Research Article | Vol. 29, No. 15 / 19 July 2021 / Optics Express 22973
Optics EXPRESS

Convolutional autoencoder for exposure effects equalization and noise mitigation in optical camera communication

CRISTO JURADO-VERDU,* VICTOR GUERRA, VICENTE MATUS, JOSE RABADAN, AND RAFAEL PEREZ-JIMENEZ

[1] Institute for Technological Development and Innovation in Communications (IDeTIC), Universidad de Las Palmas de Gran Canaria (ULPGC), 35017 Las Palmas de Gran Canaria, Canary Islands, Spain
*cjurado@idetic.eu

Abstract: In rolling shutter-based optical camera communication (OCC), the camera's exposure time limits the achievable reception bandwidth. In long-exposure settings, the image sensor pixels average the incident received power, producing inter-symbol interference (ISI), which is perceived in the images as a spatial mixture of the symbol bands. Hence, the shortest possible exposure configuration should be selected to alleviate ISI. However, in these conditions, the camera produces dark images with impracticable light conditions for human or machine-supervised applications. In this paper, a novel convolutional autoencoder-based equalizer is proposed to alleviate exposure-related ISI and noise. Furthermore, unlike other systems that use artificial neural networks for equalization and decoding, the training procedure is conducted offline using synthetic images for which no prior information about the deployment scenario is used. Hence the training can be performed for a wide range of cameras and signal-to-noise ratio (SNR) conditions, using a vast number of samples, improving the network fitting and the system decoding robustness. The results obtained in the experimental validation record the highest ISI mitigation potential for Manchester encoded on-off keying signals. The system can mitigate the ISI produced by exposure time windows that are up to seven times longer than the transmission symbol duration, with bit error rates (BER) lower than 10^{-5} under optimal SNR conditions. Consequently, the reception bandwidth improves up to 14 times compared to non-equalized systems. In addition, under harsh SNRs conditions, the system achieves BERs below the forward error correction limit for 1dB and 5 dB while operating with exposure times that are 2 and 4 times greater than the symbol time, respectively.

1. Introduction

Optical camera communication (OCC) is considered an extension of visible light communication (VLC), which replaces high-bandwidth photodiodes with image sensors (IS) to detect time and spatial variations in light intensity for enabling data communications. This technology emerges from the interest of reusing conventional cameras embedded in an increasing number of end-user devices (such as mobile phones, laptops, vehicle dashcams) to capture intensity-modulated (IM) light signals from a wide range of lighting sources, ultimately paving the way for VLC to break the market's entry barriers imposed by utilizing specific hardware. Furthermore, it has recently been included in the IEEE 802.15.7 [1] standard, which reveals the interest in this technology.

However, the handicap of IS-based receivers is their relatively low bandwidth inherently limited by the camera's frame rate [2–4], which makes them ideal for applications with low data rates, such as internet of things (IoT) applications, smart farming, indoor location, advertising, or vehicle-to-vehicle (V2V) communications among others. This restriction affects differently depending on the camera's acquisition mechanism. In global shutter (GS) cameras, the whole IS is exposed simultaneously. Therefore, the light signal is sampled once per acquired frame.

#433053 https://doi.org/10.1364/OE.433053
Journal © 2021 Received 3 Jun 2021; revised 25 Jun 2021; accepted 25 Jun 2021; published 6 Jul 2021

Optics EXPRESS

Consequently, the data rate is upper-bounded by the highest frame rate, restraining the transmission frequency, f_{tx}. Hence, the switching speed of light sources is constrained, producing, in some cases, a noticeable flicker that must be mitigated to prevent discomfort and health-related issues in human users [5]. This flickering can be alleviate by using under-sampled modulation (USM) schemes, such as under-sampled frequency shift on-off keying (UFSOOK) or under-sampled phase shift OOK (UPSOOK) [6] at the expense of decreasing the data rate. On the other hand, rolling shutter (RS) cameras scan the image progressively row by row of pixels. Each row of pixels is activated sequentially, sampling the light source at different instants during the frame capture. This acquisition mechanism produces different illuminated bands for the transmitted symbols within the image [7]. In this case, the sampling period (significantly lower than GS cameras) coincides with the time that elapses between the activation of two consecutive rows [8], which is also limited, albeit indirectly, by the camera's frame rate. Another parameter that further restricts the signal bandwidth is the exposure time (row exposure time for RS cameras), the duration in which a pixel remains exposed to light. During this exposition, the pixel integrates light, acting as a low pass filter, producing significant inter-symbol interference (ISI). This ISI is perceived as a spatial mixture of the symbol bands within the image, and its effects begin to be relevant after the exposure time is longer than half the symbol time.

Therefore, from a communications perspective, the exposure must be as short as possible to prevent ISI on high-speed signals at the cost of reducing the received signal strength. This trade-off must be addressed in any OCC system design [7,9]. However, improving the receiver bandwidth by reducing the exposure time will eventually result in impracticable light conditions for either human or machine-supervised applications. As it can be seen in Fig. 1 short exposure times would produce dark images, in which objects cannot be acceptably recognized (Mandrill picture). This energy impairment due to the reduction of the integration window can be mitigated by increasing the camera's analog gain, which can significantly improve the signal-to-noise ratio (SNR), as concluded in previous works [10–12].

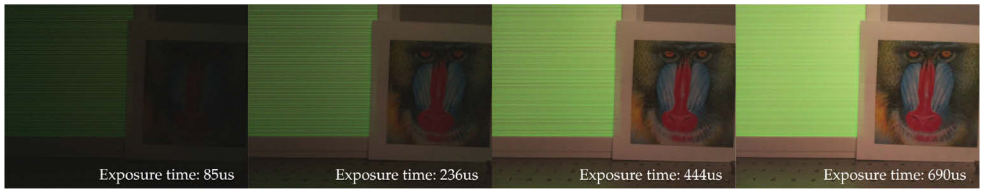


Fig. 1. Effects of increasing the exposure time in RS-cameras.

To alleviate this ISI effect, in [13] authors proposed a one-dimensional artificial neural network (ANN) equalizer with promising results. The neural network performance has been validated against Manchester encoded on-off Keying (OOK) [13], and constant power 4-PAM symbols [14]. In [13] this equalization allows to decode data with bit error rates (BER) below the forward error correction (FEC) limit of 3.8×10^{-3}, for exposure times up to 4 times greater than the symbol time. Translated in terms of bandwidth, it can recover signals whose bandwidth exceeds up to approximately nine times the low-pass filter's cutoff frequency that models the effect of the exposure time. However, the proposed multilayer perceptron (MLP) network performance was evaluated exclusively under optimal signal-to-noise ratio (SNR) levels. Furthermore, the network training was conducted online, with the receiver operating under the deployment scenario conditions. Up to the authors' knowledge, only these works try to mitigate the exposure-related ISI in RS cameras by using artificial neural networks.

However, artificial networks are acquiring a relevant role in solving other OCC technology challenges. In [15] the use of a Logistic Regression Machine Learning (LRML) algorithm is proposed for decoding signals transmitted with the backlight of advertising panels. In [16] a

Optics EXPRESS

1D-ANN architecture is proposed for the same purpose. Both works aim to decode signals affected exclusively by the interference produced by the frontal image content of the panel. Therefore, the camera is configured with short exposure times for the optimal reception of the transmitted symbols. In [17], a convolutional neural network (CNN), which combines convolutional layers with a fully connected classification network, is used for source detection and pattern recognition of LED-based sources in V2V communications. This network decodes spatially multiplexed streams under partial occlusion and/or harsh weather conditions. However, in this case, the RS acquisition mechanism is not exploited to increase the data rate, and hence the signal does not need prior equalization of the exposure-related effects. Instead, the transmitted symbol time is longer than the frame acquisition duration, so it can be considered that the system operates under GS conditions. In contrast, [18], uses a CNN for pattern detection and classification in V2V relying on the RS mechanism, recovering data from car rear taillights. Moreover, in [19], the use of CNNs is proposed for RS-symbol decoding. However, in previous works, the receiver operates on the premise that the exposure-related ISI is negligible since the cameras are configured with exposure times much shorter than the transmission symbol time.

This work proposes a novel two-dimensional convolutional autoencoder (CAE) for simultaneous exposure-related ISI equalization and noise mitigation, in which the training is conducted offline using synthetically generated images. These images are produced using exclusively time-related parameters from any chosen camera and transmitter: the configurable exposure time, the sampling period, and the transmission symbol time. These three parameters produce relevant training samples that, after prior standardization, enable the network to decode real captured images. This standardization makes the synthetic and real signals comparable from the point of view of the artificial network, regardless of the temporal average power received, as long as the camera's gamma transformation is precisely compensated. Hence, the training of the network becomes independent of the deployment scenario. These training samples can also be synthetically corrupted by a zero-mean additive white Gaussian noise (AWGN) to train the system for its operation under harsh SNR conditions. Therefore, the training can be carried using different synthetic noise levels and with a significant amount of samples stored in large databases, which considerably increases the robustness of the network.

Moreover, the use of CAE is justified since it has proven particularly useful in image denoising [20–22] presenting outcomes that outperform the capabilities of MLP architectures in this task [23], either in terms of efficiency and performance. This is, in part, because of the use of convolutional layers, but also because of the operation's nature of a CAE, which consists in extracting a latent representation or feature maps (generally of lower dimensionality) from the input (encoding part), and reconstructing it at the output using this representation (decoding part). In this encoding-decoding paradigm, noise-corrupted inputs might even be beneficial since they allow the network to deinterlace hidden useful features from the input. Furthermore, the use of two-dimensional inputs helps noise mitigation in RS acquired signals. Since the IS columns sample the light at the same sampling instants, the received signal is replicated across all the columns. This redundancy can be exploited to increase the SNR, thus facilitating the network task of noise mitigation.

The remainder of the paper is organized as follows. Section 2 introduces the RS theoretical modeling used for the synthetic image generation. Section 3 presents and describe the CAE-based proposed receiver system. Section 4 details, in the first place, the network training, including the synthetic image generation routines and the metrics used to assess the synthetic image similarity with real images and the network goodness of fit. In the second place, this section presents the procedures and metrics to evaluate the system's communication performance, including a detailed description of the experimental setup. Section 5 presents the results, to be compared in section 6 with those reported in recent literature. Finally, the conclusions of this work are summarized in section 7.

Optics EXPRESS

2. Rolling-shutter modelling

RS cameras scan the scene by progressively triggering each row of pixels rather than exposing the whole IS at once, as GS cameras do. Hence, light variations can be sampled up to H times per frame, where H is the IS's vertical resolution (height). This sampling produces different illuminated bands within the image that corresponds to the transmitted symbols. Figure 2(a) shows the temporal scheme of this acquisition mechanism. The IS starts exposing the first row of pixels to light during a configurable exposure time, t_{exp}. Then, the subsequent rows are activated sequentially every T_{s} seconds, which is the row sampling period. This acquisition can be modeled using a system with two separable stages: a filtering and a sampling stage, as shown in Fig. 2(c). The filtering part models the effects of the exposure time on the received signal for each pixel. The following stage manages the sampling instants for each pixel based on its position within the IS. Figures 2(b,d) illustrates how the incident light is converted into the final discrete values for each pixel (for a given column). Figure 2(b) represents the incident power reaching the j-th pixel over time, t, $P(t,j)$. On the one hand, Fig. 2(d) shows the pixel signals after the filtering stage, $v(t,j)$ (colored dashed lines). On the other hand, it depicts the one-dimensional discrete pixel values obtained after the sampling stage, $v[n]$ (black dotted vertical lines). It should be highlighted that, as it is shown in Fig. 2(b), the temporal evolution of the incident power follows the same shape for all the pixels. Nonetheless, the received signals are affected by different factor depending on several link parameters, such as the source's radiation pattern, the relative configuration between the transmitter and the camera, the camera lenses, the channel losses, and the scenario's reflections, among others. In this figure example (Fig. 2(b)), the average incident power gradually decreases from top to bottom. Moreover, the signal power is not enough to provide a suitable SNR for decoding in some cases. For this reason, in OCC, the light source projection within the image is generally considered as the ROI because it corresponds to the image area where the SNR is significantly higher. However, data can also be recovered from reflections as examined in [24–26]. Finally, it is important to mention that the following modeling is presented for a generic IS column. Therefore if the incident power for the pixel located at the i-th column and the j-th row is expressed with $\mathbf{P}(t,i,j)$, $P(t,j)$ satisfies the relation $P(t,j) = \mathbf{P}(t,m,j) = P^{(m)}(t,j)$, where $m \in [0, W)$ (the selected m column), and W is the IS's horizontal resolution (width).

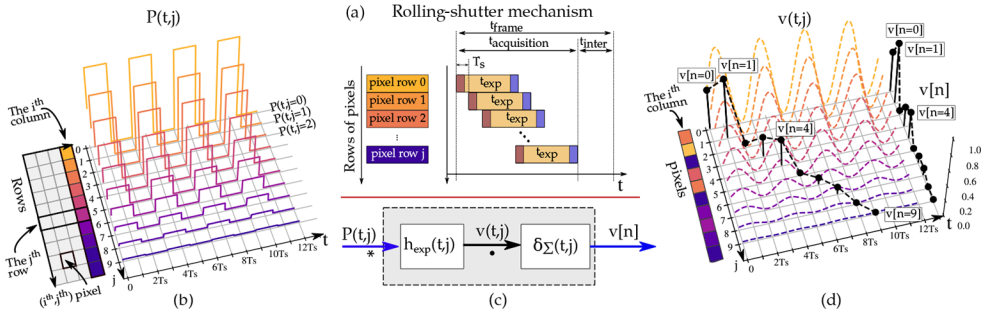


Fig. 2. RS acquisition mechanism. **(a)** RS Temporal scheme. **(b)** Normalized optical power $P(t,j)$ reaching each pixel of the i-th IS column. **(c)** RS system modeling (impulse response, h_{exp} and sampling function, δ_{Σ}). **(d)** Filtered curves, $v(t,j)$, and the discrete signal, $v[n]$ after sampling.

2.1. Filtering stage

The output value for the j-th pixel in the i-th column, $v(t,j)$, depends on the accumulated charge on the photodiode during the time it is exposed to light, the exposure time, t_{exp} (direct integration) [4]. This time extends from reset, in which the pixel's photodiode is biased with reverse voltage,

Optics EXPRESS

until its readout. During readout, in passive pixel sensors (PSS), the charge is transferred to a floating diffusion amplifier (FDA) (shared for all the IS columns), where the charge is converted to voltage with a conversion gain of G_{conv}. Finally, this voltage is amplified at the general output amplifier and quantized by the analog-to-digital converter (ADC). In active pixel sensors (APS), the charge-to-voltage conversion occurs at the pixel level, and the voltage is transferred to the output amplifier using source follower amplifiers. In short, the output voltage of the pixel at the ADC input is given by Eq. (1).

$$v(t,j) = \frac{G}{C_{\text{f}}} \int_{t}^{\text{t+t}_{\text{exp}}} P(t,j) \cdot \Re(\text{j}) dt \tag{1}$$

where $P(t,j)$ is the incident optical power, $\Re(j)$ and C_{f}, the equivalent photodiode's responsivity and capacity, respectively. The latter is approximately equal to the FDA's capacitor, and G the output amplifier's gain. This windowed integration of the input signal over the exposure time, can be modeled with a finite impulse response (FIR) low pass filter, h_{exp} given by Eq. (2), with its corresponding transfer function (Eq. (3)) [27].

$$v(t,j) = P(t,j) \circledast h_{\text{exp}}(t,j) \quad \text{where:} \quad h_{\text{exp}}(t,j) = h(t) = \frac{G}{C_{\text{f}}} \cdot \Big(u(t+t_{\text{exp}}) - u(t)\Big) \tag{2}$$

$$H(w) = \mathcal{F}\{h(t)\} = t_{\text{exp}} \frac{G}{C_{\text{f}}} \frac{sin(w \cdot t_{\text{exp}}/2)}{w \cdot t_{\text{exp}}/2} e^{jwt_{\text{exp}}} \tag{3}$$

where $u(t)$ is the unit step function. From Eq. (3) it follows that the filter DC gain is proportional to the exposure time. Regarding the available reception bandwidth, to compute the cutoff frequency, it is necessary to rely on numerical methods such as Newton-Raphson's algorithm. However, to get an approximate idea of how the reception bandwidth is related to the exposure time, the first null frequency can be examined, which is inversely proportional to the exposure time. Therefore, a trade-off between the gain and the available bandwidth must be considered for the configuration of the camera's exposure settings. Light signals captured with shorter exposure times are affected by lower ISI, but also the received power decreases, as shown in Fig. 1. In those cases, it is still possible to improve the received signal quality by increasing the analog gain G [10,11].

2.2. Sampling stage

The family of curves obtained after the filtering stage, $v(t,j)$, shown in Fig. 2(d), is ideally sampled using a two-dimensional Dirac delta train function, $\delta_{\sum}$ (Eq. (4)), generating a one-dimensional discrete signal, $v[n]$.

$$\begin{aligned} v[n] &= v(t,j) \cdot \delta_{\sum}(t,j) \qquad \text{where:} \\ \delta_{\sum}(t,j) &= \sum_{n=0}^{\infty} \delta\Big(t - \Big\lfloor \frac{n}{H} \Big\rfloor \cdot t_{\text{Frame}} - \text{mod}(n,H) \cdot T_{\text{s}}, \text{mod}(n,H)\Big) \end{aligned} \tag{4}$$

where $\lfloor \cdot \rfloor$ is the floor function, $\text{mod}(a,b)$, the modulo operation that returns the remainder of the division a/b. The floor division ($\lfloor \frac{n}{H} \rfloor$) returns the number of generated frames from the start. The modulo division ($\text{mod}(n,H)$) returns the pixel index (j-th) that contributes to the n-th sample of the discrete signal $v[n]$. This equation can be further simplified in Eq. (5) under the assumption that t_{inter} is zero and the scanning operation is continuous. In other words, there are

Optics EXPRESS

no periods in which the sensor becomes blind to transmission [7].

$$\delta_{\sum}(t,j) = \sum_{n=0}^{\infty} \delta\Big(t - n \cdot T_s, \mathrm{mod}(n,H)\Big) \tag{5}$$

Furthermore, the Eq. (6) introduced in [27] can be derived from Eq. (5) under the condition that the IS pixels are affected by the same signal power. Therefore, the discrete signal, $v_{eq}[n]$ (Eq. (6)) can be interpreted as the signal that would be acquired from a single equivalent pixel.

$$v_{eq}[n] = v(t) \cdot \sum_{i=0}^{\infty} \delta(t - n \cdot T_s) \tag{6}$$

The equations Eq. (4) and Eq. (5) indicate that each sample, $v[n]$, depends on the signal for the j-th pixel activated at the sampling instant and, consequently, the pixel position within the IS. Therefore, the sampling function relates the evolution of the signal over time with different image locations. In other words, this function express mathematically the space-time duality of OCC systems.

Finally, following the ideal sampling theory, the number of pixels (samples) per transmitted symbol (using the nomenclature introduced in [14]), N_{pps}, can be computed knowing the symbol time, t_{sym} and the sampling period, T_s ($N\mathrm{pps} = t_{sym}/T_s$).

3. Communications scheme

The proposed system architecture, and the functional blocks, are shown in Fig. 3. Regarding the transmitter, it emits non-return to zero (NRZ) Manchester encoded pulses to avoid flickering. Pseudo-random data sequences are grouped into packets with a header consisting of five consecutive ones and a zero-bit trailer. A redundant bit is inserted every three bits to prevent a header sequence from appearing within the payload. This stuffed bit is set to one if the preceding bit is zero and zero otherwise. This coding strategy eliminates the use of forbidden codes for synchronization, reducing the system's complexity and easing error detection at reception. The symbol time, t_{sym} is selected according to the camera's row sampling period T_s, to generate the desired N_{pps} ($N_{pps} = 5$). The transmitting source consists of a 20x20cm RGB LED flat panel that uniformly distributes the light across its surface. The operating link distance will depend exclusively on the lamp's size in the image and not on the optical emitted power (as long as the projection of the lamp occupies more than one pixel). As detailed in [28] the use of image-forming optics compensates the power losses due to spherical propagation with the projected size of the optical source on the IS. The receiving side consists of a RS-camera attached to a computing unit that performs the following routines for data acquisition as shown in Fig. 3. First, it selects M columns from the image's central region where the source is expected to be located (M equals 16). Next, it performs an equalization procedure to adjust the pixel values along the vertical dimension. In this work, no prior equalization is conducted. Then, the ROI is segmented into s overlapping windows according to the CAE's input dimensions (256x16 pixels) (with s equals 6). These image segments are standardized using the z-score function, which subtracts the image's mean μ to each sample x_i and divides the difference by the image's standard deviation, σ. This standardization is essential since it allows the images captured with the camera to be comparable with the synthetic training images, as will be discussed in section 4.1. The outputs generated by the CAE (with the exact dimensions of the input) for each segment are merged using a linear combination at the edges where overlap occurs. This linear merging helps mitigate the slight edge imperfections that appear near the top and the bottom of the output images. The size of this overlap depends on the number of selected segments. Increasing the number of segments will lightly improve the system's performance, but it will increase the

Optics EXPRESS

computational load. Experimentally, it has been concluded that the imperfections affect a small area with 10 pixels height under the worst conditions (long exposure settings). Hence splitting a 1080-pixel image into six segments, generating 38-pixel overlaps for 256-pixel windows, is a reasonably conservative solution. Finally, the packet synchronization within the reconstructed image is conducted using the Pearson correlation with a header searching template.

The fundamental element of the proposed system is the CAE that performs both the equalization and denoising of the ROI. An autoencoder is a neural network that attempts to reconstruct the original input using a lower-dimensional latent representation [29]. It consists of a trained encoding network (encoder) that extracts relevant features from the input whilst its counterpart (decoder) is tuned to reconstruct the original input from this representation through the minimization of a loss function and a back-propagation algorithm for updating the weights of the architecture. This process is mathematically described in Eq. (7).

$$\bar{\mathbf{x}} = \mathcal{D}\left(\mathcal{E}\left(\mathbf{x}\right)\right) \tag{7}$$

where $\mathbf{x}$ is the input signal, which can be multi-dimensional, $\bar{\mathbf{x}}$ is the autoencoded version of $\mathbf{x}$, $\mathcal{E}(\cdot)$ is the encoding operation, and finally $\mathcal{D}(\cdot)$ is the decoding procedure. In this work, the loss training function $L(\mathbf{x}, \bar{\mathbf{x}})$ is the L_2-norm (mean squared error) with a regularization term to prevent over-fitting (Eq. (8)).

$$L(\mathbf{x}, \bar{\mathbf{x}}) = E\left[(\mathbf{x} - \bar{\mathbf{x}})^2\right] + \lambda R\left(\mathcal{E}, \mathcal{D}\right) \tag{8}$$

where $E[\cdot]$ denotes expected value, λ is the regularization coefficient, and $R(\cdot)$ is the regularization function, which in this work corresponds to a combination of the $L1$ and the $L2$ weights regularization penalties. On the other hand, a denoising autoencoder (DAE) is a specific type of AE that exploits the presence of noise in inputs to de-interlace useful properties, eventually mitigating the noise corruption in the output. In this case, it minimizes Eq. (9).

$$L\left(\mathbf{x}, \mathcal{D}\left(\mathcal{E}\left(\tilde{\mathbf{x}}\right)\right)\right) \tag{9}$$

where $\tilde{\mathbf{x}}$ is a copy of $\mathbf{x}$ that has been corrupted, in this case, by an a zero-mean additive white Gaussian noise (AWGN). Finally, a CAE uses convolutional layers (*Conv*) and transposed convolutional layers (*TConv*) [23] to encode and decode the input, respectively.

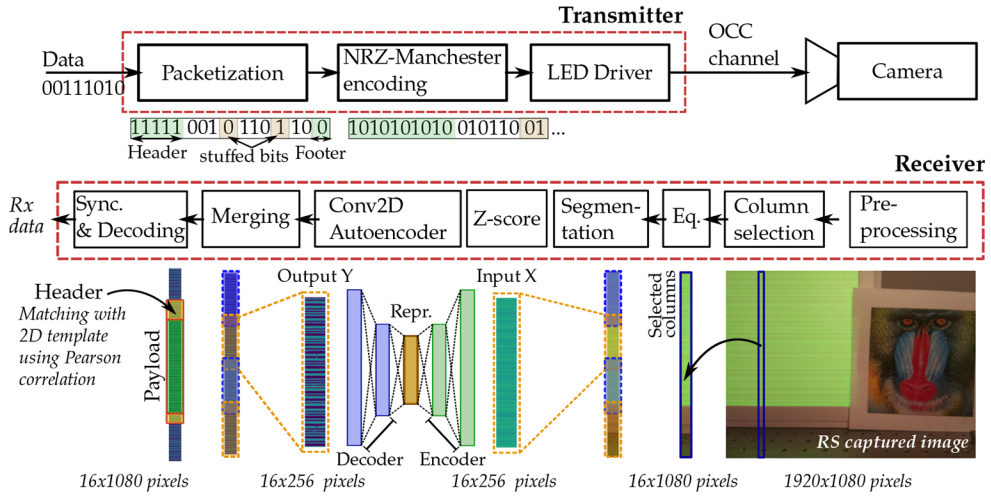


Fig. 3. Proposed system based on 2D convolutional autoencoder (CAE).

Optics EXPRESS

The *Conv* layers utilize a set of convolutional filters, which generates feature maps, F (one per filter), or activation maps by applying the discrete convolution operation. Considering the case in which the input consists of 2-dimensional gray-scale images, I, the discrete convolution uses two-dimensional kernels, K. The convolution result, $Z[i,j]$, at the i,j position for each filter is computed using Eq. (10).

$$Z[i,j] = (I \circledast K)[i,j] = \sum_{m=0}^{k_w-1} \sum_{n=0}^{k_h-1} I[m,n] \cdot K[i-m, j-n] \tag{10}$$

However, when working with RGB images, the input to the *Conv* layer consists of a three-dimensional tensor, where two dimensions are used for pixel position (width and height) and the last dimension for the three RGB color channels (depth). In this case, the convolution kernel is three-dimensional. Hence, generalizing for a number of D channels, the kernel's depth, k_d will match the number of channels of the input tensor, and the convolution result for each filter at the i,j position is computed using Eq. (11).

$$Z[i,j] = (I \circledast K)[i,j] = \sum_{m=0}^{k_w-1} \sum_{n=0}^{k_h-1} \sum_{l=0}^{k_d-1} I[m,n,l] \cdot K[i-m, j-n, l] \tag{11}$$

It should be highlighted that the filter translation over the image happens exclusively on the vertical and horizontal dimensions, summing up all the weighted contributions for all the channels to generate a two-dimensional tensor. The number of trainable weights per kernel will depend on its vertical and horizontal size and the input tensor channels. The result of the convolution, $Z[i,j]$, is then biased ($B[i,j]$) and transformed using a non-linear activation function, ψ, generating the corresponding features map, $F[i,j]$ (Eq. (12)).

$$F[i,j] = \psi(Z[i,j] + B[i,j]) \tag{12}$$

The nonlinear activation functions used in this work are the Sigmoid and the Rectified Linear Unit (ReLU) functions that work optimally in this type of architecture as demonstrated experimentally in [30].

The total trainable parameters of the l-th layer is the sum of the kernel's weights and biases considering all the filters. The latter coincides with the number of this layer's outputs, which can be computed knowing the horizontal and vertical dimensions of the output matrix, $O[i,j,l]$ using the Eq. (13).

$$dim(O[i,j,l]) = \left(\left\lfloor \frac{n_H + 2p_H - k_H}{s_H} \right\rfloor + 1, \left\lfloor \frac{n_W + 2p_W - k_W}{s_W} \right\rfloor + 1, D \right) \tag{13}$$

where n_H, n_W are the vertical and horizontal lengths of the input, p is the number of padding values added at boundaries (to control the output size), k_H, k_W are the vertical and horizontal lengths of the filter's kernel, and s the stride, the step translation of the kernel when traversing the input, and D, the number of filters.

In this CAE architecture, *Conv* layers are usually followed by a pooling layer, which replaces the layers' outputs in specific locations with a statistical summary of the outputs at the vicinity. In this model, max-pooling layers (*MaxPool*) are used, which return the maximum output of a rectangular group of outputs. This ultimately contributes to increasing the non-linearity of the output (in addition to the nonlinear activation functions) and reduces the total number of network parameters.

Conv layers have proven especially effective for extracting useful features from images, and they are widely used in object detection and classification as well as image segmentation and denoising. This convolution operation can help improve the efficiency of deep learning systems.

Optics EXPRESS

Furthermore, it allows reducing the number of network parameters by making better use of the spatial similarities in the vicinity of an input sample (sparse connectivity). In addition, in this type of network, the same filer kernel's weights would be applied across all the inputs (parameter sharing), tying the weights for different samples. This is contrary to what would happen in a dense network, in which each neuron assigns a specific weight for each input, and consequently, a separate set of parameters for every location is generated. Instead, in *Conv* layers, just a single set of parameters is learned (those concerning the filters). In this way, trained kernels would search for shared activation patterns across the image. *Conv* networks are thus dramatically more efficient than dense networks, reducing the total trainable parameters significantly.

On the other hand, the *TConv* layer reverses the spatial transformation produced by a *Conv* layer. Even though it is also (wrongfully) known as a deconvolutional layer, it does not perform the deconvolution operation. Instead, it carries out a regular convolution on an upsampled version of the input tensor to obtain an output tensor with the dimensions of the expected input of its reverse *Conv* layer. In this CAE architecture, these layers are usually followed by two-dimensional upsampling layers (*UpSampling*) that expand the input tensor by repeating samples at each position. This ensures that the output tensor has the exact dimensions of the input.

4. Methodology

This section starts by describing the CAE training procedures, including the generation of training synthetic image datasets. Next, the algorithm utilized for optimizing the network's hyperparameters and the selected search space is detailed. Finally, the communications performance evaluation procedure and metrics are introduced alongside the details of the experimental setup. This evaluation aims to demonstrate the ability of the system as a whole to equalize and decode overexposed signals with robustness in low to moderate SNR conditions, allowing simultaneous data acquisition and image visualization. Furthermore, that the network can, once it has been trained for a specific exposure time, adapt to slightly longer or shorter exposures. Finally, it aims to validate the use of synthetic (scenario-independent) images for network training.

4.1. Network training

The supervised training of the CAE is performed offline only once, using synthetically generated images. The four parameters considered for the generation are the exposure time, the sampling period, the symbol time, and the SNR. The procedure for generating a synthetic image, shown in Fig. 4, is described below.

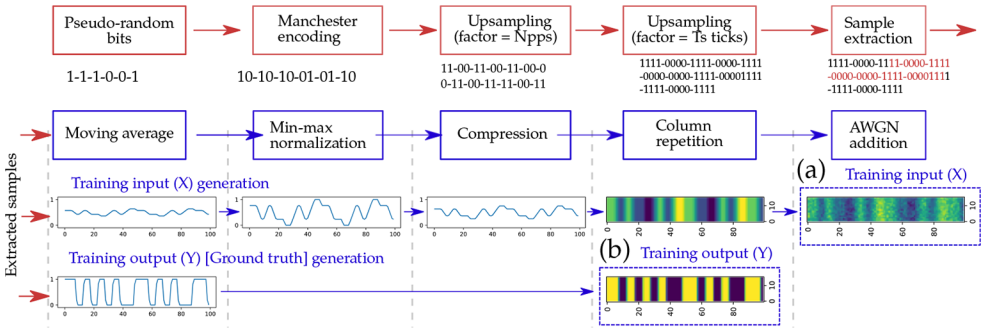


Fig. 4. Synthetic image generation.(a) Training input (X). (b) Training output (Y).

First, pseudo-random, one-dimensional bit sequences are encoded using the NRZ-Manchester line code. Then, they are upsampled by two factors: the number of row pixel samples (N_{pps}) expected per symbol and the number of ticks per sampling period T_s. The number of ticks depends

Optics EXPRESS

on the selected resolution time. In this work, the selected time equals the clock period, which is approximately 100 nanoseconds. Hence since the sampling period is $18.9\mu s$, its corresponding number of ticks is $T_s^{\text{ticks}} = 189$. Then, a section of the signal is extracted following a random starting offset to simulate non-perfect synchronization between the transmitter and the camera. This offset varies uniformly between zero (perfect synchronization) and one symbol time. Next, a moving average window is applied to model the effect of the camera's exposure time. The length of this window would depend on the requested exposure time. Particularly is equal to the number of ticks of the exposure time (for $t_{\text{exp}} = 444\mu s$, $t_{\text{exp}}^{\text{ticks}} = 4440$). The output is then normalized using the min-max normalization, resulting in a signal with values between zero and one. Then, it is compressed with a constant factor (0.5) to prevent clipping effects after adding the noise. The obtained one-dimensional signal is repeated along the horizontal dimension to generate a two-dimensional image. The resulting image has the dimensions of the CAE's input layer (256 rows and 16 columns). Finally, a zero-mean additive white Gaussian noise (AWGN) is added.

The training dataset collects sets of two synthetic images for a given random binary sequence: the input (X) and output (Y) images. The input image is made using the selected training exposure time. Figure 4(a) shows the procedure for generating the input image. The output image, that represents the ground truth, is generated similarly but selecting the shortest possible exposure time according to the time resolution (this time must be at least shorter than half of the symbol time). In this procedure, the min-max normalization, the compression, and the noise addition routines are discarded. Figure 4(b) shows the generation of the output image. The datasets generated in this work contain 35500 sets per exposure time (71000 images). From these datasets, 10% of the images are reserved for validation, while the remaining 90% are used for training.

Regarding the network training, the standardization of the input images is important. The z-score standardization applied to both the training and the real images makes them comparable from the point of view of the CAE. This eliminates the necessity to consider the expected average received power (and some camera parameters such as the analog and the digital gains) for the generation of the synthetic images. Therefore, the training is independent of the deployment scenario. However, it is mandatory to perform a prior compensation of the spatial power differences and the camera's gamma transformation. The similarity between synthetic and real images is measured using the Pearson's correlation coefficient. The training goodness-of-fit is quantified with the mean square error cost obtained for the training and the validation datasets.

Finally, regarding the network model, two different topologies are used in this work. The first topology has two stages composed of one *Conv* and one pool layer for the encoder part, and one *UpSampling* and *TConv* layer for the decoder part. The second topology adds another stage (with a total of three stages). For both topologies, efficient optimization of the network hyperparameters is conducted, following the hyperband algorithm detailed in [31] using the training exposure time of 444 μs. The considered parameters that constitute the search space are summarized in Table 1. The best architecture for each topology is used later for the system evaluation.

Table 1. Hyperparameter's search space

Hyperparameter	Value
Number of stages (*Conv*/*MaxPool* or *UpSampling*/*TConv*)	2, 3
Number of *Conv* (or *TConv*) filters	8, 16, 24, 32, 40, 48, 56, 64
Conv (or *TConv*) kernel's width and/or height	2, 3, 4, 5
MaxPool (or *UpSampling*) kernel's width and/or height	1, 2, 4
Learning rate	0.01, 0.001

Optics EXPRESS

4.2. Performance evaluation

Regarding communications, the CAE's performance is evaluated using real images captured with a conventional camera. The evaluation under different SNRs is performed by adjusting the light transmitted power through the control of the voltage source. Images taken under these conditions are used to estimate the SNR. It should be highlighted that since the received signal is affected by the camera's exposure time, it is not feasible to estimate the SNR by analyzing the image mean and variance. Hence, the following procedure is used. First, the pixel rows are averaged across all the image columns (1920 columns). This averaging increases the SNR by a factor of $N = 1920$ [14] (assuming that images are corrupted with AWGN). Next, the obtained averaged signal, $\bar{s}$, affected by a significantly low noise power, is subtracted from the signal at the desired decoding column, s, resulting in a noise signal, n. Finally, the signal power, S, is estimated using the maximum value of the autocorrelation of s (the same procedure is used for estimating the noise power N).

The selected metric to evaluate the communications' performance is the BER. In addition, to quantify the degree to which the signal is affected by exposure-related ISI, a new metric is introduced, the exposure-to-symbol ratio (ESR), the ratio between the cameras' exposure time, and the symbol time. For example, an ESR of 7 indicates that the exposure time exceeds seven times the symbol duration.

Figure 5 depicts the experimental setup used to capture the real images. It consists of an RGB flat panel pointing towards an RS-Camera separated by a distance of 50 cm. At this distance, the transmitter occupies approximately 3/4 of the image's vertical size. The transmitter signal is generated using an arbitrary signal generator and a power supply to control the voltage level of the light source.

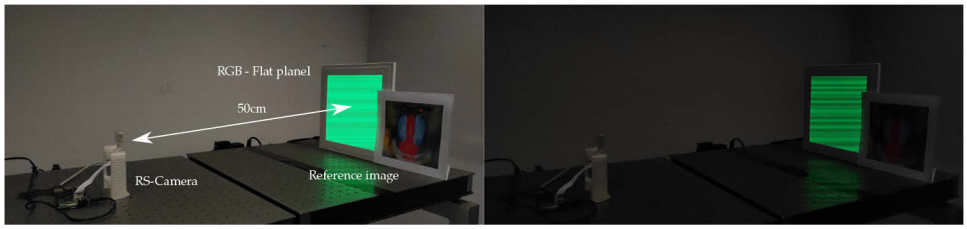


Fig. 5. Experimental setup

Table 2 summarizes the key parameters of the experiment setup.

Optics EXPRESS

Table 2. Experiment's key parameters

Parameter	Value
Camera	
Hardware	PiCamera version 2 (Sony IMX586) [32]
Aperture lens | Focal length (equivalent)	f/2 | 3 mm
Image resolution	1920x1080 pixels (Video mode - 3)
Clock time, t_{clk}	10 MHz
Sampling period, T_s	18.904 μs (Measured)
Exposure times, t_{exp}	from 85 μs to 1500 μs in steps of 19 μs
Transmitter	
Color channel used	Green channel
Voltage	from 25V to 36V in steps of 0,5V
Symbol time, t_{sym} (N_{pps})	94, 5 μs (5 pixels per symbol)
Packet's header, payload and trailer lengths	5, 42, 1 bits
Random seed	31415

5. Results

This section presents the results obtained for the training and the generation of synthetic images and the communications performance.

5.1. Network training

Figure 6 shows synthetic and real images for different exposure times to provide a visual comparison between them. Despite being placed horizontally, each segment corresponds to a vertical rectangle extracted from the image. The 24 examples are arranged into four groups based on the selected exposure time (161, 312, 444, or 520 μs). In each group, the reference template corresponds to the ground truth, the expected theoretical signal if the exposure time were infinitely short, and the incident power for all pixels, the same. The filtered template is generated for each exposure time using the reference template and normalized with the min-max function. This template is used to quantify the degree of similarity between the synthetic and real images. The following two examples correspond to the synthetic and real images without preprocessing as captured by the camera. As it can be seen, in those images, it is hard to distinguish light variations. This occurs because as the exposure time increases the dynamic range for the pixel values decays abruptly, reducing the ratio between the largest and smallest values. The last two images correspond to the standardized version (z-score) of synthetic and real images used as inputs for the network. As it was aforementioned in section 4, under the right conditions, both standardized images are comparable. Differences in the averaged received light in real images will minimally affect the standardization. Nevertheless, the camera's gamma transformation must be precisely compensated.

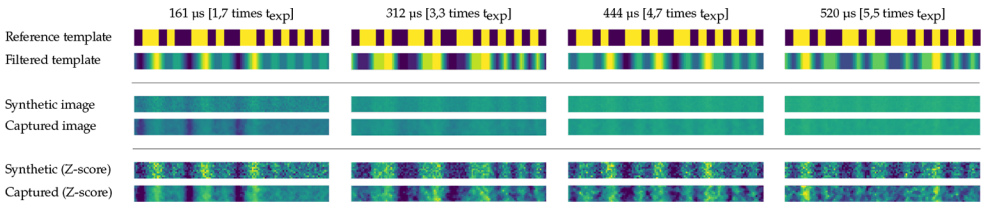
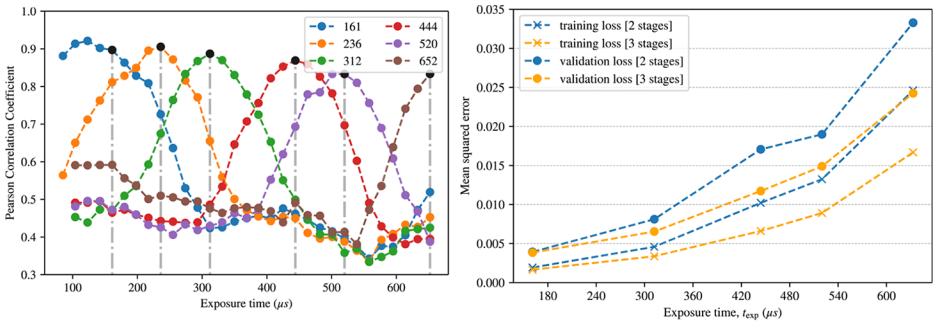


Fig. 6. Visual comparison of synthetic generated images versus real captured images

Optics EXPRESS

The degree of similarity between the synthetic and real samples is quantified by the maximum Pearson's correlation coefficient between the filtered template and the real images. Figure 7(a) presents the results obtained after matching the synthetic templates (shown in the legend) with images taken with a wide range of exposures. It should be mentioned that since unavoidable noise in the real images, the Pearson correlation coefficient does not reach its maximum value (1.0). However, it exceeds 0.85 in all cases in which both the synthetic and real exposures times are the same. As shown in this graph, as the exposure time slightly increases or decreases, the correlation coefficient rapidly decays to values around 0.4. The apparent symmetry of these curves reveals that non-similarities in the vicinity have a similar impact on the correlation. This could imply that the CAE could face approximately similar challenges when equalizing longer and shorter exposure times than the selected for the training.



(a) (b)

Fig. 7. (a) Pearson correlation coefficient between the filtered templates and the real images. (b) Training and validation losses for different training exposure times.

Table 3 summarizes the best model's parameters for both topologies as described in section 4. It details the number and type of layers (with their corresponding activation functions), the number of filters and kernel sizes, and, finally, the shape of the outputs and the total trainable parameters for each layer. In all cases, the learning rate that performed best was 0.001.

Figure 7(b) shows the training and validation losses for both topologies for different training exposure times. The use of a three-stage topology reduces losses for higher training exposures by an improvement factor of approximately 25%.

5.2. *Performance evaluation*

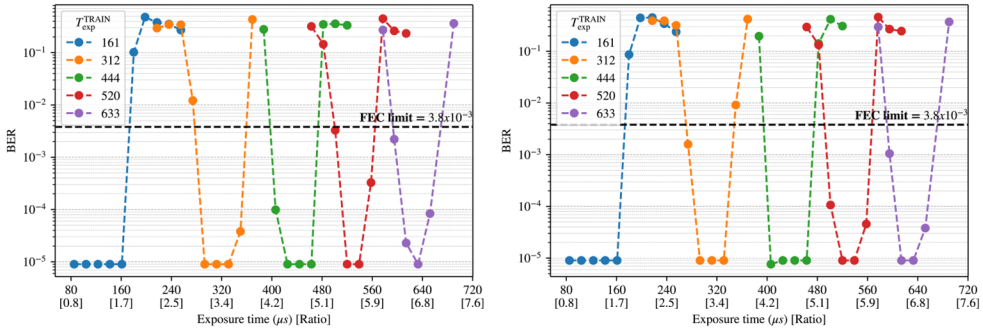
Figure 8 presents the BER results obtained after training the CAE with different exposure times of 161, 312, 444, 520, 633 μs. The x-axis shows the exposure time set by the camera (in microseconds) when taking images. In addition, the ESR (section 4.2), the ratio between the exposure time and the symbol time (94, 5 μs) is shown between brackets below its corresponding exposure. From these results, it can be extracted that the CAE can easily adapt to exposure times that are up to 10% higher or lower than the selected one for training, maintaining the BER below 10^{-5}. However, after this margin, the BER increases steeply. An exceptional case occurs when the training time is almost twice the symbol time ($T_{\text{exp}}^{\text{TRAIN}} = 161 \approx 1.8 \cdot 94.5\mu$s), as it can be seen in Fig. 8. In this particular case, the system can decode signals for all the exposures that are lower than the selected for training, maintaining the BER below 10^{-5}. The graph shown in Fig. 7(a) helps to explain this result. In this graph, it is observed that the correlation of the training template (161 μs) with images exposed with shorter exposure times remains relatively high. Hence, the neural network can extract from the training images a set of features that differ minimally from those of the images affected by lower ISI. Finally, the comparison between both

Optics EXPRESS

Table 3. Optimized model's summaries for the training dataset of 444 μs.

	Two stages topology				Three stages topology			
Layer (activ.)	Kernel	Filters	Out.shape	Params.	Kernel	Filters	Out.shape	Params.
Encoder								
Conv (ReLU)	(4,5)	16	(256,16,16)	336	(5,5)	24	(256,16,24)	624
Maxpool	(2,2)	16	(128,8,16)	0	(2,2)	16	(128,8,24)	0
Conv (ReLU)	(5,3)	56	(128,8,56)	13496	(5,4)	48	(128,8,48)	23088
Maxpool	(2,2)	56	(64,4,56)	0	(1,2)	48	(128,4,48)	0
Conv (ReLU)	*Not applicable*				(4,3)	32	(128,4,32)	18464
Maxpool	*Not applicable*				(2,2)	56	(64,2,32)	0
Decoder								
TConv (ReLU)	(5,3)	56	(64,4,56)	47096	(5,3)	32	(64,2,32)	12320
Upsamling	(2,2)	56	(128,8,56)	0	(2,2)	32	(128,4,32)	0
TConv (ReLU)	(4,5)	16	(128,8,16)	17936	(4,5)	48	(128,4,48)	30768
Upsamling	(2,2)	56	(256,16,16)	0	(2,2)	48	(256,8,48)	0
TConv (ReLU)	*Not applicable*				(4,5)	24	(128,8,24)	28824
Upsamling	*Not applicable*				(2,2)	56	(256,16,24)	0
Conv (Sigm.)	(5,3)	1	(256,16,1)	321	(5,3)	1	(256,16,1)	601
	Total trainable parameters			79185	**Total trainable parameters**			114689

topologies agrees with the results shown in Fig. 7(b). The BER decreases less steeply for the three-stage topology, especially for the high exposures.



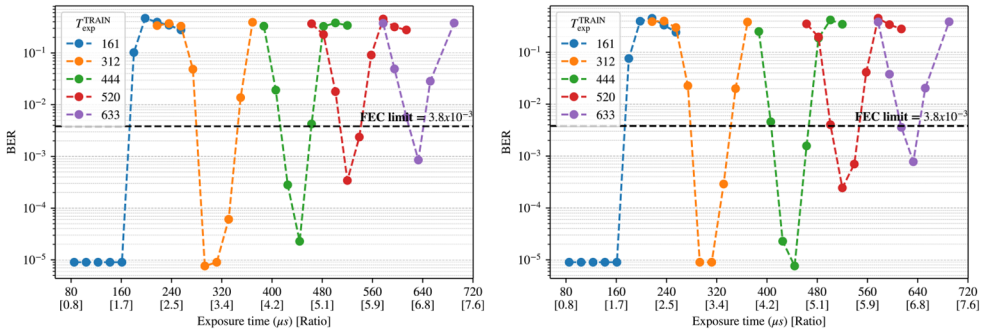
(a) Two stages (b) Three stages

Fig. 8. BER results for training exposures, T_{exp}^{TRAIN}, under high SNR conditions.

Figure 9 shows the results obtained under harsh SNR conditions (between 12 and 18 dB). In this case, using a 3-stage topology is justified for cases where the ESR is greater than 3. Under these conditions, it is possible to decode signals with BER below the FEC limit for ESR values up to 7. Fig. 10 shows the behavior of the system under different SNR conditions. As it can be seen, the BER decreases approximately linearly with the SNR (in dB), with a comparable slope for all cases (except for the blue curve). This slope is approximately 200 dB^{-1}. In addition, it indicates that as the training is conducted longer exposure times, the CAE cannot correctly minimize the error at the output, regardless of the noise level. The exceptional case occurs when the ESR is approximately equal to 2. In this case, the network has succeeded at deinterlacing the hidden features of the training signal from the added noise, enabling the system to achieve BERs

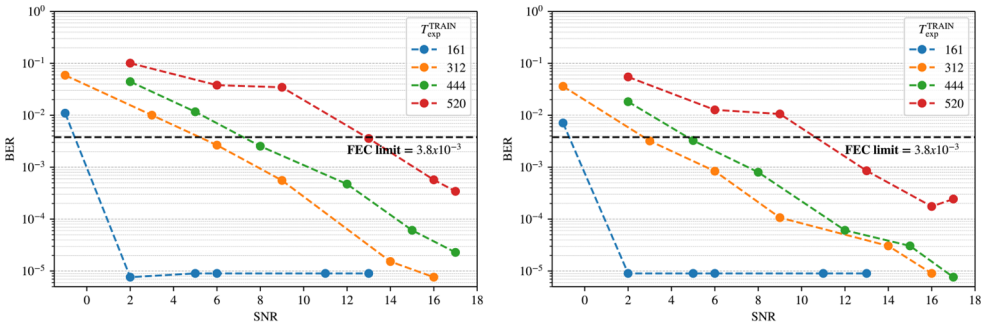
Optics EXPRESS

below 10^{-5} for SNR greater than 3 dB. Figure 11 shows a visual representation of the received signals for exposures of 444 and 520 μs, under the maximum (yellow curve) and minimum SNR (purple curve) conditions, with different average incident power. The estimated SNR is shown within the legend for each curve. To help the accurate visualization of the estimated SNR, a non-scaled version of the filtered template (black dashed curve) is included within the graph. It corresponds to the expected signal to be received in the absence of noise. Clarify that the average power level of the signal changes since it is being adjusted through the voltage source of the light to vary the SNR.



(a) Two stages (b) Three stages

Fig. 9. BER results for training exposures, T_{exp}^{TRAIN}, under moderate SNR (12 to 18 dB).



(a) Two stages (b) Three stages

Fig. 10. BER results for different training exposures, T_{exp}^{TRAIN} under SNR conditions.

Optics EXPRESS

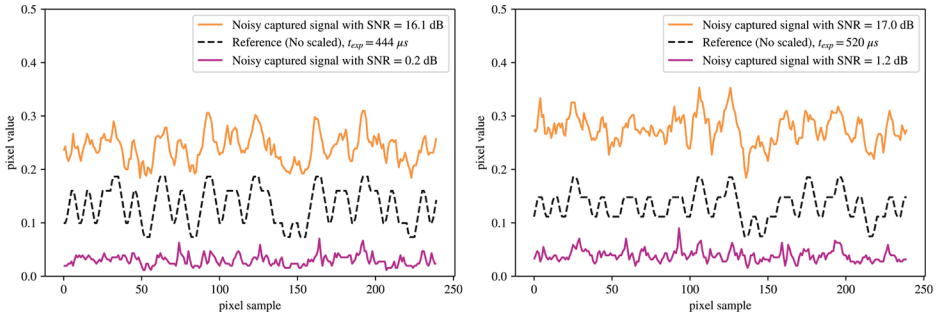


Fig. 11. Real signal examples obtained under different SNR conditions.

6. Discussion

In this section, the experimental results shown before are discussed. The proposed system is evaluated in the context of the available literature up to the authors' knowledge. Finally, a method to estimate the theoretical achievable data rate of the state-of-the-art systems reported is detailed and used to compare their performance based on the parameters of their experimental setups.

Table 4 shows the results comparison against the state-of-the-art. In this table, the fairest comparison results are the ESR (section 4.2) and the bandwidth improvement ratio. The latter refers to the ratio between the signal bandwidth and the cut-off frequency restriction imposed by the camera's exposure time. The proposed system improves the results in both cases.

Table 4. Comparison with the state-of-the-art.

Architecture	Output	Training	Max. ESR [a]	Max. BW ratio[a]	Min. SNR (ESR ≈ 4)[a]	Trainable params.	Theor. data rate[a]
1D-MLP[b]	1	Real	≈ 5 times	≈ 9 times	> 30 dB	100200[c]	2.584 kbps
2D-CAE	256x16	Synthetic	≈ 7 times	≈ 14 times	≥ 5 dB	79185	3.072 kbps

[a] With BER below the FEC limit.
[b] Younous *et al.* [13]
[c] Considering 500 input, 200 hidden and 1 output neurons (non-biased). [14]

Regarding the SNR conditions, this system can decode data with BER below the FEC, with ESR around 4, under SNR as low as 5 dB. However, it can also decode signals with ESR equals 2, under SNR as low as 2 dB, with BER below 10^{-5}. Regarding the number of training parameters, a notable reduction is also observed, which improves network efficiency by up to 20%.

The last column detail the achievable theoretical data rate. To conduct a fair comparison, both systems must meet the following requirements: the number of frames per second, fps is 30 fps, and the vertical resolution of the sensor, H, 1024 pixels. The assumptions considered for estimating the theoretical achievable data rate are detailed below. First, the whole vertical resolution of the IS is utilized for data recovery. Second, perfect synchronization between the transmitter and the receiver is assumed, without blind times at the reception, which means that the camera is operating continuously without stopping between frames. Finally, preamble or postamble bits are not considered. Under these assumptions, a fair comparison of the achievable data rate is computed using Eq. (14) (notice the factor of two since the signal is encoded using Manchester). In the case of 1D-MLDP, the data rate has been computed based on the sampling frequency of the Thorlabs CMOS camera model DCC1645C (detailed in the datasheet [33]) (13,315 kHz), and the maximum transmission frequency reported in [13] (2,240 kHz), resulting in $N_{\text{pps}} = 5.94 \approx 6$.

Optics EXPRESS

Under these conditions, the proposed CAE reaches a binary rate that is up to 500 bps higher than the achieved by the 1DMLP. Furthermore, the camera used in this work has a higher vertical resolution (1080 instead of 1024) and can be configured with 60 fps, so the maximum achievable rate is 6,480 kbps.

$$Rb = \frac{H}{2N_{\text{pps}}} \cdot \text{fps} = H \cdot \frac{T_{\text{s}}}{2t_{\text{sym}}} \cdot \text{fps} = H \cdot \frac{f_{\text{tx}}}{2f_{\text{s}}} \cdot \text{fps} \tag{14}$$

However, if previous requirements are not fulfilled, to ensure the successful packet detection within a frame, it is necessary to send the packet repeatedly, at least during the acquisition of two consecutive frames, and to restrict the packet length (in pixels) to at most half of the vertical size of the ROI [34]. Thus, the theoretical data rate must be divided by a factor of 4. In addition, preambles and postambles must be included in experiments, and the ROI is generally a fraction of the IS vertical resolution and might be considered in the computation of the data rate.

Finally, a series of comments should be added concerning the output size. Since the MLP network has a single output neuron, it is necessary to sweep the image from top to bottom pixel by pixel, which is computationally expensive. Besides, it is necessary to reserve m samples before (or after) with respect to the output sample, depending on whether the equalization weights the inputs backward or forwards (or a mixture of both). This would affect the equalization at the edges, introducing some artificial errors that further reduces the exploitable region for recovering data. In this work, the CAE's output has the same dimensions as the input, allowing equalizing the image by segments (not by pixels). Furthermore, the undesirable effects on the edges are effectively mitigated by the neural network that conveniently evaluates both backward and forward pixel values.

7. Conclusions

This work demonstrated that using a 2D CAE for exposure-related ISI mitigation outperforms the state-of-the-art one-dimensional MLP networks. The proposed system could decode signals with BER below 10^{-5} for exposure times that exceed seven times the transmission symbol time (ESR$\geq$ 7) in optimal SNR conditions. This implies a bandwidth improvement at the reception of approximately 14 times compared to a non-equalized receiver. Moreover, the system can decode signals under low SNR conditions. For example, BER values lower than the FEC limit can be obtained for SNR greater than 9, 5, or 1 dB for ESR of 7, 4, or 2, respectively. In addition, the network is capable of decoding signals with exposure times 10% longer or shorter than the selected one for training, which favors its flexibility to operate with different camera devices. On the other hand, the proposed architecture reduces up to 20% the total trainable network parameters. Finally, the procedure for generating synthetic RS training images was validated. The network's input standardization allows the synthetic procedure to rely exclusively on time-related parameters independent of the deployment scenario: the camera's exposure time, the sampling period, and the transmitted symbol time. These images can also be distorted with artificial noise. Therefore the network training can be conducted offline, for an endless number of cameras and SNR conditions, using vast training databases, ultimately improving the network fitting and the overall decoding robustness.

Funding. This work was funded by the Spanish Research Administration (MINECO project: OSCAR, ref.:TEC 2017-84065-C3-1-R). This project has received funding from the European Union's Horizon 2020 Marie Skłodowska-Curie grant agreement No. 764461.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Optics EXPRESS

References

1. I. of Electrical and E. Engineers, "IEEE standard for local and metropolitan area networks–part 15.7: Short-range optical wireless communications," IEEE Std 802.15.7-2018 (Revision of IEEE Std 802.15.7-2011) pp. 1–407 (2019).
2. W. Liu and Z. Xu, "Some practical constraints and solutions for optical camera communication," Phil. Trans. R. Soc. A **378**(2169), 20190191 (2020).
3. N. Saeed, S. Guo, K.-H. Park, T. Y. Al-Naffouri, and M.-S. Alouini, "Optical camera communications: Survey, use cases, challenges, and future trends," Phys. Commun. **37**, 100900 (2019).
4. T. Kuroda, *Essential Principles of Image Sensors* (CRC Press, 2017).
5. S. Rajagopal, R. D. Roberts, and S.-K. Lim, "IEEE 802.15.7 visible light communication: modulation schemes and dimming support," IEEE Commun. Mag. **50**(3), 72–82 (2012).
6. P. Luo, M. Zhang, Z. Ghassemlooy, S. Zvanovec, S. Feng, and P. Zhang, "Undersampled-based modulation schemes for optical camera communications," IEEE Commun. Mag. **56**(2), 204–212 (2018).
7. H. Aoyama and M. Oshima, "Line scan sampling for visible light communication: Theory and practice," in *2015 IEEE International Conference on Communications (ICC)*, (IEEE, 2015), pp. 5060–5065.
8. N. Saha, M. S. Ifthekhar, N. T. Le, and Y. M. Jang, "Survey on optical camera communications: challenges and opportunities," IET Optoelectron. **9**(5), 172–183 (2015).
9. N. T. Le, M. Hossain, and Y. M. Jang, "A survey of design and implementation for optical camera communication," Signal Proc.: Image Commun. **53**, 95–109 (2017).
10. V. Matus, V. Guerra, S. Zvanovec, J. Rabadan, and R. Perez-Jimenez, "Sandstorm effect on experimental optical camera communication," Appl. Opt. **60**(1), 75–82 (2021).
11. V. Matus, E. Eso, S. R. Teli, R. Perez-Jimenez, and S. Zvanovec, "Experimentally derived feasibility of optical camera communications under turbulence and fog conditions," Sensors **20**(3), 757 (2020).
12. V. Matus, V. Guerra, C. Jurado-Verdu, S. R. Teli, S. Zvanovec, J. Rabadan, and R. Perez-Jimenez, "Experimental evaluation of an analog gain optimization algorithm in optical camera communications," in *2020 12th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, (IEEE, 2020), pp. 1–5.
13. O. I. Younus, N. B. Hassan, Z. Ghassemlooy, P. A. Haigh, S. Zvanovec, L. N. Alves, and H. Le Minh, "Data rate enhancement in optical camera communications using an artificial neural network equaliser," IEEE Access **8**, 42656–42665 (2020).
14. O. I. Younus, N. B. Hassan, Z. Ghassemlooy, S. Zvanovec, L. N. Alves, and H. Le-Minh, "The utilization of artificial neural network equalizer in optical camera communications," Sensors **21**(8), 2826 (2021).
15. Y.-C. Chuang, C.-W. Chow, Y. Liu, C.-H. Yeh, X.-L. Liao, K.-H. Lin, and Y.-Y. Chen, "Using logistic regression classification for mitigating high noise-ratio advisement light-panel in rolling-shutter based visible light communications," Opt. Express **27**(21), 29924–29929 (2019).
16. K.-L. Hsu, C.-W. Chow, Y. Liu, Y.-C. Wu, C.-Y. Hong, X.-L. Liao, K.-H. Lin, and Y.-Y. Chen, "Rolling-shutter-effect camera-based visible light communication using RGB channel separation and an artificial neural network," Opt. Express **28**(26), 39956–39962 (2020).
17. A. Islam, M. T. Hossan, and Y. M. Jang, "Convolutional neural networkscheme–based optical camera communication system for intelligent internet of vehicles," Int. J. Distributed Sens. Networks **14**(4), 155014771877015 (2018).
18. M. R. Soares, N. Chaudhary, E. Eso, O. I. Younus, L. N. Alves, and Z. Ghassemlooy, "Optical camera communications with convolutional neural network for vehicle-tovehicle links," in *2020 12th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, (IEEE, 2020), pp. 1–6.
19. L. Liu, R. Deng, and L.-K. Chen, "47-kbit/s RGB-LED-based optical camera communication based on 2D-CNN and XOR-based data loss compensation," Opt. Express **27**(23), 33840–33846 (2019).
20. V. Jain and S. Seung, "Natural image denoising with convolutional networks," Advances in neural information processing systems **21**, 769–776 (2008).
21. L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* pp. 241–246 (2016).
22. M. Nishio, C. Nagashima, S. Hirabayashi, A. Ohnishi, K. Sasaki, T. Sagawa, M. Hamada, and T. Yamashita, "Convolutional auto-encoder for image denoising of ultra-low-dose CT," Heliyon **3**(8), e00393 (2017).
23. V. Turchenko, E. Chalmers, and A. Luczak, "A deep convolutional auto-encoder with pooling-unpooling layers in caffe," arXiv preprint arXiv:1701.04949 (2017).
24. W.-C. Wang, C.-W. Chow, L.-Y. Wei, Y. Liu, and C.-H. Yeh, "Long distance non-line-of-sight (NLOS) visible light signal detection based on rolling-shutter-patterning of mobile-phone camera," Opt. Express **25**(9), 10103–10108 (2017).
25. N. B. Hassan, Z. Ghassemlooy, S. Zvanovec, P. Luo, and H. Le-Minh, "Non-line-of-sight $2 \times N$ indoor optical camera communications," Appl. Opt. **57**(7), B144–B149 (2018).
26. J.-K. Lain, F.-C. Jhan, and Z.-D. Yang, "Non-line-of-sight optical camera communication in a heterogeneous reflective background," IEEE Photonics J. **11**(1), 1–8 (2019).
27. X. Li, N. B. Hassan, A. Burton, Z. Ghassemlooy, S. Zvanovec, and R. Perez-Jimenez, "A simplified model for the rolling shutter based camera in optical camera communications," in *2019 15th International Conference on Telecommunications (ConTEL)*, (IEEE, 2019), pp. 1–5.
28. Y. Goto, I. Takai, T. Yamazato, H. Okada, T. Fujii, S. Kawahito, S. Arai, T. Yendo, and K. Kamakura, "A new automotive VLC system using optical communication image sensor," IEEE Photonics J. **8**(3), 1–17 (2016).

Optics EXPRESS

29. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016). [Online; accessed 1-June-2021] Available: www.deeplearningbook.org.
30. M. Khalid, J. Baber, M. K. Kasi, M. Bakhtyar, V. Devi, and N. Sheikh, "Empirical evaluation of activation functions in deep convolution neural network for facial expression recognition," in *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, (2020), pp. 204–207.
31. L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," The Journal of Machine Learning Research **18**(1), 6765–6816 (2017).
32. Sony Corporation, *IMX219PQH5-C, Diagonal 4.60 mm (Type 1/4.0) 8 Mega-Pixel CMOS Image Sensor with Square Pixel for Color Cameras, Datasheet* (Sony Corporation, 2014).
33. Thorlabs Scientific Imaging, "Thorlabs DCx camera functional description and SDK manual," (2018). [Online; accessed 1-June-2021] Available: www.thorlabs.com/thorProduct.cfm?partNumber=DCC1645C.
34. C. Jurado-Verdu, V. Matus, J. Rabadan, V. Guerra, and R. Perez-Jimenez, "Correlation-based receiver for optical camera communications," Opt. Express **27**(14), 19150–19155 (2019).