UNIVERSITÀ
DI TRENTO
Department of
Information Engineering and Computer Science

Doctoral School in
Information and Communication Technology

ADVANCED TECHNIQUES FOR SPAD-BASED CMOS D-TOF SYSTEMS

Alessandro Tontini

Advisor
Prof. Roberto Passerone
University of Trento

Co-Advisor
Dott. Ing. Leonardo Gasparini
Fondazione Bruno Kessler

January 2024

Abstract

The possibility to enable spatial perception to electronic devices gave rise to a number of important development results in a wide range of fields, from consumer and entertainment applications to industrial environments, automotive and aerospace. Among the many techniques which can be used to measure the three-dimensional (3D) information of the observed scene, the unique features offered by direct time-of-flight (d-ToF) with single photon avalanche diodes (SPADs) integrated into a standard CMOS process result in a high interest for development from both researchers and market stakeholders. Despite the net advantages of SPAD-based CMOS d-ToF systems over other techniques, still many challenges have to be addressed. The first performance-limiting factor is represented by the presence of uncorrelated background light, which poses a physical limit to the maximum achievable measurement range. Another problem of concern, especially for scenarios where many similar systems are expected to operate together, is represented by the mutual system-to-system interference, especially for industrial and automotive scenarios where the need to guarantee safety of operations is a pillar. Each application, with its own set of requirements, leads to a different set of design challenges. However, given the statistical nature of photons, the common denominator for such systems is the necessity to operate on a statistical basis, i.e., to run a number of repeated acquisitions over which the time-of-flight (ToF) information is extracted. The gold standard to manage a possibly huge amount of data is to compress them into a histogram memory, which represents the statistical distribution of the arrival time of photons collected during the acquisition. Considering the increased interest for long-range systems capable of both high imaging and ranging resolutions, the amount of data to be handled reaches alarming levels.

In this thesis, we propose an in-depth investigation of the aforesaid limitations. The problem of background light has been extensively studied over the years, and already a wide set of techniques which can mitigate the problem are proposed. However, the trend was to investigate or propose single solutions, with a lack of knowledge regarding how different implementations behave on different scenarios. For such reason, our effort in this view focused on the comparison of existing techniques against each other, highlighting each pros and cons and suggesting the possibility to combine them to increase the performance. Regarding the problem of mutual system interference, we propose the first per-pixel implementation of an active interference-rejection technique, with measurement results from a chip designed on purpose. To advance the state-of-the-art in the direction of reducing the amount of data generated by such systems, we provide for the first time a methodology to completely avoid the construction of a resource-consuming *histogram of timestamps. Many of the results found in our investigations are based on preliminary investigations with Monte Carlo simulations, while the most important achievements in terms of interference rejection capability and data reduction are supported by measurements obtained with real sensors.*

Keywords

*Direct time-of-flight (**d-ToF**); Light Detection and Ranging (**LiDAR**); Single Photon Avalanche Diode (**SPAD**); Complementary Metal-Oxide Semiconductor (**CMOS**).*

Contents

List of Tables

List of Figures

Glossary

μC Micro controller.

3D Three-dimensional.

ADAS Advanced Driver Assistance Systems.

APD Avalanche Photo Diode.

AQC Active Quenching Circuit.

AR Augmented Reality.

CDMA Code Division Multiple Access.

CIS CMOS Image Sensor.

CMOS Complementary Metal Oxide Semiconductor.

d-ToF direct-Time-of-Flight.

DCR Dark Count Rate.

DNL Differential nonlinearity.

FLIM Fluorescence Lifetime Imaging.

FPGA Field Programmable Gate Array.

INL Integral nonlinearity.

LiDAR Light Detection and Ranging.

NIR Near InfraRed.

PDE Photon Detection Efficiency.

PDF Probability Density Function.

PDP Photon Detection Probability.

PQC Passive Quenching Circuit.

PSK Phase-Shift Keying.

SNR Signal to Noise Ratio.

SPAD Single Photon Avalanche Diode.

SPI Serial Peripheral Interface.

TAC Time to Amplitude Converter.

TCSPC Time-Correlated Single Photon Counting.

TDC Time to Digital Converter.

ToF Time-of-Flight.

VR Virtual Reality.

Chapter 1

Introduction

Among the 5 human senses, sight is a pillar. More than 80% of the information we collect from the surrounding comes from vision [1], and it is a common belief to be our most valued sense [2], which is also confirmed by its domination over the research interest in the field of sensory modalities [3]. Among the many properties of the surrounding environment which can be elaborated thanks to vision, the awareness of the third dimension is the most important one, as it affects the physical interaction we have with the external world. For such reasons, it is straightforward and intuitive to extend such possibilities to our technological means, through the implementation of devices which can measure the three-dimensional (3D) information from the observed scene.

The applications of three-dimensional (3D) imaging span a wide range of industries and disciplines, each presenting challenges and opportunities. Perhaps one of the most challenging and prominent applications of 3D imaging is in the domain of automotive LiDAR, where the ability to accurately perceive the surrounding environment in three dimensions is critical for ensuring the safety and autonomy of vehicles and pedestrians. From detecting obstacles to navigating complex urban environments, LiDAR systems represent one of the most promising technologies which is currently being deeply developed. Similar challenges, in particular concerning safety of operations, are posed by the possibility to use 3D imaging systems within industrial automation and robotics, where applications such as object tracking, quality control and robotic navigation in dynamic environments are nowadays possible. Another interesting field of application is represented by the entertainment industry leveraging on 3D imaging technologies to create immersive augmented reality (AR) and virtual reality (VR) experiences, where the capture and rendering of highly detailed 3D environments and objects enhances the sense of presence and realism for users.

In this work, in particular, we focus on sensors for 3D imaging implemented in Complementary Metal-Oxide Semiconductor (CMOS) technology based on the direct Time-of-Flight (d-ToF) measurement technique with Single Photon Avalanche Diode (SPAD) technology.

SPADs offer several distinctive advantages over traditional 3D imaging technologies, making them particularly performing in such field. Perhaps the most notable feature of SPADs is their high sensitivity, allowing for the capture of faint signals in low-light conditions (down to the single photon), essential for all applications where the signal of interest may be dramatically low in terms of power, such as in long-range LiDAR applications. Furthermore, another key characteristic of SPADs in this view is the ultra-fast response time, typically in the order of picoseconds, which is an essential requirement to guarantee the desired level of precision in the estimation of the time-of-flight. In addition, SPADs can nowadays provide sufficiently low noise values, in terms of both dark count rate and timing jitter, to ensure the expected performance to a considerable extent. In addition to their intrinsic properties, SPADs offer a crucial advantage in 3D imaging applications through the possibility to be fully integrated in standard CMOS processes. This allows for high data processing capabilities to be implemented, expanding the possibilities given by such high performance detectors (in particular with the most advanced technology nodes). Moreover, a standard CMOS process represents a platform suitable for industrial production, making them an economically viable choice.

In automotive LiDAR applications, SPAD-based imaging systems must adhere to strict criteria across several fronts. They must ensure an extended range capability, enabling the detection of objects at distances exceeding 200 meters. This long range capability is crucial for ensuring early identification of potential hazards and obstacles on the road. Moreover, these systems require high resolution capabilities, in the range of centimeters, essential for precise localization and identification of objects, especially in densely cluttered urban environments. Alongside this, they must operate at frame rates high enough to ensure timely updates of the surrounding environment, facilitating the smooth operation of autonomous driving systems. Furthermore, robustness is paramount, necessitating the ability to function in various environmental conditions. Among this, the capability to operate in a wide range of external illumination conditions is crucial, in particular considering extremely challenging cases of direct glare from other similar systems, giving rise to potential failures due to interferences. Unfortunately, many of the requirements for successful operations of SPADs within an automotive LiDAR application contrast to each other. For such a reason, the investigation must take into account various aspects, so as to consider the application as a whole, bearing in mind that the goal is to provide the system with as many features as possible to be adaptable to a wide range of scenarios.

In particular, we concentrate on the study of techniques to advance the state-of-the art in the field, focusing on three aspects: mitigation of the negative effects of background light, rejection of mutual interference and reduction of the amount of generated data. The results are supported by numerical (Monte Carlo) simulations and measurements with real sensors. The main outcomes of our work can be summarized as follows. First, we provide an in-depth investigation of how different background-rejection techniques compare to each other, demonstrating that the combination of different techniques applied on top of each other can improve the performance. Additionally, we identify the technique [4] which, alone, can guarantee the best performance over a wide range of operating conditions. Then, we demonstrate for the first time the in-pixel implementation of an active interference-rejection technique, with measurement results from a sensor designed in a 110 nm CIS (CMOS Image Sensor) process. Concerning the problem of data reduction, we provide results from the aforementioned sensor, which implements a smart readout technique able to output only valid data from the sensor. We also demonstrate, for the first time, the possibility to obtain the Time-of-Flight (ToF) information without the need to build a memory and power-hungry histogram of timestamps, paving the way for sensors with unprecedented resolutions and compactness.

This thesis is organized as follows. In the remainder of this chapter, we provide a high-level overview of the structure and main features of a Single Photon Avalanche Diode (SPAD), as it represents the fundamental photon-sensitive element on which our results are based. In Chapter 2, we provide a comprehensive review of the most important related works in the field of SPAD-based CMOS sensors for d-ToF applications, spanning from the first milestones where a SPAD is used in d-ToF measurements up to the most recent architectures. As several of our investigations and results are based on simulations, Chapter 3 provides in-depth insights on the structure and development of the numerical Monte Carlo simulator we used through our work. In Chapter 4 we report on a thorough discussion of how different background-rejection techniques compare to each other, focusing on the impact of circuital implementation and identifying the technique which can guarantee the best performance over a range of operating conditions. Chapter 5 is dedicated to the implementation details and measurement results of a 1x64 pixel linear sensor implemented in 110nm CIS process, which integrates interference-rejection capability at pixel level with an optimized readout architecture. Our last contribution, which focuses on the possibility to obtain a histogram-free SPAD-based

LiDAR with reduced amount of hardware resources is deeply investigated in Chapter 6. A conclusive summary with remarks and future perspectives, along with recommendations for further research, is provided in Chapter 7.

1.1 Single Photon Avalanche Diode (SPAD)

A SPAD is a p-n junction designed to be operated in reverse bias well above its breakdown voltage [5]. In such conditions, a single electron-hole pair generated in the depletion region is accelerated by the electric field and by impact ionization a macroscopic self-sustained avalanche is created. If the electron-hole pair is generated by an impinging photon (by means of the photoelectric effect), the rising edge of the current avalanche indicates its arrival time. Due to the sub-nanosecond scale of such phenomenon, the arrival time of photons can be determined with very high temporal resolution. After the avalanche current has been detected, the current flow needs to be stopped for mainly two reasons. First, no additional photons can be detected during an ongoing avalanche. Second, the amount of current flowing in the reverse-biased junction would lead to permanent damage of the structure. For such reasons, the SPAD is always coupled to a so-called quenching circuit. The quenching process, which can be either passive or active, stops the ongoing avalanche by temporarily setting the junction voltage below the breakdown. Once the current flow ceases, the junction reverse voltage is set again above breakdown to enable the detection of further photons. There are a number of physical parameters characterizing the performance of a SPAD. In this section, we will first introduce the main concept behind passive and active quenching techniques. Then, the most important SPAD parameters are considered, i.e. the Photon Detection Efficiency (PDE), Dark Count Rate (DCR) and afterpulsing.

1.1.1 Passive quenching

Passive Quenching Circuits (PQCs) represents the earliest and simplest way of controlling the behavior of a SPAD avalanche. In this scheme, the SPAD is connected in series with a high resistive load (called *quenching resistor*) and whenever an avalanche occurs, the voltage drop developed across the quenching resistor by the avalanche current reduces the excess bias voltage below a level where the probability of impact ionization of the electron-hole pairs is insufficient to self-sustain the current avalanche [5]. A principle schematic of a passively quenched SPAD is represented in Figure 1.1-(a). Considering a quenching resistor of value R_D, the quenching time can be computed as follows [5]:

$$T_q = (C_d + C_s)R_d /\!/ R_Q\,, \tag{1.1}$$

where C_d is the SPAD junction capacitance, C_s is the parasitic capacitance (with respect to ground of the diode terminal) and R_d is the diode resistance. In Figure 1.1-(b), a passive quenching scheme for a SPAD integrated into a CMOS process is shown. In this case, the quenching resistor is substituted with a transistor controlled by a voltage V_Q. By tuning the V_Q control voltage it is possible to adjust the junction discharge time, commonly referred to as the SPAD *dead-time (T_{dead})*. Between the SPAD anode and the digital comparator, an additional transistor, controlled by a voltage V_C, can be required to clamp the SPAD anode voltage to be compliant with the following logic, which usually runs at lower supply voltage.

The main drawback of passively-quenched SPADs is that photon sensitivity (as defined and discussed in Section 1.1.3) is not constant during an avalanche recovery. As soon as the avalanche occurs, the diode current falls below the self-sustaining value, resulting in a maximum voltage value at the SPAD anode $V_A \approx V_{EX}$, so that $V_S - V_A \leq V_{BD}$ and the capacitances C_d and C_s start to be recharged by the current flowing in the quenching resistor (Figure 1.1-(a)) or the quenching transistor (Figure 1.1-(b)). In such a situation, the SPAD sensitivity is nearly zero at the beginning of the recovery process ($V_S - V_A \leq V_{BD}$) and exponentially rises to the maximum value as the SPAD voltage is restored. During this process, any other impinging photon have a progressively higher probability of being detected and triggering a new avalanche. Any other additional detection occurring during the SPAD recovery is of particular concern considering that the voltage pulse which is created by the passive quenching circuit needs to cross a voltage threshold to be properly detected by the following circuits. If the new avalanche occurs when the voltage signal has not back-crossed the comparator threshold, the quenching process is restarted and the output voltage pulse is extended. If the photon flux is too high, the SPAD can be eventually paralyzed impairing the photon detection process. In Figure 1.2, the temporal evolution of the SPAD voltage, V_A, is shown. After the second photon detection, a third photon triggers an avalanche, restoring the SPAD anode voltage to V_{EX} before it re-crosses the comparator threshold, resulting into an elongated electrical pulse output. Despite the issues which may occur due to the retriggering behavior, passively quenched SPADs are still used in today's applications due to their compactness, ease of implementation and reliability [6–10].

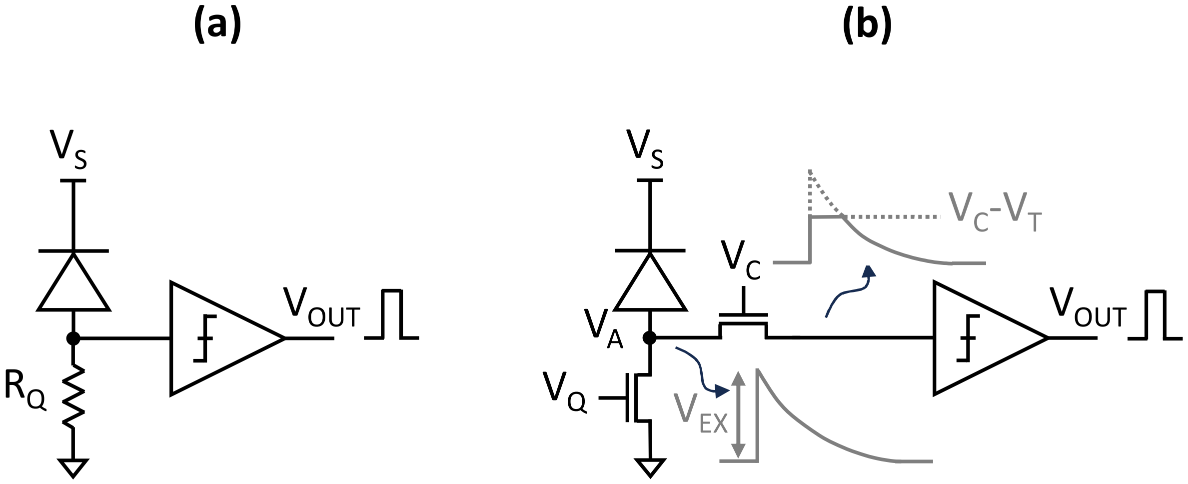


Figure 1.1: Passive quenching scheme. In (a), a principle schematic with a resistor R_Q as the quenching element, while in (b) a solution commonly used for SPADs integrated into a CMOS process is shown. In this case, the quenching operation is exploited by means of a transistor controlled by a voltage V_Q. By tuning the control voltage it is possible to change the transistor conductivity, effectively controlling the SPAD quench time. Between the SPAD anode and the digital comparator, a clamp transistor, controlled by voltage V_C, is required to limit the voltage swing.

1.1.2 Active quenching

The limitations imposed by PQCs can be overcome with Active Quenching Circuits (AQCs) [5, 11], where the onset of an avalanche is sensed and an active control loop reacts back to the SPAD junction controlling the bias voltage. With this circuit topology it is possible to drive the bias voltage below the breakdown, so that the quenching duration is not affected by statistical fluctuations of avalanche multiplication as in PQCs [12]. A principle schematic of an AQC is shown in Figure 1.3-(a). In such implementation, the avalanche is sensed by a fast comparator which triggers a monostable connected in feedback to the SPAD anode, actively controlling the quench process. A more detailed implementation is shown in Figure 1.3-(b). The onset of an avalanche is sensed at the SPAD anode (node V_A). The voltage pulse is first rectified by means of an inverter, running at a supply voltage V_{HV} (usually 3.3V). The rectified voltage pulse is level-adjusted by means of a level shifter and enters a first monostable M1. The first monostable creates a pulse V_Q of duration T_{dead} which actively quenches the SPAD below the breakdown voltage, i.e., $V_{SPAD} = V_S - V_{HV}$, with $V_{HV} > V_{EX}$. After the pulse from monostable M1, a second fast pulse by monostable M2 is used to reset the SPAD voltage back to $V_S = V_{BD} + V_{EX}$, closing the quench phase. ACQs allows external control of the SPAD state, to realize gated SPAD operations. In this case, an

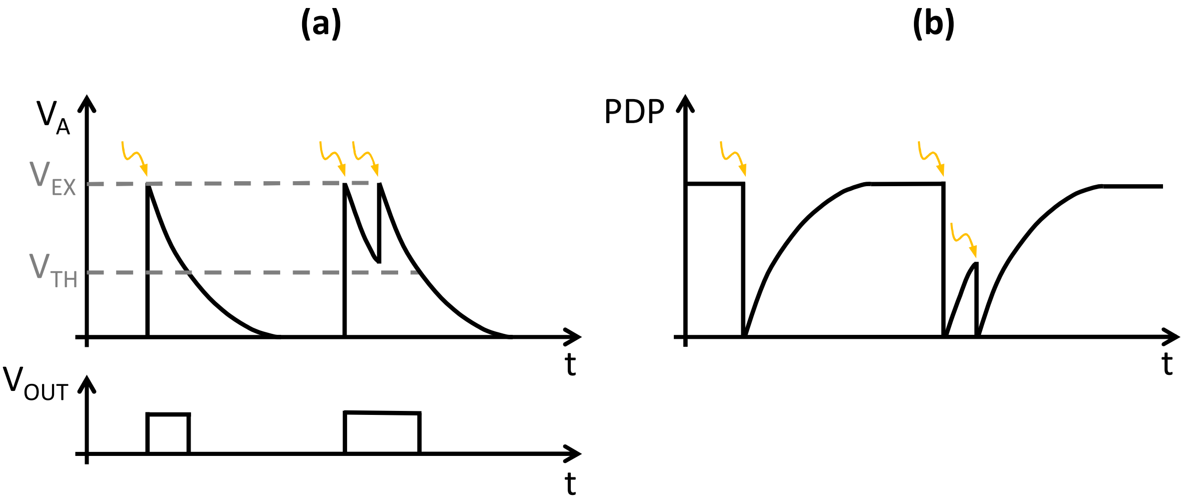


Figure 1.2: Re-triggering issue in passively quenched SPADs. In (a), the SPAD anode voltage, V_A, is shown. As soon as a photon triggers an avalanche, the anode voltage rises to V_{EX}, effectively quenching the avalanche. The SPAD capacitances C_d and C_s are recharged by means of the quenching resistor/transistor, gradually restoring the SPAD voltage to the nominal value of $V_S = V_{BD} + V_{EX}$. If a photon impinges during the SPAD recharge phase, it may trigger a second avalanche which is not sensed by the following electronics, as the comparator threshold is not back-crossed again. In (b), the SPAD sensitivity (in terms of PDP) over time is shown. Even though the sensitivity is lower soon after an avalanche, a re-trigger may occur in case of high photon flux operations.

external signal can control the state of the quenching voltage V_Q, forcing the SPAD in an OFF state for the desired amount of time. This functionality is of particular interest in gated Fluorescence Lifetime IMaging (FLIM) [13, 14] or whenever any prior information is known about the measurement conditions, for example in gated applications for d-ToF measurements [6]. Gated operations can also be achieved with PQCs by masking the output of the front-end based on an external gating signal. The key contrast with ACQs lies in how they handle the SPAD: in ACQs, the SPAD is consistently turned off, ensuring it always begins in an active (ON) state once the gating ends. In contrast, PQCs merely mask the SPAD's output during gating without ensuring its state upon gating signal deassertion.

Despite the clear advantage over PQCs, AQCs are more complex and require a bigger silicon area, as more transistors are required for the quench/reset phases and for the monostable circuits which drive them. Therefore, their implementation in sensors with imaging capabilities needs to be carefully evaluated, and in principle it is advantageous when non-idealities of SPADs such as afterpulsing may be of concern for a particular measurement process, as for example in photon-starved applications like the detection of entangled photons for quantum physics experiments [15, 16], or in high precision, single point d-ToF measurements [17]. A comparison between passive/active quenching schemes in terms of the capability to estimate the correct photon flux is shown in Figure 1.4. The limit of a passively quenched approach is evident when the photon flux is too intense that the re-triggering phenomenon occurs, limiting the effective count rate.

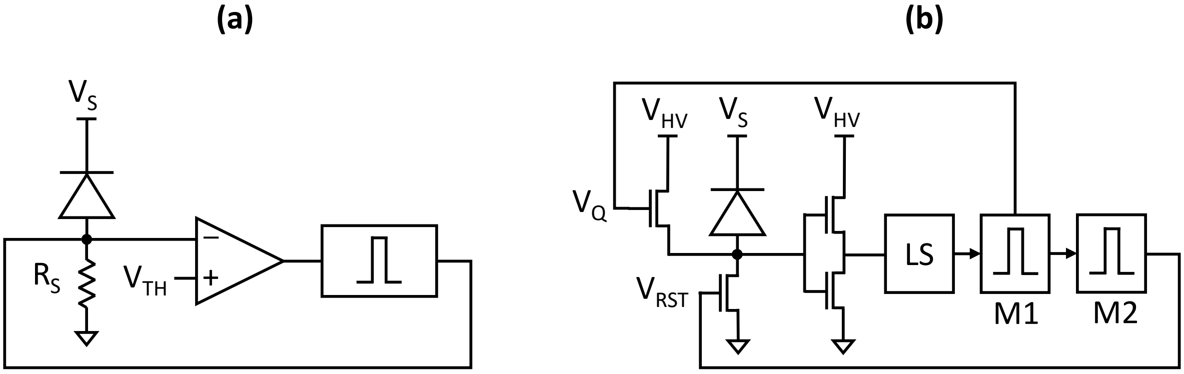


Figure 1.3: Active quenching principle schematic (a) and detailed implementation (b) with quench and reset transistors.

1.1.3 Photon Detection Efficiency (PDE)

The Photon Detection Efficiency (PDE) is defined as the ratio between the number of electrical pulses at the SPAD output and the number of impinging photons on the SPAD active area [18, 19], at a given wavelength. The PDE defines the behavior of a SPAD in terms of light sensitivity, which is of paramount importance in all applications where the intensity of the signal of interest is particularly low, as, for example, in flash Light Detection and Ranging (LiDAR) systems where a diffuse illumination scheme is employed, resulting into a low returning echo intensity.

The PDE can be expressed in two ways. With a first definition, it is expressed as the product between the photon absorption efficiency, η, and the avalanche triggering probability P_T [18]. Zappa et al. [18], defines the absorption efficiency as

$$\eta = (1 - R)e^{-\alpha D}(1 - e^{-\alpha W}), \tag{1.2}$$

where α, W, D and R are physical parameters of the SPAD structure. In particular, α is the silicon absorption coefficient, W indicates the thickness of the depletion region, D is the junction depth and R is the reflection coefficient for the interface between air and silicon. The avalanche triggering probability P_T also depends on a number of physical parameters intrinsic to the specific SPAD structure [18] and can be computed

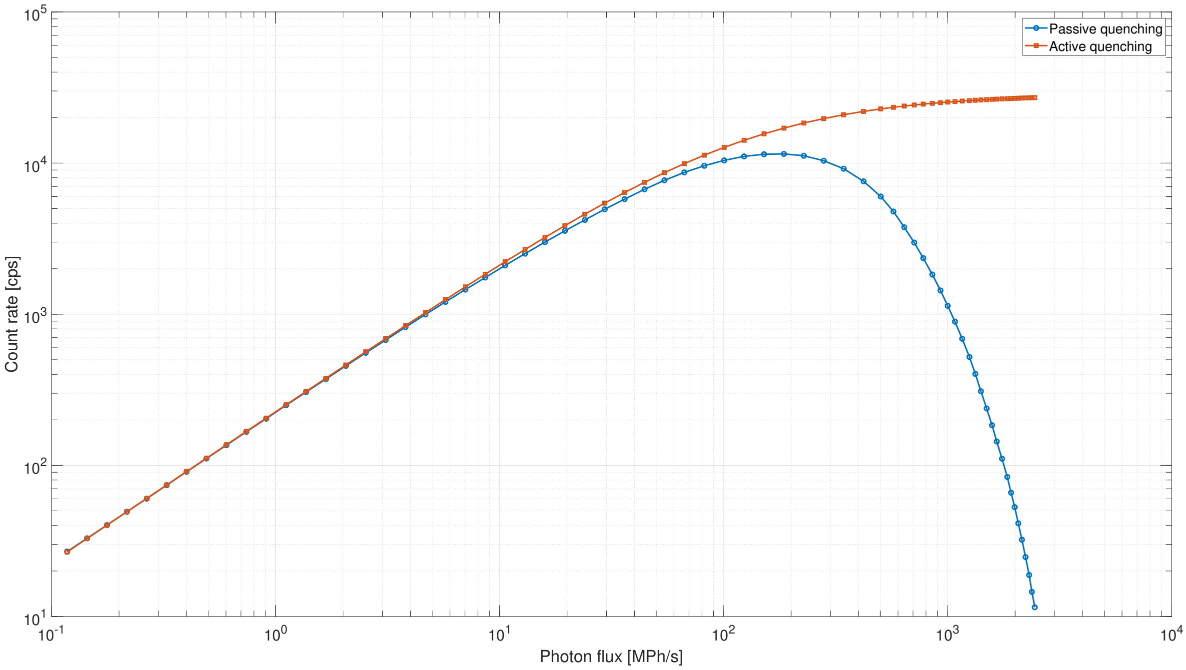


Figure 1.4: Comparison between passive/active quenching schemes performance in terms of output count rate. For high photon flux, the actively quenched SPAD reaches a plateau, while the passive quench approach result in a decrease of the output count rate, due to the paralysis effect occurring from the SPAD re-triggering during recharge.

by integrating the derivative of the probability $P_n(x)$ and $P_p(x)$ that an electron (n) and a hole (p) generates an infinite number of descendants while travelling in the depletion region from a position x, thus creating an avalanche, as defined by McIntyre [20]. With such a definition, it is possible to estimate the PDE beforehand, with the knowledge of all the required physical parameters of the SPAD.

An alternative definition is given in the work from Moreno-García et al. [19], where no knowledge of intrinsic SPADs parameter is required, as it is based on a laboratory characterization approach:

$$PDE = \frac{A - DCR}{P}, \tag{1.3}$$

where A is the rate of avalanches observed at the SPAD output, DCR is the measured dark count rate and P is the rate of impinging photons on the active area.

Being the PDE dependent on the absorption coefficient of silicon, the sensitivity over the spectrum of different light wavelengths is not constant. Figure 1.5 shows the measured SPAD PDE versus wavelength for different breakdown voltages from the work of Moreno-García et al [19]. The PDE depends upon the depth of the junction, and a typical SPAD exhibits its peak in the visible region around $\lambda = 450$ nm [19, 21, 22], gradually decreasing when approaching the beginning of the Near InfraRed (NIR) region.

Considering the increasing interest in the application of SPADs for automotive applications, there have been several studies to enhance the sensitivity in the non-visible region of the spectrum, which achieved more than 20% PDE at 905nm wavelength [23–25].

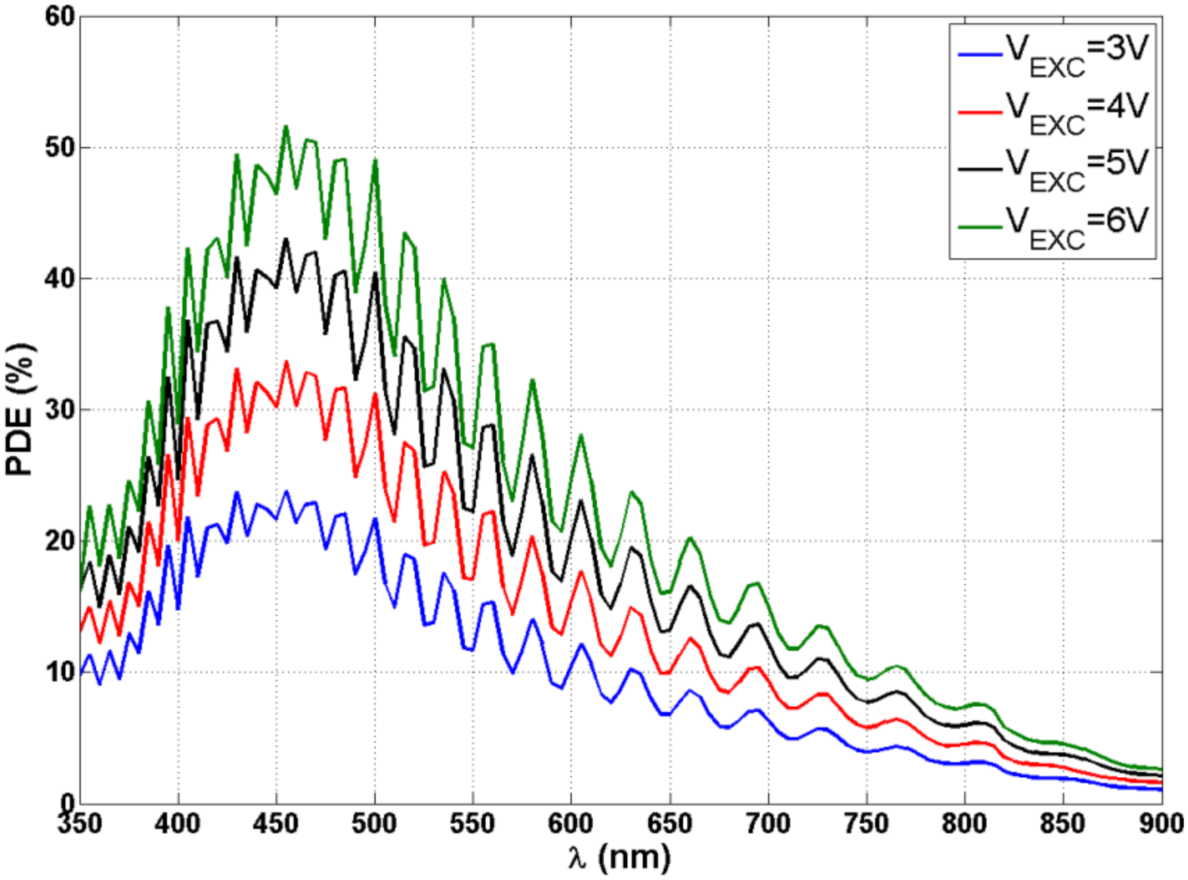


Figure 1.5: SPAD PDE versus wavelenght with SPADs designed in a 110nm CIS process, from the work of Moreno Garcia et al. [19].

1.1.4 Dark Count Rate (DCR) and afterpulsing

An ideal SPAD would provide electrical pulses only upon the detection of photons impinging on the active area. Real devices, however, exhibits avalanches even in completely dark environments, thus not generated by an absorbed photon, resulting into a baseline detector noise. There are two sources of noise in SPAD detectors: dark count events and afterpulsing events.

Dark count events can be generated by carriers produced by thermal, tunnelling and trap-assisted effects by means of generation-recombination processes within the semiconductor lattice, that are able to trigger an avalanche indistinguishable from a photon-generated one. Dark count events are Poissonian distributed and constitutes the primary source of SPAD internal noise. The main performance-limiting factor of a SPAD used for LiDAR applications is given by a high DCR, which results into a reduction in the average amount of time available to detect photons. Except in the case of high intensity, DCR is usually a well manageable non-ideality as its statistical distribution is the same as for background light. For such reason, DCR can also be exploited as a source of entropy for the generation of true random numbers [26]. Being mainly driven by thermal effects, DCR is temperature-dependent, increasing exponentially as temperature rises [27, 28]. Another DCR source, which is less dependent on temperature, is via carrier injection by band-to-band tunneling. This source of DCR has been shown to be non-negligible especially during the transition to more scaled CMOS technology nodes [29,30]. A thorough examination of the phenomenon is provided by Xu et al. [31]. Typical values of low-noise SPADs designed to be compliant with CMOS process (considering a typical 3V excess bias) lie in the range $[0.18, 0.83]$ $Hz/\mu m^2$ [19, 21, 22].

Afterpulsing, beside DCR, is the second source of noise in SPAD detectors. Differently from DCR, afterpulsing-related events occur when carriers travelling in the depletion region during an avalanche are trapped in local defects of the lattice structure (also called "traps") and released after a certain amount of time. Differently from DCR (which obeys the rules of a Poisson process), afterpulsing is a correlated noise, as the carriers which eventually re-trigger an avalanche, belongs to a previous detection event. An effective way to reduce the impact of afterpulsing is to increase the SPAD dead-time [18], in order to increase the number of carriers released from the traps before the SPAD can be retriggered. In particular, active quenching schemes are more effective than passive approaches, as in the former the SPAD is forced below breakdown for the entire duration of the dead-time, while in the latter the SPAD voltage gradually recovers toward breakdown, thus impairing the traps release process [18]. As afterpulsing is an unwanted behavior, it is desirable to keep its probability (P_{AP}) as low as possible. In the work from Ziarkash et al. [32], several commercial SPADs were tested and characterized in terms of afterpulsing probability, resulting in a P_{AP} in the range $[0.01, 8.5]\%$. Concerning the impact of the type of quenching scheme employed, it is very interesting to notice that in the work from Farina et al. [33], the impact of a well designed AQC enables the use of aggressive dead times down to 10 ns with a $P_{AP} \simeq 1.5\%$, as opposed to the work from Xu et al., where a similar performance ($P_{AP} \simeq 0.85\%$) has been obtained with a far more longer dead time of 150 ns.

Chapter 2

Related work

In this chapter, we provide an overview of the literature to properly place our work in context. We analyze several architectures proposed over the years concerning SPAD-based d-ToF sensors, highlighting the most important achievements and milestones in the field.

2.1 Pioneering results

Pioneering results with SPAD-based CMOS d-ToF sensors have been obtained by Cristiano Niclass et al. in the research group lead by Professor Edoardo Charbon at the École Polytechnique Fédérale de Lausanne (EPFL). In one of the very first implementations [34], the possibility to obtain the 3D map of objects with a simple geometrical structure has been demonstrated. Measurements have been achieved by means of an 8×4 SPAD array, with external time-to-digital conversion and resistor-based passive quenching scheme. A substantial resolution improvement has been achieved by reducing the pixel pitch by means of a quenching transistor, which also enables a fine control of the SPAD dead-time, resulting in a 32×32 SPAD array [35] capable of 3D mapping complex objects. On a similar sensor with the same physical resolution [36], an imaging resolution of 64×64 pixels is obtained by micro-scanning the array matrix to effectively double the resolution. With a further integration effort, a physical resolution of 64×48 pixels has been demonstrated [37], where a further advance in the state-of-the-art is obtained thanks to an event-driven readout architecture which improves the readout time if compared to standard solutions based on a sequential scanning.

Despite sub-millimeter ranging resolution has been demonstrated with CMOS integrated SPAD arrays, research efforts focused on the on-chip or even in-pixel integration of photon timestamping circuits, paving the way for true fully integrated SPAD-based d-ToF sensors.

In the work from Stoppa et al. [38], an early attempt at integrating photon timestamping capability is demonstrated by means of a 1×64 SPAD array where each pixel integrates a Time-to-Analog Converter (TAC), with a reported maximum range of 5 m with centimeter precision. A substantial improvement in terms of imaging resolution is presented by Niclass et al. [39], with an array of 128×128 actively-quenched SPADs embedding a bank of 32 on-chip TDCs, with a reported maximum range of 3.75 m and sub-cm accuracy. Richardson et al. [40] propose for the first time a fully integrated SPAD array, with a resolution of 32×32, where each pixel, of 50 μm pitch, integrates a SPAD and a ring oscillator-based Time to Digital Converter (TDC) plus memory elements. A similar approach, with the same 32×32 pixel resolution, is proposed by Stoppa et al. [41], with the possibility to obtain both photon intensity and timing information by means of a per-pixel Time to Analog Converter.

2.2 Main challenges

Once the possibility to integrate per-pixel timing capability has been demonstrated, the research interest in SPAD-based CMOS d-ToF sensors moved toward the investigation of techniques aimed at reducing the negative impact of background light, as it represents one of the biggest performance limiting factor of such systems.

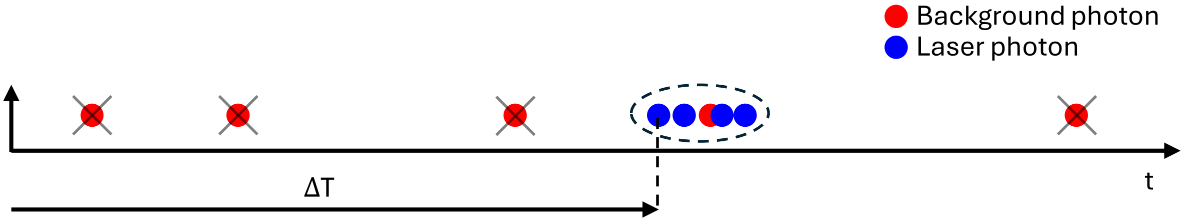
One of the most effective and widespread techniques which can effectively reduce the negative impact of background light is known as photon coincidence technique. With the photon coincidence technique, the impact of background photons is reduced by exploiting their temporal sparseness as measured at the output of several detectors. This technique requires the use of at least two SPADs, with a dedicated circuitry that is able to provide a trigger when at least two photons are detected within a temporal frame called coincidence window. In that case, the event is considered as part of the foreground, and a timestamp is extracted to measure the distance. Conversely, the isolated photons are considered background, and discarded. The first attempt to implement this detection strategy can be found in the work from Niclass et al. [42], where the technique is implemented in the analog domain. However, the full potential of photon coincidence detection is revealed in another work from the same authors [43], where the output of 12 neighboring SPADs is combined into a digital adder, achieving 3D images with 340×96 pixels resolution for a maximum range of 100 m under 80 klux of background illumination intensity. While that detector is designed to work in a scanning system (thus relying on a collimated laser source), a novel work has been presented by Perenzoni et al. [10] with a 64 × 64 pixel array designed for a more challenging flood illumination of the scene. Background rejection capability is guaranteed by the presence of 8 SPADs per pixel, building a digital-silicon photomultiplier (d-SiPM) with digital logic to exploit temporal proximity of photon detections. The reported maximum range is 300 m with centimeter precision under background intensity of up to 100 MPhotons/s/pixel. In the literature, several implementations of this technique have been proposed [4, 6, 10, 43], and by properly balancing the number of SPADs, the detection threshold and the detection time window, it is possible to adapt to a wide range of background light intensities.

Another technique is known as time-gating, where by means of a search procedure several sub-ranges of the scene are measured, increasing the SNR at the expense of an increased acquisition time. With this technique, the SPAD is enabled only during the current sub-range, thus neglecting any other previous and following photon detections [6,44,45]. The efficacy of this technique has been demonstrated of particular interest for specific applications like Fluorescence Lifetime Imaging (FLIM) [13, 14, 46–50] and Raman spectroscopy [51–53], where the measurement range is limited to some tens of nanoseconds and a reliable guess of the localization in time of the phenomena of interest is available. Concerning ranging applications, however, the benefit of this technique in terms of SNR increase is counter-balanced by the negative impact on measurement complexity and integration time, as the range is typically in the order of microseconds and no previous knowledge of an approximate target location is available to properly address the gating mechanism. A descriptive representation of both the photon coincidence and time gating techniques is shown in Figure 2.1.

A more recent approach which proved to be effective in mitigating the impact of background light is represented by the use of retriggerable, multi-event TDCs in combination with multi-event histogramming techniques [54,55]. Thanks to the possibility to acquire more than one photon per laser shot, the probability to capture signal photons even under high background levels is increased. Effectively, the negative impact of background light is mitigated without the need to actively reject photons. An even more advanced implementation, making use of the recently presented Synchronous Summation Technique (SST) by Patanwala et al. [56], in combination with a multi-event TDC, proved to outperform other pulse combining techniques which are commonly used in coincidence detection implementations.

In many scenarios, it is expected to have several systems operating together, thus encompassing the possibility of mutual system-to-system interference. For such a rela-

(a) – photon coincidence



(b) – time gating

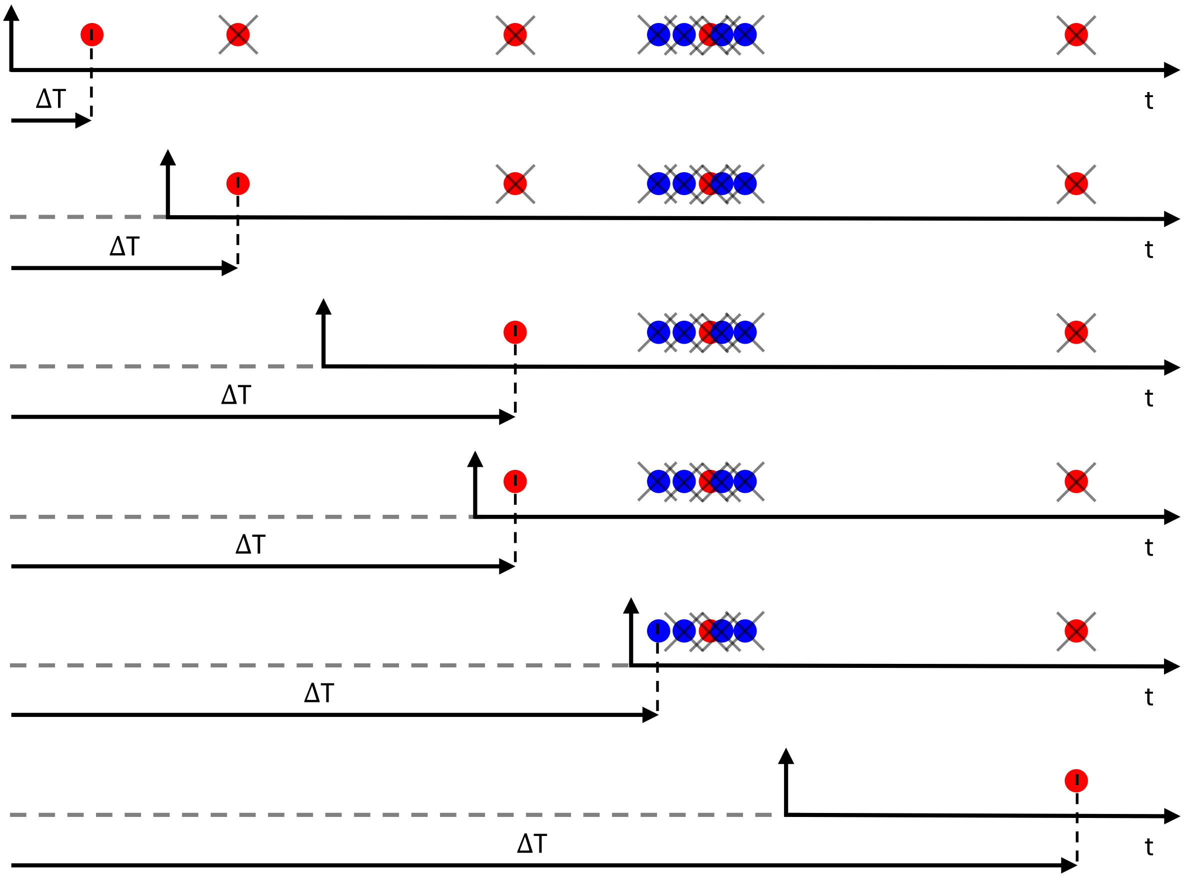


Figure 2.1: Example representation of the photon coincidence (a) and time gating (b) techniques. In (a), background photons can be discarded thanks to their temporal sparseness. Exploiting the temporal proximity of photons arrival times result in the increase of probability of detection of the correct arrival time, representative of the target ToF. In (b), a time gated detection approach is shown. Differently from the photon coincidence, the first photon is always considered and timestamped. By shifting in time the starting instant of the acquisition window, it is possible to detect laser photons even in presence of strong background light. After every temporal shift, the SNR increases, with a direct benefit on the probability of detection of the laser photons. As opposed to the photon coincidence technique, however, the acquisition process requires longer time, as several iterations are needed to span the entire acquisition window.

tively young challenge, already several solutions have been proposed in the literature. In the work from Ximenes et al. [8], interference is attenuated by means of a Phase-Shift Keying (PSK) modulation of the laser pulse emission time. A similar solution is presented in the work by Grollius et al. [57], where an adaptive Pulse Position Modulation (PPM) based on a random generator is used. With both approaches the laser emission time is randomized in time and any other source of interference with a different phase modulation is spread in the histogram of timestamps below the level of the signal of interest. Another technique for interference rejection is based on the concept of Code Division Multiple Access (CDMA). With this technique, at least two laser pulses are emitted with a known timing signature, which differs from device to device and shall be recognized to collect the signal of interest. This technique has first been implemented on analog devices with low integration capability [58] or on indirect time-of-flight systems (i-ToF) using either commercial modules [59] or bulky experimental setups based on avalanche photodiodes (APD) [60]. Recent integration of a CDMA interference rejection architecture in SPAD-based d-ToF systems in standard CMOS process has been proposed by Seo et al. [7], where the coding information of two successive laser pulses is exploited in the histogram of timestamps. In further recent developments, the capability to detect the laser timing signature has been integrated directly in-pixel, thus avoiding the need to transmit potentially information-less data and build a histogram of timestamps with highly reliable data [61]. With a PSK modulation of the emitted laser pulse [8], interference is not eliminated, but only attenuated, where the efficiency of attenuation depends on the number of phase shifts of the modulation. On the other hand, it has the advantage of requiring only one laser pulse, as opposed to the CDMA-based approach [7,61], where (at least) two laser pulses with a known timing signature are emitted. This approach, however, allows for a complete cancellation of interfering sources, as unknown detections are eliminated rather than attenuated. An example of behavior of both PSK/PPM and CDMA-based approaches is shown in Figure 2.2.

2.3 Integration challenges

In the recent years, the researcher's interest in SPAD-based d-ToF systems focused on improving the performance in mainly two directions. First, they focused on the possibility to integrate as much functionality as possible at pixel level, in particular concerning the timing measurement capability (TDC or TAC). Then, huge developments have been obtained in terms of resistance against external factors such as background light and interferences. In view of this, the maturity of such systems and the reliability

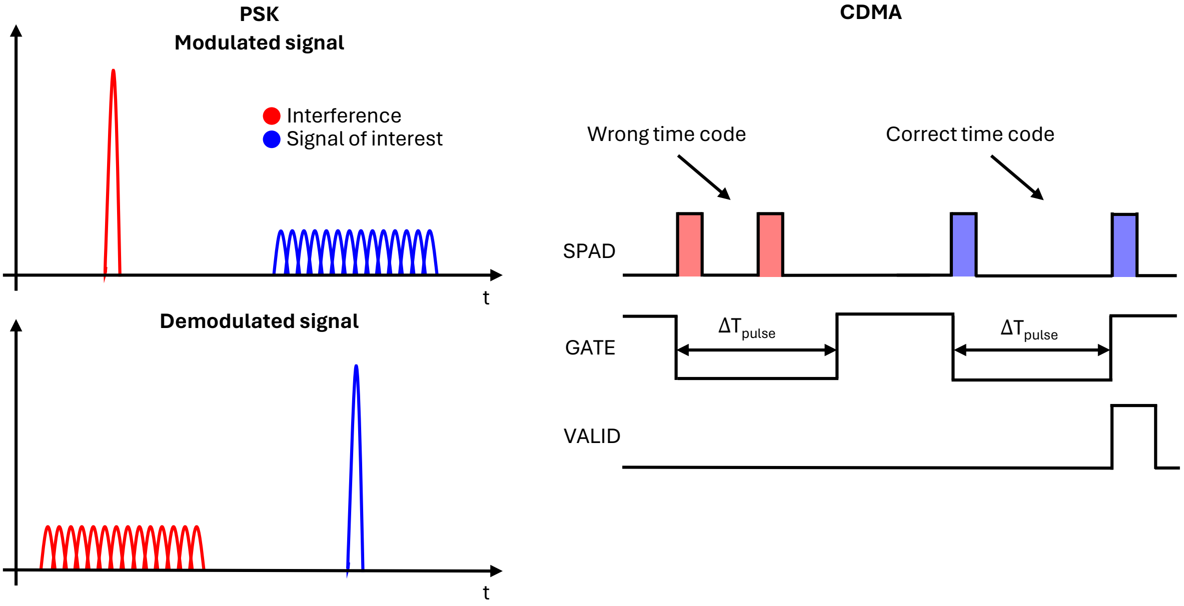


Figure 2.2: Comparison between PSK/PPM (left) and CDMA (right) approaches to reduce the impact of interferences in SPAD-based LiDAR systems. The main advantage of the PSK/PPM approach is that full laser power can be employed, but the interference is detected anyway and only with a sufficient modulation depth it can be attenuated below the level of the signal of interest. With a CDMA approach, a laser power penalty occurs as at least two laser pulses per acquisition are required. On the other hand, interference is actively eliminated, resulting in a better SNR with respect to the PSK/PPM approach and reduced amount of data to be transferred.

demonstrated through successful integration into commercial CMOS processes started to attract the interest from several market segments, from industrial [62], automotive [63], space [64] and consumer applications, asking for increased imaging, ranging and depth resolutions. As a natural consequence, the amount of generated data has started to hinder the development of such systems, making it necessary to investigate new solutions aimed at reconciling resolution and speed of operation.

The main strategy which was devised by many researchers to reduce the bandwidth requirement on the amount of data which is transferred from the chip to the controller (usually a Field Programmable Gate Array (FPGA) or micro controller (μC)) is to integrate the histogram, or part of it, directly on chip [9, 55, 65–72]. Two approaches have been proposed, which can integrate either a portion of the histogram (*partial* approach) or the entire histogram (*full* approach) on silicon.

With the so-called *partial* approach [7,66–68,70–73], a reduced histogram memory is available on-chip, therefore requiring a search procedure to identify the location of the ensemble of histogram bins containing the laser peak. In the literature, two techniques

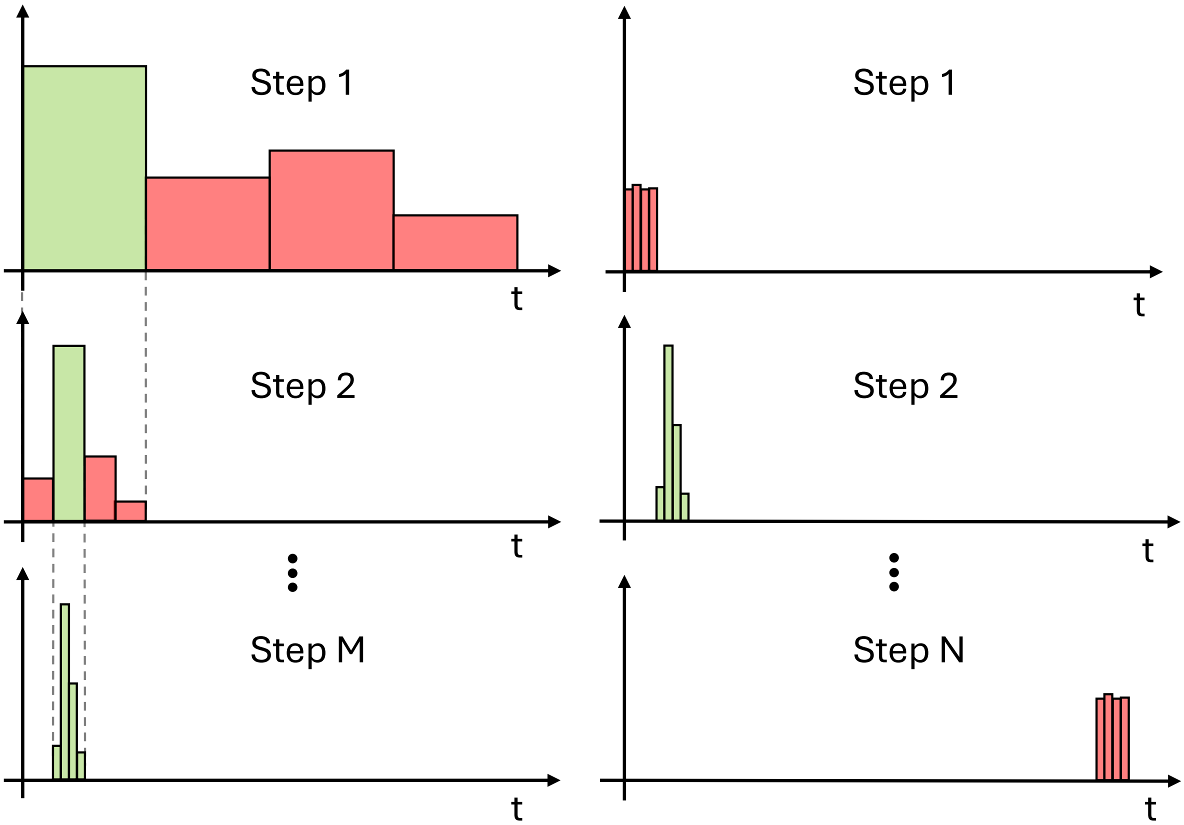


Figure 2.3: Comparison between *zooming* (left) and *sliding* (right) partial histogramming approaches.

have been described to implement a *partial* histogram behavior. With the so-called *zooming* technique [7, 66, 67, 71], at the beginning of the measurement the reduced set of histogram bins is spread across the entire distance range. By counting the number of photons detected on each bin, the reduced set of histogram bins is concentrated over several iterations on a shorter range, thus achieving the desired resolution on the estimated target distance. With the other technique, called *sliding*, the subset of histogram bins is already set to the desired resolution, thus covering only a small portion of the range. Again by means of several iterations, the subset of histogram bins slides across the entire range, and the number of photons at each iteration is used to estimate the target distance. A graphical representation of the main differences between a *zooming* and *sliding* partial histogramming approaches is shown in Figure 2.3. Finally, a *full* histogram approach has been demonstrated with resource sharing and the use of a 3D-stacked design approach by reallocating the same histogram circuitry to several pixels, as described by Kumagai et al. [69].

In the literature, other techniques to reduce the amount of data generated and handled by SPAD-based d-ToF systems have been proposed. Ingle et al. [74] propose a novel approach based on race logic which can provide an equi-depth histogram where the photon counts from several SPAD detections are divided into quantiles (in particular, the median value is used). By cascading several binner circuits, it is possible to reconstruct the equi-depth histogram of photon arrival times from which the *ToF* is extracted. This approach, however, has been only tested with simulated data and no physical realization of the binner circuit implemented with race logic has been provided. In the work from Gutierrez-Barragan et al. [75], an on-the-fly compressive histogram approach is presented. In this work, the amount of memory reduction is the same as for partial histogram approaches with the advantage of ensuring a better accuracy in the localization of the histogram peak. However, it requires a high in-pixel compute effort, since several multiply-add operations must be executed in real-time for each new timestamp. Sheehan et al. [76] propose a sketching approach based on the Fourier transform, which does not scale with either the number of photons or the timestamp resolution and does not effectively build a histogram of timestamps. However, this approach, which also was not validated with hardware, requires the real-time computation of the Fourier features which is hardly integrable at the pixel level.

Chapter 3

Numerical modelling of SPAD-based CMOS d-ToF sensors

Direct Time-of-flight (d-ToF) systems extract the distance estimating the time the light takes to travel from a pulsed emitter to the target and then back to a time-resolved photon detector. Such systems can build highly accurate and precise 3D images at a relatively low cost using solid-state image sensors in Complementary Metal-Oxide Semiconductor (CMOS) technology. Recent advances in Single-Photon Avalanche Diode (SPAD) detectors in CMOS technology show that high-sensitivity, low-noise devices can be combined with dense logic in small areas taking advantage of advanced process nodes, down to 40 nm [77], even in combination with 3D stacking [55], and the race towards a fully autonomous car driving system has raised an enormous interest in SPAD-based LiDAR technology.

Figure 3.1 shows an example structure of a SPAD-based d-ToF system, which includes the pulsed light emitter, the time-resolved photon detector, the optical elements to illuminate the target and collect the light echo, the target to be measured and the environment in which the target is placed.

The core of such a 3D imaging system is represented by a SPAD-based CMOS image sensor where each pixel is able to extract a distance measurement from the time elapsed between the emission of a laser pulse and the detection of the reflected light. Each application, however, has its own set of specifications which leads to the design of very different systems concerning both the architecture of the CMOS sensor and the set of external components such as optical elements and laser source. The performance of such systems is not easily predictable, in particular when the design of the single pixel is not trivial and aims at introducing smart features like background rejection [10, 43, 78–

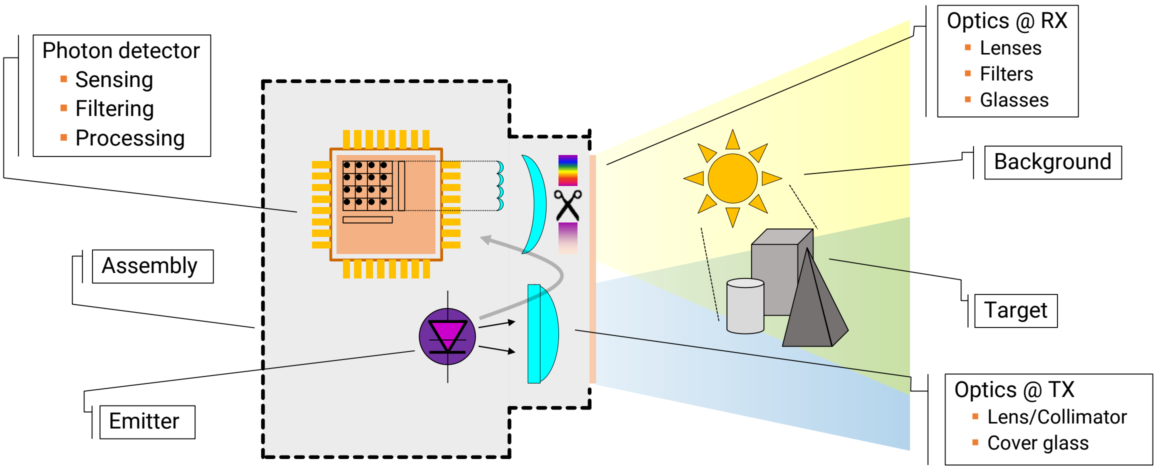


Figure 3.1: Block diagram of a d-ToF system.

80], interference suppression [58] or partial on-chip memory/histogramming [66, 81–83]. Additionally, expensive solutions like 3D stacking [8, 55, 77] and microlenses [84] are to be carefully evaluated before being employed. Consequently, it is desirable to tailor the design to meet the required specifications avoiding the risk to obtain either poor or overkilling performance. For the aforesaid motivations, the availability of a system simulator to help the preliminary design phase is of great importance [85], [86]. In this chapter, we present the design of a Monte Carlo simulator with the aim to emulate the output of a d-ToF SPAD-based CMOS flash LiDAR sensor. The model is based both on analytical equations modeling the optical concept and the target topology including the background light and a numerical Monte Carlo engine that generates Poissonian events modeling the impinging photons. Starting from the timestamps of the generated photons, many pixel structures can be simulated giving the designers the freedom to foresee the most suited architecture to fullfil the requested specifications.

3.1 Simulator architecture overview

A block diagram representing the simulator structure with its main components is reported in Figure 3.2. The simulator models the flash LiDAR system by generating random arrival times of detected photons. The first step toward the simulation of a single ToF measurement is the computation of the optical power budget, which mainly depends on the physical properties of the system and the environment: optics (a lens with a given $F_{\#}$, a certain transmittance τ_{opt} and a bandpass filter with bandwidth λ_{filt}), emitter (with a given power envelope $P(t)$, a wavelength λ_e and a beam divergence θ_e), target features (assuming Lambertian surface with reflectivity ρ_{target}) and spectrum of background light $E_{e,\lambda}$. Through the computation of the optical power budget it is possible to obtain the optical flux on the array focal plane. Starting from the knowledge of the optical flux, it is possible to generate the time envelope of detected events. In particular, photons can be modeled as a Poisson process, where the number of events in a time T follows a Poisson distribution, and the inter-arrival times are exponentially distributed. According to the number of detectors per pixel, a train of SPAD detection events is therefore created, which will be processed by the pixel architecture. The SPAD detector is identified by its fundamental parameters: photon detection probability (PDP), dark count rate (DCR) and, according to the quenching scheme, a certain dead-time T_{dead}. The distance information is retrieved by the pixel through a timestamping circuit, usually a time-to-digital converter (TDC) [87] or a time-to-amplitude converter (TAC) [88], that measures the timing information of the impinging photons. However, other features are usually implemented to reduce the impact of background light such as time gating, photon correlation in time domain and multiple laser shots within the same acquisition window to increase the frame rate. Then, the Monte Carlo engine will produce a single ToF value by processing the incoming train of photons detected by the pixel structure under investigation. Eventually, a histogram of ToF values is created and a distance extraction algorithm is applied to produce the final distance measurement.

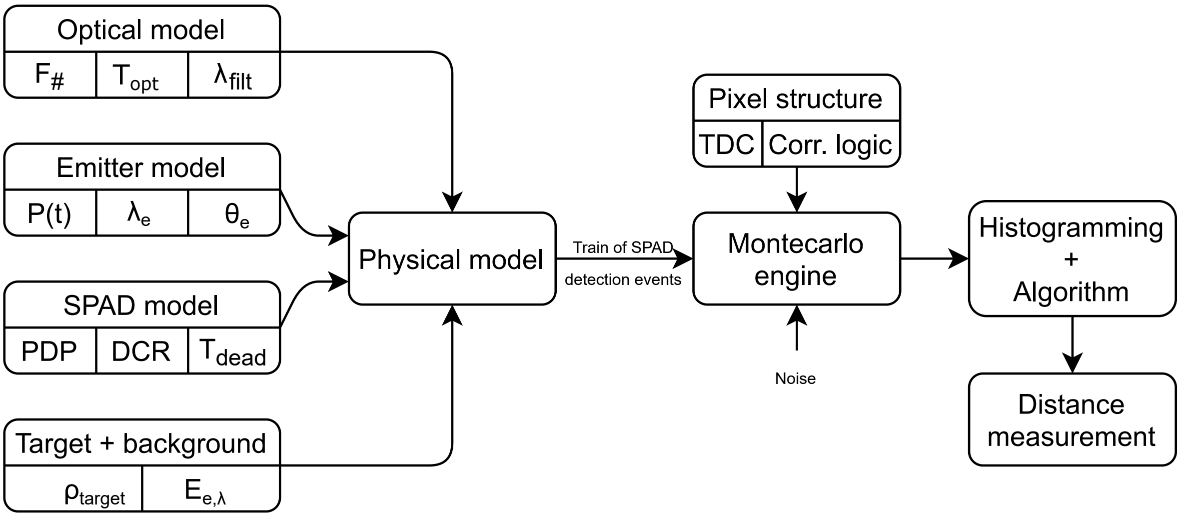


Figure 3.2: Numerical model architecture and main building blocks.

3.2 System features modeling

The aim of this Section is to provide details and mathematical background regarding the modeling of the building blocks of a d-ToF LiDAR system. First, the optical system is modeled to provide analytical expressions for the computation of the impinging optical power on each detector pixel. Then, the modeling of the illumination source is investigated and a methodology is proposed to reconstruct the laser emission power envelope.

3.2.1 Optical model

The amount of optical power impinging on the SPAD-based detector is due to the contribution of the illumination source, background light, target reflectivity and target distance. The optical model is developed considering the computation of the optical power budget for a single pixel, which can be easily extended to the entire array upon the knowledge of the illumination pattern employed. The standard case considers the illumination pattern to be matched along the detector field of view in both horizontal and vertical direction. Different or more sophisticated approaches can be employed and easily taken into account in the optical model through the computation of the optical power density on the target surface. The developed optical model takes into account the following parameters:

- Target topology, assumed to be a Lambertian surface with a given reflectivity value.
- Illumination source, a laser with emitted pulse power P_{tx}, divergence θ_e and square or circular spot shape.
- Optical system, composed of a lens with f-number $F_{\#} = f_{lens}/d_{lens}$ and a field of view over the horizontal (FOV_H) and vertical (FOV_V) direction.
- Ambient optical power density due to the background light level.

The power reaching the pixel can be expressed as the amount of radiant exitance M_{lens} collected by the lens multiplied by the area which is observed by the pixel on the target side A_{scene}:

$$P_{pix} = M_{lens} \cdot A_{scene}, \tag{3.1}$$

where the term M_{lens} of (3.1) is computed considering the target to diffuse a certain radiant exitance M_{target} according to Lambert's cosine law. Supposing the target is observed with a lens subtending an angle 2α, the amount of radiant exitance which is collected is obtained from the integration in spherical coordinates:

$$\begin{aligned} M_{lens} &= M_{target} \cdot \frac{\int_0^{2\pi} \int_0^{\alpha} cos\theta sin\theta d\theta d\varphi}{\int_0^{2\pi} \int_0^{\pi/2} cos\theta sin\theta d\theta d\varphi} \\ &= M_{target} \cdot sin^2\alpha. \end{aligned} \tag{3.2}$$

The term M_{target} of (3.2) is due to the contribution of both the illumination source and the background light. Concerning the illumination source, we consider it to be at a certain distance z with a given power P_{tx} and a given divergence θ_e with a square or round spot shape. Concerning the background light, we consider the fraction of the sun irradiance spectrum $E_{e,\lambda}$ through an optical bandpass filter with bandwidth $[\lambda_2 - \lambda_1]$ which is placed on top of the sensor. Since the spectrum of the solar radiation is not constant, the amount of solar radiant flux which can be detected by the sensor is obtained by integration. Consequently, M_{target} equals to:

$$M_{target} = \begin{cases} \frac{P_{tx}}{4 \cdot (z \cdot tan(\theta_e/2))^2} & \text{square spot} \\ \frac{P_{tx}}{\pi \cdot (z \cdot tan(\theta_e/2))^2} & \text{circular spot} \\ \int_{\lambda_1}^{\lambda_2} E_{e,\lambda} d\lambda = P_{D,BG} & \text{background light} \end{cases} \tag{3.3}$$

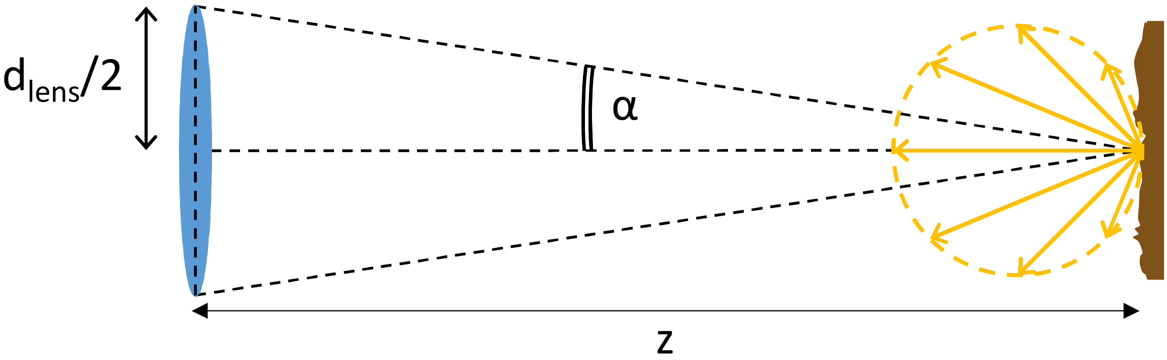


Figure 3.3: Optical system concept. The light belonging to both the laser emitter and the background is assumed to be diffusely scattered and observed within an aperture of 2α from the collecting lens.

From the analysis of the geometry of Figure 3.3, the term $sin^2\alpha$ of (3.2) can be expressed as a function of the lens diameter d_{lens} and the target distance z as:

$$sin^2\alpha = \frac{d_{lens}^2}{4z^2 + d_{lens}^2}. \tag{3.4}$$

The radiant exitance collected by the lens is then given by:

$$M_{lens} = M_{target} \cdot \frac{d_{lens}^2}{4z^2 + d_{lens}^2}. \tag{3.5}$$

The term A_{scene} of (3.1) can be computed considering the lens to project the area of a $M \times N$ array over the target with an aperture angle given by the lens field of view FOV over the horizontal (FOV_H) and the vertical (FOV_V) direction:

$$A_{scene} = \frac{2z \cdot tan(FOV_H/2)}{M} \cdot \frac{2z \cdot tan(FOV_V/2)}{N}. \tag{3.6}$$

Equation 3.6 can be rearranged to show the pixel area A_{pix} instead of the array dimensions considering that the lens field of view can be expressed as:

$$FOV = \begin{cases} 2 \cdot arctan(M \cdot Pitch_H/2f_{lens}) & \text{horizontal} \\ 2 \cdot arctan(N \cdot Pitch_V/2f_{lens}) & \text{vertical} \end{cases} \tag{3.7}$$

The terms $Pitch_H$ and $Pitch_V$ refer to the physical size of the pixel. By substituting each contribution of the field of view from (3.7) into (3.6), we obtain the following expression for the term A_{scene}:

$$\begin{aligned} A_{scene} &= Pitch_H \cdot Pitch_V \cdot \left(\frac{z}{f_{lens}}\right)^2 \\ &= A_{pix} \cdot \left(\frac{z}{f_{lens}}\right)^2, \end{aligned} \tag{3.8}$$

which is no more dependent on the array size. Consequently, the computation of the optical budget for each single pixel is valid until the pixel is completely contained within the lens focal plane. While the array size is meant to properly evaluate the impinging power, the simulator works on a single pixel basis: on the single pixel, all the different conditions of the optical power budget can be set, such as different reflectivity, distance, background, optics parameters, etc. The amount of optical power impinging on the pixel is reduced considering the lens transmittance τ_{opt} and the target reflectivity ρ_{target}, the latter being supposed uniform over the portion of target area which is subtended by the pixel. Alternatively, to model a non-uniform target surface, it is possible to consider an average reflectivity value $\overline{\rho_{target}}$, which is observed by the pixel projection over the target surface. Furthermore, the amount of optical power impinging on the SPAD surface is reduced considering the pixel fill-factor FF. Considering the illumination source with a circular spot shape we obtain the following expression:

$$P_{pix,source} = \frac{\tau_{opt} \cdot \rho_{target} \cdot FF \cdot P_{tx} \cdot A_{pix}}{\pi \cdot F_{\#}^2 \cdot tan^2(\theta_e/2) \cdot (4z^2 + d_{lens}^2)}. \tag{3.9}$$

Considering the background light:

$$P_{pix,bg} = \frac{\tau_{opt} \cdot \rho_{target} \cdot FF \cdot P_{D,BG} \cdot A_{pix} \cdot z^2}{F_{\#}^2 \cdot (4z^2 + d_{lens}^2)}. \tag{3.10}$$

Regarding the background light, it is possible to obtain more accurate results by converting the amount of power belonging to each lambda component of the solar radiation spectrum at sea level. With the knowledge of the impinging optical power, it is possible to compute the average photon rate, i.e. the λ parameter of the Poisson process used to statistically model the photon arrival times. The average photon rate is obtained by dividing the power by the photon energy, given by $E_p = \frac{h \cdot c}{\lambda_c}$, where h is the Planck's constant, c is the speed of light (assumed to be in vacuum) and λ_c is the laser central wavelength. The average photon rate is translated into an average event rate at the output of the SPAD detector considering two additional parameters: the SPAD PDP and the SPAD dead-time. In particular, the SPAD dead-time sets an upper limit to the maximum achievable event detection rate which may be orders of magnitude lower than the impinging photon detection rate. The effect of SPAD dead-time have been included and extensively discussed in Section 3.3.

The presence of a scattering medium and its implications on both the 2D intensity image and the 3D depth map has not been considered in this analysis. Regarding the 3D depth map, it can be partly included as an attenuation factor [89] in (3.9) and (3.10) and as a spread in time of the received laser echo [90]. The potential blurring effect on the 2D image, however, can not be included as the simulator works considering a single pixel.

3.2.2 Illumination source - modeling of the laser emission profile

At a first approximation, it is possible to consider the illumination source to provide a constant power during the pulse time interval. More realistic results, however, can be obtained by modelling the laser illumination source to provide the true time envelope of the photons emission time. The underlying inhomogeneous Poisson process has been simulated starting from the laser power emission profile (which can be either provided with the emitter datasheet or measured with a conventional Time-Correlated

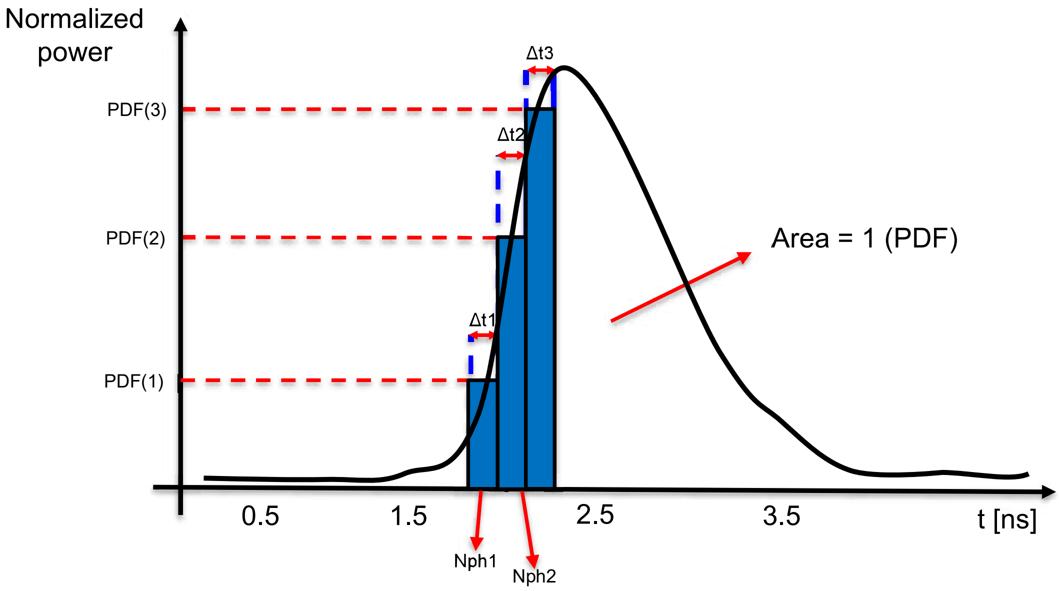


Figure 3.4: The area of the laser power envelope represents the emitted pulse energy, which is modeled as a probability density function giving the integrated photon flux for each discretized time bin.

Single Photon Counting (TCSPC) setup). Figure 3.4 shows an example of the temporal evolution of the power of a laser emitter as a function of time.

The area of this curve represents the amount of energy emitted per single laser pulse, E, which can be turned into the total number of emitted photons: $N_{ph} = \frac{E \cdot \lambda}{h \cdot c}$. The curve has to be normalized to have unitary area, consequently being considered as a probability density function (PDF). Then, the obtained PDF is discretized into N bins, each of them being sufficiently narrow (in the order of some tens of picoseconds) to have sufficient granularity in the reconstruction process. For each bin, we consider the source to emit photons with a rate:

$$RATE_i = \frac{PDF(i)}{N_{ph}}. \tag{3.11}$$

Then, we generate one random photon timestamp according to an exponential distribution and we consider it valid if it falls into the corresponding time bin, otherwise it is discarded. By iterating this procedure for a high number of trials, it is possible to reconstruct the laser pulse time envelope. Since we are modeling the photon detection process of a SPAD, it is important to take into account the pile-up effect due to the detector dead time. Being the laser pulse width typically shorter than the SPAD dead time, the true time envelope can be reconstructed if pile-up is avoided, thus being in the single photon regime. Figure 3.5 shows an example of the experiment run for different

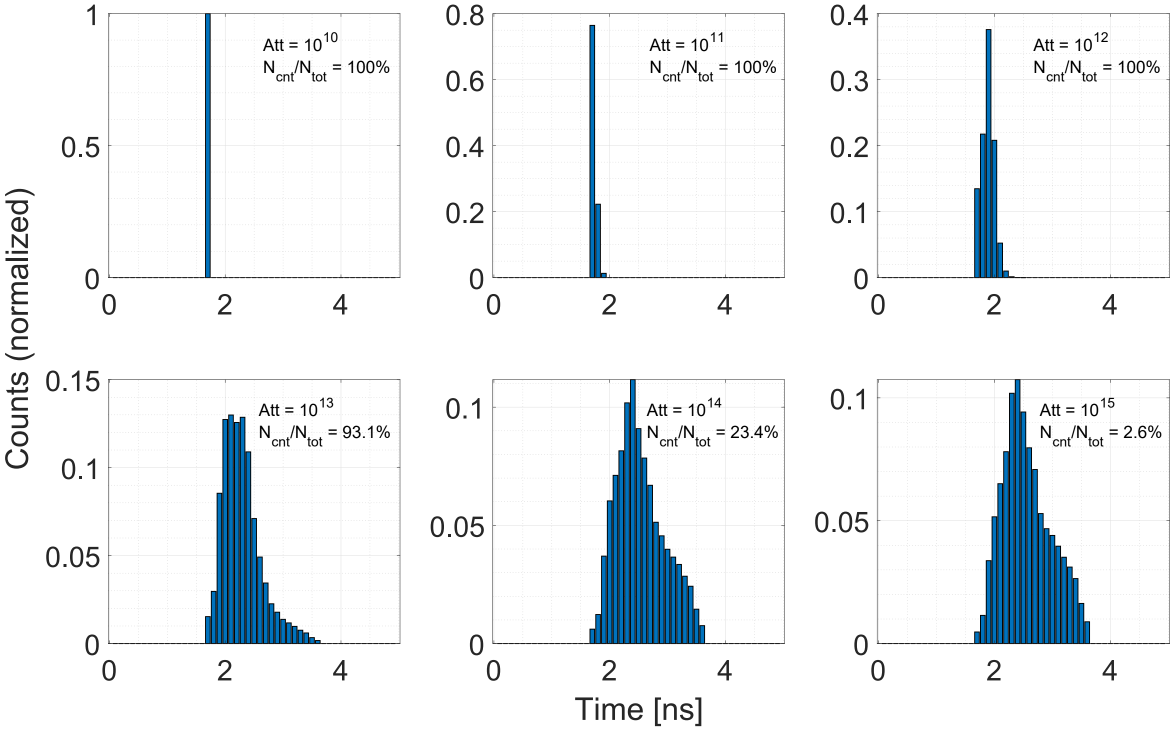


Figure 3.5: Example of the reconstruction of a laser time envelope. For each sub-figure, the attenuation value and the ratio between detected events and total trials (triggering rate) is reported. Up to $Att = 10^{13}$, the reconstructed time envelope is heavily distorted by pile-up effect, where all timestamps are compressed into few time bins. For higher attenuation values, the pile-up distortion becomes negligible and it is possible to recover the true time envelope.

power attenuation values. The number of histogram counts have been normalized to better show the pile-up effect, where most of the recorded timestamps are compressed toward early values. The attenuation value is calculated as the ratio between the total emitted power and the fraction of power returning back to the pixel (as specified in Section 3.2.1), and the intensity of pile-up effect is reported for each sub-figure as the average number of detections over the total number of trials. The simulator is capable of properly reproducing a variety of detection scenarios, spanning from strong pile-up to photon starved conditions.

The considerations regarding the pile-up effect and the power attenuation values we made in this section are oriented to demonstrate that the reconstruction method we proposed is able to generate the true laser pulse shape. In a real scenario, however, it is not always possible to recover the true laser pulse envelope. Depending on the distance extraction algorithm that is employed, the amount of distortion of the recorded laser pulse echo may introduce systematic errors in the distance detection process. This is of particular importance when distance extraction algorithms based on histogram peak detection are used. Referring to Figure 3.5, a simple peak detection algorithm would provide a detection at $t \simeq 1.7ns$ with the lowest attenuation value of 10^{10}, while with the highest attenuation value 10^{15} the detection would be located at $t \simeq 2.5ns$, resulting in a distance error between the two cases of $\simeq 0.12m$.

3.3 Monte Carlo simulation

In this section, we first provide a description of the methodology used to efficiently generate timestamps from a Poisson process. Then, a discussion on how to time-compensate the generated timestamps to meet a precise detection paradigm is provided. Eventually we discuss important aspects to properly evaluate the results of the simulation process.

3.3.1 Generation of SPAD-related events

When the intensity of the photon flux is known it is possible to generate a train of SPAD detection events from the Poisson process. The Poisson approximation is considered valid since the SPAD dead-time T_d is much longer than the coherence time τ_c of the light source, which can not therefore be resolved [91]. The photon interarrival times are generated from the Poisson process (of parameter λ) considering the probability that the first arrival time T is greater than t, which is given by:

$$\begin{aligned} P(T > t) &= P(N(t) = 0) \\ &= e^{-\lambda \cdot t}, \end{aligned} \tag{3.12}$$

where $P(N(t) = 0)$ expresses the probability of observing zero events in the interval $[0, t]$, derived from the more general formulation:

$$P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}. \tag{3.13}$$

The interarrival time t is obtained inverting (3.12):

$$t = \frac{-ln(P(T > t))}{\lambda}, \tag{3.14}$$

where $P(T > t)$ is replaced with a random number from the uniform distribution in the unit interval $\mathcal{U}((0, 1])$. SPAD detection events are then generated from the cumulative sum of N interarrival times, up to the desired simulation time. Those events originate from both photon detection mechanisms of illumination source and background light and the generation of dark counts from the SPAD itself. From the simulation perspective, and considering the additivity property of Poisson processes, it is possible to generate and superimpose two trains of photon arrival times. The first one contains the background timestamps, from both the environment illumination and dark count events of the detector, with a duration equal to the entire measurement window and a parameter $\lambda = \lambda_{bg} + \lambda_{dcr}$. The second one is needed for the photons belonging to the illumination source (with a duration equal to the laser pulse width) with a parameter $\lambda = \lambda_{src}$.

The photon arrival times generated from the Poisson process needs additional adjustments to account for the SPAD dead-time. This effect can be included by removing all events with a time difference smaller than the SPAD dead-time. This method, however, is not computationally efficient as it requires the generation of a certain number of photon arrival times which would be be eventually decimated. Additionally, it is not possible to vectorize the decimation process, thus the efficiency drops due to the need to employ a for-loop. A different method, which is computationally efficient and avoids the decimation process making use of vectorized software operations, can be employed. With this method we generate a maximum number N_{max} of detectable events (i.e. photons detected by the SPAD) within a time interval ΔT considering the dead time T_d:

$$N_{max} = \Delta T / T_d. \tag{3.15}$$

Then, a sequence of N_{max} photon arrival times is generated using the result found in (3.14):

$$t = \{t_0, t_1, ..., t_n\}, \quad n = N_{max} - 1 \tag{3.16}$$

Eventually, each photon arrival time is shifted in time to account for the SPAD dead-time as follows:

$$t_i \to t_i + i \cdot T_d\,, \tag{3.17}$$

where $i = \{0, 1, ..., N_{max} - 1\}$.

From a mathematical point of view, the proposed method achieves the same result one would obtain by considering the distribution of interarrival times in the case of a detector deadtime T_d:

$$e^{-\lambda\cdot(t-T_d)}, \quad t \geq T_d \tag{3.18}$$

In particular, (3.18) is valid from the second timestamp on, since (only) the first event can be detected before the detector deadtime T_d. Interarrival times are again obtained by inversion:

$$t = \frac{-ln(P(T > t)) + \lambda \cdot T_d}{\lambda}, \tag{3.19}$$

where also in this case, $P(T > t)$ is replaced with a random number from the uniform distribution in the unit interval $\mathcal{U}((0, 1])$. At this point, the simulator provides the timestamps of the photons detected by the SPAD detector. Usually, a time-to-digital converter (TDC) is used to measure the time elapsed between photon detections and provide digital codes that are collected into a histogram. TDCs are mainly characterized by two metrics: quantization error (dependent on the timing resolution) and linearity, in terms of differential nonlinearity (DNL) and integral nonlinearity (INL). The quantization error is included in the model with a rounding operation of the ratio between the photon timestamp and the TDC timing resolution. Regarding the linearity information, it is possible to include their effects upon the knowledge (or a prediction) of the TDC performance that will be used in the final system. In this work, we have included the effect of the differential nonlinearity by applying a non-uniform rounding to the generated timestamps, thus artificially increasing the probability to obtain an even rather than an odd digital code, as this is the most common scenario for a TDC based on a ring-oscillator, which is the most common structure employed at pixel-level. With the availability of a complete DNL characterization it is also possible to model the behavior of each single TDC code, tailoring the timestamping behavior to the actual implementation. Additional parameters modeling timing uncertainties are included as well. In particular, we consider a global jitter value representative of the entire measurement setup which is drawn from the normal distribution and is added to the final timing measurement.

3.3.2 Synchronous and asynchronous SPAD model

In SPAD-based d-ToF systems, there exists mainly two ways to drive the single-photon detectors and recover the target range information. The choice between the two SPAD driving criteria has to be made according to several parameters, such as the maximum expected range and the type of sensor (single-point or 2D imager). The first method, called *synchronous*, allows for easier and more compact implementation, but it is limited to short range applications. For long-range operations, the *asynchronous* method is usually employed. In this section, we provide a comparative evaluation between the two, focusing in particular on the *asynchronous* one, as it is the method of choice for long-range, challenging scenarios.

Synchronous SPAD

In the *synchronous* driving scheme, the SPAD is enabled synchronously with the start of the measurement window (typically in coincidence with the laser pulse emission), and the time of flight measurement is stopped with the first detected photon. Then, the SPAD is forced in an OFF state by the quenching circuit, and eventually recharged for the next measurement cycle. Such driving scheme is typically employed for short range applications with moderate background illumination. A pictorial representation of a *synchronous* SPAD driving scheme is reported in Figure 3.6-(a). This driving method has both advantages and disadvantages. Concerning the advantages, it allows for a simpler hardware implementation and low average power consumption, as independently of the intensity of photon flux only one photon (thus, only one avalanche) can be detected per laser cycle. Regarding hardware implementation, it enables the utilization of small transistors for actively recharging all SPADs through a rapid synchronous global signal. This stands in contrast to passive quenching implementations, which typically requires larger transistors (in terms of channel length L). Another benefit is the 100% sensitivity of the sensor at the beginning of the acquisition, as all SPADs are activated synchronously with the emission of the laser pulse. On the other hand, it is limited by several disadvantages. In case of very high background photon flux, it is more likely that the first photon detection belongs to a background photon, thus severely limiting the probability of detection of a laser photon. For the same reason, as the method allows for up to one detected photon per laser cycle, it is not compatible with strategies (as the photon coincidence technique) which aims at reducing the probability of being blinded by background light, as the SPAD can not be further recharged within the same laser cycle. Another disadvantage arises from the simultaneous activation of all SPADs, leading to a potential (temporally localized) surge in power consumption.

Asynchronous SPAD

In the *asynchronous* driving method each SPAD is driven in a free-running fashion, independently from the timing of the acquisition window, and after each photon detection

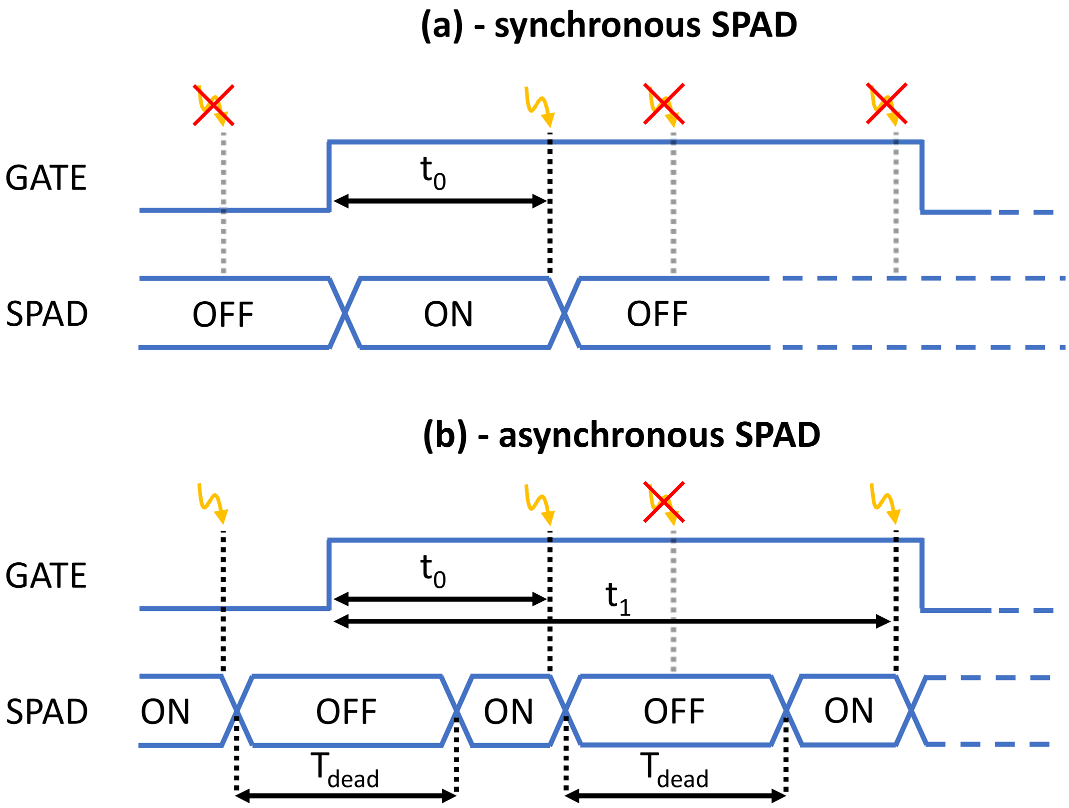


Figure 3.6: Graphical representation of a synchronous and asynchronous SPAD detection paradigm.

is automatically recharged (after a dead-time T_{dead}) allowing for a multitude of photons to be detected for each laser cycle. By employing this driving method, it is possible to effectively utilize background rejection schemes like the photon coincidence technique. In situations where a detection is likely caused by background light and needs to be discarded, having a rechargeable SPAD becomes essential. For such reasons, it is very well suited for long range LiDAR operation, where the laser echo could be potentially located at any time during the acquisition window. Another advantage, as opposed to the *synchronous* method, is that it is compatible with multi-event TDCs [54–56], allowing for on-chip histogramming capabilities. The main drawback of this driving method is the increased power consumption, as the SPADs are free-running and can detect photons even during the sensor readout phase (which shall be, in principle, avoided to save power). An example of asynchronous SPAD detection is depicted in Figure 3.6-(b).

Careful attention is required to properly simulate a train of SPAD detection events considering the *asynchronous* scheme. In particular, there exist two scenarios with this method. In the first scenario, the acquisition starts when the SPAD is active, i.e., not in dead-time. In this case, the generation procedure described in Section 3.3.1 can be used as it is. In the opposite scenario, the acquisition starts when the SPAD is within its dead-time period, as a photon detection occurred before. In this case, the time of arrival of the first useful photon (which is exponentially distributed) would be delayed for the amount of remaining dead-time period. In this case, then, the simulation of the train of SPAD detection events would follow the same rules as described in Section 3.3.1, time-shifted by a correction offset which is drawn from a uniform distribution $\mathcal{U}([0, T_{dead}])$. Whether to add or not the correction offset depends on the probability that the SPAD is within its dead-time period, which depends on the combination between the dead-time itself (T_{dead}) and the magnitude of the background photon flux λ_B. This probability can be computed as the ratio between the amount of time T_{OFF} during which the SPAD is within its dead-time period (which coincides with T_{dead}), and the average rate of events in the *asynchronous* scheme, which equals to $\overline{T} = \frac{1}{\lambda_B} + T_{dead}$. Then, the probability that the SPAD is within its dead-time period at a given time instant is given by:

$$\begin{aligned} P_{OFF} &= \frac{T_{OFF}}{\overline{T}} \\ &= \frac{\lambda_B T_{dead}}{1 + \lambda_B T_{dead}} \end{aligned} \tag{3.20}$$

Hybrid method

For the sake of discussion, and to make a clearer distinction between the *synchronous* and *asynchronous* SPAD driving methods, we discuss a further one, which we call *hybrid* method. This third method has little or even no practical use in real-world cases; however, the discussion helps provide a deep understanding of the differences between the two former methods.

The main point with an *asynchronous* driving method is the (asynchronous) timing relationship between the always-on, free-running SPAD (whose timing is dictated by the impinging flux of photons) and the regular time-base which drives the sensor acquisition phases. Therefore, the SPAD is always active, even during such sensor phases where it is not strictly necessary to detect photons. A plausible inference suggests that the SPAD should be driven in a free-running fashion only within each laser cycle, i.e., it can be synchronously enabled by the front-end circuitry and then let in free running for the whole duration of the acquisition window (where the laser echo return is expected to be). In such a case, power consumption can be reduced, as the SPAD would be kept OFF during other sensor acquisition phases (as during data readout) where there is no interest in detecting photons. This approach, which we call *hybrid*, blends elements from both pure *synchronous* and *asynchronous* driving methods. However, it can not be employed due to a significant limitation. In this case, the SPAD would be activated at the beginning of the acquisition window and thus ready to detect a photon; then, it would eventually trigger and after the dead-time T_{dead} it would be recharged, thus being ready to detect another photon. If we consider, on a time basis, the probability that the SPAD is active (i.e., not in dead-time), we have an oscillating behavior as shown in Figure 3.7, which reports this probability for each time instant, in the hypothesis that the SPAD is enabled at time $t = 0$ and in presence of a constant flux of background light with different intensities. As shown, this probability has an oscillating behavior which eventually sets. The magnitude of oscillation and the settling time depends upon the incoming photon flux. In case of a high photon flux, as it would be easily expected for an outdoor automotive scenario, the settling time is in the range of few microseconds, where, with a periodicity equal to the dead-time, the SPAD would be periodically unable to detect further photons. This amount of time corresponds to a distance range of hundreds of meters, which is typical for automotive applications, where the SPAD would be periodically impaired. After the initial settling time, the SPAD eventually gathers a constant probability of being active for each time instant, making it suitable for reliable operation.

For this reason, only a pure *asynchronous* SPAD driving scheme can ensure a constant sensitivity over time and proper operation over the whole measurement range.

3.4 Experimental results

The model has been validated against a d-ToF system based on the detector described in [10]. Distance measurements have been carried out indoor in controlled conditions. The simulation parameters, reported in Table 3.1, have been tuned to match the characteristics of the detector together with the selected optics and laser source. The laser pulse shape has been estimated using a TCSPC setup to obtain the temporal envelope of the emitted optical power, as shown in Figure 3.8. The total emitted energy of the laser pulse has been estimated using a fast photodiode [92].

The Monte Carlo simulator implements the pixel structure of the detector of choice, which exploits temporal correlation between photon detections to improve background rejection. The digital Silicon Photomultiplier (d-SiPM) structure has been accurately modeled to include the effect of bandwidth limitation due to the OR-tree pulse combiner [10], together with the smart triggering logic that generates the rolling time window to detect temporal correlation within the stream of pulses. We also model the nonlinearity of the TDC. Due to the very low laser pulse energy available, distance

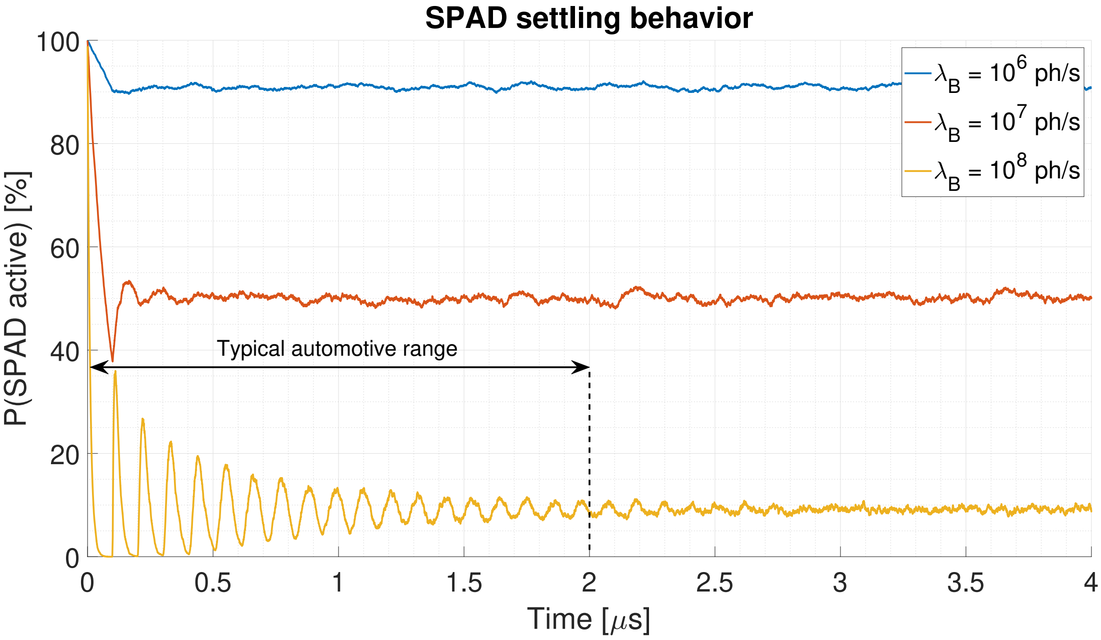


Figure 3.7: Settling behavior of the probability that a SPAD is active over time under three different photon flux intensity for the *hybrid* driving method. At time $t = 0$, the SPAD is turned on, and thus it is always active and ready to detect a photon. The probability becomes stable after an initial settling time, which depends on the photon flux intensity. The periodicity of the oscillations is given by the SPAD deadtime, which has been set to 100 ns. In the case of a very high background photon flux (e.g., $10^8 ph/s$), the SPAD would be periodically unable to see further photons during the settling time, severely impairing the possibility to detect the laser return echo during this timeframe.

measurements have been carried out within a maximum range of $\simeq$ 1.9 m using polytetrafluoroethylene (PTFE) tape as target with $\simeq$ 75 % Lambertian reflectivity [94]. A comparison that validates the power budget and the Monte Carlo photon generation model is shown in Figure 3.9. The first subplot reports the average number of detected pulses, which is a direct indication of the impinging optical flux. The second subplot reports the measurement precision, computed as the standard deviation of a collection of 250 distance measurements. Each distance measurement, obtained without photon temporal correlation, is extracted from a population (histogram) of 1000 single d-ToF values by applying a centroid-based algorithm [10] to both the measured data and simulated data.

A second comparison between measured and simulated data is provided considering background illumination to exploit the photon correlation logic. In some cases, the signal power level may be orders of magnitude lower with respect to the background power, as reported in Figure 3.10 where we show the relationship between signal and background for three different background power density on the target side of 1 W/m^2,

Table 3.1: Simulation parameters

Parameter	Value	Unit
Pixel		
PDP[†]	25	%
Fill factor	26.5	%
Pixel Area	3600	μm^2
Dead time	15	ns
Median DCR[‡]	6.8	kHz
TDC LSB	250	ps
Emitter		
Central wavelength	405	nm
Pulse energy	6.2	pJ
Pulse FWHM	$\simeq$250	ps
Beam divergence	$\simeq$1.7	°
Optical elements		
Filter bandwidth FWHM	10	nm
Transmittance	66	%
Focal length	6	mm
Diameter	5	mm
Environment		
Reflectivity	75	%
Background flux	$\simeq$90	MPh/s/pixel
System		
Nr. histogram points	1000	
Global jitter FWHM[††]	1500	ps

[†] Value from [93].
[‡] Value from [10].
[††] estimated from the entire measurement setup.

7 W/m^2 and 50 W/m^2. The possibility to exploit photon correlation to drastically reduce the intensity of the background noise is fundamental to recover the signal information. Figure 3.11 reports one histogram for both measured (a) and simulated (b) data, with a background flux of $\simeq$ 90 MPh/s/pixel (corresponding to a power density on the target of $\simeq$ 6.9 W/m^2) and no correlation between detected photons. Most common algorithms to extract the distance result from the histogram would fail because, as it can be noticed, the signal peak is below the noise level. The repetitive noise in the histogram is due to the DNL of the TDC, resulting in a systematic error between even and odd codes [10]. This effect has been included in the model by applying a non-uniform rounding to the generated timestamps before being transformed into TDC codes. By exploiting the temporal correlation between photons it is possible to reduce

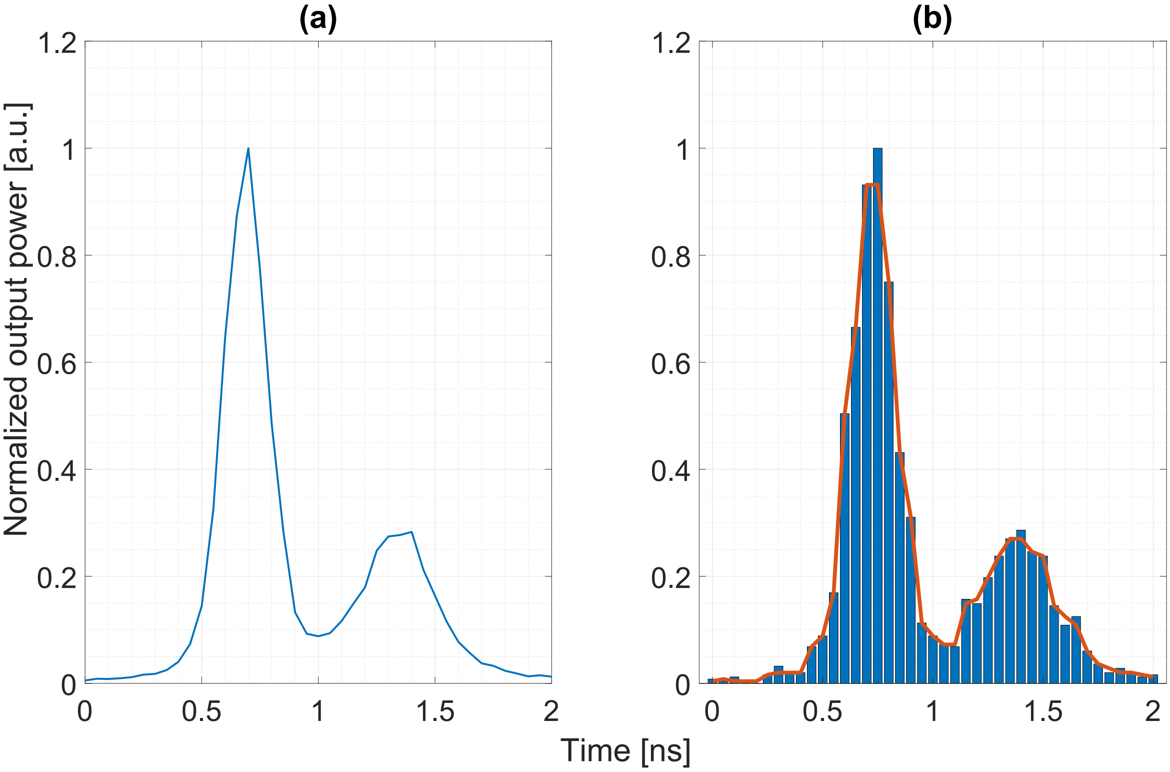


Figure 3.8: Comparison between the measured (a) and one realisation (b) of the simulated laser pulse envelope (250 ps FWHM, measured energy of $\simeq$ 6.2 pJ).

the intensity of the background noise and recover the signal peak. Figure 3.12 shows the resulting histogram after the temporal correlation of $N_{ph} = 2$ photons when a time window $T_{win} \simeq 2.3$ ns is applied.

3.5 Summary

A comprehensive, complete, Monte Carlo simulator for d-ToF systems has been described and validated against an existing setup. All the components of a complete ToF detection system can be modeled to provide a solution to foresee the final product in terms of both costs and achievable performance. Many different pixel structures and detection paradigms can be easily included in the simulation flow enabling the possibility to implement the most promising solution already in the design phase of the detector. The capability of the simulator has been proven against an existing setup, with good match between simulated and measured data.

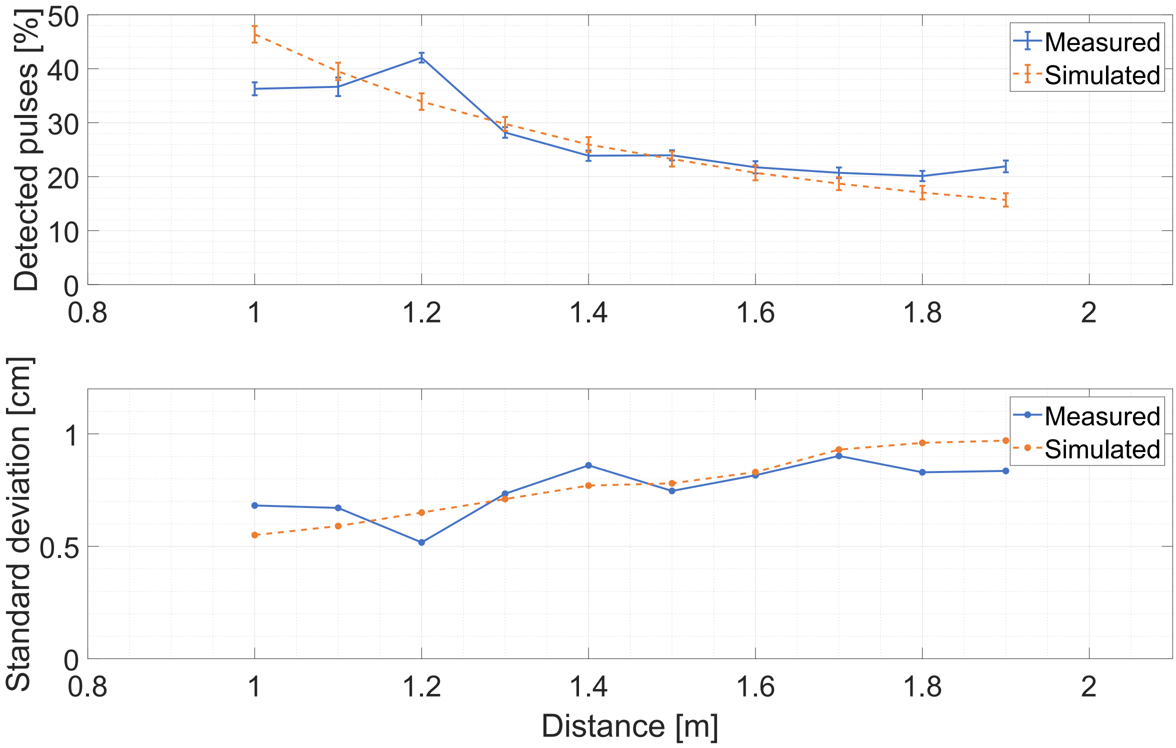


Figure 3.9: Comparison between measurements and simulations in terms of percentage of detected pulses and standard deviation.

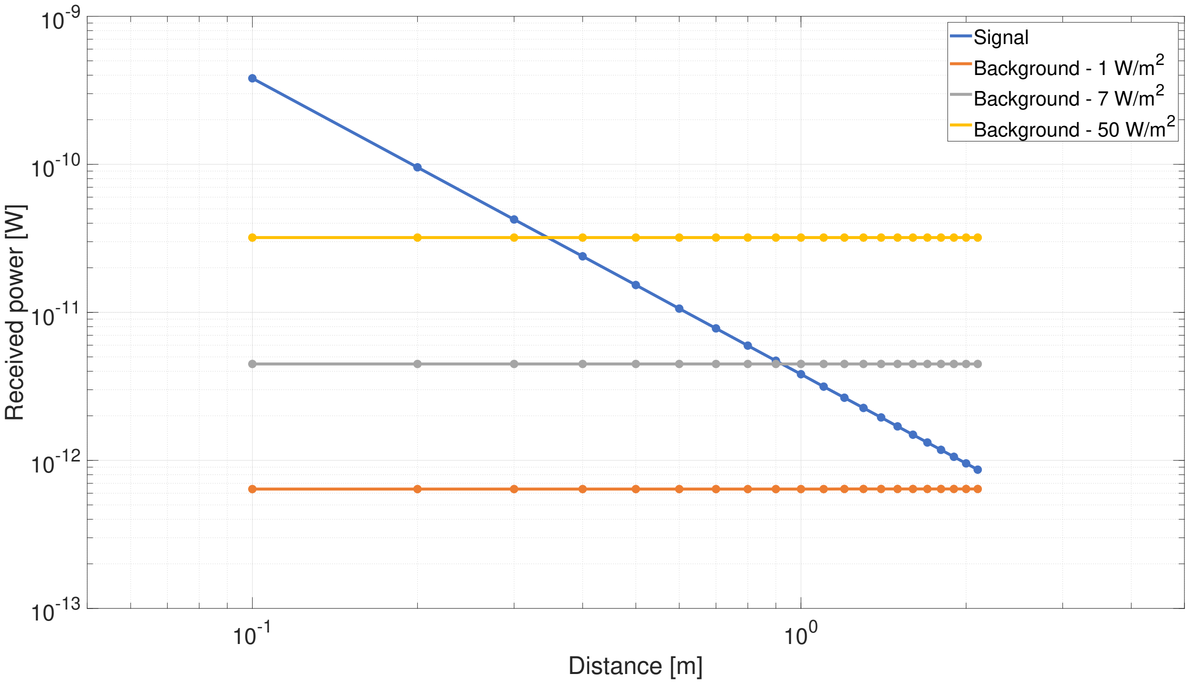


Figure 3.10: Comparison between the received optical power from signal and four different background illumination intensity.

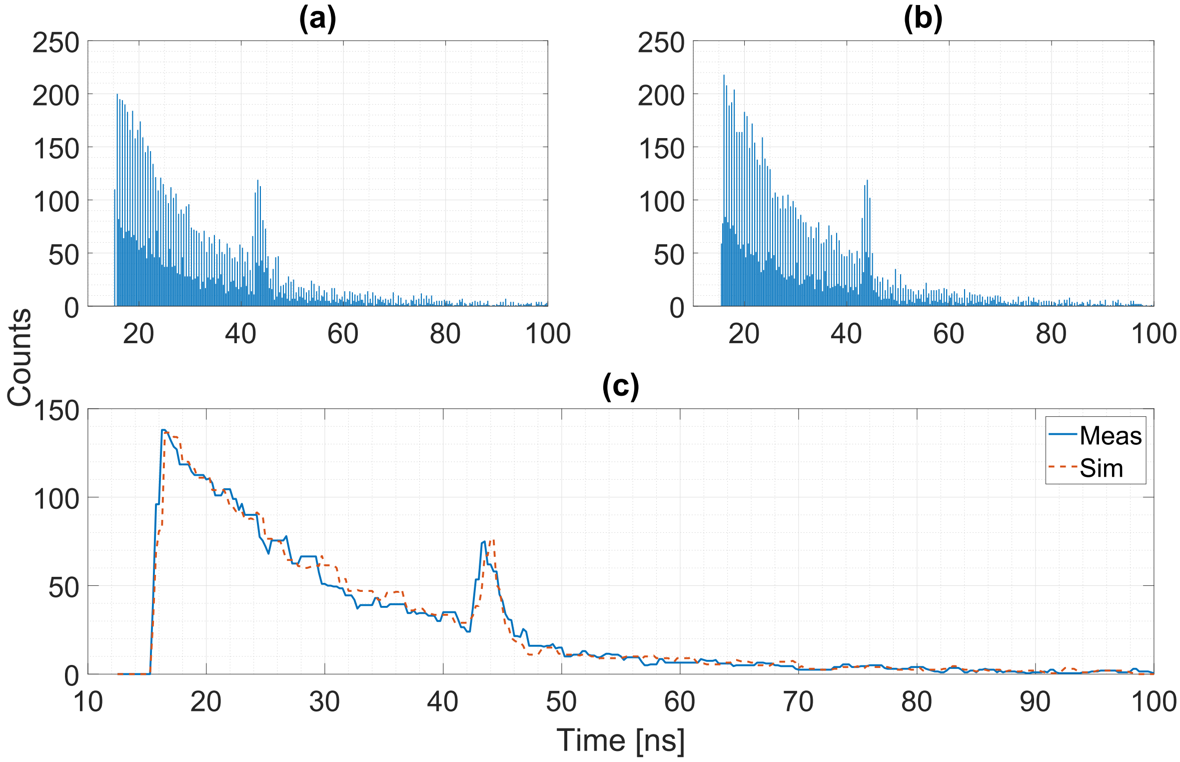


Figure 3.11: Example of histogram in presence of intense background: measured data (a), simulated data (b), comparison of the two histograms after median filtering (c). The systematic difference between even and odd codes in the histogram originates from the TDC differential nonlinearity of the detector used for comparison [10], and has been included in the model with good final match.

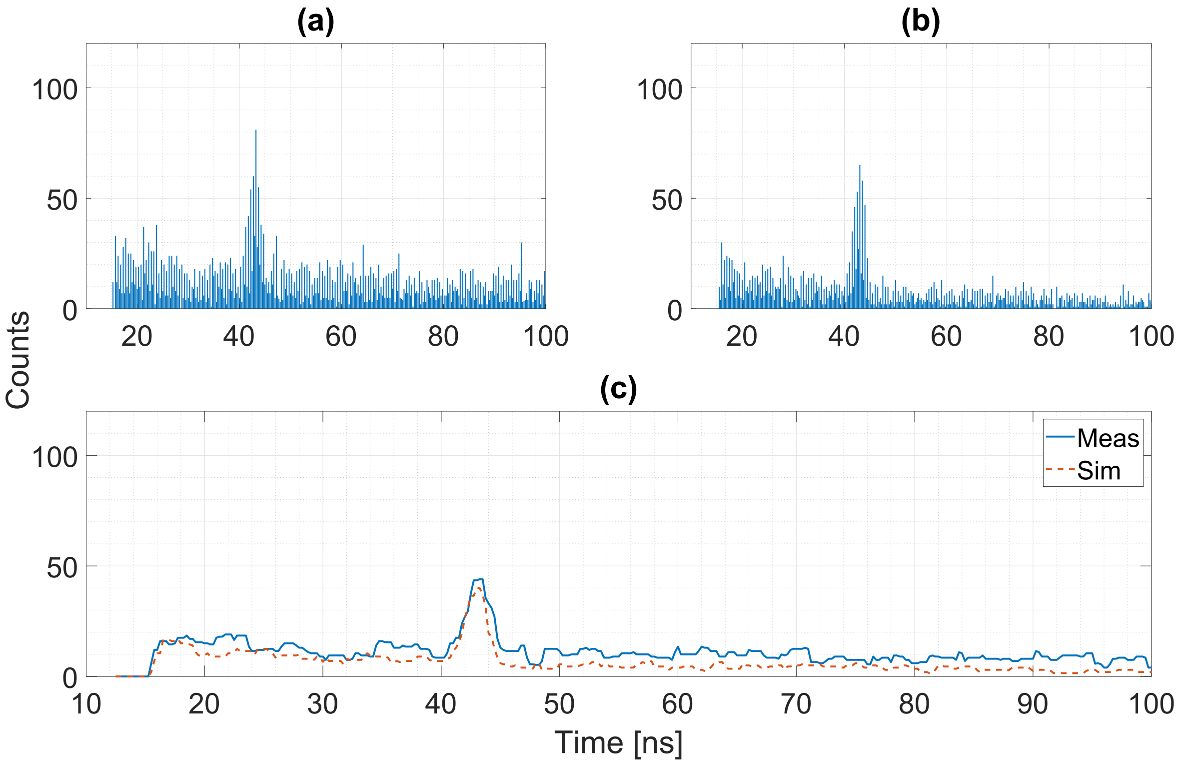


Figure 3.12: Histograms with temporal correlation of 2 photons within a time window of $\simeq 2.3$ ns: measured data (a) and simulated data (b), comparison of the two histograms after median filtering (c).

Chapter 4

Analysis and comparative evaluation of background rejection techniques

In Chapter 3, we provide deep insights of the structure of a mixed physical/numerical simulator we developed to aid the design of modern SPAD-based d-ToF systems. The effectiveness of a system simulator proved to be fundamental in the definition of a LiDAR system architecture by providing designers with the flexibility to predict the performance in a wide range of possible configurations [86, 95, 96], which would be otherwise extremely difficult if not impossible. In this chapter, we make extensive use of simulations in the context of one of the most important aspects of modern SPAD-based LiDAR systems: the capability to deal with uncorrelated light from the external environment.

Being d-ToF an active sensing technique, the presence of photons from background illumination with uncorrelated timing relationship with respect to the emission time of the laser pulse represents a severe performance-limiting factor. In the design of a modern, competitive SPAD-based d-ToF sensor, it is mandatory to implement effective solutions to mitigate the negative effect of this phenomenon ensuring proper sensor operation in a wide range of scenario.

While already several solutions have been deployed over the years [6,8,10,43], there is a lack of knowledge concerning if (and how) different techniques could be possibly operated together to increase their efficacy in rejecting background-related photons and maximize the detection of photons from the laser source.

In this chapter we try to fill this gap of knowledge by presenting an analysis and comparative evaluation of three background-rejection techniques implemented in CMOS processes for SPAD-based LiDAR systems, using both Monte Carlo simulations and laboratory measurements. The first technique, known as *photon coincidence technique*, uses the temporal proximity of multiple detections to reject background light and maximize the detection of photons belonging to the reflected laser pulse. With the second technique, named *Auto-Sensitivity (AS) technique*, background light is rejected by automatically reducing the SPAD photon-detection probability (PDP), in order to avoid sensor saturation due to the intense environment illumination. The third technique we consider is the *last-hit* detection, which is able to detect and timestamp the last event impinging on the sensor over the acquisition window rather than the first, maximizing the system performance for long distance targets.

More specifically, in our analysis we first compare the photon coincidence against the AS technique and show that the reduction of SPAD sensitivity, or photon detection probability (PDP), required by the AS technique to achieve the same level of performance as that of photon coincidence is unfeasible with the current technology. Then, we operate the same two techniques together, showing an increase in the maximum distance range with respect to each technique considered in isolation. Measurements taken in a laboratory setup confirm the results obtained with the Monte Carlo simulator. We then focus on three different implementations found in the literature for the photon coincidence detection [4,10,43,55], showing how performance is affected based on how the photon coincidence is exploited. Finally, we present a comparison between the standard first-hit detection paradigm with the last-hit detection strategy, confirming the validity of the proposed technique in favoring the detection of long distance targets. The results also show that last-hit, on average, outperforms the first-hit detection, and thus it is a potential candidate to be the primary detection strategy in d-ToF SPAD-based LiDAR systems.

The results that we present have been obtained with a Monte Carlo simulator that uses an analytical model for the computation of the optical power budget and a numerical engine to emulate a train of photon detections [97]. This approach is justified, since several models for SPAD-based d-ToF LiDAR systems have been proposed in the literature [86, 95–97], showing a high level of confidence and thus giving the possibility to study several features of the system without the need for instrumentation and physical setups. Nevertheless, we do complement the simulations with actual laboratory measurements, where possible with our setup, which support and confirm the analytical results.

This chapter is organized as follows. In Section 4.1, we provide an in-depth analysis of the three considered background-rejection techniques. Results from both simulation and measurements are detailed in Section 4.2. Considerations on the possibility to

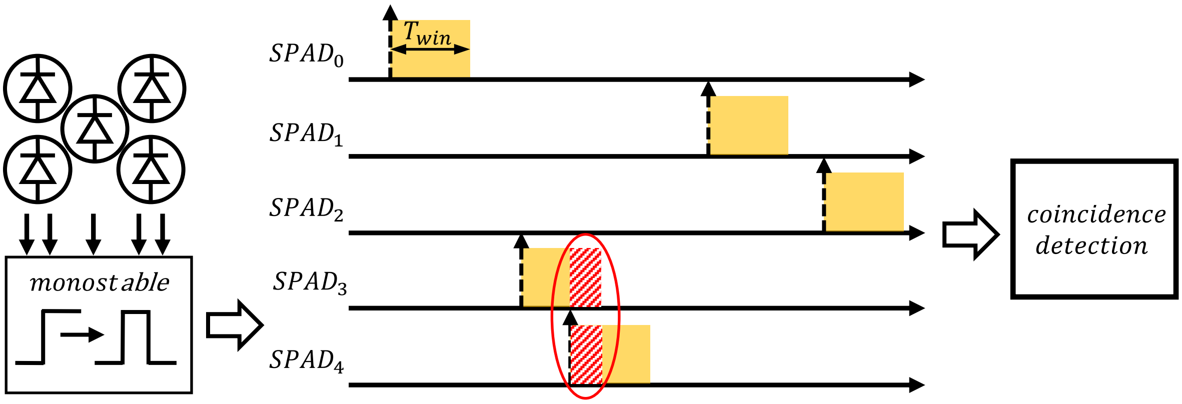


Figure 4.1: Example of photon coincidence detection with five SPADs and an event validated with two triggers.

automatically adapt the settings for the coincidence detection technique are discussed in Section 4.3, while conclusive remarks are provided in Section 4.4.

4.1 Background rejection techniques

In this section, we illustrate the working principle, key parameters and circuit implementation for the photon coincidence, the Auto-Sensitivity and last-hit detection techniques.

4.1.1 Photon coincidence technique

With the photon coincidence technique, the impact of background photons is reduced by exploiting their temporal sparseness as measured at the output of several detectors. This technique requires the use of at least two SPADs, with a dedicated circuitry that is able to provide a trigger when at least two photons are detected within a temporal frame called *coincidence window*. In that case, the event is considered part of the foreground, and a timestamp is extracted to measure the distance. Conversely, the isolated photons are considered background, and discarded. An example of implementation of this process is depicted in Fig. 4.1, using a total of five SPADs.

From a logical point of view, the output of each SPAD triggers a pulse upon each photon detection. The width of the pulse, indicated with T_{win}, can be set by the user and defines the temporal window of validity during which photon detections are considered coincident. Overlapping pulses indicate that detections have occurred close enough in time to be considered coincident. In the example, a photon-detection event

Table 4.1: Summary of photon coincidence circuits.

Ref/year	Type	Coincidence levels	full coincidence
OR-based [10] - 2017	OR tree + counter	$N \geq 2$, $N \geq 3$, $N \geq 4$	no
XOR-based [55] - 2019	XOR tree + dual-edge counter	$N = 0, 1, 2, 4, 8$	no
Adder-based [43] - 2013	Full and half adders	$N \geq 2$	yes
Neighbor-based [4] - 2021	Primary + N Secondary pixels	$P + 1$, $P + 2$	yes

is validated on SPADs nr. 3 and 4, assuming that coincidence requires at least two SPADs, since their pulses overlap in time. The total number of SPADs employed in the coincidence detection process, as well as the threshold N of concurrent outputs required for coincidence, depend upon the implementation and the desired level of background rejection capability to be achieved. In the following section, we discuss several ways this method can be implemented in practice.

Comparison of photon coincidence circuits

In the literature, this technique has been implemented in many ways. While the underlying idea is the same, the performance could be highly affected by the way in which the coincidence of events is exploited by means of CMOS circuits.

We consider several implementations, each with its own pros and cons. In the following, we will refer to them as *OR-based* [10], *XOR-based* [55], *Adder-based* [43] and *Neighbor-based* [4]. A summary reporting the type of logic employed, the achievable coincidence levels and a short description is reported in Table 4.1.

The main difference between the considered coincidence detection circuits consists in how multiple photon detections are managed and how the coincidence is exploited. With the *OR-based* and *XOR-based* implementations, the detections from a multitude of SPADs are converted into electrical pulses, and conveyed at the clock input of a digital counter by means of a tree of digital gates. Then, according to the number of pulses within the user-defined temporal window T_{win}, an event can be validated or not. The main difference between the implementations of the *OR-based* and *XOR-based* solution is the type of digital tree used to convey the electrical pulses. The *OR-based* implementation uses a tree of OR gates to derive a single signal from several SPADs. Because of the inclusive way the OR gate works, if the SPAD signal were to stay active after the detection, the output of the tree would remain at the high level, and none of the following detections would be able to propagate through the tree and be counted. To avoid this, the output from each SPAD has to be transformed into a narrow pulse beforehand. In the referenced implementation, a monostable circuit with a pulse width of approximately 400 ps is employed. The output of the tree, i.e., the combined 400 ps pulses from each SPAD, is then conveyed to the clock input of a digital counter to keep track of all the events. Conversely, in the *XOR-based* implementation, the combination tree is made with XOR gates: in this case, the arrival of each new photon causes the output of the tree to always change state (from 1 to 0, or from 0 to 1), thus encoding the detection into a rising or falling edge of the signal. With the implementation of a dual-edge digital counter, the monostable circuit is then no longer necessary.

Both the *OR-based* and *XOR-based* approaches present advantages and disadvantages if compared to each other. The main disadvantage of the *OR-based* implementation is that if the average photon rate from the multitude of SPADs is comparable to the equivalent bandwidth implied by the monostable pulse width, different pulses can merge with each other in the OR-tree, reducing the effectiveness of the coincidence detection scheme as some pulses are missed by the counter. With the *XOR-based* approach, thanks to the absence of the monostable circuit, a much higher bandwidth can be guaranteed if compared to the *OR-based approach.* On the other hand, in case of very strong photon flux, the *OR-based* approach would give at least one detection event (even though no coincidence can be exploited), as opposed to the *XOR-based* approach in which nearly simultaneous events would cancel out each other, completely impairing the measurement process.

A potential drawback of both the *OR-based* and *XOR-based* implementations is that the photon coincidence is not exploited over the entire set of detected photons. In particular, the validation window of duration T_{win} is opened by the first detected photon. Whether or not the validation threshold N, i.e., the number of detections required to consider the event valid, is reached, none of the other pulses which are present within the current coincidence window is able to open a new coincidence evaluation window. A new coincidence window can then be re-opened only after the current one is closed. In other words, the coincidence windows do not overlap. With this implementation scheme, we can say that the circuit is not able to exploit the *full coincidence.*

A different approach is employed in the *Adder-based* implementation, which relies on a combinational circuit made of half-adders (HA) and full-adders (FA). In this implementation, each photon detection produces a fixed pulse of either 4 or 8 nanoseconds (thus T_{win} is fixed), which is fed into a network of full-adders (FA) and half-adders (HA). In practice, the network *counts* how many pulses are active at any given time, in a way similar to the logical working principle discussed in Section 4.1.1. In the referenced implementation, the threshold is fixed to $N \geq 2$ and the circuit is able to exploit the *full coincidence* since there is no prioritization of pulses as in the *OR-based* and *XOR-based* implementation.

While in the *OR-based*, *XOR-based* and *Adder-based* implementations all of the SPADs involved in the photon coincidence exploitation are peers, a new concept has been developed in the *Neighbor-based* implementation. In this latter case, each pixel contains a single SPAD, and the coincidence is exploited by considering the output of its neighbors. In particular, each valid event requires a prior detection in the current, or *Primary (P)* pixel, followed by a defined number of detections within the T_{win} window in the neighboring, or *Secondary (S)* pixels. The main advantage of this technique is that each pixel needs to accommodate only a single SPAD, resulting in a more compact design without sacrificing the capability to exploit the photon coincidence.

Another implementation, designed to operate the sensor using a continuous-sampling TDC, is called *Synchronous Summation Technique (SST)* and proved to outperform both the OR-tree and XOR-tree based approaches in terms of detected photon rate [56]. The technique was implemented on FPGA by Patanwala et al. [56], with a proposed integration in silicon by Van Blerkom [98]. Even though the results are promising, we decided not to further consider it in this work, as we are focusing on approaches which are integrated and characterized at the CMOS level.

A timing diagram showing a detailed example of the *neighbor-based*, *XOR-based* and *OR-based* implementations is shown in Fig. 4.2.

4.1.2 Auto-Sensitivity (AS) technique

One way to reject background light is to selectively adjust the photon detection probability (PDP) of the detector, a measure of its sensitivity. In the Auto-Sensitivity technique [4], the SPAD PDP is reduced when the number of detected photons falling within a time calibration window (T_{calib}) exceeds a specific threshold.[1] The circuit is implemented at the pixel level by means of a dual excess bias SPAD front-end realized with 3.3 V thick oxide transistors and a photon counter. Fig. 4.3 illustrates the implementation schematics at the transistor-level, showing two SPAD recharging paths that can be selected according to the desired sensitivity level. Fig. 4.4 shows on the left how the main signals involved (V_{ph}, V_{inv} and R_H/R_L) change depending on the level of sensitivity, high or low, that is selected. The same figure on the right shows the detection probability (PDP) as a function of the wavelength for several excess bias voltages V_{ex}. For high sensitivity, the SPAD anode is discharged to ground during the reset phase

---

[1] Going the other way, i.e., increasing the PDP when the number of detected photons is below the threshold, would not work, since the technique is designed to avoid saturating the sensor before the reflected laser pulse is back. If the background is low, sensitivity would be increased, possibly resulting in high background noise, without the possibility to then reduce it.

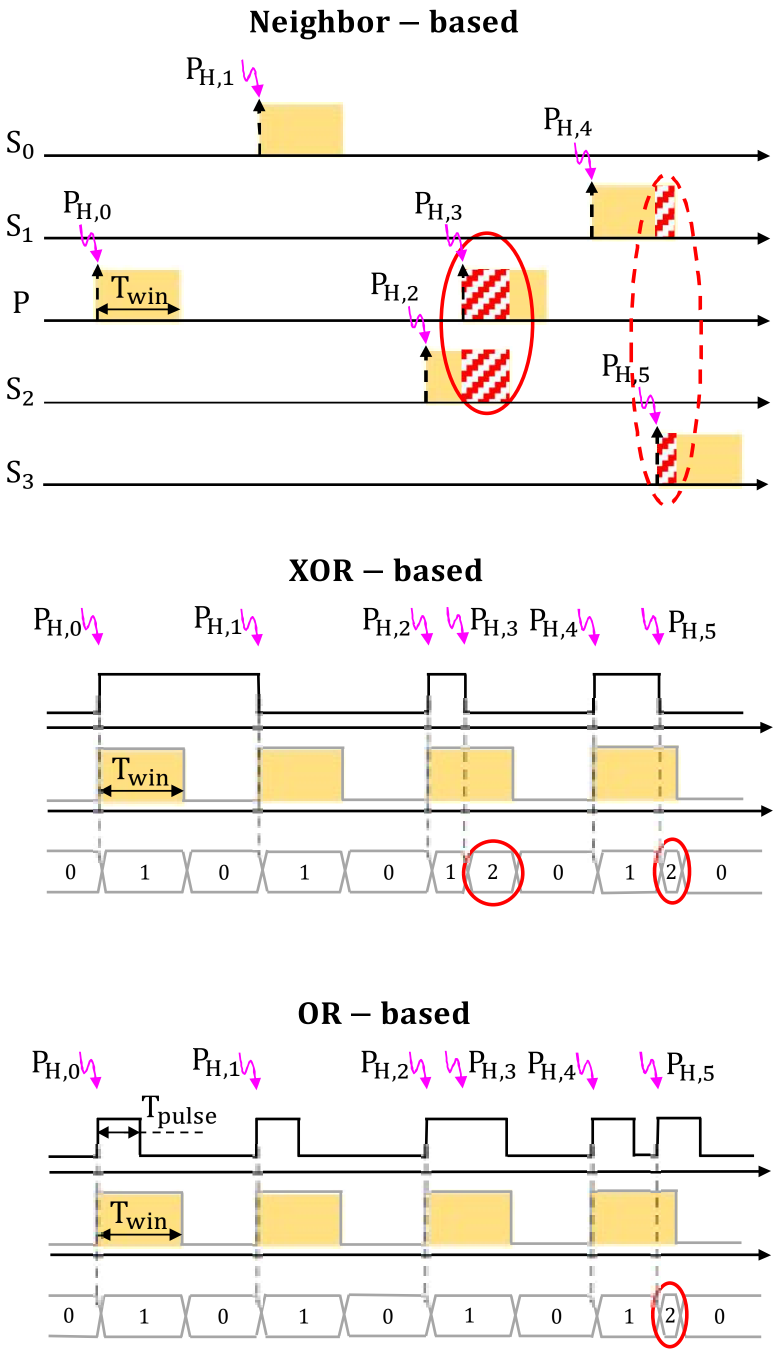


Figure 4.2: Timing diagram showing a detailed example of behavior of each of the three implementations selected for comparison, assuming a threshold of $N \geq 2$ photons within a time window T_{win}. In the *neighbor-based* implementation, as the coincidence exploitation is constrained to a detection in the *Primary (P)* pixel, only photons P_2 and P_3 will trigger a detection event, while photons P_4 and P_5 are not considered (even though they have temporal proximity within T_{win}). With the *XOR-based* implementation, however, each photon detection is considered and thus both photon pairs $P_2 - P_3$ and $P_4 - P_5$ trigger a coincidence event. With the *OR-based* implementation, each photon detection is converted into a single electrical pulse of duration T_{pulse}, resulting in a bandwidth limitation. Because of that, only photon pair $P_4 - P_5$ is recognized as a coincidence event, while the coincidence information from the pair $P_2 - P_3$ is lost.

by transistor M1. Conversely, to set a lower sensitivity, the SPAD is reset to a tunable reference voltage V_{sens} through the M2-M3 pair. Therefore, depending on the selected path (through signals $R_H, R_L, \overline{R_L}$), the SPAD excess bias voltage V_{ex}, i.e., the potential above the breakdown voltage, is set either to $V_{spad} - V_{bd}$ or to $V_{spad} - V_{bd} - V_{sens}$, maximizing or minimizing the SPAD PDP (where V_{bd} is the SPAD breakdown voltage).

When the AS is activated, the pixel starts counting the number of photons in high sensitivity, and switches to low sensitivity if $N \geq 15$ photons are detected within T_{calib} [4]. The high sensitivity setting can be restored with the pixel reset signal (not shown in the schematics for simplicity). We point out that the level of PDP reduction is limited by the maximum applicable V_{sens}, which also affects the speed of the circuit. In fact, despite the large transistor sizes, a higher V_{sens} results in lower conductivity for the nMOS transistor M0, which might be unable to set the recharging voltage within a short time. Moreover, in case of low sensitivity, the pMOS transistors of the level shifter (needed to interface the 3.3 V domain with the 1.2 V domain of the downstream logic) become slower because of the reduced input voltage swing $[V_{sens}, 3.3\ V]$. At the same time, the output pulse time delay variation between high and low sensitivity must not exceed the Time-to-Digital-Converter (TDC) resolution in order to preserve the photon timestamp information. Because of these reasons, the maximum applicable V_{sens} is limited to 2.4 V. Therefore, the PDP can be scaled by only a factor of 1.25, as reported in the plot in Fig. 4.4, which shows the PDP as a function of the wavelength and of the excess bias voltage V_{ex} (which in turn depends on V_{sense}).

While the level of background light intensity is the primary factor which triggers the AS control circuit to switch the sensitivity setting, also the selected wavelength of operation has a non-negligible impact with this technique. Considering the higher sensitivity of SPADs in the visible region of the spectrum, saturation (and so the AS setting change) is expected earlier than in an equivalent system operating in the near infrared (NIR) region.

The AS technique was originally proposed for a flash LiDAR scenario [4]. Because the amount of background light is independent of the acquisition configuration, the same considerations apply also for a scanning scenario, which is the one selected for the analysis of this work.

4.1.3 Last-hit detection

In SPAD-based d-ToF systems, the reduced pixel area limits the memory available for storing timestamps to a few, typically one, events and the conventional acquisition

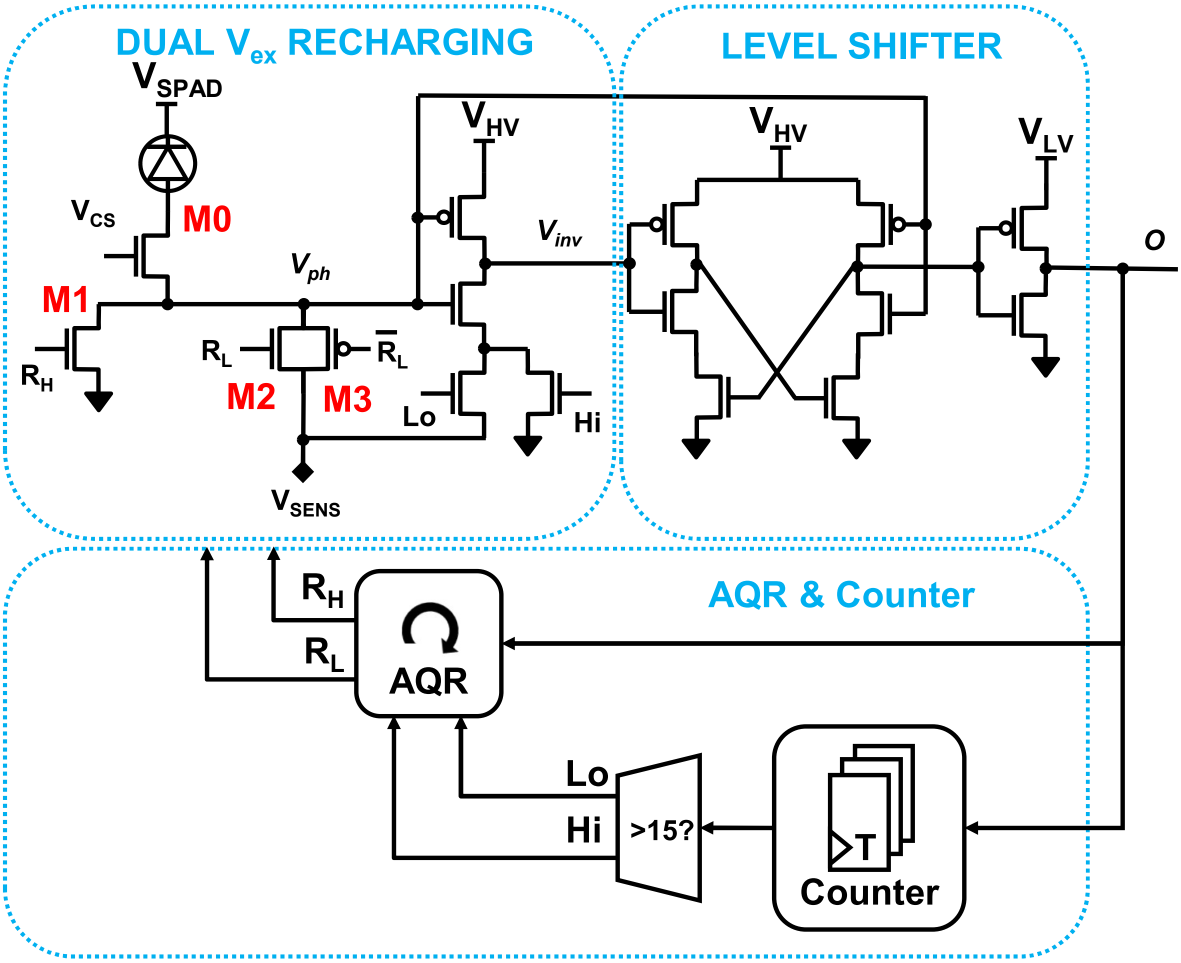


Figure 4.3: Transistor-level implementation of a dual-sensitivity front-end circuit, showing the SPAD recharge path selected for maximum sensitivity (through transistor M1) and the SPAD recharge path for the reduced sensitivity (through transistor M2/M3).

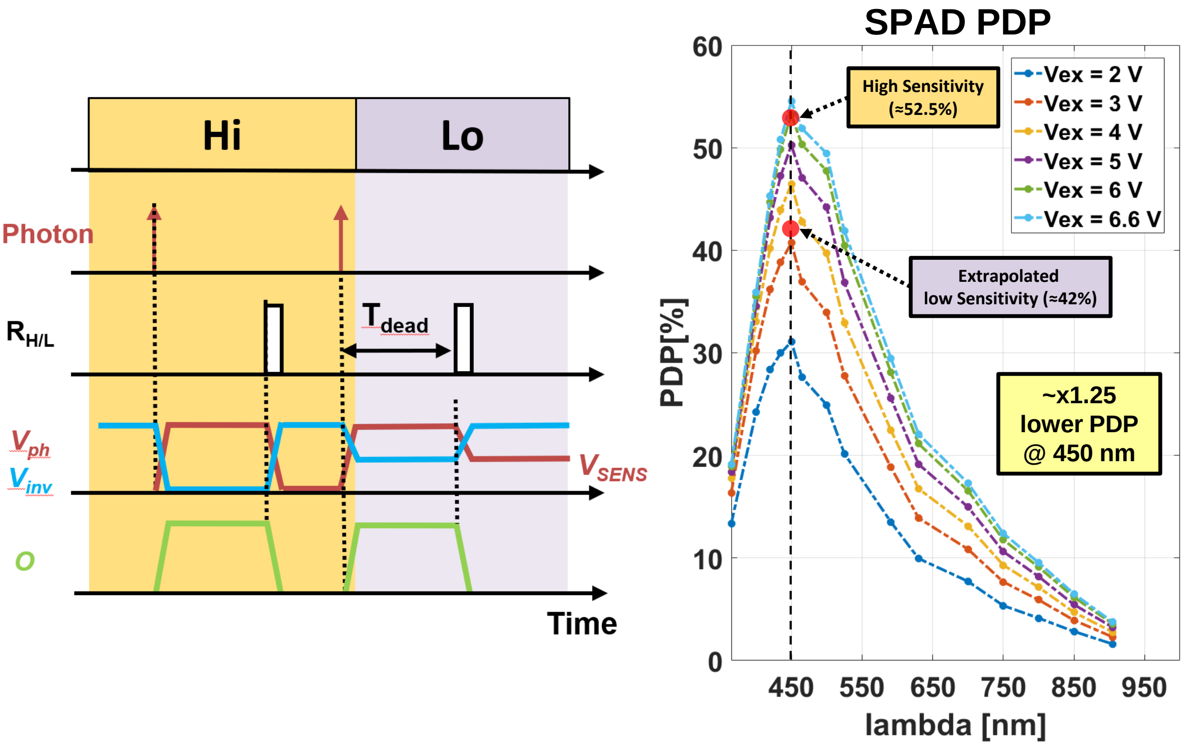


Figure 4.4: Timing diagram showing the main signals of the SPAD front-end circuit in high (Hi) and low (Lo) sensitivity and the behavior of the SPAD PDP for different excess bias voltages V_{ex} at different wavelengths. The V_{ex} corresponding to the low sensitivity setting is given by the difference between the high sensitivity setting (6 Volts in the example) and the maximum V_{sense} of $\simeq$2.4 Volts, resulting in a lower value of V_{ex} of $\simeq$3.5 Volts (the actual low sensitivity PDP value is thus interpolated).

scheme considers the first detected event to be stored into the histogram of timestamps for the ToF extraction. In the work proposed by Manuzzato et al. [4], however, a novel acquisition scheme is proposed which considers the *last* detected event. This criterion has been developed to favor the detection of long-distance targets, i.e., when the returning laser echo is expected towards the end of the acquisition window. In that work, the last-hit detection strategy is implemented on a per-pixel basis by means of a special memory controller. Upon a detection, the current timestamp is saved into an in-pixel memory, and the TDC is quickly reset to measure the next event. If a new event is detected, the memory is overwritten with the latest information, and the process is repeated up to the end of the acquisition window. This acquisition scheme can be operated in conjunction with other background-rejection techniques such as the photon coincidence and Automatic-Sensitivity, as it operates at the end of any event validation circuitry, or even with pile-up reduction techniques such as the asynchronous imaging method proposed in the work by Gupta et al. [99].

Due to the underlying Poissonian nature of the incoming photons, the background-related histogram of timestamps that results when a sensor is operated under the last-hit detection strategy is a mirrored version of the "first-hit" histogram. In addition, the peak of the probability of detection will be found towards the right hand side of the horizontal axis, i.e., towards the end of the acquisition window, if we assume the target is far. This strategy increases the signal to noise ratio of the system. Intuitively, this can be justified by the fact that, during the continuous process of detection and overwrite of the last-hit detection, background-related events are overwritten continuously and the probability that the last detected event belongs to the target reflection is higher as the target distance increases. For this reason, the detection of targets which have their echoes near the end of the acquisition is more robust. A comparison of histograms for a target located at 1/3 and 2/3 of the measurement range using both first-hit and last-hit detection is shown in Fig. 4.5, with an increase in the signal peak when the last-hit detection is considered with a long-distance target. Even though the last-hit detection has been designed to favor long-distance targets, the SNR benefits also for short distance targets, as shown in Fig. 4.5-b. The selection criteria between the *first* or *last* detection strategy depends upon several variables, such as target displacement, reflectivity, amount of optical power and background intensity, resulting in a complex decision scheme which makes an in-pixel real-time decision hardly feasible. Consequently, the decision boundary between *first* or *last* detection strategy is determined at the system level, upon the analysis of the histogram of timestamps to properly select the most suitable configuration. If an a-priory decision has to be taken, the *last* detection strategy is more advantageous, as higher priority is given to the signal of a long distance target, as it will be shown in the results presented Section 4.2.3.

4.2 Results

In this section, we study the behavior of the background rejection techniques under investigation through the Monte Carlo simulator [97] as described in Chapter 3. Simulation helps us evaluate the performance over a wide range of parameters. In particular, the simulator produces a train of simulated photon-detections upon the computation of the optical power budget, given a list of parameters describing the LiDAR system. We consider a set of real-world parameters inspired by a typical automotive LiDAR scenario in scanning mode [43, 100], reported in Table 4.2.

Regarding the optical configuration of the system, in particular referring to the photon coincidence technique, we assumed a uniform pulse energy over the selected

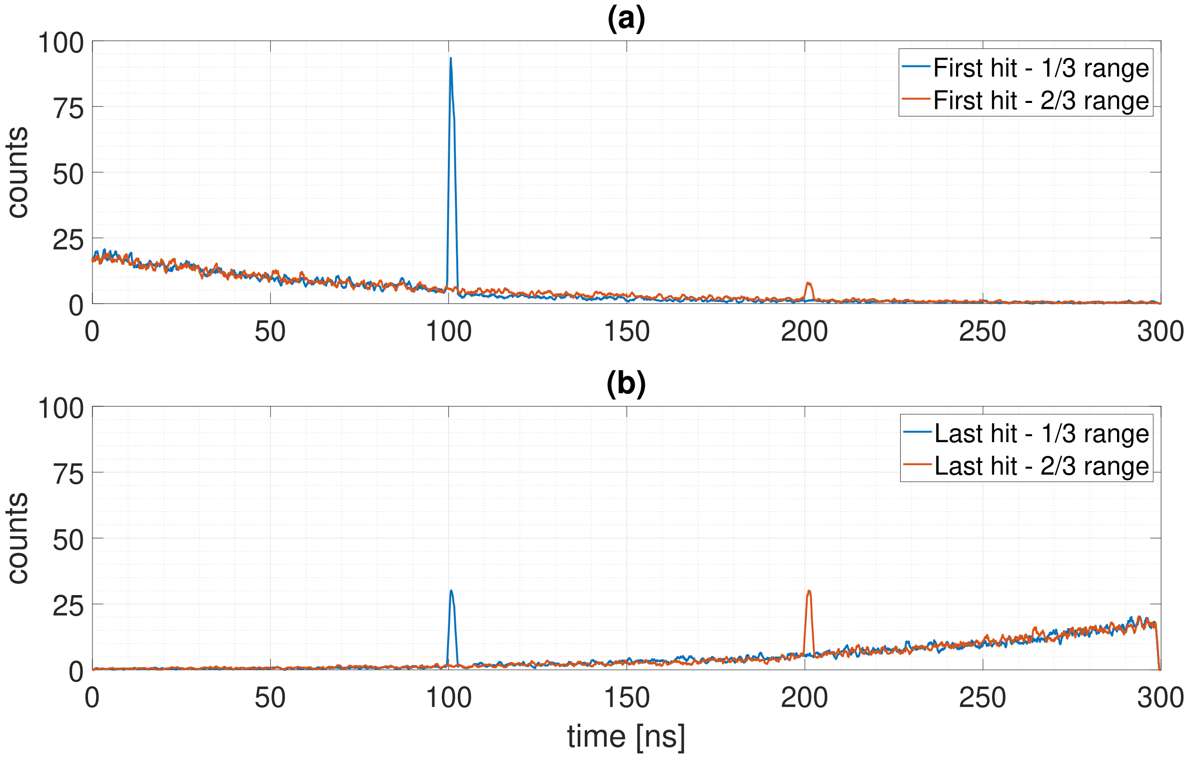


Figure 4.5: Qualitative comparison of simulated histograms with first-hit (a) and last-hit (b) detection strategies for a target at $1/3$ (100 ns) and $2/3$ (200 ns) of the measurement range (300 ns). The SNR for the long-distance target is in favor of the last-hit detection strategy. Additionally, even considering the whole range, the SNR with the last-hit detection strategy is more uniform, as opposed to the SNR in first-hit.

pixels for all of the implementations selected for our comparison [4, 10, 55].

In Section 4.2.1, we compare the performance of the AS technique against the photon coincidence, showing the latter to be outperforming. Additionally, we show that the overall performance can increase if the AS and photon coincidence techniques are applied together.

A deeper analysis which focuses on the comparison of different implementations of the photon coincidence is reported in Section 4.2.2, showing the overall system performance can improve by up to 50% with the best performing coincidence detection scheme. The analysis of the last-hit detection is presented in Section 4.2.3, showing that this technique can be potentially selected as the main detection strategy due to its intrinsic advantages with respect to the classical first-hit detection.

Considering one of the photon coincidence implementations selected in this work, i.e., the *Neighbor-based* implementation from [4], we have included both implementations available in the sensor, to show that both can be used with successful results.

In particular, the configuration with one *Primary (P)* pixel and four *Secondary (S)* pixels is selected for the analysis of Section 4.2.1. The second configuration, with eight *Secondary (S)* pixels is selected for the other two set of analysis, in Sections 4.2.2 and 4.2.3.

4.2.1 Auto-Sensitivity vs. photon coincidence

The first set of simulations compares the performance of a system implementing the AS technique against a system implementing the photon coincidence. In particular, we simulated scenarios where the target is not visible in standard conditions (i.e., with the detection of the first photon) due to the combination of distance (50 m and 100 m), reflectivity ($\rho = 50\%$ and 90%) and background illumination (50 kLux) saturating the histogram of timestamps from which the ToF is extracted. The AS technique works by automatically decreasing the SPAD PDP if a background intensity over a certain threshold is detected. For this reason, we executed a sweep on the PDP parameter to understand by how much it shall be decreased to reach the same performance (in terms of percentage of correct measurements) of a system implementing the photon coincidence technique. In the evaluation of results, a measurement is considered correct if its deviation from the ground truth is within a given threshold, which was set equal to ± 10 cm. The results are depicted in Fig. 4.6 for the different combination of parameters, and show that the performance of the AS technique reaches (and in some cases, exceeds) the performance of the coincidence technique *only for extremely low values of PDP (the required PDP would be $\simeq$0.05%)*. This is equivalent to a decrease factor of $\simeq$ 100 times with respect to the nominal PDP value of $\simeq$ 5%, which, as discussed, cannot be implemented with the architecture of Section 4.1.2.

The second set of simulations explores if and by how much the performance can be improved if the AS technique is used *together* with the photon coincidence. The simulation envisages a very high target reflectivity ($\rho = 90\%$) with a moderate (50 kLux) and low (10 kLux) background intensity and target distance in the range from 10 m to 130 m and from 140 m to 340 m, respectively. The results are illustrated in Fig. 4.7, showing an increase in the maximum measurement range from 50 m to 100 m and from 270 m to 310 m in the two cases, considering a threshold at 80% of probability of correct measurement. This result is significant, and shows that despite its low performance when used alone, decreasing the SPAD sensitivity reduces the probability of being blinded by the reflected background light and thus increases the measurement range when combined with photon coincidence. In this case, the application of the AS

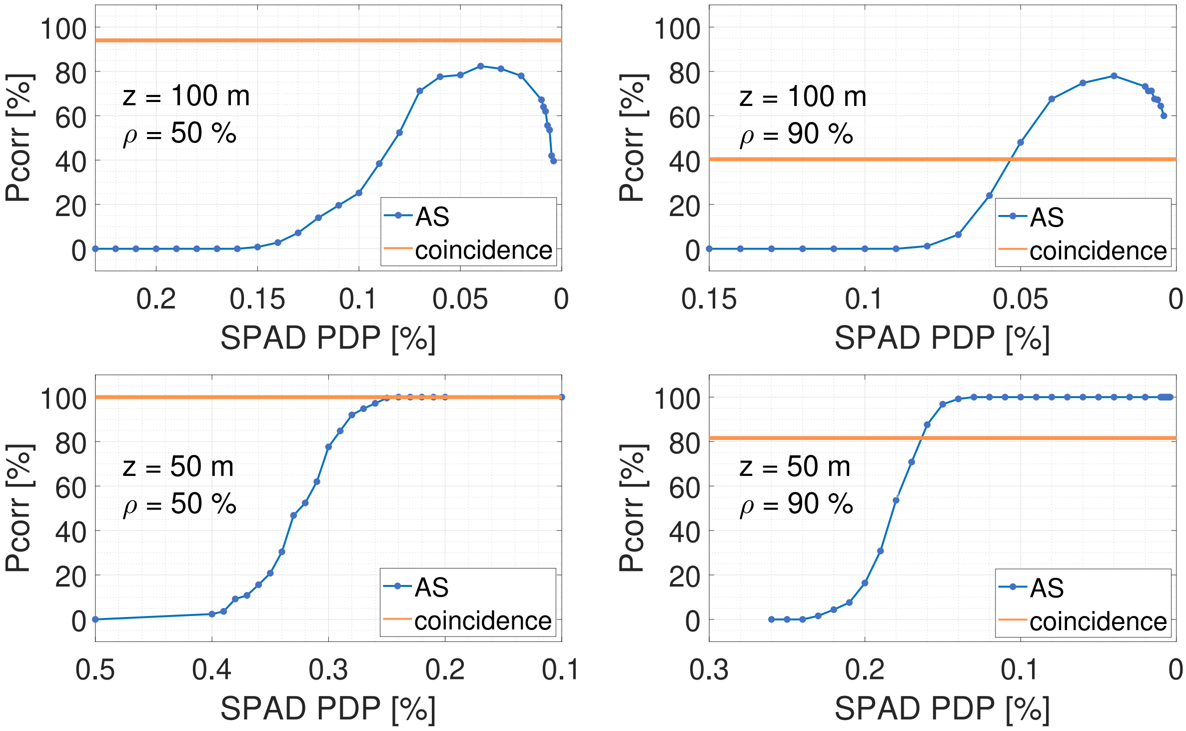


Figure 4.6: Probability of correct measurement as a function of the SPAD PDP. The amount of PDP decrease required to reach the performance of the photon coincidence technique is unfeasible with the proposed AS technique, since the minimum PDP which can be reached would be $\simeq 4\%$ (considering a reference PDP of $\simeq 5\%$ at $\lambda = 940$ nm).

technique decreased the SPAD PDP from the reference value of $\simeq 5\%$ to $\simeq 4\%$, which is in line with the proposed architecture of Section 4.1.2.

The obtained results have been qualitatively verified in a lab setup employing the detector described in [4]. The setup is shown in Fig. 4.8, and consists of two targets where the second is illuminated by a collimated fiber optics halogen lamp. Measurement results are shown in Fig. 4.9. In Fig. 4.9-(a), no background rejection technique is used, and the target illuminated by the halogen lamp shows a hole in the final 3-D image, since the laser return was completely covered by background photons, resulting in the complete loss of the ToF information. In Fig. 4.9-(b), the AS technique alone was activated, showing almost no improvement due to the too low decrease of the SPAD PDP (from $\simeq 5\%$ to $\simeq 3\%$). In Fig. 4.9-(c), the photon coincidence technique was applied, considering at least two photons within a temporal window T_{win} of $\simeq 2$ ns with the *Neighbor-based* implementation. In this case, the ToF information from the illuminated target area is only partially recovered. Finally, in Fig. 4.9-(d), the AS technique was also applied, and the concurrent decrease of the SPAD PDP with the application of the photon coincidence technique allowed for a complete recovery of

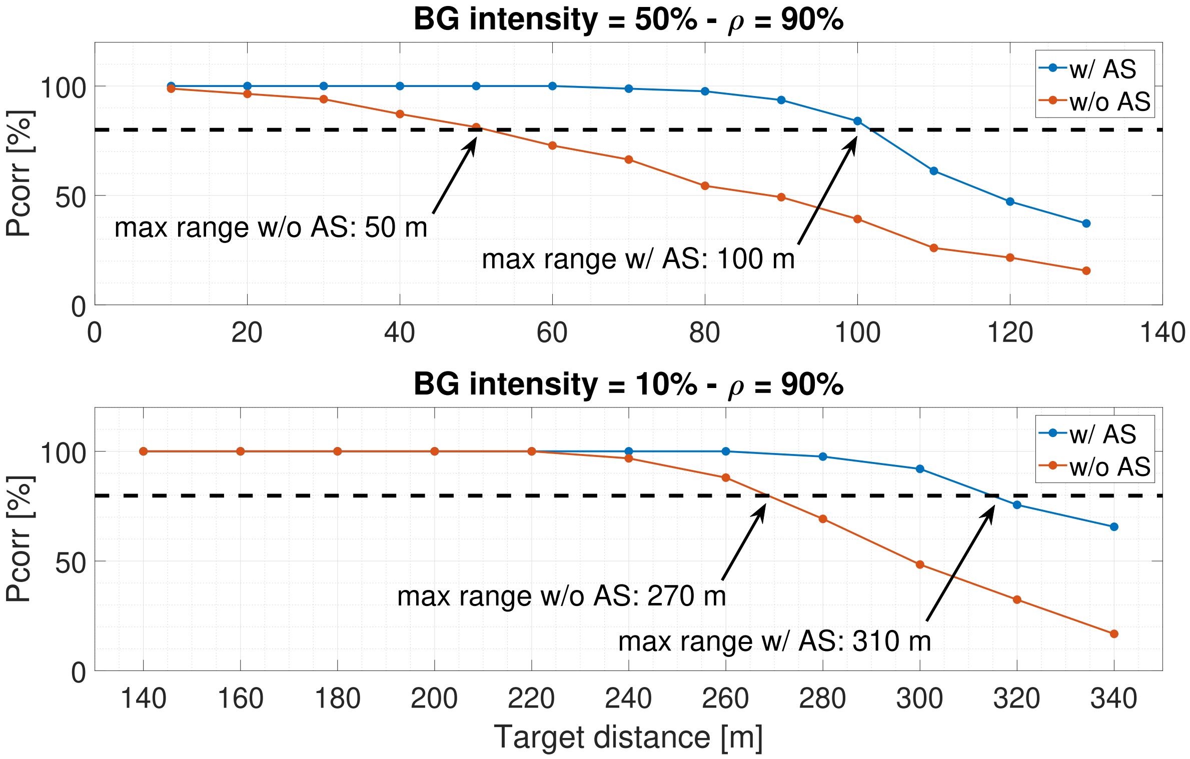


Figure 4.7: Probability of correct measurement as a function of target distance for two different background intensities. The AS technique, if used together with the photon coincidence, improves the maximum measurement range. Considering a threshold at 80% of probability of correct measurements, the range can be increased from 50 m to 100 m and from 270 m to 310 m in the two cases, respectively.

the ToF information from the histogram of timestamps. The combination is therefore shown to work better also in practice. Besides, Auto-Sensitivity does not degrade the information when not needed. In fact, AS works on a per-pixel basis, therefore the PDP is reduced only for those pixels with the strongest illumination. Conversely, the photon coincidence technique is a global setting of the array, thus preventing a complete image reconstruction if several illumination levels are present at the same time on the observed target area.

4.2.2 Comparison of photon coincidence circuits

With the photon coincidence technique, photon detection events are qualified to trigger a distance measurement only if their number and their temporal proximity exceeds a user-defined threshold. With this criteria, background-related events which are more likely to be sparse in time can be rejected improving the quality of the measurement. This detection process has been implemented in many ways and performance can be

Table 4.2: List of parameters used in the Monte Carlo simulation.

Parameter	Value	Unit	Description
T_{dead}	50	[ns]	SPAD dead-time
DCR^{a}	0.4	[Hz/μm^2]	SPAD dark-count rate
PDP^{b}	5	[%]	SPAD PDP
z^{b}	100	[m]	target distance
T_{win}^{b}	1	[ns]	coinc. window width
$SPAD_n$	4+1^{c}	n.a.	nr. of SPADs
Pix_{pitch}	50	[µm]	pixel size
FF	10	[%]	pixel fill-factor
P_{tx}	50	[W]	laser peak power
T_w	5	[ns]	laser pulse width
ϕ_e	0.28	[deg]	laser beam divergence
BW	10	[nm]	bandpass filter width
τ_{opt}	90	[%]	optics transmittance
$F\#$	1.2	n.a.	optics F-number
ρ_{target}^{b}	10	[%]	target reflectivity
$BG_{level}^{b,d}$	50	[kLux]	background level
λ	940	[nm]	wavelength
N_s	500	n.a.	nr. of simulation points

[a] Value from [19].
[b] Default values, changed in the simulation.
[c] Coincidence configuration from [4].
[d] Background power obtained from ASTM G173 standard.

greatly affected by design choices. In this section we focus on simulation results trying to enhance the differences between three implementations of the coincidence detection circuit. In particular, we consider the *OR-based*, *XOR-based* and the *Neighbor-based* as already described in Section 4.1.1. The fourth implementation, *Adder-based* has shown very similar results to the *XOR-based* and thus it will not be reported in the plots. The main reason for this finding is that in both cases the pulses received from the various SPADs are not merged together as in the *OR-based* approach. This, however, occurs when considering the range of parameters used in simulations, as even the *XOR-based* approach, for extremely high values of flux, would suffer from the issue of merging between nearly simultaneous SPAD pulses. As previously outlined, the main difference between the *Adder-based* and *OR/XOR-based* approaches is the possibility to exploit the full coincidence. In this case, however, this has not been observed as typically the coincidence window is as long or even shorter than the laser pulse, leaving no possibility of restarting the coincidence within the same laser pulse.

The simulator settings for this comparison have been tuned to the worst-case sce-

Figure 4.8: Laboratory setup for the qualitative evaluation of AS and coincidence techniques. Target nr. 2 was illuminated by a collimated fiber optics halogen lamp, emulating a highly reflective target reflecting an intense background illumination. In order to place the system close to a realistic scenario, the optical bandpass filter was removed.

nario for d-ToF LiDAR systems. In particular, we set a background intensity of 100 kLux with a target reflectivity of 10 % and a distance sweep between 20 m and 200 m. In order to perform a fair comparison and enhance the differences between the selected photon coincidence circuits, we set the same number of SPADs for all of the three selected circuits (thus not following the real implementation), i.e., 8 SPADs over which the coincidence is exploited (8+1 for the *Neighbor-based*). Also, for the sake of comparison, we did not consider hardware limitations of rise and fall times of

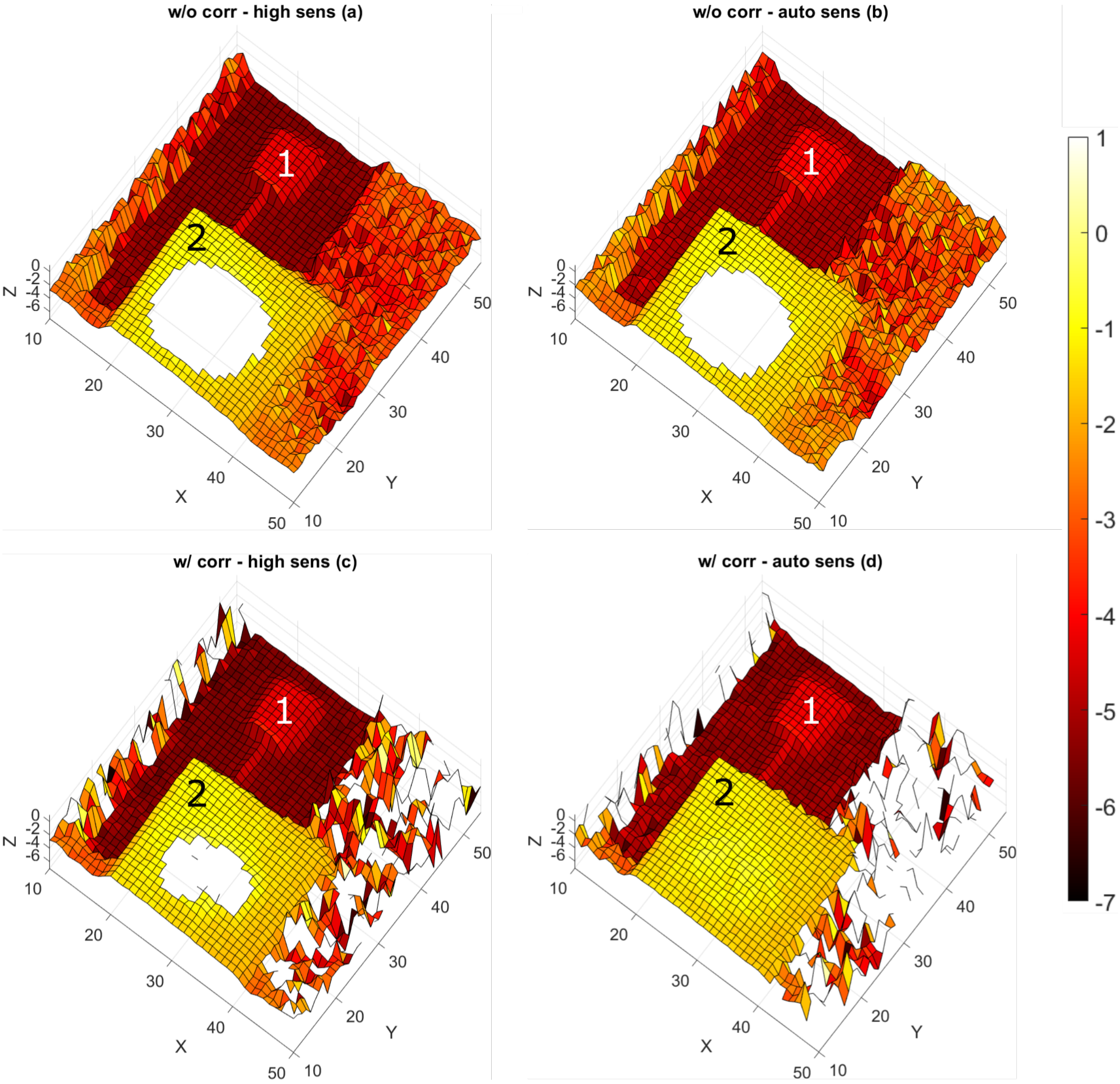


Figure 4.9: Qualitative measurement showing the reference 3D image in four cases. In (a), no background-reduction techniques are applied, in (b) and (c), the AS technique and photon coincidence are applied independently, and in (d) the two techniques are applied together, allowing for a full recovery of the 3D information.

electrical signals (being a common denominator to all of them), but we only considered the bandwidth limitation of the *OR-based* implementation which comes from the monostable circuit.

In the simulations, for each distance point in the range from 20 m to 200 m, we selected three different values of the coincidence window width, T_{win}, of 1 ns, 2 ns and 3 ns and for each value we selected three different thresholds for the coincidence detection ($N = 2$, $N = 3$, $N = 4$ for the *OR-based* and *XOR-based* and $P+1$, $P+2$, $P+3$

for the *Neighbor-based*). Consequently, we have a total of 27 simulation scenarios over which the performance is evaluated in terms of the probability of correct measurement. For the sake of compactness, we show only the most interesting cases which are worth reporting, highlighting the differences between the different implementations. For the final evaluation of the different implementations, we decided to use a threshold equal to 80% of measurement success rate, which we identify with the symbol P_{ok}.

In the first setting ($T_{win} = 1$ ns), the only implementation providing results above P_{ok} with all the selected threshold values ($N = P + 1$, $N = P + 2$ and $N = P + 3$) is the *Neighbor-based*, while the other two provide results above this threshold only with $N = 3$ and $N = 4$. The overall best performance, however, comes from the *XOR-based* implementation with a threshold of $N = 4$, yielding a minimum probability of correct measurement of 75%, as opposed to the minimum provided by the *Neighbor-based* of 40%. Considering the *OR-based* and *XOR-based* implementations with a threshold of $N = 3$, for distances up to $\simeq$120 m the *XOR-based* performs better, while the opposite holds afterward (as shown in Figure 4.10). This is explained by the bandwidth reduction given by the monostable circuits required by the *OR-based* implementation, which, in this case, filters also background light, providing a performance increase. However, the bandwidth limit given by the monostable circuit is overall a performance-degrading factor: in this configuration ($T_{win} = 1$ ns), the *OR-based* implementation does not work with a threshold of $N = 4$, due to the limit posed by the 400 ps monostable pulses. Additionally, as shown in Figure 4.10, the *OR-based* implementation with $N = 3$ shows a reduced probability of correct measurement also at short ranges. This misbehavior can be again attributed to the bandwidth limitation given by the monostable pulses propagating through the compression OR-tree. If the flux of photons is too high, the pulses are too close and merge, losing the information on the actual number of detected photons.

In the second setting ($T_{win} = 2$ ns), the overall best performance is yielded by the *Neighbor-based* implementation with a minimum of 80% of success rate. Also in this case, it is interesting to notice the effect of the monostable bandwidth limitation of the *OR-based* implementation. As shown in Figure 4.11, with a threshold of $N = 3$ photons, the bandwidth limit helps in rejecting also background light, yielding better performance than the *XOR-based* one. The opposite holds when a threshold of $N = 4$ photons is used: in this case the *XOR-based* outperforms the *OR-based*.

In the third setting ($T_{win} = 3$ ns), the overall system performance begins to degrade due to the higher integration time over which the coincidence is exploited, letting more background-related events to be falsely validated. In this case, the implementation scheme which overall performs better is the *Neighbor-based*: this is explained by the fact that, with this scheme, SPADs are not peers, but the validation of photon-detection events is constrained to a detection in one specific SPAD, that is the primary one. Consequently, with the highest coincidence window width considered in this work, the additional filtering level given by this particular implementation results in an overall performance that is higher than the other two schemes.

To evaluate the overall performance of the three coincidence detection implementations, we consider for each scheme the number of times the probability of correct measurement is greater than or equal to P_{ok}. This serves as an overall indication of which circuit implementation performs better over a wide range of settings, both in terms of coincidence window width T_{win} and the desired level of coincidence (the threshold N). Results are shown in Figure 4.12, indicating that the implementation scheme which overall performs better is the *Neighbor-based*, with a performance $\simeq 2.08$ times higher than the *OR-based*. Considering the absolute performance, the *Neighbor-based* results in the $\simeq 55.5\%$ of the total amount of distance points with a probability of success measurement higher than or equal to P_{ok}, followed by the *XOR-based* ($\simeq 37.7\%$) and the *OR-based* ($\simeq 26.6\%$).

4.2.3 Last-hit detection characterization

The last-hit detection strategy has been specifically designed to deal with long distance targets [4]. With the classical detection paradigm (first-hit), the probability of detecting a background event increases with the target distance. Conversely, with the last-hit technique, the opposite holds, since the priority is given to the last detected event which is more likely to belong to the target as its distance increases, as explained in Section 4.1.3.

Similarly to the results shown in Section 4.1.1, the combination of the last-hit detection with other background-rejection techniques outperforms the performance obtained with each individual technique. In particular, we show the results from the combination with the photon coincidence.

The last-hit has been characterized against the first-hit technique by considering the same simulation environment as reported in Table 4.2, with a modified setting to emulate the worst conditions for the detection of long distance targets, i.e., a background intensity BG_{level} of 100 kLux and a target reflectivity ρ_{target} of 10%. We characterize the performance over distance considering three different values for the coincidence window width T_{win} with the architecture of [4]. The complete list of modified parameters is

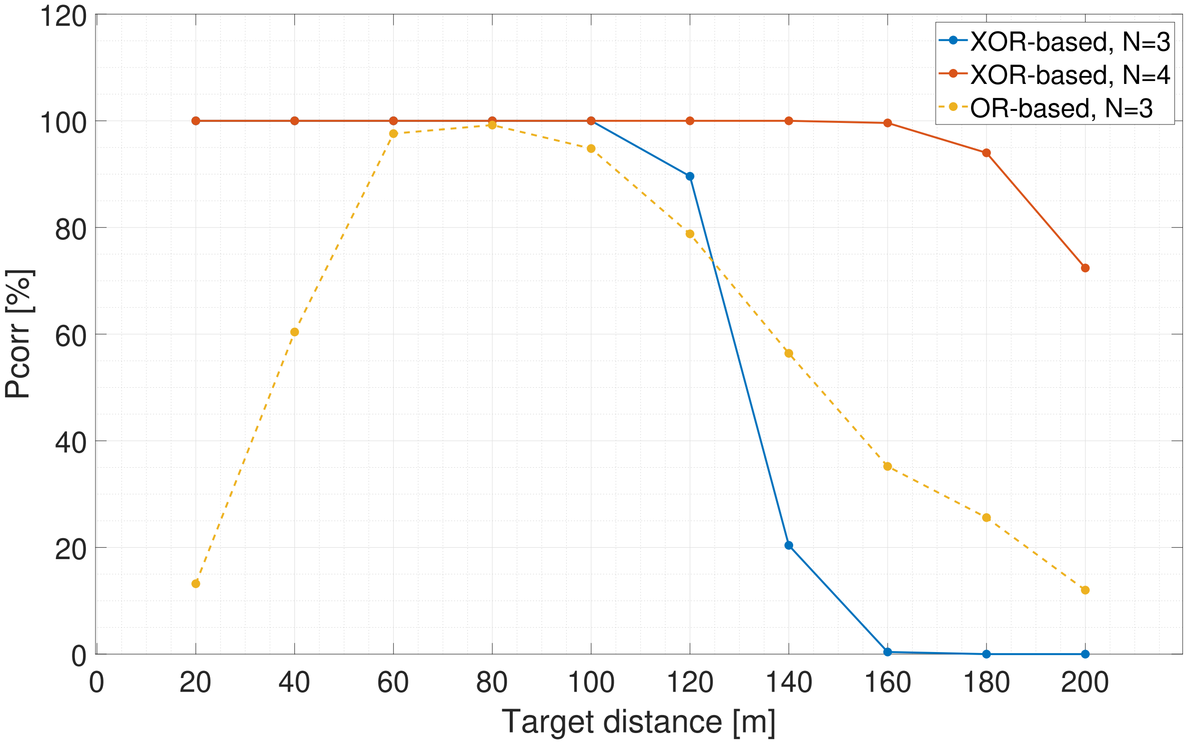


Figure 4.10: Comparison between *OR-based* and *XOR-based* implementations showing how, for the same threshold ($N = 3$), the bandwidth limitation of the monostable pulses of the *OR-based* helps as an additional background-rejection factor. On the other hand, for $N = 4$, this technique never works with the selected coincidence window width T_{win} of 1 ns.

reported in Table 4.3.

The results are shown in Figure 4.13, where the probability of correct measurement (success rate) is reported for the three different values of T_{win}. The benefit of the last-hit detection in rejecting background light is clear: almost the same success rate of near targets in first-hit detection can be obtained for the long distance targets, even though the received optical power is lower. Concerning the performance with different values of T_{win}, we observe an almost specular behavior for shorter T_{win} values. In those cases, it is expected that the control unit of the ToF system is capable to detect a change in the SNR and switch from first to last-hit accordingly. With a 2 ns long T_{win}, however, the last-hit mode reaches a success rate of more than 80% over almost the whole range. While the probability of detecting a background photon follows an exact specular behavior with first and last-hit detections, the signal gains from the use of the last-hit mechanism, and in particular it equalizes the SNR over the measurement range, as shown in Fig. 4.14. This can be explained by the fact that the last-hit mechanism gives priority to late photons, somehow compensating for the exponential decay of the signal power as the target moves away from the LiDAR system. This is clearly visible

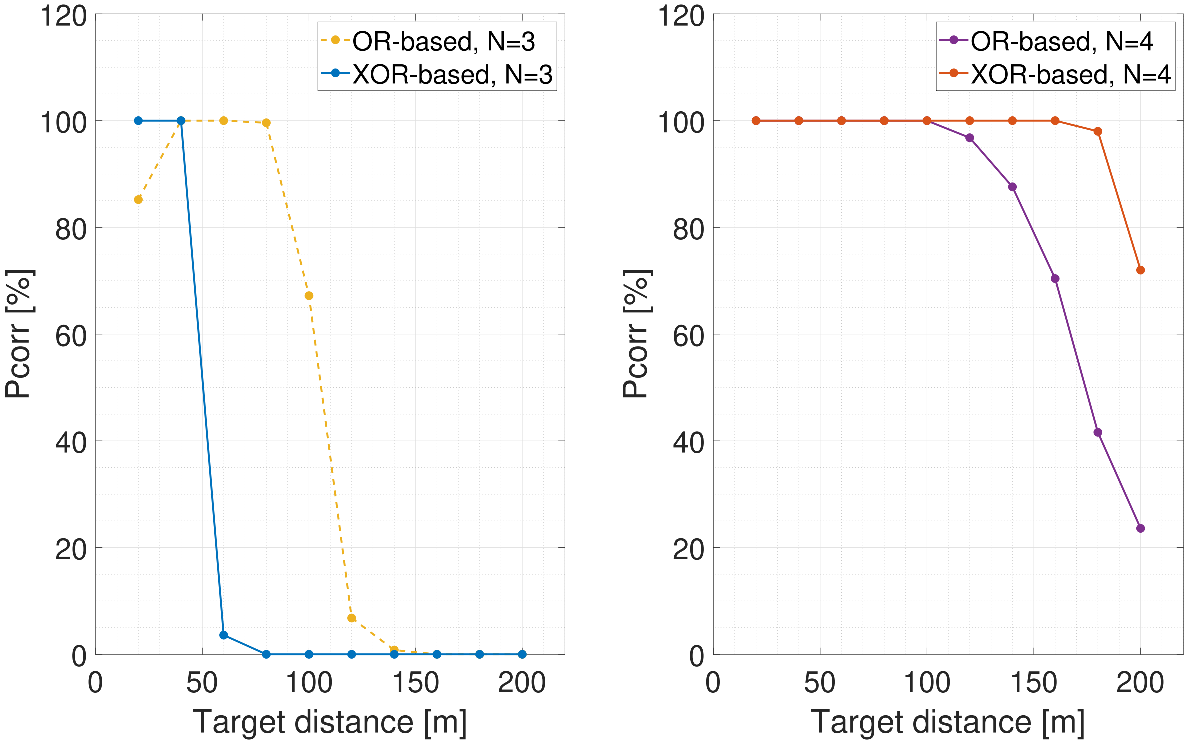


Figure 4.11: Comparison between *OR-based* and *XOR-based* implementations with a coincidence window width $T_{win} = 2$ ns, for a threshold of $N = 3$ (left) and $N = 4$ (right). The bandwidth limitation of the *OR-based*, as a collateral effect, brings a little performance improvement with $N = 3$, but it is overwhelmed by the *XOR-based* implementation with $N = 4$.

in Fig. 4.5: using the last hit mode, the peak height remains the same as the target moves from $1/3$ to $2/3$ of the range, even if the echo power at the receiver end is reduced due to the increased distance. Consequently, for particular design constraints where the system cannot implement too many options, the last-hit mode could be eligible to be the only detection scheme available.

The behavior in terms of SNR has been analyzed employing a formulation inspired by Koerner [101]

$$SNR = \frac{\Gamma \cdot \sum\limits_{i=ToF}^{N} H(i)}{\sqrt{1/N \cdot \sum\limits_{i=ToF}^{N} H(i)}}, \tag{4.1}$$

where H is the histogram of timestamps, N is the number of bins between the time of

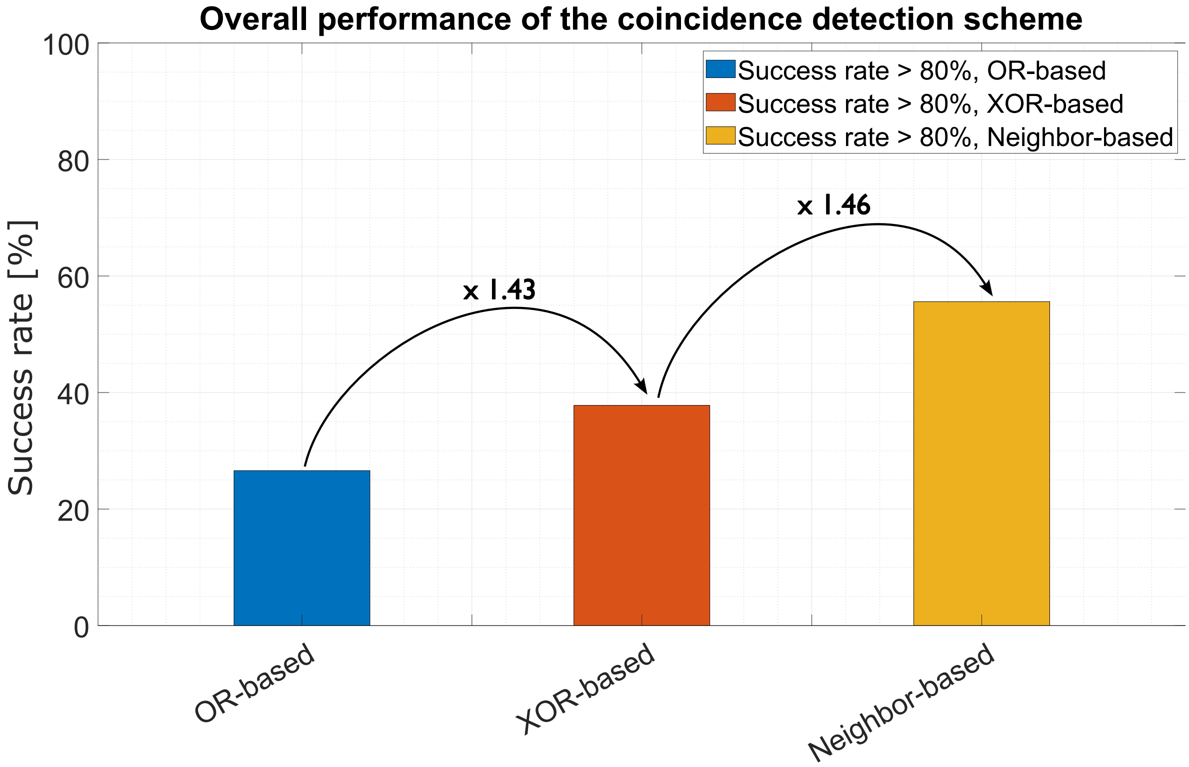


Figure 4.12: Global performance of the three different coincidence detection implementations considered. The *Neighbor-based*, thanks to the additional rejection capability given by the constraint that the detection must happen on the Primary pixel ($N = P + 1$, $P + 2$, etc) results in a 1.46 times higher performance with respect to the *XOR-based*, which in turn outperforms the *OR-based* approach by a factor of 1.43.

flight (ToF) and the end of the laser pulse ($ToF + T_W$), and:

$$\Gamma = \frac{\lambda_S}{\lambda_B + \lambda_S} \tag{4.2}$$

is the ratio between the signal intensity (λ_S) and the total light intensity over the sensor ($\lambda_B + \lambda_S$). In the employed formulation, the signal is computed as the total number of signal photons, while the noise as the standard deviation of the average number of total counts per time bin, to encompass for the shot noise of photon counts.

The performance in terms of SNR in the configuration of Figure 4.13-c is shown in Figure 4.14. In the plot, an SNR threshold which guarantees a measurement success rate of at least 80% is drawn with a dotted black line, showing that the system SNR in last-hit detection is almost flat and enough to guarantee the performance over almost the entire range, as opposed to the SNR in first-hit detection which drops below the threshold at approximately half range.

Table 4.3: List of modified parameters used in the Monte Carlo simulation for the characterization of the last-hit technique.

Parameter	Value	Unit	Description
N	3	n.a.	Coinc. threshold
T_{win}	[1; 1.5; 2]	[ns]	coinc. window width
$SPAD_n$	8+1	n.a.	number of SPADs [4]
ρ_{target}	10	[%]	target reflectivity
BG_{level}	100	kLux	background level
Coinc. scheme	[4]-2021	n.a.	Selected coinc. scheme

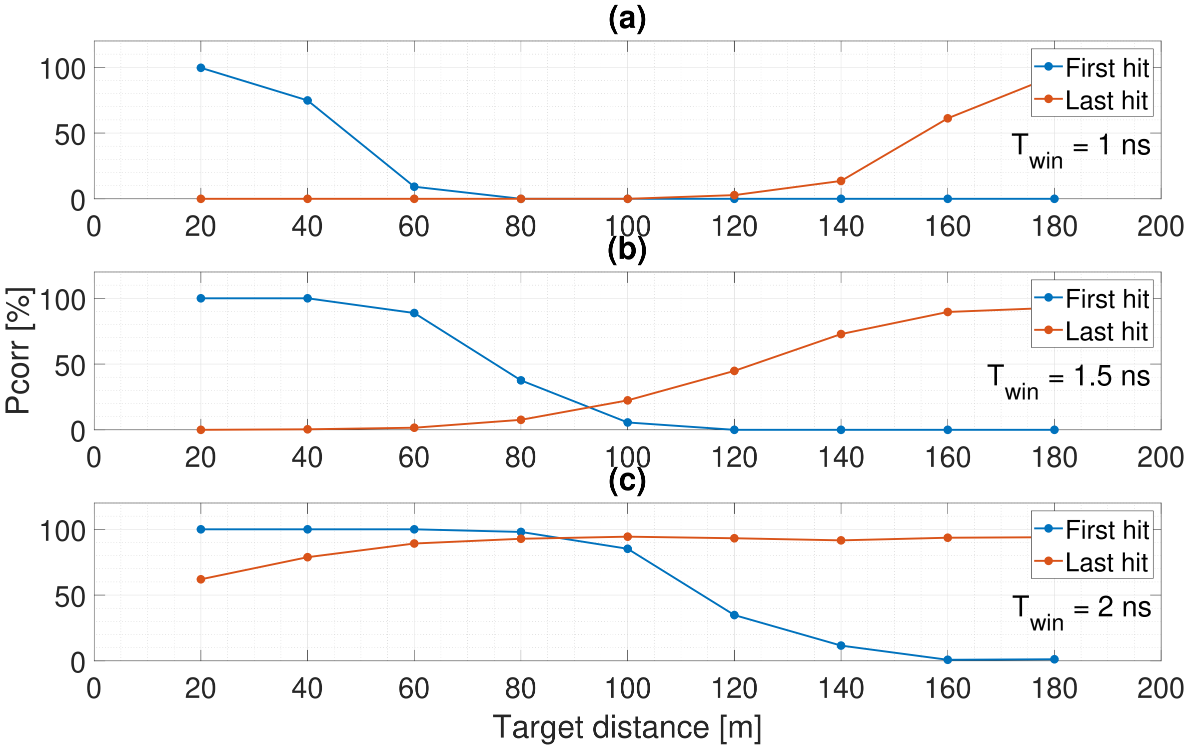


Figure 4.13: Performance comparison of first and last-hit detection techniques with T_{win} = 1 ns (a), 1.5 ns (b) and 2 ns (c).

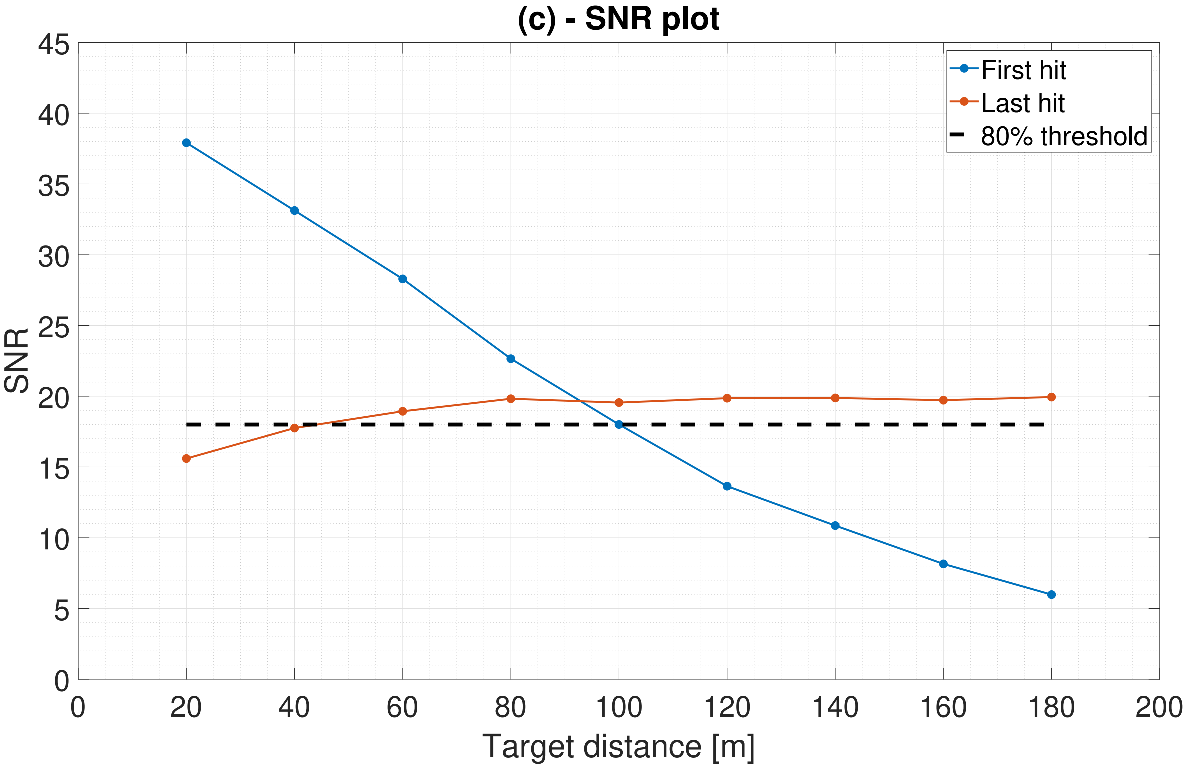


Figure 4.14: SNR behavior of first and last-hit detection with T_{win} of 2 ns, showing an almost flat SNR for the last-hit detection that guarantees a success rate of at least 80% over almost the whole range.

Considering the information obtained from the whole dataset, we show in Figure 4.15 the percentage of simulation points where the measurement success rate is higher than 80%, categorized for each of the (a), (b) and (c) cases of Figure 4.13. Considering the two techniques independently from each other (i.e., assuming it is not possible to switch them to increase the success rate), the global performance is higher with the last-hit detection strategy by approximately 7.4%. Considering, however, a ToF system capable to select the best setting for each target distance, the overall system performance increases by a substantial +77.8% with the implementation of the last-hit detection.

4.3 Automatic adaptation of pixel parameters

SPAD-based LiDAR systems are expected to operate in a wide range of scenarios, from photon-starved up to strong pile-up conditions even within the same observed scene. For this reason, it would be desirable that each pixel can select the best configuration in terms of automatic sensitivity and coincidence technique.

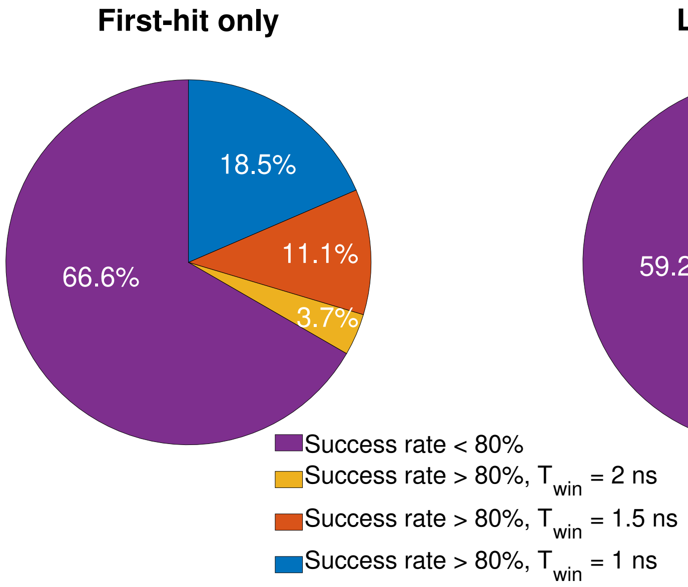


Figure 4.15: Global performance of first and last-hit detection strategies. Considering the two techniques independently from each other (i.e., assuming it is not possible to switch from one to the other), the last-hit detection achieves anyway global performance higher than 7.4% than the first-hit.

Concerning the AS technique, it has already been demonstrated [4] that a per-pixel, real-time adaptation is possible, even with a fixed threshold for the number of counted photons to threshold the intensity of background light.

The photon coincidence technique prevents sensor saturation from background-related events occurring before the arrival time of the useful signal. As detailed earlier in this chapter, the technique involves detecting coincident events within a user-defined temporal window. Ideally, the duration of the temporal window should match the width of the laser pulse, as only within this specific time frame coincident events are expected to occur. In the case where the intensity of background illumination exceeds a certain threshold, false coincident events may be triggered. While shortening the temporal window would decrease the likelihood of false triggers caused by background light, it would also lead to the rejection of photons from the reflected laser light, further compromising the ToF measurement. For such a reason, it is reasonable to assume the coincidence window as a fixed parameter, and therefore set the intensity of the coincidence filtering by means of the number of photons N. In existing many implementations [4, 6, 43, 55], the setting of the coincidence parameters is fixed and applied uniformly to the whole array of pixels. Considering that usually SPAD-based LiDAR systems need to operate in high dynamic range conditions, an optimal solution would envisage per-pixel settings, possibly with an automatic, real-time approach to set the coincidence threshold N. A similar approach has been developed by Beer et al. [80], where a per-pixel adjustment of the coincidence threshold is implemented. In this approach, however, the per-pixel adjustment require a sensor data readout, resulting in potentially long (fraction of seconds) time scale to adjust the coincidence parameter and thus hardly adaptive to fast changes in the scene illumination. A real-time, per-pixel approach in this sense can be implemented by taking inspiration from the implementation of the AS technique by Manuzzato et al. [4], which uses a pre-measurement calibration phase in which photons from background light are counted directly in pixel within a user-defined integration window to automatically decrease the SPAD sensitivity. A similar approach can be implemented to provide a pixel-level awareness of the magnitude of background illumination, to automatically adjust the coincidence by tuning the number of events to be considered (N), without the need to transfer data out from the sensor. To optimize the measurement duty cycle, the automatic calibration procedure can also be executed in parallel to the sensor data readout, thanks to the possibility to operate SPADs in a free-running fashion.

4.4 Summary

In this analysis, three background-rejection techniques for SPAD-based d-ToF systems have been analyzed and compared with each other. If considered alone, the AS technique proves effective only with a drastic reduction of the SPAD PDP, which is unfeasible with the proposed circuit [4]. On the other hand, a significant improvement in both range and probability of correct detection can be obtained by combining the two techniques. An in-depth comparison of three different circuit implementations of the photon coincidence technique shows how performance can be affected by design choices, with the *Neighbor-based* approach being overall the winning approach. Eventually, the analysis on the last-hit technique shows a significant performance increase if the proper detection strategy is dynamically selected by the system. Additionally, we show that this technique could anyway be selected as the main one in a SPAD-based LiDAR system if required by design constraints. The results from the Monte Carlo simulator have been qualitatively verified with a laboratory setup using a sensor implementing two of the techniques.

Chapter 5

A SPAD-based linear sensor with in-pixel temporal pattern detection for interference and background rejection with smart readout scheme

In Chapter 4, several background-rejection techniques from both other researcher's works [10, 43, 55] and works from our research activity [4] have been compared against each other in simulations, focusing in particular on the possibility to properly combine them to improve the sensor performance. While background light rejection remains one of the most concerning issues of SPAD-based LiDAR systems, the increased interest in their application for Advanced Driver Assistance Systems (ADAS) introduced a new challenge, i.e., the capability to reject interfering information from similar devices operating in the same environment.

In this chapter, we report on preliminary characterization results from a 1x64 pixel SPAD-based linear sensor for d-ToF applications, designed to efficiently reject both background light and interferences from similar devices. The sensor has per-pixel, real-time interference and background rejection capability and it is coupled with a smart readout scheme. Each pixel is composed by 4 SPADs with passive quenching, a digital logic circuit to exploit photon temporal coincidence with a threshold of up to 3 photons for background rejection, a finite state machine for the detection of temporal laser patterns for the rejection of interfering signals generated by other similar devices and a 16-b time-to-digital converter with 150 ps timing resolution that can be repurposed for intensity measurements. The sensor implements a smart readout scheme capable to output only pixels with meaningful data, i.e., detection events that have been validated by the photon temporal coincidence circuit and/or the laser pattern detection circuit.

The interference rejection is based on the emission of two laser pulses with a known timing signature as in [7], but in this sensor it is implemented directly on a pixel basis in a compact form and operates in real time, thus no post-processing is required on the histogram of timestamps. The benefit of an active, in-pixel interference rejection is twofold. First, also background light can be rejected, resulting in an increased signal to background ratio in the final histogram. Second, power consumption can be reduced as the TDC is activated only when two photon detections occur within the expected time frame, resulting also in a more robust operation against pile-up distortion, as the probability of saturating the TDC channel is reduced.

This chapter is organized as follows. A detailed description of the array architecture is provided in Section 5.1, focusing in particular on the pixel architecture (Section 5.1.1) and on the readout scheme (Section 5.1.2). Preliminary characterization results are then reported in Section 5.2, while in Section 5.3 we discuss on the best operating conditions where the proposed scheme is more effective and its potential limitations. Conclusions and perspectives for future improvements are discussed in Section 5.4.

5.1 Architecture

In this section, we describe the sensor in detail, focusing on the pixel architecture and on the readout scheme.

5.1.1 Pixel architecture

The pixel, designed in a 110 nm 4M CMOS Image Sensor (CIS) technology, is composed by 4 SPADs arranged as a mini digital silicon photomultiplier. Each SPAD is passively quenched by 2 thick-oxide transistors to recharge the SPAD and properly clamp the voltage to 1.2 V to be compliant with the following circuitry. Each SPAD is paired with a monostable circuit to create a temporal window for the coincidence detection circuit, which is realized in pure digital logic. A threshold of N=1/4, N=2/4 and N=3/4 events can be selected via Serial Peripheral Interface (SPI) programming. The output from the coincidence detection is fed into the measurement control circuit, which implements a finite state machine for the detection of the laser signature. The laser temporal signature can be set with 4-bit granularity, i.e., up to 16 combinations are possible. The per-pixel TDC is based on a fine-coarse architecture, where the coarse timing measurement is given by an 8-bit counter with 100 MHz clock delivered by the FPGA controller and the fine timing by a ring oscillator with 150 ps timing resolution. The TDC 8-bit counter can be repurposed to count the number of detected photons for intensity measurements. When the sensor is operated to recognize the laser pattern to reject interference, the TDC is triggered only when the second laser pulse is correctly detected, reducing unnecessary power consumption.

The measurement control circuit generates a *VALID* flag whenever a photon-detection event occurs. If the laser pattern detection feature is disabled, a *VALID* flag is generated by the first incoming event, which can be either the first detected photon (if no coincidence threshold is applied) or the first 2 or 3 photons detected within the coincidence window generated by the monostable circuit. The *VALID* flag is needed by the smart readout scheme to optimize the bandwidth by reading only pixels with validated data. The chip micrograph, array architecture and pixel block diagram are shown in Figure 5.1.

5.1.2 Readout architecture

The trend for the latest SPAD-based d-ToF sensors is to increase range, depth and imaging resolution to achieve long-range measurements with a finer depth and geometrical detail [6, 77, 102]. This, however, result in an increase on the total amount of time required to transfer the data generated by the sensor. In contrast to this, additional features for background and interference rejection, as implemented in the proposed chip design, result in a reduction of the amount of generated data, as information from background-related photons or interferences can be discarded. For those reasons, a classical scheme where the entire array is read out is not optimized, as pixels with either non-validated data or background data are anyway considered, resulting in an increased readout time, negatively affecting the sensor frame rate.

Previous works tried to solve those problems with different approaches. Zarghami et. al [15] proposed a row skipping feature and a threshold-based approach to enable the sensor readout only when a minimum selectable number of pixels have triggered. A different approach from Field et al. [103] has been proposed which makes use of a valid flag and pixel address to completely skip pixels with invalid data during the readout phase. With this approach however, the additional overhead over the number of timestamp bits is significant, adding 11 bits of overhead for 10 bits of timestamp data.

Within the design of the chip, we propose a new readout scheme which consists of

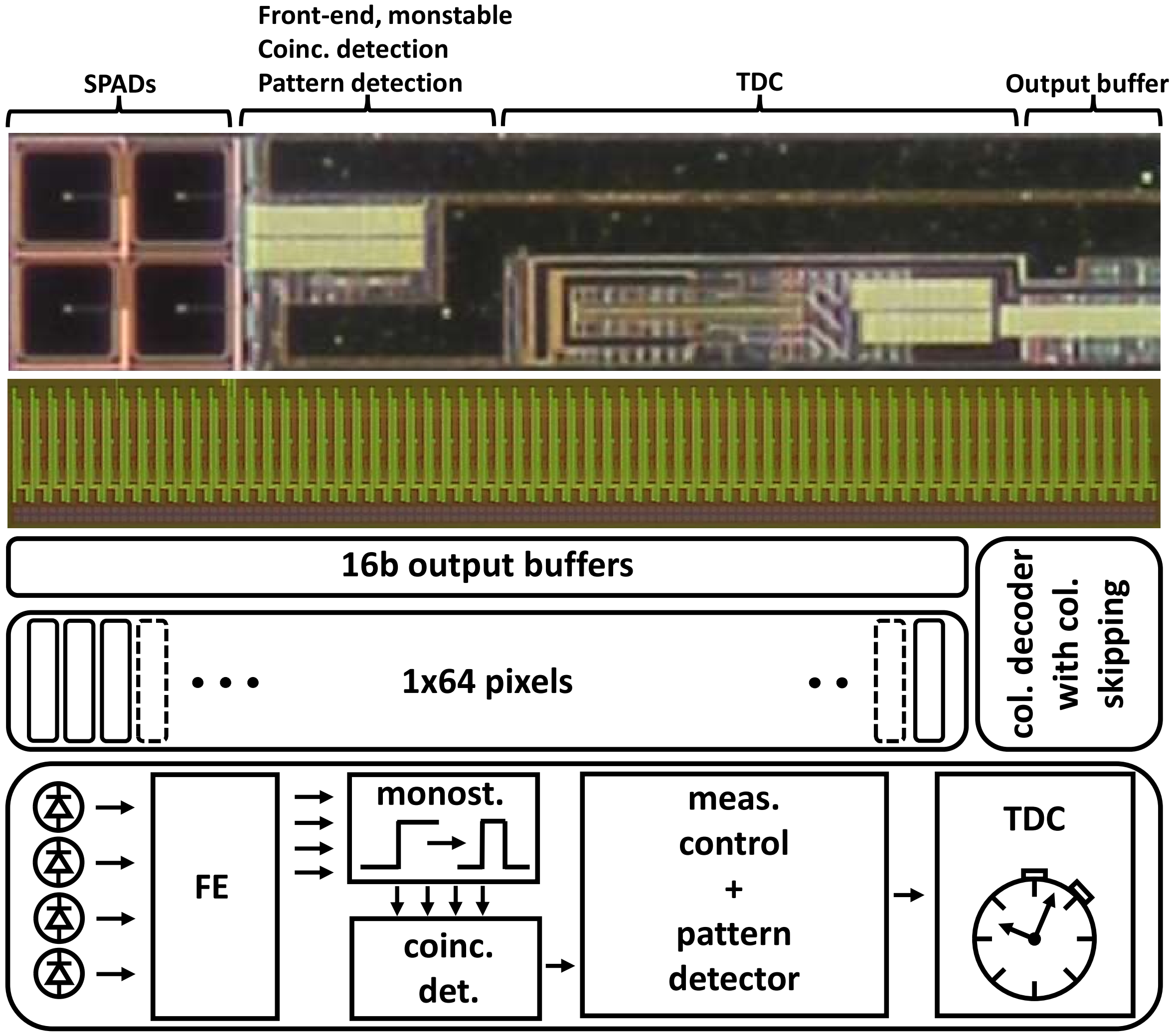


Figure 5.1: Chip micrograph, array architecture and pixel block diagram. The array of pixels is implemented in a 110 nm 4M CIS technology within a multi-project chip. Due to the reuse of the TDC from a previous project, pixel size is not optimized, resulting in a final pitch of 40 x 180 µm^2. The in-pixel measurement control block, with laser signature detection capability, has an area occupation of 28 x 14 µm^2. By considering the device from Manuzzato et al., [4], which is realized in the same technology node with a pixel pitch of 48 x 48 µm^2, the occupation of this block takes 17% of the total pixel area, thus allowing its integration also in a 2D array.

two phases: the first phase is meant to transfer 1 bit per pixel to inform the controller FPGA about which pixels contain valid data. In the second phase, only pixels with validated data are read out, thus suppressing zeros. The first readout phase is therefore only needed for the controller FPGA to associate each data values with the pixels that generated it, with a minimum overhead of only 1 bit per pixel. A graphical representation of the proposed readout scheme is shown in Figure 5.2.

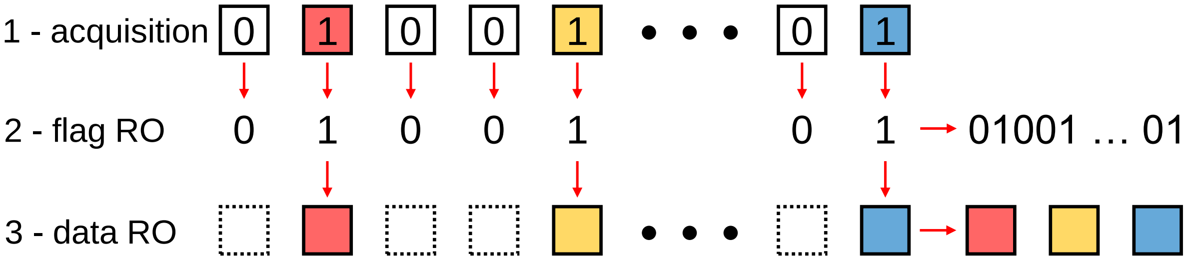


Figure 5.2: Graphical representation of the proposed readout scheme considering a row of pixels. In the first readout phase a 1 bit flag information is transferred for each pixel. In the second phase, only pixels with flag equal to '1' are actually transferred, minimizing the readout time.

We compare our approach against two address-based approaches as follows:

- Full address-based readout: the sensor implements pixel skipping feature enabling the readout only for pixels containing valid data. The pixel position is encoded by adding the full address coordinates to identify row and column.
- Half address-based readout: similarly to the full address-based readout, skipping feature is implemented. However, only half bits encoding the pixel address are needed, since the row selection information is provided by means of dedicated signalling directly in the readout phase.
- Binary mask-based readout: in our approach, pixel skipping feature is implemented. However, no address is required, as for each pixel a binary flag indicates with the value of '1' if valid data is contained, otherwise the value is '0'. Only if the binary flag has value '1', the selected pixel outputs its timestamp data, otherwise the next pixel is selected. The binary flag is readout anyway to assign each pixel data to the correct array coordinates.

For the comparison, we consider a 64×64 pixel array with 16 bits timestamp. Table 5.1 provides details about the additional overhead required for each scheme and specifies the total amount of readout bits. We compare our methods against the other by using two different references: a complete readout (100 % of the data is readout regardless of the true sensor activity) and an ideal readout, where only the values of triggered pixels are readout without any additional overhead to account for their position. The results are provided in Figure 5.3, where the percentage of readout data is shown with respect to the percentage of triggered pixels. Due to their variable amount of overhead, the two address-based readout schemes perform better up to $\simeq 8$ % and $\simeq 16$ % of triggered pixels, respectively. From this point on, the binary mask-based

Table 5.1: Comparison parameters for readout schemes considering a 64×64 pixel array with 16-bit timestamp. The number of triggered pixels is indicated with N_{trig}. The address-based readout schemes feature a variable quantity overhead to encode the valid pixels, while the proposed approach provides a lighter fixed-quantity overhead which is globally optimum over the whole range of triggered pixels.

Readout type	Overhead		Readout bits
	Type	Quantity	
Full address	variable	$N_{trig}\times 12$	$N_{trig} \times (12 + 16)$
Half address	variable	$N_{trig}\times 6$	$N_{trig} \times (6 + 16)$
Binary mask	fixed	64×64	$64\times 64 + N_{trig} \times 16$
Full readout	–	0	64×64×16

scheme outperforms thanks to its limited overhead encoding only one additional bit per-pixel. Additionally, the advantage of the two address-based schemes with respect to a complete readout ends when the percentage of triggered pixels reaches $\simeq$ 56 % and $\simeq$ 72 %, respectively. From that point on, the additional overhead makes a complete readout convenient, thus limiting the dynamic range of array activity where such scheme is beneficial. On the other hand, the proposed binary mask-based readout is convenient up to $\simeq$ 94 % of triggered pixels, adding in the worst case of 100 % array activity a modest +6 % of readout data. At 100 % array activity, the address-based methods add a considerable +37,5 % and +75 % of readout data, respectively, thus completely jeopardising any previous advantage. In the case of even larger array sizes, which are nowadays becoming more and more common due to the advancements in 3D stacking and newer technology nodes, the advantage of the proposed readout scheme becomes even higher if compared to the other approaches. This is due to the fact that the per-pixel overhead is constant (1 bit per pixel), while for the address-based readout it scales logarithmically with the size of the array.

The proposed readout scheme is the one that better approaches the ideal case, making it a good candidate to be employed and always used in a wide range of scenario, from photon-starved to saturated-array conditions. Moreover, it well matches with the in-pixel implementation of the laser signature detection, as the generation of data from the sensor can be highly reduced with such a strong filtering. Consequently, the benefit in terms of performance is twofold: on one side, the reliability of the timestamp detected with the in-pixel finite state machine for the laser pattern detection is increased, and on the other side the reduced amount of data (mainly due to the filtered events) results in a decrease of the required readout time, with benefits in terms of frame rate and power consumption.

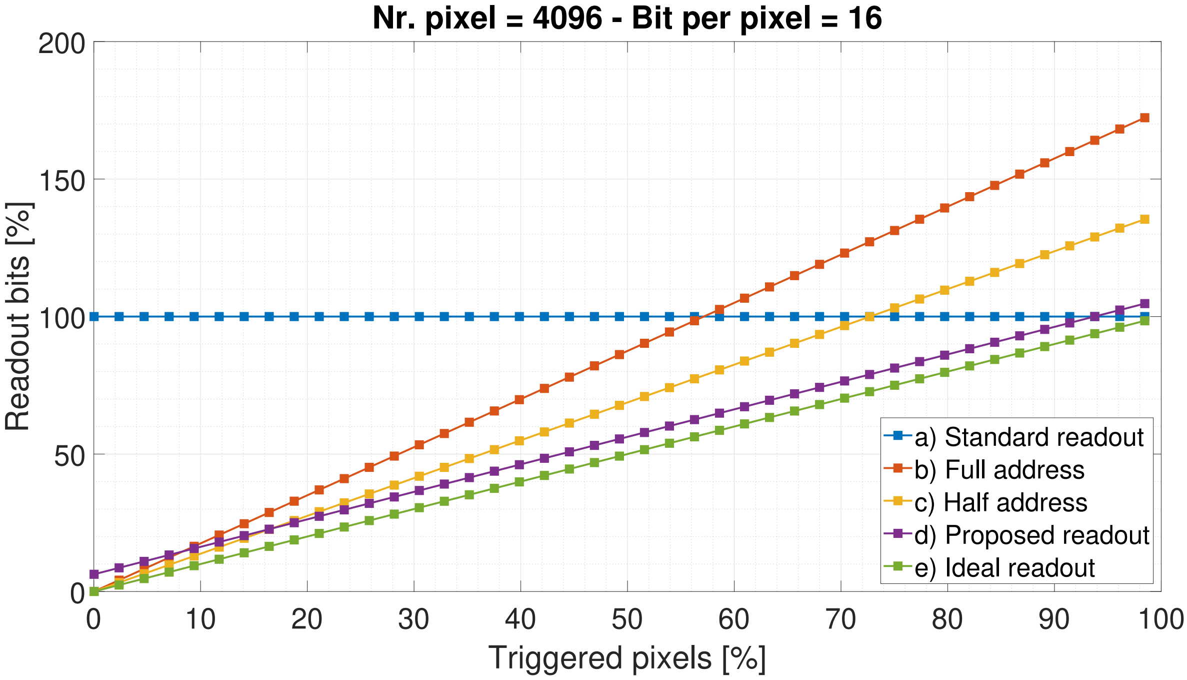


Figure 5.3: Comparison of different readout schemes in terms of percentage of readout data for a range of triggered pixels between 0 % and 100 %.

5.2 Characterization

In this section, characterization results focusing on the in-pixel laser pattern detection and readout circuit performance are shown.

5.2.1 In-pixel laser pattern detection characterization

The in-pixel laser pattern detection feature has been first characterized on a single pixel basis using only the coarse TDC (100 MHz counter) information by means of two low-power picosecond pulsed lasers. The first laser was meant to be the signal source, while the second was used as an interferer. A picture describing the experimental setup is shown in Figure 5.4. The sensor is operated in the presence of background light, without any optical bandpass filter in front of the detector and with a coincidence threshold of N=2 photons.

The results of the first characterization are shown in Figure 5.5, demonstrating that an interference with a histogram peak 18.5 dB higher than the signal of interest is almost completely eliminated, with a suppression of 42.5 dB. On the other hand, the signal of interest gained 10 dB with respect to the case where no laser pattern matching detection was applied, demonstrating that the in-pixel active detection of laser temporal

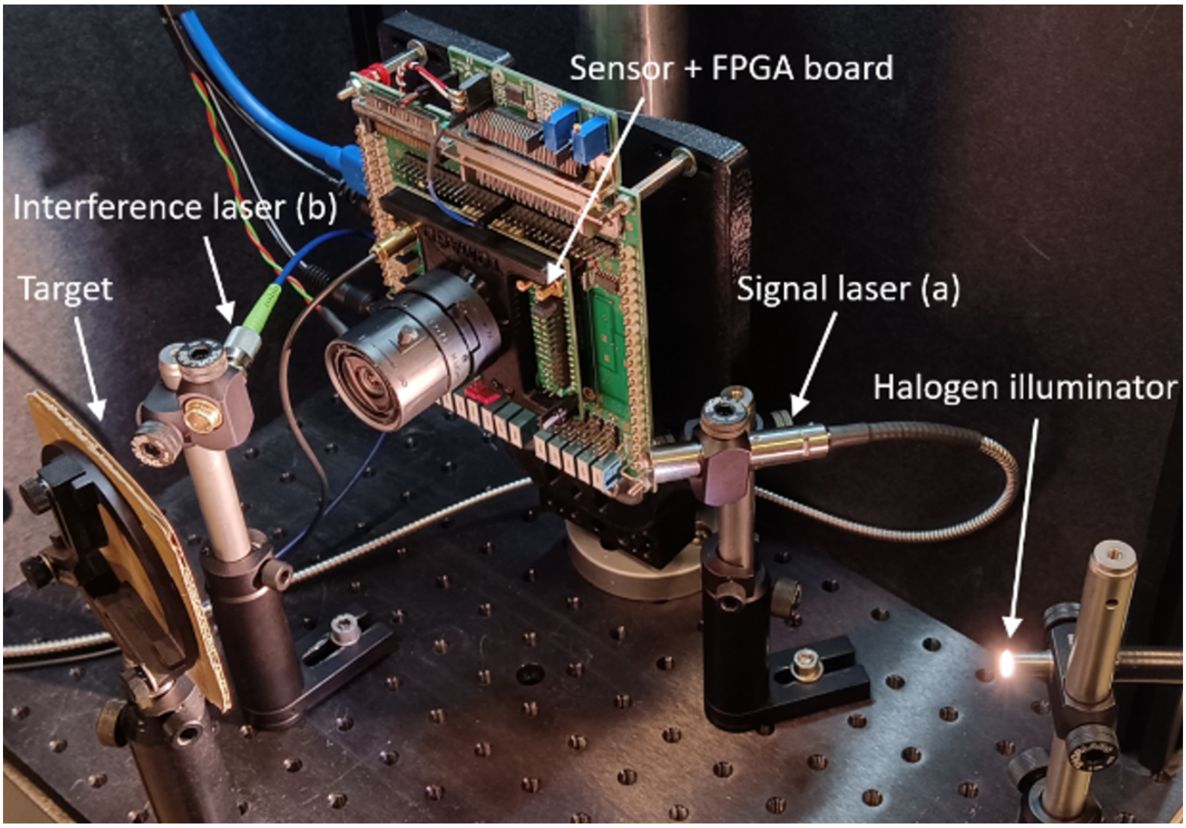


Figure 5.4: Picture of the experimental setup to test the in-pixel interference-rejection capability with the signal laser (a) and the interfering laser (b). Each laser has been set with its own timing signature: for the signal of interest, the two laser pulses are separated 80 ns from each other, while for the interfering signal the pulses separation is 90 ns.

patterns also helps in reducing uncorrelated background light.

Then, the entire array has been characterized with the 3D measurement of a scene profile, using the fine resolution given by the 150 ps in-pixel ring-oscillator and employing two 25 W 905 nm lasers for the emission of the temporal laser pattern. A third laser was pointed toward a portion of the scene, to emulate the interference from a second device. Results are shown in Figure 5.6.

With a third measurement we targeted the background-rejection capability of the sensor by focusing on the combination between the photon coincidence technique and the in-pixel laser pattern detection. The sensor was operated without optical bandpass filter and a 180 W halogen illuminator was used to flood the scene with background photons. Without any background/interference rejection features, it was impossible to reconstruct the 3D profile of the scene. The 3D information was recovered completely with the photon coincidence technique (operated with a threshold of N=2/4 photons), but still several background events are allowed to trigger a coincidence, increasing the

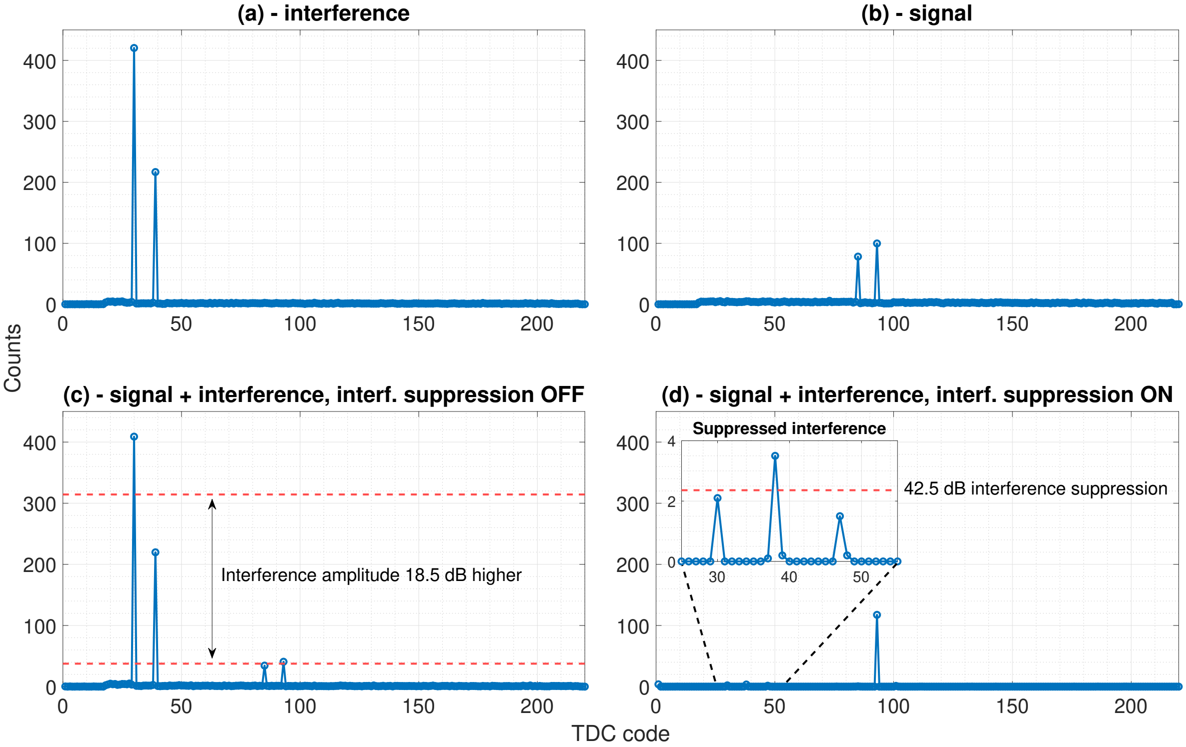


Figure 5.5: First experimental validation of the interference-rejection capability of the device. Histograms (a) and (b) show the contribution of interference (a) and signal (b) alone, while in (c) the joint effect of the two sources is shown (with laser signature detection disabled). To stress the interference-rejection capability, the interference histogram peak is 18.5 dB higher in amplitude than the peak of the signal of interest. When the in-pixel interference-rejection is enabled (d), the interfering signal is almost completely suppressed by 42.5 dB, whereas the signal of interest gains 10 dB with respect to case (c), enabling the possibility to build an interference-free histogram directly from pixel data.

amount of data generated by the sensor and thus the required readout time. By enabling the laser pattern detection on top of the photon coincidence, it was anyway possible to reconstruct the 3D profile of the scene and at the same time dramatically reduce the amount of false triggers, with a reduction of data of $\simeq 98\%$ with respect to the previous case. By considering the laser peak to background ratio of the histogram, the combination of the laser pattern matching on top of the photon coincidence detection allowed to gain $\simeq 23$ dB, thus increasing the robustness of the measurement. Results are shown in Figure 5.7.

5.2.2 Readout performance assessment

The readout architecture has been characterized in two ways. The first characterization aimed at showing the capability of the sensor to output only pixels with meaningful data.

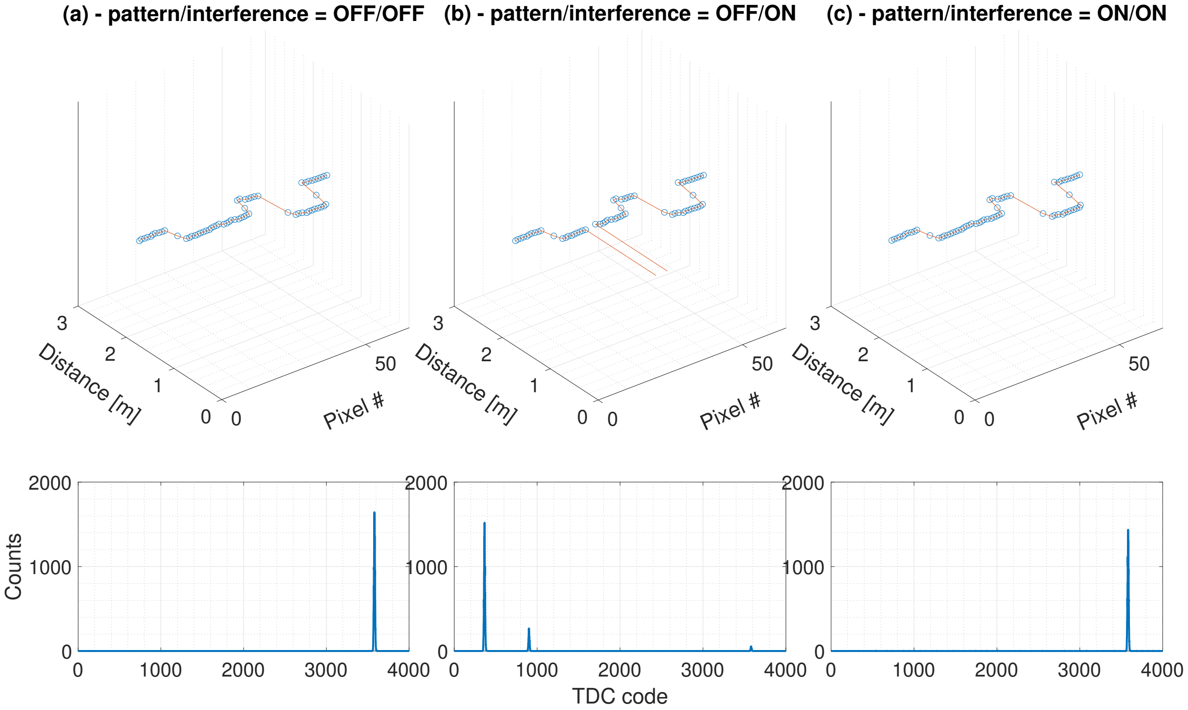


Figure 5.6: Measurement of the profile of a scene composed by 2 boxes with different size and distance (range 0-2m). The first measurement (a) is a reference obtained without any interfering signal and with pattern detection disabled. In the second measurement (b), the interfering laser is pointed toward the first box with an earlier timing with respect to the signal laser, resulting in a complete loss of information from the illuminated portion of the target. In the third measurement (c), the in-pixel laser pattern detection was enabled allowing to completely recover the lost information, as the interfering laser timestamps are actively discarded in favor of the signal laser timestamps, which have the correct timing. For each measurement, the histogram of one pixel under the interfering portion of the target is shown.

For this measurement, a target (flat panel) was illuminated completely by background light and a collimated laser source was used to illuminate only a portion of it, resulting in only 2 pixels of the array to be shined with laser photons. First, no background-rejection technique was applied nor laser pattern matching detection, resulting in a probability of detection of almost 1 for all pixels. Then, photon coincidence detection was enabled with a threshold of N=2 photons, and a general reduction of the probability of detection is observed, but still allowing background-only pixels to be triggered. In the last measurement, the in-pixel laser pattern detection was enabled on top of the photon coincidence, reducing the probability of detection on pixels not illuminated by the laser source to a negligible level, allowing to only output data from the subset of pixels illuminated by the laser. Results are shown in Figure 5.8.

In the second characterization, only background light was considered and the sensor

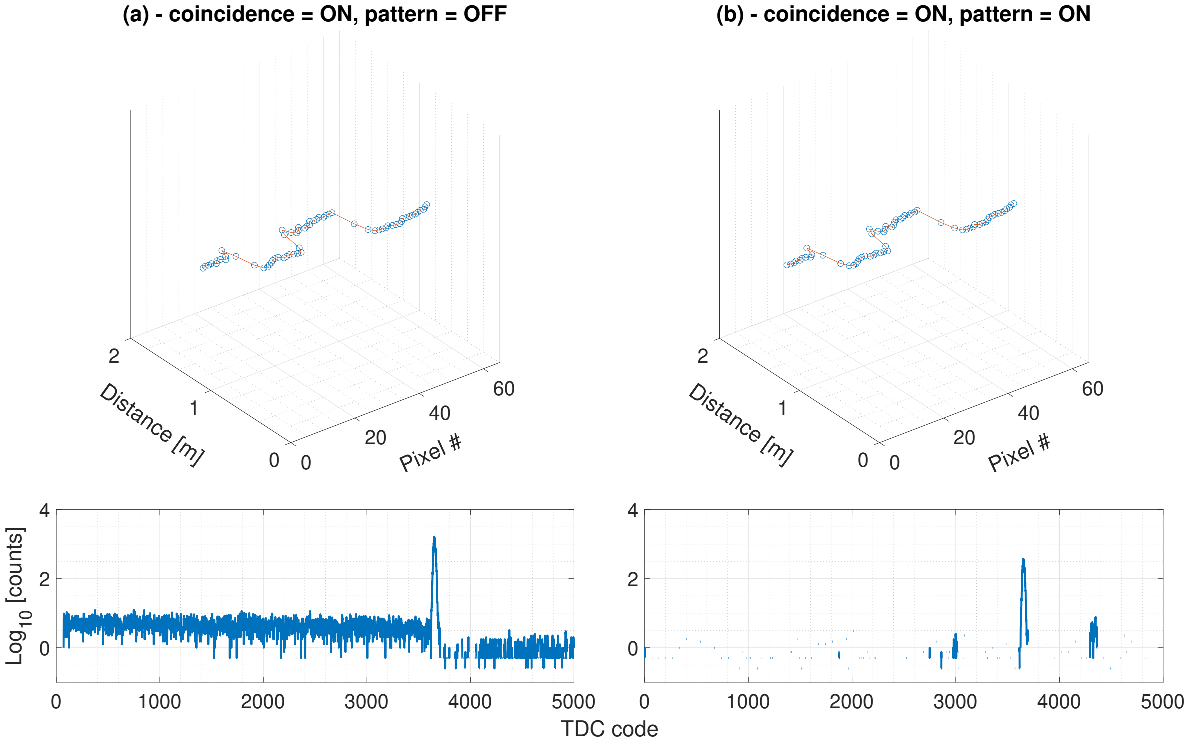


Figure 5.7: Measurement of the 3D profile of a scene with high background light. In (a), only the photon coincidence technique was enabled, with a threshold of N=2/4 photons. In (b), the laser pattern detection was enabled on top of the photon coincidence, resulting in ≃ 98% data reduction with respect to (a). The laser peak to background ratio is ≃ 50 dB in (a) and increased up to ≃ 73 dB in (b).

internal frame rate was measured by controlling the intensity of the light source. Results are shown in Figure 5.9, showing the capability of the sensor to adapt to different levels of photon activity.

5.3 Operating conditions and limits

As discussed previously, the proposed interference rejection scheme works by emitting two consecutive laser pulses which must be detected within the same measurement cycle to validate the ToF measurement against background light or interfering sources. This architecture poses a well defined limit over the operating conditions where it is beneficial as opposed to the conditions where performance may be degraded even more.

Concerning the emission of two laser pulses, this may result in a potential violation of eye-safety limits, thus requiring to either lower the peak power or emission frequency. On the receiver side, the detection of two consecutive laser pulses may be unlikely if the signal strength is not sufficiently high. For such reasons, the proposed scheme is best

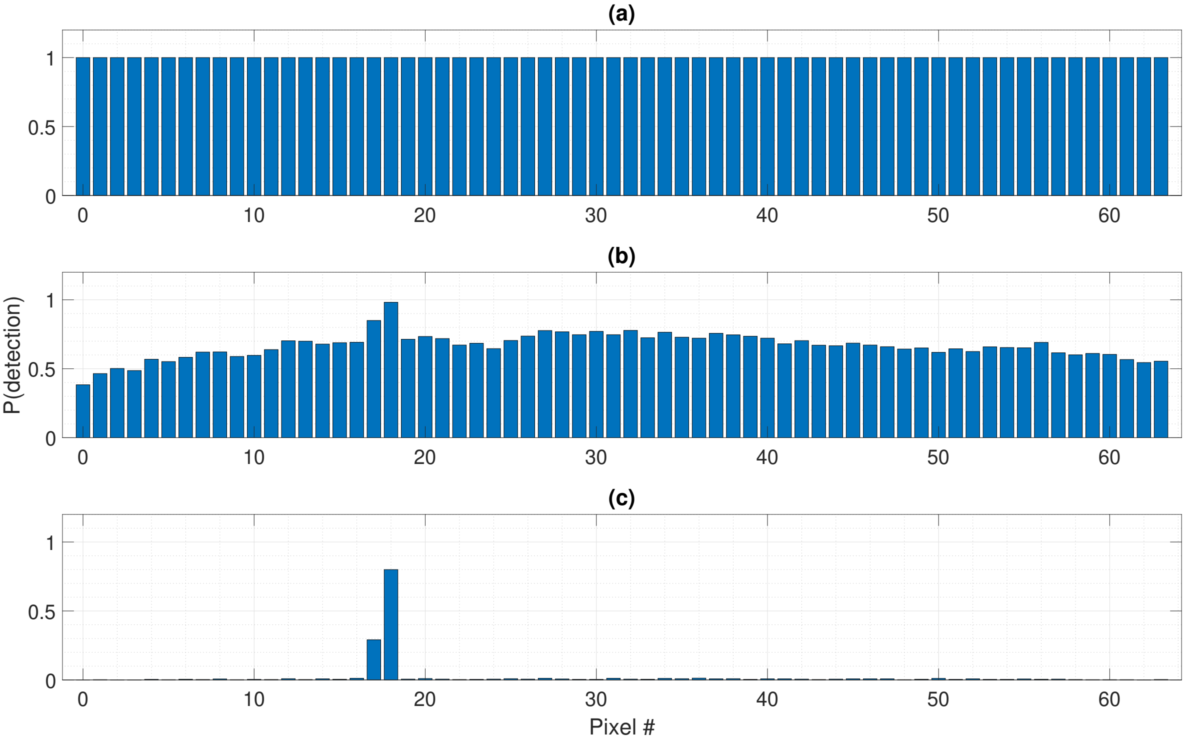


Figure 5.8: Measured per-pixel activity (in terms of probability of detection) with the proposed readout scheme. With a minimum 1-bit overhead per pixel, it is possible to output only pixels with validated data. In (a), the readout ratio is almost 100% as no data filtering technique was applied. In (b), the enabling of the photon coincidence technique allows to reduce the readout ratio down to 66%, with a visible readout peak activity over pixels number 17 and 18, as they coincide with the reflected laser footprint. In (c), the additional enabling of the in-pixel laser pattern matching detection allows to recover the information from the only two illuminated pixels, further reducing the readout ratio down to a minimum of 2.25%.

suited to be implemented within a scanning LiDAR architecture, as the possibility to use a collimated laser source (resulting in a favorable signal to noise ratio) maximizes the probability of signals detection. Concerning the measurement range, for similar reasons the interference rejection feature is expected to provide the best results in the medium-low range, as the expected return of signal intensity degrades quadratically with the distance.

Nevertheless, the proposed sensor architecture leaves the user complete freedom in the selection of the best combination of background/interference rejection features, as every functionality (number of threshold photons, coincidence window, interference rejection) can be used and tuned independently from each other, maximizing the possibility to find the optimal setting for a given scenario.

As discussed in Chapter 4, the proper combination of two (or more) detection techniques can improve the performance of the system. Considering the wide range of

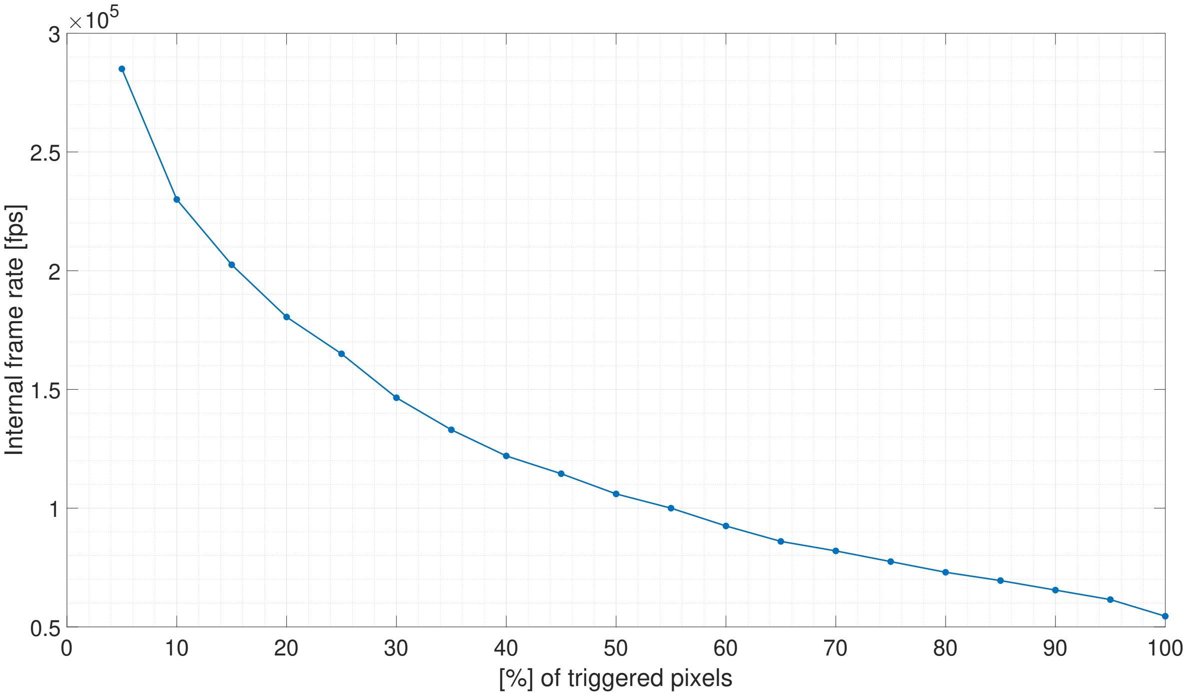


Figure 5.9: Measured sensor internal frame rate (thus not considering PC processing time) for different number of triggered pixels, demonstrating the capability of the proposed readout scheme to adapt to the level of photon activity.

scenarios in which a SPAD-based LiDAR system is expected to work, from photon-starved up to strong pile-up conditions, it is advisable to design such systems with the maximum number of features. This would provide them with the highest flexibility to adapt to different conditions, increasing the possibility of successful measurements.

Concerning the readout scheme, as discussed in Section 5.1.2, we demonstrate that it represents the solution that best approaches an ideal readout for the widest range of sensor activity. Therefore, only when the sensor is nearly saturated in terms of active pixels (i.e., pixels with valid data), it is advisable to disable the feature and use a standard readout approach.

5.4 Summary

In this chapter, a SPAD-based d-ToF linear sensor with the first in-pixel interference rejection capability has been demonstrated. In combination with the dedicated readout architecture, it enables the acquisition of interference-free data with an optimized, adaptive, readout time. The sensor capability to reject interference from other laser sources has been demonstrated by preliminary results in a laboratory setup, as well as the opportunity to use the interference rejection feature on top of the classical photon coincidence to reduce background light to a negligible level, resulting in a measured 42.5 dB interference suppression and up to 23 dB signal gain with a measured data compression ratio of $\simeq 98\%$. The pixel pitch in its actual form factor is not optimized for the evolution into a 2D array and for such reason we envisage a further development with a more compact TDC to extend the proposed pixel architecture into a 2D imager array.

Chapter 6

SPAD response linearization: histogram-less LiDAR and high photon flux measurements

In Chapter 4 and Chapter 5, we focused on two major limitations of modern SPAD-based LiDAR systems, i.e., the presence of uncorrelated background light and interferences from other devices. In this chapter, we deal with another performance-limiting factor, that is the continuously increasing amount of generated data. By means of extensive use of system modelling (both analytical and numerical), we define a new concept for SPAD-based d-ToF systems, which is eventually supported by real measurement data from an existing sensor.

The demand for SPAD-based d-ToF sensors with continuously higher imaging and depth resolutions directly conflicts with the amount of data which is generated and must be handled. A typical d-ToF measurement operation includes the construction of a per-pixel histogram of timestamps from the repetition of several laser shots. Therefore, the amount of data from this process scales linearly with the imaging resolution (number of pixels) and the depth range and resolution (number of bits per timestamp and number of histogram bins). Without the need to go for sensors with megapixel-level resolution, even relatively low resolutions can demand significant amounts of memory. For instance, a low-resolution sensor (64x64 pixels) with 16-bit timestamps [10] width and 8-bit histogram depth, would require $\simeq 2.15 \times 10^9$ bit of memory, which is hardly feasible to have and manage even on high-end μC and FPGAs. In addition to the memory requirement, also the data rate could reach very high levels. In the context of a typical flash LiDAR architecture featuring 128x128 imaging resolution [6], 14-bit TDC depth and a modest imaging rate of 25 FPS, the resulting data rate would be $\approx 1.4 \times 10^9$ bits per second, significantly exceeding the capabilities of commonly used communication standards.

In addition to the challenges of data quantity and data rate, and considering that ADAS is currently one of the most interesting and emerging field of application for SPAD-based LiDAR systems, guaranteeing reliable and safe operation under high photon flux conditions is the other key enabling factor for the successful deployment of this technology.

The main issue of a very high photon flux, which can be easily present in outdoor scenario with very high sunlight illumination, is the distortion of the histogram of timestamps due to the well-known pile-up phenomena, resulting in an intensity-dependent ToF computation. This intensity-dependent distortion in the histogram of timestamps is due to the non-linear response of the system to the incident flux of photons over time, which in turn depends on the dead-time of SPAD detectors and timestamping circuits.

In this chapter, we propose a SPAD response linearization method that can address both the issue of data rate and the problem of pile-up distortion arising from high photon flux. The proposed SPAD linearization method, which is supported by a rigorous mathematical model, can extract the 3D information from a set of acquired timestamps without the need to build a histogram and sustain high photon fluxes enabling the possibility to operate beyond the standard limit of 5% detection rate [104] for pile-up distortion. The method can be implemented using only two registers and one accumulator for each pixel. With such a low amount of resources, the per-pixel memory requirement is reduced by more than 3 orders of magnitude compared to standard d-ToF architectures (off-chip histogram) [6, 8], and by a factor of $\simeq 5$ compared to architectures with on-chip, full histogramming capability [105]. The goal is reached in two steps. First, we propose and evaluate an algorithm to efficiently extract the target distance from a set of timestamps based on a simple on-the-fly average operation, which does not require the allocation of a histogram memory. Then, since the proposed algorithm works on the assumption that the detector response is linear, we present two acquisition schemes that can be easily implemented on chip and emulate the behavior of a single photon detector with no dead-time, providing the desired linear response to the input flux of photons. The proposed method is supported by analytical and numerical (Monte Carlo) models and has been validated experimentally up to a distance of 3.8 meters (mainly limited by the sensor used for characterization [17]) under a background light equivalent to 85 klux and beyond the standard 5% rule for pile-up distortion [104].

Table 6.1: List of parameters for a typical d-ToF acquisition.

Parameter	Unit	Description
λ_B	s^{-1}	Background events rate
λ_S	s^{-1}	Reflected laser events rate
T_W	s	Laser pulse duration
ToF	s	Time of flight

6.1 Preliminary validation

In this section, we present the principle of operation of a SPAD-based d-ToF system with preliminary considerations and Monte Carlo simulations on the histogram-less approach which will be further developed in the paper.

6.1.1 Typical d-ToF operation

A typical d-ToF image acquisition requires a pulsed laser and a time-resolved, single photon image sensor with photon timestamping capabilities. It works by sending periodic laser pulses and then measuring the arrival time, or *timestamp*, of the first detected photons reflected by the target following each pulse. Due to space and bandwidth limitations, the number of photon timestamps generated per laser pulse is typically limited to one.

In principle, a single laser pulse, and thus a single timestamp, would be sufficient to estimate the time of flight. However, due to the presence of uncorrelated background events (from both external light sources or internal SPAD DCR) and of shot noise, the first detected photon may not be from the laser pulse, so that several repetitions are needed to discriminate the different contributions. To do so, the timestamps measured during the acquisition process are collected in a histogram memory that records how many times each timestamp has been observed. This provides a convenient representation of the temporal distribution of the arrival times, as shown in the example of Figure 6.1. In a system capable of acquiring only one photon (the first), the distribution of the arrival times is a piece-wise exponential curve, where each segment is described by:

$$P_i(t) = A_i \cdot e^{-\lambda_i \cdot t} \tag{6.1}$$

whose *intensity* (rate) λ_i depends on the intensity of background light, dark counts and laser echo. Table 6.1 summarizes the most important parameters of the detection process.

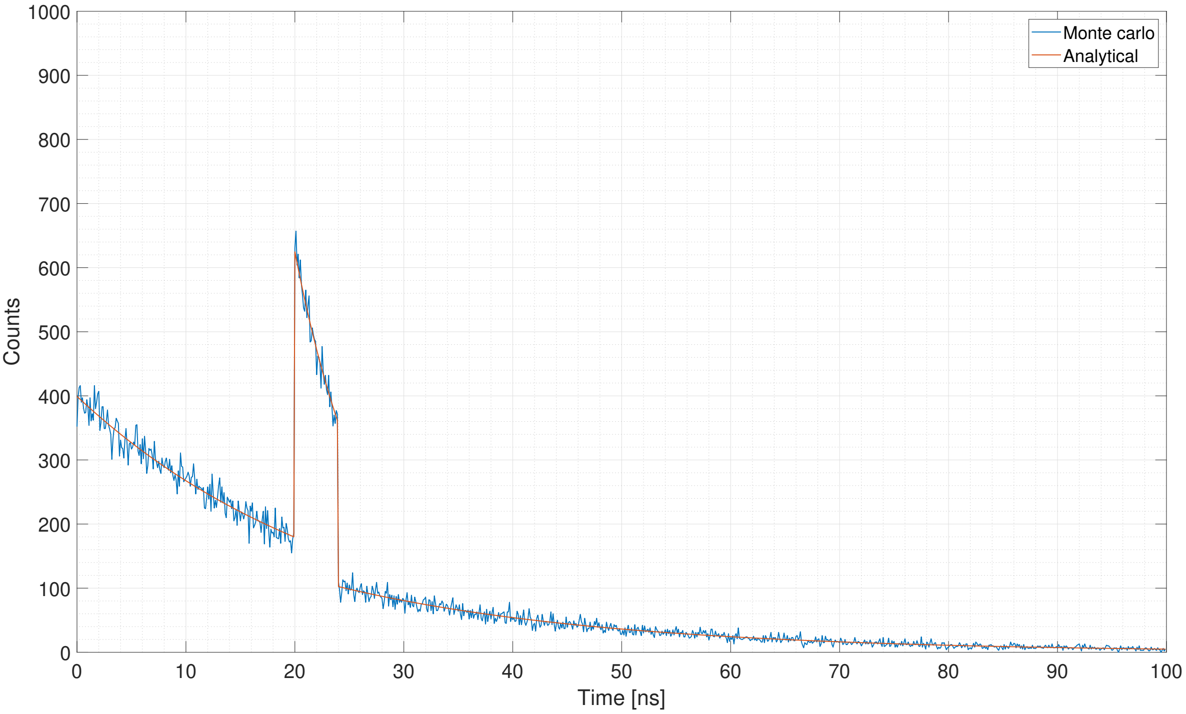


Figure 6.1: Simulated distribution of timestamps in a typical d-ToF system able to record 1 photon per acquisition, with a *ToF* of 20 ns and a laser pulse duration T_W of 4 ns. The histogram is composed of 10^5 timestamps, with a bin size of 100 ps. Superimposed to the Monte Carlo simulation, we show also the analytical exponential distribution.

The *ToF* is typically estimated from the histogram by finding the location of its peak or of a sharp rising edge, which likely belongs to the reflected laser pulse. The histogram of timestamps contains all the relevant information to properly estimate the time of flight, and represents the gold standard processing technique in the field of SPAD-based d-ToF systems. Unfortunately, the histogram requires a considerable amount of resources in terms of memory, bandwidth and power, as it requires the readout of every timestamp from the sensor for processing by an external controller (FPGA or μC). Even with the latest implementations where the histogram is available on-chip, the required amount of resources is considerable. As an example, a 10-m-range 128 × 128 LiDAR system with a 100 ps time resolution and 8 bit histogram depth requires approximately 10 MB of memory.

6.1.2 Histogram-less approach

Intuitively, if no background events are present, and neglecting the width of the laser pulse, we could estimate the time-of-flight without the need to build a histogram. This can be achieved by simply calculating the average of the continuous stream of laser-only timestamps. To extend the above method to scenarios where background events are also present, we need to eliminate their contribution to the average. Again, intuitively, this can be accomplished by dividing the measurement into two acquisitions. The first is performed with the laser turned off, and is used to estimate the contribution of the background light only, by computing the average $\bar{t}_{bg}$ of the recorded timestamps. The same operation is repeated in the second acquisition with the joint contribution of background and laser timestamps, resulting in a total average time $\bar{t}_{tot}$. In principle, the time of flight could be estimated as a linear combination of the averages[1] as

$$ToF \propto k_{tot} \cdot \bar{t}_{tot} - k_{bg} \cdot \bar{t}_{bg}\,, \tag{6.2}$$

with the contribution of the background canceling out. The coefficients k_{tot} and k_{bg} depend on the number of detected photons, hence they are affected by shot noise, with a direct impact on the measurement precision.

This approach, however, relies on the superposition property which does not hold, as SPADs are non-linear detectors.

More specifically, the problem lies in the dead time of the detection process, which depends upon the SPAD dead-time itself, on the limited bandwidth of the timestamping circuit and also on the limited memory available, blinding the measurement channel for some time after each detection. With this limitation, the detection of a photon belonging to the laser echo does prevent a later photon from being potentially detected, resulting in a distortion of the statistics. In particular, the amount of background photons which contribute to $\bar{t}_{tot}$ in the second acquisition (with the laser turned on) is underestimated. This behavior can be observed on the histograms of Figure 6.2, where the distribution of timestamps in different scenarios are compared, emphasizing the estimation error. Furthermore, the greater the laser echo intensity, the higher the number of background photons that are underestimated.

This can also be seen analytically by computing explicitly the cumulative distribution function F of the random variable T associated to the first photon detection time, defined as $F(t) := \mathbb{P}(T \in [0,t])$. If $t \in [0, ToF]$ then the incoming photon necessarily belongs to the background, hence

$$F(t) = \mathbb{P}\left(T \in [0,t]\right) = 1 - e^{-\lambda_B t}$$

---

[1] For the precise formulation, see equation (6.6).

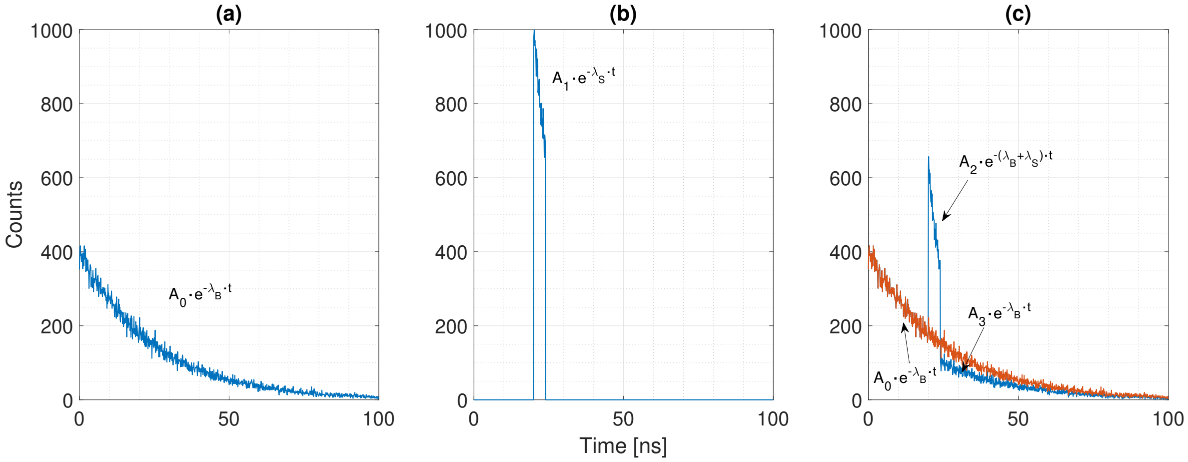


Figure 6.2: Simulated distribution of timestamps in a typical d-ToF acquisition. In (a) and (b), the distributions of background-only events (with rate λ_B) and laser-only events (with rate λ_S), respectively, are shown. In (c), the distribution of the combination of background and laser events is reported, graphically showing that the superposition property does not hold due to the non-linear behavior of the detection process. In particular, the portion of background events after the laser peak are underestimated, as only one photon per acquisition can be detected. For each contribution, the amplitude terms of the exponential (A_0, A_1, A_2 and A_3) are reported, with $A_3 < A_0$ due to the SPAD non-linearity.

If $t \in (\mathit{ToF}, \mathit{ToF} + T_W]$, then

$$\begin{aligned} F(t) &= \mathbb{P}(T \in [0, \mathit{ToF}]) + \mathbb{P}(T \in (\mathit{ToF}, t]) \\ &= 1 - e^{-\lambda_B \mathit{ToF}} + e^{-\lambda_B \mathit{ToF}} \left(1 - e^{-(\lambda_S+\lambda_B)(t-\mathit{ToF})}\right) \end{aligned}$$

Finally, for $t > \mathit{ToF} + T_W$ we have:

$$\begin{aligned} F(t) &= \mathbb{P}(T \in [0, \mathit{ToF}]) + \mathbb{P}(T \in (\mathit{ToF}, \mathit{ToF} + T_W]) + \mathbb{P}(T \in (\mathit{ToF} + T_W, t]) \\ &= 1 - e^{-\lambda_B \mathit{ToF}} + e^{-\lambda_B \mathit{ToF}} \left(1 - e^{-(\lambda_S+\lambda_B)T_W}\right) + e^{-\lambda_B \mathit{ToF}} e^{-(\lambda_S+\lambda_B)T_W} \left(1 - e^{-\lambda_B(t-\mathit{ToF}-T_W)}\right) \end{aligned}$$

The probability density f of the distribution of T, where $F(t) = \int_{-\infty}^{t} f(u)du$, can be easily computed by means of the formula $f(t) = F'(t)$, yielding:

$$f(t) = \begin{cases} \lambda_B e^{-\lambda_B t} & t \in [0, \mathit{ToF}] \\ (\lambda_S + \lambda_B) e^{\lambda_S \mathit{ToF}} e^{-(\lambda_S+\lambda_B)t} & t \in (\mathit{ToF}, \mathit{ToF} + T_W] \\ \lambda_B e^{-\lambda_S T_W} e^{-\lambda_B t} & t > \mathit{ToF} + T_W \end{cases}$$

while the average first arrival time is given by

$$\mathbb{E}[T] = \int_0^{+\infty} t f(t) dt = \frac{1 - e^{-\lambda_B ToF}}{\lambda_B} + \frac{e^{-\lambda_B ToF}(1 - e^{-(\lambda_S + \lambda_B)T_w})}{\lambda_S + \lambda_B} + \frac{e^{-\lambda_S T_W} e^{-\lambda_B (ToF + T_W)}}{\lambda_B} \tag{6.3}$$

As shown in Equation (6.3), the average detection time depends non-linearly on two parameters (λ_S and ToF). This confirms that it is not possible to uniquely extract the ToF from the aforementioned acquisition method, since the average of the timestamps acquired in the second acquisition, $\bar{t}_{tot}$, which is governed by the probability of detection of laser photons, also depends on the λ_S parameter, which depends not only on the ToF, but also on the target reflectivity. One could try to compensate for the error, however this requires measuring also the intensity of the received laser light (which affects the error), introducing an extra variable which is hard to estimate, invalidating the procedure.

Conversely, a linear detector, with no dead-time, is able to timestamp every photon which falls within the acquisition window. In this case, the histograms shown in Figure 6.2 become linear in time, as shown in Figure 6.3.

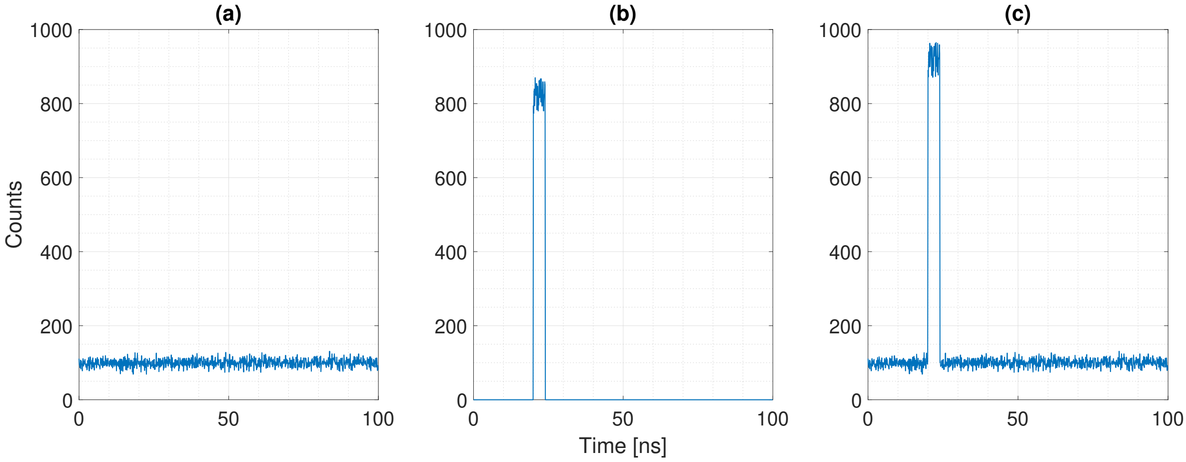


Figure 6.3: Distribution of timestamps obtained with a linear detection process, i.e., with no dead-time limitation due to both SPAD and TDC. The distributions are uniform, since we are now considering the absolute arrival time of detected photons with respect to the beginning of the acquisition window. This approach enables the secure subtraction of the background contribution in (a) from the combined measurement in (c). The result is the isolated contribution from the laser light alone (b), which carries the ToF information. As the superposition property holds in this case, there no longer is an under-weighting condition of background counts after the laser pulse peak in the histogram (c).

One can then extract the ToF with the proposed two-step procedure. In the first step, we measure the total number of *background* events, N_{bg}, and their average absolute arrival time, $\bar{t}_{bg}$. In the second step, with the combination of both background and laser events, we measure the *total* number of events and their average absolute arrival time, denoted N_{tot} and $\bar{t}_{tot}$. Because the superposition property holds, the difference $N_{tot} - N_{bg}$ is equal to the amount of photons from the reflected laser source. We can then extract the *ToF* by properly weighting each average timestamp measurement with the relative photon count contribution:

$$N_{tot} \cdot \bar{t}_{tot} = N_{bg} \cdot \bar{t}_{bg} + (ToF + \bar{t}_l) \cdot (N_{tot} - N_{bg}) \tag{6.4}$$

where $\bar{t}_l$ is the average arrival time of the laser photons referred to the laser emission time in the absence of background light (i.e., $\lambda_B = 0$ and $ToF = 0$). The value $\bar{t}_l$ is a characteristic parameter of the laser source, which can be experimentally estimated by means of an initial calibration. The proposed extraction method does not require the allocation of histogram memory, and needs only two counters to store N_{bg} and N_{tot}, and two accumulators to compute $\bar{t}_{bg}$ and $\bar{t}_{tot}$ reducing the memory requirements by more than three orders of magnitude compared to recent long-range high-resolution d-ToF sensors [6, 8]. We can further reduce the amount of resources down to a single accumulator, needed to store $\bar{t}_{tot}$, and two counters for N_{bg} and N_{tot}, because a constant background throughout the acquisition window leads to an average background time $\bar{t}_{bg}$ of $T_{acq}/2$. In this case, Equation (6.4) turns into the simpler form:

$$N_{tot} \cdot \bar{t}_{tot} = N_{bg} \cdot \frac{T_{acq}}{2} + (ToF + \bar{t}_l) \cdot (N_{tot} - N_{bg}) \tag{6.5}$$

which yields:

$$ToF = \frac{N_{tot}\bar{t}_{tot} - N_{bg}\frac{T_{acq}}{2}}{N_{tot} - N_{bg}} - \bar{t}_l \tag{6.6}$$

We have simulated this extraction method with a Monte Carlo simulator [97] by sweeping the parameters λ_S and λ_B in the range $[10^6 - 10^8]$ and $[10^5 - 10^9]$, respectively. For each pair of λ_S and λ_B values, 10^4 measurements have been acquired with the *ToF* value set to 25 ns. The resulting *ToF*, obtained from Equation (6.4), is shown in Figure 6.4 with the correct estimation over a wide range of λ_B, λ_S pairs, failing only when the λ_S/λ_B ratio is too low even for a classic histogram-based approach.

In the next section, we provide a rigorous mathematical analysis which proves the results briefly introduced with Equation (6.6) from the underlying statistical distribution of photons.

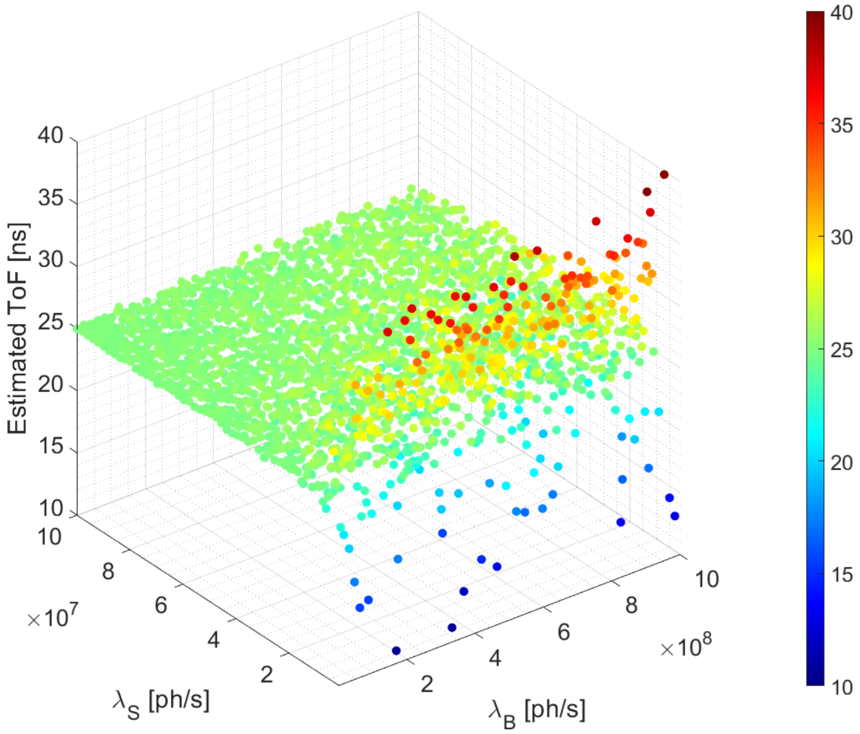


Figure 6.4: Preliminary Monte Carlo simulation results showing the *ToF* computed with the proposed acquisition method with the hypothesis of an ideal linear detector over a full scale range (acquisition window) of 100 ns. The *ToF* can be properly estimated over a wide range of λ_B, λ_S pairs. With the minimum λ_S value considered for the simulation (10^6 events/s), the *ToF* estimation becomes very noisy, but still centered around the correct value of 25 ns. If the signal intensity were zero, the *ToF* estimation would be centered around zero, as the contributions from the two measurements (with and without background) cancel out reciprocally.

6.2 Mathematical analysis

This section proves the validity of the method described above analytically. In the following, we shall denote the duration of the acquisition window by T_{acq}, assume the laser echo to entirely occur within it, i.e., $T_{acq} \gg ToF + T_W$, and denote the time-dependent intensity of the laser pulse by the function $\lambda_S : [0, T_W] \to \mathbb{R}$. The flux of photons can be modeled by a counting process $(N_t)_{t\in[0,T_{acq}]}$ obtained as the sum of two independent Poisson processes: $(N_t^B)_{t\in[0,T_{acq}]}$, with intensity λ_B, describing the background flux of photons, and $(N_t^S)_{t\in[ToF, ToF+T_W]}$, describing the signal. In particular, the process $(N_t^S)_{t\in[ToF, ToF+T_W]}$ is modeled through an inhomogeneous Poisson process with time dependent intensity $(\tilde{\lambda}_S(t))_{t\in[ToF, ToF+T_W]}$ given by $\tilde{\lambda}_S(t) := \lambda_S(t - ToF)$. Hence, the overall flux of photons reaching the SPAD is given by an inhomogeneous

Poisson process $(N_t))_{t\in[0,T_{acq}]}$ with varying intensity λ given by

$$\lambda(t) = \begin{cases} \lambda_B & t \in [0,\, \mathit{ToF}) \cup [\mathit{ToF} + T_W, T_{acq}] \\ \lambda_B + \tilde{\lambda_S}(t) & t \in [\mathit{ToF},\, \mathit{ToF} + T_W) \end{cases}$$

Hence, we can prove that, given n photons detected in the interval $I = [0, T_{acq}]$, the n detection times are independent and distributed on $[0, T_{acq}]$ with a distribution density function $f : [0, T_{acq}] \to \mathbb{R}$ given by

$$f(t) = \begin{cases} \dfrac{\lambda_B}{\lambda_B T_{acq} + \displaystyle\int_0^{T_W} \lambda_S(u)du} & t \in [0,\, \mathit{ToF}) \cup [\mathit{ToF} + T_W, T_{acq}] \\ \\ \dfrac{\lambda_B + \lambda_S(t - \mathit{ToF})}{\lambda_B T_{acq} + \displaystyle\int_0^{T_W} \lambda_S(u)du} & t \in [\mathit{ToF},\, \mathit{ToF} + T_W) \end{cases} \tag{6.7}$$

Indeed, by considering a partition $\{I_j\}_{j=1,\dots,m}$ of the interval $I = [0, T_{acq}]$, the independence of the increments of the inhomogeneous Poisson processes $(N_t)_{t\in[0,T_{acq}]}$ yields

$$\mathbb{P}(N(I_1) = n_1, \dots, N(I_m) = n_m \mid N(I) = n) = \frac{\displaystyle\prod_{j=1}^{m} e^{-\lambda_{I_j}} \frac{(\lambda_{I_j})^{n_j}}{n_j!}}{e^{-\lambda_{[0,T_{acq}]}} \dfrac{\left(\lambda_{[0,T_{acq}]}\right)^n}{n!}} \tag{6.8}$$

where, for an interval $I = [a, b] \subset \mathbb{R}$ we adopted the notation $\lambda_I := \int_a^b \lambda(t)dt$ and we assumed $\sum_{j=1}^m n_j = n$. By a straightforward computation, the right hand side of (6.8) can be written as

$$\mathbb{P}(N(I_1) = n_1, \dots, N(I_m) = n_m \mid N(I) = n) = \frac{n!}{n_1! \cdots n_m!} \prod_{j=1}^{m} \left(\frac{\lambda_{I_j}}{\lambda_{[0,T_{acq}]}} \right)^{n_j}$$

the latter being equivalently obtained in terms of n independent and identically distributed continuous random variables $T_1, \dots, T_n$ with density f given by (6.7). In other words, the arrival times $\{T_i\}_{i=1,\dots,n}$ of the n photons provide a statistical sample for the distribution (6.7).

The distribution (6.7) has a mean μ given by

$$\mu = \int_0^{T_{acq}} tf(t)dt = \frac{1}{\lambda_B T_{acq} + \int_0^{T_W} \lambda_S(t)dt} \left(\frac{\lambda_B T_{acq}^2}{2} + ToF \int_0^{T_W} \lambda_S(t)dt + \int_0^{T_W} t\lambda_S(t)dt \right) \tag{6.9}$$

Clearly, since μ is a linear function of ToF, it can easily be inverted. By denoting with α the ratio

$$\alpha := \frac{\lambda_B T_{acq}}{\lambda_B T_{acq} + \displaystyle\int_0^{T_W} \lambda_S(t)dt} \tag{6.10}$$

we get an equation providing the time-of-flight as a function of the other characteristic parameters of the process:

$$ToF = \frac{1}{1-\alpha} \left(\mu - \alpha \frac{T_{acq}}{2} \right) - \bar{t}_l \tag{6.11}$$

where $\bar{t}_l = \frac{\int_0^{T_W} t\lambda_S(t)dt}{\int_0^{T_W} \lambda_S(t)dt}$ is the average arrival time of the laser photons referred to the laser emission time in the absence of background light (i.e., $\lambda_B = 0$ and $ToF = 0$).

If n photons are detected within the interval $[0, T_{acq}]$, the sample mean $\bar{t}_{tot}$ of their detection times provides an unbiased estimator for μ. The main issue with this approach is the estimation of the parameter α, which depends on λ_B and λ_S, the latter being affected by a high level of uncertainty since it is related to the intensity of the laser echo. Because the total number of photons detected in the time interval $[0, T_{acq}]$ is a Poisson random variable with average $\lambda_B T_{acq} + \int_0^{T_W} \lambda_S(t)dt$ and, analogously, the number of background photons detected in the time interval $[0, T_{acq}]$ is a Poisson random variable with average $\lambda_B T_{acq}$, we can estimate both parameters by observing a realization of both processes. More precisely, let us first switch off the laser source and collect the number N_{bg} of photons arriving during the interval $[0, T_{acq}]$, then let us switch on the laser source and collect the total number N_{tot} of photons arriving during the interval $[0, T_{acq}]$. The observed values N_{bg} and N_{tot} are respectively an estimate for $\lambda_B T_{acq}$ and $\lambda_B T_{acq} + \int_0^{T_W} \lambda_S(t)dt$, while their ratio is an estimate $\hat{\alpha} := \frac{N_{bg}}{N_{tot}}$ for the parameter α defined in (6.10). By replacing the parameters μ and α with their estimates $\bar{t}_{tot}$ and $\hat{\alpha}$

in formula (6.11), we obtain the following estimator for ToF:

$$\widehat{ToF} = \frac{1}{1-\hat{\alpha}}\left(\bar{t}_{tot} - \hat{\alpha}\frac{T_{acq}}{2}\right) - \bar{t}_l \tag{6.12}$$

$$= \frac{N_{tot}\bar{t}_{tot} - N_{bg}\dfrac{T_{acq}}{2}}{N_{tot} - N_{bg}} - \bar{t}_l \tag{6.13}$$

which coincides with (6.6).

6.3 Acquisition schemes

The simulation results obtained in Section 6.1.2 are based on the assumption that the photon detection process is ideal, i.e., with no dead-time and with a linear response over the incoming flux of photons. In a real-world scenario, however, detectors are limited by the dead-time between subsequent detections, resulting in a non-linear response. To implement the proposed extraction method, we propose a novel SPAD acquisition scheme which emulates the behavior of a linear detector. More in detail, we propose two ways to obtain a linearized SPAD response from a real SPAD. Both methods are based on the assumption that the underlying statistical processes are stationary and ergodic. In particular, we assume that there are no major fluctuations of the characteristic parameters of the process during the acquisition time. Similarly to an equivalent-time sampling oscilloscope, both methods rely on repeating the observation multiple times to emulate the response of a SPAD detector with no dead time. In Section 6.3.1 and 6.3.2, we describe the working principle of each method and propose a possible implementation. Then, in Section 6.3.3, we provide the mathematical proof that both acquisition methods are capable of correctly sampling the distribution of photon arrival times.

6.3.1 Acquisition scheme #1: Acquire or discard

The first acquisition scheme relies on a simple (albeit inefficient) mechanism which requires no additional resources in terms of the SPAD driving circuit. The acquisition works over multiple runs, each requiring multiple observations. The first timestamp of every run is considered valid, memorized, and used to increment either N_{bg} or N_{tot}, depending on the current phase of the acquisition, and update $\bar{t}_{tot}$. Then, in the next observations, timestamps are considered valid, and used to update the algorithm pa-

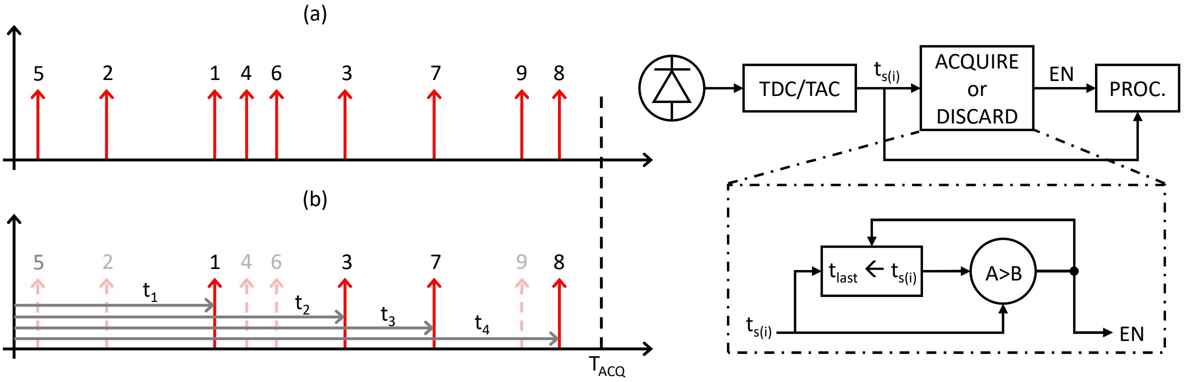


Figure 6.5: Example of SPAD response linearization with the *acquire or discard* acquisition method and a synchronous SPAD operation. In (a), each photon arrival time is represented by a red arrow and the order of arrival is indicated. The first run starts with the acquisition of photon #1, resulting in timestamp t_1. Photon #2 is discarded, since its arrival time is earlier than photon #1. The next recorded information comes instead from photon #3, which is later than photon #1, and sets the new minimum time. The run proceeds with the same criteria resulting in the stream of photon arrival times t_1, t_2, t_3 and t_4 from photons #1, #3, #7 and #8, which is a single realization of the emulated response of the linearized SPAD detector. On the right, a principle schematic is proposed, showing the lightweight usage of resources, with only one comparator and one register required on top of the processing circuit. The *acquire or discard* acquisition method is simple but inefficient, as most of the photons arrival times are discarded, resulting in longer acquisition times.

rameters, only if they are higher than the largest previous timestamp, otherwise they are discarded. The procedure is iterated until the last photon within the acquisition window T_{acq} is timestamped. The necessary condition to conclude the current linearization cycle is the absence of detected photons before T_{acq}. This implies that the next photon will occur after the acquisition window, thereby validating that the last recorded photon marks the conclusion of the current linearization cycle. The process is then repeated multiple times to increase statistics. Figure 6.5 shows an example of a run, including all the discarded events, and a possible implementation. While the implementation is straightforward, the method is inefficient because the majority of the detected photons may end up being discarded.

6.3.2 Acquisition scheme #2: Time-gated

The time-gated acquisition scheme works by delaying the activation of the SPAD to start from the previously acquired timestamp, until no photon is detected before the end of the acquisition window. With this approach, there is no need to discard timestamps, allowing for a faster acquisition. This, however, comes at the expense of a more complex

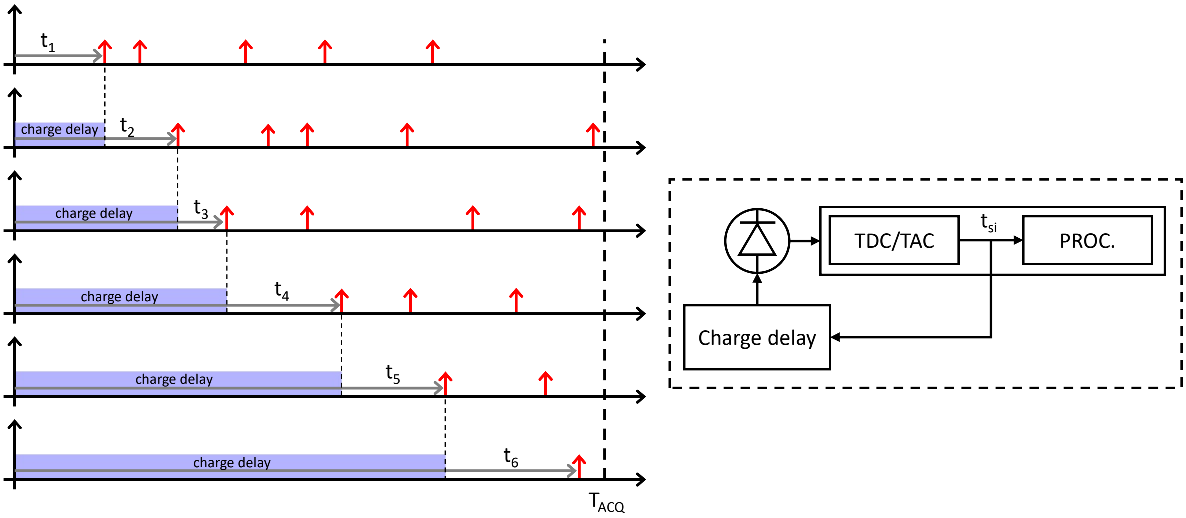


Figure 6.6: Example of SPAD response linearization with the *time-gated* acquisition method. With this approach, no photon timestamp is discarded thanks to the delayed activation of the SPAD for each timing measurement. During the *charge delay* phase, the SPAD front-end is forced OFF, thus photons can not be detected. For each measurement, the first timestamp is detected and used to increment either N_{bg} or N_{tot} and update the average time $\bar{t}_{tot}$. At the same time, the *charge delay* phase value is updated accordingly for the next measurement. As opposed to the *acquire or discard* method, more hardware resources are needed to build the delay element which controls the activation of the SPAD.

hardware implementation, which needs a time-activated gating scheme, for instance using a programmable delay line. An example of acquisitions is shown in Figure 6.6, together with a possible implementation.

6.3.3 Discussion on implementation, expected performance and mathematical analysis

While providing the same result, it is clear that the implementation cost and the performance of the two acquisition schemes are different. With the *acquire or discard* scheme, almost no hardware modification is required to an already existing SPAD sensor. However, because of the decimation process, the efficiency of the acquisition could be very low. This also depends on the intensity of the incoming rate of events: the higher the rate, the higher the probability to have smaller timestamps which block the detection process. On the other hand, the *time-gated* scheme requires a delay line and the SPAD-gating, but the efficiency is much higher since no decimation process occurs. To show the difference in terms of efficiency of the two proposed acquisition schemes, we run a Monte Carlo simulation with background light flux in the range $[10^6, 10^8]$ events/s (corresponding to approx. $[0.64, 64]$ klux) with the aim to linearize the SPAD response over an acquisition window T_{acq} of 100 ns. As shown in Figure 6.7-(a), at the highest rate of 10^8 events/sec, the amount of timestamps to be acquired to cover the acquisition window T_{acq} for the *acquire or discard* scheme is more than 3 orders of magnitude higher compared to the more efficient *time gated* scheme. From the simulation, we can also identify the maximum number of measurements which can be executed by the two acquisition schemes to sustain an operation frame rate of 30 Frames Per Second (FPS). The *time-gated* scheme can average the linearized SPAD response up to $N = 3 \cdot 10^4$ times over the whole range of background light event rate. On the other hand, with the same number N of acquisitions, the *acquire or discard* scheme can only support up to $\approx 2.4 \cdot 10^7$ events/s of background rate. Figure 6.7-(b) shows a comparison in relative terms against an ideal SPAD coupled to an ideal TDC, thus being capable of timestamping all photons impinging on the sensor surface over the acquisition window at once. Both the *acquire or discard* and *time gated* schemes are meant to work with a real SPAD, which in this case can timestamp at most one photon per acquisition window. For such a reason, if the acquisition schemes are compared against an ideal system, a loss of photons occurs, which can be limited by the *time gated* scheme at ≈90% in the worst case of photon flux.

Concerning the implementation complexity, the *time-gated* scheme requires a finely controlled delay element, which may be expensive on a per-pixel basis. Proper balance between the linearization efficiency and implementation complexity can be obtained with a mixture between the two acquisition schemes. This can be accomplished by employing a coarse but simple delay architecture for the gating mechanism (e.g., a clock-based counter), which activates the SPAD one clock cycle before the last recorded timestamp. Then, the linearization process is concluded by relying on the *acquire or discard* scheme to cover the remaining amount of time. The efficiency is still guaranteed, as the *acquire or discard* scheme starts at most one clock cycle before the last recorded timestamp, and not from the beginning as in the standalone implementation.

From a mathematical point of view, both acquisition methods allow sampling the correct distribution of the photon arrival times $(T_n)_{n\geq 1}$. Let T_n denote the occurrence time of the $n-$th event, i.e., the arrival of the n-th photon. This can be defined as the infimum of the set of times t such that the number of arrivals N_t in the interval $[0, t]$ is greater than or equal to n:

$$T_n := \inf\{t \geq 0 \,:\, N_t \geq n\}, \qquad n \geq 1$$

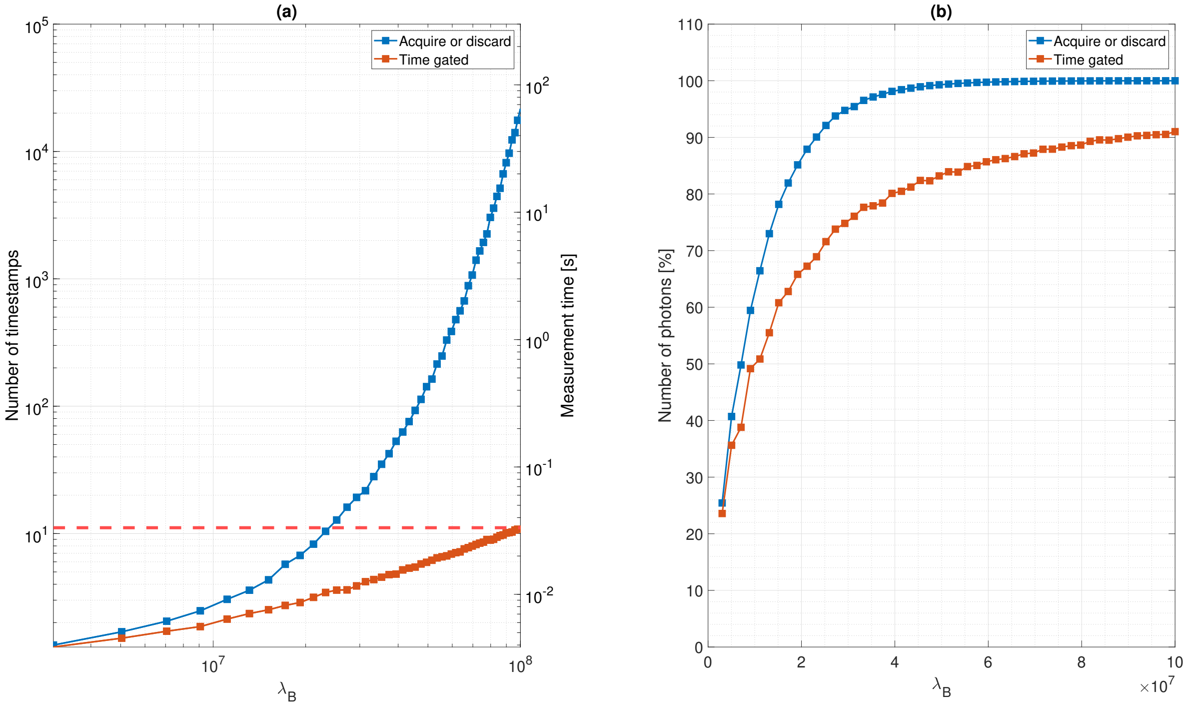


Figure 6.7: Result of Monte Carlo simulation comparing the two proposed acquisition schemes in terms of efficiency for increasing values of background light flux (λ_B) in the range $[10^6, 10^8]$ events/s. In (a), we compare the acquisition schemes in absolute terms. On the left axis, we show the number of timestamps required to linearize the SPAD response over an acquisition window T_{acq} of 100 ns. On the right axis, we show the total time required for the two methods to collect $N = 3 \cdot 10^4$ measurements to average the linearized response of the SPAD N times. The horizontal line indicates a limit of ≈ 33.3 ms, for an equivalent operation frame-rate of 30 FPS. With the *time gated* scheme, the required frame rate can be guaranteed over the entire range of background light flux, while considering the *acquire or discard* scheme, the maximum sustainable flux is limited to $\approx 2.4 \cdot 10^7$ events/s. In (b), we compare the acquisition schemes in relative terms against an ideal system composed by an ideal SPAD with an ideal TDC, showing the percentage of photons lost during the linearization process over the acquisition window T_{acq}. The loss of photons occurs as both acquisition schemes works with a real SPAD, which in this case can timestamp at most one photon per acquisition.

By definition, $N_{T_n} = n$. We can therefore derive an equivalent representation of the random variables T_n. Indeed, for $n \geq 2$, the time T_n of occurrence of the $n-$th event can be obtained as the infimum of the set of times t greater than T_{n-1} (the time of occurrence of the $(n-1)$-th event) such that the increment $N_t - N_{T_{n-1}}$, the number of events occurring in the interval $(T_{n-1}, t]$, is greater than 1. Therefore

$$
\begin{aligned}
T_1 &:= \inf\{t \geq 0 \,:\, N_t \geq 1\} \\
T_n &:= \inf\{t \geq T_{n-1} \,:\, N_t - N_{T_{n-1}} \geq 1\}, \quad n \geq 2
\end{aligned}
$$

which corresponds to the results of the acquisition schemes described previously.

6.3.4 Comparison with state-of-the-art

In this section, we compare our histogram-less acquisition method with state-of-the-art SPAD-based LiDAR sensors in terms of memory requirement, scalability and tolerance to high background light flux. For all comparisons in this section, we consider for our method 16 bits of counters depth (i.e., up to 65535 counts for N_{bg} and N_{tot}, respectively) and then 3 times the number of TDC bits (1xTDC bits required for the TDC word itself, and then 2xTDC bits to properly size the accumulator memory). First, we compare against standard sensors, i.e., sensors that require the raw timestamps to be read out to build the necessary histogram of timestamps off-chip. To provide a fair comparison, we do not consider the sensor resolution, which changes from chip to chip, but only the amount of memory required to build the histogram for one pixel. In the works we consider for our comparison [4,6,8,43,63], we extrapolate the total amount of per-pixel memory based on the number of reported TDC bits and on an 8-bit histogram depth for all of them. We then compare our solution to sensors that offer full on-chip histogram capability [54,105,106]. Also in this case, we consider the amount of memory reported in each work necessary to build the histogram of timestamps for one pixel. Results are reported in Figure 6.9, where an average memory reduction factor of ≈ 2129 and ≈ 136 for standard and full on-chip histogram sensors, respectively, is obtained. More comparison details, including minimum and maximum memory reduction factors are reported in Table 6.2.

Similarly to our method, partial histogram approaches [55,66,67,72] are also quite effective in reducing the memory requirements. Nevertheless, our approach not only outperforms them with a memory reduction that ranges from 67% [55] to 3% [72], but also performs better in many other important aspects. In fact, unlike previous work, our approach does not have any of the following needs:

- The need to find the laser peak in time using a zooming or a sliding search procedure, which is at the basis of every partial histogram approach [107].

- The need to share hardware resources (TDCs, memory) among pixels in the same column [66] or in the same cluster [72], to reduce the area usage.
- The need for area consuming processors to manage the algorithm underlying the partial histogram technique and that can only be implemented using advanced 3D integrated technologies with single-pixel access [55].

All these translate into higher measurement time and laser power penalty as more acquisitions are needed than a standard full-histogram approach [107], and higher costs.

Our method, given the very limited amount of required memory resources, is also advantageous in terms of scalability to higher sensor resolutions, and also in terms of range extension. As an example, a standard histogram-based sensor with 15-bit TDC requires memory to store up to 32767 histogram bins per pixel. If the measurement range is doubled, the additional TDC bit results in an increase of 100% on the memory requirement. Conversely, with our approach, the amount of memory increase to double the range is limited to only $\approx 3.9\%$.

While the advantage in terms of reduction of memory/hardware resources is evident, careful considerations are needed regarding performance in terms of accuracy and precision. Unlike a standard full-histogram approach, the proposed method does not have direct access to the full statistics of photon arrival times. This results, when considering the same number of timestamps (i.e., laser pulses), in a degradation of accuracy and precision. The approach, however, is advantageous in any case, thanks to the benefit provided by the linearization process, especially in conditions of strong background light. This can be observed in Figure 6.8, where we present the results of a simulation set on a real setup [43]. While accuracy error can be compensated in post processing (especially with the proposed linearization method, which is almost immune against signal pile-up distortion), precision worsen if considering the same amount of timestamps or laser pulses to extract the ToFinformation, as the full statistic of photon arrival times is available only with a histogram-based approach. On the other hand, the linearization process allows us to detect photons from the reflected laser pulse at higher distances than a standard approach, resulting in correct distance estimation over the whole range. The reason for this behavior is that pile-up distortion due to strong background light is completely compensated, making it possible to detect signal photons even at long distances. In the context of the proposed simulation, this result in $\approx$+53% range increase.

Concerning the tolerance to high background light flux, both detection processes can sustain very high flux regimes, with a limit determined by the finite resolution T_{TS}

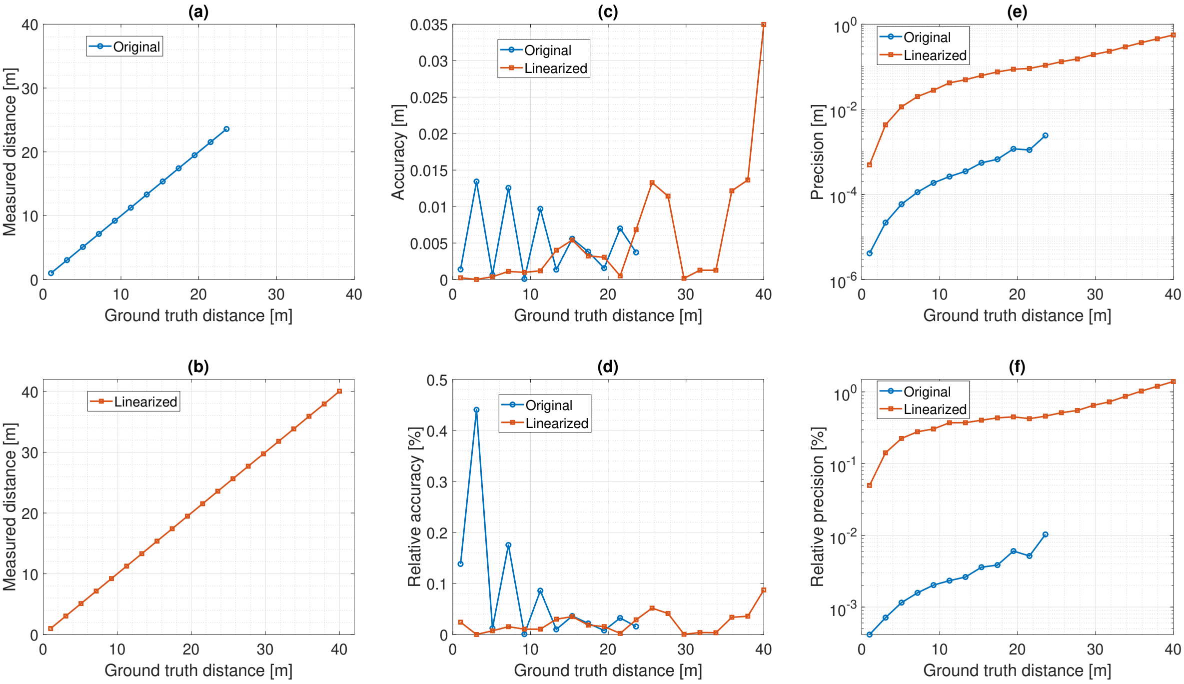


Figure 6.8: Comparison of accuracy and precision performance between a standard full-histogram and the proposed histogram-less approaches. In (a) and (b) we show the maximum range achieved by the full-histogram and histogram-less approaches, respectively. In (c) and (d) we show the accuracy error in absolute and relative terms, respectively, while in (e) and (f) the absolute and relative precision errors are shown in logarithmic scale. By considering the same number of timestamps (i.e., laser pulses) a standard full-histogram approach outperforms the proposed histogram-less method in terms of precision, yielding 0.01 % vs 1.4 % at the longest achievable distance of ≈ 23.58 m and ≈ 40.03 m, respectively. On the other hand, the advantage offered by the time linearization of the SPAD response reflect in an increase of the measurement range of up to ≈ 53 %.

of the timestamping circuit. This limit translates to the requirement $\frac{1}{\lambda_B+\lambda_S} \gg T_{TS}$, i.e., T_{TS} must be small enough that the probability of more than one photon falling into the same time bin is negligible. By considering a threshold on this probability, we can extract the maximum flux of photons λ_{max} which can be sustained by our detection process.

The probability to have more than one photon per time bin of duration T_{TS} can be computed as $P(n > 1) = 1 - P(n = 0) - P(n = 1)$. Considering that events belongs to a Poisson distribution of parameter $\lambda_{max}t$, the probability is expressed as $P(n > 1) = 1 - e^{-\lambda_{max} \cdot T_{TS}} \cdot (1 + \lambda_{max} \cdot T_{TS})$. By setting a threshold of less than 1%, and considering $T_{TS} = 100$ ps, the maximum photon flux that can be sustained is equal to $\lambda_{max} \simeq 1.48 \cdot 10^9$ ph/s. Compared to the maximum flux required by a standard system which must comply with the 5% rule, and with the hypothesis of an acquisition window

Table 6.2: Minimum, average and maximum memory reduction factor of the proposed histogram-less acquisition method against standard d-ToF sensors (off-chip histogram) [4, 6, 8, 43, 63] and sensors with on-chip full histogram capability [54, 105, 106].

Standard sensors			**Full on-chip histogram**		
Min.	Avg.	Max.	Min.	Avg.	Max.
69.4 [63]	2129.5	6553.6 [4]	4.7 [105]	135.9	331.3 [106]

T_{acq} of 100 ns, our detection process can sustain a photon flux $\simeq$ 3000 times higher.

As the *ToF* is extracted without direct access to the complete statistics of photon timestamps (i.e., the histogram), the contributions to the *ToF* due to multiple targets or multipath reflections cannot be distinguished as they linearly contribute to the average timestamp value extracted after the linearization process. A foreseen countermeasure is the implementation of an additional system-level gating scheme (thus, on top of the one required for the SPAD linearization), where by means of an approximate knowledge of the target(s) location, the timing localization of the LiDAR acquisition is properly steered to avoid the mixture of multiple contributions. On the other hand, the two acquisition schemes proposed in this work do not prevent the construction of a histogram of (linearized) timestamps, supporting the complete distinction of multiple targets and multipath reflections.

6.4 Measurement results

The proposed acquisition scheme has been validated with measurements using real data from an existing single-point SPAD-based d-ToF sensor, with an architecture similar to the one from Perenzoni et al. [17], which in addition offers on-chip histogramming capability. The sensor is fabricated in a standard 150 nm CMOS process with the SPAD technology developed in the work by Xu et al. [93]. The histogram features 1024 bins with 10-b depth, and a TDC resolution of 100 ps. The SPADs are enabled synchronously with the beginning of the acquisition window and the first measured timestamp for each acquisition increments the corresponding histogram bin. After a user-selectable number of acquisitions, the histogram is read out and unpacked. Then, the unpacked data is shuffled to recover a realization vector of the arrival times of the detected photons.

Background events were generated by means of a ≈180 W fiber-coupled halogen illuminator pointed directly toward the sensor, while a black matte panel with low reflectivity (≈10%) was selected as target, with a distance range from 1 m up to 3.8 m.

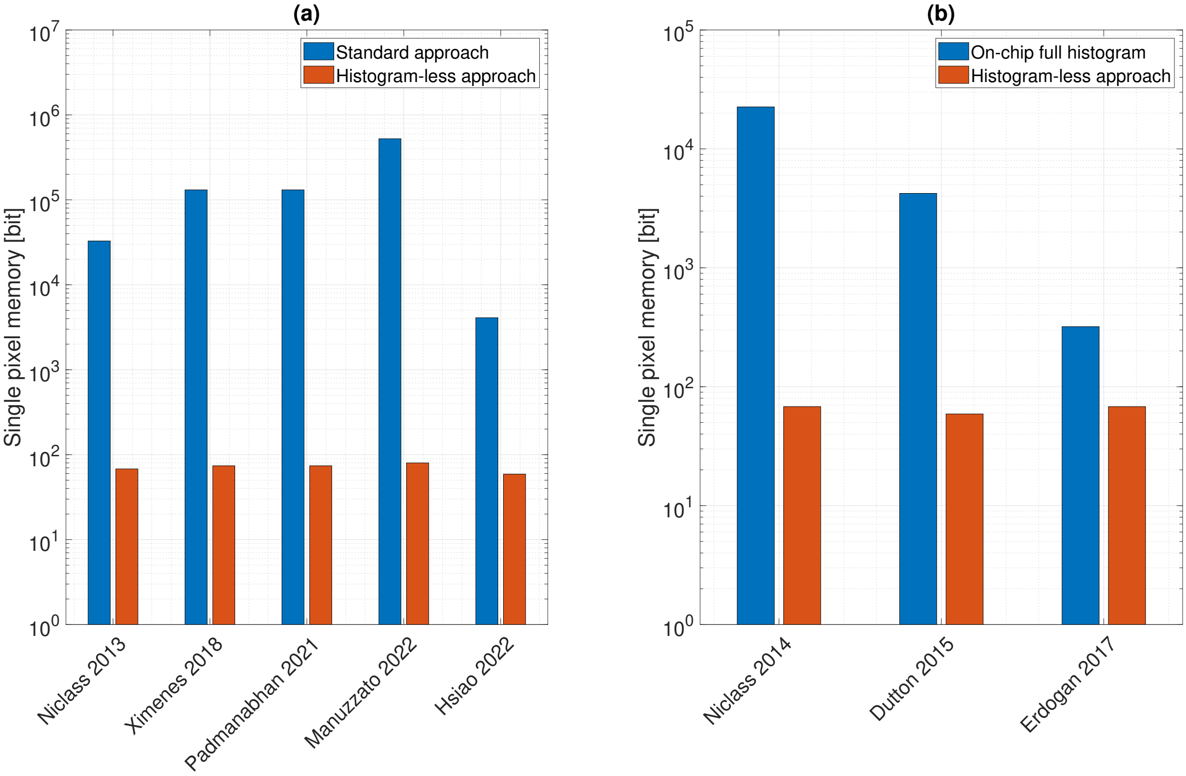


Figure 6.9: Comparison of the amount of per-pixel memory required by our histogram-less acquisition method against histogram-based d-ToF sensors. In (a), we consider standard sensors where every timestamp is read out and the histogram is built off-chip [4, 6, 8, 43, 63]. In (b), we consider sensors with full on-chip histogram capability [54, 105, 106].

A picture of the setup is shown in Figure 6.10, with indications on the main components.

First, we focus on the validation of the linearization behavior of the proposed acquisition scheme by considering only background light. Then, we consider the combination of background and laser together, as in a real scenario, and we compute the ToF with the proposed histogram-less acquisition scheme.

6.4.1 Preliminary considerations

As we base our measurements on the re-engineering of an existing d-ToF sensor, preliminary considerations are needed before providing further details on measurement results. The sensor measures the arrival time of the first detected photon for each laser pulse, as described in Section 6.1, which is stored in an on-chip histogram memory. Since the sensor measures the arrival time of the first photon, the statistical distribution is exponential, thus we are considering *relative* arrival times. A statistically valid realization of the incoming timestamps is obtained by unpacking and randomly shuffling the content

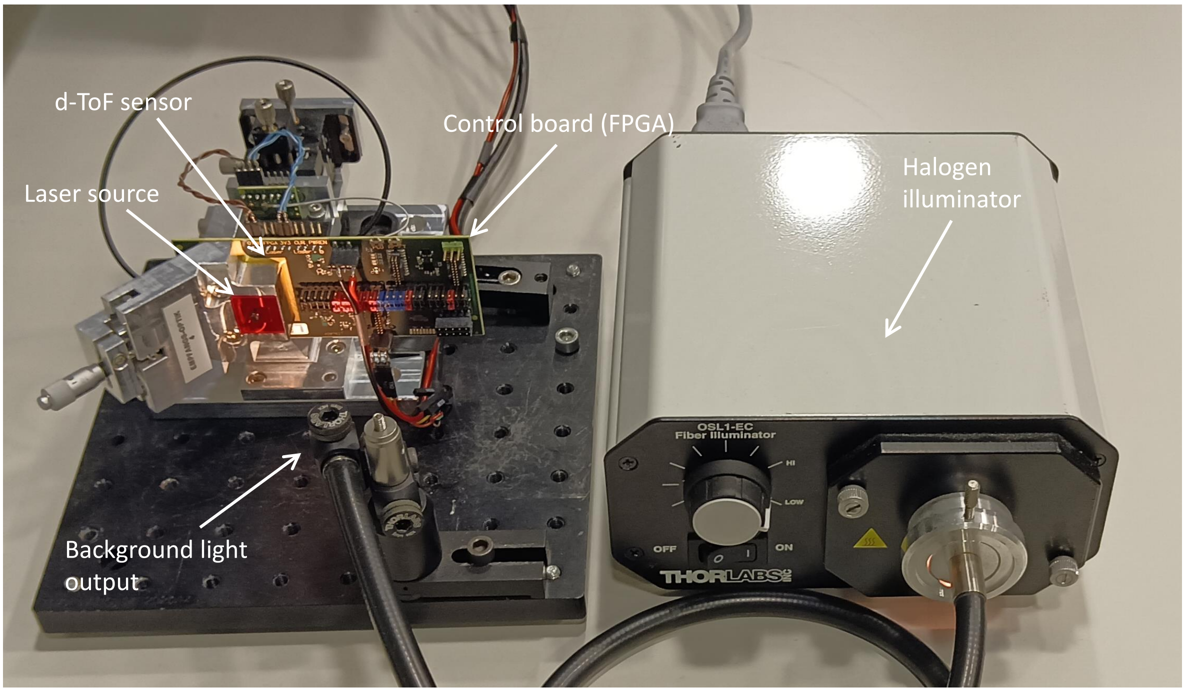


Figure 6.10: Measurement setup with the FPGA control board, d-ToF system and halogen illuminator for the generation of background light pointed directly toward the sensor.

of the histogram memory. The obtained realization is a vector of *relative* arrival times, which is the starting point of our measurement analysis.

In Section 6.3, we described two possible acquisition schemes. The *acquire or discard* scheme, even though is intrinsically inefficient, can be straightforwardly used with our dataset as it requires no hardware modification over the already existing SPAD-based d-ToF system. The *time-gated* scheme, while more efficient, requires a time-gating circuit which is not implemented in our sensor.

The first set of measurements focuses on background events only. In this case, there is a single source of events with intensity λ_B, so we can apply the *time-gated* scheme by computing the cumulative sums of timestamps to obtain *absolute* arrival times from *relative* ones. On the other hand, when events from background and laser are combined, as in a real measurement scenario, it is not possible to mimic the behavior of the *time-gated* scheme by means of the cumulative sum operation. In that case we rely on the *acquire or discard* scheme.

6.4.2 Measurements with background light only

We set the intensity of background light from a minimum of $\approx 6.5 \cdot 10^6$ up to $\approx 133 \cdot 10^6$ events/s. This is the rate of events at the output of the SPAD, which therefore takes into account all physical parameters of concern of a typical d-ToF system [97]. Considering an acquisition window of 100 ns, specific from the sensor [17], the equivalent average number of detections within T_{acq} equals ≈ 0.65 and ≈ 13.3 for the minimum and maximum background light intensity, respectively. In both cases, this is much higher than the conventional limit of 5% events [104] (13 and 266 times higher, respectively), showing the high resistance of our method against pile-up distortion. By considering the equation which links the intensity of background events, λ_B, with the physical parameters of the system [97], it is possible to derive the equivalent background illumination level, in kilolux, up to a maximum of $\approx$ 85 kilolux. Measurement results are shown in Figure 6.11, showing a relative deviation from the reference background intensity extracted from the exponential fit of the original histogram of less than ±0.5 % over the whole range of values.

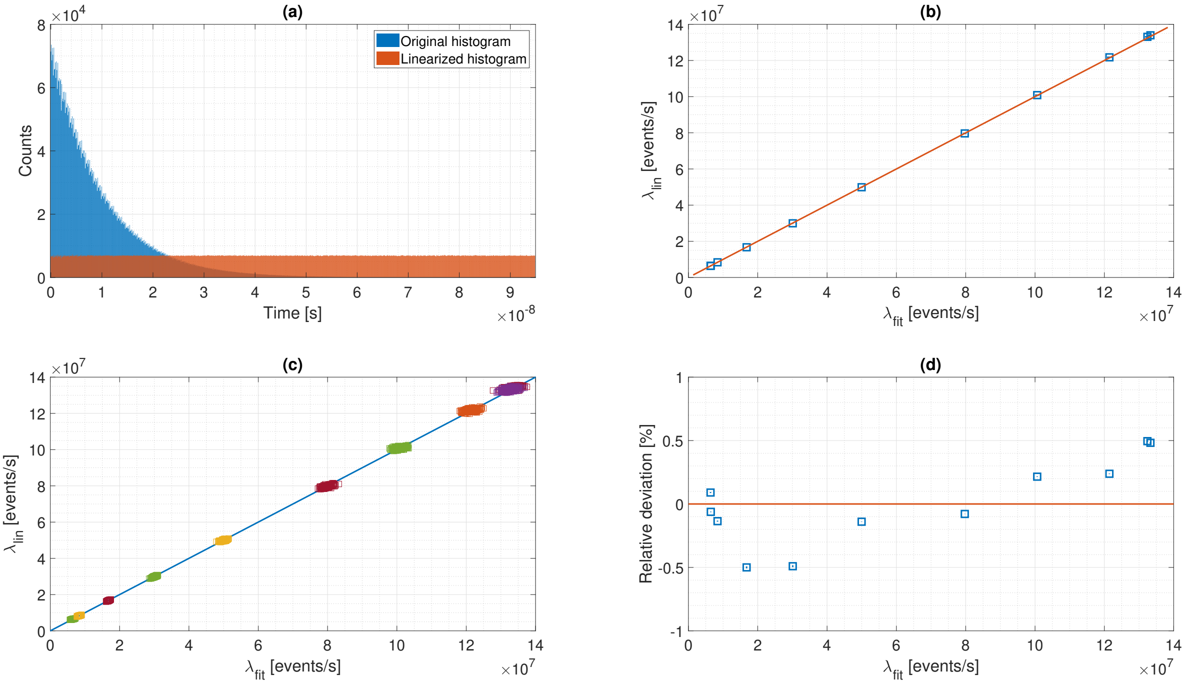


Figure 6.11: Linearization of the SPAD response with background events only. For each value of background flux, $8 \cdot 10^6$ timestamps are acquired from the sensor. In (a), an example of linearized histogram is shown together with the original one (exponentially distributed) for a background flux of $\approx 100 \cdot 10^6$ events/s. In (b), we show the flux of background events estimated from the linearized histogram of timestamps, λ_{lin} against the flux estimated from an exponential fit on the original histogram of timestamps, λ_{fit}. In (c), for each value of background flux, the entire dataset was split in 200 subsets to analyze the homogeneity of the linearization process, while in (d), the relative deviation from the background flux measured from the original histograms is shown, used as a reference, demonstrating a relative deviation below ±0.5 % over all data subsets.

6.4.3 Measurements with background and laser light and extraction of the *ToF*

Our first goal is to show that the underestimation of background counts which occurs in a standard d-ToF system can be completely recovered with our acquisition scheme. This is demonstrated in the first measurement, displayed in Figure 6.12, which compares a traditional acquisition with the *acquire or discard* scheme, qualitatively showing the linearization process by means of the linearized histogram of timestamps.

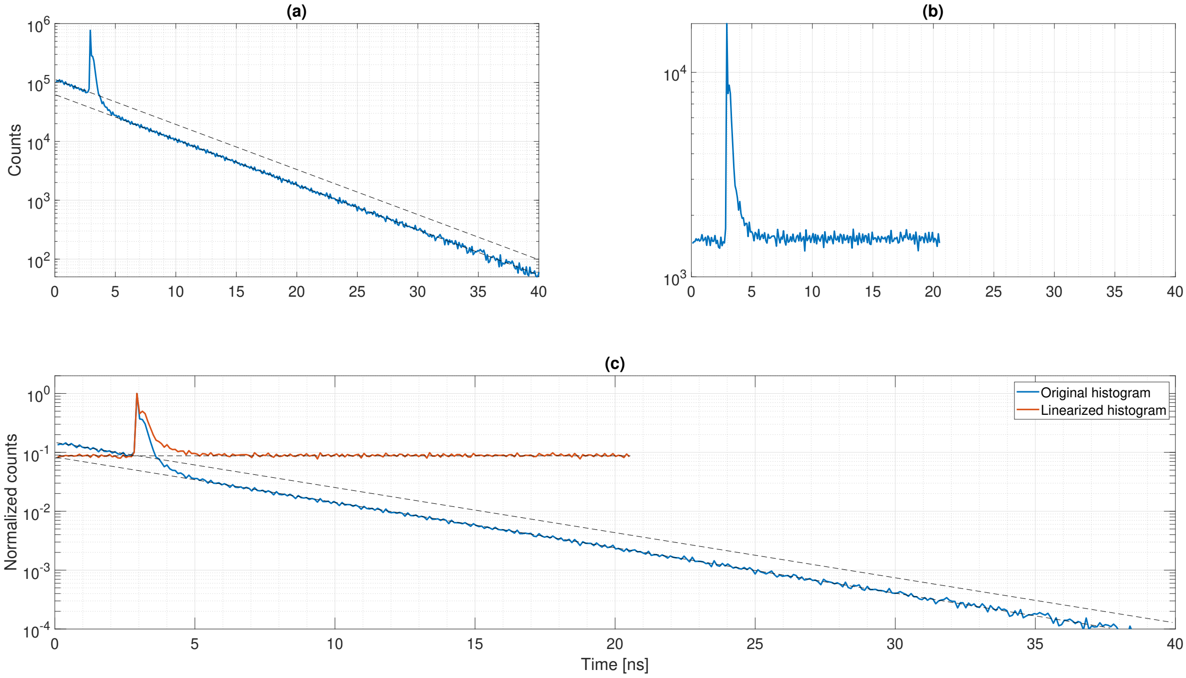


Figure 6.12: Qualitative measurement showing the linearization process of the proposed acquisition scheme. In (a), the original histogram of timestamps is shown in logarithmic scale, where the drop of counts which occurs after the laser peak is clearly visible. In (b), the histogram obtained with the *acquire or discard* scheme proves the efficacy of the linearization process, which fully compensates for the non-linearity of the detector. The length of the linearized histogram is shorter than the original dataset, as we decided to stop the linearization earlier to reduce the data loss which naturally occurs with the *acquire or discard* scheme. Due to the intrinsic inefficiency of this scheme, the histogram peak in (b) is attenuated by ≈ 34 dB with respect to the original dataset in (a). In (c), the two histograms are shown together after normalization.

We then quantitatively evaluated the linearization process by estimating the intensity of background light from both portions of the histogram, i.e., before and after the laser peak. For this characterization, we used the ≈10% reflectivity target (black matte panel) at 2.5 m distance from the sensor. The results are depicted in Figure 6.13, showing a relative deviation from the ground truth (estimated from an exponential fit on the original dataset) below ±4 %.

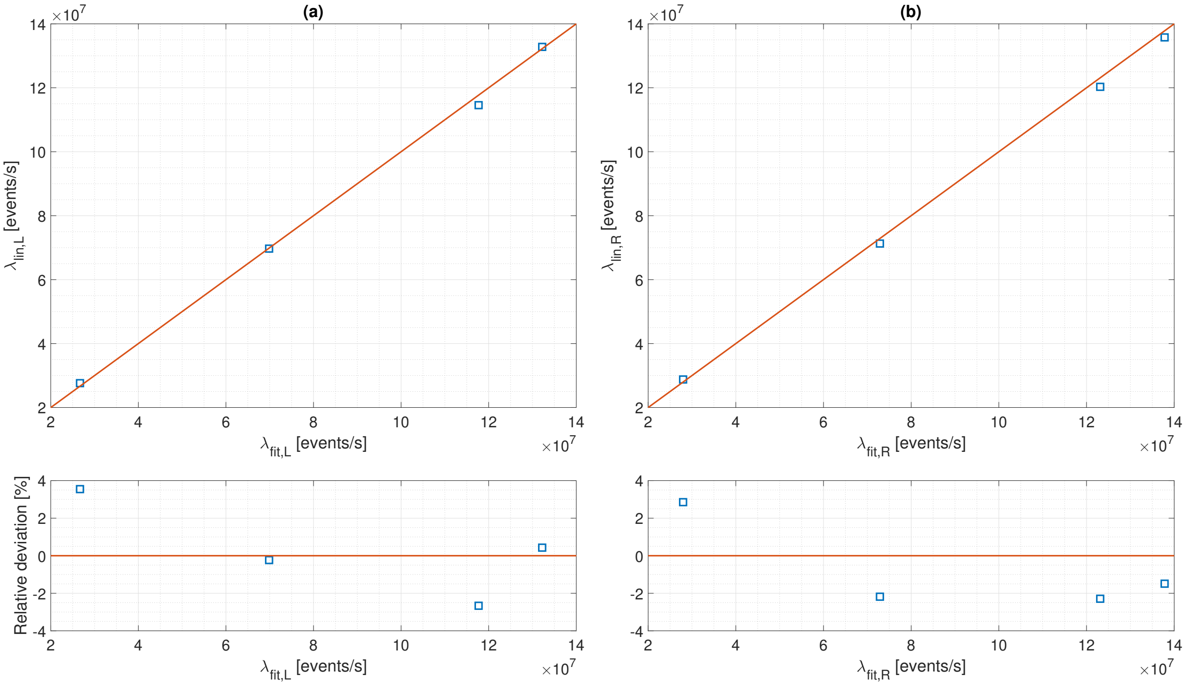


Figure 6.13: Quantitative characterization of the linearization process considering four different values of background flux, from $\approx 27.6 \cdot 10^6$ events/s up to $\approx 133 \cdot 10^6$ events/s (corresponding to approx. [17.6,85] klux) with a target distance of 2.5 m. For each value of background flux, $2.5 \cdot 10^6$ timestamps are acquired from the sensor. In (a), the relationship between the background fluxes computed before the laser peak is shown, where $\lambda_{fit,L}$ comes from an exponential fit on the original histogram, while $\lambda_{lin,L}$ comes from the linearized histogram. In (b), the same relationship is shown but considering the portion of background events after the histogram peak. For each portion, the relative deviation of the flux extracted from the linearized histogram of timestamps is shown, demonstrating an estimation error below ±4 % over the whole range. The application of the *acquire or discard* acquisition scheme results in a data reduction factor of ≈ 7.5 and ≈ 165 for the minimum and maximum background light flux, respectively.

In a different measurement, we verify the resistance of the proposed SPAD linearization method against pile-up distortion. To do so, we acquire several timestamps from the reflected laser pulse with a detection rate of 90%, which is 18 times higher than the conventional limit of 5%. The results, shown in Figure 6.14, proves the efficacy of our linearization method in challenging pile-up conditions where a standard sensor would fail. A measurement acquired with a conventional Time-Correlated Single Photon Counting (TCSPC) setup is shown as reference.

The last set of measurements shows the extracted *ToF* without the need to build a histogram of timestamps. For each measured distance, we run the linearization algorithm 250 times to have sufficient statistics to compute accuracy and precision. For

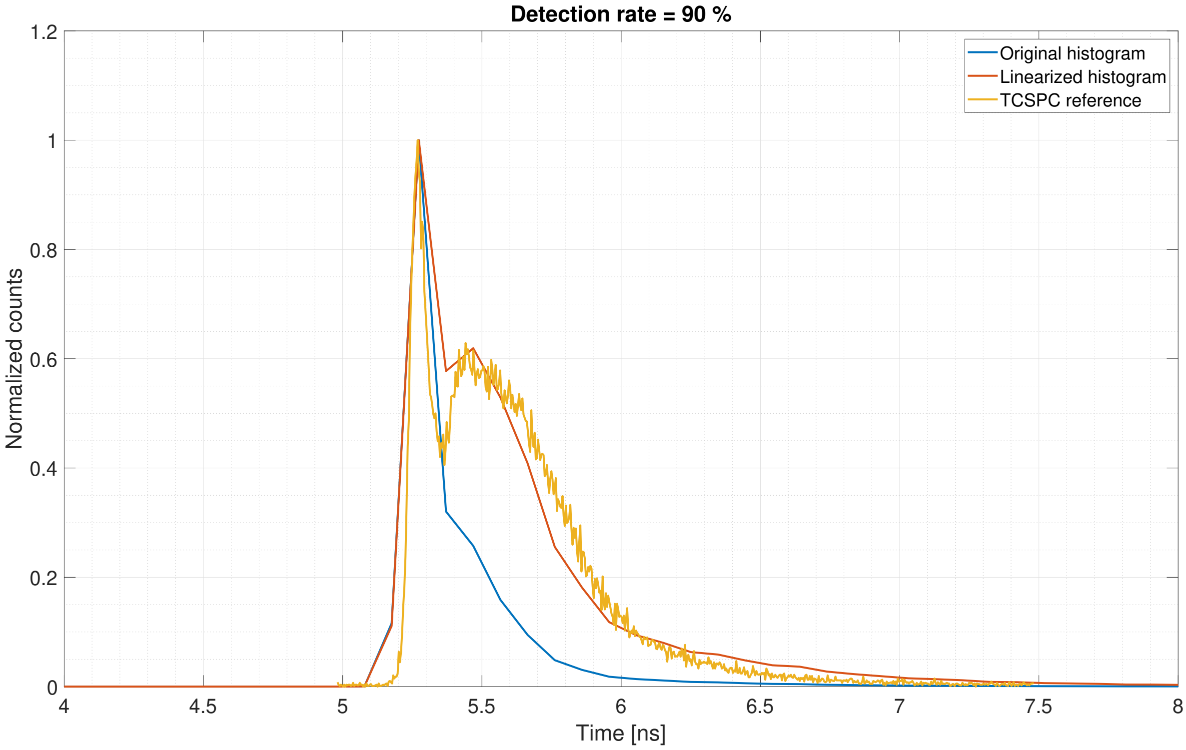


Figure 6.14: Characterization of the behavior of the proposed SPAD linearization method under strong pile-up conditions. The histogram obtained from the linearized vector of timestamps is compared against the original histogram (built from the detection of the first arrival time) and against a reference measurement obtained with a conventional TCSPC setup. In the histograms obtained from our sensor timestamps, the bin width is 100 ps, while the reference measurement from the TCSPC setup has 4 ps timing resolution. The proposed SPAD linearization method allows us to recover the full shape of the laser envelope even if the detection rate is 18 times higher than the conventional limit of 5%.

each run of the algorithm, we average the results from $N = 1.5 \cdot 10^4$ vectors of linearized SPAD timestamps, to emulate an equivalent 30 FPS operation rate, as outlined in Section 6.3.3 and with Figure 6.7. For all measurements, the same $\simeq 10\%$ reflectivity target (black matte panel) was used, in the range from 1 m to 3.8 m, to emulate a challenging scenario for a typical SPAD-based d-ToF system. First, we evaluate the behavior of the *ToF* extraction process without background light. The results, depicted in Figure 6.15, show good agreement between the extracted *ToF* and the ground truth. Then, we repeat the measurements with the inclusion of background light by setting the halogen illuminator to generate a background light flux of $7.7 \cdot 10^6$ events/s and $120 \cdot 10^6$ events/s. The values of background light flux are considered at the output of the SPADs of the sensor, and they correspond to an illumination level of $\simeq$ 15 kilolux and $\simeq$ 75 kilolux, respectively. Results are shown in Figure 6.16 and 6.17, demonstrating the validity of the proposed histogram-less *ToF* estimation in a real setup.

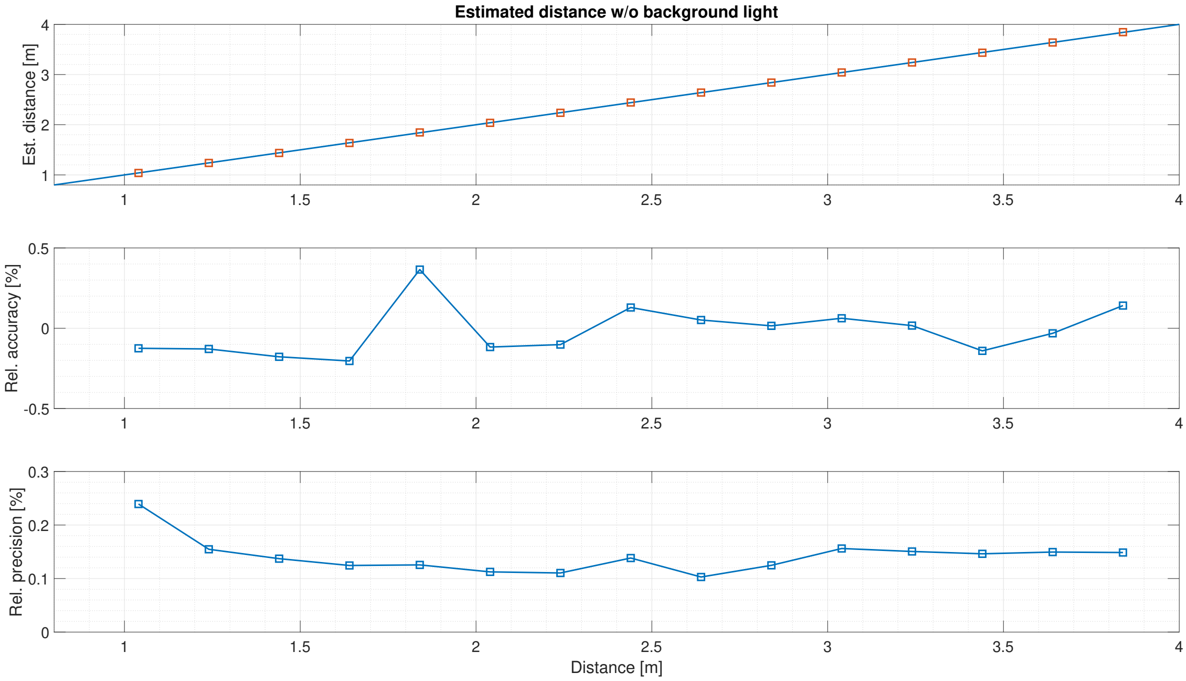


Figure 6.15: Measurement results with no background light, showing the *ToF* extracted without the need to build a histogram of timestamps. The relative accuracy is below ±0.5%, while the relative precision is below 0.25% for all measurements.

6.5 Summary

In this chapter, we demonstrate how to extract the time-of-flight information in a SPAD-based direct time-of-flight system without the need to build a resource and bandwidth-hungry histogram of timestamps. Moreover, the proposed method is resistant against high photon fluxes and can withstand detection rates three orders of magnitude higher than the conventionally recognized limit of 5%. The acquisition method, which is based on the linearization of the SPAD response, is suitable for integration in CMOS technology using low resources and is therefore scalable to large arrays, since it can be easily integrated per-pixel. The proposed extraction method has been completely characterized, first with Monte Carlo numerical simulations. The method is also mathematically justified, and we demonstrated its validity with real measurements, by repurposing an existing d-ToF sensor and using real data to extract the ToF. The proposed extraction method can be implemented at least in two ways, by means of the *acquire or discard* or *time-gated* detection schemes. While the *acquire or discard* scheme allows for the

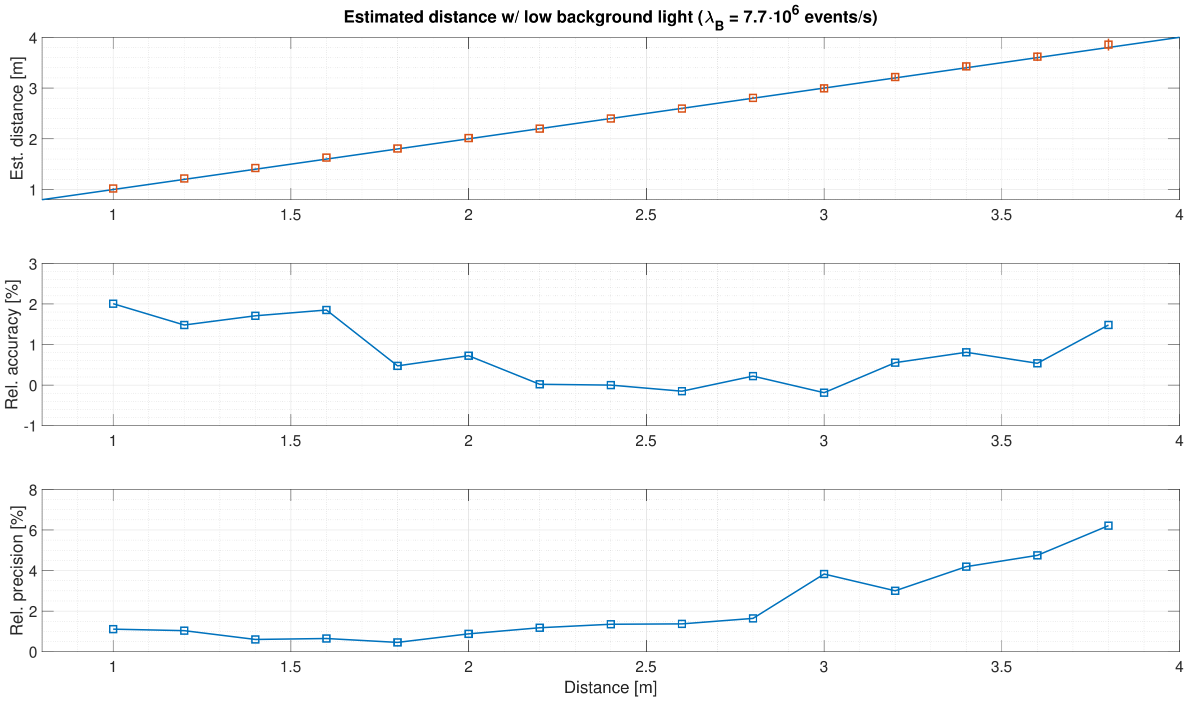


Figure 6.16: Measurement results with low background light flux ($\lambda_B = 7.7 \cdot 10^6$ events/s), showing the extracted *ToF* without the need to build a histogram of timestamps. The relative accuracy is in the range [-0.2,2] %, while the worst relative precision is 6% at the highest distance of 3.8 m.

least usage of resources, it suffers from long integration times especially when the flux of photons is too high. On the other hand, the *time-gated* scheme can guarantee a more efficient acquisition at the expense of a per-pixel controllable delay element. Concerning the ToF extraction method, we demonstrated its validity by using an extremely low amount of resources, as only two counters and one accumulator are required.

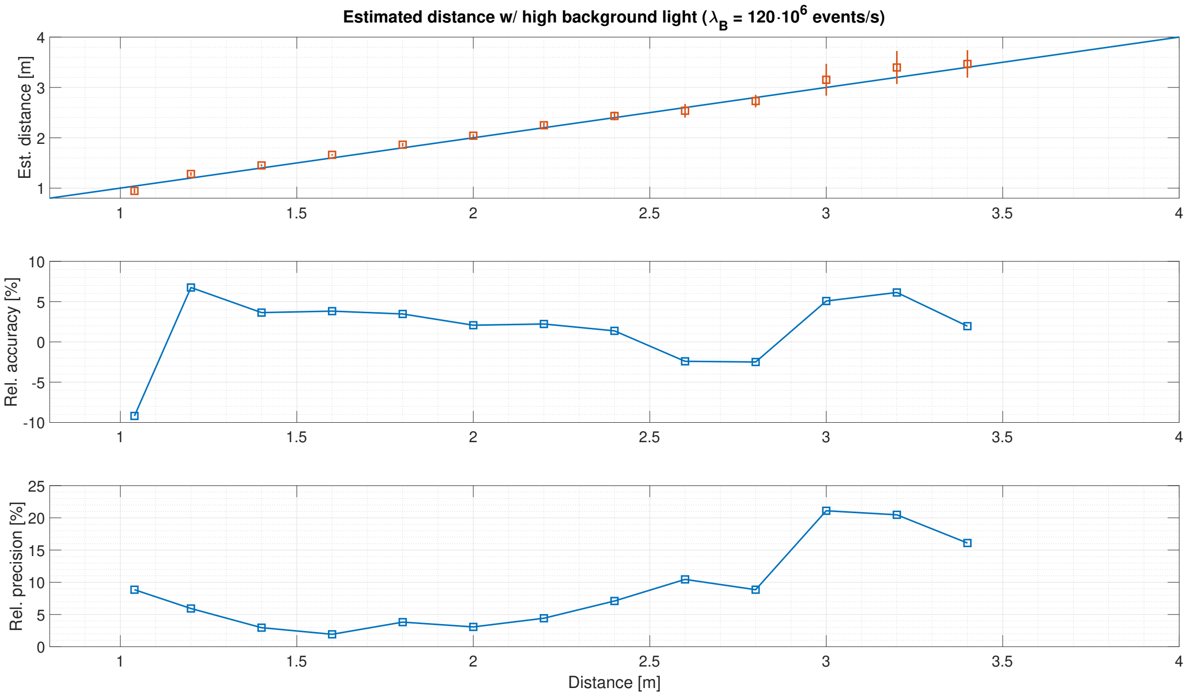


Figure 6.17: Measurement results with high background light flux ($\lambda_B = 120 \cdot 10^6$ events/s), showing the extracted *ToF* without the need to build a histogram of timestamps. The relative accuracy is in the range [-9,6.7] %, while the worst relative precision is 21% at 3 m. With this high background light flux (corresponding to ≈ 75 kilolux), and the decision to use a low reflectivity target (≈ 10%), the maximum achieved range decreased to 3.4 m.

Chapter 7

Conclusion

In this thesis, we investigate on techniques, architectures and methods to advance the state-of-the-art in the field of SPAD-based CMOS d-ToF systems, focusing in particular on three aspects: mitigation of the negative effects of background light, rejection of mutual interference and reduction in the amount of generated data. The main achievements are summarized following the order of chapters in Section 7.1, while final considerations and perspectives for future work are provided in Section 7.2.

7.1 Results

7.1.1 Modelling of SPAD-based d-ToF systems

SPAD-based CMOS d-ToF systems present a high degree of complexity and their performance depend upon a number of physical parameters. For such a reason, a system simulator represents a doable method to foresee the most suited architecture avoiding the risk to obtain either insufficient or overkilling performance. Additionally, some of the results presented in this thesis required the comparison of many architectures in different operating conditions, which could not be easily possible without the use of a system simulator.

The system simulator we thoroughly described in Chapter 3 employs a mixed physical-numerical model to emulate the response of a SPAD-based d-ToF sensor [97]. The physical model is used to estimate the so-called optical power budget, i.e., the total amount of optical power returning back to the detector during a time-of-flight measurement, which depends upon the parameters of the receiving and emitting optics, the laser power, the SPAD sensitivity and fill-factor and the target reflectivity, assuming a Lambertian diffusion of the scattered light. With the knowledge of the impinging optical power then, a numerical Monte Carlo engine emulates a train of SPAD detection events, assuming photon detection times follow a Poisson distribution. The simulator can successfully emulate the behavior of both *synchronous* and *asynchronous* SPAD driving mechanisms, together with the possibility to model an arbitrary laser pulse shape. The model has been validated against an existing SPAD-based d-ToF sensor [10], showing a good agreement in matching results from a real system.

7.1.2 Comparative evaluation of background-rejection techniques

Background light represents a limiting factor for SPAD-based d-ToF systems, in particular concerning their implementation in the automotive sector for ADAS. Over the years, already several approaches have been proposed to reject background light. However, the trend was to propose and evaluate different techniques independently from each other, with a lack of knowledge whether a possible combination of them could improve the performance.

For such reason, in Chapter 4 our effort focused on understanding whether or not two or more techniques can be combined with each other, and what the effect on the final measurement is [108]. Additionally, we tried to understand the impact of hardware implementation in the performance of a widely-known background rejection technique (photon coincidence) [109]. Several of the results from this investigation have been obtained by using the system simulator presented in Chapter 3.

The results we obtained can be summarized as follows:

- If compared one against the other, the *photon coincidence* technique outperforms the *Automatic-Sensitivity (AS)* technique, as the required level of PDP decrease goes beyond what can be reasonably obtained with the current technology.

- The combination of the *AS* technique on top of the *photon coincidence* can improve the performance by increasing the measurement range. In the context of the system setup selected for this study (a scanning d-ToF architecture), the range can be increased by up to 50 m with moderate background light intensity.

- Among the three selected hardware solutions to implement the *photon coincidence* technique, the *neighbor-based* approach [4] provides the best performance.

- The *last-hit* detection technique demonstrated to outperform both the *AS* and the *photon coincidence* techniques, in particular when the target is located toward the end of the measurement range, achieving an almost flat SNR response over the distance range. Nevertheless, also this technique can be used together with both *photon coincidence* and *AS* techniques.

7.1.3 Interference rejection

The diffusion of SPAD-based d-ToF systems in both industrial and automotive scenario, where several similar systems are expected to operate simultaneously, posed a great challenge as mutual system-to-system interference could lead to a decrease in performance and safety of operations.

Within the context of this thesis, we addressed this issue with the design of a prototype sensor in a 110nm CIS process [61], with a linear array of 64 pixels, described in Chapter 5. Despite at least two interference-rejection systems have been proposed in the literature [7, 8], the proposed design, to the best of our knowledge, proposes for the first time cancellation of interferences directly in-pixel, with a series of benefits in terms of measurement reliability and generation of data. Moreover, a smart readout scheme has been implemented to take advantage of the reduced amount of generated data when interference and background-rejection techniques are operated.

The main achievements from the designed sensor are summarized as follows:

- For the first time, interference can be rejected directly in-pixel. This enables the construction of an interference-free histogram of timestamps, reducing at the same time the amount of generated data and increasing the reliability of the measurement. From the measurements on the fabricated device, interference can be almost completely suppressed by -42.5 dB in the final histogram of timestamps.

- The interference-rejection technique implemented in the device can also be used to reject background light, in particular if combined with the photon coincidence technique. The measured laser peak to background ratio in challenging background illumination conditions increased by up to 23 dB, almost completely eliminating the background contribution in the final histogram of timestamps.

- To take advantage of the general reduction of generated data, a dedicated readout architecture has been implemented, which is quasi-optimum in terms of efficiency and allowed up to a measured 97.75% of readout bandwidth saving in the most challenging scenario.

7.1.4 Histogram-less and high-flux LiDAR

The increasing demand for high depth and imaging resolutions and extended measurement range, pushes the classical histogram-based d-ToF approach to its limit, due to the enormous amount of data which has to be transferred, stored and elaborated to extract the time-of-flight information.

In this thesis, an extensive investigation aimed at overcoming this limitation has been presented in Chapter 6. Thanks to the linearization of the SPAD response over time, it is possible to achieve a histogram-less approach to compute the time-of-flight, which can be implemented with a modest amount of hardware resources [110]. Additionally, the linearized SPAD response naturally overcomes the pile-up distortion problem, enabling safe operation of LiDAR even under high photon flux conditions.

The most important achievements are summarized as follows:

- A SPAD linearization process has been studied and verified using real data. The linearization process can be carried out in at least two ways. The first method guarantees the least amount of resources by using a comparator and a single slot of memory, but results in a potentially long acquisition time. The second method requires more hardware, as a finely-controlled delay element is required, allowing for a much shorter acquisition time. A third method, blends elements of the other two, providing the best trade-off between complexity and acquisition time.

- A histogram-less approach for SPAD-based LiDAR has been studied and verified using real data from an existing d-ToF sensor. The computation of the time-of-flight can be executed by using only an accumulator and a pair of counters, thanks to the assumption of a linearized SPAD response over time. The time-of-flight has been computed from real measurement data up to a distance of 3.8 m, guaranteeing 30 FPS of acquisition rate.

- The linearized SPAD response enables the acquisition even under prohibitive pile-up conditions, overcoming the standard limit of 5% detection rate, providing benefits in terms of acquisition time and ease of experiment setup for TCSPC applications. According to measurement results, it is possible to recover the true envelope of the laser profile even under detection rates higher than 90%.

7.2 Future work and research

In Chapter 4, we demonstrate that the performance of SPAD-based LiDAR systems (in terms of both probability of correct measurement and SNR) can increase with the proper combination of two background-rejection techniques. Additionally, we show that another technique, especially when the target is located toward the far-end of the measurement range, can provide the best performance over the others. Given the diverse operational environments expected for LiDAR systems —including variations in target distance, reflectivity, background illumination, illumination type, etc.—, a specific technique could provide better performance than a different one for any combination of the aforesaid factors. Consequently, an ideal SPAD-based LiDAR should be capable of dynamically selecting the most suitable background-rejection technique, ensuring maximum performance across a broad spectrum of external conditions.

At the current state-of-the-art, this kind of decision is taken at a higher hierarchical level, usually at the controller level by analysing the resulting histogram of timestamps or final 3D image, making it difficult to quickly and dynamically react to a change in external conditions. Moreover, it is a common choice to select the same background-rejection technique for the entire array of pixels (via SPI programming), even though different portions of the imaged area may require an ad-hoc setting.

For such reasons, a promising direction for future research involves the implementation of dynamic and per-pixel selection techniques, enabling the sensor to quickly and autonomously adjust its settings based on external conditions, possibly taking advantage of the advancements in 3D-stacked solutions and machine learning techniques.

In Chapter 5, we demonstrate the possibility to integrate at pixel level interference-rejection capability, together with an optimized readout architecture which takes advantage of the natural decrease in the amount of generated data when such technique is applied. The principle was tested in a real environment by means of a dedicated prototype sensor. The prototype sensor is a linear array of 64 pixels, thus limiting the possibility to obtain 3D images only with the implementation of a scanning setup.

The natural continuation of this research, which has been already planned for the upcoming activities, is the extension of the prototype chip to a bidimensional array of pixels, with a resolution of at least 64x64 pixels, to take advantage of a flash illumination of the scene. The extension from a linear to a 2D configuration requires the re-design of the readout architecture, to include the concept of vertical scanning of the array, and the re-design of the in-pixel control circuit, to be able to include the required functionality within a 2D array. The envisaged pixel pitch will be $< 45\mu m$.

In Chapter 6, we show a novel approach to SPAD-based LiDAR, where by means of the linearization of the SPAD response, a histogram-less approach can be achieved. The approach is verified by means of models and supported by real data, obtained by repurposing an existing d-ToF sensor.

The proposed linearization method has been studied and verified by considering a synchronous SPAD detection approach. Future planned research will investigate the possibility (and challenges) to obtain the same result but considering an asynchronous SPAD detection scheme. We also envisage additional measurements in different conditions of SNR and background intensity and an in-depth comparison with other existing techniques (e.g., partial/full on-chip histogramming, time-gating, etc.). Moreover, future work will include the design and fabrication of a device where both the SPAD linearization procedure and the histogram-less ToF extraction are implemented in pixel by dedicated circuitry.

Bibliography

[1] Dariusz Man and Ryszard Olchawa. Brain Biophysics: Perception, Consciousness, Creativity. Brain Computer Interface (BCI). In Wojciech P. Hunek and Szczepan Paszkiel, editors, *Biomedical Engineering and Neuroscience*, pages 38–44, Cham, 2018. Springer International Publishing.

[2] Jamie Enoch, Leanne McDonald, Lee Jones, Pete R. Jones, and David P. Crabb. Evaluating Whether Sight Is the Most Valued Sense. *JAMA Ophthalmology*, 137(11):1317, November 2019.

[3] Fabian Hutmacher. Why Is There So Much More Research on Vision Than on Any Other Sensory Modality? *Frontiers in Psychology*, 10, 2019.

[4] Enrico Manuzzato, Alessandro Tontini, Andrej Seljak, and Matteo Perenzoni. A 64×64-pixel flash LiDAR SPAD imager with distributed pixel-to-pixel correlation for background rejection, tunable automatic pixel sensitivity and first-last event detection strategies for space applications. In *2022 IEEE International Solid-State Circuits Conference (ISSCC).*

[5] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa. Avalanche photodiodes and quenching circuits for single-photon detection. *Appl. Opt.*, 35(12):1956–1976, Apr 1996.

[6] Preethi Padmanabhan, Chao Zhang, Marco Cazzaniga, Baris Efe, Augusto R. Ximenes, Myung-Jae Lee, and Edoardo Charbon. 7.4 a 256×128 3D-stacked (45nm) SPAD FLASH LiDAR with 7-level coincidence detection and progressive gating for 100m range and 10klux background light. In *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, volume 64, pages 111–113, 2021.

[7] Hyeongseok Seo, Heesun Yoon, Dongkyu Kim, Jungwoo Kim, Seong-Jin Kim, Jung-Hoon Chun, and Jaehyuk Choi. Direct TOF scanning lidar sensor with

two-step multievent histogramming TDC and embedded interference filter. *IEEE Journal of Solid-State Circuits*, 56(4):1022–1035, 2021.

[8] Augusto Ronchini Ximenes, Preethi Padmanabhan, Myung-Jae Lee, Yuichiro Yamashita, D. N. Yaung, and Edoardo Charbon. A 256×256 45/65nm 3D-stacked spad-based direct TOF image sensor for LiDAR applications with optical polar modulation for up to 18.6db interference suppression. In *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, pages 96–98, 2018.

[9] Scott Lindner, Chao Zhang, Ivan Michel Antolovic, Martin Wolf, and Edoardoseo Charbon. A 252 × 144 SPAD pixel flash lidar with 1728 dual-clock 48.8 ps TDCs, integrated histogramming and 14.9-to-1 compression in 180nm CMOS technology. In *2018 IEEE Symposium on VLSI Circuits*, pages 69–70, 2018.

[10] Matteo Perenzoni, Daniele Perenzoni, and David Stoppa. A 64 × 64-pixels digital silicon photomultiplier direct TOF sensor with 100-Mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6 km for spacecraft navigation and landing. *IEEE Journal of Solid-State Circuits*, 52(1):151–160, 2017.

[11] A. Antognetti, Sergio Cova, and Antonio Longoni. A study of the operation and performances of an avalanche diode as a single photon detector. 1975.

[12] Andrea Gallivanoni, Ivan Rech, and Massimo Ghioni. Progress in quenching circuits for single photon avalanche diodes. *IEEE Transactions on Nuclear Science*, 57(6):3815–3826, 2010.

[13] Arin Can Ulku, Claudio Bruschini, Ivan Michel Antolović, Yung Kuo, Rinat Ankri, Shimon Weiss, Xavier Michalet, and Edoardo Charbon. A 512 × 512 SPAD image sensor with integrated gating for widefield FLIM. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(1):1–12, 2019.

[14] Arin Ulku, Andrei Ardelean, Michel Antolovic, Shimon Weiss, Edoardo Charbon, Claudio Bruschini, and Xavier Michalet. Wide-field time-gated SPAD imager for phasor-based FLIM applications. *Methods and Applications in Fluorescence*, 8(2):024002, February 2020.

[15] Majid Zarghami, Leonardo Gasparini, Luca Parmesan, Manuel Moreno-Garcia, Andrè Stefanov, Bänz Bessire, Manuel Unternährer, and Matteo Perenzoni. A

32 × 32-pixel CMOS imager for quantum optics with per-SPAD TDC, 19.48% fill-factor in a 44.64-μm pitch reaching 1-Mhz observation rate. *IEEE Journal of Solid-State Circuits*, 55(10):2819–2830, 2020.

[16] Leonardo Gasparini, Majid Zarghami, Hesong Xu, Luca Parmesan, Manuel Moreno Garcia, Manuel Unternährer, Bänz Bessire, André Stefanov, David Stoppa, and Matteo Perenzoni. A 32×32-pixel time-resolved single-photon image sensor with 44.64μm pitch and 19.48% fill-factor with on-chip row/frame skipping features reaching 800khz observation rate for quantum physics applications. In *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, pages 98–100, 2018.

[17] Matteo Perenzoni, Nicola Massari, Leonardo Gasparini, Manuel Moreno Garcia, Daniele Perenzoni, and David Stoppa. A fast 50 × 40-pixels single-point DTOF SPAD sensor with photon counting and programmable ROI TDCs, with σ <4 mm at 3 m up to 18 klux of background light. *IEEE Solid-State Circuits Letters*, 3:86–89, 2020.

[18] F. Zappa, S. Tisa, A. Tosi, and S. Cova. Principles and features of single-photon avalanche diode arrays. *Sensors and Actuators A: Physical*, 140(1):103–112, 2007.

[19] Manuel Moreno-García, Hesong Xu, Leonardo Gasparini, and Mattero Perenzoni. Low-noise single photon avalanche diodes in a 110nm CIS technology. In *2018 48th European Solid-State Device Research Conference (ESSDERC)*, pages 94–97, 2018.

[20] R.J. McIntyre. On the avalanche initiation probability of avalanche diodes above the breakdown voltage. *IEEE Transactions on Electron Devices*, 20(7):637–641, 1973.

[21] Tomer Leitner, Amos Feiningstein, Renato Turchetta, Rebecca Coath, Steven Chick, Gil Visokolov, Vitali Savuskan, Michael Javitt, Lior Gal, Igor Brouk, Sharon Bar-Lev, and Yael Nemirovsky. Measurements and simulations of low dark count rate single photon avalanche diode device in a low voltage 180-nm CMOS image sensor technology. *IEEE Transactions on Electron Devices*, 60(6):1982–1988, 2013.

[22] Chockalingam Veerappan and Edoardo Charbon. A low dark count p-i-n diode based SPAD in CMOS technology. *IEEE Transactions on Electron Devices*, 63(1):65–71, 2016.

[23] K. Ito, Y. Otake, Y. Kitano, A. Matsumoto, J. Yamamoto, T. Ogasahara, H. Hiyama, R. Naito, K. Takeuchi, T. Tada, K. Takabayashi, H. Nakayama, K. Tatani, T. Hirano, and T. Wakano. A back illuminated 10μm SPAD pixel array comprising full trench isolation and cu-cu bonding with over 14% PDE at 940nm. In *2020 IEEE International Electron Devices Meeting (IEDM)*, pages 16.6.1–16.6.4, 2020.

[24] Angelo Gulinatti, Francesco Ceccarelli, Massimo Ghioni, and Ivan Rech. Custom silicon technology for SPAD-arrays with red-enhanced sensitivity and low timing jitter. *Opt. Express*, 29(3):4559–4581, Feb 2021.

[25] Edward Van Sieleghem, Gauri Karve, Koen De Munck, Andrea Vinci, Celso Cavaco, Andreas Süss, Chris Van Hoof, and Jiwon Lee. A backside-illuminated charge-focusing silicon SPAD with enhanced near-infrared sensitivity. *IEEE Transactions on Electron Devices*, 69(3):1129–1136, 2022.

[26] Alessandro Tontini, Leonardo Gasparini, Nicola Massari, and Roberto Passerone. SPAD-based quantum random number generator with an N^{th}-order rank algorithm on FPGA. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66(12):2067–2071, 2019.

[27] Shingo Mandai, Matthew W. Fishburn, Yuki Maruyama, and Edoardo Charbon. A wide spectral range single-photon avalanche diode fabricated in an advanced 180 nm CMOS technology. *Opt. Express*, 20(6):5849–5857, Mar 2012.

[28] E. Charbon. Single-photon imaging in complementary metal oxide semiconductor processes. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2012):20130100, March 2014.

[29] Robert K Henderson, Justin Richardson, and Lindsay Grant. Reduction of band-to-band tunneling in deep-submicron cmos single photon avalanche photodiodes. In *International Image Sensor Workshop (IISW 2009), Bergen, Norway*, pages 26–28, 2009.

[30] Mohammad Azim Karami, Cristiano Niclass, and Edoardo Charbon. Random telegraph signal in single-photon avalanche diodes. In *Int. Image Sensor Workshop*, pages 1–4, 2009.

[31] Yux Xu, Ping Xiang, and Xiaopeng Xie. Comprehensive understanding of dark count mechanisms of single-photon avalanche diodes fabricated in deep submicron cmos technologies. *Solid-State Electronics*, 129:168–174, 2017.

[32] Abdul Waris Ziarkash, Siddarth Koduru Joshi, Mario Stipčević, and Rupert Ursin. Comparative study of afterpulsing behavior and models in single photon counting avalanche photo diode detectors. *Sci. Rep.*, 8(1), March 2018.

[33] S. Farina, I. Labanca, G. Acconcia, M. Ghioni, and I. Rech. 10-nanosecond dead time and low afterpulsing with a free-running reach-through single-photon avalanche diode. *Review of Scientific Instruments*, 93(5):053102, 05 2022.

[34] C.L. Niclass, A. Rochas, P.A. Besse, and E. Charbon. A CMOS single photon avalanche diode array for 3D imaging. In *2004 IEEE International Solid-State Circuits Conference (IEEE Cat. No.04CH37519)*, pages 120–517 Vol.1, 2004.

[35] C. Niclass, A. Rochas, P.-A. Besse, and E. Charbon. Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes. *IEEE Journal of Solid-State Circuits*, 40(9):1847–1854, 2005.

[36] C. Niclass and E. Charbon. A single photon detector array with 64/spl times/64 resolution and millimetric depth accuracy for 3D imaging. In *ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005.*, pages 364–604 Vol. 1, 2005.

[37] Cristiano Niclass, Maximilian Sergio, and Edoardo Charbon. A CMOS 64×48 single photon avalanche diode array with event-driven readout. In *2006 Proceedings of the 32nd European Solid-State Circuits Conference*, pages 556–559, 2006.

[38] D. Stoppa, L. Pancheri, M. Scandiuzzo, M. Malfatti, G. Pedretti, and L. Gonzo. A single-photon-avalanche-diode 3D imager. In *Proceedings of the 31st European Solid-State Circuits Conference, 2005. ESSCIRC 2005.*, pages 487–490, 2005.

[39] Cristiano Niclass, Claudio Favi, Theo Kluter, Marek Gersbach, and Edoardo Charbon. A 128×128 single-photon image sensor with column-level 10-bit time-to-digital converter array. *IEEE Journal of Solid-State Circuits*, 43(12):2977–2989, 2008.

[40] Justin Richardson, Richard Walker, Lindsay Grant, David Stoppa, Fausto Borghetti, Edoardo Charbon, Marek Gersbach, and Robert K. Henderson. A

32×32 50ps resolution 10 bit time to digital converter array in 130nm CMOS for time correlated imaging. In *2009 IEEE Custom Integrated Circuits Conference*, pages 77–80, 2009.

[41] David Stoppa, Fausto Borghetti, Justin Richardson, Richard Walker, Lindsay Grant, Robert K. Henderson, Marek Gersbach, and Edoardo Charbon. A 32x32-pixel array with in-pixel photon counting and arrival time measurement in the analog domain. In *2009 Proceedings of ESSCIRC*, pages 204–207, 2009.

[42] Cristiano Niclass, Mineki Soga, and Satoru Kato. A 0.18μm CMOS single-photon sensor for coaxial laser rangefinders. In *2010 IEEE Asian Solid-State Circuits Conference*, pages 1–4, 2010.

[43] Cristiano Niclass, Mineki Soga, Hiroyuki Matsubara, Satoru Kato, and Manabu Kagami. A 100-m range 10-frame/s 340×96-pixel time-of-flight depth sensor in 0.18-μm CMOS. *IEEE Journal of Solid-State Circuits*, 48(2):559–572, 2013.

[44] Francescopaolo Mattioli Della Rocca, Tarek Al Abbas, Neale A. W. Dutton, and Robert K. Henderson. A high dynamic range SPAD pixel for time of flight imaging. In *2017 IEEE SENSORS*, pages 1–3, 2017.

[45] Henna Ruokamo, Lauri W. Hallman, and Juha Kostamovaara. An 80 × 25 pixel CMOS single-photon sensor with flexible on-chip time gating of 40 subarrays for solid-state 3-D range imaging. *IEEE Journal of Solid-State Circuits*, 54(2):501–510, 2019.

[46] Lucio Pancheri and David Stoppa. A SPAD-based pixel linear array for high-speed time-gated fluorescence lifetime imaging. In *2009 Proceedings of ESSCIRC*, pages 428–431, 2009.

[47] Yuki Maruyama and Edoardo Charbon. A time-gated 128x128 CMOS SPAD array for on-chip fluorescence detection. 2011.

[48] Lucio Pancheri, Nicola Massari, and David Stoppa. SPAD image sensor with analog counting pixel for time-resolved fluorescence detection. *IEEE Transactions on Electron Devices*, 60(10):3442–3449, 2013.

[49] Kazuhiro Morimoto, Andrei Ardelean, Ming-Lo Wu, Arin Can Ulku, Ivan Michel Antolovic, Claudio Bruschini, and Edoardo Charbon. Megapixel time-gated

SPAD image sensor for 2D and 3D imaging applications. *Optica*, 7(4):346–354, Apr 2020.

[50] Michael Wayne, Arin Ulku, Andrei Ardelean, Paul Mos, Claudio Bruschini, and Edoardo Charbon. A 500 × 500 dual-gate SPAD imager with 100% temporal aperture and 1 ns minimum gate length for FLIM and phasor imaging applications. *IEEE Transactions on Electron Devices*, 69(6):2865–2872, 2022.

[51] Alberto Dalla Mora, Alberto Tosi, Franco Zappa, Sergio Cova, Davide Contini, Antonio Pifferi, Lorenzo Spinelli, Alessandro Torricelli, and Rinaldo Cubeddu. Fast-gated single-photon avalanche diode for wide dynamic range near infrared spectroscopy. *IEEE Journal of Selected Topics in Quantum Electronics*, 16(4):1023–1030, 2010.

[52] I. Nissinen, J. Nissinen, A-K. Länsman, L. Hallman, A. Kilpelä, J. Kostamovaara, M. Kögler, M. Aikio, and J. Tenhunen. A sub-ns time-gated CMOS single photon avalanche diode detector for raman spectroscopy. In *2011 Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, pages 375–378, 2011.

[53] I. Nissinen, A-K. Länsman, J. Nissinen, J. Holma, and J. Kostamovaara. 2×(4×)128 time-gated CMOS single photon avalanche diode line detector with 100 ps resolution for Raman spectroscopy. In *2013 Proceedings of the ESSCIRC (ESSCIRC)*, pages 291–294, 2013.

[54] Neale A. W. Dutton, Salvatore Gnecchi, Luca Parmesan, Andrew J. Holmes, Bruce Rae, Lindsay A. Grant, and Robert K. Henderson. 11.5 a time-correlated single-photon-counting sensor with 14GS/s histogramming time-to-digital converter. In *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pages 1–3, 2015.

[55] Sam W. Hutchings, Nick Johnston, Istvan Gyongy, Tarek Al Abbas, Neale A. W. Dutton, Max Tyler, Susan Chan, Jonathan Leach, and Robert K. Henderson. A reconfigurable 3-D-stacked SPAD imager with in-pixel histogramming for flash LiDAR or high-speed time-of-flight imaging. *IEEE Journal of Solid-State Circuits*, 54(11):2947–2956, 2019.

[56] Sarrah M. Patanwala, Istvan Gyongy, Hanning Mai, Andreas Aßmann, Neale A. W. Dutton, Bruce R. Rae, and Robert K. Henderson. A high-throughput

photon processing technique for range extension of SPAD-based LiDAR receivers. *IEEE Open Journal of the Solid-State Circuits Society*, 2:12–25, 2022.

[57] Sara Grollius, Simon Grosse, Manuel Ligges, and Anton Grabmaier. Optimized interference suppression for tcspc lidar. *IEEE Sensors Journal*, 22(24):24094–24101, 2022.

[58] Bernhard Buttgen and Peter Seitz. Robust optical time-of-flight range imaging based on smart pixel structures. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 55(6):1512–1525, 2008.

[59] Achuta Kadambi, Refael Whyte, Ayush Bhandari, Lee Streeter, Christopher Barsi, Adrian Dorrington, and Ramesh Raskar. Coded time of flight cameras: Sparse deconvolution to address multipath interference and recover time profiles. *ACM Trans. Graph.*, 32(6), November 2013.

[60] Thomas Fersch, Robert Weigel, and Alexander Koelpin. A CDMA modulation technique for automotive time-of-flight LiDAR systems. *IEEE Sensors Journal*, 17(11):3507–3516, 2017.

[61] Alessandro Tontini, Leonardo Gasparini, and Roberto Passerone. A SPAD-based linear sensor with in-pixel temporal pattern detection for interference and background rejection with smart readout scheme. In *Proceedings of the 2023 International Image Sensor Workshop*, Crieff, Scotland, UK, May 21–25, 2023.

[62] Vincenzo Sesta, Klaus Pasquinelli, Renato Federico, Franco Zappa, and Federica Villa. Range-finding SPAD array with smart laser-spot tracking and TDC sharing for background suppression. *IEEE Open Journal of the Solid-State Circuits Society*, 2:26–37, 2022.

[63] An-Tai Hsiao, Chun-Hsien Liu, Po-Hsuan Chen, Yao-Lun Liu, Wei-Chi Wang, Tzu-Hsien Sang, Chia-Ming Tsai, Gray Lin, Jiun-In Guo, and Sheng-Di Lin. Real-time LiDAR module with 64x128-pixel CMOS SPAD array and 940-nm PCSEL. In *2022 IEEE Sensors Applications Symposium (SAS)*, pages 1–6, 2022.

[64] Federica Villa, Fabio Severini, Francesca Madonini, and Franco Zappa. SPADs and SiPMs arrays for long-range high-speed light detection and ranging (LiDAR). *Sensors*, 21(11), 2021.

[65] Francesco Mattioli Della Rocca, Hanning Mai, Sam W. Hutchings, Tarek Al Abbas, Kasper Buckbee, Andreas Tsiamis, Peter Lomax, Istvan Gyongy, Neale A. W. Dutton, and Robert K. Henderson. A 128 × 128 SPAD motion-triggered time-of-flight image sensor with in-pixel histogram and column-parallel vision processor. *IEEE Journal of Solid-State Circuits*, 55(7):1762–1775, 2020.

[66] Chao Zhang, Scott Lindner, Ivan Michel Antolović, Juan Mata Pavia, Martin Wolf, and Edoardo Charbon. A 30-frames/s, 252 × 144 SPAD flash LiDAR with 1728 dual-clock 48.8-ps TDCs, and pixel-wise integrated histogramming. *IEEE Journal of Solid-State Circuits*, 54(4):1137–1151, 2019.

[67] Bumjun Kim, Seonghyeok Park, Jung-Hoon Chun, Jaehyuk Choi, and Seong-Jin Kim. 7.2 a 48×40 13.5mm depth resolution flash LiDAR sensor with in-pixel zoom histogramming time-to-digital converter. In *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, volume 64, pages 108–110, 2021.

[68] David Stoppa, Sargis Abovyan, Daniel Furrer, Radoslaw Gancarz, Thomas Jessening, Robert Kappel, Manfred Lueger, Christian Mautner, Ian Mills, Daniele Perenzoni, Georg Roehrer, and Pierre-Yves Taloud. A reconfigurable QVGA/Q3VGA direct time-of-flight 3d imaging system with onchip depth-map computation in 45/40 nm 3D-stacked BSI SPAD CMOS. In *Proc. Int. Image Sensor Workshop, 2021.*, 2021.

[69] Oichi Kumagai, Junichi Ohmachi, Masao Matsumura, Shinichiro Yagi, Kenichi Tayu, Keitaro Amagawa, Tomohiro Matsukawa, Osamu Ozawa, Daisuke Hirono, Yasuhiro Shinozuka, Ryutaro Homma, Kumiko Mahara, Toshio Ohyama, Yousuke Morita, Shohei Shimada, Takahisa Ueno, Akira Matsumoto, Yusuke Otake, Toshifumi Wakano, and Takashi Izawa. 7.3 a 189×600 back-illuminated stacked SPAD direct time-of-flight depth sensor for automotive LiDAR systems. In *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, volume 64, pages 110–112, 2021.

[70] Istvan Gyongy, N.A.W. Dutton, H. Mai, F. Mattioli Della Rocca, and R.K. Henderson. A 200kFPS, 256 x 128 SPAD dToF sensor with peak tracking and smart readout. In *Proc. Ing. Image Sensor Workshop, 2021.*, 2021.

[71] Seonghyeok Park, Bumjun Kim, Junhee Cho, Jung-Hoon Chun, Jaehyuk Choi, and Seong-Jin Kim. An 80×60 flash LiDAR sensor with in-pixel histogramming TDC based on quaternary search and time-gated δ-intensity phase detection for

45m detectable range and background light cancellation. In *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, volume 65, pages 98–100, 2022.

[72] Chao Zhang, Ning Zhang, Zhijie Ma, Letian Wang, Yu Qin, Jieyang Jia, and Kai Zang. A 240 × 160 3D-stacked SPAD dToF image sensor with rolling shutter and in-pixel histogram for mobile devices. *IEEE Open Journal of the Solid-State Circuits Society*, 2:3–11, 2022.

[73] Istvan Gyongy, Ahmet T. Erdogan, Neale A.W. Dutton, Germán Mora Martín, Alistair Gorman, Hanning Mai, Francesco Mattioli Della Rocca, and Robert K. Henderson. A direct time-of-flight image sensor with in-pixel surface detection and dynamic vision. *IEEE Journal of Selected Topics in Quantum Electronics*, pages 1–12, 2023.

[74] Atul Ingle and David Maier. Count-free single-photon 3D imaging with race logic. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2023.

[75] Felipe Gutierrez-Barragan, Atul Ingle, Trevor Seets, Mohit Gupta, and Andreas Velten. Compressive single-photon 3D cameras. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17833–17843, 2022.

[76] Michael P. Sheehan, Julián Tachella, and Mike E. Davies. A sketching framework for reduced data transfer in photon counting lidar. *IEEE Transactions on Computational Imaging*, 7:989–1004, 2021.

[77] Robert K. Henderson, Nick Johnston, Sam W. Hutchings, Istvan Gyongy, Tarek Al Abbas, Neale Dutton, Max Tyler, Susan Chan, and Jonathan Leach. 5.7 a 256×256 40nm/90nm CMOS 3D-stacked 120db dynamic-range reconfigurable time-resolved SPAD imager. In *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, pages 106–108, 2019.

[78] M. Beer, C. Thattil, J. F. Haase, W. Brockherde, and R. Kokozinski. 2×192 pixel CMOS SPAD-based flash LiDAR sensor with adjustable background rejection. In *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pages 17–20, Dec 2018.

[79] Maik Beer, Olaf M. Schrey, Jan F. Haase, Jennifer Ruskowski, Werner Brockherde, Bedrich J. Hosticka, and Rainer Kokozinski. SPAD-based flash LiDAR

sensor with high ambient light rejection for automotive applications. In Manijeh Razeghi, Gail J. Brown, Jay S. Lewis, and Giuseppe Leo, editors, *Quantum Sensing and Nano Electronics and Photonics XV*, volume 10540, pages 320 – 327. International Society for Optics and Photonics, SPIE, 2018.

[80] Maik Beer, Jan Haase, Jennifer Ruskowski, and Rainer Kokozinski. Background light rejection in SPAD-based LiDAR sensors by adaptive photon coincidence detection. *Sensors*, 18(12):4338, 2018.

[81] D. Stoppa, L. Pancheri, M. Scandiuzzo, L. Gonzo, G. Dalla Betta, and A. Simoni. A CMOS 3-D imager based on single photon avalanche diode. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 54(1):4–12, Jan 2007.

[82] A. T. Erdogan, R. Walker, N. Finlayson, N. Krstajić, G. Williams, J. Girkin, and R. Henderson. A CMOS SPAD line sensor with per-pixel histogramming TDC for time-resolved multispectral imaging. *IEEE Journal of Solid-State Circuits*, 54(6):1705–1719, 2019.

[83] I. Vornicu, A. Darie, R. Carmona-Galán, and Á. Rodríguez-Vázquez. Compact real-time inter-frame histogram builder for 15-bits high-speed ToF-imagers based on single-photon detection. *IEEE Sensors Journal*, 19(6):2181–2190, 2019.

[84] Istvan Gyongy, Amy Davies, Benjamin Gallinet, Neale A.W. Dutton, Rory R. Duncan, Colin Rickman, Robert K. Henderson, and Paul A. Dalgarno. Cylindrical microlensing for enhanced collection efficiency of small pixel SPAD arrays in single-molecule localisation microscopy. *Opt. Express*, 26(3):2280–2291, Feb 2018.

[85] M. Beer, O. M. Schrey, B. J. Hosticka, and R. Kokozinski. Modelling of SPAD-based time-of-flight measurement techniques. In *2017 European Conference on Circuit Theory and Design (ECCTD)*, pages 1–4, Sep. 2017.

[86] Preethi Padmanabhan, Chao Zhang, and Edoardo Charbon. Modeling and analysis of a direct time-of-flight sensor architecture for LiDAR applications. *Sensors*, 19(24), 2019.

[87] Alessandro Tontini, Leonardo Gasparini, Lucio Pancheri, and Roberto Passerone. Design and characterization of a low-cost FPGA-based TDC. *IEEE Transactions on Nuclear Science*, 65(2):680–690, 2018.

[88] Luca Parmesan, Neale A. W. Dutton, Neil J. Calder, Andrew J. Holmes, Lindsay A. Grant, and Robert K. Henderson. A 9.8 μm sample and hold time to amplitude converter CMOS SPAD pixel. In *2014 44th European Solid State Device Research Conference (ESSDERC)*, pages 290–293, 2014.

[89] Al Naboulsi, Sizun, and D. Fornel. Propagation of optical and infrared waves in the atmosphere. 2005.

[90] Wenhua Song, JianCheng Lai, Zabih Ghassemlooy, Zhiyong Gu, Wei Yan, Chunyong Wang, and Zhenhua Li. The effect of fog on the probability density distribution of the ranging data of imaging laser radar. *AIP Advances*, 8(2):025022, 2018.

[91] Michael Wahl, Matthias Leifgen, Michael Berlin, Tino Röhlicke, Hans-Jürgen Rahn, and Oliver Benson. An ultrafast quantum random number generator with provably bounded output bias based on photon arrival time measurements. *Applied Physics Letters*, 98(17):171105, 2011.

[92] THORLABS. DET10A/M - Si Detector, 200 - 1100 nm, 1 ns Rise Time. *www.thorlabs.com/thorProduct.cfm?partNumber=DET10A/M*.

[93] Hesong Xu, Lucio Pancheri, Gian-Franco Dalla Betta, and David Stoppa. Design and characterization of a p+/n-well SPAD array in 150nm CMOS process. *Opt. Express*, 25(11):12765–12778, May 2017.

[94] M. Janecek. Reflectivity spectra for commonly used reflectors. *IEEE Transactions on Nuclear Science*, 59(3):490–497, June 2012.

[95] Maik Beer, Olaf M. Schrey, Bedrich J. Hosticka, and Rainer Kokozinski. Modelling of SPAD-based time-of-flight measurement techniques. In *2017 European Conference on Circuit Theory and Design (ECCTD)*, pages 1–4, 2017.

[96] Alfonso Incoronato, Mauro Locatelli, and Franco Zappa. Statistical model for SPAD-based time-of-flight systems and photons pile-up correction. In *2021 Conference on Lasers and Electro-Optics Europe & European Quantum Electronics Conference (CLEO/Europe-EQEC)*, pages 1–1, 2021.

[97] Alessandro Tontini, Leonardo Gasparini, and Matteo Perenzoni. Numerical model of SPAD-based direct time-of-flight flash LIDAR CMOS image sensors. *Sensors*, 20(18), 2020.

[98] Daniel Van Blerkom. Modelling TDC circuit performance for SPAD sensor arrays. In *Proceedings of the International SPAD Sensor Workshop*, ISSW 2020, Edinburgh, Scotland, UK, June 2020.

[99] Anant Gupta, Atul Ingle, and Mohit Gupta. Asynchronous single-photon 3D imaging. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7908–7917, 2019.

[100] Kentaro Yoshioka, Hiroshi Kubota, Tomonori Fukushima, Satoshi Kondo, Tuan Thanh Ta, Hidenori Okuni, Kaori Watanabe, Masatoshi Hirono, Yoshinari Ojima, Katsuyuki Kimura, Sohichiroh Hosoda, Yutaka Ota, Tomohiro Koizumi, Naoyuki Kawabe, Yasuhiro Ishii, Yoichiro Iwagami, Seitaro Yagi, Isao Fujisawa, Nobuo Kano, Tomohiko Sugimoto, Daisuke Kurose, Naoya Waki, Yumi Higashi, Tetsuya Nakamura, Yoshikazu Nagashima, Hirotomo Ishii, Akihide Sai, and Nobu Matsumoto. A 20-ch TDC/ADC hybrid architecture LiDAR SoC for 240 × 96 pixel 200-m range imaging with smart accumulation technique and residue quantizing SAR ADC. *IEEE Journal of Solid-State Circuits*, 53(11):3026–3038, 2018.

[101] Lucas J. Koerner. Models of direct time-of-flight sensor precision that enable optimal design and dynamic configuration. *IEEE Transactions on Instrumentation and Measurement*, 70:1–9, 2021.

[102] Toru Okino, Shota Yamada, Yusuke Sakata, Shigetaka Kasuga, Masato Takemoto, Yugo Nose, Hiroshi Koshida, Masaki Tamaru, Yuki Sugiura, Shigeru Saito, Shinzo Koyama, Mitsuyoshi Mori, Yutaka Hirose, Masayuki Sawada, Akihiro Odagawa, and Tsuyoshi Tanaka. 5.2 a 1200×900 6μ m 450fps geiger-mode vertical avalanche photodiodes CMOS image sensor for a 250m time-of-flight ranging system using direct-indirect-mixed frame synthesis with configurable-depth-resolution down to 10cm. In *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, pages 96–98, 2020.

[103] Ryan M. Field, Simeon Realov, and Kenneth L. Shepard. A 100 fps, time-correlated single-photon-counting-based fluorescence-lifetime imager in 130 nm CMOS. *IEEE Journal of Solid-State Circuits*, 49(4):867–880, 2014.

[104] Wolfgang Becker. Advanced time-correlated single photon counting techniques. *Springer science & business media*, 2005.

[105] Ahmet T. Erdogan, Richard Walker, Neil Finlayson, Nikola Krstajic, Gareth O. S. Williams, and Robert K. Henderson. A 16.5 giga events/s 1024 × 8 SPAD line sensor with per-pixel zoomable 50ps-6.4ns/bin histogramming TDC. In *2017 Symposium on VLSI Circuits*, pages C292–C293, 2017.

[106] Cristiano Niclass, Mineki Soga, Hiroyuki Matsubara, Masaru Ogawa, and Manabu Kagami. A 0.18-μ m CMOS SoC for a 100-m-range 10-frame/s 200 $\times$96-pixel time-of-flight depth sensor. *IEEE Journal of Solid-State Circuits*, 49(1):315–330, 2014.

[107] Filip Taneski, Tarek Al Abbas, and Robert K. Henderson. Laser power efficiency of partial histogram direct time-of-flight LiDAR sensors. *Journal of Lightwave Technology*, 40(17):5884–5893, 2022.

[108] Alessandro Tontini, Leonardo Gasparini, Enrico Manuzzato, Matteo Perenzoni, and Roberto Passerone. Comparison of background-rejection techniques for SPAD-based LiDAR systems. In *2022 17th Conference on Ph.D Research in Microelectronics and Electronics (PRIME)*, pages 45–48, 2022.

[109] Alessandro Tontini, Leonardo Gasparini, Enrico Manuzzato, Matteo Perenzoni, and Roberto Passerone. Comparative evaluation of background-rejection techniques for SPAD-based LiDAR systems. *Integration*, 90:1–10, 2023.

[110] Alessandro Tontini, Sonia Mazzucchi, Roberto Passerone, Nicolò Broseghini, and Leonardo Gasparini. Histogram-less LiDAR through SPAD response linearization. *IEEE Sensors Journal*, 24(4):4656–4669, 2024.

[111] Alessandro Tontini, Sonia Mazzucchi, Nicol'o Brosegini, Roberto Passerone, and Leonardo Gasparini. A histogram compensation process for spad-based d-tof lidar systems for high photon flux measurements. *Proceedings of SPIE Photonics Europe, Strasbourg, France, April 7-11, 2024.*