Roadmap to Free Space Optics
Abderrahmen Trichili, Mitchell A. Cox, Boon S. Ooi, Mohamed-Slim Alouini

Abstract
With the ever-increasing demand for data and radio frequency spectrum becoming congested, Free Space Optics (FSO) may find a niche for situations where fiber is too expensive or too difficult to install. FSO is a cross-disciplinary field that draws from radio and fiber communication, astronomy, and even quantum optics, and it has seen major advances over the last three decades. In this tutorial-style review, we provide a broad overview of many of the important topics required to design, develop, and research the next generation of FSO technology.

I. INTRODUCTION
Free Space Optics (FSO), otherwise known as optical wireless communication in the infrared band, has received tremendous attention due to the broad unlicensed spectrum that it utilizes, and is proposed to complement the congested radio spectrum. In a sense, it is “fiber without the fiber” [1]. The first instance of FSO communication was when Alexander Graham Bell transmitted sound modulated onto a beam of light using his photo-phone (US patent 235199A) over a distance of 213 m in 1880.

Modern commercial FSO technology can be used to transmit tens of Gbps (Gigabits per second) over several kilometers. As such, FSO is a possible solution to the so-called “last mile” and “last meter” connectivity problems, when fiber installation is too expensive [1]. State-of-the-art experimental systems are able to sustain Petabits per second over several meters, and Terabits per second over several kilometers in the atmosphere and more in the vacuum of space.

The technology is also eminently suitable for brown-field situations, where it is impossible or impractical to install fiber, while at the same time, it is conducive to rapid installation and commissioning. Accordingly, FSO is expected to be a significant part of the future 6G era [2], [3]. Providing high-quality internet connectivity at a reasonable cost will enable developing economies to access the so-called knowledge economy and help create a “digital opportunity” [4]. This high-impact use case for FSO comes with significant, somewhat novel challenges and will be discussed in more detail later.

Unsurprisingly, FSO propagation is subject to several technical challenges. These include beam divergence over long distances, pointing errors (precise alignment is required), strong atmospheric attenuation when the weather is not clear, and atmospheric turbulence. There are solutions to divergence, alignment, and attenuation, but the problem of atmospheric turbulence is largely unsolved.

We have come a long way since 1880. Given the broad background and multi-disciplinary knowledge (for example, adaptive optics from astronomy) that has been distilled into modern FSO systems, this paper presents a convenient, tutorial-style review of the salient aspects of state-of-the-art optical wireless communications. We address this from different perspectives involving communication systems, optics, modeling, and design considerations. Importantly, mitigation strategies for the various propagation challenges are discussed.
An illustration of the use of FSO communications is in Fig. 1 with a graphical outline of the various components of a system. This paper is broadly arranged as follows. Basic principals of an optical communication system are discussed in Sec. II with the particulars of an FSO system (divergence, atmospheric attenuation, turbulence-induced fading, and pointing errors) discussed in Sec. III. An overview of more advanced techniques to mitigate the channel effects are discussed in Sec. IV which are generally only found in experimental links. Finally, we discuss broader FSO communication systems that incorporate multiple links, hybrid FSO/RF, and even quantum technologies in Sec. V.

A. Bridging the Digital Divide

As of January 2020, only 59% of Earth’s population is connected to the Internet. The distribution of connectivity is not uniform: in different parts of the world, there is a significant disparity between the number of people that are connected and those that are not. The internet penetration rate in Africa, for instance, is 39.3%. On the contrary, this rate exceeds 98% in some countries of Northern Europe [5].

Reliable connectivity enables developed communities to share knowledge and contribute to improving the economies of less fortunate communities. In addition, during pandemics and ensuing lock-downs or physical distancing (such as during COVID-19), fast and reliable internet connectivity is especially crucial to maintain online education to students, enable remote work opportunities, and access to online healthcare services.

The installation of optical fibers is typically expensive due to the high level of manual labor required in addition to the cost of way-leaves from municipalities [4]. Microwave radio is currently the technology of choice to avoid these issues but is bandwidth-limited compared to the capabilities of FSO. It doesn’t help to have 3G or even 4G connectivity to a mobile device if the backhaul cannot sustain high throughput for multiple users.

A possible application of FSO communication is to increase the capacity of existing microwave infrastructure. FSO systems can be installed on existing high sites in hybrid configurations for reliability, as presented in Sec. V-B. FSO can also provide ad-hoc support in the case of natural disasters by providing high capacity emergency links so that people can continue using their mobile devices to contact the outside world while the damaged infrastructure is repaired.

A recent yet novel proposal is to use photo-voltaic panels to provide connectivity to remote areas while simultaneously harvesting energy [6]. Off-the-shelf solar cells fixed on top of buildings for off-grid energy generation can simultaneously be used to receive modulated signals - although the bandwidth would be somewhat limited due to the high capacitance of the cells.

B. Existing FSO Systems

There are a number of commercially available FSO systems with a diverse range of data rates and ranges, shown in Fig. 2. The highest data rate commercially available system available is 30 Gbps, at a
range of approximately 1.5 km. The longest range system is just over 5 km with a data rate of 150 Mbps, but the best overall system claims to sustain 10 Gbps over 5 km.

These systems typically make use of 850 or 1550 nm wavelengths. The longer-range systems use 1550 nm due to it’s “eye-safe” nature, as the transmit power can be increased significantly compared to shorter wavelengths to overcome attenuation and turbulence without the system surpassing regulatory safety limits. Presumably, some systems make use of Wavelength Division Multiplexing (WDM) to achieve high data rates, such as 30 Gbps.

Numerous experimental FSO demonstrations have been reported in the literature using various degrees of freedom of light, including phase, polarization, wavelength, and space. Figure 2 also shows the experimental systems reported here, which are a selection of seminal or state-of-the-art works. A more in-depth description of the various techniques referred to here can be found in later sections.

One of the first demonstrations using multiple wavelengths was reported in 1999 by Nykolak et al. over 4.4 km using four channels, each carrying a 2.5 Gbps signal [7]. In 2006, a 5.625 Gbps transmission based on homodyne coherent detection was reported over a distance of 142 km between two Canary Islands [8]. Cvijetic et al. conducted a dual-polarization QPSK (Quadrature Phase Shift Keying), 112 Gbps FSO transmission test with coherent detection [9]. A 64 Gbps PAM-4 transmission over 100 m clear atmosphere environment has been demonstrated [10], and a 400 Gbps PAM-4 transmission over a short propagation distance of 1.2 m has been achieved [11]. A 200 Gbps quadrature amplitude modulation (32-QAM) FSO transmission over 55 m using a direct-detection Kramers-Kronig (KK) receiver-based on a 35 GHz photodiode has also been reported [12]. Coherent and non-coherent systems, as well as modulation schemes, are discussed in Sec. II-A.

Extremely high data rate systems have also been demonstrated. In 2009, a 1 Tbps IM/DD-based (Intensity Modulation/Direct Detection) WDM transmission over 210 m between two buildings in Pisa, Italy, was reported [13]. An inter-building 1.6 Tbps dual-polarization transmission over 80 m has been demonstrated using 16 WDM channels [14], and a 1.08 Tbps error-free transmission across a distance of 100 m subject to a light dust storm is reported in [15]. Impressively, in 2016, a team of researchers at the German Aerospace Center demonstrated a ground to mountain FSO transmission over a distance of
10.45 km with a data rate of 1.72 Tbps [16]. Mode Division Multiplexing (MDM) harnesses the space degree of freedom, and a record-setting data rate of 1.036 Pbps has been achieved over a lab-bench [17], with the longest range MDM demonstration to-date being 260 m [18] using two spatial modes and a data rate of 80 Gbps.

FSO systems benefit from the use of Multiple-Input-Multiple-Output (MIMO) configurations to increase the transmission capacity [19]. MIMO systems have been demonstrated [20] to achieve 80 Gbit/s in a system with two transmitters and two receiving apertures. The MIMO concept is based on the use of multiple transmit and/or receive apertures that are spatially separated (physically or via orthogonal modes), and have been equally proposed as a way to mitigate propagation effects [21], with more detail in Sec. IV.

FSO demonstrations involving mobile terminals have also been conducted. A 130 Mbps full-duplex communication link between two balloons at an altitude of 20 km over distances exceeding 100 km has been demonstrated [22], and so has an 80 Gbps transmission between a ground terminal and a moving Unmanned Aerial Vehicle (UAV) 50 m away [23]. The UAV was equipped with a retro-reflector and no transmitting or receiving components, resulting in a propagation round trip of 100 m.

II. THE BASICS OF FSO COMMUNICATIONS

A typical FSO communication system is illustrated in Fig. [1]. On a system level, a modulated beam produced by a laser diode or high-intensity LED is expanded to a size so that beam divergence is minimized and transmitted in the direction of the receiver. At the receiver plane, a large receive aperture collects as much light as possible, and a telescope arrangement is used to reduce the size of the beam for detection by a photodiode. In essence, the received electrical signal with time, $r_e$, in this setup can be expressed mathematically as

$$ r_e = \eta ([h_d h_p h_a h_t] I_s + I_a) + n,$$

(1)

where $I_s$ is the modulated, transmitted intensity signal and $\eta$ is the optical-to-electrical conversion efficiency given by

$$ \eta = R(\lambda) \eta_\lambda,$$

(2)

where $R(\lambda)$ is the detector (photodiode) responsivity at the system wavelength, $\lambda$, and $\eta_\lambda$ is the efficiency of the optics before the detector, which may include a band-pass filter for the system wavelength. $I_a$ is ambient light interference, but we assume that this is perfectly filtered and is therefore ignored in the model. The noise term, $n$, may be assumed to be Additive White Gaussian Noise (AWGN), predominantly from the photodiode and receiver electronics, and is discussed in Sec. II-B.

The channel response factors, $h_x$, are convolved with the transmitted signal, but the models are often such that this becomes simple multiplication, as shown here. Many of these factors change with time, but they can be assumed constant for the short duration of a symbol. The dominant factors are beam divergence, $h_d$, pointing errors, $h_p$, attenuation $h_a$, and turbulence-induced fading, $h_t$. These “channel impairments” are each discussed in detail in Sec. III.

While this model appears simple, there are numerous considerations associated with each term and, indeed, each component of the system - as expected in any complex system. In this section, we ignore channel impairments (i.e., $h_x = 1$) and focus on the essence of an optical communication system: transmission, modulation, detection, demodulation, and noise.

A. Modulation and Demodulation

1) Intensity Modulation / Direct Detection (IM/DD): Intensity Modulation (IM) is the simplest method to modulate a laser source using an electrical (typically high-speed) modulating signal and is widely used when data rates up to about 10 Gbps are required. The intensity of an optical source is varied in accordance with the amplitude of a modulating signal, but phase information cannot be directly encoded onto the
laser beam with this method. Direct Detection (DD) is where the intensity of the optical signal governs the amplitude of a resulting electrical signal through the use of a detector element such as a photodiode. [24]

The most widespread modulation format is known as On-Off Keying (OOK), where a binary one corresponds to the laser being "on", and when the laser is "off" the signal represents a zero. OOK requires at receiver a decision threshold to make an optimal decision. A more general format is Pulse Amplitude Modulation (PAM), whereby different signal amplitude levels are used to encode the information (OOK is a subset of PAM). In general, the number of bits of data that can be transmitted per modulated symbol is \( \log_2 N \), where \( N \) is the number of unique symbols in the modulation scheme. Therefore, PAM-N with \( N > 2 \) allows higher transfer rates than OOK; however, it has some issues at higher spectral efficiency values. In addition, since the information is encoded directly onto the amplitude of the signal, schemes such as PAM are adversely affected by intensity fluctuations brought about by the channel [25].

In IM/DD links, information can also be encoded using a Pulse Position Modulation (PPM) format. In a PPM scheme, the amplitude of the pulse is kept constant, but the positions of the pulses, which have the same width, are varied in accordance with the amplitude of the modulating signal. The information is encoded in the time dimension rather than amplitude, without requiring a threshold at the detector to make an optimal decision, therefore, making PPM more robust in comparison to PAM in the presence of intensity fluctuations. A challenge is that precise time synchronization between the transmitter and the receiver is critical for this modulation scheme.

Sub-carrier Intensity Modulation (SIM) is a scheme for IM/DD systems, which enables modulation formats that also make use of phase, with some caveats. In SIM modulation, the data is first modulated onto an RF sub-carrier, and then this is used to drive the intensity of the laser. The resulting optical envelope corresponds to the modulated RF sub-carrier. Critically, a DC bias must be added to the SIM signal before it can be used to drive the laser since negative voltages are invalid [26]. Thus, this scheme suffers from a lower dynamic range (modulation depth), but it is otherwise very flexible. A similar but more refined approach to SIM is to modulate the data using real-valued Orthogonal Frequency Division Multiplexing (OFDM) signals, also known as discreet multi-tone (DMT). Typically, OFDM signals are used for radio, but by harnessing Hermitian symmetry, it is possible to sacrifice some spectral efficiency in return for satisfying the non-negative real-valued criteria of optical signals. Real valued OFDM is sensitive to nonliterary, but this is more of a problem in fiber than in free-space.

Recently, an alternative to sub-carrier based methods for the use of advanced modulation formats requiring complex values (i.e., phase and amplitude) has been invented. A so-called Kramers-Kroning (KK) receiver is able to reconstruct a complex-valued signal such as QPSK or QAM using only intensity detection by performing a digital phase retrieval [27].

2) Coherent Systems: In contrast to IM/DD systems, various complex modulation formats can be implemented in coherent systems making use of the amplitude, phase, frequency, and polarization, for example QPSK and QAM. The transmitter uses a narrow line-width continuous wave (CW) laser to generate the optical carrier signal. A Mach-Zehnder intensity and / or phase modulator is then used on this carrier signal to apply an arbitrary modulation.

A coherent system receiver mixes the received optical signal coherently with a narrow line-width continuous wave (CW) laser, known as the local oscillator (LO) as depicted in Fig 3.

There are two detection techniques in coherent systems known as homodyne and heterodyne. In homodyne detection, the frequency of the LO coincides with the carrier signal frequency, so that the total frequency of the mixed signal is zero. While in the heterodyne detection, the difference between the carrier signal frequency and the LO frequency, known as the intermediate frequency, is in the microwave region (\( \sim 1 \text{ GHz} \)). Heterodyne detection is simpler than homodyne because digital signal processing can be used to demodulate the signal from the intermediate frequency signal, even if the frequency is not perfectly stable. In homodyne detection, the LO must be extremely precise, which is difficult to
achieve and also typically more expensive. In terms of signal-to-noise (SNR), homodyne detection has a 3 dB improvement compared to the heterodyne case. The receiver sensitivity in a coherent system is up to 20 dB better than an IM/DD system. A coherent detection system is clearly more complicated than direct detection. Still, the increase in sensitivity and the requirement for more signal processing may be necessary for long-range or very high-speed systems.

B. Detector Noise

Noise in FSO systems is mainly introduced at the receiver, with the type of receiver, for example, PIN photodiode or Avalanche Photodiode (APD), influencing the makeup of the total noise. The main sources of noise at the receiver are summarised below.

The noise power due to these various sources is dependent on the bandwidth \( \Delta f \) in Hertz of the detection system, with wider bandwidths resulting in more noise. Each of the noise terms may be lumped into an overall AWGN term as in Eq. (1), according to Eq. (5).

**Thermal noise** (also known as the Johnson or Johnson-Nyquist noise) occurs in conducting materials and is caused by the thermal fluctuations of the electrons in the receiver circuit. Thermal noise is frequency independent (i.e., white noise) and has a Gaussian power spectral density with zero mean and variance given by

\[
\sigma^2_{th} = \frac{4KT}{R},
\]  

(3)

where \( K \) is Boltzmann’s constant, \( T \) is the absolute temperature in Kelvin, and \( R \) is the resistance under consideration. Every resistance introduces noise, which in this case would predominantly be the shunt resistor connected to the photodiode, but the load resistor of a circuit or even internal resistance of a photodiode are also significant contributors.

**Shot noise** is associated with the quantum particle nature of light and describes the random fluctuations over time of the number detected photons. Similarly, shot noise is also caused by the movement of electrons in a circuit (or through the photodiode). With large signals, shot noise follows Gaussian statistics with zero mean and variance given by

\[
\sigma^2_{sn} = 2q(R(\lambda)P_{Rx} + i_d),
\]  

(4)

where \( q = 1.6 \times 10^{-19} \) is the electron charge and \( P_{Rx} \) is the received optical power (intensity). \( i_d \) is the dark current, which is the current that flows through the photodiode in the absence of light - at non-zero temperature, random electron/hole pairs are produced, which induce a small current. The dark current is also a function of the reverse bias voltage, with higher voltage resulting in higher dark current. For APDs, the shot noise variance is slightly different as it includes a multiplication factor and an additional dark current term.
Since these noise types are all approximately Gaussian, the total noise variance can be written as follows
\[ \sigma_n^2 = \sigma_{th}^2 + \sigma_{sn}^2. \] (5)

The total noise power for the receiver without amplification is therefore
\[ P_n = \sigma_n^2 \sqrt{\Delta f}. \] (6)

**Noise Equivalent Power (NEP)** is a convenient metric that includes thermal and shot noise contributions of a device such as a photodiode. The minimum Noise Equivalent Power (NEP\(_{\text{min}}\) or just NEP) is typically available on device datasheets and at the wavelength of interest is given by
\[ \text{NEP}(\lambda) = \text{NEP}_{\text{min}} \left( \frac{R_{\text{max}}}{R(\lambda)} \right), \] (7)

where \( R_{\text{max}} \) is the maximum responsivity of the detector regardless of wavelength. The noise power due to the NEP in a specific bandwidth is simply
\[ P_{\text{NEP}} = \text{NEP}(\lambda) \sqrt{\Delta f}. \] (8)

Since we require a Signal to Noise Ratio (SNR) greater than one for the signal to not be hidden in the noise, the minimum detectable power is also given by Eq. (8). This is a best-case estimation of the receiver sensitivity in the absence of other noise sources, such as the amplifier.

**Amplifier noise** stems from many of the same sources as described in this section. Not only is noise added to the signal by the amplifier, but the noise already present in the signal is also amplified. The total output noise power is, therefore, a function of the input noise power, \( P_n \), the gain of the amplifier, \( G \), and the noise contribution of the amplifier, \( P_a \), as
\[ N = P_n G + P_a. \] (9)

Thus, this noise power, \( N \), may be used to determine the signal to noise of the receiver or used in Monte-Carlo simulations as the AWGN power.

### III. Channel Impairments and Models

In this section we summarise the significant factors that must be taken into account for an effective FSO communication system. They are beam divergence (Sec. III-A), atmospheric attenuation (Sec. III-B), atmospheric turbulence (Sec. III-C) and pointing errors (Sec. III-D).

At the end of each of these subsections, we provide what the resulting channel response factor, \( h_x \), is (as per Eq. (1)), as well as gradually develop an expression for the received power, \( P_{Rx} \), in analogy to the Friis equation used in radio. The reader should use whichever form is most convenient. The channel responses are useful for a signal-level simulation of a link, but the received powers are useful for estimating the system SNR and link budget.

In Sec. III-E we provide a simple example to demonstrate how each of these factors may be used in a design situation.

#### A. Beam divergence

As it propagates, a laser beam tends to diverge, as illustrated in Fig. 4. The Rayleigh length, \( z_R \), is the distance from the beam waist, along the propagation direction, where the beam radius is increased by a factor of the square root of two and the cross-sectional area by two. The Rayleigh length can also be used to determine whether a plane or spherical wave approximation is best, especially for turbulence fading models.
Fig. 4. A schematic illustration of beam divergence. Less divergence is possible with a larger beam waist, allowing for longer links.

For a Gaussian beam, the Rayleigh length is expressed as a function of the beam waist $w_0$ and the wavelength $\lambda$:

$$z_R = \frac{\pi w_0^2}{\lambda}.$$  \hfill (10)

The radius of a beam at a distance $z$ (along the propagation direction) from the Gaussian beam waist is

$$w(z) = w_0 \sqrt{1 + \left(\frac{z}{z_R}\right)^2}.$$  \hfill (11)

The plane wave approximation for a propagating Gaussian beam is valid when

$$\frac{\lambda z}{\pi w_0} \ll 1,$$  \hfill (12)

and when this expression is $\gg 1$, then the spherical wave approximation is better. The half-angle spread of a Gaussian beam in radians can also be expressed as a function of the Rayleigh length as

$$\theta_{\text{div}} = \frac{w_0}{z_R},$$  \hfill (13)

and is labeled in Fig. 4.

Since the beam has expanded when reaches the receive aperture, only a clipped portion is captured and so the received power, $P_{R_x}$ can be written as

$$P_{R_x} = P_{T_x} \frac{D_{R_x}^2}{(D_{T_x} + 2\theta_{\text{div}} L)^2},$$  \hfill (14)

where $P_{T_x}$ is the transmitted power. $D_{T_x}$ and $D_{R_x}$ are the diameters of the aperture of the transmitter and receiver, respectively, and $L$ is the link distance. Similarly, Eq. (14) can be rewritten as a gain for Eq. (1) as

$$h_d = \frac{D_{R_x}^2}{(D_{T_x} + 2\theta_{\text{div}} L)^2}.$$  \hfill (15)

B. Atmospheric Attenuation

As a laser beam propagates through a medium such as the atmosphere, it not only diverges, but it is absorbed and scattered by particles in that medium, depending on their size and concentration. This overall attenuation is described by Beer’s law:

$$P_R(z) = P_T \exp(-\gamma(\lambda)z),$$  \hfill (16)

where $P_R(z)$ is the power at position $z$ and $P_T$ is the initial transmitted power. The exponential decay is ruled by a wavelength-dependent extinction coefficient $\gamma(\lambda)$ in $[\text{km}^{-1}]$, known also as the attenuation coefficient. $\gamma(\lambda)$ accounts for the absorption and scattering from the interaction of a laser beam with
various gaseous molecules and aerosol particles in the atmosphere. As such, the attenuation coefficient can be expressed as follows:

\[ \gamma(\lambda) = \alpha_{ml}(\lambda) + \alpha_{al}(\lambda) + \beta_{ml}(\lambda) + \beta_{al}(\lambda), \]  

(17)

where \( \alpha_{ml} \) and \( \alpha_{al} \) are the molecular and aerosol absorption coefficients respectively, and \( \beta_{ml} \) and \( \beta_{al} \) are the molecular and aerosol scattering coefficients, respectively.

FSO system wavelengths are chosen in accordance with the atmospheric absorption spectra and generally fall into so-called “windows” where the absorption is minimal. Therefore the absorption can be ignored in comparison to the scattering effects.

Depending on the weather conditions, the attenuation in the atmosphere varies significantly. When the air is clear, attenuation can be as low as 0.2 dB/km, to more than 270 dB/km under heavy fog for the 1550 nm wavelength window. For convenience, several common visibility ranges and corresponding attenuation losses are summarized in Tab. I. However, for a more precise calculation, Eq. (17) should be used.

There are two well-known scattering types, which depend on the size of the scattering particles compared to the wavelength of the incident beam:

**Rayleigh scattering** is caused by air molecules smaller than the wavelength. The Rayleigh scattering coefficient is inversely proportional to the fourth power of the wavelength and can be approximated by the following empirical model (28, and references therein):

\[ \beta_{ml}(\lambda) = 0.827N_p A_p^3 \lambda^{-4}, \]  

(18)

where \( N_p \) is the number of particles per unit volume, and \( A_p \) is the cross-sectional area of scattering. Rayleigh scattering is negligible for wavelengths beyond 800 nm [29].

**Mie scattering** is caused by particles that are larger or comparable to the operating wavelength. In Mie scattering, the light tends to scatter in the forward direction. Fog and haze are the major causes of Mie scattering, and since these conditions affect the weather “visibility”, this metric is used to determine the scattering losses. A commonly used empirical model for the attenuation due to Mie scattering is expressed as a function of the visibility range, \( V \) [km], as follows [30]:

\[ \beta_{al}(\lambda) = 3.91 \frac{V}{\lambda_0} \left( \frac{\lambda}{\lambda_0} \right)^{-q}, \]  

(19)

where \( \lambda_0 = 550 \text{ nm} \) is the visibility range reference wavelength, and the coefficient \( q \) is the size distribution of the scattering particles. \( q = 1.6 \) for \( V = 50 \text{ km} \), \( q = 1.3 \) for \( 6 < V < 50 \text{ km} \), and \( q = 0.585V^{1/3} \) for \( V < 6 \text{ km} \). Different forms of \( q \) are also given in [30], which leads to different attenuation values, in particular at low visibility \( V < 500 \text{ m} \). Note that there are other models to predict aerosol attenuation in the literature [31]. To obtain the attenuation coefficient in [dB/km], \( \beta_{al}(\lambda) \) in Eq. (19) should be multiplied by \( 10 \log(e) \).

It is worth mentioning that a third scattering type, induced by particles that are significantly larger than the operating wavelength, may occur, which is typically caused by rain, snow, and dust. An approximate expression for the attenuation due to rain \( \gamma_{rain} \) [dB/km], is given by [32]:

\[ \gamma_{rain} = K_R R^{\alpha_R}, \]  

(20)

where \( R \) [mm/hr] is the precipitation intensity. \( K_R \) and \( \alpha_R \) are the model parameters whose values depend on raindrop size and rain temperature. The attenuation typically ranges from 1 to 24 dB/km for precipitation intensity varying from 2.5 mm/hr (light rain) and 100 mm/hr (very heavy rain).

For snow induced scattering the attenuation coefficient, \( \gamma_{snow} \) [dB/km], can be estimated with the following approximate expression [32]:

\[ \gamma_{snow} = a_S S^{b_S}, \]  

(21)
TABLE I
ATMOSPHERIC ATTENUATION AT 850 AND 1550 NM FOR VARIOUS WEATHER CONDITIONS.

<table>
<thead>
<tr>
<th>Atmospheric Condition</th>
<th>Visibility [km]</th>
<th>Attenuation [dB/km]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>850 nm</td>
</tr>
<tr>
<td>Clear air</td>
<td>23</td>
<td>0.42</td>
</tr>
<tr>
<td>Haze</td>
<td>4</td>
<td>2.8</td>
</tr>
<tr>
<td>Mist</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Light fog</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>Moderate fog</td>
<td>0.5</td>
<td>28</td>
</tr>
<tr>
<td>Dense fog</td>
<td>0.2</td>
<td>73</td>
</tr>
<tr>
<td>Heavy fog</td>
<td>0.05</td>
<td>309</td>
</tr>
</tbody>
</table>

where \( S \) is the snowfall rate [mm/hr], \( \alpha_S \) and \( b_S \) are snow parameters whose values are different if the snow is wet or dry. For a snowfall rate of 4 mm/hr, the approximate attenuation is 11 dB/km for wet snow and 38 dB/km for dry snow.

For dust, an empirical model to estimate the attenuation, \( \gamma_{dust} \) [dB/km], is given by \[33\]

\[
\gamma_{dust} = K_d V^{b_d},
\]  

where \( K_d \) and \( b_d \) are model parameters, whose values can be chosen as 52 and -1.05, respectively. The typical attenuation loss due to dust ranges from 50 to 300 dB/km for visibility ranges from 0.2 to 1 km, which is considerably higher than losses due to fog.

The total path loss equation for an FSO link involving all of the attenuation factors described here, or \( h_a \) in Eq. (1), is given by

\[
h_a = \exp(-\gamma(\lambda)L),
\]

where \( \gamma \) could be from Eqs. (17), (20), (21) or (22), if similar weather conditions prevail over the length of a link. Otherwise, the attenuation should be calculated in a piece-wise manner. When this is included with divergence (Eq. (14)) the overall received power is now

\[
P_{Rx} = P_{Tx} \left( \frac{D_{Tx}^2}{D_{Rx}^2 + 2\theta_{div} L} \right)^2 \exp(-\gamma(\lambda)L).
\]

C. Atmospheric Turbulence

Optical turbulence in the atmosphere is a result of random, spatially varying fluctuations in the air temperature. The widely-accepted Kolmogorov model for these fluctuations relates these temperature fluctuations to refractive index fluctuations. When a laser beam propagates through these random regions (called cells), it is aberrated, leading to wavefront distortions and intensity fluctuations. Two boundary values specify the average size of these turbulent cells: the inner scale, \( l_0 \), which is on the order of millimeters and the outer scale, \( L_0 \), which is on the order of meters (typically twice the height of a link above the ground if that is less than about 100 m) \[34\].

Inner (small) scale turbulence leads to scintillation, which appears as “speckle” on the received beam, shown in Fig. 5. The resulting random fluctuations in intensity are the leading cause of turbulence-induced fading in FSO communications and is therefore described in detail in this section. Outer (large) scale turbulence, over longer distances, causes the beam to “corkscrew” through the air, resulting in beam wander at the receive aperture, also leading to fading, which is described in Sec. III-C2. Beam wander can be considered and dealt with in a similar manner to pointing errors.

A convenient approach to modeling atmospheric turbulence is by using the Power Spectral Density (PSD) of the refractive index fluctuations. The Kolmogorov spectrum, which is a simple model that assumes the inner scale is zero and the outer scale is infinity is given by

\[
\Phi_n^K(K) = 0.033C_n^2K^{-11/3},
\]
where $C_n^2$ is the refractive index structure parameter and $\kappa$ is a spatial frequency coordinate. A more general form of the Kolmogrov spectrum is the modified von Kármán spectrum that accounts of the inner and outer scales of turbulence: 

$$\Phi_{n}^{MK}(\kappa) = 0.033 C_n^2 \exp\left(-\frac{\kappa^2}{\kappa_m^2}\right) (\kappa^2 + \kappa_0^2)^{-11/6},$$

where $\kappa_m = 5.92/l_0$ and $\kappa_0 = 2\pi/L_0$. The refractive index structure parameter can be estimated using the Huffnagle-Valley model as follows [35]:

$$C_n^2(h) = 0.00594\left(\frac{v_{wind}}{27}\right)^2(10^{-8}h)^{10} \exp(-h/1000) + 2.7 \times 10^{-16} \exp(-h/1500) + A \exp(-h/100),$$

where $h$ is the altitude in [m], $v_{wind}$ is the RMS wind speed [m/sec], and $A$ is the nominal value of $C_n^2(0)$ at the ground.

Numerous methods to simulate laser beam propagation through atmospheric turbulence have been developed [36]–[39]. One of the more physically realistic approaches is called the Split-Step Fourier method, shown schematically in Fig. 5 [40]. In this method, a beam is propagated through multiple phase screens generated using PSDs such as Eqs. (25) and (III-C) and spaced to simulate a real link of several kilometers accurately. It is clear from the figure that the received beam is distorted, and depending on the position of the beam on the detector, random intensity fading will occur. A set of phase screens with a resolution of $1024 \times 1024$, generated following the previously described turbulence models, is depicted in 6. The outer and inner turbulence scale values are chosen as $L_0 = 100$ m, and $l_0 = 10$ cm following [36].

Of course, links may also be physically tested in a laboratory environment using rotating heated pipes [41], computer-generated holograms loaded on spatial light modulators [42], glass plates [43] and controllable climate chambers [33], [44].

It should be noted that there has been some investigation into the anisotropy of turbulence, especially in boundary layers of the atmosphere [45]. This is important for ground-to-space FSO links, but will not be discussed further.

1) Statistical Models for Turbulence: Performing a full simulation to design an FSO link is impractical. Several convenient metrics and statistical techniques exist to accurately predict the performance of FSO links and will be described below.

Intuitively, the further a beam propagates through turbulence, the higher the impact of that turbulence. The Fried parameter, $r_0$ (also called atmospheric coherence length), is a measure of the turbulence strength, which takes into account the length of propagation. As a rule of thumb, smaller $r_0$ lengths result in stronger turbulence. For a plane wave (approximately a collimated Gaussian beam) in unspecified turbulence, $r_0$ is given by

$$r_0 = \left(0.423 k^2 \int_0^L C_n^2(z')dz'\right)^{-3/5},$$

(27)
Fig. 6. A set of turbulence phase screens generated following the Kolmogorov, and the modified van Kármán models at different turbulence regimes with Fried parameter values of 0.12 (weak), 0.06 (moderate turbulence), and 0.01 (strong turbulence). The phase in each of the screens wraps from 0 to $2\pi$.

where $k = 2\pi/\lambda$ is the wave-number, and the propagation distance is incorporated by the integral over path length $L$. When assuming a horizontal FSO propagation, Eq. (27) simplifies to

$$r_0 = 1.68 \left( C_n^2 L k^2 \right)^{-3/5}. \quad (28)$$

The Fried parameter is related to another quantity known as Strehl Ratio (SR) defined as the ratio of the on-axis intensity with, to the initial intensity without turbulence, $I(0,0,L)$ and $I_0(0,0,L)$ respectively. The SR is useful to roughly estimate the average attenuation due to turbulence, for example as $h_t$ in Eq. (1) and can be expressed as

$$SR = \frac{\langle I(0,0,L) \rangle}{I_0(0,0,L)} \approx \frac{1}{1 + (D_{Rx}/r_0)^{5/3} r_0^{6/5}}. \quad (29)$$

The scintillation index, $\sigma^2$, is a common parameter similar to the SR but is useful where knowledge of the intensity in the absence of turbulence is unknown, which is typically the case in non-lab environments, where

$$\sigma^2 = \frac{\langle I^2 \rangle - \langle I \rangle^2}{\langle I \rangle^2} = \frac{\langle I^2 \rangle}{\langle I \rangle^2} - 1. \quad (30)$$

Many probabilistic fading models can be used to calculate a scintillation index. These may be used in Monte-Carlo simulations of FSO links with less computational intensity than turbulence propagation simulations. Also, closed-form solutions are available for several models for capacity, outage probability, and bit error rate - important metrics for an FSO communication link.

One of the first fading models for FSO channels was the log-normal channel distribution [46]. It is restricted to the weak turbulence regime and sometimes underestimates the behavior at the peak and in the tails of the probability distribution. The negative exponential channel distribution is applicable for a very strong turbulence regime, but it is not consistent with small-scale turbulence effects [47].

A more recent model for FSO channels is the gamma-gamma, which provides an excellent fit with measurement data for a wide range of turbulence conditions, both weak and strong, with some example curves shown in Fig. 7. This gamma-gamma model is also analytically tractable, and so expressions for the outage link probability, error probability (i.e., bit error rate), and capacity are available in the literature [48]. The Málaga distribution, also known as the $M$-distribution, unifies most of the existing statistical models and covers a wide range of atmospheric turbulence conditions from weak to strong [49]. In particular, log-normal and gamma-gamma are particular cases of the Málaga distribution.
Here we provide a brief description of the gamma-gamma fading model. In this model, the irradiance, $I_a$, is modeled as the product of two statistically independent processes: one for the small scale effects and another for the large scale effects, each of which is modeled by a Gamma distribution. The resulting Probability Distribution Function (PDF) for the fading (plotted in Fig. 7) is [35]

$$f_{I_a}(I_a) = \frac{2(\alpha\beta)^{\alpha+\beta/2}}{\Gamma(\alpha)\Gamma(\beta)} I^{(\alpha+\beta)/2-1} K_{\alpha-\beta} \left(2\sqrt{\alpha\beta I}\right),$$

(31)

where $\Gamma(\cdot)$ is the gamma function and $K_v(\cdot)$ is modified Bessel function of the second kind and order $v$. $\alpha$ and $\beta$ are the turbulence fading parameters, and under the plane wave approximation expressed as follows [28]:

$$\alpha = \left[\exp \left(\frac{0.49\sigma_R^2}{1 + 1.11\sigma_R^{12/5}L^{7/6}}\right) - 1\right]^{-1},$$

(32)

$$\beta = \left[\exp \left(\frac{0.51\sigma_R^2}{1 + 0.69\sigma_R^{12/5}L^{5/6}}\right) - 1\right]^{-1},$$

(33)

where $\sigma_R = 0.5C_n^2k^{7/6}L^{11/6}$ is the Rytov variance, which is also an indicator of turbulence strength and is approximately equal to the scintillation index in weak turbulence. The scintillation index for the gamma-gamma model is

$$\sigma_I^2 = \frac{1}{\alpha} + \frac{1}{\beta} + \frac{1}{\alpha\beta}.$$

(34)

One way to mitigate the effect of turbulence is to use aperture averaging. If the size of the receiving aperture is increased to at least $3l_0$ (typically), the fluctuations over the aperture will be averaged, which decreases the scintillation index and thus boosts the overall received power. Fried first proposed aperture averaging as a way to reduce the effect of atmospheric turbulence on laser beams [50]. Churnside defined an aperture averaging factor $A_g$ to measure the reduction of scintillation index defined as the ratio of the
scintillation index of the signal for a receiver with diameter $D_{Rx}$ to a receiver with an infinitesimally-small aperture (i.e., can be approximated by a point $D = 0$) as [51]

$$A_g = \frac{\sigma_I^2(D_{Rx})}{\sigma_I^2(0)}.$$  \hspace{1cm} (35)

In Kolmogorov turbulence and assuming a plane-wave approximation, $A_g$ is given as follows [34]:

$$A_g = \left[1 + 1.062 \left(\frac{kD_{Rx}^2}{4L}\right)\right]^{-7/6},$$  \hspace{1cm} (36)

Following our model nomenclature in Eq. (1), $h_t = I_a$. A closed-form solution for the mean fading, that is available for a wide range of atmospheric turbulence regimes, is not available, but the Strehl Ratio can be used approximately for link budget purposes. We also note that under some particular turbulence regimes, an approximated expression for $I_a$ can be derived, as shown in [52].

Revisiting the received power equation (24) and taking into account loss due to atmospheric turbulence and the aperture averaging in case it is applied, the equation becomes

$$P_{Rx} = P_{Tx} \frac{D_{Rx}^2}{(D_{Tx} + 2\theta_{div}L)^2} \exp\left(-\gamma(\lambda)L\right)I_aA_g.$$  \hspace{1cm} (37)

2) Turbulence-Induced Beam Wander: One of the effects of large scale atmospheric turbulence is a random lateral displacement of the beam at the receiver plane that changes with time. This may result in deep fading if the beam wanders too far from the receive aperture. While beam wander is accounted for in the gamma-gamma fading distribution, the effect has been modeled in isolation.

Beam wander is typically modeled as if there is a pointing error at the transmitter, as shown in Fig. 8. The long term average of this wandering spot (called the short term beam - which is, in fact, the beam predicted after taking into account divergence and turbulent scintillation) at the receiver forms a larger Gaussian-shaped beam due to the central limit theorem [53].

The radius of this long term beam can be expressed as the sum of the radius of the short term beam, $\omega_{ST}$, and the average radial variance, $\langle r_c^2 \rangle$, as follows:

$$\omega_{LT} = \omega_{ST} + \langle r_c^2 \rangle.$$  \hspace{1cm} (38)

Andrews and Philips provided a beam wander model that is valid for all turbulence conditions [35]

$$\langle r_c^2 \rangle = 4\pi^2k^2W^2 \int_0^L \int_0^\infty k\Phi_n(k)H_{LS}(1 - e^{\lambda L\nu^2(1-z/L)^2/k})dkdz,$$  \hspace{1cm} (39)
There are various simplifications of Eq. (39), depending on the turbulence model and beam wavefront. For a collimated beam in Kolmogorov turbulence with infinite outer scale the expression simplifies to

\[ \langle r^2 \rangle = 2.42C_n^2L^3\omega_0^{-1/3}. \]  

(40)

Knowledge of the radial variance is useful in determining the required receive aperture size and for estimating the average number of fades (and errors) in a specific period, which is required for channel code design. In addition, recent research has shown that beam wander and related angle-of-arrival fluctuations may be modeled more accurately as a memory process \[55\]. Over long distances, it may be beneficial to use an automated tracking system at the transmitter (or receiver - with less benefit), and the radial variance helps inform this design.

D. Pointing Errors

FSO systems require continuous alignment between the communicating terminals. Maintaining this alignment is challenging if the terminals are fixed on structures that move or vibrate (for example swaying due to wind loads \[56\]), or if the link must operate over long distances. Small earthquakes, thermal expansion as well as other vibration sources, can also introduce pointing errors.

Common mitigation strategies for pointing errors are active alignment systems and / or ensuring the beam is suitably expanded by the time it reaches the receiver. Active systems typically require precise and stable control systems as well as micro-radian accuracy. This allows the beam to be well collimated (or even focused) at the receiver, ensuring high SNR and thus high data rate. A further benefit of an active system, depending on its design, is the possibility of actively correcting turbulence-induced tip and tilt aberrations \[57, 58\]. An expanded beam is a cheap and simple solution but is unlikely to be feasible over long distances due to receiver sensitivity issues.

The pointing error can be approximated by a Gaussian form where there is an average pointing error radius, \( r \), as follows \[59\]:

\[ h_p = A_0 \exp \left( -\frac{2r^2}{w_{eq}^2} \right), \]  

(41)

where \( A_0 = [\text{erf}(v)]^2 \), \( w_{eq}^2 = w(z)^2 \frac{\sqrt{\pi \text{erf}(v)}}{(2v \exp(-v^2))}, \) and the parameter \( v = \sqrt{\pi} (D_{Rx}/2)/(\sqrt{2}w(z)) \).

Here, \( A_0 \) denotes the fraction of power collected by the receiver in the absence of pointing errors and \( w_{eq} \) refers to the equivalent beam width at the receiver with pointing errors. \( \text{erf}(.) \) is the error function. The radial displacement, \( r \), is typically modeled using a PDF, for example, a Rayleigh distribution for building sway \[59\].

E. Summary

In summary, the Friis equation, taking into account the different losses and gains, can take the following form:

\[ P_{Rx} = P_{Tx}L_dL_aL_pL_tG_g, \]  

(42)

with \( L_d \) is a loss term, which can takes \( h_d \) (defined in [15]). \( L_a \) is the atmospheric attenuation loss coefficient that can take \( h_a \) (defined in [23]), \( L_p \) denotes the loss due to pointing errors, and \( L_t \) represents the loss due to the atmospheric turbulence. \( G_g \) can be viewed as the aperture averaging gain that can take the value of the aperture averaging factor \( A_g \) (defined in [35]). Due to the statistical nature of the atmospheric turbulence and pointing errors, no constant loss terms could be given to those quantities. Additional losses like window attenuation or optical component reflection loss could be added to Eq. (42) as multiplication terms.
IV. FURTHER TECHNIQUES

In this section, we discuss a range of additional techniques that can be harnessed in an FSO system. Some of these techniques are still experimental and under significant research, and others are well understood but dependent on the situation due to their additional complexity or cost.

A. Channel coding

Channel coding, also known as Forward Error Correction (FEC), and involves error coding schemes to improve the robustness of FSO links against the randomly varying propagation effects by inserting redundant bits in the data streams. Various FEC schemes have been proposed for FSO systems, including Convolutional codes, Reed-Solomon codes, Turbo codes, and Low Density Parity Check (LDPC) codes [60], [61].

Rateless codes, whose code rate is not fixed at the transmitter, have also been investigated in the case of FSO. Anguita et al. considered the use of rateless codes in an adaptive coding scheme where the rate of the coding could be changed depending on the channel state [62]. Djordjevic also proposed the use of an adaptive coding and modulation scheme to cope with the effect of turbulence [63]. The idea here is to estimate the channel conditions at the receiver and feed it back to the transmitter using an RF feedback signal so that the transmitter can adapt the modulation and coding scheme, in particular, through an LDPC coded modulation. Such a technique has proven to be efficient for weak and strong turbulent conditions but could be ideal for system implementation due to peak power constraints [64]. Very recently, an adaptive coded modulation approach incorporating probabilistic constellation shaping has been reported [65].

The coherence time for FSO channels subject to turbulence varies between 100 $\mu$s and 10 ms [66]. In order to benefit from time diversity by some channel coding schemes, long interleavers are needed, which may require large memory sizes in the case of high bit rate transmissions and may lead to severe latencies [67]. Besides the memory constraint, before employing any channel coding technique, another issue to consider is the receiver complexity, which is sometimes overlooked in some reports in the literature.

B. Diversity

Spatial diversity refers to the use of multiple transmitting lasers and (or) multiple receiving detectors, as shown in 9 [68]. The simplest form of spatial diversity consists of sending the same copy of the signal using spatially separated transmitters. The transmitters must be spaced by at least $r_0$ to ensure statistical independence; otherwise, a diversity gain will not be achieved. The concept of sending the same information signal on different apertures is referred to as repetition coding (RC) [48]. At the receiver, the signals collected by different receiving apertures are combined [21]. Another form of spatial diversity is called modal diversity, where multiple, co-propagating higher-order spatial modes are used instead of physical aperture separation. There have been some recent works demonstrating this for Orbital Angular Momentum (OAM), and other mode sets [69]–[72]. Spatial diversity can also be an effective way to cope with pointing errors, particularly if multiple receiving apertures are used to detect a beam from a transmitting laser. In the case of multiple independent receivers, different combining techniques can be considered, including Maximum-Ratio Combing (MRC), Selective Combing (SC), and Equal Gain Combing (EGC) [73]. In an MRC scheme, the received signals by the different apertures are weighted according to their SNR and summed. The SC technique corresponds to the link with the highest SNR, and the EGC scheme corresponds to the coherent sum of all the received signals.

The performance of an IM/DD MIMO-FSO system subject to turbulence conditions following a gamma-gamma distribution is studied in [74]. The authors assumed EGC and MRC combining schemes. The analysis revealed that MRC has little to moderate gain compared to EGC at different MIMO settings in terms of the number of transmitting and receiving apertures, which makes the EGC more attractive for
Fig. 9. A block diagram of a MIMO FSO system with multiple transmitters and receivers. These apertures must be separated by at least $r_0$ for a diversity gain, but modal diversity, where orthogonal spatial modes are used instead of physical separation, may also be possible.

practical implementation due to implementation simplicity. Various space-time coding schemes have been proposed to improve the system performance in the presence of fading. Simon and Vilnrotter proposed a modified Alamouti code for IM/DD FSO systems, which involves a DC bias to avoid transmitting negative copies of the signal similar to the original Alamouti code proposed for RF systems [75].

It has been shown that space-time coding is unnecessary for IM/DD FSO systems and is outperformed by repetition codes, with analysis conducted under the assumption of log-normal distributed turbulent channels [76]. The authors also reported that employing space-time codes could even be detrimental in some cases. These claims were revisited in [77]; considering a coherent FSO system with heterodyne detection over gamma-gamma channels, where the authors show that space-time coding is better than repetition codes.

C. Adaptive Optics

Adaptive Optics (AO) is an optical technology that aims to minimize the effect of wavefront distortions caused by atmospheric turbulence. The correction is typically performed by an optomechanical system formed by a wavefront sensor, wavefront correcter, and a control unit. The aberrations of the incoming wavefront are measured, the control unit computes a suitable inverse or correction phase, which is then applied to the wavefront corrector device. This device can be an array of deformable mirrors, liquid crystals, or even a digital micro-mirror device. An illustration of a possible AO system configuration is depicted in Fig. 10.

A simpler, far less expensive, and remarkably effective system uses only a tip/tilt mirror to correct only the lower order aberrations of turbulence and a simple quadrant photodiode to measure the incoming tilt. The overwhelming majority of the effect or turbulence is tip and tilt, and so this simplified AO approach is usually sufficient [78].

The use of wavefront sensors to estimate the impact of turbulence on laser beams was demonstrated in [79], [80]. Using wavefront sensors for FSO systems could be expensive for large scale deployment, but sensor-less AO solutions have been proposed [81]. In particular, the use of the Gerchberg-Saxon (GS) algorithm for computational phase retrieval and correction has shown promise in this area [82], [83]. The GS algorithm is also useful for correcting or compensating for aberrations in a complex optical system [84], which is particularly useful for systems employing higher-order spatial modes.

By using the reciprocity of the turbulent atmosphere, it is also possible to perform turbulence pre-compensation using AO kits, as demonstrated in [85]. In bidirectional communication, the beam from each transceiver can serve as a beacon to the other, assuming that the effect of the turbulent atmosphere is reciprocal in both superimposed links.
D. Higher-Order Spatial Modes

In 2004, Gibson et al. conducted an FSO transmission using Mode Division Multiplexed (MDM) Orbital Angular Momentum (OAM) modes [86], heralding a new wave of research [87], [88]. This wave is due to the fact that optical communication technology is nearing the Shannon limit, a fundamental maximum capacity for a communication channel [89]. This issue is commonly called the “capacity crunch” [90]. There is no physical way to surpass this limit unless an additional degree of freedom is used. Unfortunately, modern optical communication technologies (in fiber) already make substantial use of all the available degrees of freedom, except one: space. The underutilized space degree of freedom, harnessed via optical modes such as OAM in a very compact way, enables scientists and engineers to surpass the channel capacity limits of existing technology with Tbps data rates having been demonstrated [91]–[97]. As with any new technology, there are many challenges, but with time these challenges are being solved. An obvious disadvantage of higher-order modes is more rapid divergence compared to the fundamental Gaussian [98]. This section provides a brief overview of the use of structured light in FSO.

OAM modes have twisted phase-fronts in the form of $\exp(i\ell\phi)$ where $\ell$ is an unbounded integer known as the topological charge, and $\phi$ is the azimuth in cylindrical coordinates [99]. OAM modes, as well as other orthogonal mode sets, such as Laguerre-Gauss (LG) (where the OAM modes are a subset), Hermite-Gauss (HG), and Ince-Gauss (IG), are all candidates for MDM [100]–[102]. Examples of these modes are shown in Fig. 11. When polarization is considered, modes with a uniform polarization are called scalar modes, and vector modes are those with a spatially varying polarization [103].

Modal bases other than OAM are less well studied in FSO but show promise under different conditions and applications. The resilience of spatial modes in turbulence is of primary concern: spatial modes are very sensitive to aberrations, which cause crosstalk and mode-dependent loss (fading of a specific mode). While adaptive optics are typically an expensive solution, higher-order spatial modes generally benefit significantly from their use [104]. Vector modes have been put forward for some time for FSO propagation [103], but studies both for and against their resilience to turbulence have been published [105], [106].

Information theoretical studies for channel capacities when higher-order modes are used indicate that they are not a silver bullet to solve the capacity crunch [107]–[110], however, this is not irrefutable. Recent research indicates that there appears to be a basis dependence to the effects of turbulence on spatial modes [111]. This is supported by several studies into the resilience of different spatial mode bases in turbulence through multiplexing as well as modal diversity [71], [72], [112]–[116].

In addition to inter-modal crosstalk, there have been relatively few studies into accurate fading probability distribution functions for higher-order modes, but experiments indicate that they are quite different from those of plane or spherical waves [87], [117]–[119]. This knowledge gap must be filled before we can effectively engineer FSO links that make use of higher-order modes.

Fig. 10. A simple diagram for an adaptive optics system for FSO links.
E. Machine Learning Algorithms

Machine Learning (ML) algorithms have recently been successfully applied in various optical communication systems [120]. In the context of FSO communication, a wide range of ML algorithms have been proposed as a way to detect structured light beams after propagating through turbulent atmosphere [121]–[123]. Those algorithms allowed the identification of the structured light intensity profiles without the need to use mode sorters or mode demultiplexing devices [88]. ML algorithms have also been proposed to compensate for turbulence effects at the beam level without the need for AO kits [124], [125]. Other applications of ML algorithms in FSO involve channel estimation over a wide range of atmospheric turbulence, as demonstrated in [126]. Other applications of ML that have been demonstrated in optical fiber systems such as modulation format identification [120], will equally find potential uses in FSO.

V. FSO Systems

A. Relay-assisted transmission

The overall range and resilience of an FSO system can be improved by using multiple relays to link two communicating terminals. There are two possible relay configurations for a relay-assisted transmission; serial and parallel [127]. A schematic illustrating the two relay configurations connecting a source “S” and a destination “D” is depicted in Fig. 12. With analogy to wireless communication, the serial configuration is known as “multihop,” and the parallel configuration is known as “cooperative diversity”. Note that within a single branch in the cooperative diversity configuration, it is possible to have multiple hop relay nodes.

At the relay node, there are two possible relaying schemes: Decode and Forward (DF) and Amplify and Forward (AF). A relay with a DF scheme decodes the received optical signal and regenerates a new signal and forwards it to the next node. The AF relay amplifies the received signal and forwards it to the next node. DF is preferred over AF in terms of system performance. However, AF is less complex to implement. For AF relays, the amplification is performed in the electrical domain, i.e., at the relay node, the received optical signal is converted to an electrical signal by a photo-detector and amplified by an electrical amplifier before being converted to an optical signal and forward it to the next node. Signal

Fig. 11. Examples of LG_{p=2, p=2, m=6} and HG_{n=2, m=2} modes respectively, showing intensity (top) and phase (bottom). The phases are 0 to 2\pi (blue to yellow).
amplification can also be performed optically without any electrical-to-optical and optical-to-electrical conversions at the relay node [128].

It is worth mentioning that the all-optical amplifying approach is not available for all the wavelength windows. However, this technique can take advantage of the progress in amplification in optical fiber communication in particular in using erbium-doped fiber amplifiers (EDFAs).

Several studies have investigated the performance of relay-assisted systems employing DF and AF relays over various turbulence conditions. The error performance analysis of a multihop FSO transmission with DF relays reported in [129], showed that using DF relay decreases the mean error rate compare to a single hop for an FSO channel subject to attenuation.

Darsikas et al. studied the end-to-end performance of multihop FSO systems employing EDFA-based AF relays over turbulent gamma-gamma modeled channels [130]. By studying the performance of a multihop transmission subject to atmospheric turbulence, Bayaki et al. showed that besides being simpler to implement up to a certain number of hops, all-optical relays outperform AF relays with electrical amplifiers [128]. All-optical amplification was also demonstrated experimentally in a 10 Gbit/s transmission over a turbulent channel [131].

B. Hybrid RF/FSO

A possible scenario for the deployment of FSO is when it is combined with RF transmission systems to fill the connectivity gap between the RF access network and the fiber backbone network [132], [133]. A schematic illustration of a dual-hop RF/FSO link is depicted in Fig. 13 (a). The data initially encoded on an RF channel is received by an antenna at a relay node. At the relay, the received signal is forwarded to the FSO transmitter. In an RF/FSO hybrid scheme, the data can be modulated using sub-carrier intensity modulation, as discussed in Sec. II [133], [134].

Another scenario of heterogeneous technologies is when an FSO link is co-installed with an RF link [135]. The RF link, often operating in the millimeter waves band, can serve as a backup when the FSO link is unavailable, although probably at a lower capacity. A block diagram of a hybrid RF/FSO system is illustrated in Fig. 13 (b).

The feedback from the receiver to the transmitter can be used to coordinate the link selection. The hybrid configuration is of interest if the RF and FSO links are subject to channel effects, especially when
both technologies are affected differently by weather conditions. In particular, FSO links are severely
degraded under the effect of heavy fog, but not to rain, while RF is mainly affected by rain and resilient
to fog [136].

The switching between the RF and the FSO links could be either hard-switching or soft-switching
depending on the system configuration [137]. In the hard-switching approach, the transmitter and receiver
jointly select the FSO or the RF link, which may result in loss of rate. Soft-switching is based on channel
coding techniques that allow transition between the two links.

A feedback-free architecture for hybrid RF/FSO systems was presented in [138]. The authors proposed
sending the same information over a millimeter-wave RF link and an FSO link, assuming that both systems
operate at the same data rate and employing the same modulation formats. Both signals are combined in
a symbol by symbol basis in the electrical domain. The main drawback of duplicating the data signals in
the millimeter-wave and FSO links is the bandwidth restriction on the FSO link imposed by the limitation
on the millimeter-wave system.

Note that commercially available solutions for co-installed RF/FSO links already exist [139].

C. Satellite communication

A comparison between RF microwave and optical communication system for space applications showed
that optical links are best suited for near-Earth applications, including satellite-to-satellite communication
[140]. SpaceX Starlink and Telesat companies aim to establish FSO satellite crosslinks in their satellite
cstellations installed to provide broadband internet access all over the globe [141], [142]. Many earlier
successful inter-satellite optical links have been reported [143], [144]. An inter-satellite demonstration
has reported an error-free 5.6 Gbps transmission between two Low Earth Orbit (LEO) satellites (NFIRE
and TerraSar-X) over several seconds [144]. An illustration of optical satellite crosslinks is depicted in
Fig. 14(a). Pointing errors arising from satellite vibrations are one of the main issues for inter-satellite
optical communication, which requires implementing advanced acquisition and tracking mechanisms to
ensure transmission reliability [145].
Ground-to-satellite (uplink) and satellite-to-ground (downlink) FSO links have also been demonstrated [146]-[148]. The first uplink and downlink demonstrations date back to 1994 by the National Institute of Information and Communications Technology (NICT) (formerly known as the Communications Research Laboratory) with a two-way 1 Mbps communication between the Japanese Engineering Test Satellite-VI (ETS-VI) and a ground station [146]. In 2013, NASA’s Lunar Laser Communication Demonstration (LLCD) reported a series of full-duplex communications between a satellite in lunar orbit (400,000 km from the Earth) and multiple ground stations (in Spain and USA) with a maximum uplink throughput of 20 Mbps and downlink throughput of 622 Mbps [147]. In 2016, the SOTA (Small Optical TrAnspender) system, developed by NICT, was used to establish an optical communication link between a LEO satellite (SOCRATES) to ground connection with a ground station affiliated to the German aerospace center [148]. NASA, in collaboration with MIT Lincoln Laboratory, is currently developing a 200 Gbps optical communication system to be installed on an LEO Cubesat (small satellite) to deliver more than 50 Terabytes of information per day to a ground station, as apart of the TeraByte InfraRed Delivery (TBIRD) program [149]. Beyond the vast offered bandwidth, optical satellite links require small footprints devices that consume less energy compared to those operating in the microwave bands typically used for Earth-satellite communications.

In addition to pointing errors, ground-to-satellite (as well as satellite-to-ground) are mainly subject to weather conditions and turbulence in the lower layers of the atmosphere. A potential solution of this issue would be to rely on high-altitude platforms that could be connected with ground stations using microwave signals and with satellites with optical signals to minimize the impact of turbulence that could significantly affect the optical signals as depicted in Fig. 14(b). For a detailed description of the challenges and the mitigation techniques of FSO inter-satellite and ground-satellite as well as satellite-to-ground, we refer the reader to [145].

D. Toward quantum FSO communication

The field of quantum FSO communication is growing rapidly and seems to be promising, particularly for quantum key distribution (QKD) applications, aimed to significantly improve the traditional cryptography methods. Here we provide a brief summary of some of the seminal works in this field to guide the interested reader.

Seminal experiments on outdoor QKD FSO were reported in the 90s and early 2000s [150]-[152]. In 2018, Pang et al. reported a 10 Mbps quantum transmission on a laboratory test-bench (for a distance of 1 m) [153]. It is also worth mentioning that satellite-to-Earth quantum FSO has been a topic of interest that has known major achievements in the last three years [154]-[156]. A team from NICT demonstrated a quantum communication between the SOCRATES LEO satellite and a ground station [154]. Liao et al. demonstrated a free space satellite-to-ground QKD over a distance of 1200 km (from Micius satellite to Xinglong ground station near Beijing in China) [155]. In [156], Günthner and his co-workers reported the first measurement of quantum signals sent from a satellite in the geostationary Earth orbit, propagated over a total distance of 38,600 km to reach a ground station. The primary motivation of these demonstrations is to create a global QKD network to increase security further and move toward the “quantum Internet”. Many other ongoing projects, by different international institutions further, aim to launch their customized CubeSats for quantum communication [157].

Deploying quantum FSO systems needs addressing several practical challenges. In [158], photons in a single mode fiber were entangled with higher-order modes in free space. This may enable existing fiber infrastructure to be used effectively for QKD, in a hybrid system with FSO from satellite or long-range ground links. Naturally, the problem of turbulence arises, and turbulence damages quantum states. Fortunately, it has been shown that the effect of turbulence on quantum states is similar to the effect on classical vector modes (see Sec. IV-D), and so it stands to reason that classical “beacons” could somehow be used to improve the resilience of quantum communications through turbulence [159].
In 1880, before radio, FSO could be used to transmit audio hundreds of meters wirelessly. In modern times, radio is the wireless technology of choice for line-of-sight communication due to its long history of extensive research and development. A consequence of this is that the radio spectrum is now congested.

With the rise of 5G, for example, there is a requirement for high capacity, short-range links between base-stations. In many cities, the cost of installing fiber is prohibitive, and so point-to-point radio links are used instead. Millimeter-wave radio devices can sustain gigabit bandwidths over hundreds of meters, but with their proliferation, the resulting interference is likely to degrade their viability. An alternative technology is FSO communication in the sense that it can be “fiber without the fiber”.

While much of the developed world eagerly awaits high speed, low latency mobile connectivity, a significant percentage of the world’s population struggles to download a text-based email. This digital divide stems from the fact that it is too expensive to upgrade existing radio-based backhaul to provide...
affordable, reliable, and high-speed internet to “disconnected” communities. Optical fiber is ultimately
the best solution, but could FSO be used as an interim technology in its absence? By retrofitting existing
microwave high sites with FSO to create hybrid systems, in fair weather, the overall system capacity
could be increased dramatically, while maintaining a (slower) microwave backup.

FSO communication is a broad topic that draws from many different fields, such as astronomy,
photonics, and communications theory. A broad perspective on the various aspects of FSO communication
- both well understood and on the edge of current knowledge - is difficult to acquire and so this
comprehensive tutorial is an invaluable resource for a large community.

There are many outstanding challenges in FSO, in particular the problem of atmospheric turbulence
over long distances. With new research drawing from and adapting techniques developed for radio, the
advent of structured modes of light, and progress in photonics in general, high capacity and long-range
FSO is the next frontier in wireless communication.

ACKNOWLEDGEMENTS

KAUST Office of Sponsored Research. M.A.C. acknowledges NRF grant 121908.

REFERENCES

Jan 2020.
Kelly, C. Marquardt, S. Matlhanle, B. Ndagano, F. Petruccione, R. Slavik, F. Romanato, C. Rosales-Guzmán, F. S. Roux,
2020-04-19.
WDM free-space optical link at 1550 nm,” in Optical Fiber Communication Conference and the International Conference
[8] R. Lange, B. Smuty, B. Wandernoth, R. Czichy, and D. Giggenbach, “142 km, 5.625 Gbps free-space optical link based
6105, International Society for Optics and Photonics. SPIE, 2006, pp. 68 – 76.
transmission based on polarization modulation and direct detection.” IEEE Photonics Technology Letters, vol. 31, no. 10,
optics transmission using a Kramers-Kronig receiver,” in Optical Fiber Communication Conference (OFC) 2019. Optical
(32x40 Gbit/s) WDM transmission system for free space optical communications,” IEEE Journal on Selected Areas in
(16x100 Gbit/s) for next-generation convergent urban infrastructures.” Optical Engineering, vol. 52, no. 11, pp. 1 – 6,
2013.
1–12, 2017.


M. Chen, K. Dholakia, and M. Mazilu, “Is there an optimal basis to maximise optical information transfer?” *Scientific reports*, vol. 6, p. 22821, mar 2016.


