finite-element mesh-adaptation algorithm [6].

Finally, a comprehensive study of the influences of the fabrication errors for far-field beam-shaping elements has been conducted. The conclusions drawn from simulations have been confirmed by measurements realised on fabricated elements. The differences of behaviour between diffusing grating-type elements, re-mapping type elements and para-geometric grating-type elements have been emphasised. This study has led to propose a possible improvement for rotation-symmetrical phase zone plates and to identify the tolerances of the various element types.

Acknowledgments

This work would not have been achieved without the help and support of many people. I would like to thank

- Professor René Dändliker, my director of thesis, for having given me the opportunity to realise this work in the Applied Optics group of the IMT.
- Professor Hans Peter Herzig, who gave me, five years ago, the opportunity to make my national service as a trainee in the Applied Optics group. He directly supervised my work all over these years and has spent much energy supporting me.
- The other members of the jury: Professor Markku Kuittinen, Doctor Manfred Maul and Doctor Ville Kettunen, for the many advice and suggestions they have given me.
- The two companies which have provided financial support for this work, namely CSEM for the first two years and Colibrys SA for the last two years.
- The people outside the IMT who have provided data and advice, namely Dr. Philippe Regnault, Kirstin Becker and Nicolas Faure.
- The post-doctoral fellows who have supervised my work at the IMT, Dr. Peter Kipfer and Dr. Ville Kettunen. They have being valuable sources of information about optics.
- The IMT members whom I worked with on the project, Philippe Nussbaum, Irène Philipoussis and Joëlle Vuille.
- The post-doctoral fellows who, while not supervising my thesis, have provided me with suggestions and information, Dr. Martin Salt and Dr. Manuel Flury.

- Dr. Ken Weible and Dr. Andreas Schilling for the many answers they have kindly provided to my questions.
- The fellows which have shared their office with me, Dr. Antonello Nesci, Dr. Yves-Alain Peter, Dr. Frédéric Gonté, Stéphanie Clément and Patrick Ruffieux.
- The fellows which have shared their laboratory space with me, Dr. Karim Haroud, Dr. Omar Manzardo and Iwan Maerki.
- The administrative and technical staff of the IMT. I would like to express very special thanks to the secretaries Mary-Claude Gauteaub, Sandrine Piffaretti, Joëlle Banjac and Sylvie Baillod, to our administrator Martial Racine, and to our system administrator and electronics technician Marcel Groccia.
- Olivier Scherler for having been my student and my successor in webmaster functions.
- The former members of the Applied Optics group, with special thanks to Dr. Étienne Rochat, Dr. Yves Salvadé, Dr. Manuel Bouvier, Dr. Manuel Teijido, Dr. Peter Blattner and Dr. Jean-Christophe Roulet.
- The present members of the Applied Optics group for their help and the pleasant work ambiance, the beers and the technical help.
- The members of other IMT groups who have provided technical support or just agreeable presence namely Christian Robert, Patrick Stadelmann, Laura Ceriotti and Sylviane Pochon.
- The many many people I have been lucky to count as friends, from the Swiss, French, Spanish communities, or from other nationalities. Their support has been a *sine qua non* condition for the success of this work.
- My family in France, whose weekly emails and phone calls have been a very precious support in these last five years.

Bézier curves

A

Bézier curves were invented in the 1960's to fulfill the needs of the automotive industry, providing curves and surfaces versatile enough for car designers. Pierre Bézier, engineer at Renault and Paul de Casteljau, engineer at Citroën, are usually both credited to have invented independently, at the same time, the bases of the Bézier curves. From the car-body design, the use of this tool has spread in computer graphics, where they are encountered in fonts, two and three-dimensional drawing softwares and interpolation techniques.

Definition and construction of Bézier curves

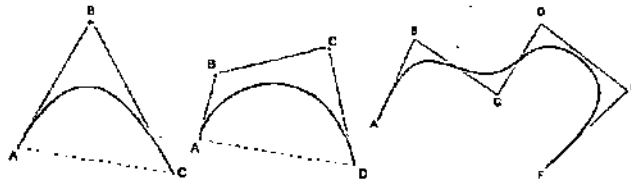
A Bézier curve of order n is defined by $n+1$ control points $\mathrm{A}_0, \ldots, \mathrm{A}_n$. Given a parameter t, usually $0 \leqslant t \leqslant 1$, a point $\mathrm{P}(t)$ on the curve is given by

$$\mathrm{P}(t) = \sum_{i=0}^{n} \mathrm{A}_i B_{i,n}(t), \tag{A.1}$$

where $B_{i,n}(t)$, the Bernstein polynomials, are defined as

$$B_{i,n}(t) = \frac{n!}{i!\,(n-i)!} t^i (1-t)^{n-i}. \tag{A.2}$$

Figure A.1 shows examples of Bézier curves. As can be seen, a Bézier curve passes through its first and last control points (for $t = 0$ and $t = 1$, respectively), but does not pass through the other control points. Figure A.1(a) shows the most simple Bézier curve, the quadratic one. This curve is used for instance in true type fonts (TTF), invented by Apple and Microsoft as a cheap alternative to Adobe's postscript fonts, based on the cubic Bézier curves of



(a) (b) (c)

Figure A.1: Quadratic (a), cubic (b) or higher degree (c) Bézier curves are defined by their control polygon.

Fig. A.1(b). Metafonts of Donald Knuth are also based on the cubic Bézier curves. The cubic Bézier curve definition may be expressed in matrix form

$$\mathrm{P}(t) = \begin{bmatrix} \mathrm{P}_0 & \mathrm{P}_1 & \mathrm{P}_2 & \mathrm{P}_3 \end{bmatrix} \begin{bmatrix} 1 & -3 & 3 & -1 \\ 0 & 3 & -6 & 3 \\ 0 & 0 & 3 & -3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ t \\ t^2 \\ t^3 \end{bmatrix}. \tag{A.3}$$

Another equivalent definition of the Bézier curves is given by the theorem of de Casteljau, describing the geometric construction of the point $\mathrm{P}(t)$. This construction is illustrated in Fig. A.2 for a cubic Bézier curve and the point $\mathrm{P}(t) = 2/3$. From the four control points A, B, C and D, the three points I, J and K are built with AI = 2/3AB, BJ = 2/3BC and CK = 2/3CD. Then the points M and N are built with IM = 2/3IJ and JN = 2/3JK. Finally, O = $\mathrm{P}(t = 2/3)$ is obtained on the segment MN with MO = 2/3MN. Mathematically, this algorithm is expressed as

$$\mathrm{P}(t) = \mathrm{P}_n^{(n)}(t), \tag{A.4}$$

where

$$\mathrm{P}_i^{(j)}(t) = \begin{cases} (1-t)\,\mathrm{P}_{i-1}^{(j-1)}(t) + t\mathrm{P}_i^{(j-1)}(t) & \text{if } j > 0, \\ \mathrm{P}_i & \text{otherwise.} \end{cases} \tag{A.5}$$

All Bézier curves share many interesting properties:

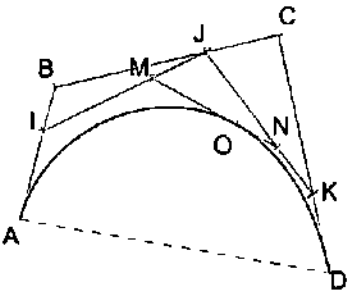


Figure A.2: A geometric construction of a Bézier curve can be obtained via de Casteljau algorithm, illustrated here for the point $P(t) = \frac{2}{3}$ of a cubic Bézier curve.

1. They are invariant by affine transformations (rotation, scaling, translation, shearing or parallel projection). This means that the knowledge of the control points is sufficient in order to perform such transformations.

2. They are always contained in the convex hull of the control polygon.

3. A Bézier curve of order n has at most $n-2$ inflection points.

4. At A_0, the Bézier curve is tangent to $\mathbf{A_0A_1}$; at A_n, the curve is tangent to $\mathbf{A_{n-1}A_n}$.

5. De Casteljau algorithm also divides a Bézier curve into two adjacent Bézier curves. In Fig. A.2, the two resulting curves are defined by the points (A,I,M,O) and (O,N,K,D).

Reference [149] contains an in depth presentation of the Bézier curves and the splines.

Bézier curves, splines and data interpolation

Because a change of one control point will influence the whole Bézier curve, only quadratic or cubic Bézier curves are used in practice. For longer lines, quadratic or cubic Bézier curves can be linked piecewise together. These piecewise curves, that we have used in our applications, are a subset B-splines which in turn are a subset of splines. The concept of spline curves were originally studied in the XIXth century with wooden or metallic splines constrainted to pass by given points by "ducks" (metal weights). The elastic nature of the spline was making the curve smooth.

It is possible to convert B-splines to piecewise cubic Bézier curves thanks to the "knot insertion" algorithm [149]. Splines are commonly used for interpolation and may be usefull to infer smooth curves from isolated points. In the same field of application, there exist algorithms to extract contours from black and white images [150] or to interpolate level curves from continuous profiles [151] which may be used to refiné tessellated structures, as shown in Fig. 3.15.

The complex error function

B

The error function $\mathrm{erf}(x)$ can be extended to the complex plane ($z = x + iy$)

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left(-t^2\right) \mathrm{d}t. \tag{B.1}$$

From there, we can define the complex error function $w(z)$, also called Faddeeva function or plasma dispersion function, as

$$w(z) = \exp\left(-z^2\right) \mathrm{erfc}(-iz) = \exp\left(-z^2\right) \cdot \left(1 + \frac{2i}{\sqrt{\pi}} \int_0^z \exp\left(t^2\right) \mathrm{d}t\right). \tag{B.2}$$

It has the interesting property

$$w(-z) = 2\exp\left(-z^2\right) - w(z), \tag{B.3}$$

which allows to determine its value on one half of the complex plane from the knowledge of its value on the other half. Hence, Eq. (3.18) is reduced to the computation of

$$\int_0^z \exp\left(t^2\right) \mathrm{d}t = i\frac{\sqrt{\pi}}{2} \left(1 - w(z) \cdot \exp\left(z^2\right)\right). \tag{B.4}$$

This equation can be either computed from tabulated polynomial approximations [152], or by using an algorithm recently proposed by Weideman [153]. This algorithm is only defined on half of the complex plane, thus the interest of Eq. (B.3). Its precision can be tuned by a parameter $N = 2^p$ and it can achieve accurate results in a short time. Moreover, the implementation given by Weideman is only eight lines in Matlab, and not much more is needed to extend it to the whole complex plane.

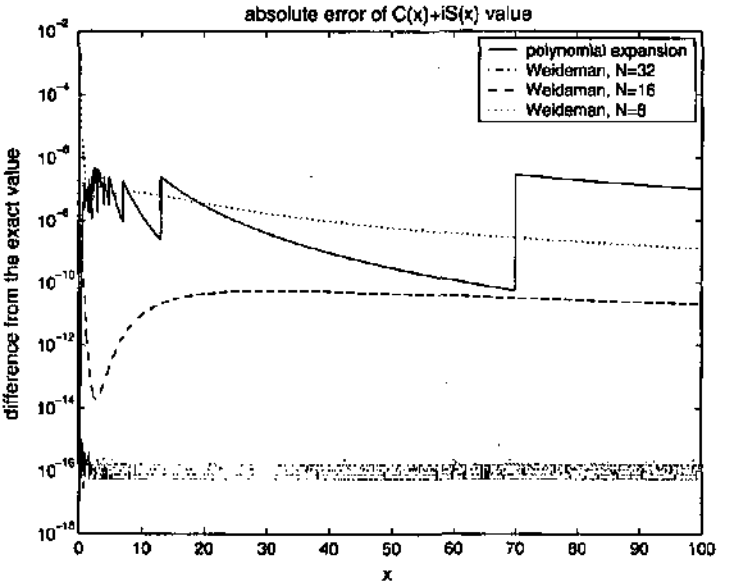


Figure B.1: Comparison of the absolute error for the computation of $C(x) + iS(x)$using the polynomial expansion or the Weideman complex error function algorithm .

This complex error function $w(z)$ is related to the Fresnel integrals $C(z)$ and $S(z)$ by

$$C(z) + iS(z) = \frac{1+i}{2}\left(1 - w\left(\frac{1+i}{2}\sqrt{\pi}z\right)\cdot \exp\left(i\frac{\pi}{2}z^2\right)\right), \qquad (B.5)$$

which gives a possibility to compare both methods.

This comparison is illustrated by Fig. B.1. For $N \geq 16$, the accuracy is drastically increased by using the Weideman algorithm in most parts of the complex plane, especially far from zero. For $N = 32$, the error is so small that it reaches the limits of precision of Matlab, as can be seen from the typical numerical quantised noise pattern. However, we can see that for small values of x, the algorithm has the lowest precision and may be beaten by the polynomial expansion. Table B.1 gives the values of the absolute error at $x = 0$ for different values of the algorithm precision parameter.

For $N \geq 16$, the absolute error of the Weideman algorithm is worse than the polynomial expansion for $x \leq x_\epsilon$. From Eq. (3.22), this implies the algorithm is better as soon as

$$a - y' \geqslant \sqrt{\lambda z}\cdot x_\epsilon. \qquad (B.6)$$

N	error at $x = 0$	x_ϵ	$\frac{time_{polynom}}{time_{algorithm}}$
8	$1.5 \cdot 10^{-4}$	NC	5.5
16	$1.25 \cdot 10^{-7}$	0.523	4.5
32	$3 \cdot 10^{-14}$	0.235	3.3
64	$2 \cdot 10^{-16}$	NC	2.3

Table B.1: Comparison of between the polynomial expansion and the Weideman algorithm for various values of the parameter N. x_ϵ is the value for x where the accuracy of both formulas is equal.

For a wavelength of 633 nm, a parameter $N = 16$ and a distance $z = 150$ mm, Eq. (B.6) yields about 150 μm. Hence, any element whose size is larger than this value will benefit from the Weideman algorithm in terms of accuracy.

Bibliography

[1] J. Finders, J. van Schoot, M. Mulder, A. Hunter, M. Dusa, B. Socha, and P. Jenkins. DUV lithography (KrF) for 130 nm using off-xis illumination and assisting features. In M. Messe, editor, *Semicon Japan 99*, Chiba, Japan, 1999.

[2] A. Kwot-Kit Wong. *Resolution Enhancement Techniques in Optical Lithography*, volume TT47 of *Tutorial texts in Optical Engineering*. SPIE Press, Bellingham, 2001.

[3] S. Kawata, I. Hikima, Y. Ichihara, and S. Watanabe. Spatial coherence of KrF excimer lasers. *Applied Optics*, 31(3):387–396, 1992.

[4] Y. Lin and J. Buck. Numerical Modeling of the Excimer Beam. In B. Singh, editor, *Metrology, Inspection, and Process Control for Microlithography XIII*, volume 3677, pages 700–710. SPIE, 1999.

[5] Y. Lin, G. N. Lawrence, and J. Buck. Characterization of excimer lasers for application to lenslet array homogenizers. *Applied Optics*, 40(12):1931–1941, 2001.

[6] T. Dresel, M. Beyerlein, and J. Schwider. Design of computer-generated beam-shaping holograms by iterative finite-element mesh adaptation. *Applied Optics*, 35(35):6865–6874, 1996.

[7] D. Maystre. Integral Methods. In R. Petit, editor, *Electromagnetic Theory of Gratings*, pages 63–100. Springer-Verlag, Berlin, 1980.

[8] J. Chandezon, C. Raoult, and D. Maystre. A new theoretical method for diffraction gratings and its numerical application. *Journal of Optics*, 11(4):235–241, 1980.

[9] L. Li, J. Chandezon, G. Franet, and J.-P. Plumey. Rigorous and efficient grating-analysis method made easy for optical engineers. *Applied Optics*, 38(2):304–313, 1999.

[10] P. Lalanne and P. Chavel, editors. *Perspectives for Parallel Optical Interconnects*. ESPRIT Basic Research series. Springer Verlag, Berlin Heidelberg New York, first edition, 1991.

[11] C. Heine. *Thin Film Coated Submicron Gratings: Theory, Design Fabrication and Application*. PhD Thesis, University of Neuchâtel, 1996.

[12] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes : The Art of Scientific Computing*. Cambridge University Press, Cambridge, 1986.

[13] J. Turunen. Diffraction Theory of Microrelief Gratings. In H. P. Herzig, editor, *Micro-Optics ; Elements, systems and applications*, pages 31–52. Taylor and Francis, London, 1 edition, 1997.

[14] M. Schmitz and O. Bryngdahl. Rigorous concept for the design of diffractive microlenses with high numerical apertures. *Journal of the Optical Society of America A*, 14(4):901–906, 1997.

[15] J. W. Goodman. *Introduction to Fourier Optics*. Electrical and computer engineering. McGraw-Hill, second edition, 1968-1996.

[16] E. W. Marchand and E. Wolf. Consistent Formulation of Kirchhoff's Diffraction Theory. *Journal of the Optical Society of America*, 56(12):1712–1722, 1966.

[17] J. J. Stamnes and H. A. Eide. Exact and approximate solutions for focusing of two-dimensional waves. I. Theory. *Journal of the Optical Society of America A*, 15(5):1285–1291, 1998.

[18] H. A. Eide and J. J. Stamnes. Exact and approximate solutions for focusing of two-dimensional waves. III. Numerical comparisons between exact and Rayleigh-Sommerfeld theories. *Journal of the Optical Society of America A*, 15(5):1308–1319, 1998.

[19] H. A. Eide and J. J. Stamnes. Exact and approximate solutions for focusing of two-dimensional waves. II. Numerical comparisons among exact, Debye, and Kirchhoff theories. *Journal of the Optical Society of America A*, 15(5):1292–1307, 1998.

[20] E. Wolf and E. W. Marchand. Comparison of the Kirchhoff and the Rayleigh-Sommerfeld Theories of Diffraction at an Aperture. *Journal of the Optical Society of America*, 54(5):587–594, 1964.

[21] J. J. Stamnes and B. Spjelkavik. Focusing at small angular apertures in the Debye and Kirchhoff approximations. *Optics Communications*, 40(2):81–85, 1981.

[22] L. A. DeAcetis, F. S. Einstein, R. A. J. Juliano, and I. Lazar. Single strip diffraction: comparison of Kirchhoff theory and geometrical theory with the exact solution in the limit of small glancing angle and width; perpendicular polarization. *Applied Optics*, 15(11):2866–2870, 1976.

[23] G. C. Sherman. Application of the Convolution Theorem to Rayleigh's Integral Formulas. *Journal of the Optical Society of America*, 57:546–547, 1967.

[24] E. Lalor. Conditions for the Validity of the Angular Spectrum of Plane Waves. *Journal of the Optical Society of America*, 58(9):1235–1237, 1968.

[25] A. S. Marathay. Fourier transform of the Green's function for the Helmholtz equation. *Journal of the Optical Society of America*, 65(8):964–965, 1975.

[26] G. C. Sherman, A. J. Devaney, and L. Mandel. Plane-wave expansions of the optical field. *Optics Communications*, 6(2):115–118, 1972.

[27] Y. Shono and T. Inuzuka. Representation of a diffracted wave field by the band-limited angular spectrum. *Journal of the Optical Society of America*, 68(11):1579–1586, 1978.

[28] K.-H. Brenner and W. Singer. Light propagation through microlenses: a new simulation method. *Applied Optics*, 32(26):4984–4988, 1993.

[29] G. J. Swanson. Binary optics technoloy: theoretical limits on the diffraction efficiency of multilevel diffractive optical elements. Technical Report 914, Massachusetts Institute of Technology, 1 March 1991.

[30] B. Kress and P. Meyrueis. *Digital Diffractive Optics : An Introduction to Planar Diffractive Optics and Related Technology*. John Wiley and Sons, Chichester, 6rst edition, 2000.

[31] D. A. Pommet, M. G. Moharam, and E. B. Grann. Limits of scalar diffraction theory for diffractive phase elements. *Journal of the Optical Society of America A*, 11(6):1827–1834, 1994.

[32] T. Hessler. *Continuous-Relief Diffractive Optical Elements: Design, Fabrication and Applications.* PhD Thesis, University of Neuchâtel, 1997.

[33] V. Kettunen, M. Kuittinen, and J. Turunen. Effects of abrupt surface-profile transitions in nonparaxial diffractive optics. *Journal of the Optical Society of America*, 18(6):1257–1260, 2001.

[34] T. Vallius, V. Kettunen, M. Kuittinen, and J. Turunen. Step-discontinuity approach for non-paraxial diffractive optics. *Journal of Modern Optics*, 48(7):1195–1210, 2001.

[35] M. Born and E. Wolf. *Principles of Optics.* Cambrigde University Press, London, seventh edition, 1999.

[36] J. M. Teijido. *Conception and design of illumination light pipes.* PhD Thesis, University of Neuchâtel, 2000.

[37] O. Ripoll, V. Kettunen, and H. P. Herzig. Low-data DOE simulation based on the zones geometry. In J. Turunen and F. Wyrowski, editors, *Diffractive Optics 2001*, volume 30 of *Topical meeting digest series*, pages 78–79, Budapest, 2001.

[38] O. Ripoll, V. Kettunen, and H. P. Herzig. Low-data simulation of diffractive optical elements based on the zones geometry. *Journal of Modern Optics*, 49(11):1801–1809, 2002.

[39] M. A. McCord and M. J. Rooks. Electron Beam Lithography. In P. Rai-Choudhury, editor, *Handbook of Microlithography, Micromachining, and Microfabrication. Volume 1: Microlithography*, volume PM39 of *SPIE Press*, pages 139–249. SPIE, 1997.

[40] V. Arrizón and M. Testorf. Efficiency limit of spatially quantized Fourier array illuminators. *Optics Letters*, 22(4):197–199, 1997.

[41] C.-L. Chen and T. R. Osborne. Quantization effects on the fields of electron-beam generated cylindrical zone plates. *Applied Optics*, 26(12):2343–2347, 1987.

[42] E. Carcolé, J. Campos, and S. Bosch. Diffraction theory of Fresnel lenses encoded in low-resolution devices. *Applied Optics*, 33(2):162–174, 1994.

[43] E. Carcolé, J. Campos, I. Juvells, and S. Bosch. Diffraction efficiency of low-resolution Fresnel encoded lenses. *Applied Optics*, 33(29):6741–6746, 1994.

[44] E. Carcolé, J. Campos, I. Juvells, and J. d. F. Moneo. Diffraction theory of optimized low-resolution Fresnel encoded lenses. *Applied Optics*, 34(26):5952–5960, 1995.

[45] E. Carcolé, J. Campos, and I. Juvells. Phase quantization effects on Fresnel lenses encoded in low resolution devices. *Optics Communications*, 132:35–40, 1996.

[46] S. M. Rubin. *Computer Aids for VLSI Design*. Addison-Wesley VLSI Systems Series. Addison-Wesley Publishing Company, second edition, 1997.

[47] J. Fan, D. Zaleta, K. S. Urquhart, and S. H. Lee. Efficient encoding algorithms for computer-aided design of diffractive optical elements by the use of electron-beam fabrication. *Applied Optics*, 34(14):2522–2533, 1995.

[48] H. Aagedal, F. Wyrowski, and M. Schmid. Paraxial beam splitting and shaping. In J. Turunen and F. Wyrowski, editors, *Diffractive Optics for Industrial and Commercial Applications*, pages 165–188. Akademie Verlag, Berlin, 1997.

[49] D. R. Brown. Beam Shaping with Diffractive Diffusers. In F. M. Dickey and S. C. Holswade, editors, *Laser Beam Shaping ; Theory and Techniques*, Optical Engineering, pages 249–271. Marcel Dekker, New York Basel, 1 edition, 2000.

[50] A. E. Siegman. Quasi fast Hankel transform. *Optics Letters*, 1(1):13–15, 1977.

[51] H. P. Herzig and P. Kipfer. Aperture modulated diffusers (AMDs). In T. Asakura, editor, *International Trends in Optics and Photonics ICO IV*, volume 74 of *Series in Optical Sciences*, pages 247–257. Springer, Berlin, 1999.

[52] W. Sillitto. Fraunhofer diffraction at straight-edged apertures. *Journal of the Optical Society of America*, 69(5):765–770, 1979.

[53] J. Komrska. Simple derivation of formulas for Fraunhofer diffraction at polygonal apertures. *Journal of the Optical Society of America*, 72(10):1382–1384, 1982.

[54] I. Kallioniemi, J. Saarinen, K. Blomstedt, and J. Turunen. Polygon approximation of the fringes of diffractive elements. *Applied Optics*, 36(28):7217–7223, 1997.

[55] G. R. Gindi and A. F. Gmitro. Optical feature extraction via the Radon transform. *Optical Engineering*, 23(5):499–506, 1984.

[56] S. N. Dixit, I. M. Thomas, B. W. Woods, A. J. Morgan, M. A. Henesian, P. J. Wegner, and H. T. Powell. Random phase plates for beam smoothing on the Nova laser. *Applied Optics*, 32(14):2543–2554, 1993.

[57] I. N. Ross, D. A. Pepler, and C. Danson. Binary phase zone plate designs using calculations of far-field distributions. *Optics Communications*, 116:55–61, 1995.

[58] C. Kopp, L. Ravel, and P. Meyrueis. Efficient beamshaper homogenizer design combining diffractive optical elements, microlens array and random phase plate. *Journal of Optics A*, 1(3):398–403, 1999.

[59] V. Kettunen, J. Lautanen, R. Silvennoinen, J. Turunen, and P. Vahimaa. Pixel-approximation effects in resonance-domain interferogram-type diffractive elements. In J. Turunen and F. Wyrowski, editors, *Diffractive Optics 97*, volume 12 of *Topical meeting digest series*, pages 150–151, Savonlinna, Finland, 1997.

[60] H. Aagedal and F. Wyrowski. Consequence of high resoltuion lithography for the design in the paraxial domain. In J. Turunen and F. Wyrowski, editors, *Diffractive Optics 97*, volume 12 of *Topical meeting digest series*, pages 166–167, Savonlinna, Finland, 1997.

[61] P. Blair, H. Lüpken, M. R. Taghizadeh, and F. Wyrowski. Multilevel phase-only array generators with a trapezoidal phase topology. *Applied Optics*, 36(20):4713–4721, 1997.

[62] M. Madou. *Fundamentals of Microfabrication*. CRC Press, Boca Raton, 1st edition, 1997.

[63] M. T. Gale and M. Rossi. Continuous-relief diffractive lenses and microlens arrays. In J. Turunen and F. Wyrowski, editors, *Diffractive Optics for Industrial and Commercial Applications*, pages 105–145. Akademie Verlag, Berlin, first edition, 1997.

[64] M. Flury. *Design and fabrication of Diffractive Optical Elements (DOE) for high power laser beam shaping*. PhD Thesis, University of Strasbourg I, 2002.

[65] M. Gruber, D. Hagedorn, and W. Eckert. Precise and simple optical alignment method for double-sided lithography. *Applied Optics*, 40(28):5052–5055, 2001.

[66] J. Turunen, A. Vasara, J. Westerholm, and A. Salin. Stripe-geometry two-dimensional Dammann gratings. *Optics Communications*, 74(3-4):245–252, 1989.

[67] A. Vasara, M. R. Taghizadeh, J. Turunen, J. Westerholm, E. Noponen, H. Ichikawa, J. M. Miller, T. Jaakkola, and S. Kuisma. Binary surface-relief gratings for array illumination in digital optics. *Applied Optics*, 31(17):3320–3336, 1992.

[68] J.-N. Gillet and Y. Sheng. Irregular spot array generator with trapezoidal apertures of varying heights. *Optics Communications*, 166:1–7, 1999.

[69] W. Brünger, E. B. Kley, B. Schnabel, I. Stolberg, M. Zierbock, and R. Plontke. Low energy lithography; energy control and variable energy exposure. *Microelectronic Engineering*, 1995(1-4):135–138, 1995.

[70] B. Niemann, T. Wilhein, T. Schliebe, R. Plontke, O. Fortagne, I. Stolberg, and M. Zierbock. A special method to create gratings of variable line density by low voltage electron beam lithography. *Microelectronic Engineering*, 30(1-4):49–52, 1996.

[71] S. Y. Popov, A. T. Friberg, M. Honkanen, J. Lautanen, J. Turunen, and B. Schnabel. Apodized annular-aperture diffractive axicons fabricated by continuous-path-control electron beam lithography. *Optics Communications*, 154(5-6):359–367, 1998.

[72] A. G. Poleshchuk, E. G. Churin, V. P. Koronkevich, V. P. Korolkov, A. A. Kharissov, V. V. Cherkashin, V. P. Kiryanov, A. V. Kiryanov, S. A. Kokarev, and A. G. Verhoglyad. Polar coordinate laser pattern generator for fabrication of diffractive optical elements with arbitrary structure. *Applied Optics*, 38(8):1295–1301, 1999.

[73] S. Reichelt, M. Daffner, H. J. Tiziani, and R. Freimann. Wavefront aberrations of rotationally symmetric CGHs fabricated by a polar coordinate laser plotter. *Journal of Modern Optics*, 49(7):1069–10287, 2002.

[74] J. P. Bowen, R. L. Michaels, and G. G. Blough. Generation of large-diameter diffractive elements with laser pattern generation. *Applied Optics*, 36(34):8970–8975, 1997.

[75] S. Y. Chou, P. R. Krauss, W. Zhang, L. Guo, and L. Zhuang. Sub-10 nm imprint lithography and applications. *Journal of Vacuum Science and Technology B*, 15(6):2897–2904, 1997.

[76] P. Nussbaum, R. Völkel, H. P. Herzig, M. Eisner, and S. Haselbeck. Design, fabrication and testing of microlens arrays for sensors and microsystems. *Pure and Applied Optics: Journal of the European Optical Society Part A*, (6):617–636, 1997.

[77] A. Schilling. *Diffractive and Refractive Optical Microstructures: Theory, Design and Applications*. PhD Thesis, University of Neuchâtel, 2000.

[78] A. Schilling, R. Merz, C. Ossmann, and H. P. Herzig. Surface profiles of reflow microlenses under the influence of surface tension and gravity. *Optical Engineering*, 39(8):2171–2176, 2000.

[79] P. Nussbaum and H. P. Herzig. Low numerical aperture refractive microlenses in fused silica. *Optical Engineering*, 40(7):1412–1414, 2001.

[80] P. Ehbets, M. Rossi, and H. P. Herzig. Continous-relief fan-out elements with optimized fabrication tolerances. *Optical Engineering*, 34(12):3456–3464, 1995.

[81] T. Hessler and R. E. Kunz. Relaxed fabrication tolerances for low-Fresnel-number lenses. *Journal of the Optical Society of America A*, 14(7):1599–1606, 1997.

[82] O. Bryngdahl. Geometrical transformations in optics. *Journal of the Optical Society of America*, 64(8):1092–1099, 1974.

[83] C. Frère and O. Bryngdahl. Computer-generated holograms: reconstruction of curves in 3-D. *Optics Communications*, 60(6):369–372, 1986.

[84] S. Bará, C. Frère, Z. Jaroszewicz, A. Kolodziejczyk, and D. Leseberg. Modulated on-axis zone plates for a generation of three-dimensional focal curves. *Journal of Modern Optics*, 37(8):1287–1295, 1990.

[85] Z. Jaroszewicz, A. Kolodziejczyk, D. Mouriz, and S. Bará. Analytic design of computer-generated Fourier-transform holograms for plane curves reconstruction. *Journal of the Optical Society of America A*, 8(3):559–565, 1991.

[86] F. Roux. Intensity distribution transformation for rotationally symmetric beam shaping. *Optical Engineering*, 30(5):529–536, 1991.

[87] C. Frère, D. Leseberg, and O. Bryngdahl. Computer-generated holograms of three-dimensional objects composed of line segments. *Journal of the Optical Society of America A*, 3(5):726–730, 1986.

[88] D. Leseberg. Computer generated holograms: cylindrical, conical, and helical waves. *Applied Optics*, 26(20):4385–4390, 1987.

[89] T. Dresel, M. Beyerlein, and J. Schwider. Design and fabrication of computer-generated beam-shaping holograms. *Applied Optics*, 35(23):4615–4621, 1996.

[90] L. L. Doskolovich, N. L. Kazanskiy, S. I. Kharitonov, and G. V. Uspleniev. Focusators for Laser-Branding. *Optics and Laser Engineering*, 15:311–322, 1991.

[91] C.-Y. Han, Y. Isbii, and K. Murata. Reshaping collimated laser beams with Gaussian profile to uniform profiles. *Applied Optics*, 22(22):3644–3647, 1983.

[92] J. Sochacki, A. Kolodziejczyk, Z. Jaroszewicz, and S. Bara. Nonparaxial design of generalized axicons. *Applied Optics*, 31(25):5326–5330, 1992.

[93] Z. Jaroszewicz, A. Kolodziejczyk, D. Mouriz, and J. Sochacki. Generalized zone plates focusing into arbitrary line segments. *Journal of Modern Optics*, 40(4):601–612, 1993.

[94] Z. Jaroszewicz, J. Sochacki, A. Kolodziejczyk, and L. R. Staronski. Apodized annular-aperture logarithmic axicon:smoothness and uniformity of intensity distributions. *Optics Letters*, 18(22):1893–1895, 1993.

[95] N. C. Roberts. Beam shaping by holographic filters. *Applied Optics*, 28(1):31–32, 1989.

[96] M. A. Golub, I. N. Sisakyan, and V. A. Soifer. Infra-red Radiation Focusators. *Optics and Lasers in Engineering*, 15:297–309, 1991.

[97] W. Singer, H. P. Herzig, M. Kuittinen, E. Piper, and J. Wangler. Diffractive beamshaping elements at the fabrication limit. *Optical Engineering*, 35(10):2779-2787, 1996.

[98] H. Aagedal, M. Schmid, S. Egner, J. Müller-Quade, T. Beth, and F. Wyrowski. Analytical beam sbaping with application to laser-diode arrays. *Journal of the Optical Society of America A*, 14(7):1549–1553, 1997.

[99] A. Hermerschmidt, H. Eichler, S. Teiwes, and J. Schwartz. Design of diffractive beam-shaping elements for non-uniform illumination waves. In *Diffractive and Holographic Device Technologies and Applications V.*, volume 3291, pages 40–48, San Jose, CA, USA., 1998. SPIE.

[100] A. Hermerschmidt, S. Teiwes, M. Ferstl, and R. Steingrüber. Design of diffractive beam-shaping elements and experimental investigation of binary test elements. In *Diffractive Optics '99*, pages 278–179, Jena, 1999. European Optical Society.

[101] M. Ferstl, R. Steingrüber, S. Krüger, S. Teiwes, and H. Andreas. Computer-generated DOEs for laser beam shaping and beam splitting applications. In *EOS Topical Meeting on Diffractive Optics*, volume 22 of *Topical Meeting Digest Series*, pages 223–224, Jena, 1999.

[102] E. W. Weisstein. Eric Weisstein's World of Mathematics, 1999.

[103] S. Y. Popov and A. T. Friberg. Apodization of generalized axicons to produce uniform axial line images. *Pure and Applied Optics: Journal of the European Optical Society Part A*, 7(3):537–548, 1998.

[104] N. Davidson, A. A. Friesem, and E. Hasman. Diffractive elements for annular laser beam transformation. *Applied Physics Letters*, 61(4):381–383, 1992.

[105] F. Wyrowski. Design theory of diffractive elements in the paraxial domain. *Journal of the Optical Society of America A*, 10(7):1553–1561, 1993.

[106] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[107] M. A. Seldowitz, J. P. Allebach, and D. W. Sweeney. Synthesis of digital holograms by direct binary search. *Applied Optics*, 26(14):2788–2798, 1987.

[108] N. Yoshikawa, M. Itoh, and T. Yatagai. Quantized phase optimization of two-dimensional Fourier kinoforms by a genetic algorithm. *Optics Letters*, 20(7):752–754, 1995.

[109] P. M. Hirsch, J. A. Jordan, and L. B. J. Lesem. Method of making an object independent diffuser, November 9, 1971 1970.

[110] J. R. Fienup. Iterative method applied to image reconstruction and to computer-generated holograms. *Optical Engineering*, 19(3):297–305, 1980.

[111] J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982.

[112] L. Ingber. Very Fast Simulated Re-Annealing. *Mathematical and Computer Modelling*, 12(8):967–973, 1989.

[113] J. Turunen, A. Vasara, and J. Westerholm. Kinoform phase relief synthesis: a stochastic method. *Optical Engineering*, 28(11):1162–1167, 1989.

[114] M. R. Feldman and C. C. Guest. Iterative encoding of high-efficiency holograms for generation of spot arrays. *Optics Letters*, 14(10):479–481, 1989.

[115] A. G. Kirk and T. J. Hall. Design of binary computer generated holograms by simulated annealing: coding density and reconstruction error. *Optics Communications*, 94:491–496, 1992.

[116] N. Yoshikawa and T. Yatagai. Phase optimization of a kinoform by simulated annealing. *Applied Optics*, 33(5):863–868, 1994.

[117] V. Arrizón and L. A. González. Optimization of quantized diffractive elements with symmetry constraints. *Optics Communications*, 180(4-6):247–254, 2000.

[118] L. Ingber and B. Rosen. Genetic algorithms and very fast simulated reannealing: A comparison. *Mathematical and Computer Modelling*, 16(11):87–100, 1992.

[119] L. Ingber. Simulated annealing: Practice versus theory. *Mathematical and Computer Modelling*, 18(11):29–57, 1993.

[120] J.-N. Gillet and Y. Sheng. A new trapezoidal topology for designing diffractive optical elements with the iterative simulated quenching optimization. In OSA, editor, *DOMO 2000*, Québec (Canada), 2000.

[121] J.-N. Gillet and Y. Sheng. Iterative simulated quenching for designing irregular-spot-array generators. *Applied Optics*, 39(20):3456–3465, 2000.

[122] P. Birch, R. Young, M. Farsari, D. Budgett, J. Richardson, and C. Ghatwin. A Comparison of the Iterative Fourier Transform Method and Evolutionary Algorithms for the Design of Diffractive Optical Elements. *Optics and Laser Engineering*, 33(6):439–448, 2000.

[123] C.-h. Chen and A. A. Sawchuk. Nonlinear least-squares and phase-shifting quantization methods for diffractive optical element design. *Applied Optics*, 36(29):7297–7306, 1997.

[124] S. Bühling and F. Wyrowski. Improved transmission design algorithms by utilizing variable strength projections. *Journal of Modern Optics*, 49(11):1871–1892, 2002.

[125] F. Wyrowski. Diffractive optical elements: iterative calculation of quantized, blazed phase structures. *Journal of the Optical Society of America A*, 7(6):961–969, 1990.

[126] H. Schwarzer, S. Teiwes, and F. Wyrowski. "Non-pixelated" design of computer-generated diffractive elements for increased diffraction efficiency. In J. Turunen and F. Wyrowski, editors, *Diffractive Optics 97*, volume 12 of *Topical meeting digest series*, pages 164–165, Savonlinna, Finland, 1997.

[127] G.-Y. Yoon, S. Matsuoka, T. Jitsuno, M. Nakatsuka, and Y. Kato. Wavefront design algorithm for shaping a quasi-far-field pattern. *Applied Optics*, 37(8):1386–1392, 1998.

[128] D. Prongué, H. P. Herzig, R. Dändliker, and M. T. Gale. Optimized kinoform structures for highly efficient fan-out elements. *Journal of the Optical Society of America A*, 31(26):5706–5711, 1992.

[129] M. Johansson and J. Bengtsson. Robust design method for highly efficient beam-shaping diffractive optical elements using an iterative-Fourier-transform algorithm with soft operations. *Journal of Modern Optics*, 47(8):1385–1398, 2000.

[130] V. Arrizón, M. Testorf, S. Sinzinger, and J. Jahns. Iterative optimization of phase-only diffractive optical elements based on a lenslet array. *Journal of the Optical Society of America A*, 17(12):2157–2164, 2000.

[131] D. Schäfer. Design concept for diffractive elements shaping partially coherent laser beams. *Journal of the Optical Society of America A*, 18(11):2915–2922, 2001.

[132] M. Skeren, I. Richter, and P. Fiala. Iterative Fourier transform algorithm: comparison of various approaches. *Journal of Modern Optics*, 49(11):1851–1870, 2002.

[133] H. Aagedal, M. Schmid, T. Beth, S. Teiwes, and F. Wyrowski. Theory of speckles in diffractive optics and its application to beam shaping. *Journal of Modern Optics*, 43(7):1409–1421, 1996.

[134] P. Senthilkumaran and F. Wyrowski. Phase synthesis in wave-optical engineering: mapping- and diffuser-type approaches. *Journal of Modern Optics*, 49(11):1831–1850, 2002.

[135] D. W. Ricks. Scattering from diffractive optics. In S. H. Lee, editor, *Diffractive and Miniaturized Optics*, volume CR49 of *Critical Reviews of Optical Science and Technology*, pages 187–211, San Diego, 1993. SPIE Optical Engineering Press.

[136] D. A. Pommet, E. B. Grann, and M. G. Moharam. Effects of process errors on the diffraction characteristics of binary dielectric gratings. *Applied Optics*, 34(14):2430–2435, 1995.

[137] V. Kettunen, J. Simonen, O. Ripoll, M. Kuittinen, and H. P. Herzig. Diffractive elements designed to suppress unwanted zeroth order due to surface depth error. *Submitted to Journal of Modern Optics.*

[138] J. A. Cox, T. Wernet, J. Lee, S. Nelson, B. Fritz, and J. Bergstrom. Difraction efficiency of binary optical elements. In *Computer and Optically Formed Holographic Optics*, volume 1211, pages 116–124. SPIE, 1990.

[139] J. A. Cox, B. Fritz, and T. Werner. Process error limitations on binary optics performance. In I. Cindrich and S. H. Lee, editors, *Computer and Optically Generated Holographic Optics*, volume 1555, pages 80–88. SPIE, 1991.

[140] M. B. Stern, M. Holz, S. S. Medeiros, and R. E. Knowlden. Fabricating binary optics: Process variables critical to optical efficiency. *Journal of Vacuum Science and Technology B*, 9(6):3117–3121, 1991.

[141] J. A. Cox. Processing Error Limitations on Performance of Diffractive Optical Elements. In *Diffractive Optics: Design, Fabrication, and Applications Technical Digest*, volume 9, pages 143–145. Optical Society of America, Washington DC, 1992.

[142] M. Ferstl, B. Kuhlow, and E. Pawlowski. Effect of fabrication errors on multilevel Fresnel zone lenses. *Optical Engineering*, 33(4):1229–1235, 1994.

[143] V. Kettunen, O. Ripoll, and H. P. Herzig. Beam shaping in deep UV: comparison of methods. In J. Turunen and F. Wyrowski, editors, *Diffractive Optics 2001*, volume 30 of *Topical meeting digest series*, pages 80–81, Budapest, 2001.

[144] T. H. Bett, C. Danson, P. Jinks, D. A. Pepler, I. N. Ross, and R. M. Stevenson. Binary phase zone-plate arrays for laser-beam spatial-intensity distribution conversion. *Applied Optics*, 34(20):4025–4036, 1995.

[145] W. H. Welch, J. E. Morris, and M. R. Feldman. Iterative discrete on-axis encoding of radially symmetric computer-generated holograms. *Journal of the Optical Society of America A*, 10(8):1729–1738, 1993.

[146] M. Kuittinen and H. P. Herzig. Encoding of efficient diffractive microlenses. *Optics Letters*, 21(21):2156–2158, 1995.

[147] V. Arrizón and L. A. González. Non-paraxial array illuminator based on a single low-resolution pixelated lens. *Optics Communications*, 199(5-6):345–353, 2001.

[148] F. M. Dickey and S. C. Holswade, editors. *Laser Beam Shaping ; Theory and Techniques*. Optical Engineering. Marcel Dekker, New York Basel, first edition, 2000.

[149] G. Farin. *Curves and Surfaces for Computer-Aided Geometric Design*. Academic Press, San Diego, 4th edition, 1997.

[150] L. Shao and H. Zhou. Curve Fitting with Bézier Cubics. *Graphical Models and Image processing*, 58(3):223–232, 1996.

[151] M. Froumentin, F. Labrosse, and P. Willis. A Vector-based Representation for Image Warping. *Computer Graphics Forum*, 19(3), 2000.

[152] M. Abramovitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover Publications, Inc., New York, eighth edition, 1972.

[153] J. A. C. Weideman. Computation of the Complex Error Function. *SIAM Journal of Numerical Analysis*, 31(5):1497–1518, 1994.