Review Article

Taimoor Talpur* and Alois Herkommer

Review of freeform TIR collimator design methods

DOI 10.1515/aot-2016-0003
Received January 5, 2016; accepted February 29, 2016; previously published online March 25, 2016

Abstract: Total internal reflection (TIR) collimators are essential illumination components providing high efficiency and uniformity in a compact geometry. Various illumination design methods have been developed for designing such collimators, including tailoring methods, design via optimization, the mapping and feedback method, and the simultaneous multiple surface (SMS) method. This paper provides an overview of the different methods and compares the performance of the methods along with their advantages and their limitations.

Keywords: design methods; extended source design; illumination design; optimization methods; TIR collimator.

1 Introduction

Most light sources in our everyday life employ one or the other form of collimation optics. With increasing efficiency and output flux of light-emitting diode (LED) sources, adequate collimator designs are vitally important to completely utilize the LED power and to shape the output. Optimized collimators not only improve the efficiency but can also increase the quality of lighting in terms of color temperature, color rendering, and homogeneity.

Total internal reflection (TIR) collimators provide high efficiency and low losses due to total internal reflection and, therefore, are most widely used. Geometrical design criteria for such collimators typically include collimator dimensions, target field distance, and target field size. The main performance criteria are efficiency within the target field, light output homogeneity, light mixing, source imaging, and so on. The final limitations usually result from etendue restrictions, especially from finite source sizes. It is the general task during the optical design process to find the optimum mapping of the source radiance to the target field. Various design methods exist to provide this mapping, and a certain nomenclature for the methods has been established, which we will adopt for this paper, even so all of them provide some sort of mapping: the tailoring method [1], design via optimization, the mapping and feedback method [2], and the simultaneous multiple surface (SMS) method [3]. These different design methods allow improved control of one or the other parameter; therefore, selection of the design method is crucial in achieving the desired performance.

In this paper, the different methods are compared in terms of performance along with their benefits and drawbacks for the design of TIR collimators. First, an optical design achieved by the point source mapping method is presented. The design performs well with the point source, but the performance drops with the use of the extended source. The deviation in performance with the size of the extended source is determined, and feedback and optimization methods are then used to optimize the design for an extended source. Finally, the results are compared to designs achieved with the simultaneous multiple surface (SMS) method. Those methods have already briefly been compared in a short proceeding [4] by the authors. In this work, we extend this work and give a detailed overview as well as a discussion of design examples.

2 Tailoring method

The tailoring method, first proposed by Maes and Janssen in 1991 [5], is based on a mathematical relation between the acceptance angle and the target distribution. This relationship between the source and the target field results in a set of differential equations. The shape of the optical surface is obtained by solving this set of differential equations. Maes and Janssen in reference [5] designed a linear symmetric reflector for a linear light source to generate a prescribed intensity distribution. In 1993, Winston and Ries extended the method to several designs for small sources in reference [1] and later for extended sources

***Corresponding author: Taimoor Talpur,** Institut für Technische Optik, Universität Stuttgart, Pfaffenwaldring 9, D 70569 Stuttgart, Germany, e-mail: talpur@ito.uni-stuttgart.de
Alois Herkommer: Institut für Technische Optik, Universität Stuttgart, Pfaffenwaldring 9, D 70569 Stuttgart, Germany

www.degruyter.com/aot

using edge-ray principles in reference [6]. In 1994, Rabl and Gordon utilized the edge ray principle to obtain additional solutions for the extended sources in reference [7]. Ong in 1996 used the tailored edge-ray principle to extend the designs of Rabl in reference [7] for an extended lambertian source [8].

In 2002, Ries and Muschaweck extended the method to three-dimensional freeform surfaces [9]. The two-dimensional tailoring method leads to ordinary differential equations; however, the three-dimensional tailoring leads to very complicated and highly nonlinear partial differential equations. These nonlinear partial differential equations were solved numerically in reference [9] to yield the shape of the freeform optical surface. Zehnrong in 2009 used Runge-Kutta method to solve the differential equations to the design freeform LED lens [10]. These nonlinear partial differential equations require significant efforts to determine theoretical solutions, and the corresponding numerical algorithms are non-trivial. Usually, the Runge-Kutta method was used to get an approximate solution. Oliker in 2005 proposed geometric and variational methods to solve such complex nonlinear problems [11]. As the numerical codes are not publicly available and are quite complex, this method was not further investigated within this paper. Generally, tailoring is a promising method for developing source size-adapted design types.

3 Mapping method

The mapping method was first proposed by Parkyn in 1998 [2] to design general illumination lenses. The method involves two steps: ray-mapping and surface generation. The schematics of the mapping method are shown in Figure 1.

The ray mapping involves parameterization of the source and the target fields. Various approaches are defined for optimum parameterization targeting smooth transition and better mapping. W. Parkyn proposed two grids on unit spheres based on constant flux and intensity distribution of the source [2]. In 2006, B. Parkyn defined a pseudo-rectangular spherical grid, which establishes correspondence between spherical geometry of source emission and the rectangular cells of the target [12]. In 2007, Wang defined a variable separation mapping method [13]. The mapping is obtained by solving the energy conservation differential equation. These mappings methods result in normal vectors, which lead to discontinued curves. Forcing surface continuity causes slope errors leading to an inaccurate target distribution. Florian in 2010 defined integrability constraints for continuous optical surface and proposed methods to achieve mappings fulfilling the integrability conditions [14]. In 2015, Ma described a composite ray mapping, which is a combination of traditional spherical and modified spherical (u,v) coordinate systems [15]. The source intensity distribution is parameterized in the peripheral region using the traditional spherical coordinate system and in the central region using the modified spherical coordinate system. The mapping reduces the surface error resulting in improved target distribution uniformity.

After mapping, the surface normals are computed using laws of refraction and reflection to direct the light from each grid to the corresponding target location. The surface is generated by either numerical integration of the partial differential equation [2, 13], or the normals and surface points are computed iteratively via intersection of light ray with local tangent planes [13]. In the end, a surface fit is applied on the computed points.

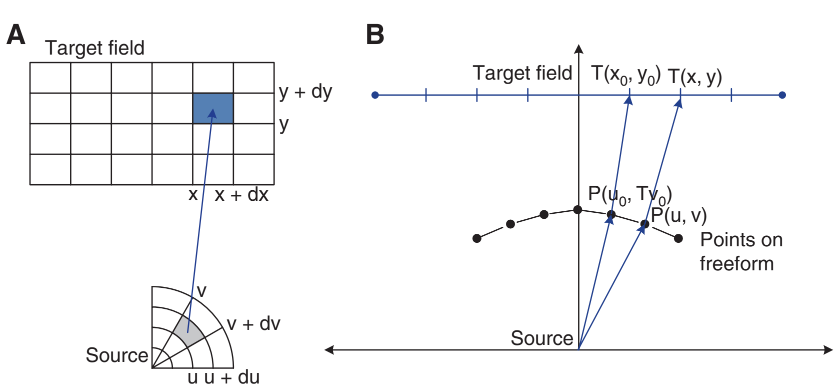

Figure 1: Visualization of the mapping method. (A) Variable separation mapping of source to target plane and (B) freeform surface construction.

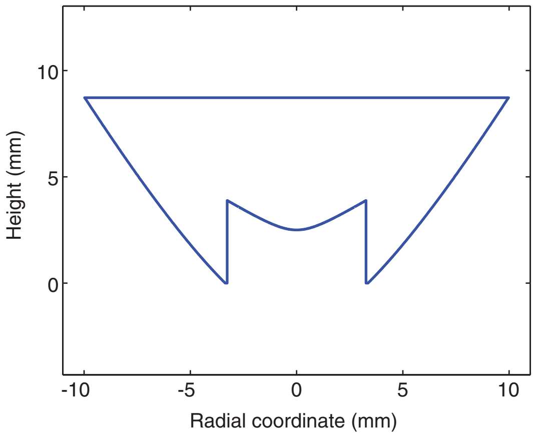

Figure 2: TIR collimator dimensions.

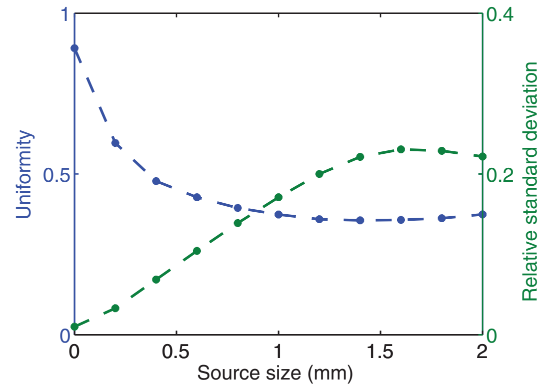

Figure 4: Variation in performance of the design achieved by the mapping method with increasing radial size of the source disk.

The method described by Wang [13] is implemented in MATLAB to design a rotational symmetric collimator for a point source and a homogenous circular output. The dimensions for the target field are set to a radius of 250 mm at a distance of 1000 mm. The dimensions and shape of the collimator are shown in Figure 2. The top surface is selected to be flat with a total height of 10 mm and diameter of 18 mm. The simulation results in Zemax at the far-field target plane are presented in Figure 3 and show a homogeneous output for the point source. However, as the mapping method is based on a point source, the performance drops as the design is used in combination with an extended source. This performance drop with increase in the radial size of the source disk from 0 mm (point source) to 2 mm is depicted in Figure 4.

Uniformity and relative standard deviation are used as performance metrics in this research work, and are are defined as

$$\text{Uniformity, } U = \frac{E_{min}}{E_{avg}}, \tag{1}$$

$$\text{Relative standard deviation, } RSD = \frac{E_{sd}}{E_{avg}}, \tag{2}$$

where E_{min} is the minimum value of the illuminance distribution, E_{avg} is the average value of the illuminance distribution, and E_{sd} is the standard deviation of the illuminance distribution.

In order to recover the performance, the design is adapted to the extended source in the following sections

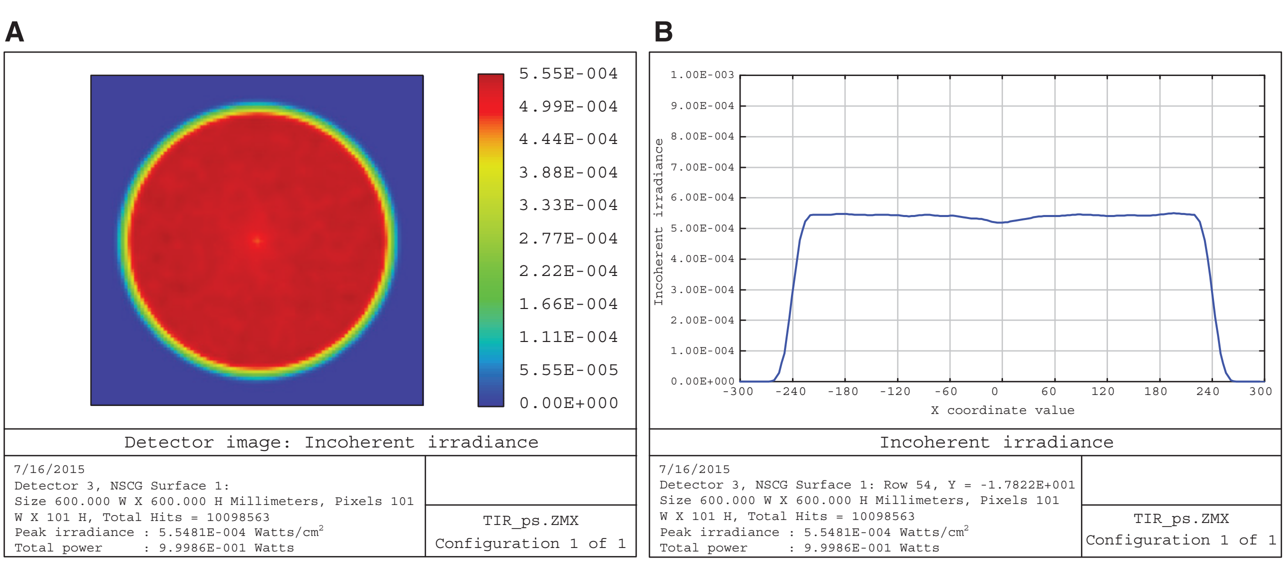

Figure 3: Simulation results of the design achieved with the mapping method for a point source at the target plane at 1-m distance. (A) Two-dimensional detector view, and (B) cross-sectional view.

with either modification of the output distribution via a feedback method or via optimization.

4 Feedback method

The feedback method is an iterative procedure in which the design is adapted iteratively based on a feedback from the last iteration (Figure 5). Typically, the initial design is created by a simple method with a point source, a lambertian source, or an isotropic source. The effects of a real source in terms of radiation distribution or dimensions are then compensated by the iterative feedback. The required illuminance distribution in the current iteration depends on the actual illuminance output from a real source in the previous iteration. The illuminance distribution in the current iteration is given by

$$E_{(i+1)}(x)=\eta_i(x)E_o(x), \tag{3}$$

where $E_{(i+1)}$ is the illuminance distribution in the current iteration, η is the feedback function, and E_o is the desired illumination distribution. The feedback function, η, is defined as

$$\eta_i(x)=\frac{E_i(x)}{E_i'(x)}, \tag{4}$$

where E_i is the required illumination distribution of the ith iteration, and E_i' is the simulation result with a real source of the ith iteration.

This method has been used by several groups in combination with various design methods including tailored edge design [16], functional method [17], and variable separation mapping method [18] to improve the performance. In general, the results of the feedback method are good if the image of the source is much smaller than the target field dimensions. Typically, the results improve with a few iterations; however, the design is still an adaptation of a point source design to a real source; therefore, the improvement is limited.

We have implemented the feedback design method in MATLAB in combination with the mapping method mentioned in the last section. We have employed the algorithm to adapt the TIR collimator design for a point source to an extended source. Division feedback method was used iteratively to improve the output uniformity of the collimator for an extended lambertian disk source of 1 mm diameter. Figures 6 and 7 show the simulation results in Zemax before and after the feedback optimization, respectively. The figures clearly show that the feedback improves the system performance for the extended source quite dramatically, and a homogeneous output is achieved after

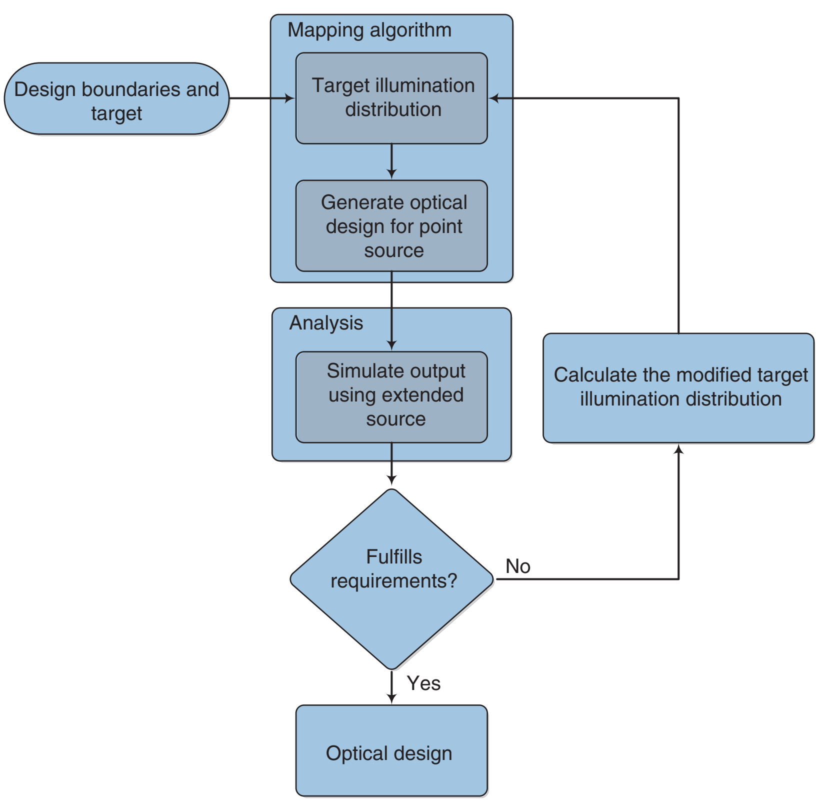

Figure 5: Flow diagram of the feedback design method.

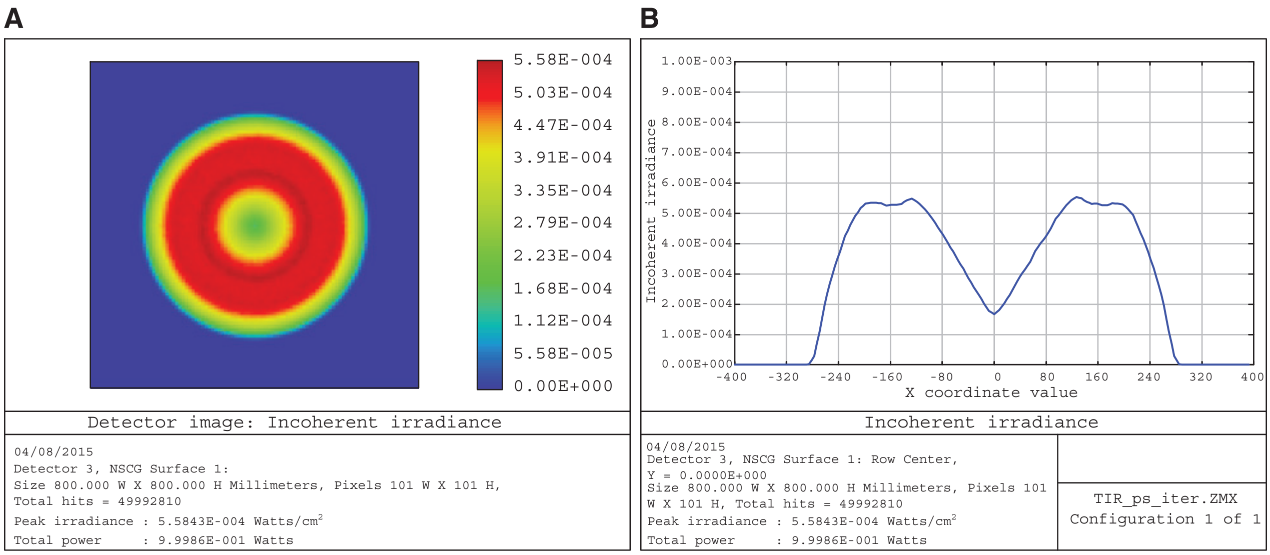

Figure 6: Simulation results at the target field at 1-m distance for the design achieved with the mapping method and a point source, but used with an extended source disk of diameter 1-mm. (A) Two-dimensional detector view and (B) cross-sectional view.

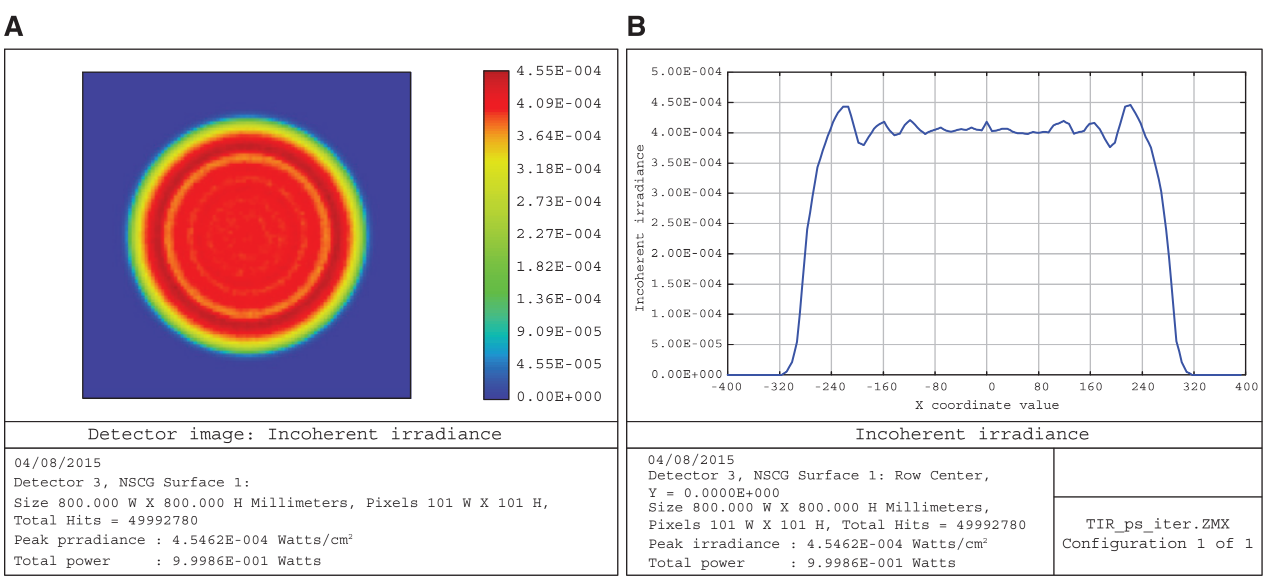

Figure 7: Simulation results at the target field at 1-m distance for the design achieved by the feedback method with an extended source disk of diameter 1 mm. (A) Two-dimensional detector view and (B) cross-sectional view.

15 iterations of the feedback method. The improvement in uniformity and relative standard deviation per iteration is shown in Figure 8 and proves that the system improves in only a few iterations, and after about eight cycles, the performance stays almost constant.

5 Optimization method

The above-mentioned illumination design methods are deterministic, i.e. an exact surface shape is calculated for a given input. However, many illumination design methods employ one or the other form of approximation to ease or fasten the design method, resulting in a compromised performance. To improve the performance further, or to adapt the design to real sources or ray-files, often, optimization methods are employed as the final design step.

Optimization contains three sub-factors: parameterization, merit function, and optimization algorithms. Parameterization is the method to define the surface shape with a finite set of variables. Most illumination surface profiles are freeform, which make parameterization difficult;

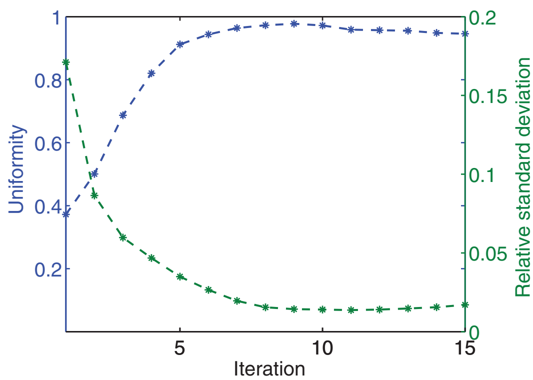

Figure 8: Performance improvement with feedback method vs. iteration.

also, the number of free variables can get quite large. The choice of the variables plays a critical role during optimization. It determines the distribution of the merit function minima as well as the convergence time. With good parameterization, the design can converge quickly to an optimum value. Typically, profile-based parameterization is used, which can be direct profile parameterization [19] or curve fitting such as by non-uniform rational B-splines (NURBs) [20]. Zhang recommends to use an indirect profile parameterization to overcome range definition, range overlapping, and parameter-dependent problems, by the slope angle-based parameterization [21].

The merit function is a set of functions and constraints along with their target values, which defines the goals of system optimization. Merit functions are typically a measure of uniformity, optical efficiency, and/or the concentration. Optimization changes system variables systematically to achieve the target values in the merit function. Various optimization algorithms are already developed and are a part of commercial illumination design tools, such as Brent's method, simulated annealing, and downhill simplex in ASAP [22] as well as damped least square, orthogonal descent, hammer, and global optimization in Zemax [23]. Various illumination design-specific optimization algorithms are also developed to improve the illumination design optimization such as edge-ray optimization and direct optimization method [24].

The merit function is typically computed via Monte Carlo ray tracing, which traces randomly selected rays. The statistical nature of the merit function causes some of the derivative-dependent optimization methods such as damped least square to be unstable. In such situations, derivative-free optimization algorithms are more favorable, such as orthogonal descent, simulated annealing, and direct optimization algorithm. The accuracy of the merit function increases with increase in the number of rays traced along with the decrease in the statistical noise. However, with an increase in the number of rays, the convergence time increases dramatically. Kudaev developed an edge-ray optimization algorithm [24], which traces only edge rays and, thus, reduces the number of rays and, in turn, the convergence time.

We also have employed an optimization method within Zemax to enhance the same collimator design

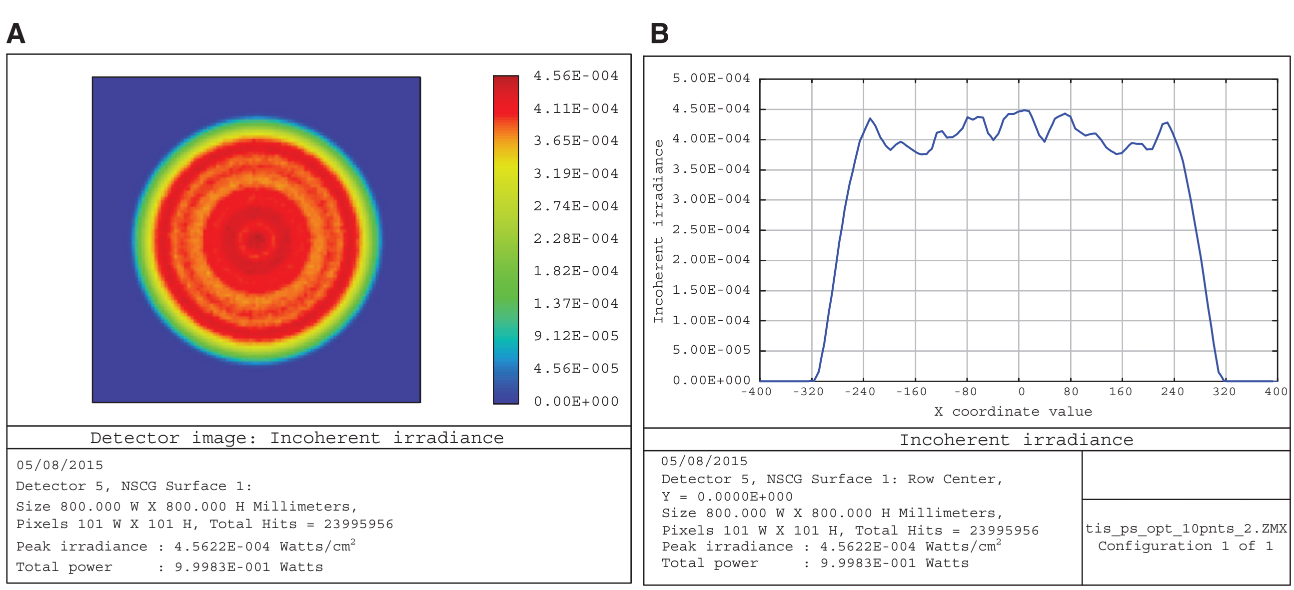

Figure 9: Simulation results at the target field at 1-m distance for the design achieved by the optimization method for an extended source disk of diameter 1 mm. (A) Two-dimensional detector view and (B) cross-sectional view.

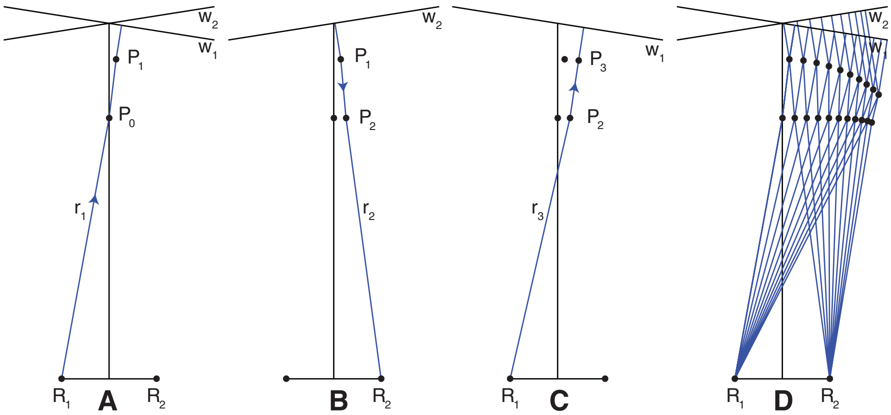

Figure 10: Steps of the SMS method to design a lens that couples two parallel wavefronts w1 and w2 to two spherical wavefronts R1 and R2 (picture taken from Ref. [27]). (a) P0 is freely chosen and P1 is determined to couple R1 to w1, (b) point P2 is determined to couple w2 to R2 satisfying the optical path length requirement. In step (c) the step (a) is repeated to obtain a new point P3 using the constant optical path length condition. This process is repeated to calculate alternate points, constructing an SMS chain of points, as shown in step (d).

resulting from the point source mapping method. Zemax freeform surfaces are used to model the TIR collimator. These freeform surfaces connect a series of points via cubic spline curve. The optimization method used is orthogonal descent, which does not compute numerical derivatives of the merit function. Therefore, it suits better for non-sequential systems with noisy merit functions. The results, depicted in Figure 9, show that the output is quite homogeneous. The method improves the uniformity from 37% to 93% and provides similar performance as the feedback method.

6 Simultaneous multiple surface (SMS)

The simultaneous multiple surface (SMS) method designs multiple surfaces simultaneously, hence, the name. Other first-order design generation methods utilize only one surface, which allows coupling of only one input wavefront to one output wavefront. SMS uses multiple surfaces to couple more than one input wavefronts to output wavefronts. This added degree of freedom allows the SMS method to consider extended sources. The SMS method considers source edge points; therefore, the designs do not suffer from loss of efficiency or smearing out of the output pattern with the extended source.

The SMS method was initially developed in 1990 as a two-dimensional method [3] primarily by Minano. In the two-dimensional designs, the designs are limited to linear or rotational symmetric systems, which inherently limits the method to symmetric optical problems. Nevertheless, the method is powerful and is able to create ultra-compact, highly efficient optical designs. To overcome the symmetry limitations, the method was extended by Benitez to a 3D design method in 1999 [25] and later generalized [26]. Three-dimensional designs are more difficult than two-dimensional designs, but they have more degrees of freedom and consequently provide better control over the rays.

The SMS design method is based on the determination of the multiple surfaces iteratively. The first surface couples the first pair of input-output wavefronts in the presence of the second surface and, in turn, the second surface is determined in a way that it couples the second pair of input-output wavefronts in the presence of the first surface as shown in Figure 10, which is taken from [27]. The whole procedure is computed point by point in an iterative way resulting in a final design, which perfectly couples the two pairs [26].

Light Prescriptions Innovators (LPI) has designed a variety of highly efficient illumination systems using the SMS design method including RR [28], XR [28], RX [28], XX [28], TIR-R Solar concentrator [29], RXI LED collimator [30–32], RXI LED Headlamp designs [33, 34], and imaging systems [35]. The acronyms R, X, and I stand for refractive, reflective, and total internal reflection, respectively.

Despite the tremendous results from the SMS method, there are some inherent limitations. SMS requires input and output wavefront definitions for the design. For a defined irradiance or intensity distribution, there is no deterministic method to derive the appropriate output wavefronts; however, approximate methods are used [36]. The output wavefronts need to fulfill the differential etendue requirements, which means that the local input etendue at a

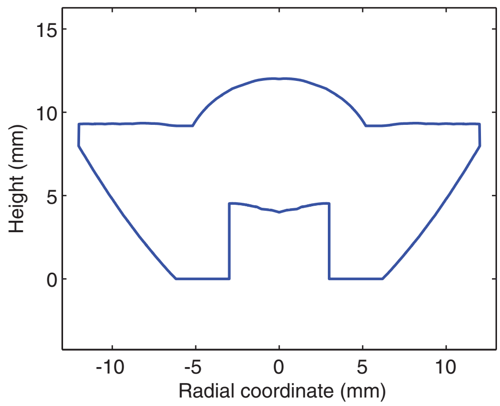

Figure 11: TIR collimator dimensions.

surface should match the required local output etendue. Violations of this will lead to surface discontinuities. Determination of such wavefronts is vital for the SMS design to be successful. Additionally, the extended sources are taken into account only by the edge points or edge wavefronts, and the resulting design is assumed to be well behaving [33]. For a non-homogeneous source, the inhomogeneities on the source will appear in the output.

Again, we have created our own version of the SMS method within MATLAB and used it to develop a similar TIR design as discussed in the previous sections. The design parameters, however, need to be modified to fulfill the etendue requirements of the SMS procedure. The dimensions for target field are set to a radius of 40 mm and at a distance of 1000 mm, which corresponds to an improved concentration compared to the previous designs. The dimensions and shape of the collimator are shown in Figure 11, with a height of 12 mm and diameter of 24 mm. The results simulated in Zemax (Figure 12) show quite homogeneous output with uniformity of 88% for a 1-mm diameter extended source.

7 Conclusion

In conclusion, we have given an overview of the different illumination design methods for freeform illumination components. In particular, we have created algorithms to test and compare these methods at the example of a TIR collimator. It turns out that an efficient TIR collimator design requires an adapted and well-chosen design method. The various design methods employ several approximations, and typically, a subsequent combination of methods is required to allow a deterministic design process toward a good optimum. Moreover, the different methods allow improved control of one or the other parameter; therefore, selection of the design method is crucial in achieving the desired performance. The mapping design method quickly provides design starting points using a point source; however, it lacks performance with the extended sources. The feedback design method improves the mapping method performance and delivers effective results in just a few iterations. In general, the point source and feedback method require the source image to be much smaller than the target field. The optimization method should be used in combination with all methods for the final tweaking of the optical system

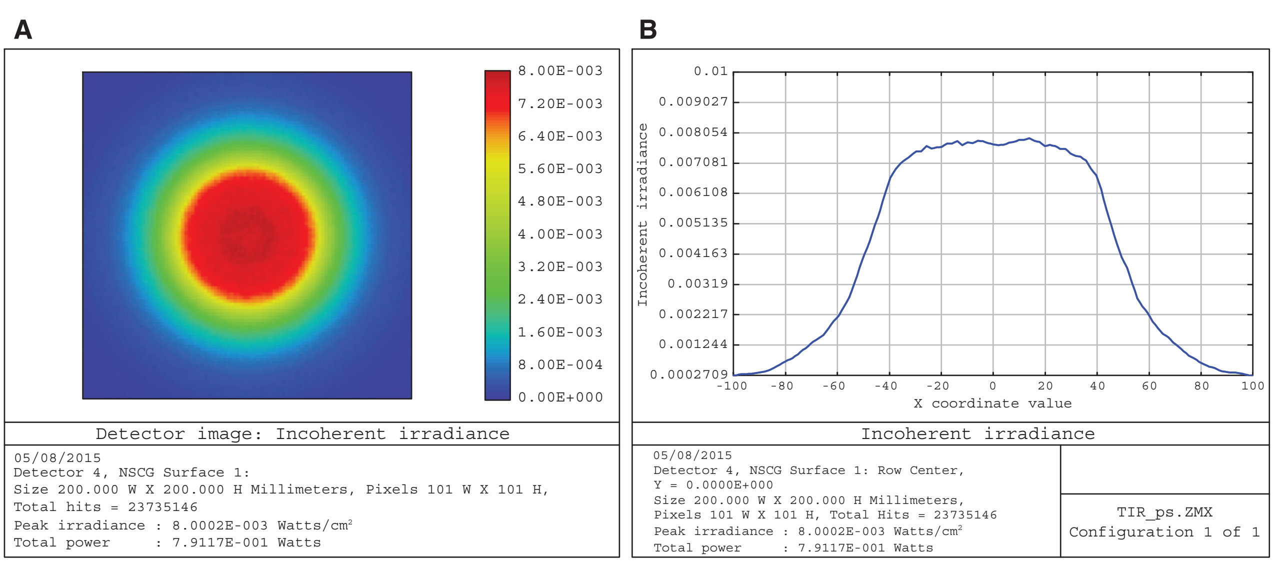

Figure 12: Simulation results at the target field at 1-m distance for the SMS design with an extended source disk of diameter 1 mm. (A) Two-dimensional detector view and (B) cross-sectional view.

to improve the system performance or to adapt the design to the real source. The performance and convergence highly depends on the initial system, number of free parameters, and the optimization algorithm used. The SMS method, on the other hand, provides a good starting point for extended sources. It creates etendue-limited designs but is based on a finite number of source/target points and requires definitions of etendue-based wavefronts.

Acknowledgments: The authors are grateful to the Zentrales Innovationsprogramm Mittelstand (ZIM) program for the support of this research under project number KF 2281403DF3.

Author contributions: T. T. performed software implementations, numerical calculations, design and simulations, and wrote the manuscript. A. H. participated in discussions, provided ideas, verified the calculations, designs and simulations, and contributed to editing the manuscript.

References

[1] R. Winston and H. Ries, J. Opt. Soc. Am. A 10, 1902–1908 (1993).
[2] W. A. Parkyn, Proc. SPIE 3428, 154–162 (1998).
[3] J. C. Miñano and J. C. González, Appl. Optics 31, 3051–3060 (1992).
[4] T. Talpur and A. Herkommer, Proc. SPIE 9629, 962906–962906–10 (2015).
[5] M. Maes and A. Janssen, Optik 88, 177–181 (1991).
[6] H. R. Ries and R. Winston, J. Opt. Soc. Am. A 11, 1260–1264 (1994).
[7] A. Rabl and J. M. Gordon, Appl. Optics 33, 6012–6021 (1994).
[8] P. T. Ong, J. M. Gordon and A. Rabl, Appl. Optics 35, 4361–4371 (1996).
[9] H. Ries and J. Muschaweck, J. Opt. Soc. Am. A 19, 590–595 (2002).
[10] Z. Zhenrong, H. Xiang and L. Xu, Appl. Optics 48, 6627–6634 (2009).
[11] V. Oliker, Proc. SPIE 5942, 594207–594207–12 (2005).
[12] B. Parkyn and D. Pelka, Proc. SPIE 6338, 633808–633808–7 (2006).
[13] L. Wang, K. Qian and Y. Luo, Appl. Optics 46, 3716–3723 (2007).
[14] F. R. Fournier, W. J. Cassarly and J. P. Rolland, Opt. Express 18, 5295–5304 (2010).
[15] D. Ma, Z. Feng and R. Liang, Appl. Optics 54, 498–503 (2015).
[16] A. Rabl, P. T. Ong, J. M. Gordon and W. Cai, Proc. SPIE 2538, 16–23 (1995).
[17] J. Bortz and N. Shatz, Proc. SPIE 6670, 66700A–66700A–15 (2007).
[18] Y. Luo, Z. Feng, Y. Han and H. Li, Opt. Express 18, 9055–9063 (2010).
[19] T. L. R. Davenport, T. A. Hough and W. J. Cassarly, Proc. SPIE 5456, 81–90 (2004).
[20] M. G. Turner and K. J. Garcia, Proc. SPIE 7061, 70610H–70610H–10 (2008).
[21] W. Zhang, Q. Liu, H. Gao and F. Yu, Opt. Eng. 49(6), 063003–063003–7 (2010).
[22] Breault Research Organization. ASAP. "The optical design software". Available at: http://www.breault.com/software/about-asap. Accessed 28.10.2015.
[23] Zemax, LLC. "Optical and illumination design software". Available at: http://www.zemax.com/. Accessed 28.10.2015.
[24] S. Kudaev and P. Schreiber, Proc. SPIE 5962, 59620B–59620B–9 (2005).
[25] P. Benitez, R. Mohedano and J. C. Miñano, Proc. SPIE 3781, 12–21 (1999).
[26] P. Bentez, J. C. Miñano, J. Blen, R. Mohedano, J. Chaves, et al., Proc. SPIE 5185, 18–29 (2004).
[27] F. Duerr, New free-form optics design methods for solar concentrators and imaging applications (PhD thesis, Vrije Universiteit Brussel, 2013).
[28] R. J. Koshel, ed., in 'Illumination engineering: design with nonimaging optics', (Wiley-IEEE Press, Piscataway, NJ, 2012).
[29] J. L. Alvarez, M. Hernandez, P. Bentez and J. C. Miñano, Proc. SPIE 4446, 32–42 (2001).
[30] J. C. Miñano, J. C. Gonźlez and P. Bentez, Appl. Optics 34, 7850–7856 (1995).
[31] F. Muñoz, P. Bentez, O. Dross, J. C. Miñano and B. Parkyn, Opt. Eng. 43, 1522–1530 (2004).
[32] D. Grabovickic, J. C. Miñano, P. Bentez, C. Cocho and M. Buljan, Proc. SPIE 8124, 812404–812404–10 (2011).
[33] O. Dross, R. Mohedano, P. Benitez, J. C. Minano, J. Chaves, et al., Proc. SPIE 5529, 35–47 (2004).
[34] J. C. Miñano, P. Bentez and A. Santamara, Opt. Rev. 16, 99–102 (2009).
[35] J. C. Miñano, P. Bentez, W. Lin, J. Infante, F. Muñoz, et al., Opt. Express 17, 24036–24044 (2009).
[36] P. Bentez, J. C. Miñano, J. Blen, R. Mohedano, J. Chaves, et al., Opt. Eng. 43, 1489–1502 (2004).

Taimoor Talpur
Institut für Technische Optik, Universität Stuttgart, Pfaffenwaldring 9, D 70569 Stuttgart, Germany
talpur@ito.uni-stuttgart.de

Taimoor Talpur is a PhD research scholar at the Institute for Technical Optics, University of Stuttgart. He finished his post-graduation in Photonics from Abbe School of Photonics, at Friedrich Schiller University Jena, Germany in 2013. He is interested in the domain of optical modeling and design and has worked on Delano diagrams during his master thesis, in Carl Zeiss, Germany. Currently, he is working on TIR collimators for illumination systems for his PhD research.

Alois Herkommer
Institut für Technische Optik, Universität Stuttgart, Pfaffenwaldring 9, D 70569 Stuttgart, Germany

Alois Herkommer received his PhD in Physics from the University of Ulm in the area of quantum optics. In 1996, he joined Carl Zeiss in Oberkochen as an optical designer for lithography, inspection, and metrology systems. After 2000, he was senior optical designer at the Carl Zeiss Laser Optics GmbH, developing laser and spectral components, as well as X-ray optics. From 2006, he headed the optical design group at the Carl Zeiss SMT GmbH, responsible for the design of high-performance lithography systems. Since 2011, he is the professor for 'optical design and simulation' at the Institute for Technical Optics (ITO) at the University of Stuttgart. His main research interests are complex surfaces in optical design, design methods for illumination and imaging systems, and optical systems for biomedical applications. Prof. Alois Herkommer is a member of the SPIE, the OSA, and the German DGaO. He has published several articles on optical design and holds over 20 patents.