Holographic optical tweezers

Gabriel C. Spalding
Department of Physics, Illinois Wesleyan University, Bloomington, IL 61701, United States

Johannes Courtial
SUPA, Department of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, United Kingdom

Roberto Di Leonardo
CRS-SOFT INFM-CNR c/o Dipartimento di Fisica, Università di Roma “La Sapienza”, P.le A. Moro, 2, 00185, Roma, Italy

Abstract.

The craze for miniaturization
has swept o’er most every nation,
but should a hand c’er so slightly tremble
no micro-machine can it assemble:
and so all those really small bits
leave the technicians in fits
and the hope for a lab on a chip
might seem frightfully flip.

Yet, while optical forces are weak
they provide the control that we seek.
When light’s holographically defined,
HOT opportunities can be mined!
and so this review
contains quite a rich stew,
as we hope that you will soon find.

So dinnae fear phase modulation
nor the details of plane conjugation:
it’s simple, you’ll see,
we’ll guide you for free,
to make optical landscapes extended,
so your science can truly be splendid.
1. Background

All around the world there are hobbyists and collectors who take great pride in the workmanship put into miniaturizations, producing small train sets with clear windows and seats in the cars, tiny buildings with working doors and detailed interiors. Clearly though, as the size is further reduced, model building becomes more and more a painstaking task, and we are all the more in awe of what has been produced. Ultimately, as we move towards the microscopic scale, we find that we can no longer use, for the creation and assembly of machines – or even simple static structures – the same approaches that we would naturally use on the macroscopic scale: entirely new techniques are needed.

Luckily, Newton’s second law dictates that as the mass of the object involved is reduced (as we shrink the size of our components), then a correspondingly weaker force will suffice for a given, desired acceleration. So, while we would not think to use light to assemble the engine of a train, it turns out that optical forces are not only sufficient for configuring basic micro-machines, but that the optical fields required can be arrayed in ways that offer up the possibility of complex system integration (with much greater ease than, say, magnetic fields). Indeed, within clear constraints, optics provides a natural interface with the microscopic world, allowing for imaging, interrogation and control.

There is no doubt that these are powerful technologies. Optical forces have convincingly been calibrated [1] down to 25 fN, and recent efforts have similarly extended the range of optical torque calibration [2]. Careful studies of the optical torques exerted upon micro-scale components [3, 4, 5, 6], provided a much clearer understanding of wave-based spin and orbital angular momentum (both optical [7, 8] and, by extension and analogy, the quantum mechanical angular momentum of electrons in atoms). This work clearly deserves a place in the standard canon of the physics curriculum. As for the science that has resulted from the use of optical forces, consider the experiments done by Steven Block’s group at Stanford, whose apparatus (using 1064-nm light) now has a resolution of order the Bohr radius (!!) and who have used this exquisitely developed technique to directly observe the details of error correction in RNA transcription of DNA [9], a true tour de force. Another prime example might be the results from the Carlos Bustamante group at Berkeley, whose experimental confirmations [10, 11] of new fluctuation theorems have allowed recovery of RNA folding free energies, a key feat, given that biomolecular transitions of this sort occur under nonequilibrium conditions and involve significant hysteresis effects that had previously been taken to preclude any possibility of extracting such equilibrium information from experimental data. In fact, work now being done on fluctuation theorems (both theoretically and experimentally) is among the most significant in statistical mechanics in the past two decades. These theorems have great general importance, and include extension of the Second Law of Thermodynamics into the realm of micro-machines and biomolecules. Clearly such work opens up vast new intellectual opportunities.
2. Example rationale for constructing extended arrays of traps

For some while now there have been available basic optical techniques used to manipulate one or two microscopic objects at a time, but the focus of this review is upon recently emerging technical means for creating arrays of optical traps. The statistical nature of the experiments described above – which involve samples that are in the diffusive limit, where Brownian motion is significant – suggests that there might be a significant benefit to simultaneously conducting an array of optical trap experiments, by creating independent sets of traps across the experimental field of view.

At the same time, the very issue of whether or not such an array of experiments may be treated independently points also towards an altogether different class of studies, aimed at either probing or exploiting the wide array of physical mechanisms that might serve to couple spatially separated components. For example, biological studies of cell-cell signaling change character when, instead of dealing with a pair of cells isolated from all others, one deals with an ensemble, as this changes the conditions required for quorum sensing. In such cases, the use of an array of optical traps can ensure well-controlled, reproducible ensembles for systematic studies. As a specific instance, the early stages of biofilm formation are studied using various mutant strains of bacteria, with different biofilm-related genes deleted [12]; here, optical trap arrays are used to ensure geometric consistency from ensemble to ensemble. In fact, quite a wide variety of many-body problems are amenable to study using trap arrays [13, 14, 15, 16, 17, 18, 19].

It should be added that optical trap arrays determine not only the equilibrium structures that assemble, but also the dynamics of particles passing through the lattices [20, 21, 22, 23]: the flow of those particles that are influenced most by optical forces tends to be channelled along crystallographic directions in a periodic trap array (often referred to as an optical lattice). In such a lattice, the magnitude of the optical forces is an oscillatory function of particle size [24], which means that it is possible to tune the lattice constant so as to make any given particle type either strongly interacting with the light or essentially non-interacting [25]. Therefore, it is possible to perform all-optical sorting of biological/colloidal suspensions and emulsions [24] and strong claims have been made as to the size selectivity associated with this type of separation technology [26]. In any case, this approach does allow massively parallel processing of the particles to be sorted, and so the throughput can be much higher than with active microfluidic sorting technologies (e.g., Ref. [27], where particles are analyzed one by one and then a deflection control decision is based upon feedback from that analysis). Moreover, while the laser power delivered to the optical lattice may be, in total, significant, the intensity integrated over each biological cell can be a fraction of what is used in conventional optical tweezers, so all-optical sorting does relatively little to stress the extracted cells [28]. There remain many opportunities for studies of colloidal traffic through both static and dynamic optical lattices [29, 30], and real potential for the use of optical trap arrays in microfluidic, lab-on-a-chip technologies.

While trap arrays have been put to a number of good uses in microfluidics (e.g., for
Holographic optical tweezers

multipoint micro-scale velocimetry [31]), it bears repeating that optical forces are, in the end, still relatively weak, and so for some purposes it would be reasonable to combine the use of optical forces (to allow sophisticated, integrated assembly of components) with the use of other forces for actuation. That said, the use of optical forces does allow for construction of simple micro-machines: components may be separated or brought together (even allowing ship-in-a-bottle-type assembly, e.g., of axel-less micro-cogs in a microfluidic chamber [32]), oriented (e.g., for lock-and-key assembly of parts), and even actuated.

Here, we have mentioned only a few of the many reasons one might be interested in constructing extended arrays of traps. Clearly, this is a field where a number of new applications are expected to emerge over the coming years. Our next task, then, is to describe in detail some of the techniques used for generation – and dynamic reconfiguration, in three dimensions – of optical trap arrays.

Before going on, though, we should note that the holographic techniques we describe are not limited to array formation, but also enhance the sorts of control one can exert over a single trap, through their ability to shape the optical fields in three dimensions; e.g., to create an optical bottle beam [33, 34], or even arrays of bottle beams [35]. A bottle beam is akin to a bubble of light (a dark region of space completely enclosed by a skin of high intensity fields), and is intended for trapping cold atom clouds (“The Atom Motel ... where atoms check in but they don’t check out!”).

3. Experimental details

3.1. The standard optical train

The majority of experiments involving optical traps use the single-beam gradient trap geometry [36] (referred to as optical tweezers), and we will focus upon this particular sort of setup. The most important details of experimental realizations of optical tweezers are discussed elsewhere, for example Ref. [37], but for our purposes it is sufficient to note that in optical tweezers, the laser is very tightly focused, producing strong gradients in the optical fields in the region surrounding the focal spot, and an associated dipole force on polarizable media in that region. This dipole force, which is usually called the gradient force, dominates over radiation pressure for the sorts of samples usually studied with this technique.

There are alternative trapping geometries that may be desired. For gold nanoparticles and for transmissive particles with an index of refraction much higher than that of the surrounding medium, radiation pressure is significant. A counterpropagating beam trap is strongly preferred over optical tweezers in those instances [38, 39]. While a significant radiation pressure would tend to knock particles out of optical tweezers, radiation pressure actually plays a helpful role in both counterpropagating beam traps and in levitation traps [40], each of which is seeing a resurgence in the literature.

Figure 1 shows a schematic representation of a standard holographic optical
Holographic optical tweezers (HOT) setup. Beginning at the laser, we first have a beam-expanding telescope, so that the beam fills the holographic element, which is shown here as a reflective, programmable spatial light modulator (SLM) but may be replaced with any (reflective or transmissive) diffractive optical element (DOE). The size of the illuminating Gaussian beam is a compromise between power efficiency and resolution: On the one hand, the larger the beam, the better use it makes of the area of the SLM, which in turn means higher resolution of the resulting light pattern in the trapping plane. On the other hand, if the Gaussian beam illuminating the SLM is too large, a fraction of the incident power misses the active area of the SLM and is therefore either lost or goes into the zeroth order. The standard compromise is a beam diameter that is roughly matched to that of the SLM. But it is not only for the sake of improved efficiency that the beam diameter is matched to that of the holographic element: when an SLM is used any significant incident beam power must be distributed, so as to avoid boiling the liquid-crystal active element, permanently destroying the device. Following the SLM, there is a second telescope, which ensures that the beam diameter is appropriate given the diameter of the back aperture of the lens to be used for tweezing (usually a standard microscope objective lens).

In practice, the two lenses comprising this second telescope are separated by the sum of their focal lengths; the SLM is separated from the first lens in that telescope, the Fourier lens, by the focal length of that lens, and the second lens in the telescope is separated by its own focal length from the back aperture of the objective. In this way, this second telescope actually serves four separate roles. First, it adjusts the beam diameter so as to fill the entrance pupil (i.e., the back aperture) of the objective.
lens. Second, as the placement of the SLM is equivalent to that of the steering mirror in conventional optical tweezers, the telescope projects the hologram onto the back aperture of the objective lens, thereby ensuring that any beam deflections created by the hologram do not cause the beam to walk off of the entrance pupil of the objective lens. Third, the fact that the hologram and the back aperture of the objective are conjugate image planes also creates a simple relationship between the beam at the output of the hologram and the beam in the focal plane. \( P \): the beam’s complex amplitude in the trapping plane is the Fourier transform of that in the SLM plane. Fourth, because the telescope used is Keplerian, rather than Galilean, the amount of real estate taken up on the optical table is somewhat greater, but the benefit that comes with this cost is that it allows for spatial filtering to be done in the plane labeled \( F \) in Figure 1. Plane \( F \) is conjugate to the focal plane \( P \), so an enlarged image of the trapping pattern is there, where it may be manipulated [41]. (Commonly, a spot block is introduced at plane \( F \), to remove the undeflected, zeroth-order spot. An alternative is to add a blazing to the hologram, shifting the desired output pattern array away from the zeroth order, with all subsequent optics aligned along the path of the centroid of the first-order beamlets.)

In optical tweezers, the strong field gradient along the direction of propagation is produced by the peripheral rays in the tightly focused beam, and not by those rays along the optic axis. For that reason, it is essential to use a final, tweezing lens with a high numerical aperture (NA), and to ensure that the input beam delivers significant power to those peripheral rays. Therefore, a simple telescope is usually used to match the beam diameter to that of the entrance pupil (back aperture) of a high-NA microscope objective or, in the case of a Gaussian beam, where the intensity of peripheral rays is weak, to slightly overfill the back aperture of the objective. (Excessive overfilling simply throws away light that would otherwise go towards the desired mechanical effect, and also produces an undesired Airy pattern around each trap site created.)

It should be noted that the original work using holographic methods for optical trapping did not require that the DOE be projected onto the back focal plane of the tweezing lens. Fournier et al. [42, 43] simply illuminated a binary phase hologram with a quasi-plane wave which, through Fresnel diffraction, generates self-images of the grating in planes that are periodically positioned along the direction of propagation (a phenomenon known as the Talbot effect). Already in 1995, very strong trapping of 3-micron spheres was observed in these Talbot planes, and the means of creating a variety of lattices was discussed. In their 1998 paper, Fournier et al. also proposed the use of programmable “spatial light modulators to obtain a time-dependent optical potential that can be easily monitored.” Fresnel holographic optical trapping is now seeing a revival for use in holographic optical trapping [44], for the generation of large, periodic array structures. The first demonstration of sorting on an optical lattice [24] used a DOE that was conjugate to the image plane, rather than the Fourier plane. Because of this conjugacy condition, the use of a fan-out DOE results in the convergence of multiple beams in the trapping volume, and the associated formation of an interference pattern. This setup can produce high-quality 3D arrays of traps over a large region
of space, which can be tuned in ways that include the lattice constant, the lattice envelope, and the degree of connectivity between trap sites [45]. That said, compared to Fresnel holographic optical trapping, the Fourier-plane holography that we describe here offers a number of benefits, e.g. the ability to fully utilize such trade-offs as, for example, restricting the area in the trapping-plane over which the beam is shaped in order to gain resolution [46]. Moreover, generating a 3D array of traps intended to deflect the flow of particles passing through (i.e., optical sorting) is less of a challenge than filling a 3D array of traps so as to create a static structure: in the latter case, particles filling traps in each layer perturb the optical fields in all subsequent layers. Direct manipulation of the beam’s Fourier-space properties can allow the creation of optical fields in the trapping volume that are self-healing (or self-reconstructing) in a number of arbitrarily chosen directions [47], a significant feature for the generation of some types of 3D filled arrays. Primarily, though, the benefits of Fourier-plane HOTs are the ability to generate generalized arrays without any special requirements regarding symmetry, and the ability to provide flexible, highly precise individual trap positioning [48].

Also in 1995, the group of Heckenberg and Rubinsztein-Dunlop at the University of Queensland produced phase-modulating holograms for mode-conversion of conventional optical tweezers into traps capable of transmitting orbital angular momentum [3]. Because the Laguerre-Gaussian modes utilized are structurally stable solutions of the Helmholtz equation, the DOE/SLM need not be positioned in any particular plane, and so this work in Australia, which is compatible with Fourier-plane holographic trapping, established (along with work in Scotland [5]) the techniques that are still in use today for the generation of traps carrying optical angular momentum.

The basic optical train shown in Figure 1, projecting the hologram onto the back focal plane of a microscope objective, was first demonstrated (using a pre-fabricated, commercial DOE) at the end of the decade, in 1998 [49], with a description of the most common algorithm for creating tailored holograms for optical trapping following in 2001 [50] and again in 2002 [51]. In this work, and in much of what has followed, it is assumed that the phase-modulating hologram will be positioned as we have described, so that the complex amplitude in the DOE/SLM plane and the complex amplitude in the central trapping plane form a Fourier-transform pair. With this relationship established, the complex amplitude distribution in one plane can be calculated very efficiently from the other using a Fast Fourier Transform [52].

Given this simple relationship, it may seem surprising that there is any need to discuss algorithms at all: given a desired intensity distribution in the focal plane of the microscope objective one need only take the inverse Fourier transform to determine the appropriate hologram – but the result of that simple operation would be a hologram that modulates both the phase and amplitude of the input beam. The required amplitude modulation would remove power from the beam, with catastrophic consequences for the efficiency in many cases. If, then, for the sake of efficiency, we constrain the hologram to phase-only modulation, there is usually not any analytical solution that will yield the
desired intensity distribution in the trapping plane.

Some of the required phase modulations are obvious. The simplest DOE would be a blazed diffraction grating, which is the DOE equivalent of a prism, and the Fresnel lens, which is the DOE equivalent of a lens. With these basic elements, we see that we can move the optical trap sideways (with the blazing), but also in and out of the focal plane (with the Fresnel lens), and by superposition of such gratings and lenses we can create multiple foci, which can be moved and, through the addition of further phase modulation, even be shaped individually [51].

Clearly we are not limited to a superposition of gratings and lenses. Because a standard HOT setup has holographic control over the field in the trapping plane, it can shape not only the intensity of the light field but also its phase. For example, there are very simple holograms that turn individual traps into optical vortices – that is, bright rings with a phase gradient around the ring. Because light will repel particles with a lower index of refraction than the surrounding medium, holographically produced arrays of vortex beams can be used to trap and create ordered arrays of low-index particles [53]. Other simple holograms can create non-diffracting and self-healing (over a finite range) Bessel beams [54, 55] (though holographic generation of Bessel beams is not compatible with Fourier-plane placement of the DOE). Still, in all of the cases described above, the phase profile required for the DOE is something you could guess, with good results.

However, while it is possible to intuit a phase modulation that will form extended periodic structures [56], matters become much more complex when considering arbitrary arrangements of particles, and the simplest guesses do not always yield the best results, as is shown in Section 4. Luckily, as we will describe in detail, the quality of the trap array can be greatly improved through the selection of an optimized iterative algorithm, though what might be optimal within the context of a particular experiment may involve trade-offs in terms of efficiency and computational speed.

3.1.1. The diffractive optical element or spatial light modulator. In this section we discuss the sorts of physical specifications which are required of the diffractive optical element to be used for the generation of optical traps. This DOE can either be static (etched in glass or, inexpensively, stamped into plastic), but is often in the reconfigurable form of a phase-only spatial light modulator (SLM) – a phase hologram under real-time computer control.

A discussion of the issues involved in lithographically manufactured DOE, such as the influence of the number of phase levels created and of surface roughness upon the resulting trap array, is contained in Ref. [50]. A sense of these issues may be gained by noting that while the light intensity directed to each spot in a $10 \times 10$ array was predicted to vary by $\pm 10\%$ for a binary DOE (given the specific algorithm used to calculate them), in an actual experiment it varied by $\pm 23\%$, with the additional variation attributed to manufacturing defects associated with the process of etching the phase modulation profile into glass [57]. As more phase levels are added, the lithographic challenges increase (though, with careful alignment, $2^N$ phase levels can be created by
Holographic optical tweezers

N etch steps). The greater the extent of the array, the more sensitive the uniformity of trap intensity becomes to fabrication errors.

Although holograms etched into glass clearly produce static images, one can rapidly raster the laser between a tiled array of such holograms, thereby producing a dynamic hologram with a refresh rate limited only by the speed of rastering. However, because of the heightened opportunity for real-time reconfiguration and optimization, our emphasis here will be upon the special case where the DOE is a programmable SLM.

There are various types of SLMs [58]: they can, for example, be divided according to which aspect of light they modulate, which mechanism they use to do this, or whether they work in reflection or in transmission. Most HOTs use phase-only, liquid-crystal (LC), reflective SLMs: phase-only because of the efficiency advantage mentioned in section 3.1; liquid crystal because of the cheap availability and maturity of this technology; and reflective because of a speed (and efficiency) advantage (the switching time depends upon the thickness of the liquid-crystal layer, which can be halved for reflective devices).

The LC character of most SLMs used in HOTs determines many of their properties. We will discuss here a few of these; more detailed discussions can be found, for example, in Ref. [59]. They work by applying locally defined voltages to an array of areas spread across a LC layer. The liquid crystal molecules re-align themselves in response to the voltage, thereby changing their optical properties. In parallel-aligned nematic liquid crystals, for example, they tilt relative to the substrate [59]. LC SLMs were first used as holograms after correction of thickness nonuniformities [60]. Nowadays a number of different models are commercially available, for example from Boulder Nonlinear Systems [61], Holoeye Photonics AG [62], and Hamamatsu [63].

Liquid-crystal-based SLMs use either nematic or ferroelectric liquid crystals. Nematic SLMs offer a large number of phase levels (typically 256, though the sigmoidal grayscale-to-phase-level function of some of these systems compresses the number of useful levels to a number well below the nominal rating), but are slow (in practice, an update rate for a Near-IR nematic SLM would typically be 20 Hz or less). Ferroelectric SLMs have only two phase levels (phase shift 0 and π), which limits the choice of algorithms for the calculation of the hologram patterns and the diffraction efficiency (by always creating a symmetric -1st order of the same brightness as the +1st order) [64]. On the other hand, with update rates of typically tens of kilohertz, ferroelectric SLMs are significantly faster. While nematic SLMs are the usual choice, ferroelectric SLMs [65, 66] have also been used in HOTs [64, 67].

The maximum phase delay a LC SLM can introduce grows with the thickness of its LC layer. Depending on the LC-layer thickness, a LC SLM has a maximum wavelength for which it can achieve a full 2π phase delay; this is often taken as the upper limit of an SLM’s specified wavelength working range. On the other hand, the response time is proportional to the square of the LC layer’s thickness [68]. Again, in reflective SLMs, light passes through the LC layer twice, halving the required thickness.

Due to lack of surface flatness, reflective SLMs usually aberrate the light beam more
than transmissive SLMs. Like the spherical aberration introduced by the objective, this can deteriorate the quality of the traps and limit the trapping range [37]. Aberrations in the flatness of the SLM can – together with other aberrations in the optics – be corrected by displaying a suitable phase hologram on the SLM [69, 70, 71, 72]. This is perhaps not surprising, as SLMs are also used in other adaptive-optics systems [73, 69].

Commercial LC SLMs are either optically (Hamamatsu) or electrically (Boulder, Holoeye) addressed. Electronically addressed SLMs contain (square or rectangular) pixels, with a small gap – and associated dead area – between pixels. The convolution theorem dictates that the field in the SLM’s Fourier plane (the trapping plane) is the Fourier transform of the field of coherent point light sources centered in each pixel (the phase and intensity is that of the pixel), multiplied by the Fourier transform of the field of a single, centered, pixel. The first term (the Fourier transform of the point light sources) is periodic in $x$ and $y$. The copies of the central field are essentially diffraction orders of the grating formed by the point light sources. The second term (the Fourier transform of a single pixel) is usually close to a sinc function in $x$ and $y$ (whereby the sinc function gets wider when the pixel gets narrower, and vice versa), which means that the intensity of the field in the Fourier plane falls off to zero at the nodes of the sinc function. This limits the power that is wasted into higher diffraction orders. By optimizing the pixel shape, the power diffracted into higher grating orders can be further reduced. This is essentially what is happening in the optically addressed SLMs by Hamamatsu, where a “write light” pattern produced with the help of a pixelated liquid-crystal display is projected onto a photoconductive layer. The projected pattern is slightly out of focus, resulting in an image consisting of pixels with smoothed edges. The photoconductive layer controls the voltage across the LC layer, which modulates the phase of “readout light” passing through it. The resulting phase modulation is such that diffraction into the higher grating orders is almost completely suppressed.

It should once again be noted that a light beam with too much power can destroy a LC SLM, by boiling the liquid crystal. This limits the number of optical traps that has been achieved with LC-SLM-based, Fourier-plane HOTs to $\approx 200$ [74], a limit which can perhaps be overcome by using cooled SLMs, SLMs with a larger area over which the beam’s power can be distributed, or other types of SLMs, for example deformable mirrors [75].

4. Algorithms for holographic optical traps

The intensity distribution usually required in HOTs consists of a set of diffraction-limited bright spots surrounded by darkness. There are algorithms for specialized applications that require other light distributions, which we briefly review in section 4.9. Our main emphasis, however, is on reviewing the algorithms that are presently available for optical trapping using the standard optical train described in section 3.1, discussing both why and how they work. Closely following Ref. [76], we compare algorithm performance in terms of efficiency, that is fraction of overall power in the traps, and uniformity, a
measure of how evenly power is distributed between the different traps, for a $10 \times 10$ array of traps in the Fourier plane, created using an SLM with $768 \times 768$ pixels and 256 (8 bit) phase levels under uniform illumination.

Some of the algorithms we review here are fast and therefore particularly well suited to interactive use (e.g. Ref. [77]); other algorithms create better traps. The two classes are somewhat mutually exclusive, sometimes so much so that when a higher degree of control in trap intensities is required holograms calculated *prior* to the experiment are used. This is the case when, for example, one requires a set of traps with *uniform* (or precisely controlled) potential well depths, or when ghost traps cannot be tolerated (for example when they prevent filling of the intended trapping sites).

![Figure 2. Geometry of pixel and trap positions relative to the (effective) Fourier lens's focal planes. The transverse position of the $j$th pixel in the SLM plane is $(x_j, y_j)$, the position of the $m$th trap relative to the centre of the Fourier plane is $(x_m, y_m, z_m)$. $x$ and $y$ are the two transverse coordinates, $z$ is the longitudinal coordinate.](image)

All the algorithms reviewed here utilize the phase shifts $\Delta^m_j$ picked up by the light (of wavelength $\lambda$) as it travels from pixel $j$ to trap $m$. These phase shifts can be calculated as follows. The complex amplitude of the electric field at the $j$th pixel is
\[
u_j = |u| \exp(i\phi_j),\tag{1}\]
where $\phi_j$ is the corresponding phase shift. We can use scalar diffraction theory to propagate the complex electric field from the $j$th pixel to the location of the $m$th trap in the image space [78]. Summing up the contributions from all of the $N$ pixels we obtain the complex amplitude $v_m$ of the electric field at the position of trap $m$,
\[
u_m \propto \sum_{j=1}^{N} |u| \exp(i(\phi_j - \Delta^m_j)).\tag{2}\]
The phase shifts $\Delta^m_j$ are given by
\[
\Delta^m_j = \frac{\pi z_m}{\lambda f^2} (x_j^2 + y_j^2) + \frac{2\pi}{\lambda f} (x_j x_m + y_j y_m),\tag{3}\]
where \(x_j, y_j\) are the \(j\)th SLM pixel’s coordinates (projected into the back focal plane of the objective) and \(x_m, y_m, z_m\) are the \(m\)th trap’s coordinates relative to the centre of the Fourier plane (Fig. 2).

These propagation phase shifts can be used to calculate a hologram pattern that creates a single focus at the position of the \(m\)th trap: by choosing the phase \(\phi_j\) at each pixel to cancel out the phase shift as the light propagates from pixel \(j\) to trap \(m\), that is for \(\phi_j = \Delta_j^m\) for all pixels \((j = 1, \ldots, N)\), the contribution from all pixels interferes constructively at the position of trap \(m\). We call this choice of hologram pattern the single \(m\)th trap hologram.

The normalized field at trap \(m\) is given by the equation

\[
V_m = \sum_{j=1}^{N} \frac{1}{N} \exp(i(\phi_j - \Delta_j^m));
\]  

(4)

its modulus squared is the normalized intensity at trap \(m\), that is

\[
I_m = |V_m|^2.
\]  

(5)

For the single \(m\)th trap hologram, all terms in the sum (4) are real and equal to \(1/N\), so \(I_m = |V_m|^2 = 1\).

| Algorithm | \(e\) \(|\%\)| | \(u\) \(|\%\)| | \(\sigma\)% | \(K\) | Scaling |
|-----------|----------------|-----------------|---------|------|--------|
| RM        | 0.01 (0.07)    | 0.58 (0.79)     | 16 (13) | 1 (1) | \(N\)  |
| S         | 0.29 (0.69)    | 0.01 (0.52)     | 257 (40)| 1 (1) | \(N \times M\) |
| SR        | 0.69 (0.72)    | 0.01 (0.57)     | 89 (28) | 1 (1) | \(N \times M\) |
| GS        | 0.94 (0.92)    | 0.60 (0.75)     | 17 (14) | 30 (30) | \(K \times N \times M\) |
| AA        | 0.93 (0.92)    | 0.79 (0.88)     | 9 (6)   | 30 (30) | \(K \times N \times M\) |
| DS        | 0.68 (0.67)    | 1.00 (1.00)     | 0 (0)   | 7.5 \times 10^5 (1.7 \times 10^8) | \(K \times P \times M\) |
| GSW       | 0.93 (0.93)    | 0.99 (0.99)     | 1 (1)   | 30 (30) | \(K \times N \times M\) |

Table 1. Summary of the theoretical performance of some algorithms for shaping of a highly symmetric 2D and a less symmetric 3D structure (in brackets). The algorithms are random-mask encoding (RM), superposition (S), random superposition (SR), Gerchberg-Saxton (GS), adaptive-additive (AA), direct search (DS), and weighted Gerchberg-Saxton (GSW). The 2D target trap structure is a \(10 \times 10\) square grid of bright spots, the 3D target structure consists of 18 traps located on the sites of a diamond-lattice unit cell. The efficiency, \(e\), and uniformity measures, \(u\) and \(\sigma\), are calculated after \(K\) iterations. The right-most column refers to computational cost scaling, where \(M\) is the number of traps, \(N\) is the number of pixels in the hologram, and \(P\) is the number of gray levels (256 here). (From Ref. [76].)

The performance of different algorithms is quantified by three measures: efficiency \((e)\), uniformity \((u)\) and percent standard deviation \((\sigma\), another measure of how equal power is distributed between different traps):

\[
e = \sum_m I_m; \quad u = 1 - \frac{\max[I_m] - \min[I_m]}{\max[I_m] + \min[I_m]}, \quad \sigma = \sqrt{\langle(I - \langle I \rangle)^2 \rangle} / \langle I \rangle.
\]  

(6)
denotes the average over all trap indices, \( m \). Table 1 summarizes the results of the performed benchmark test on the highly symmetric, 2-dimensional, \( 10 \times 10 \) square target grid mentioned above, and on a less symmetric, 3-dimensional, target structure. It demonstrates that many algorithms – most notably the superposition algorithms, but also random-mask encoding and the Gerchberg-Saxton and adaptive-additive algorithms, but not direct search and the weighted Gerchberg-Saxton algorithm – give worse uniformity for symmetric patterns. It also demonstrates the huge differences in the number of iterations different algorithms require to reach a good hologram, and how the computational cost scales with various parameters. For most applications, the recent weighted Gerchberg-Saxton algorithm appears to be a good choice.

4.1. Random mask encoding

One approach for creating more than one trap which is particularly appealing by virtue of its simplicity and speed is random mask encoding [79]. In the single-\( m \)-th-trap hologram all SLM pixel phases are chosen to be \( \Delta_m^{m_j} \), thereby cancelling out the propagation phase shifts between all pixels and the \( m \)th trap, so that the contributions from all pixels interfere constructively at the trap position. In random-mask encoding, for each SLM pixel, say pixel number \( j \), the index \( m \) of a trap is randomly chosen, and the SLM pixel phase is set to the corresponding propagation phase shift \( \Delta_m^{m_j} \):

\[
\phi_j = \Delta_j^{m_j},
\]

where \( m_j \) is a number between 1 and \( M \) randomly chosen for each \( j \). The contributions from the fraction of pixels that displays the propagation phase shift corresponding to trap \( m \), and therefore the corresponding fraction of the light illuminating the SLM, then interferes constructively at the position of trap \( m \); the contributions from the other pixels similarly interfere constructively at the positions of the other traps.

The technique is very fast, and performs remarkably well as far as uniformity is concerned. However, the overall efficiency can be very low when \( M \) is large. In fact, on average, for each \( m \) only \( N/M \) pixels will interfere constructively, all the others giving a vanishing contribution. Therefore \( |V_m|^2 \simeq 1/M^2 \) and \( e \simeq 1/M \), which can be significantly smaller than one when \( M \) is large. In our benchmark case, where \( M = 100 \), Ref. [76] numerically obtained \( u = 0.58 \) but \( e = 0.01 = 1/M \).

Still, random-mask encoding is particularly useful to quickly generate either one or a few additional traps on top of a complex light structure obtained via a pre-calculated hologram. Such “helper tweezers” are useful for filling in the pre-calculated array of traps, allowing the user to interactively trap, drag and drop initially free particles into the desired locations. For such purposes, a fraction of the SLM pixels can be randomly chosen and temporarily used to display the “service-trap” hologram.
4.2. Superposition algorithms

The field immediately after the SLM displaying a single \( m \)th trap hologram is given by \( u_j = |u| \exp(i\Delta_j^m) \). As discussed above, this field leads to a bright spot at the position of the \( m \)th trap. If the superposition, or complex sum, of the fields due to several \( m \)th trap holograms could be created, it would give bright spots at several trap positions. However, this superposition field cannot easily be created with a phase-only SLM: the intensity immediately after the SLM is simply not that of the superposition field.

Superposition algorithms ignore this intensity mismatch and use the phase pattern of the superposition field as the phase-hologram pattern, so \([80, 81]\)

\[
\phi_j = \arg \left( \sum_m \exp(i\Delta_j^m) \right). \tag{8}
\]

This algorithm has been called “superposition of prisms and lenses” (SPL) [76]. It is slower than random-mask encoding (due to the extra \( N \arg \) function evaluations), but produces useful light distributions with reasonable efficiencies but very poor uniformities. In our benchmark case, the efficiency is \( e = 0.29 \) and the uniformity is \( u = 0.01 \). Moreover, a significant part of the energy is diverted to unwanted ghost traps; this happens particularly in highly symmetrical trap geometries [82].

The algorithm described above is actually just the simplest superposition algorithm. It can be improved if we add random phases \( \theta_m \) (uniformly distributed in \([0, 2\pi]\)) to each single trap hologram, and therefore to each trap. The phase of SLM pixel \( j \) then becomes

\[
\phi_j = \arg \left( \sum_m \exp(i(\Delta_j^m + \theta_m)) \right). \tag{9}
\]

This algorithm, usually called Random Superposition [83], has the same computational cost as the SPL algorithm, produced similar uniformity but much better efficiency in our benchmark case (\( e = 0.69, u = 0.01 \)).

When dealing with low-symmetry geometries, holograms calculated using superposition algorithms can also produce good uniformity levels and no further refinement is needed. If precise trap positioning is not an issue one can even deliberately reduce the pattern symmetry by adding a small amount of random displacement to trap locations, as demonstrated in Ref. [82].

Although slower than random-mask encoding, the computational speed of superposition algorithms still allows for interactive manipulation. They can be speeded up further by computing the hologram using a computer’s Graphic Processing Unit (GPU) for both computational and rendering tasks [100]. As the entire hologram is continuously recalculated, superposition algorithms are usually preferred to random-mask encoding for interactive applications requiring dynamic deformation of the entire trapping pattern.
4.3. Gerchberg-Saxton algorithm

The Gerchberg-Saxton algorithm [83, 84, 85, 86, 87, 88] was developed by the crystallographers Ralph Gerchberg and Owen Saxton to infer an electron beam’s phase distribution in a transverse plane, given the intensity distributions in two planes. It can also be applied to light, specifically to find a phase distribution that turns a given input intensity distribution arriving at the SLM plane into a desired intensity distribution in the trapping plane. In the Gerchberg-Saxton algorithm the complex amplitude is propagated back and forth between the two planes, replace the intensity in the trapping plane with the target intensity and that in the SLM plane with the illuminating laser beam’s actual intensity profile.

The algorithm can be extended to 3D trap geometries where multiple planes are considered for forward propagation. The back-propagated field is then obtained as the complex sum of the corrected and back-propagated fields from the target planes. Generalization to full 3D shaping [89, 90] is currently far too slow for interactive use (taking days to calculate the desired phase modulation).

In the original implementation, forward and backward propagation were performed with Fast Fourier Transforms (FFTs). However, when the target intensity is an array of bright spots surrounded by darkness, it is not necessary to calculate the complex field in points whose intensity will be replaced by zero before back propagation. The FFT has also the drawback of discretizing the transverse coordinates of traps. A much faster and more versatile implementation of the Gerchberg-Saxton algorithm for HOTs only computes the field at the trap locations. This is exactly what equation (4) does, which calculates the fields $V_m$ at the trap positions from the SLM phases, $\phi_j$. By weighting the contribution from each SLM pixel with the same factor, $1/N$, equation (4) also effectively replaces the intensity distribution in the SLM plane with the uniform intensity distribution of the illuminating laser beam. (This can, of course, easily be generalized to non-uniform illumination intensities.) A phase-hologram pattern that takes into account the phases of the field at the trap locations can then be calculated using a superposition algorithm in which the single-trap holograms are superposed with the relative phase of the corresponding trap,

$$\theta_m = \arg(V_m),$$  \hspace{1cm} (10)

using equation (9). Like all algorithms that incorporate the superposition algorithm, the Gerchberg-Saxton algorithm can be speeded up by using the computer’s graphic processing unit [100]. One iteration of this optimized Gerchberg-Saxton algorithm [91, 92] then consists of the successive application of equations (4), (10), and (9). It converges after a few tens of iterations. After thirty iterations, Ref. [76] obtained $e = 0.94$ and $u = 0.60$ for our benchmark case.
4.4. Adaptive-additive algorithm

It can be shown mathematically that the Gerchberg-Saxton algorithm maximizes the sum of the amplitudes of \( V_m, \sum_m |V_m| \). It therefore has no bias towards uniformity [76]. It is possible to modify this algorithms so that it maximizes other quantities, resulting in a bias towards uniformity. The most important example is the maximization of the function \((1 - \xi) \sum_m |V_m| + \xi \sum_m \log |V_m|\), in which the strength of the uniformity bias is controlled by the parameter \( \xi \). This leads to the generalized adaptive-additive algorithm [50, 51], in which the calculation of the phase hologram is performed according to

\[
\phi_j = \arg \left[ \sum_m \exp(i(\Delta^m_j + \theta_m)) \left(1 - \xi + \frac{\xi}{|V_m|}\right) \right] \tag{11}
\]

instead of equation (9). The generalized adaptive-additive algorithm includes the Gerchberg-Saxton algorithm as the special case \( \xi = 0 \). For \( \xi > 0 \), it yields improved uniformity when compared to the Gerchberg-Saxton algorithm. For \( \xi = 0.5 \), for example, the uniformity is \( u = 0.79 \) and efficiency \( e = 0.93 \) for the benchmark case.

4.5. Curtis-Koss-Grier algorithm

The adaptive-additive algorithms in Ref. [51] contains further, very useful, generalizations. By subtracting “kernel phases” \( \kappa_j^m \) in the calculation of the field at the trap positions, that is by using

\[
V_m = \sum_{j=1}^N \frac{1}{N} \exp(i(\phi_j - \kappa_j - \Delta^m_j)) \tag{12}
\]

instead of equation (4), and by adding \( \kappa_j \) again in the calculation of the phase hologram, that is by using

\[
\phi_j = \arg \left[ \sum_m \exp(i(\kappa^m_j + \Delta^m_j + \theta_m)) \left(1 - \xi + \frac{\xi}{|V_m|}\right) \right] \tag{13}
\]

instead of equation (11), the foci corresponding to the traps can be individually shaped. For example, the choice

\[
\kappa^m_j = \ell \varphi_j, \tag{14}
\]

where \( \varphi_j \) is the azimuthal angle (in radians) of pixel \( j \) with respect to the center of the SLM and \( \ell \) is an integer, turns the focal spot at the position of trap \( m \) into an optical vortex of charge \( \ell \). This very popular and versatile algorithm is sometimes called the Curtis-Koss-Grier algorithm. For our benchmark case, it produces the same results as the adaptive-additive algorithm.

4.6. Weighted Gerchberg-Saxton algorithm

Another variation of the Gerchberg-Saxton algorithm that is biased very strongly towards a uniform distribution of the intensity in the different traps has been derived
Holographic optical tweezers

by introducing the $M$ additional degrees of freedom $w_m$ and maximizing the weighted sum $\sum_m w_m |V_m|$ with the constraint that all $|V_m|$s are all equal. In the corresponding variation of the Gerchberg-Saxton algorithm, the weights for the current ($k$th) iteration, $w_m^k$, are calculated from those for the previous (($k-1$)th) iteration, $w_m^{(k-1)}$, according to

$$w_m^k = w_m^{(k-1)} \frac{|V_m^{k-1}|}{|V_m^{k-1}|},$$

before the phase hologram is calculated using

$$\phi_j = \arg \left[ \sum_m w_m^k \exp(i(\Delta_m^j + \theta_m)) \right].$$

One iteration of this algorithm, called the Weighted Gerchberg-Saxton Algorithm [76], comprises successive application of equations (4), (10), (15) and (16). The algorithm converges with a speed typical of the Gerchberg-Saxton and Adaptive-Additive algorithms.

In the benchmark case, and starting from a random-superposition phase hologram ($\phi_j$ calculated according to equation (9)) and setting all initial weights to 1 ($w_0^m = 1$), the Weighted Gerchberg-Saxton Algorithm produces a hologram having the almost optimal performance of $e = 0.93$, $u = 0.99$ [76].

4.7. Direct search

Direct-search algorithms make direct use of the possibility of computers to try out a vast number of different holograms in the search for the best one. The first direct-search algorithm [93] defined an error between the target intensity pattern and the intensity pattern that was calculated to result from a specific hologram. Starting with a random binary intensity hologram, it then changed the intensity of the first pixel, compared the errors corresponding to the unchanged and changed holograms, and kept the hologram with the lower corresponding error. Starting with this kept hologram, it then altered the intensity of all pixels in lexicographic order (line by line), after each change keeping the better hologram. When the last pixel was reached, the algorithm altered the first pixel again and so on.

Most direct-search algorithms are slight variations on this original algorithm. Firstly, they do not alter pixels in lexicographic order, but in random order. Exhaustive random order – making sure that all pixels are visited once during each “round of iterations” before any pixel is re-visited – is most efficient [94]. Secondly, unlike the original algorithm they usually deal with more than one available phase or intensity level. There are different optimization strategies for that case: some algorithms compare the error of the original hologram with that of one hologram in which the intensity (or phase) has been randomly altered, keeping the better hologram; other algorithms compare the original hologram with the holograms in which the randomly selected pixel takes on all possible values. We concentrate here on the latter strategy. Thirdly, the choice of the error function (or cost function) which is minimized – or, alternatively, of a quality
function which is maximized – depends on exactly what it is to be achieved, and many possibilities have been implemented. For the purposes of benchmarking, we follow Refs. [95, 96, 76] and concentrate here on a quality function that is a linear combination of the efficiency metric $e$ and the uniformity metric $\sigma$, namely

$$e/M - f\sigma,$$

where $f$ is a factor that allows their relative importance to be adjusted.

The benchmark algorithm starts from an initial hologram obtained from the random-superposition algorithm. It picks one pixel at random, cycles through all the $P = 256$ phase levels, and then sets the pixel phase to the one that gives the highest quality function. As suggested in Ref. [96], for $f = 0.5$, the algorithm achieves a perfect uniformity ($u = 1.00$) after $1.3N$ steps, with the computational cost scaling as $M \times P$, although the overall efficiency is diminished to $e = 0.68$. Better holograms can be obtained by giving more bias to efficiency (e.g. $f = 0.25$) and waiting for a substantially longer time ($\approx 10N$ steps – that is, about a hundred times longer than GS), whereby this number of iterations can be decreased with only a moderate performance decrease by reducing the number of gray levels, $P$ [95]. With eight “greyscale” phase levels and all other parameters set as before, Ref. [76] obtained $e = 0.84$ and $u = 1.00$ after $7N$ steps (that is, about three times longer than GS).

4.8. Simulated annealing

A number of articles on holographic optical tweezers refer to simulated annealing (usually together with genetic algorithms) as some sort of last-ditch attempt towards achieving the best possible hologram.

Direct-search algorithms always move towards better holograms. In the corresponding quality-function landscape they always move upwards, which makes such algorithms vulnerable towards being caught in local maxima of the quality function. Simulated annealing [97, 98], also known as the Metropolis algorithm, is an optimization strategy that attempts to simulated the characteristics of the process of crystallization of a slowly cooled liquid that results in a perfect crystal – the lowest-energy state. It was first applied to holography in Ref. [99].

Simulated-annealing algorithms tend to measure the quality of a hologram in terms of its energy – the (dimensionless) cost function, re-named to emphasize the analogy with crystallization. Like direct-search algorithms, simulated-annealing algorithms always accept energy decreases. Unlike direct-search algorithms, however, they sometimes accept energy increases (and therefore decreases in quality), whereby the probability of accepting an energy increase $\Delta E$ is given by

$$\exp(-\Delta E/T).$$

The temperature $T$ (also dimensionless) is lowered while the algorithm is running, whereby the “annealing schedule” – the choice of the initial temperature, and the way
in which it is lowered – requires experimentation to achieve a good result in a reasonable number of iterations.

Simulated annealing is a trade-off between the time it takes to run the algorithm and the quality of the resulting hologram. This is again analogous to crystallization of a cooling liquid: if a liquid metal is cooled too quickly (or “quenched”), it does not reach the lowest-energy state but a higher-energy, polycrystalline or amorphous, state [98]. Simulated annealing is therefore not suitable for interactive use, but is the algorithm of choice if the best hologram quality is required and calculation speed is not an issue.

4.9. Algorithms for specialized applications

Gabe’s section: vortices, vortex arrays, shaped vortices, lines (Bessel beams), ...

5. Alternative means of creating extended optical potential energy landscapes

Large arrays of optical traps can be generated in several ways. While holographic optical tweezers are very flexible, there are situations where alternative approaches might be considered.

For example, multi-beam interference is simple, produces high-quality optical lattices over extended 3D volumes, and can tolerate high beam powers. Extensive work on laser-induced freezing and other novel phase transitions has utilized this approach [13, 19]. Burns et al. used the standing optical field resulting from the interference of several beams to trap polystyrene spheres, thereby producing a 2D colloidal crystal, and to propose the existence of optically-mediated particle-particle interactions in this system (optical binding) [101, 102]. However, such approaches are limited to symmetric patterns.

Alternatively, galvan mirrors [103] or piezoelectrics [37, 104] have served as the basis for designs involving scanning laser tweezers. These approaches are briefly summarized in Section 5.0.1. Another – SLM-based – strategy that allows flexible generation of trap arrays uses the Generalized Phase Contrast (GPC) method. In addition, arrays covering large areas have now been produced using evanescent waves. In the remainder of Section 5 we discuss these promising alternative techniques.

5.0.1. Time-sharing of traps. For generating a simple, smooth potential, such as a ring trap, the use of analog galvo-driven mirrors might be preferred over either HOTs or acousto-optic deflectors (AODs). Coupled with an electro-optical modulator, this can yield smoothly varying intensity modulations in a continuous optical potential [105, 106]. Galvo systems provide much higher throughput of the incident light than either AODs or Fourier-plane HOTs, and have been put to good use in many experiments (e.g., the Bechinger group uses galvo-driven mirrors to create simple optical “corrals” that
Holographic optical tweezers

controllably adjust the density of 2D colloidal ensembles). However, inertia limits the scan speed of any macroscopic mirror to a fraction of what is available via AODs.

Acousto-optic deflectors are simply another class of reconfigurable diffractive optic element, one which is limited to simple blazings, but which has a much higher refresh rate: AODs can be scanned at hundreds of kilohertz, repositioning the laser on such a short time scale that, under some circumstances, the trapped particles experience only a time-averaged potential. This short time constant allows for the creation of multiple “time-shared” traps using AOD deflection of the same (first-order) diffraction spot [107, 108].

In order for each trapped particle to feel only the time-averaged potential, the maximal time that the laser can spend away from any one trap would be something like a tenth of the time scale set by the corner frequency in the power spectrum of particle displacements. The inverse of this corner frequency indicates, essentially, the time it takes the particle to diffuse across the trap. The smaller the particle, the shorter this time scale becomes. Less viscous environments also present challenging time scales: for aerosols, the corner frequency can be 2 kHz even for an 8-micron sphere [109]. Because of these high corner frequencies, it could be a challenge to use AODs to generate large arrays in such samples and, indeed, HOT-based array generation has been preferred for such work [110]. It should be emphasized that, for any type of sample, “while the laser’s away, the beads will stray!” That is, over a time interval \( t \) when the laser is elsewhere (at other trap sites, or traveling between traps), a microbead will diffuse away from its nominal trap site, a distance \( d = \sqrt{2Dt} \). For a 1-micron diameter bead in water, the diffusion coefficient is \( D = k_B T / 6 \pi \eta r = 0.4 \mu m^2 / s \). So if the laser is absent for 25 microseconds, the bead is expected to diffuse 5nm. Clearly, smaller spheres would diffuse further during the same time and, even for micron-scale spheres, as the number of trapping sites increases, the demand for speed from all components of the control system (both hardware and software) becomes significant. This, coupled with the requirement that the laser spend sufficient time at each trap site to produce the time-averaged power required for the desired trap strength and the fact that while trap strength depends only on the time-average power, sample damage due to two-photon absorption and local heating contain a dependence upon the peak power [111], represents a maximum limit to both the type of arrays that can be constructed and the accuracy to which the spheres can be positioned when using time-shared trapping.

That said, impressive results have been obtained. A 20 × 20, two-dimensional array of traps was constructed and (mostly) filled with 1.4\( \mu m \) silica spheres, by the van Blaaderen group [112]. Moreover, by physically splitting the beam and creating two optical paths with offset image planes, the van Blaaderen group was able to make small arrays of traps in two separate planes. Because nearly all of the colloidal particles in their sample were index-matched to the surrounding medium, they were able to tweeze a dilute concentration of non-index-matched tracer particles, so as to controllably seed nucleation of 3D order in a concentrated colloidal sample [112], and used fluorescent confocal microscopy to image the results. Primarily, though, work utilizing AODs has
been limited to the generation of 2D arrays of traps.

For experiments involving just one or two spheres, the positioning resolution of AODs driven by low-jitter digital frequency synthesizers cannot be beaten. It is not unheard of for such systems to claim a positioning resolution of less than one-tenth of a nanometer though, at this level, the accuracy of positioning is not only affected by diffusion of the particle in the optical trap, but also by the pointing stability of the trapping laser, and by hysteresis and drift of the sample stage and of the objective lens. There are a number of technical points to be aware of. For analog AOD systems there can remain problems with “ghost traps” (as the beam is often sequentially repositioned in $x$ and then in $y$, so the generation of two traps along the diagonal of a square yields an unintended spot at one of the other corners). Moreover, because the AOD efficiency falls off as a function of the deflection angle, for applications requiring uniform arrays, one must compensate, either by spending more time at peripheral traps or by increasing the power sent to those traps. If such steps are taken, AODs offer good trap uniformity, precise positioning in 2D, and fast updates. Also, in some sense, AOD-generated arrays can be thought of as being made of incoherent light (different beams do not interfere, being present only one at a time).

More details, including references, can be found in Ref. [37]. For generating arrays of traps using AODs, pages 2964-2965 of our Ref. [112] clearly and concisely describe many of the relevant parameters that must be considered.

Unlike the SLM-based techniques, AOD-based systems cannot normally do mode conversion, aberration correction, or generate arrays of traps dispersed in 3D.

5.0.2. Generalized phase-contrast method. Frits Zernike won the 1953 Nobel Prize for developing an imaging technique that could turn a phase modulation caused by a transparent sample (e.g., a biological cell) into an intensity modulation at the plane of a camera or the eye. While Zernike’s approach was valid for weak phase modulation, workers at Risø National Laboratory in Denmark have created a Generalized Phase-Contrast technique (GPC) which provides a very nice, simple method for using the SLM to produce arbitrary, user-defined arrays of traps in 2D [113], but which can also be made to work in 3D [114, 115].

In the GPC approach, the SLM is conjugate to the trapping plane, and no computations are required in order to convert the phase-only modulation of the SLM into an intensity modulation in the image plane; instead there is a direct, one-to-one correspondence between the phase pattern displayed on the SLM and the intensity pattern created in the trapping plane. In the Fourier plane, a small $\pi$-phase filter shifts focused light coming from the SLM so that at the image plane it will interfere with a plane-wave component. The result is a system that only requires the user to write their desired 2D patterns on the SLM. The downside to this is that $xy$ positions are limited to pixel positions, meaning that ultra-high-precision trap positioning is not possible to the degree it is with the other techniques we have described. Also, beam-shaping Fourier-holography tricks are not applicable to the GPC approach.
Extension of the GPC method to 3D requires the use of counterpropagating beam traps, rather than optical tweezers, and so three-dimensional control is, in some sense, more involved. For this reason the Risø team has developed an automated alignment protocol [116] for users interested in 3D control. It is not possible to controllably place traps behind each other with this method, but there have now been many impressive demonstrations of 3D manipulation using the GPC technique.

Notably, in this “imaging mode,” the SLM efficiency is much higher, providing a throughput of up to 90% of the light, as there is no speckle noise and no diffraction losses (i.e., there are no ghost orders; there is no zero-order beam) [117]. The version of GPC using counterpropagating beam traps can also use low-NA optics, which can have a large field of view, and a large Rayleigh range. So, while Fourier-plane holographic optical traps can provide only a small range of axial displacements, limited by spherical aberration, the low-NA GPC trap arrays are sometimes called “optical elevators” because of the large range over which the traps can be displaced along the optic axis. Moreover, the working distance can be up to one centimeter, which is 100 times that of a conventional optical tweezers setup. This long working distance even makes it possible to image the trapped structure from the side [118]. Because no immersion fluid is required for low-NA imaging optics, one could imagine performing experiments using this approach in extreme environments, such as vacuum or zero gravity. That said, the use of a high-NA, oil-immersion trapping lens is necessary in order to provide high trap stiffness along the axial direction.

5.0.3. *Evanescent-wave optical trap arrays.* Evanescent fields hold promise for future generation of trap arrays, primarily for two reasons. One is not subject to the free-space diffraction limit, and can therefore create significantly sub-wavelength structures in the optical fields. Also, it has been shown that patterned evanescent fields can create large numbers of traps spanning macroscopically large areas [119]. Interestingly, in an *unpatterned*, but resonantly enhanced, evanescent field, arrays of trapped particles have recently been observed to self-organize, due to the onset of nonlinear optical phenomena (optical solitons) [120]. Several schemes have been proposed for holographic control of evanescent fields, and a tailored algorithm for doing this sort of light shaping has recently appeared [121].

The disadvantages of evanescent fields are that one *must* obviously work very close to the surface, create patterns using only the range of incident angles beyond the critical angle for total internal reflection, and allow for the strong variation in penetration depth as a function of incident angle (which, on the other hand, allows 3D shaping of the evanescent field [121]). Taken together, these necessarily constrain the shapes that can be holographically achieved in the near field, and clearly require the development of specialized algorithms.
6. The future of holographic optical tweezers

Good, fast algorithms currently exist for Fourier-plane holographic optical tweezers consisting of arbitrary 3D trap configurations. Already, today, HOTs can be bought commercially [74]. A number of researchers have begun to use HOTs as the centerpiece of integrated biophotonic workstations [122, 123, 124]. With the capability of doing holographic work established, it is reasonable to combine HOTs with digital holographic microscopy [125, 126, 127]. Moreover, as the traps are already under computer control, it is relatively easy to combine HOTs with pattern recognition to automate particle capture and sorting [128], and to trigger key events, localized in space and time, in systems near instabilities [129]. Related techniques are also undergoing significant development. So, while a great deal of science has been accomplished with one or two point-like traps, there clearly remains enormous potential associated with extended arrays of optical traps.

Acknowledgments

G.C.S. was supported by an award from the Research Corporation and by the Donors of the Petroleum Research Fund of the American Chemical Society. J.C. acknowledges the support of the Royal Society.

References

Holographic optical tweezers


Holographic optical tweezers


Holographic optical tweezers

1996.


Holographic optical tweezers


