Published in
Gradient Index, Miniature, and Diffractive Optical Systems II - Proceedings of SPIE 4437, 1-13, 2001
which should be used for any reference to this work

Design and modeling of a miniature system containing micro-optics

Norbert Lindlein
Lehrstuhl für Optik, University of Erlangen-Nürnberg, Staudtstr. 7/B2, D-91058 Erlangen, Germany
e-mail: norbert.lindlein@optik.uni-erlangen.de

Hans Peter Herzig
Institute of Microtechnology, University of Neuchâtel, Rue A.-L. Breguet 2, CH-2000 Neuchâtel, Switzerland
e-mail: hanspeter.herzig@unine.ch

ABSTRACT

The design and modeling of micro-optical systems is still a challenging task because classical methods like ray tracing do not take into account diffraction effects and other coherent effects which appear e.g. in the presence of micro-optical array systems. On the other side, there exist scalar or rigorous diffraction theories to model optical systems. But they are also limited in their applications because they either neglect non-paraxial effects or the calculation time is too high for a practical use.
In this paper we will therefore give an overview about existing (scalar) theories to model optical systems, especially systems containing micro-optics: a simple paraxial matrix theory, ray tracing, Gaussian beam propagation and the propagation of a wave using the angular spectrum of plane waves. The advantages and disadvantages of these theories will be shown and compared.
At the end we will describe a combination of ray tracing and wave propagation methods to give a more realistic simulation of micro-optical systems.

Keywords: ray tracing, diffraction theories, micro-optics, optical design, modeling of optical systems, Gaussian beams, angular spectrum of plane waves

1. INTRODUCTION

Micro-optical elements are widely used in modern optical systems. Some applications are the collimation of laser diodes, the coupling of light into fibers or the focusing of light into a very small spot like in CD players. Also complete arrays of micro-optical elements are used. In a Shack-Hartmann wave front sensor an array of microlenses is used to detect the local wave front slope at each microlens. Light homogenizers use arrays of microlenses to homogenize the light distribution and arrays of micro-mirrors deflect a plane wave into several waves with different deflection angles.
So, the small dimensions of micro-optical elements and the array aspects require other simulation methods as in the case of classical macroscopic optical systems. Nevertheless, the simulation methods for micro-optical systems are based on classical methods and in many cases classical methods give very good approximations.
Therefore, this paper will review some classical methods and develop new methods based on these traditional methods. We will only consider scalar theories and free-space optical systems, i.e. no integrated optical systems. Nearly all of the described methods are either implemented in the optical design software RAYTRACE [1], which was developed by the first author at the University of Erlangen during the last ten years, or in the MATLAB based wave-optical simulation software HAPPY [2], which was developed at the University of Neuchâtel in the group of the second author.
The first simulation method to be described (section 2) is the simple paraxial matrix theory [3] which allows to calculate the paraxial parameters of a system by the multiplication of some elementary matrices. In the case of an array of elements each element of the array has to be calculated separately. An extended theory using 3x3 matrices is proposed [4].

This method allows also to describe off-axis systems as long as the off-axis parameters are in the paraxial range. Of course, this method does not take into account non-paraxial effects like aberrations or diffraction effects. Therefore, it is only useful to make a first paraxial design.
The second method (section 3) is the well-known ray tracing [5] which is the most important method for the simulation and optimization of classical optical systems. Also for micro-optical systems ray tracing offers some possibilities. Arrays of elements can e.g. be calculated by using the so called non-sequential ray tracing. Ray tracing takes into account non-paraxial effects but it does of course not take into account diffraction effects. In classical focusing systems diffraction effects are considered by using ray tracing to calculate the wave aberrations in the exit pupil of the system and then using the Debye integral [6] for the calculation of the point spread function (PSF). In micro-optical array systems there exists in general no simple PSF so that this special combination of ray tracing and wave propagation cannot be used in several cases.
The third method (section 4) is the also well-known theory of the propagation of fundamental Gaussian beams by using the Fresnel diffraction approximation [3]. Although, this is a quite special method which can be only applied to Gaussian beams it is interesting for practical purposes because Gaussian beams are very important. This method takes into account diffraction effects during free-space propagation but no diffraction effects at apertures or lens frames. Additionally, it is only a paraxial theory so that no aberrations (with the exception of astigmatism) are taken into account. We will shortly describe an extended method for the propagation of Gaussian beams in quite arbitrary off-axis optical systems based on differential ray tracing [7-9].
The fourth and last classical method (section 5) is the propagation of a wave front by using the angular spectrum of plane waves [10]. Of course, this method describes only the free-space propagation of a wave front and not the refraction or diffraction at a lens or another optical element. Therefore, it has to be combined with other simulation methods to enable the simulation of complete optical systems. Nevertheless, it is very interesting because it is an exact solution of the scalar wave equation.
If we consider the principles of all these four theories we can see that each complete simulation theory requires a method for the free-space propagation of a wave front and a method for the change of a wave front by refraction, reflection or diffraction at an optical element. Therefore, from a practical point of view it is in principle possible to combine several methods if suitable interfaces for wave fronts are defined. Especially, the combination of ray tracing for the tracing through optical elements and of the angular spectrum of plane waves for free-space propagation seems promising. Some ideas to do this will be given in section 6. Of course, these ideas will not form an exact and complete theory but they are heuristic approaches to overcome practical problems in the simulation of micro-optical systems.
At the end of some sections examples for the simulation of a micro-optical system using the respective method are given.
This paper will in general show that there exists nowadays (and to our opinion in the foreseeable future) no single complete theory for the simulation of micro-optical systems which can be applied in every case but that there are several theories with a certain range of application. So, this paper can only describe the fundamentals and give some hints for a more complete theory. It is not our aim and far beyond of our reach to replace the practical experience of an optical designer who has to decide in the end which of all theories is best fitted to his special problem.

2. THE PARAXIAL MATRIX THEORY

The paraxial matrix theory in its most simple formulation describes the propagation of a paraxial ray with ray height x and ray angle φ with respect to the optical axis of an on-axis optical system [3]. In the paraxial approximation the conditions $\sin\varphi \approx \tan\varphi \approx \varphi$ and $\cos\varphi \approx 1$ are fulfilled. There is a matrix for the free-space propagation of a paraxial ray which connects the ray parameters (x,φ) before the propagation with the ray parameters (x',φ') behind the propagation along a distance d (see figure 1):

$$\begin{pmatrix} x' \\ \varphi' \end{pmatrix} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ \varphi \end{pmatrix} \tag{1}$$

There are also matrices for the refraction of such a ray at a plane surface perpendicular to the optical axis with refractive indices n and n' left and right to the surface

$$\begin{pmatrix} 1 & 0 \\ 0 & n/n' \end{pmatrix} \tag{2}$$

or the refraction at a spherical surface with radius of curvature R (see figure 2)

$$\begin{pmatrix} 1 & 0 \\ -(n'-n)/(n'R) & n/n' \end{pmatrix}. \tag{3}$$

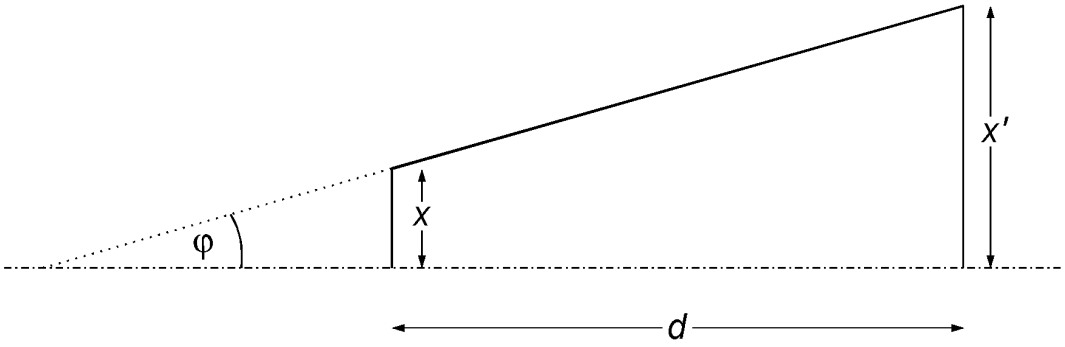

Fig. 1: Free-space propagation of a paraxial ray.

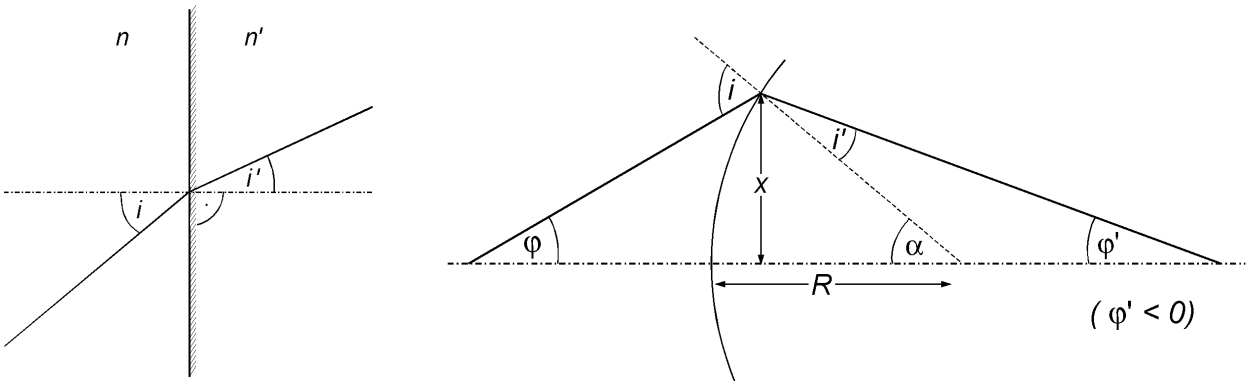

Fig. 2: Refraction at a plane surface (left) or a spherical surface (right) with radius of curvature R.

The change of the ray parameters at a thin lens with focal length f are described via:

$$\begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \tag{4}$$

There are of course also further matrices for other operations which are not described here.

The matrix M_{total} for the propagation of a paraxial ray through a complete system of optical elements is done by multiplying the matrices M_1, …, M_n for the different operations (e.g. free-space propagation or refraction at a thin lens) in the order given in equation (5):

$$M_{total} = M_n M_{n-1} ... M_1 \tag{5}$$

A disadvantage of this simple theory is that e.g. the refraction at a thin prism which adds in the paraxial approximation a constant angle $\Delta\varphi$ to the ray angle cannot be described by a simple matrix multiplication. Therefore, an extension of the method can be done by using a three by three matrix and a three component ray vector whereby the third component is always one [4]:

$$\begin{pmatrix} x' \\ \varphi' \\ 1 \end{pmatrix} = \begin{pmatrix} A & B & \Delta x \\ C & D & \Delta\varphi \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ \varphi \\ 1 \end{pmatrix} \tag{6}$$

The components *A,B,C* and *D* are hereby the components of the normal two by two matrix and the values Δx and $\Delta\varphi$ describe offsets for the ray height and the ray angle.
By using such extended matrices also the propagation through a system of tilted and laterally shifted elements can be described. This is done by introducing a tilt matrix *T* or a shift matrix *S* which transforms the ray into the local coordinate system of the tilted or shifted optical element:

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \Delta\varphi \\ 0 & 0 & 1 \end{pmatrix}, \quad S = \begin{pmatrix} 1 & 0 & \Delta x \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{7}$$

The ray data are then of course defined in the local coordinate system of the respective element and a reverse operation has to be done to transform the ray back into the global coordinate system.

Example:
The following example explains the method by calculating the data of a ray which is propagated through a system consisting of a thin lens with focal length f_1 which is on the global optical axis and a thin lens with focal length f_2 which follows at a distance *d* and is laterally shifted by the value σ and tilted by the angle δ (see figure 3). To calculate the matrix M_{syst} of the whole system the following matrices habe to be multiplied: matrix for lens 1, matrix for free-space propagation, matrix for transformation into the local coordinate system of lens 2, matrix for lens 2 and finally the matrix for back-transformation into the global coordinate system.

$$\begin{aligned} M_{syst} &= \begin{pmatrix} 1 & 0 & \sigma \\ 0 & 1 & \delta \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -1/f_2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -\sigma \\ 0 & 1 & -\delta \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & d & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -1/f_1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \\ &= \begin{pmatrix} 1-d/f_1 & d & 0 \\ -(1/f_1+1/f_2-d/(f_1 f_2)) & 1-d/f_2 & \sigma/f_2 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned} \tag{8}$$

As can be seen from equation (8) the tilt of a lens can be neglected in the paraxial domain. Of course, the tilt angle has to be so small that the change of the axial coordinate of the point of intersection between the ray and the tilted lens can be neglected.

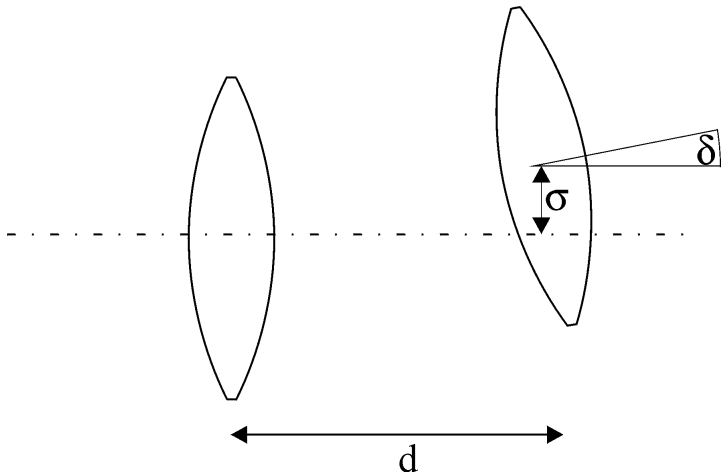

Fig. 3: Propagation through an off-axis system consisting of two thin lenses.

Let us now consider a point source with the coordinate *x* a distance d_1 in front of the first thin lens and calculate the complete matrix for the propagation of the light to a distance d_2 behind the second thin lens. Then, the complete matrix is:

$$M_{total}=\begin{pmatrix}1 & d_2 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}M_{syst}\begin{pmatrix}1 & d_1 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}=\begin{pmatrix}1-\frac{d}{f_1}-\frac{d_2}{f} & d+d_1+d_2-d\left(\frac{d_1}{f_1}+\frac{d_2}{f_2}\right)-\frac{d_1d_2}{f} & \sigma\frac{d_2}{f_2}\\ -\frac{1}{f} & 1-\frac{d}{f_2}-\frac{d_1}{f} & \frac{\sigma}{f_2}\\ 0 & 0 & 1\end{pmatrix} \tag{9}$$

Here, the abbreviation $1/f=1/f_1+1/f_2-d/(f_1f_2)$ was used. This quantity can also in the case of a shifted second lens be interpreted as the focal power of the system. To have an image of the point source all rays starting from the point source under different angles φ have to focus in the same point. So, the element $m_{total1,2}=B$ of the matrix has to be zero:

$$B=d+d_1+d_2-d\left(\frac{d_1}{f_1}+\frac{d_2}{f_2}\right)-\frac{d_1d_2}{f}=0\Rightarrow d_2=\frac{-\left(d+d_1-\frac{dd_1}{f_1}\right)}{1-\frac{d}{f_2}-\frac{d_1}{f}} \tag{10}$$

By using equation (9) it can be seen that the lateral shift σ of the second lens introduces an additional lateral shift $\Delta x=m_{total1,3}=\sigma d_2/f_2$ of the image point which has the size

$$\Delta x=m_{total1,3}=-\sigma\frac{d+d_1-\frac{dd_1}{f_1}}{f_2-d-\frac{d_1f_2}{f}} \tag{11}$$

For the special case that the point source is in the focal plane of the first lens ($d_1=f_1$) we have of course the well-known result $\Delta x=\sigma$.

3. RAY TRACING

The most important method for the calculation of the performance of classical macroscopic optical systems is ray tracing [5]. A light ray in a homogeneous medium is described by a starting point $\underline{p}=(p_x,p_y,p_z)$ and a direction vector $\underline{e}=(e_x,e_y,e_z)$ with $|\underline{e}|=1$. The position vector $\underline{r}$ along the path of the ray is given by (see figure 4)

$$\underline{r}=\begin{pmatrix}x\\ y\\ z\end{pmatrix}=\underline{p}+\mu\underline{e} \tag{12}$$

Each surface in the optical system is described by an equation of the type

$$F(x,y,z)=0 \tag{13}$$

A simple example for a surface is a plane surface with center $\underline{r}_0$ and unit vector $\underline{n}$ perpendicular to the surface, so that for a point $\underline{r}$ on the surface the equation

$$F(\underline{r})=(\underline{r}-\underline{r}_0)\cdot\underline{n}=0 \tag{14}$$

holds. Another example is a spherical surface with center of curvature $\underline{r}_0$ and radius of curvature R. There a point $\underline{r}$ on the surface is described by the equation

$$F(\underline{r})=\left|\underline{r}-\underline{r}_0\right|^2-R^2=0 \tag{15}$$

Of course, in both examples there have to be further conditions to describe only a part of the surface and not the whole plane or a full sphere. But we cannot consider these details in this paper.

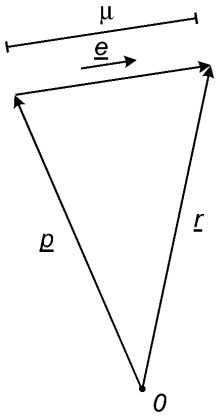

Fig. 4: Parameters of a ray. O is the origin of the coordinate system.

Ray tracing in a homogeneous medium consists of two different types of operation:

- Free-space propagation, i.e. determination of the point of intersection of a ray with the next surface, and
- refraction, reflection or diffraction of a ray at that surface.

Here, the word diffraction means the interaction of a ray with a diffractive optical element like e.g. a grating.
So, the point of intersection of a light ray with a surface *F* is determined by solving the equation

$$F(\underline{p}+\mu\underline{e})=0 \tag{16}$$

with respect to μ. For general aspheric surfaces this has to be done numerically. After having done this, the point of intersection is selected to be the starting point of the next ray section. In many cases it is useful to transform the ray in a local coordinate system in which the surface has a simple mathematical form before calculating the point of intersection. Afterwards the ray can be transformed back into the global coordinate system. In practice there may be several points of intersection with a surface. Then, the point with the smallest positive distance μ has to be taken.
The refraction, reflection or diffraction of a ray is performed by the vector equation [9]:

$$\underline{e}'=\frac{n}{n'}\underline{e}+m\frac{\lambda}{n'}\underline{G}-\frac{n}{n'}(\underline{e}\cdot\underline{N})\underline{N}\pm\operatorname{sign}(\underline{e}\cdot\underline{N})\sqrt{1+\left(\frac{n}{n'}\right)^2(\underline{e}\cdot\underline{N})^2-\left(\frac{n}{n'}\right)^2-\left(m\frac{\lambda}{n'}|\underline{G}|\right)^2-2m\frac{n\lambda}{n'^2}\underline{e}\cdot\underline{G}\,\underline{N}} \tag{17}$$

whereby *n* is the refractive index in front of the surface, *n'* the refractive index behind the surface and $\underline{N}$ the local unit vector perpendicular to the surface at the point of intersection. The local grating vector $\underline{G}$ is defined as the component parallel to the surface with $|\underline{G}|=1/\Lambda$ (Λ is the local grating period) and *m* is the diffraction order. The function sign is +1 if the argument is positive and –1 if the argument is negative. The "+" sign in front of the sign function is taken for refraction or a transmitting diffractive optical element. The "-" sign is taken in the case of reflection or for a reflecting diffractive optical element.
So, the ray tracing is performed by free-space propagation and refraction, reflection or diffraction at a surface by turns. Additionally to the ray direction the optical path length along each ray section is calculated by

$$OPD'=OPD+n\mu \tag{18}$$

In practice there are two different types of ray tracing, the so called sequential ray tracing and the non-sequential ray tracing. In the sequential ray tracing the user determines the order in which the surfaces of the optical system are hit manually. In the non-sequential ray tracing the program calculates itself the surface which is hit next by taking that surface with the smallest positive distance from the starting point of the ray traced along the ray direction. Of course, the non-sequential ray tracing needs in practice more calculation time than the sequential ray tracing. On the other side, the non-sequential ray tracing is necessary if the sequence in which the surfaces are hit is not known. If an array of microlenses is part of an optical system it is e.g. necessary to determine whether one of the microlenses is hit or whether the ray passes the space between the lenses. In the non-sequential mode it is also possible to divide a ray into several rays like e.g. a transmitted and a reflected ray or rays in different diffraction orders of a diffractive optical element. Each of these rays is then traced independently. By doing this stray light calculations due to Fresnel reflections can be taken into account.

As mentioned above ray tracing is a geometrical optical method and does not consider diffraction effects at apertures or during the free-space propagation. On the other side it is possible to calculate the ray paths through systems containing arrays of lenses which is often the case in micro-optical systems.

Example:
A light homogenizer often uses microlens arrays. The following example (see fig. 5) consists of an extended incoherent light source, an aspheric condensor lens and two refractive microlens arrays with circular apertures. The incoherent light source has a Lambertian light distribution and an elliptic area of 1.2mm x 0.24mm diameter. The condensor lens has a focal length of 6.9mm and forms together with each lens of the first microlens array (f_{array1}=1.65mm, diameter of lenses 0.595mm, hexagonal arrangement) an image of the light source (magnification factor –0.14) on each lens of the second microlens array (f_{array2}=1.36mm, diameter of lenses 0.54mm, hexagonal arrangement). Additionally, each lens of the second array images the aperture of the corresponding lens of the first array into the plane where the light has to be homogenized. The magnification factor is in this case –10 and the images of all lens apertures are on-axis, so that the homogenized field should be circular with a diameter of 5.95mm. Fig. 6 shows the result of a ray tracing simulation made with the program RAYTRACE. The number of rays is 10 million and the detector, which counts the number of rays in each detector element, has 50 x 50 elements. This example shows that ray tracing is very appropriate to simulate micro-optical systems as long as diffraction effects can be neglected.

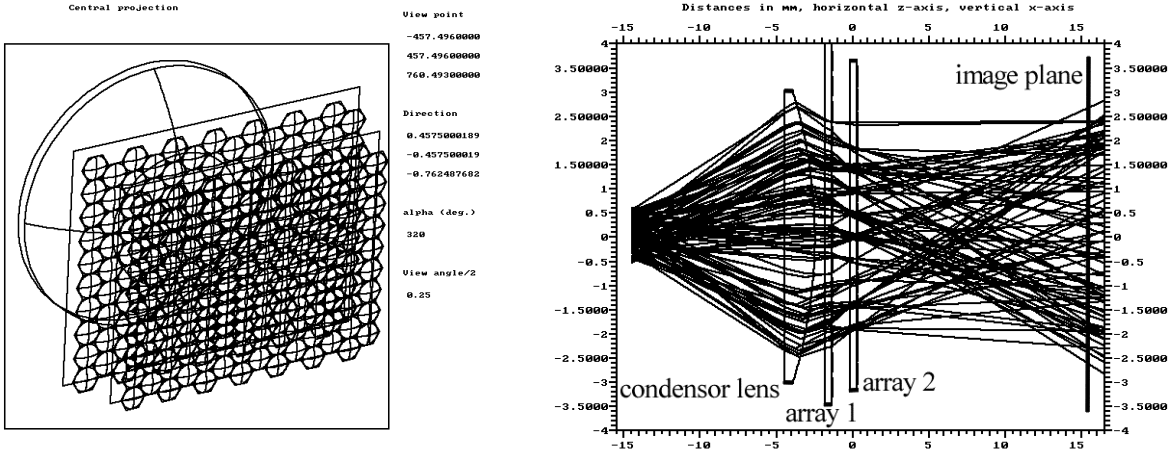

Fig. 5: Optical elements (left) and ray tracing scheme (right) of a light homogenizer using two microlens arrays.

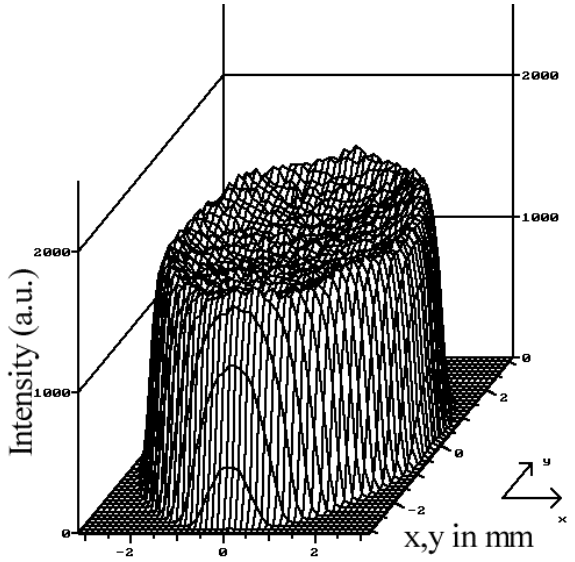

Fig. 6: Intensity distribution in the "image plane" of the light homogenizer.

4. GAUSSIAN BEAM PROPAGATION

The ground mode of a laser is often described very well with a fundamental Gaussian beam. A so-called simple astigmatic Gaussian beam has in each plane perpendicular to the optical axis (z-direction) two certain beam radii w_x and w_y and two radii of curvature R_x and R_y. The complex amplitude u of such a Gaussian beam is described by

$$u(x,y,z) = \Psi(z)\exp\left[-\left(\frac{x^2}{w_x^2(z)}+\frac{y^2}{w_y^2(z)}\right)\right]\exp\left[i\frac{\pi}{\lambda}\left(\frac{x^2}{R_x(z)}+\frac{y^2}{R_y(z)}\right)\right]\exp\left[i\frac{2\pi}{\lambda}z\right] \tag{19}$$

The smallest beam radius in the xz- or yz-plane which can be achieved due to diffraction effects is called the beam waist $w_{0,x}$ or $w_{0,y}$ of the Gaussian beam. It is well-known that the radius of curvature in the plane of the beam waist is infinity. The equations for the free-space propagation of a Gaussian beam are calculated by solving the paraxial time-independent wave equation (paraxial Helmholtz equation) and result in the well-known equations [3]:

$$w_x(z) = \sqrt{w_{0,x}^2 + \frac{\lambda^2 z^2}{\pi^2 w_{0,x}^2}} \tag{20}$$

and

$$R_x(z) = z + \frac{\pi^2 w_{0,x}^4}{\lambda^2 z} \tag{21}$$

Of course, a set of identical equations is also valid for the values w_y and R_y.

If a Gaussian beam hits a lens the radius of curvature changes according to the laws of geometrical optics. Of course, a Gaussian beam remains only a Gaussian beam if two conditions are fulfilled:

- The lens has no aberrations in the area where the intensity of the Gaussian beam is considerably different from zero.
- The lens aperture does not clip the Gaussian beam. In practice, this means that the lens aperture has to be at least so large that an area of the Gaussian beam with radii $2w_x$ and $2w_y$ can pass the lens aperture.

If these conditions are fulfilled the Gaussian beam keeps its Gaussian shape and can be traced through an optical system. There are several methods to do this. A method for on-axis systems is based on the paraxial ABCD matrix theory of section 2 [3], another method uses several rays calculated by ray tracing to propagate the Gaussian beam [11]. A third method which we will describe shortly uses the so-called differential ray tracing and the free-space propagation equations (20) and (21). The differential ray tracing, which is presented in [7-9], considers for each ray additional to the ray direction and phase also the local principal curvatures of the corresponding wave front and delivers equations for tracing these properties through an optical system.

The Gaussian beam is represented by a ray along its direction $\underline{e}$ of propagation, the two principal directions $\underline{P}_x$ and $\underline{P}_y$ of the corresponding wave front ($\underline{P}_x \cdot \underline{e} = \underline{P}_y \cdot \underline{e} = \underline{P}_x \cdot \underline{P}_y = 0$) and the two beam radii w_x and w_y and the two principal curvatures $1/R_x$ and $1/R_y$ along the two principal directions. The free-space propagation of the beam radii and the principal curvatures is calculated by using equations (20) and (21).

The change of these parameters by refraction, reflection or diffraction at a surface is performed using the following two methods.

(i) The differential ray tracing is used for the calculation of the new principal curvatures behind the surface. This is an extension of the well-known fact that e.g. a lens changes just the curvature of the Gaussian beam.

(ii) The new beam radii behind the surface are determined in such a way that the projection of the incident beam ellipse, i.e. the curve where the amplitude falls to 1/e of the maximum value, onto the surface taken along the incident beam is equal to the back projection of the beam ellipse onto the surface taken along the new beam direction.

The additional condition for this method which is also implemented in the program RAYTRACE is that the principal directions of the beam ellipse and the principal directions of the wave front curvature are parallel. Otherwise so-called general astigmatism [12] appears and the beam ellipse rotates with respect to the principal curvatures during free-space propagation. In principle, also the case of general astigmatic Gaussian beams can be treated by using a similar method. But in this case the equations become more complex.

5. WAVE PROPAGATION USING THE ANGULAR SPECTRUM OF PLANE WAVES

A complex wave amplitude $u(x,y,0)$ given in a plane perpendicular to the z-axis at the point $z=0$ can be propagated in a homogeneous medium to a parallel plane at $z=z_0$ by using the angular spectrum of plane waves. This is due to the facts that plane waves are an exact solution of the wave equation and that the wave equation is linear. The latter means that also the superposition of different plane waves is a solution of the wave equation. The resulting wave amplitude $u(x,y,z_0)$ can be calculated from $u(x,y,0)$ in the following well-known way [10].

Calculate the Fourier transform $\tilde{u}(\nu_x,\nu_y)$ from $u(x,y,0)$, i.e. decompose the wave amplitude into a spectrum of plane waves:

$$\tilde{u}(\nu_x,\nu_y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} u(x,y,0)\exp\left[-2\pi i\left(\nu_x x+\nu_y y\right)\right]dx\,dy \tag{22}$$

Multiply the function $\tilde{u}(\nu_x,\nu_y)$ with the propagator $\exp\left[2\pi i\, z_0\sqrt{1-\lambda^2\left(\nu_x^2+\nu_y^2\right)}/\lambda\right]$, which describes the phase change of a plane wave with the spatial frequency (ν_x,ν_y) by propagating a distance z_0 to a parallel plane, and make an inverse Fourier transformation:

$$u(x,y,z_0) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\tilde{u}(\nu_x,\nu_y)\exp\left[2\pi i\, z_0\sqrt{1-\lambda^2\left(\nu_x^2+\nu_y^2\right)}/\lambda\right]\exp\left[2\pi i\left(\nu_x x+\nu_y y\right)\right]d\nu_x\,d\nu_y \tag{23}$$

So, by performing two Fourier transformations and multiplying a function with another the complex wave amplitude can be propagated between two parallel planes. In the approximation of the scalar theory this method is exact.

In practice, the complex wave amplitude will be sampled at equidistant points giving an array of N x N complex numbers, whereby N is a power of two so that the Fast Fourier transformation (FFT) can be used.

The limitations of the method are caused by the limited number N of samples which can be handled in practice. If D is the diameter of the field in the spatial domain the maximum spatial frequency which can be represented is

$$\nu_{\max} = \frac{N}{2D} \tag{24}$$

So, to avoid aliasing in the spatial frequency domain the maximum spatial frequency in the wave front has to be smaller than ν_{max}.

Second, there can also occur aliasing in the spatial domain during the second FFT. This means that parts of the wave reach an edge of the field with diameter D and enter the field at the opposite edge. To avoid this, the field can be embedded in a larger field with zeros. In some cases it may be also intended that aliasing occurs during the propagation in the spatial domain. This is e.g. the case if the Talbot effect for an infinite array of microlenses has to be simulated [13]. In the following example this will be demonstrated.

It has to be emphasised that the method of the angular spectrum of plane waves is just a method to propagate a complex wave amplitude from one plane to another parallel plane. The change of the complex wave amplitude at optical elements has to be calculated separately. In section 6 we will try to combine ray tracing and the method of the angular spectrum of plane waves.

Example:

The Talbot effect of a periodic orthogonal array of refractive microlenses with an infinite number of lenses can be simulated by calculating the complex wave amplitude behind a single lens and using the aliasing effect of the FFT during propagation of the angular spectrum of plane waves in the spatial domain [13]. The simulation was made again with the program RAYTRACE. A refractive planoconvex microlens with a diameter of 200µm and a focal length of 1mm is taken. The best focus due to aberrations is at 0.965mm behind the lens vertex. The lens is illuminated from the plane side with a plane wave with 633nm wavelength. The foci in the different fractional Talbot planes behind the lens best focal plane are shown in the following figures (fig. 7). As is known from the theory of the fractional Talbot effect the number of foci is tripled in the 1/6 and 1/3 Talbot plane, whereby in the 1/6 Talbot plane a lateral shift of 1/6 period occurs and no shift in the 1/3 Talbot plane. The number of foci in the ¼ Talbot plane is doubled. In the ½ Talbot plane there is only one focus per subaperture but the foci are shifted laterally by half a period.

It can be also seen that the diameter of the foci increases with increasing distance to the 0. Talbot plane. This is due to the fact that the Talbot effect occurs exactly if the Fresnel approximation for the wave propagation is strictly valid, i.e.

if the square root of the propagator in equation (23) can be replaced by $1-\frac{1}{2}\lambda^2(\nu_x^2+\nu_y^2)$. If this is not the case the foci become larger in the fractional Talbot planes with increasing distance to the lens array. The propagation with the angular spectrum of plane waves takes this correctly into account.

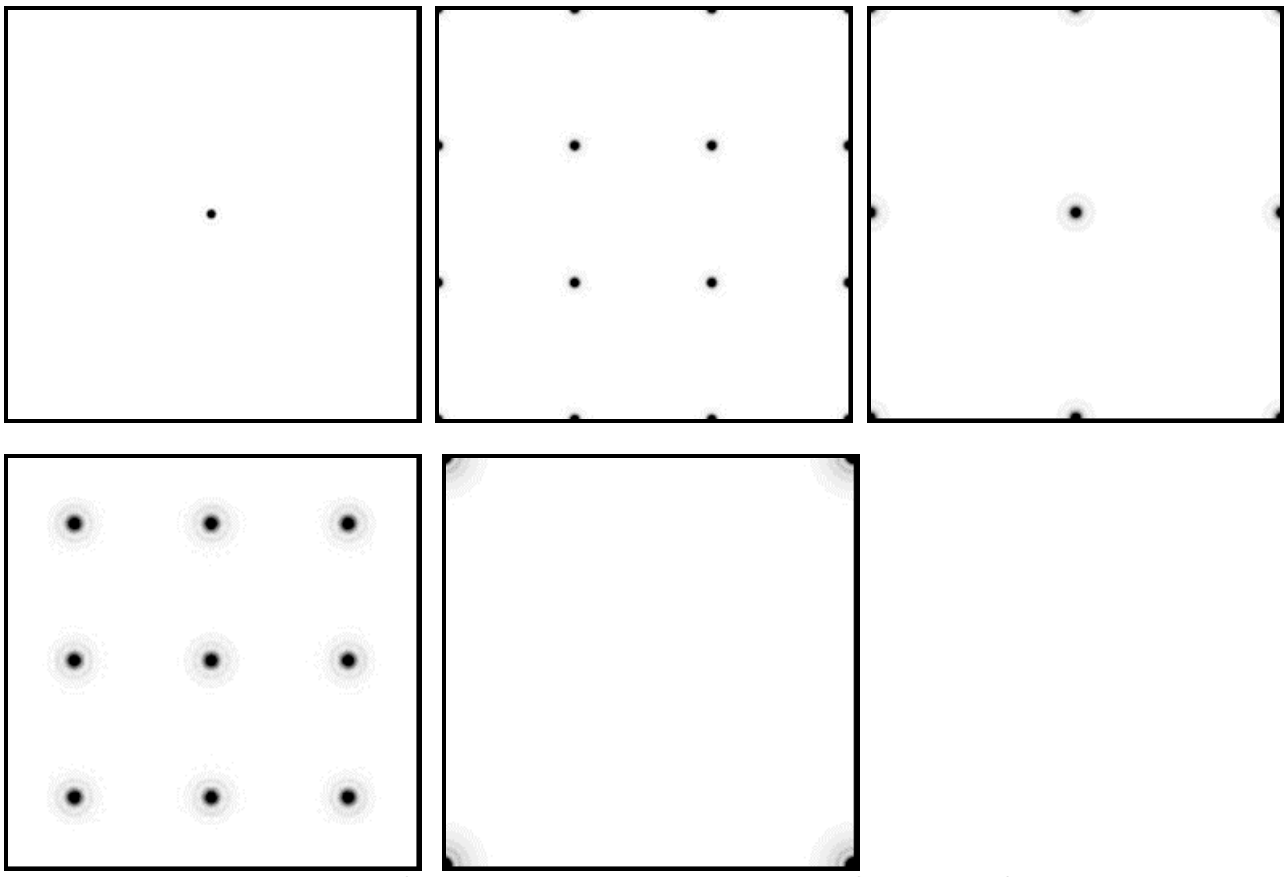

Fig. 7: Intensity distribution in the fractional Talbot planes behind the best focus plane of a lens array. Shown is in every case a subaperture of 200µm diameter which is the period of the lens array. From left to right and top to bottom is shown: 0. Talbot plane, 1/6 Talbot plane, ¼ Talbot plane, 1/3 Talbot plane and ½ Talbot plane.

6. COMBINATION OF RAY TRACING AND WAVE PROPAGATION METHODS

The standard method in wave optical calculations is that the free-space propagation is either made with Fraunhofer diffraction (for the far field), Fresnel diffraction (for the "middle field") or angular spectrum of plane waves (for the near field). So, the free-space propagation is correct if the correct propagation method is taken. In the program HAPPY the user can select one of these propagation methods.

The change of the complex wave amplitude at a lens or another optical element (e.g. diffractive optical element, aperture, filter, …) is normally taken into account by multiplying the complex wave amplitude with the complex transmission function of the element. If the paraxial approximation is valid an efficient method using the ABCD matrix of the system can be used [4]. This method calculates the complex wave amplitude behind a system with matrix ABCD by evaluating a special integral which uses the coefficients A, B, C and D of the matrix and Fresnel diffraction for the free-space propagation. If there is an aperture or another optical element (e.g. filter, general diffractive optical element, …) in the system which cannot be described by a normal ABCD matrix the complex wave amplitude has to be multiplied separately with the transmission function of this element. So, by dividing the complete optical system into several subsystems which can be described with ABCD matrices and other elements which cannot be described with such matrices the wave propagation through a complete system can be made quite efficient if the paraxial approximation is valid.

If the paraxial approximation is not valid this method cannot be used. A lens is e.g. normally described by just adding a quadratic phase term to the phase of the incident wave. This neglects of course aberrations of the lens and for increasing numerical apertures the quadratic phase term is also no more correct even if the lens has no aberrations. Of course, the latter error can be corrected easily but the first error (aberrations) cannot be corrected easily because the aberrations of a lens depend on the incident wave front.
So, it is obvious that a combination of ray tracing, which calculates the aberrations of a lens in dependence of the incident wave front, and wave-optical propagation methods like the angular spectrum of plane waves is quite promising. Although the propagation with the angular spectrum of plane waves is also implemented in the program RAYTRACE it is in the current version not possible to convert a complex wave amplitude back into rays. Since RAYTRACE is written in C it is useful to implement tested methods but not to test something if there is a faster solution. Therefore, we developed interfaces between the two programs RAYTRACE (calculation of aberrations via ray tracing) and HAPPY (wave-optical propagation). The advantage of HAPPY is that it is written in MATLAB so that the development of new procedures is quite fast.

To convert a wave, which is given in a plane, into rays two conditions have to be fulfilled with our solution:

- There is no wave front warping and the plane is far away from a focus.
- There are no discontinuities in the wave front.

For the back conversion of rays into a wave (complex wave amplitude) the following conditions have to be fulfilled:

- The rays have to form a nearly equidistant mesh of data points in the considered plane. “Nearly” means here that it must be possible to cover the data points with an orthogonal mesh so that in each subaperture of the mesh there is exactly one ray.
- It is again not allowed that wave front warping occurs. This means that it is not allowed that two rays are at the same position or that they change their places with respect to the ray mesh.

If all of these conditions are fulfilled the conversion can be made.

Conversion wave into rays:
The complex wave amplitude is given in a plane with coordinates x,y as a two-dimensional array of data points:

$$u(x_m, y_n) = A(x_m, y_n)\exp\left[i\Phi(x_m, y_n)\right] \tag{25}$$

The coordinates x_m and y_n form an equidistant mesh. Here, we assume that the phase is a continuous function. Since the wave-optical propagation methods deliver the phase only modulo 2π a phase continuation algorithm like in interferometry has to be used.
The ray direction vector is then of course given by:

$$\underline{e} = \frac{\lambda}{2\pi}\left(\frac{\partial\Phi}{\partial x} \quad \frac{\partial\Phi}{\partial y} \quad \sqrt{\left(\frac{2\pi}{\lambda}\right)^2 - \left(\frac{\partial\Phi}{\partial x}\right)^2 - \left(\frac{\partial\Phi}{\partial y}\right)^2}\right) \tag{26}$$

λ is the wavelength in the medium.
Since the phase function Φ is only known at discrete coordinates it is necessary to use equations for a numerical differentiation. So, to avoid errors the differential quotient should be calculated by using higher order equations. Otherwise, there are errors, especially at the rim of the data array where only neighbours at two or three sides are available.
The light power transported by a ray is set to A^2. This is possible since we have an orthogonal mesh of data points where each ray represents the same surface area.

Conversion rays into wave:
In this case, data $\Phi(x_m,y_n)$ and $P(x_m,y_n)$ are known. P means here the light power represented by a ray. Unfortunately, the coordinates (x_m,y_n) are not equidistant, but the phase and amplitude has to be known in an equidistant mesh.
For the determination of the phase Φ two things are made. First, an orthogonal equidistant mesh is calculated which fits best to the coordinates (x_m,y_n). This is necessary because the mesh size changes between two planes if the wave has a spherical part. Second, the phase value at the center $(\hat{x}_m, \hat{y}_n)$ of each subaperture of the mesh, which is not identical to the coordinate (x_m,y_n) in general, is calculated by

$$\Phi(x_m+\Delta x, y_n+\Delta y)=\Phi(x_m,y_n)+\Delta x\frac{\partial\Phi(x_m,y_n)}{\partial x}+\Delta y\frac{\partial\Phi(x_m,y_n)}{\partial y}=$$
$$=\Phi(x_m,y_n)+\frac{2\pi}{\lambda}\Delta x\,e_x+\frac{2\pi}{\lambda}\Delta y\,e_y \qquad (27)$$

with

$\Delta x=\hat{x}_m-x_m$, $\Delta y=\hat{y}_n-y_n$. e_x and e_y are the components of the direction vector of the corresponding ray.

To determine the amplitude A of the wave it is in general not sufficient to just take the square root of the ray light power P because the mesh distortion between the equidistant mesh before the ray tracing and the actual mesh after the ray tracing is in this case not taken into account. Therefore, we calculate the intensity $I=A^2$ of the wave by taking into account the local mesh distortion. If I_0 is the light intensity without mesh distortion, i.e. $I_0=P/a^2$ (a is the diameter of a subaperture of the best fitting orthogonal and equidistant mesh), we have

$$I=I_0/F \quad \text{with} \quad F=\left|\underline{r}_x\times\underline{r}_y\right|=\left|\frac{\partial x}{\partial\hat{x}}\cdot\frac{\partial y}{\partial\hat{y}}-\frac{\partial x}{\partial\hat{y}}\cdot\frac{\partial y}{\partial\hat{x}}\right| \quad \text{and} \quad \underline{r}_x=\left(\frac{\partial x}{\partial\hat{x}}\quad\frac{\partial y}{\partial\hat{x}}\quad 0\right),\ \underline{r}_y=\left(\frac{\partial x}{\partial\hat{y}}\quad\frac{\partial y}{\partial\hat{y}}\quad 0\right) \qquad (28)$$

The coordinates (x,y) are here the actual coordinates of the rays in the plane, whereas $(\hat{x},\hat{y})$ are the coordinates of the equidistant and orthogonal best fitting mesh.

By doing this the amplitude is calculated correctly if the intensity increases in an area because the rays are there more close to each other than in another area.

So, methods to convert a wave into rays and vice versa are known. The actual calculation of the micro-optical system is made by propagating the wave from one element to the other taking wave-optical methods and converting the wave into rays before each optical element. The optical elements are passed by ray tracing and behind the element the rays are converted back into a wave.

A disadvantage of this method is that no wave front warping is allowed. But wave front warping appears at the rim of the field if diffraction at an aperture takes place.

So, there are two possible solutions for this problem:

1. The complex wave amplitude has to be decomposed into elementary waves like e.g. a plane wave spectrum. Then, each component of the wave is converted into rays and traced trough an optical element. Behind the element all components of the back-converted waves have to be superimposed by taking the complex coefficients of the decomposition. Of course, this method needs a lot of computation time.
2. A simpler method would be to just take the spherical part (and of course tilt if it is present) of the wave to calculate the aberrations of the optical element by ray tracing and then to add the wave aberrations to the phase of the complex amplitude array, whereby the curvature of the spherical part of the wave (or torical part if the wave is astigmatic) has to be stored separately. This neglects of course the influence of the aspheric part of the incident wave to the aberrations of the optical element. But normally, these contributions should be of higher order.

There are of course a number of unsolved problems in both methods.

7. CONCLUSION

Several methods for the simulation of a micro-optical system have been discussed.

A simple matrix method which is also valid for off-axis systems (as long as the elements are in a plane) allows the calculation of paraxial properties of the system and therefore a first paraxial design.

Ray tracing can be used to simulate complete arrays of microlenses as long as diffraction effects can be neglected and if the light is incoherent. In other cases ray tracing can be used at least to calculate the aberrations of an optical element.

The propagation of Gaussian beams was discussed as long as aberrations and diffraction at apertures can be neglected.

For the free-space propagation of a wave the angular spectrum of plane waves was used since it is an exact method as long as the scalar theory is valid and as long as the numerical limitations which were discussed are no problem.

At the end a kind of synthesis between wave-optics and geometrical optics (ray tracing) has been proposed. But there is still a lot of work to be done before a reliable method for the wave-optical simulation of micro-optical systems is available.

Most of the discussed methods are also implemented in commercial optical design programs together with other concepts which we did not discuss here (e.g. polarisation effects) or which are not known to us because they are secret. Nevertheless, to our opinion also the commercial programs do not have a method which allows the exact simulation of an arbitrary micro-optical system, especially not without the intervention of an experienced user. So, also in the future the development of simulation tools will continue and the experience of an optical designer who knows the limitations of the different simulation methods will still be needed for a long time.

REFERENCES

1. N. Lindlein, F. Simon, J .Schwider, "Simulation of micro-optical array systems with RAYTRACE", *Opt. Eng.* **37**(6), 1809-1816 (1998).
2. Urs Vokinger, *Propagation, Modification and Analysis of Partially Coherent Light Fields*, (UFO Atelier für Gestaltung & Verlag, Allensbach, 2000), Band 387.
3. H. Kogelnik, T. Li, "Laser Beams and Resonators", *Appl. Opt.* **5**(10), 1550-1567 (1966).
4. A. E. Siegman, *Lasers*, (University Science Books, Mill Valley, California, 1986).
5. G. H. Spencer, M. V. R. K. Murty, "General Ray-Tracing Procedure", *J. Opt. Soc. Am.* **52**(6), 672-678 (1962).
6. M. Born, E. Wolf, *Principles of Optics*, (Cambridge University Press, Cambridge, 1980) 459-490.
7. O. N. Stavroudis, *The Optics of Rays, Wavefronts, and Caustics*, (Academic, New York, 1972), 136-179.
8. J. A. Kneisly, "Local curvature of wave-fronts in an optical system", *J. Opt. Soc. Am.* **54**(2), 229-235 (1964).
9. N. Lindlein, J. Schwider, "Local wave fronts at diffractive elements", *J. Opt. Soc. Am. A* **10**(12), 2563-2572 (1993).
10. J. W. Goodman, *Introduction to Fourier Optics*, (McGraw-Hill, New York, 1988).
11. R. Herloski, S. Marshall, R. Antos, "Gaussian beam ray-equivalent modeling and optical design", *Appl. Opt.* **22**(8), 1168-1174 (1983).
12. J. A. Arnaud, H. Kogelnik, "Gaussian Light Beams with General Astigmatism", *Appl. Opt.* **8**(8), 1687-1693 (1969).
13. B. Besold, N. Lindlein, "Fractional Talbot effect for periodic microlens arrays", *Opt. Eng.* **36**(4), 1099-1105 (1997).