2. Phase-Shift-Based Time-of-Flight Imaging Systems

Time-of-Flight (ToF) is a widely spread technique for depth imaging. The term *depth imaging* is often used as a synonym of *3-dimensional (3D) imaging*, but *sensu stricto* the former is a specific case of the latter. In general terms, 3D imaging aims to capture 3D information of a scene in a 2D array of data (an image). The addition of a third dimension does not imply that the data are dense in all three dimensions of the scene space. We speak of dense 3D imaging if we are actually able to measure the value of a certain property—e. g., reflectivity—all along the three dimensions of the scene space. Examples of this kind are active systems using radiation that is able to go through objects up to a certain extent, allowing dense full 3D reconstructions. A well-known example is the Magnetic Resonance Imaging (MRI) [228] technique, widely used as non-invasive method to detect malfunctions within alive biological systems.

In most cases, the radiation used for imaging does not penetrate the object and we get sparse data in the third dimension. In this category fall conventional depth imaging systems, which provide a collection of distances from the camera to the scene points, in the shape of a depth image. Such cameras also provide an intensity value per pixel. If a 3D intensity map is reconstructed, we obtain a surface in 3D-space, not a dense volume. For this reason, depth cameras are often called 2.5D sensors, instead of 3D sensors, since they can only capture surfaces in 3D.

More recently, the development of novel transient imaging techniques [445] has allowed recovering an approximate time profile of the returning light signal in ToF imaging, opening the possibility of sensing several depths per pixel when dealing with translucent objects in the scene, recovering its volume. This concept, in combination with the PMD chip, allows for low-cost systems able to face the ToF multipath problem [248] and sense occluded objects [220].

M. Heredia Conde, *Compressive Sensing for the Photonic Mixer Device*, DOI 10.1007/978-3-658-18057-7_2

Apart from these exceptions, most 3D imaging devices sense the scene as a discrete 2D-array, offering a single depth value per pixel. The interested reader can find a survey of the multiple possibilities that consumer depth cameras offer in the field of computer vision in [181]. In this chapter we provide an introduction to conventional depth imaging methods, pointing out the main strengths and weaknesses of each technology. We describe the principle of operation of ToF imaging and focus on the Photonic Mixer Device (PMD), as reference technology for phase-shift-based ToF imaging. The understanding of the PMD principle of operation naturally unveils the main limitations of the technology, which are enunciated.

2.1. Introduction to Depth Imaging

This section provides a general overview on 3D imaging [404]. Four fundamental systems widely used for 3D imaging [49] are introduced: laser scanners, stereo systems, light coding and ToF cameras. Some basic background on the principles of operation and comparative remarks are provided. Some techniques that, even being previous to ToF imaging, can be classified as a special type of phase-shift-based ToF technique—e. g., interferometric techniques—are considered in Section 2.2.

2.1.1. Laser Scanners

Laser scanners are the preferred option when high accuracy in depth measurements is required. Laser scanners are either based on the Time-of-Flight principle, like ToF cameras, or in the triangulation principle, which is closely related to the concept of stereo vision. Generally, laser scanners gather data sequentially, therefore no 2D-array of distances is acquired at once and their classification as 3D-*imaging* device is questionable.

ToF laser scanners measure the time the emitted light takes to reach the scene point and return to the sensor. Consequently, the accuracy of ToF-based scanners is directly dependent on the precision of their time measurements and is typically constant along the measurement range [61]. The time measurement is often substituted by a measurement of difference of phase between emitted and received light, which is also proportional to the distance traveled.

Triangulation scanners use one or two—typically CCD—cameras to receive the returning signal, i. e., the receiver is no longer a single element but a line or an array of pixels. In the simpler setup, using a single camera, the laser

beam is projected onto the scene with a known direction and is received by the camera, which is placed at a known distance from the emitter. The distance to the object can be easily determined by solving the triangle between emitter, object and reception position of the point or line formed on the camera image. Note that, in this setup, the operation range of the scanner is bounded by the length of the baseline, or distance between emitter and camera. The accuracy of triangulation scanners decreases with the square of the distance between scanner and object [61]. When measuring optically rough surfaces, speckle will be observed in the image. The specle noise limits the uncertainty of the spot localization and, consequently, the uncertainty of the distance measurement [155]. These characteristics make ToF-based scanners suitable for long distances, while triangulation-based scanners are preferred for short range measurements. For a quantitative evaluation on the accuracy and resolution of commercial laser scanners the interested reader is referred to [60].

Laser scanners make use of rotating mirrors to change the direction of the beam. The lateral resolution of laser scanners—number of pixels of the depth *image*—is typically high, much higher than most other 3D-imaging devices, and is governed by the minimum step size of the mirror rotation system. Due to the sequential acquisition, a large number of points slows down the acquisition process. Typical acquisition rates are in the order of thousands to tens of thousands of points per second. Consequently, for very high resolutions, the acquisition time might grow up to several seconds or even minutes. This gives rise to one of the main limitations of laser scanners: the need for a strictly static scene and no external motion of the scanning device. This requirement is often impossible to fulfill in real scenes and, therefore, severe motion artifacts come up when gathering dense data from non-static environments. A way to overcome this limit is using many laser sources and receivers simultaneously in the same scanner (e. g., 64 in the Velodyne lidar [224]), leading to bulky and expensive systems that can reach the frame rate of ToF cameras.

Typically, laser scanners gather the measurements following sequential raster lines, i. e., a regular acquisition pattern. In applications such as 3D modeling this may lead to an insufficient number of points in critical areas, where high detail is required, or too many points in unimportant areas. Using custom acquisition patterns, which can be adapted to the object to sense, are an appropriate solution in such cases. The laser scanner presented in [49] uses Lissajous curves as acquisition patterns. They also develop a recursive optimization method that allows obtaining high-resolution 3D reconstructions from the sparse range data.

2.1.2. Stereo Systems

One of the most widely-used methods for depth imaging is stereo vision [46]. The hardware normally consists on two cameras sensing the scene from two different and typically known positions. Using two images simultaneously taken from both cameras, it is possible to extract 3D information of the scene. Note that stereo vision is, in its basic concept, a passive method, since no energy is emitted and no camera movement is needed. Conventional stereo systems often use two cameras of the same model, pointing in the same direction but displaced horizontally a certain distance (baseline) from one another, similarly to the human binocular vision system. In such case, and in absence of any distortion, corresponding pixels between the two images lie on the same horizontal row and a disparity map can be directly computed from the positions of corresponding pixels in the images. The disparity is inversely proportional to the distance from the system to the object, therefore, a depth value can be computed from the disparity if the camera parameters are known. In most practical cases, there exist lens distortions, that might be different for each camera, and the cameras might not be perfectly aligned. Consequently, images must first be cleaned from distortions and later both images of the stereo pair must be projected to a common plane (rectification process), in order to ensure that corresponding pixels are actually in the same row in both images.

In general terms, the depth estimation is more accurate for objects close to the stereo system (higher disparity) than for far objects (disparities tending to zero). Nevertheless, for very close objects we might encounter overlapping problems if the object is being observed mostly by only one camera of the stereo pair. From the description of the principle of operation it is easy to observe one major problem of stereo depth estimation, the so-called parallax problem. For large parallaxes (cameras placed at largely different viewpoints) an accurate depth estimation can easily be computed, even for objects that are relatively far from the system. A large baseline implies a reduction of the SNR in depth estimation. On the other hand, the matching problem becomes more difficult due to the occlusion of features. Another critical point of this technique is the search for correspondences. Depending on its location in 3D-space, an object point might be only present in one camera view, in which case no match is possible. Additionally, in many real applications, the system faces areas without texture—e. g., long corridors without distinctive elements on the walls—, where useful correspondences cannot be established. The projection of structured light [331, 397], known as active stereo, is a common solution in such situations. The optimal design of the patterns to project has

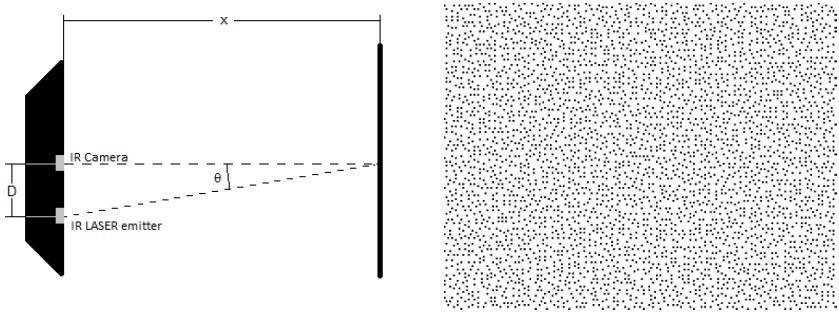
been studied in [227]. Other active stereo solutions include camera motion [126, 442], allowing fixation in different scene objects and accurate surface reconstruction. Passive stereo systems usually cannot provide a dense depth image due to lack of correspondences. Coarse-to-fine approaches and certain hypothesis on depth smoothness can partially solve the problem but cannot totally overcome the effect of occlusions.

2.1.3. Light Coding Technology: The Kinect Sensor

The so-called *light coding* technology evolves from the structured light range imaging systems [177]. In general terms, structured light systems project a light pattern on the scene and observe it through a camera. The different depths of the scene points induce different distortions in the observed projection and the depths can be estimated if univocal correspondences between the emitted and observed patterns are established. If several cameras are used, the pattern can be used to ease the search for correspondences between the different camera images and the depth can be estimated from disparity as in a stereo system. In early approaches [376] several masks featuring different spatial codes are used to generate different projection patterns. One of the most common options is to use a set of Gray codes [453], which are robust and can be efficiently generated [48]. Other binary codes can also be used as projection patterns [319]. Later approaches tend to rely on a single acquisition [379] and the code design becomes a critical issue. Using projection of color codes [97, 482] in combination with a color camera allows reducing the acquisition time or, conversely, acquiring more information, resulting in a better depth estimation.

The light coding method we refer from now on is that used in the first version of the widely-known Kinect sensor, developed by Microsoft. The hardware of the Kinect sensor is based on the range camera technology provided by PrimeSense. Focusing in the depth estimation hardware, the sensor features an infrared (IR) laser projector and an IR camera, pointing in the same frontal direction and separated a few centimeters from each other in horizontal direction. The existence of an emitter and a receiver could induce to think about a ToF-based system but the emitted light is DC and no time or phase measurements are performed. Instead of modulating in time domain, the light is modulated in spatial domain, according to a fixed pattern. The pattern, which looks like small dots randomly distributed, is projected onto the scene. The pattern is composed by a sub-pattern, which is repeated 3×3 times. Fig. 2.1b shows a schema of the sub-pattern, where each bright dot in the pattern is represented by a black square. The

reflection of the pattern is captured by the IR camera. The relative position of the dots in the infrared image, with respect to a reference image of the pattern at a known distance, depends on the distance between the sensor and the object that reflected the dot. The full description of the system is given in [188], which seems to be an implementation of the more general system presented in [418], where it is explicitly stated that groups of spots in the IR image are compared to those in the reference image, using image correlation or some other image matching technique. Fig. 2.1a is a diagram of the Kinect sensor observing a wall. It has been shown that the raw measurements delivered by the sensor depend linearly on the angle θ [447] and the actual depth, conceived as the distance from the imaginary image plane to the object, can be calculated through a tangent transformation.



(a) Kinect Operation **(b)** Kinect Pattern

Figure 2.1.: Schema of the Kinect sensor (a). A pattern of dots is projected onto the scene and the reflection is captured by an infrared camera. The disparity of the acquired IR image with respect to a reference image of the pattern depends on the angle θ and this, in turn, on the distance to measure, x. (b): the basic sub-pattern used to generate the pattern, from [386]. Black squares represent the bright spots. The pattern is obtained by repeating this sub-pattern 3×3 times.

The depth images are delivered as monochrome, with 11-bit VGA resolution (640×480 pixels), streamed at 30 Hz frame rate. The value of the sensor response is calculated internally from the disparity between the reflected pattern recorded by the IR camera and a reference pattern, which is stored within the sensor [260]. Objects whose distance to the sensor is higher or lower than the default distance used to obtain the reference pattern will

reflect the pattern points displaced in the direction of the baseline between the laser emitter and the IR camera, with respect to their original position in the reference pattern. Using simple image correlation between the captured pattern and the reference pattern the disparity image can be calculated. This means establishing correspondences between two images, for what multiple pixels are used. As a consequence, the real lateral resolution of the depth images is lower than that given by the number of pixels [282] (118 pixels in width, instead of 640). A theoretical limit of the lateral resolution is given by the size of the pattern, which is composed by 220×170 bright dots. Refined formulas to compute the depth value from raw data are given in [282].

According to the data provided by PrimeSense, the field of view (FOV) of the cameras is 58° horizontal, 45° vertical and 70° diagonal. The spatial x/y resolution is 3 mm and the depth resolution is 1 cm, both at 2 m distance from the sensor [447]. PrimeSense gives an operation range for the Kinect sensor between 0.8 m and 3.5 m, but the real range depends on the required precision. Ranges up to 5 m [447] are found in some works, but it has to be taken into account that the depth resolution is proportional to the square of the depth. The Kinect also implements an internal lower limit in depth of 0.4 m, probably to avoid a severely poor overlap with the RGB image. Known limitations of the first version of the Kinect are the inability of operating under sunlight illumination and the absence of depth estimation for dull objects, such as computer screens.

The Kinect sensor has often been used as an inexpensive substitute of laser scanners in mobile robotic applications. A comparison between the first version of the Kinect and laser scanners is given in [488]. One of the many examples of this trend is found in [233], where a Kinect sensor is mounted on a quadcopter and used to perform visual odometry and mapping. A considerable amount of research has been oriented to adapt the Simultaneous Localization And Mapping (SLAM) algorithms used in mobile robotics to the data provided by this new sensor, giving birth to the so-called RGB-D SLAM [171, 168, 143], which was first introduced in [223]. Not only the area of robot navigation has profited from this sensor. Often the interest is moved towards scene reconstruction [349, 239, 257, 167, 290], in which case the crucial point is how to fuse redundant measurements to reduce noise and increase reconstruction accuracy.

2.1.4. Time-of-Flight Cameras

Time-of-Flight (ToF) cameras [179, 282] can be seen as an extension of ToF laser scanners, where the beam has been substituted by a dense light

projection over the scene and the receiver is a planar array of pixels. As in the case of laser scanners, the principle of operation can be the measure of the time a light signal needs to travel from the camera to the object and back to the camera, or the measure of the phase shift between emitted and received signal, using some high-frequency periodic waveform. An image of the first phase-shift-based ToF camera [279] is provided in Fig. 2.2a.The main advantage of ToF cameras over laser scanners is the absence of scanning motion: all the points are acquired simultaneously. This softens the static scene restriction of laser scanners, since only movements that are fast enough to produce changes in the image during the exposure time might lead to motion artifacts in the depth map. On the other hand, the lateral resolution of commercial ToF cameras is generally low (ranging from 64×48 to 640×480 pixels), significantly lower than that achievable with laser scanners. Depth resolution strongly depends on the estimation method and the hardware capabilities and can reach subcentimetric theoretical values. Nevertheless, the depth accuracy, strongly influenced by environmental conditions, is typically of several centimeters (one to five for most commercial cameras).

When compared to stereo systems [425], ToF cameras offer several advantages. One of them is that they perform direct depth—or phase—data acquisition, using methods such as the pixel-level correlation in PMD chips (Section 2.3), instead of applying heavy algorithms on pairs of conventional intensity images. Additionally, ToF cameras can operate in textureless scenes, where stereo approaches typically fail. Another strong point of ToF imaging is that the depth image is generally dense, while stereo approaches can only offer a depth value where a pair of correspondent pixels are found. As a drawback, ToF cameras are active systems, that is, an illumination source is needed, typically IR light modulated at several megahertz frequency. Consequently, interference phenomena might come up when several ToF cameras operate simultaneously. The signal-to-background ratio, i. e., effective dynamic range, may be dramatically decreased when a source of IR light is present in the scene or when operating under sunlight illumination. For phase-shift-based ToF cameras, the modulation frequency of the illumination signal establishes a tradeoff between depth resolution and unambiguous range: high frequencies are needed to obtain good depth resolution, but the higher the frequency, the shorter the unambiguous range of the camera.

A widely-spread technology for phase-based ToF imaging is the so-called Photonic Mixer Device (PMD), developed by PMD Technologies. At the time the author started to conduct the research presented in this thesis commercial PMD sensors offered 19k (120×160) and 41k (200×200) resolutions. The manufacturer also offered complete ToF cameras based on these chips, e. g.,

the PMD CamCube (41k). Since 2015 the production of PMD sensors has been externalized to Infineon, whose sensor family REAL3^{TM} [385] offers ToF imagers based on PMD pixels. The sensitivity has been improved by means of microlenses. Differently from the previous generation of PMD chips, the control signals are internally generated in the chip, as well as the ADCs. This allows for higher modulation frequencies (up to 100 MHz) and lower latency, at the cost of reducing the flexibility of the system. Additionally, two higher resolutions are added to the 19k standard: 38k (224×172) and 100k (352×288). In fact, the recent PMD CamBoard pico flexx reference design already supports the 38k chip. Due to the importance of PMD technology both for ToF imaging in general and for this work in particular, it is analyzed independently in Section 2.3. An image of a state-of-the-art PMD camera is provided in Fig. 2.2b. Another ToF sensor that also requires special mention is the new Kinect sensor (the so-called second-generation Kinect, or Kinect v2 or Xbox One sensor [471]), which is depicted in Fig. 2.2c. This camera is—as well as the first Kinect—an RGB-D camera, delivering both RGB and depth images. It offers a depth lateral resolution of 512×424, very high for a ToF device. Problems derived from sunlight illumination and illumination interference with other ToF cameras have been solved using a complex modulation signal, whose frequencies range from 10 to 130 MHz. The demodulation contrast (68% at 50 MHz) is also high when compared to state-of-the-art technology (Section 2.3). An introduction to the new Kinect technology is presented in [362] and further detailed in [22], where numerical values of key parameters are given (Table 1).

In order to provide a visual overview of the performance of some well-known commercial solutions for depth imaging, we extend the depth camera comparison in [282] by including the recent Intel-Creative®[1] ToF-based RGB-D camera and the new Kinect sensor. The other sensors in the comparison are the first generation Kinect sensor, the PMD CamCube, featuring a 41k PMD chip, a PMDTec camera featuring an earlier 3k chip, two ZESS MultiCams featuring 3k and 41k PMD chips, and the Softkinetic sensor. In the case of the Softkinetic sensor, different light powers (denoted in the legend as percentage of the maximum illumination power) and a special *close* mode (probably using a higher modulation frequency) are tested. Fig. 2.3a shows the depth measurements provided by each one of the cameras we consider after applying a linear correction. Fig. 2.3b depicts the error with respect to the ground truth for all the linearly corrected measurements

[1] Intel® , Creative® and SwissRanger® are registered trademarks and will be denoted in the following without the ® symbol for notation simplicity.

(a) Lange's Camera **(b)** CamBoard Pico S **(c)** Kinect v2

Figure 2.2.: Evolution of phase-shift-based time-of-flight imaging. From left to right: The first ToF camera (a), presented in [279] and based on *intelligent* pixels, similar to those of the current PMD cameras (image taken from [278]). The PMD CamBoard pico S (b), the latest PMD 3D camera reference design. The second-generation Kinect sensor (c) (image from [471]), commercialized together with the Xbox One gaming console, provides state-of-the-art depth sensing at a very low cost. The basic principle of operation of these three cameras is the *in pixel* cross-correlation.

shown in Fig. 2.3a. Both graphs are in meters and 100 measurements were averaged per position. The two variants of the Kinect v1 (simple and tangent) make use of the two different formulas for depth calculus given in [282], respectively. Note that the Intel-Creative camera operates accurately at its theoretical maximum range (99 *cm*) but cannot sense over 2 m. The new Kinect sensor obtains surprisingly good results, offering a very low depth error along the whole range of operation. According to our results, the sensor delivers high accuracy depth measurements between 0.56 and 4.44 m. The average depth error within this range is 4.8 mm, one order of magnitude lower than most competitors, including the first Kinect version. Additionally, it was observed that the new Kinect provides no measurements out of its optimal range (0.50 to 4.5 m). As it is easy to deduce from Fig. 2.3b, where the Kinect v2 errors are the lower bound of all the other sensors considered, this new sensor is a serious competitor for any 3D-imaging device. Standard deviation measurements reveal that the Kinect v2 is the most stable sensor among all considered here, with an average standard deviation of 2.2 mm along its operative range. A comparison between a ToF camera (the SwissRanger®[1] SR4000, from MESA Imaging®) and the structured light Kinect sensor is provided in [213]. For a thorough

performance comparison between the first and second versions of the Kinect sensor, the interested reader is referred to [396].

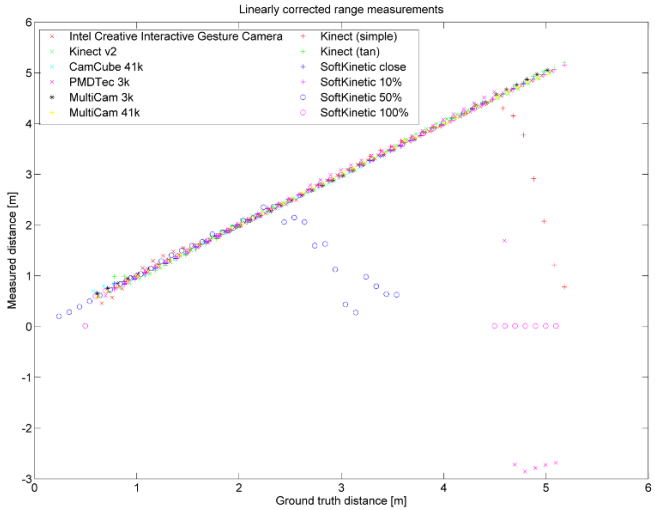
After this expanded comparison was completed, new ToF sensors from different manufacturers arrived to the market, which are, consequently, not analyzed here. The author is aware of the existence of new models of ToF depth sensors from Panasonic® and Texas Instruments®[2] .

Despite the D-IMager[3] ToF cameras from Panasonic [1] had a resolution of 120×160 pixels, i. e., equivalent to a PMD or Swiss Ranger SR-2 camera, for instance, the newest Panasonic ToF chips reach VGA resolution, namely, 640×480. The depth accuracy of the D-IMager cameras was 3 cm in absence of ambient light and as poor as 14 cm with ambient light, both for the models EKL3104 (test with 2×10^4 lx ambient illumination) and EKL3106 (test with 10^5 lx ambient illumination). Interestingly, for the model EKL3105 Panasonic reported improved accuracies of 2 cm and 5 cm, without and with ambient light (2×10^4 lx), respectively. An example of ToF depth camera featuring the recent VGA Panasonic chip MN34902BL is the Basler ToF camera (first model, named 6m) [32], presented in mid-2015. The peculiarity of the new generation of Panasonic ToF chips (series MN349XXXX) is that they are not just a NIR ToF sensor, but also a color sensor, that is, they simultaneously provide a depth image and a color image of the scene, pixelwise registered. The manufacturer of the camera claims an accuracy of ± 1 cm in a range of 0.5 to 5 m in ideal conditions, but, obviously, the real value will depend, at least, on the depth being measured. In any case, this accuracy is superior to that of the earlier D-IMager cameras. Also, according to the number of pixels in the sensor, the Basler ToF camera might be state of the art in terms of lateral resolution. Additionally, the manufacturer plans three other models implementing the Panasonic ToF chips MN34903TL, MN34922BL and MN34923TL, which are to appear under the model names 6c, 13m and 13c, respectively. We just want to point the reader's attention towards the future model 13m, implementing the Panasonic MN34922BL chip, which would deliver color and depth images, both of SXGA resolution (1280×1024), with a single image sensor (see [31]).

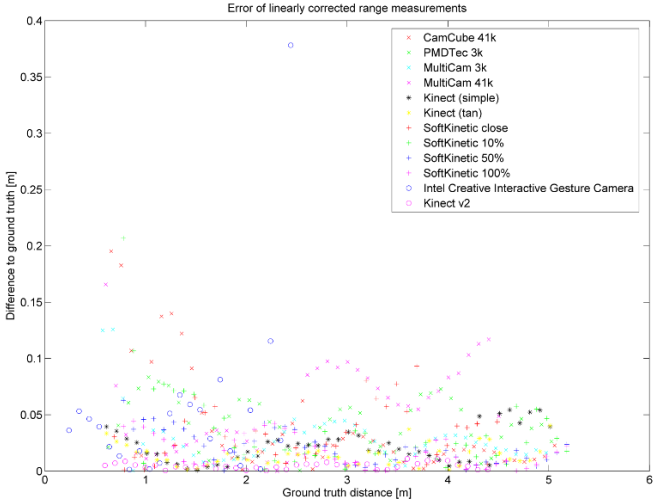
Texas Instruments offers the OPT8320 and the OPT8241 [358] ToF sensors. The first is a complete system-on-chip (SoC) that delivers depth images directly, without the need of any external circuitry. This SoC even integrates an illumination driver. Unfortunately, the resolution is as low as 80×60 pixels, i. e., QQQVGA. The second sensor is a ToF sensor that only integrates

[2] Panasonic® and Texas Instruments® are registered trademarks and will be denoted in the following without the ® symbol for notation simplicity.

[3] The production of the Panasonic D-IMager cameras was discontinued in April 1^{st} 2015



(a) Corrected Depth



(b) Depth Error

Figure 2.3.: Linearly corrected depth measurements averaged over 10 frames per position (a) and absolute error with respect to the ground truth (b). ©2015 IEEE.

the timing generator and the ADCs on chip, but offers a superior resolution of 320×240 pixels, i. e., QVGA.

2.2. Phase-Shift-Based Time-of-Flight Imaging Systems

As indicated in the previous section, a ToF imaging system, *sensu stricto* measures the time the light needs to travel from the illumination system to the scene and from the scene to the camera. Unfortunately, a depth sensor based on this principle requires picosecond resolution in the time measurements to achieve millimeter resolution in depth. Building an array of such sensors operating simultaneously as an imaging system is challenging even for low resolutions. The pulse counting ToF system of [19] is an example of such a pure ToF imaging sensor, but a crucial limitation is imposed by the maximum frequency of the clocking signal, which is suggested to be of 2 GHz frequency, leading to only 500 ps ToF resolution. This is the reason why most ToF cameras measure phase shift instead of time.

In order to measure phase shift, a periodic signal is required. This brings the concept of unambiguous range, which is the maximum distance that can be measured without ambiguity and is determined by the frequency of the signal, f. Provided that the light has to cover the depth twice, the depth d can be computed from the phase shift θ_{depth} as

$$d = \frac{c}{4\pi f}\theta_{\text{depth}} \tag{2.1}$$

and, in the simplest case, if only one frequency is considered, the unambiguous range is

$$d_u = \frac{c}{2f} \tag{2.2}$$

At this point we make a distinction between pure *optical interferometry* and what was originally called *optical radio frequency interferometry* (ORFI) [474]. These two techniques are often presented separately, being the first one named simply *interferometry* and the second one *phase-shift-based* ToF or simply *ToF*. The truth is that both techniques share the same principle of operation and should be presented together, since the ORFI mimics the optical interferometry, aiming to obtain a much larger unambiguous range by means of an intensity modulation of the light at radio frequencies. Supposing a periodic modulation scheme, in both cases the depth can only be estimated

from the phase shift and, therefore, Eq. 2.1 and Eq. 2.2 hold, being $f = f_{\text{mod}}$ the modulation frequency in the ORFI case and the actual electromagnetic frequency of the light in the optical interferometry case.

2.2.1. Interferometry

The human eye is typically sensitive to light of wavelengths between 390 and 700 nm [424], which correspond to a frequency band between 430 and 790 THz. Even light of relatively large wavelength, such as near infrared light (NIR), leads to frequencies in the range of hundreds of terahertz. For instance, a wavelength of 850 nm is equivalent to a frequency of 350 THz, which leads to an unambiguous range of 425 nm, i. e., half the wavelength.

Interferometry [214, 215] is based on the superposition of light waves (interference), which can be constructive or destructive. For instance, if at a certain point in space, which we denote by the vector $\vec{r} = [x, y, z]$, two electromagnetic waves of equal amplitude and frequency interfere, the amplitude of the resulting wave may be higher or lower than that of the interfering waves, depending on their relative phase shift. In general, consider two plane light waves, E_1, E_2, traveling along the same direction, say z, and polarized with their field vectors in the same plane:

$$\begin{aligned} E_1(\vec{r},t) &= E_1(z,t) = e_1 \cos\left(2\pi f t - \frac{2\pi}{\lambda} z\right) \\ &= \Re[e_1 \exp(-i\theta) \exp(i\omega t)] = \Re[e'_1 \exp(i\omega t)] \\ E_2(\vec{r},t) &= E_2(z,t) = e_2 \cos\left(2\pi f t - \frac{2\pi}{\lambda} z + \Delta\theta\right) \\ &= \Re[e_2 \exp(-i(\theta - \Delta\theta)) \exp(i\omega t)] = \Re[e'_2 \exp(i\omega t)] \end{aligned} \tag{2.3}$$

where e_k is the amplitude of the wave $E_k, k \in \{1, 2\}$, f and λ are the frequency and the wavelength of the radiation, respectively, and $\Delta\theta$ is the relative phase shift between waves. In the complex formulation, the following changes of variables was applied: $\theta = 2\pi z/\lambda$, $\omega = 2\pi f$. In the last expression, e'_k is the complex amplitude, dependent on the phase shift. The intensity of a wave is defined as $I_k = |e'_k|^2$ and, consequently, the resultant intensity is

$$\begin{aligned} I = |e'|^2 &= (e'_1 + e'_2)(e'^*_1 + e'^*_2) \\ &= |e'_1|^2 + |e'_2|^2 + e'_1 e'^*_2 + e'^*_1 e'_2 \\ &= I_1 + I_2 + 2\sqrt{I_1 I_2} \cos\Delta\theta \end{aligned} \tag{2.4}$$

Note that Eq. 2.4 is equivalent to the cross-correlation function between two sinusoidal signals, as far as they both result in another sinusoidal function, which depends on a relative phase shift. If the two waves share the same amplitude, the resultant intensity would vary from zero to twice that of the interfering waves, depending exclusively on the phase shift.

An interferometer is a system that captures the interference pattern, to retrieve the phase shift between the interfering waves. If both waves share the same frequency and their intensities are known, the phase shift can be calculated directly from Eq. 2.4. If there are additional unknowns, e. g., the light intensities I_1, I_2 or an additional DC offset due to ambient light I_0, more constraints are required to estimate the phase shift. If it is possible to delay one of the waves with respect to the other, custom delays can be superimposed and different resultant intensities will be obtained. Since Eq. 2.4 is a sinusoidal function, at maximum three parameters are to be estimated: the DC intensity offset $(I_0 + I_1 + I_2)$, the amplitude of the sinusoid $(I_1 I_2)$ and the phase shift $\Delta\theta$. Consequently, three acquisitions using three different custom shifts suffice to estimate the phase shift. Obviously, one can profit from gathering more than three acquisitions if noise in the measurements is to be expected. In such case, solving the overdetermined system derived from the multiple acquisitions may help to minimize the error produced by the measurement noise. The optimal solution, in the least squares sense is given by a formula of the form of Eq. 2.5 [314], which is based on the general least squares separation procedure of [206].

$$\Delta\theta = \arctan\left(-\frac{\sum_{k=1}^{N} I_k \left(A_{1,1} + A_{1,2}\cos\theta_k + A_{1,3}\sin\theta_k\right)}{\sum_{k=1}^{N} I_k \left(A_{2,1} + A_{2,2}\cos\theta_k + A_{2,3}\sin\theta_k\right)}\right) \tag{2.5}$$

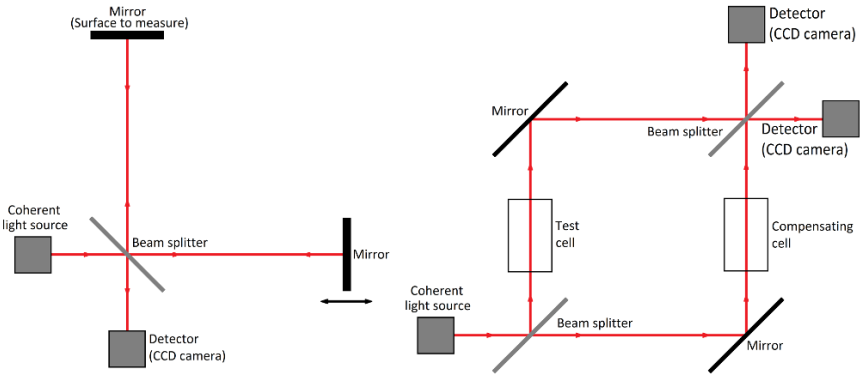
where $I_k, k \in [1, N]$ denotes the k^{th} measurement, gathered at the phase shift θ_k. The coefficients $A_{i,j}$ are constants which depend only on the sampling scheme, i. e., which θ_k are considered. For the definition of these constants, the reader is referred to Eq. 5.6-5.9 of [314]. Clearly, in this and later formulas implying the arctangent function, the specific sign of both numerator and denominator is to be taken into account in order to resolve the correct quadrant. If the measurements are taken at equally spaced intervals, Eq. 2.5 further simplifies to Eq. 2.6 [332], which is the so-called *synchronous detection algorithm* or *diagonal least-squares algorithm* [314].

$$\Delta\theta = \arctan\left(-\frac{\sum_{k=1}^{N} I_k \sin\theta_k}{\sum_{k=1}^{N} I_k \cos\theta_k}\right) \tag{2.6}$$

Before diving into implementation details, which are obviously wavelength-dependent, it is worth having a last look at Eq. 2.6, this time considering the minimalistic case in which two orthogonal phases are acquired, say $\theta_1 = 0°$ and $\theta_2 = 90°$. Then, provided that $\sin\theta_1 = \cos\theta_2 = 0$ and $\cos\theta_1 = \sin\theta_2 = 1$, the argument of the arctan boils down to the quotient $-I_2/I_1$, which is totally expectable, since this is nothing else than calculating the phase of a complex number, given its real and complex components. This special case of two-phases interferometry is the classical coherent sensing implemented in, e. g., radar systems, yielding the so-called *in-phase* and *quadrature* components, which fully define both the amplitude and the phase of the received echo. Differently from optical interferometry, radar systems operate in radio frequencies, at which silicon is insensitive and, therefore the arrays of detectors are not miniaturized dense arrays of pixels, but bulky arrays of very few antennas. This induces to think that radar technology cannot produce high-resolution depth images and is relegated to target detection and coarse ranging, thus unrelated to the core technology in this work. Nevertheless, if the antenna(s) are mounted on a moving vehicle, e. g., an aircraft, the motion can be used to synthetically increase the aperture, without actually increasing the antenna size. This is the basic idea behind Synthetic Aperture Radar (SAR), a technique that is able to produce high resolution images from comparatively small antennas. Furthermore, SAR interferometry (often shortened IFSAR) [23] makes use of two coordinated vehicles (aircrafts or spacecrafts), or, alternatively, two passes of a single vehicle over the scene to image, in order to achieve outstandingly large baselines, e. g., between 100 and 300 m. Relatively short wavelengths can be used, e. g., $\lambda = 3$ cm for the X band and very fine range resolution can be attained, e. g., around 1 m. Despite the phase images are extremely noisy, the difference of phase between the interferometric pair accurately retains the information on the scene structure. IFSAR yields high resolution interferometric images with height accuracy in the order of 1 to 10 m.

The different (optical) interferometer setups can be classified in two main groups, regarding whether they use two or a single optical path. Classic interferometers, such as the Michelson and the Mach-Zehnder interferometer

belong to the double path group. One example of common path interferometer is the Sagnac interferometer. Conventional fiber optic gyroscopes are not more than a Sagnac interferometer where the optical circuit is confined in an optical fiber. For completeness, we provide simplified schematics of the Michelson and Mach-Zehnder interferometers in Fig. 2.4. For a complete description of these and other interferometers, the reader is referred to [215].



(a) Michelson Interferometer **(b)** Mach-Zehnder Interferometer

Figure 2.4.: Michelson (a) and Mach-Zehnder (b) interferometers. In the Michelson interferometer the light traverses the light paths twice, before and after reflecting on each mirror. The Mach-Zehnder interferometer offers a more flexible setup, with separated light paths for the reference and the probe beams. Consequently, in a Mach-Zehnder interferometer each optical path is traversed only once.

One could say that the simple setup of the Michelson interferometer is adequate for measurements in reflective mode, where the sample to measure acts as one of the mirrors. In such setup, if the addition of custom phase shifts was required to retrieve the actual phase shift induced by the sample, they could be generated, e. g., by precise displacement of the mirror of the reference beam in the direction of the beam. A piezoelectric transducer (PTZ) might be used to achieve the required precision. The Mach-Zehnder interferometer is the setup of choice for working in transmissive mode. There are two independent paths, one for the reference beam and the other for the probe beam. The reference beam traverses the so-called *compensation cell*, while the other traverses the *test cell*. The function of the compensation

cell is just to mimic the properties of the test cell and assure that the final measurements are representative of the sample, exclusively. In general, regardless of the interferometer setup, if a 2D sample or scene is to be sensed, lenses are to be placed after the laser source, in order to open the beam, and before the sensing array, in order to form the image on it.

Until now we implicitly sticked to the homodyne case, i. e., the interference occurs between light of the same wavelength, resulting in light of the same wavelength whose intensity depends on the relative phase shift of the interfering radiations. In the heterodyne case, the interfering signals are of different frequencies and the result of the interference is formed by two periodic signals of frequencies equal to the difference and the sum of those of the interfering signals. In most cases, only one of those frequencies is measured, typically the lowest (beat frequency), and the other is filtered out.

Despite the good precision that can be achieved measuring depth with interferometric techniques (two to three orders of magnitude lower than the wavelength, i. e., in the nanometer range), the short unambiguous range of half a wavelength limits their fields of application. The naïve solution to this problem is to use several frequencies instead of only one [303]. This method is often called multi-wavelength interferometry (MFI) and the unambiguous range is one half of the least common multiple of the used wavelengths, i. e., half the lowest value that is simultaneously a multiple of all the wavelengths considered. The *half* is due to operation in reflective mode, as in ToF imaging, but is to be omitted in transmissive mode. Clearly, wavelengths that are multiples from each other are to be avoided, since they bring no advantage over using simply the highest. A common approach is the use of closely spaced wavelengths. A study on frequency selection and a method for selecting the optimal set of frequencies in MFI is given in [457]. Using closely spaced wavelengths leads to an unambiguous range equivalent to that that would be obtained using radiation of a frequency equal to the beat frequency, i. e., the low frequency given by the difference of frequencies of the emitted radiations. This technique often called Synthetic Wavelength Interferometry (SWI). A recent work [458] has shown that micrometric resolution with an unambiguous range of more than one meter can be achieved in an SWI setup using four closely spaced wavelengths.

Other methods exist that solve the ambiguity problem. The so-called electronic speckle pattern interferometry (ESPI) [415, 240], which was first introduced in [289, 65], is based on the fact that the interference pattern generated between a reference and a speckle wave contains the phase information of the speckle wave. It differs from holography in that there is

no need for reconstructing the 3D complex amplitude of the wave. Note that the autocorrelation function of the complex amplitude and intensity of a speckle wave is close to a Dirac delta function, due to the fact that displacements of the diffusive surface lead to rapid random variations of the speckle pattern and any random pattern is, by definition, poorly correlated with another random pattern. One of the reasons of including this short note on ESPI is to give the reader the opportunity of confronting it to recent methods based on the projection of pseudorandom patterns, such as the depth-varying pattern used in [418] for depth sensing.

A natural extension of multiple wavelength interferometry is to use a band of frequencies, instead of a finite number. If the visible spectrum is chosen, the technique is called *white light* interferometry. A review of white light interferometry is given in [469], where the possible setups are classified into three categories: diffraction grating interferometers, vertical scanning or coherence probe interferometers and white light scatterplate interferometers. From these, the second one is of special interest for being the preferred option in most industrial applications requiring surface metrology. In its simpler setup, a vertical scanning white light interferometer may adopt the form of a Michelson interferometer (Fig. 2.4a), where the mirror is located at the same distance than the probe from the beam splitter and can be precisely displaced in the beam direction, typically obtaining equally-spaced samples of the white light correlogram. A maximum in the intensity of the interference signal will be achieved when the optical paths of the reference and probe beams have equal length (constructive interference). As in the original Michelson setup, the maximum depth accuracy is bounded by the positioning accuracy of the positioning stage. If further magnification is required, a Mirau interferometer is used instead (10× to 50×). For even higher magnification (more than 50×), a Linnik interferometer, which includes additional optics also in the reference path, is the appropriate setup.

2.2.2. Optical Radio Frequency Interferometry or Modulation Interferometry

White light interferometry solves the problem of the very short unambiguous range of single wavelength optical interferometry, often in the range of hundreds of nanometers. Nevertheless, there remain some issues that restrict optical interferometry to metrology in high-end industrial applications. First, one or several sources of coherent light, providing appropriate coherence length, are required. Laser sources and, specially, the control electronics, may be too expensive for some applications. Furthermore, an interferometer

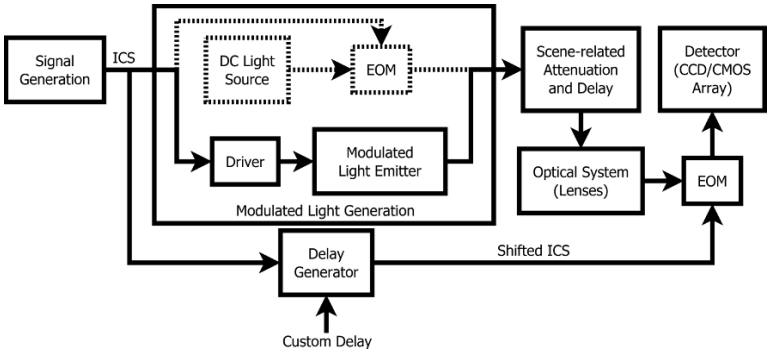
setup is required, which demands a very stable structure, isolated from any vibration or external disturbance and machined with fine tolerances. The same standards of quality and care apply to all elements of the optical path—lenses, mirrors, beam splitters, etc.—since even submicrometer irregularities affect the resulting interference pattern. In general, interferometers cannot be presented as a competitive solution to the rest of depth imaging techniques, designed to sense scenes with depths in the meter range or more. This is not only because of physical limitations of the measurement principle, but also because of the cost and the delicateness of the equipment they require.

Using larger wavelengths is an immediate way of increasing the unambiguous range. Radio wavelengths in the range of tens to hundreds of megahertz lead to unambiguous ranges of several meters. The Ultra High Frequency band (UHF) is also known as *decimeter* band, since the wavelength range goes from 1 dm to 1 m. Radiation in such band would be optimal for close range depth imaging in terms of unambiguous range. Another well-known frequency band for television broadcasting is the Very High Frequency (VHF), which includes wavelengths from 1 to 10 m, well-suited for medium range unambiguous depth imaging. Unfortunately, radiation of such *lower* frequencies does not produce photoelectric effect in silicon and, therefore, cannot be detected with a conventional CCD or CMOS camera. The longest wavelength for which photoelectric effect occurs is around 1.1 μm, i. e., still in the near infrared (NIR) range. Using radio frequencies would require building an array of antennas as detectors and would no longer be any kind of ToF *imaging* technique, but a radar system. In this context, Optical Radio Frequency Interferometry (ORFI) arises as a technique that aims to bring together the benefits of using visible or NIR light, i. e., photoelectric effect in silicon, and the large unambiguous range associated with radio frequencies. ORFI also avoids the need for coherent light sources and an actual interferometer.

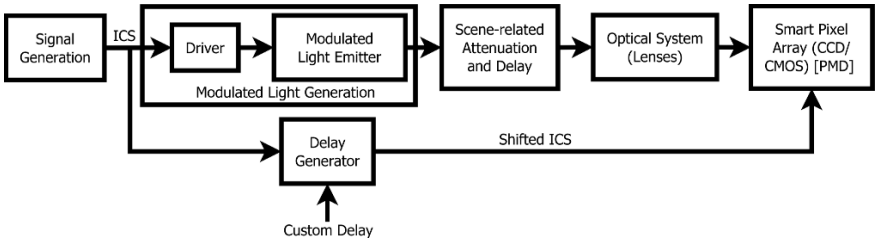
The basic idea behind ORFI is to modulate the intensity of the light which is projected onto the scene to sense. The light, normally incoherent, is intensity modulated in continuous wave mode (IM-CW), in the simplest case according to a periodic signal of the desired radio frequency. If such a signal is split in two beams and one of them is used as reference beam and the other as probe beam in an interferometer setup, no interference will take place and the false "interference" image formed in the camera would be only determined by the different reflectivity of the scene points, but independent from their depths. A *demodulation* step, coherent with the modulation scheme, is required to mimic optical interference. This can be done using an electro-optic modulator (EOM), which is an optical device able to modulate

an input beam according to an electrical signal. The same signal used to control the illumination may be used to control the EOM for demodulation of the light reflected by the scene, before focusing it on the camera chip. If the light source is DC, two EOMs can be used, one for modulation before projection to the scene and another for demodulation. For clarity, the block diagram of a general ORFI setup is given in Fig. 2.5a. In the diagram the generation of the modulated light is obtained either using an EOM over DC light or directly by driving a fast light emitter, e. g., fast LEDs or laser diodes. The latter option is preferred, since it is more efficient, compact and cheaper than the former. There exists a variety of devices that can be used as EOMs, which will be briefly discussed in the following. As well as in the light modulation block it is preferred to perform the modulation at the light generation and not after, in the demodulation part it is of interest to perform the demodulation within the detector, i. e., when the carriers are being generated by photoelectric effect, and not before. This would reduce energy losses, allowing for a reduction of the optical power of the illumination system. The bulkiness and cost of the system would also be reduced, by eliminating the EOM. For a complete comparative analysis between this setup and that using an external EOM the reader is referred to [474]. Such a system requires the use of an array of *smart* pixels, able to perform the mixing of the incoming light signal with some reference signal during the integration process. A block diagram adopting this internal demodulation method is provided in Fig. 2.5b, where the light is modulated by driving a fast emitter, as in commercial ToF cameras, and the demodulation is performed *in pixel*, by means of *smart* pixels, such as those introduced by PMD Technologies (Section 2.3). In the following, the mixing of incident light signal and reference is assumed to be perfect. In case of non-ideal mixing due to, e. g., pixel imperfections, its effect is implicitly modeled in the shape of the *effective* reference signal acting at pixel level, which will differ from the ideal one.

In both cases presented in Fig. 2.5, the reference signal used for demodulation, i. e., the signal the reflected illumination signal is mixed with, is a shifted version of the so-called illumination control signal (ICS). In other words, for each pixel a sample of the cross-correlation between the two signals is obtained at the end of the integration process, at a phase shift that depends exclusively on the phase shift induced by the depth θ_{depth} and the custom phase shift θ_k, which is different for each one of the $k \in [1, N]$ acquisitions. If the waveform of the light signal, as emitted, coincides with that of the reference signal used for demodulation, the problem of phase estimation is reduced to finding the unknown relative phase shift from samples



(a) External Demodulation



(b) PMD Demodulation

Figure 2.5.: Two possible setups for Optical Radio Frequency Interferometry (ORFI). A classic setup (a), where the demodulation is carried out by an external electro-optic modulator (EOM). In (b) the demodulation is performed internally in the pixel array, by means of *smart* pixels, able to control the integration process according to an electric signal. An example of such pixels, able to sample the cross-correlation between the modulated light arriving to the pixel and some reference signal, are the PMD pixels.

of the autocorrelation function at custom phase shifts. For simplicity and analogy with pure optical interferometry, let the emitted light signal and the reference signal be a sinusoidal signal

$$i(t) = A_0 + A\cos(\omega t - \theta_0) \tag{2.7}$$

where A_0 is some DC offset, A is the amplitude of the signal and θ_0 is the initial offset, which can be assumed zero without loss of generality. Now suppose that each point of the scene being illuminated by such signal reflects it with a certain attenuation a and produces a return that arrives to the pixel array with a certain phase shift θ_{depth}, which is exclusively given by the distance the light had to cover from the illumination system to the scene point and from it to the sensor (recall Eq. 2.1). Then we can formulate the environment response at that point as

$$e(t) = a\delta(t - t_{\text{depth}}) \tag{2.8}$$

where t_{depth} is the equivalent in time-of-flight to the depth to measure and $\delta(t - t_{\text{depth}})$ is a Dirac delta function centered at $t = t_{\text{depth}}$. Consequently, the reflected light signal that reaches the sensor is given by

$$r(t) = (i * e)(t) = A_0^r + A^r\cos(\omega t - \theta_r) \tag{2.9}$$

where $A_0^r = aA_0 + A_0^e$ accounts for the attenuated DC component of the signal, plus some unknown level of environmental light, also called background illumination, A_0^e. The phase offset is mostly due to the environmental response, $\theta_r = \theta_0 + \theta_{\text{depth}}$, but may contain additional delays produced by the electronics, wires and light emitters, which should be accounted for in θ_0. The reference signal is, at no additional phase shift, just a linear transformation of that in Eq. 2.7, that is

$$q(t) = A_0^q + A^q\cos(\omega t - \theta_q) \tag{2.10}$$

where the terms A_0^q and A^q simply model the EOM or the *in pixel* mixing, since the original signal might experience some rescaling or offset addition or subtraction. In the simplest case, $\theta_q = \theta_0$. Note that we implicitly suppose that the *residual* delays are the same in the *probe* path, i. e., the one carrying the signal for modulation, and the *reference* path, i. e., the one carrying the reference signal, similarly to an interferometer, being in both cases θ_0. In real cases this cannot be always ensured and an unknown phase offset might appear, which has to be fixed by calibration. The result

of the cross-correlation between the reflected signal and the reference signal is defined as

$$\begin{aligned} c_{q,r}(\tau) &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} q^*(t) r(t+\tau) dt \\ &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} [A_0^q + A^q \cos(\omega t - \theta_q)][A_0^r + A^r \cos(\omega(t+\tau) - \theta_r)] dt \end{aligned} \tag{2.11}$$

Eq. 2.11 can be simplified using known trigonometric equalities. For readability, we omit the mathematic demonstration here, but the interested reader is referred to Appendix A.1, where we demonstrate that Eq. 2.11 yields Eq. 2.12, which is a generalization to arbitrary sinusoidal signals of the formula already presented in [368] for the case of signals of equal offset and amplitude.

$$c_{q,r}(\tau) = A_0^q A_0^r + \frac{A^q A^r}{2} \cos\left(\omega\tau - \theta_{\text{depth}}\right) \tag{2.12}$$

where $\theta_{\text{depth}} = \theta_r - \theta_q$, by definition. Note that sampling the cross-correlation function at custom values of τ, $\tau_k = \theta_k/\omega$ is equivalent to performing optical interferometry (cf. Eq. 2.4). The cross-correlation samples are obtained by shifting the reference signal in θ_k. From the measurement point of view, we are in an equivalent case to interferometry and, consequently, the same methods for phase calculus from the measurements apply. In fact, the general multiphase depth estimation method described in [476], which seems to be that implemented in the ToF version of the Kinect sensor, is not more than that given in Eq. 2.6 and used in interferometry [332]. Obviously, the intensity values of Eq. 2.6, I_k, correspond to the cross-correlation samples $c_{q,r}(\tau_k)$ given by the pixels in an ORFI setup. If the illumination signal is not exactly sinusoidal, a careful choice of the phase shifts θ_k may be used to cancel the main harmonics at sensing. In [363] phase steps of 45° are used to cancel the third and fifth harmonics, which are the most dominant if the signals are square (recall that even harmonics are null in a square signal). Despite the method was implemented with PMD hardware, it is based on Eq. 2.6 and, therefore, also valid in ORFI setups with external EOM or pure interferometric setups.

In real cases, it is hard to generate a pure sinusoidal illumination signal using low-cost light emitters, such as LEDs. Using a coherent light source, together with an interferometric EOM, e. g., a Mach-Zehnder modulator,

yields an accurate sinusoidal modulation, at the cost of adding bulky and expensive components. Consequently, most commercial solutions rely on direct modulation of the light source. LEDs have to be driven in current and, therefore, require a driver, which is typically controlled by a binary ICS. The ICS is a square signal, alternating between a low level (typically zero) and a high level, and, in consequence, also a square signal controls the demodulation. Then, if the mixing occurs between two square signals, the main question would actually be why ToF systems based on ORFI work at all without tweaking the sampling as in [363], since the model (Eq. 2.9-2.10) is not satisfied and, consequently, Eq. 2.6 does not hold. Indeed, a square wave has a Total Harmonic Distortion (THD) of 48.3%, i. e., around half of the signal power is contained in the harmonic components. In general, for periodic, yet not sinusoidal signals, the reflected and reference signals may be written according to their Fourier representation:

$$\begin{aligned} r(t) &= A_0^r + \sum_{n=1}^{\infty} \left(A_{1,n}^r \sin(n\omega t - \theta_r) + A_{2,n}^r \cos(n\omega t - \theta_r) \right) \\ q(t) &= A_0^q + \sum_{n=1}^{\infty} \left(A_{1,n}^q \sin(n\omega t - \theta_q) + A_{2,n}^q \cos(n\omega t - \theta_q) \right) \end{aligned} \tag{2.13}$$

Substituting the expressions in Eq. 2.13 in Eq. 2.11, we obtain:

$$\begin{aligned} c_{q,r}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \Bigg\{ & A_0^q + \sum_{n=1}^{\infty} \Big[A_{n,1}^q \sin\left(n\omega t - \theta_q\right) \\ & + A_{n,2}^q \cos\left(n\omega t - \theta_q\right) \Big] \Bigg\} \\ & \Bigg\{ A_0^r + \sum_{n=1}^{\infty} \Big[A_{n,1}^r \sin\left(n\omega t + \omega\tau - \theta_r\right) \\ & + A_{n,2}^r \cos\left(n\omega t + \omega\tau - \theta_r\right) \Big] \Bigg\} dt \end{aligned} \tag{2.14}$$

which is the most general expression for mixing of two periodic signals, be by means of an external EOM or by means of a *smart* pixel. Simplifying Eq. 2.14 is not immediate and, for the sake of brevity and readability, we provide the full process in Appendix A.2. We mostly exploit the orthogonality of Fourier

basis, i. e., the product of two sinusoidal signals of different frequencies, which are multiples of some fundamental frequency, leads to a null contribution when integrated along a period of the lowest frequency, and well-known trigonometric identities. We obtain that Eq. 2.14 can be reduced to Eq. 2.15.

$$c_{q,r}(\tau) = A_0^q A_0^r + \frac{1}{2}\left\{ \left[\sum_{n=1}^{\infty} A_{n,1}^q A_{n,1}^r + A_{n,2}^q A_{n,2}^r\right] \cos\left(\omega\tau - \theta_{\text{depth}}\right) + \left[\sum_{n=1}^{\infty} A_{n,2}^q A_{n,1}^r - A_{n,1}^q A_{n,2}^r\right] \sin\left(\omega\tau - \theta_{\text{depth}}\right) \right\} \tag{2.15}$$

As already pointed out in [304], it would be sufficient to ensure that one of the signals involved in the correlation (Eq. 2.11) is a pure sinusoidal signal to completely remove the disturbing effect of the harmonics present in the other signal. In other words, if a perfect sinusoidal illumination signal is achieved, the shape of the reference signal, as used for the mixing at pixel level, is irrelevant, as far as it is a periodic signal of the same fundamental frequency.

Modulation and Demodulation at Different Frequencies Despite we have focused on mixing periodic signals of identical fundamental frequency, as well as in the case of optical interferometry, mixing two signals of different frequencies is also possible. The signal frequencies are to be kept high, in order to maintain good phase resolution, while the frequency difference has to be low enough to allow performing all the necessary acquisitions per period of the beat frequency. Obviously, the number of acquisitions per period has to be an integer and a perfect synchronization between the acquisition rate and the beat frequency is crucial. If a conventional camera is used to sample the beat signal, the frequency difference may fall on the hertz range, while typical modulation frequencies for imaging a real-world scene are in the megahertz range. The mixing can be carried out externally, e. g., using a Pockels cell or an image intensifier as external mixer, as done in [154]; or internally, profiting from the existing arrays of *smart* pixels, such as the PMD, which will be analyzed in depth in Section 2.3. A heterodyne camera using a PMD sensor for mixing is presented in [117].

Pseudonoise Modulation The pseudonoise (PN) modulation [67] is a *sui generis* modulation-demodulation scheme, since it cannot be classified according to the homodyne/heterodyne distinction. A PN sequence is used

both for modulation of the illumination and for demodulation of the reflected signal. Interestingly, a PN sequence exhibits a wide frequency spectrum, including very high frequencies, which are always a requirement for good depth precision. Additionally, this modulation scheme allows simultaneous operation of ToF cameras, since two different PN sequences have a null cross-correlation function. Nevertheless, the most significant advantage of a PN modulation is a very sharp autocorrelation function. In fact, the autocorrelation approaches a Dirac delta function as the length of the PN sequence tends to infinity. Summarizing, the sample cross-correlation between PN sequences is

$$c_{n_i,n_j}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} n_i^*(t) n_j(t+\tau) dt = \delta_{i,j}(\tau) \tag{2.16}$$

where n_i, n_j are normalized PN sequences of infinite length and $\delta_{i,j}(\tau) = \delta_{i,j}\delta(\tau)$ is an hybrid between the Dirac and Kronecker delta functions and, consequently, takes the value 1 only if $i = j$ and $\tau = 0$. Despite its attractiveness, the use of PN modulation for range measurements typically requires dense sampling of the cross-correlation function in order to find the unique maximum. Multiscale and adaptive sensing schemes may leverage this requirement by progressively restricting the feasible region where such maximum can be located.

Multiple Frequencies to Increase the Unambiguous Range The use of high modulation frequencies is a requirement to achieve low depth uncertainty. On the other hand, relatively low frequencies are required to achieve a sufficiently large unambiguous range (Eq. 2.2, where $f = f_{\text{mod}}$). As well as in the case of optical interferometry, a common solution is to use several frequencies to remove the ambiguity. Similarly, the achieved unambiguous range is the least common multiple of the unambiguous ranges corresponding to each modulation frequency. An intelligent option is to use two close frequencies, as done in [280] for long range ToF imaging. This way, high depth precision is preserved, while the unambiguous range is greatly increased. If only two close frequencies are used, the unambiguous range is given by half of the wavelength corresponding to the beat frequency of the corresponding heterodyne setup. Note that the reasoning of the beat frequency is equivalent to that of the so-called *synthetic wavelength* [474], often found in interferometry, being the synthetic wavelength the one corresponding to the beat frequency. For instance, using two high frequencies with 1 MHz difference yields an 150 m unambiguous range.

Another option would be to combine a low frequency with a high frequency. The low frequency is used to determine how many full ranges of the high frequency have passed, i. e., a coarse estimation, while the high frequency is used to get a fine—yet ambiguous—estimation of the depth. Such a setup may not bring a large increase of the unambiguous range and it also does not enhance the depth precision with respect to that obtained using the high frequency alone.

In general, in a multifrequency setup measurements are gathered sequentially [280, 152] and the number of acquisitions scales with the number of frequencies. A careful selection of the frequencies and of the sampling points allows operating with two frequencies simultaneously—or sequentially but still within the integration period, which is equivalent—, as in [364], where the phase differences between sampling points for the lower frequency is twice that for the higher frequency. This idea allows reducing the number of acquisitions when using two frequencies but, as the authors recognize, the use of a low frequency increases the observed depth standard deviation.

In the limit, using a very large number of frequencies, one can do without phase shifting, that is, one can operate in frequency domain, acquiring intensity measurements for different frequencies, at an arbitrary phase, e. g., zero. This is the framework proposed in [249], inspired by optical coherence tomography. Similarly to phase (or time) domain, where the depth is encoded in phase shift, in frequency domain the depth is encoded in frequency. While this method overcomes depth ambiguities and permits resolving several paths per pixel, it demands a large modulation frequency bandwidth.

As presented in Fig. 2.5, despite they share the same principle of operation and, in consequence, the same formulas for phase retrieval, the setups for ORFI can be classified in two different categories, regarding whether the demodulation is carried out using an external EOM or internally by *smart* pixels. Before the development of the latter, initial 3D-vision systems [405] made use of external EOMs such as Pockels cells. The different possibilities that exist for performing electrically-driven optical modulation are shortly discussed in the following, while the *in pixel* modulation is left to Section 2.3, where the PMD technology is presented.

Methods for External Optical Modulation In this section the main techniques for performing optical intensity modulation of light are introduced. We adopt the underlying physical principle as classification criterion, as in [222]. According to this principle, we distinguish three categories: electro-optic, acousto-optic and magneto-optic, which we analyze in the following.

For further information on the topic the reader is referred to [467, Chapter 3] and [222].

Mecano-optic Mechanic methods for shuttering could be used as a mean for binary modulation. Unfortunately, conventional mechanical shutters, widely used in photography, rarely can provide exposures below one millisecond. This would lead to maximum modulation frequencies in the kilohertz range, far from the megahertz frequencies typically required in ToF imaging. Nowadays, there exist cameras providing up to 125 μs exposure. Supposing that we desire a square binary modulation (50% duty cycle), this poses a frequency limit of 4 KHz. One could think of mechanical setups allowing elements moving periodically at high speeds, which could act as shutters. One example may be a rotating cogwheel, used as in the famous Fizeau experiment, moved by a high-speed and low-torque motor, after appropriate speed multiplication, e. g., by means of a gearbox. Very high speed motors, such as those used for generating vibration, can turn at speeds up to 10^5 rpm, while having millimeter-range sizes. A prototype of the *MegaNdrive* low power motor of the ETH Zurich achieved over 10^6 rpm [490] before disintegration of the ball bearings. The Fizeau's cogwheel may be substituted by a printed disc of transparent plastic or glass, similar to those used in optical rotary encoders. While such an approach may achieve the required modulation frequencies, the presence of a rotary element turning at hundreds of thousands of rpm is reason enough to avoid this method due to maintenance considerations. Still, there are mechanical methods that suit the application of fast light modulation. We contemplate here using the so-called Frustrated Total Internal Reflection (FTIR) and microelectromechanical systems (MEMS).

FTIR, recently used for optical touchpads, is not a new technique. Under the denomination of *optical tunneling*, it has been extensively discussed in [34], where typical applications are given, such as optical beam splitters, modulators and filters. The basic principle and formulas are given in [265, Section 5.5], under the name of Frustrated Total Reflection (FTR). Consider the interface between two media of different refractive indices, such that the light is transmitted from the medium of higher refractive index to that of lower. Applying Snell's law, we trivially get that, for incidence angles higher than a critical angle, which depends exclusively on the ratio of refractive indices, all light gets internally reflected in the first medium and no light gets transmitted to the second. This phenomenon is often called Total Internal Reflection (TIR). If a third medium is added after the second, with a refractive index that is higher than that of the second, so that the width

of the second medium is around the value of the wavelength of the radiation, then the light is no longer totally reflected and gets transmitted to the third medium, even when the incidence angle in the first medium is greater than the critical angle. This effect does not contradict Snell's law, since both interfaces are still totally reflecting, being the partial transmission due to an interference effect. Consequently, for widths of the second medium that are of higher order of magnitude than that of the wavelength of the radiation, the addition of the third medium after the second does not have any effect. Consequently, the modulation procedure is based on varying the width of the second medium. If the second medium is air, one way of varying the width is, e. g., displacing the interface of the third medium with respect to the first.

If NIR radiation is being used, micrometric or submicrometric displacements are enough to change the transmittance of the system from 0 to close to 100%. This makes FTIR highly recommendable for light modulation, since displacements in the range on tens of nanometers to micrometers can be achieved, even at very high repetition rates, without the need of bulky mechanical actuators. The most common way of performing this task is by means of a piezoelectric actuator. Piezoelectric actuators are based on the inverse piezoelectric effect, i. e., the ability of some materials (e. g., some crystals and ceramics) to generate a mechanical strain under the application of an electrical field. The mechanical strain may eventually deform the material and be used to generate fast short displacements. If the piezoelectric crystal is excited at frequencies close to the resonance frequency, high vibration amplitudes can be achieved with relatively low voltage levels. Since piezoelectric crystals are often presented as thin disks, two resonance frequencies are observed, corresponding to the radial and longitudinal modes, respectively, being the latter the highest, often in the megahertz range. A review of the properties of the main piezoelectric ceramics, such as the well-known lead zirconate titanate (PZT) is provided in [241]. Other suitable piezoelectric materials might include polymers, such as the polyvinylidene fluoride (PVDF). Binocular setups for a 3D camera using two coupled FTIR modulators, for modulation and demodulation, are proposed in [474, 222].

One of the principal factors to consider on any mechanical system with parts moving at high speed is the total aerodynamic drag, which depends linearly on the density of the fluid surrounding the system and the surface in contact with it and quadratically on the relative speed between the fluid and the moving part. That is, a linear increase of the speed of a moving part translates, *ceteris paribus*, into a quadratical increase in the required energy. Nevertheless, if the surface in frontal contact with the fluid, typically air, is

reduced to a negligible value, then the corresponding aerodynamic drag will be also negligible. Similarly, if the reduction of surface is accompanied by a reduction of volume of the system parts, their mass will decrease and so will do related forces and torques. That is, under miniaturization conditions, physical phenomena that are dominant at macroscopic scale, such as the effect of gravity, become irrelevant in comparison to the effect of other, e. g., electrostatic fields, which are negligible at macroscopic scale. Microelectromechanical systems (MEMS) are systems made up of components of sizes ranging from one to a hundred micrometers, which may combine mechanical microstructures, microsensors, microactuators and microelectronics. Typically the volume reduction with respect to macroscopic systems is higher than the surface reduction, leading to greater ratios surface/volume. This concede even more importance to electrostatic effects. Since the inertia of the components is very low, well-designed MEMS can produce microscopic movements (e. g., rotation, tilting) that are fully controlled by a rapidly-varying electrical field, with very low response times. For instance, MEMS engines able to rotate at speeds up to 10^6 rpm have been characterized in [380], when rotating up to 5×10^5 rpm.

Even more interesting than using high speed MEMS motors for shuttering are the arrays of micromirrors, which are able to tilt independently, driven by an electrical field. The worldwide leading technology in this area is the Texas Instruments Digital Light Processing (DLP®). The core of the DLP is the so-called Digital Micromirror Device (DMD®[3]), which was invented by Larry Hornbeck in 1987 [229]. Each DMD pixel can be in two stable micromirror states, namely, tilted $+12°$ or $-12°$ with respect to the surface normal [145]. Below each micromirror, a dual CMOS memory stores the desired state of the micromirror. Once the memory is set to the desired mirror state, a clock signal is responsible to translate the memory state into an actual mirror state. The micromirror is able to tilt thanks to a central via, which attaches it to a torsional hinge. It is clear that such a system can provide binary modulation with very large contrast. Furthermore, using two detectors would lead to a lossless modulation, according to two simultaneous binary signals displaced $180°$ between each other. The remaining issue is the driving speed. In normal operation, the main time constraint is imposed by the data transfer to the memories of the DMD pixels. Suppose that the DMD pixels can be all simultaneously set (not possible) and reset (possible) to one or another position. In that case, if all the DMDs can be driven by a

[3]The Texas Instruments DLP® and DMD® are registered trademarks and will be denoted in the following without the ® symbol for notation simplicity.

single signal without sequentially writing the memories, the minimum period is given by the mirror transition and settle time [158]. It takes about 18 μs to achieve a stable position, but the transition itself is completed in less than 5 μs. Summarizing, a DLP could provide a square binary modulation, with 100% modulation depth at a physically-limited maximum frequency of 100 KHz.

Electro-optic Three classical electro-optic effects that can be used for light modulation are presented here: the linear electro-optic effect, also known as Pockels effect, the quadratic electro-optic effect, or Kerr effect, and the electro-optic effect in liquid crystals. Additionally, the use of multiple quantum well (MQW) structures as light modulators exploiting the quantum confined Stark effect (QCSE) is briefly introduced.

The Pockels effect [62] refers to the linear variation of the refractive index (birefringence) that some materials experience under the application of an electric field. The electric field can be applied longitudinally or transversely to the light beam. Longitudinal Pockels cells require transparent electrodes. In general, Pockels cells may be used to rotate the polarization of a polarized light beam. Consequently, in combination with polarizers, they provide a technique for intensity modulation of light.

Within a Pockels cell, the refractive index depends on the polarization of the light beam, i. e., light polarized in a certain direction propagates faster through the crystal than light polarized in the perpendicular direction. In longitudinal Pockels cells, the phase difference between these two principal axes is directly dependent on the applied electric field, and not on the length of the crystal. The main drawback of Pockels cells is the relatively large voltage required to generate the electrical field. The KD*P (Potassium Dideuterium Phosphat) modulator studied in [222] required 5233 V to achieve the so-called half-wavelength point (180° phase shift) and 1850 V for the quarter-wavelength point (90° phase shift), for a wavelength of 630 nm. Another issue is that the transmittance of the Pockels cells depends largely on temperature [474], so that variations of tens of degrees Celsius translate into discrepancies of thousands of volts in the previous values.

In the transversal case, the phase difference is directly dependent on the modulator length. Transversal cells with small apertures can have lower voltages, but in the case of depth imaging, flat 2D modulators with relatively large apertures are required. For instance, a Pockels cell from KD*P crystal of 25×25 mm still requires driving voltages in the range of 1500 V.

Pockels cells behave as capacitors from the electrical point of view and, clearly, driving them with voltages of hundreds to thousands of volts and

switching speeds in the nanosecond range require very fast electronics, able to deliver very high currents. Apart from space or efficiency considerations, systems driving tens of amperes at megahertz frequencies may introduce distortions in the power network that could disturb the correct operation of other devices. Electromagnetic interference (EMI) might also become an issue, both for the camera and for other electronic systems close to it.

The Kerr effect is similar to the Pockels effect, since it is also a change in the refractive index induced by an electric field. It differs from the Pockels effect in that the variation is directly proportional to the square of the applied electric field, for which reason it is also called quadratic electro-optic effect. There exist two different types of optical Kerr effects: the electro-optic and the pure optical effect.

The first one is also called DC Kerr effect and is exploited in a setup that is close to that of a Pockels cell, in that the external electric field is applied by means of electrodes at the sides of the material. The electric field is generated perpendicularly to the light beam and two crossed polarizers are placed before and after the material. Some organic polar liquids exhibit a strong DC Kerr effect and are confined in a glass cell, forming a so-called *Kerr cell*. Kerr cells can be used to achieve very fast modulation, in the gigahertz range, at the cost of driving voltage even higher than those required for Pockels cells, often in the kilovolt range.

The optical Kerr effect is also called AC Kerr effect and studies the case when the electric field is produced by the light itself. The variation of the refractive index depends linearly on the intensity of the radiation, i. e., quadratically on the amplitude of the wave, as in the DC Kerr effect. A modulator based on the DC Kerr effect would avoid the need for electrodes and high voltages, providing a mean for pure optical mixing, since the intensity of a *reference* beam can be used to change the polarization (and, in combination with the crossed polarizers, modulate the intensity) of a *probe* beam. Unfortunately, very large intensities (irradiances in the order of GW/cm^2) are required to produce the AC Kerr effect in most materials, only achievable by a well-focused laser beam.

Regarding the electro-optical effect in fluids, liquid crystals [52, 258] deserve special mention. As said above, some organic polar liquids show strong DC Kerr effect, but the principle of operation of optical modulators based on liquid crystals is slightly different. Liquid crystals are liquids in which the molecules are arranged according to a definite structure, i. e., a crystalline state, which leads to anisotropic mechanic, electric and magnetic properties, differently from the isotropic non-crystalline liquids. As pointed out in [258], the birefringence produced in the liquid crystalline state can

be up to four orders of magnitude greater, with similar permanent dipole moments. Since the molecules of the crystal are dipoles, when an external electric field is applied, they orient themselves along the direction of the field. Applying a magnetic field produces a similar effect and the molecules tend to align with or against the field. Controlling the orientation of the molecules translates into controlling the rotation of polarized light. Together with crossed polarizers, liquid crystals can be used as optical modulators. The principal drawback is the low speed of the reorientation process, i. e., once the electric field is extinguished, the molecules need times in the millisecond range to return to the original orientation. Imposing special structures to the liquid crystal may lead to a reduction of the decay times. The nanocomposite presented in [238], which combines an anisotropic polymer with a liquid crystal, achieves a decay time of 15 μs. A different composite is presented in [237], achieved by photopolymerization of a mixture of a monomer and a liquid crystal. In that work nanometric pores are formed in the polymeric matrix and a decay time of 10 μs is reported. Still, supposing symmetric rising and falling times, 10 μs decay time means a frequency limit of 50 KHz, still far from the megahertz level.

One of the major issues of external demodulation, which is solved with *in pixel* demodulation, is to avoid an eventual image distortion produced by anisotropy of the materials and complex optical paths. It is, thus, desirable to perform the external demodulation as close to the pixel surface as possible. For this reason, spatial light modulators made of nematic ferroelectric liquid crystal (FLC) are an appealing solution, since they are fast and it is possible construct them over VLSI silicon (the so-called FLCoS or FLC-LCoS) [116], which can, in turn, be used for the image sensing. A schematic of a FLC-LCoS for *on pixel* ORFI demodulation in 3D imaging systems is found in [474].

Multiple quantum well (MQW) structures are formed by stacking quantum wells. They were intended to be a part of optically-controlled optical switches, basic components of a fully optical computer, which would provide faster processing and communications. A representative example of such devices is the so-called self-electro-optic device (SEED) [327], which is briefly a MQW-based switch with optical inputs and outputs. Briefly, a quantum well is a potential well where only discrete energy levels are allowed. One method to create quantum wells is to confine particles to a very thin layer, i. e., a 2D space, instead of 3D. When the well thickness becomes comparable to the *de Broglie* wavelength of the carriers, the so-called quantum confinement is achieved and the energy levels are quantized in energy subbands, which range from 10 to 100 meV. As a result, a staggered optical absorption spectrum

is observed, with strong absorption peaks at the edges of the steps—the so-called exciton absorption peaks—.

Quantum wells have a sandwich structure, where the external semiconductors exhibit a wider bandgap than that in the middle. Common materials for constructing quantum wells are gallium arsenide between aluminium gallium arsenide or indium gallium nitride between gallium nitride [186]. A MQW structure is not more than a sequence of alternating thin layers of each material of the quantum well pair (e. g., 10 nm width per pair). In general, quantum well modulators are based on the quantum confined Stark effect (QCSE), which describes the effect of applying an electric field to a quantum well. If an electric field is applied perpendicularly to the MQW layers, electrons and holes are pulled toward opposite barriers of the layer, but still confined in the smaller bandgap layer (middle layer of the *sandwich*), reducing the energy of the electron-hole pair. Consequently, the optical absorption associated with the creation of the pair decreases and the absorption edge shifts towards higher wavelengths (i. e., less photon energy). This effect is sensibly stronger in quantum wells than in a bulk semiconductor, since it gets enhanced by bound excitons. The effect can be used to construct an electrically-driven absorption-based fast light modulator, either in reflective or transmissive mode. As a side effect, the refractive index also changes when applying an electrical field and, therefore, refractive modulators are also possible.

MQW modulators have several advantages when compared to other optical modulation techniques. One of them is the very low energy consumption, comparable to that of a full electronic system. Electrically, the quantum well behaves as a capacitor of very low capacitance, requiring low electrical fields (ranging between 10^4 and 10^5 V/cm) for modulation. Additionally, it is of capital importance that the modulator can be located *on pixel*, over a silicon-based detector array, e. g. a CCD array. This fact implies that better fill factors than those achievable with the *smart* pixels introduced in Section 2.3 can be achieved, thanks to the decoupling between modulation and sensing processes. A schematic of a CCD array with built-in MQW modulation layer in transmissive mode for application in ToF imaging is found in [474]. Note that the modulation operates—differently from PMD pixels—in transversal direction and therefore does not impose any restriction on the size or structure of the pixels.

Acousto-optic Acousto-optic modulators (AOMs) are based on the acousto-optic effect, which is a specific case of photoelasticity, that is, a change of the permitivity of the material under a mechanical strain. In short terms,

an acoustic wave produces a periodic strain in the material, which leads to a periodic variation of the refractive index with the same frequency as the acoustic wave. Since the acoustic wave also travels through the material, the refractive index variation is a sinusoid that depends both on the time and the position. Consequently, a diffraction grating is produced, that travels through the material in the propagation direction of the sound wave, at the same speed this propagates. For simplicity, consider that the acoustic wave is of single frequency, then two different types of diffraction are produced: Raman-Nath diffraction and Bragg diffraction. The later occurs at high acoustic frequencies, often over 100 MHz, while the former occurs for lower frequencies, typically lower than 10 MHz.

AOMs based on Bragg diffraction are also called *Bragg cells* and are composed by a piece of the appropriate material, typically quartz or glass, with a piezoelectric transducer at one side and an acoustic absorber at the other. The acoustic wave travels through the material, from the piezoelectric actuator to the absorber. The light beam traverses the material and diffracted beams appear at specific angles that depend on the order of diffraction and the ratio between the wavelength of the light and that of the sound. The intensity of the diffracted beam is proportional to the amplitude of the acoustic wave [462]. Consequently, the Bragg cell can be used directly as a linear light modulator. Since the acoustic wave travels through the material, the frequency of the diffracted beam of order m will be that of the input beam plus a Doppler shift of m times the frequency of the acoustic wave; that is, a Bragg cell could also be used for optical frequency shifting [2], e. g., in a heterodyne interferometric system. Similarly, the phase of the diffracted beam is also shifted by that of the acoustic wave. Also remarkable is the fact that Bragg cells produce a rotation of the polarization angle in a linearly-polarized beam [161], an effect which is exploited by other electro-optical and magneto-optical modulation techniques.

Magneto-optic We consider four magneto-optic effects that can be used to build a magneto-optic modulator: the magneto-optic Kerr effect, the Faraday effect, the Cotton-Mouton effect and the Voigt effect. In general, magneto-optic effects are the magnetic counterpart of electro-optic effects, since they are mostly based on optical rotation, produced by an anisotropic medium. This translates into non-diagonal permitivity or permeability tensors [378].

The magneto-optic Kerr effect (MOKE) or surface magneto-optic Kerr effect (SMOKE) [459] describes the change of polarization and intensity that the light experiences when it is reflected by a magnetized surface. The effect is similar to the Faraday effect, differing in that the Faraday effect describes

the changes experienced by light when it traverses a magnetic material, while the MOKE accounts only for reflection on the surface. In short terms, the MOKE is based, as well as the Faraday effect, on the anisotropic permitivity of the material. Provided that the speed of light through the material is inversely proportional to the square root of the product of the permitivity and permeability, light propagates at different speed depending on the orientation. As a consequence, variations in the phase of polarized light are observed. Depending on the direction of the magnetization vector with respect to the surface and incidence plane, the MOKE can be classified in three categories: polar, longitudinal and transversal. In the polar setup the magnetization vector is perpendicular to the reflective surface and parallel in both the longitudinal and transversal cases. The magnetization vector is also parallel to the plane of incidence in the longitudinal setup, while perpendicular in the transversal. In the longitudinal MOKE light changes its polarization from linear to elliptical, being the change proportional to the magnetization vector. In the transversal MOKE, the surface changes its reflectivity, being the change proportional to the magnetization vector.

The Faraday effect, also called Faraday rotation, is a simple magneto-optical phenomenon that describes the rotation of the plane of polarization in presence of a magnetic field, being the rotation proportional to the component of the magnetic field that is parallel to the direction of the beam. In the Faraday effect the rotation is due to a non-diagonal permeability tensor, while the permitivity tensor is supposed diagonal. The angle of rotation is given by the product of the Verdet constant of the material, the magnetic flux density in the direction of the beam and the length of the sample. Materials with high Verdet constants can be used to construct Faraday rotators. An appropriate material to construct a Faraday modulator is the terbium gallium garnet (TGG) [27], which is interesting due to its high transparency and very high Verdet constant. It is to be taken into account that the Verdet constant of TGG strongly depends on the wavelength of the radiation, decreasing dramatically with it, i. e., the material is more suitable for operation in the visible range than in the NIR. A variety of organic and inorganic liquids exist, which exhibit high Verdet constants [456], still typically one order of magnitude lower than that of TGG. As in previous cases, placing a sample of the material between crossed polarizers, together with the appropriate driving system, yields a simple optical modulator. The main drawback of the Faraday effect, as well as of the other magneto-optic effect, is its extremely low efficiency. Large magnetic fields are required to generate moderate optical rotation. For instance, if a sample of TGG of 10 mm length is used to construct a Faraday rotator, a magnetic flux

density of approximately 0.5 T is needed to rotate the polarization 45°, when operating with 600 nm light. Furthermore, so intense magnetic fields cannot be generated at high frequencies, due to the highly-inductive coils used to generate them.

Apart from rotation of the plane of polarization (Faraday effect), the magnetic field also produces birefringence, which is known as Cotton-Mouton effect [121]. The optically isotropic probe becomes anisotropic in presence of a magnetic field applied perpendicular to the direction of light propagation through the probe. The refractive index, as observed by the light traversing the probe is different for light polarized parallel and perpendicular to the direction of the magnetic field. This induces a *retardation* or optical path difference between parallel and perpendicularly polarized light at the end of the probe which depends linearly on the difference of refractive indices and the length of the probe. The variation of refractive indices depends linearly on the material-dependent Cotton-Mouton constant, the wavelength of the light and the square of the magnetic field.

The Voigt effect [452] is often presented as an independent effect, but it is a magnetically-induced birefringence, observed in materials that are optically isotropic in absence of magnetic field. For this reason, it is common to find in the literature explicit references to the equivalence between the Cotton-Mouton and Voigt effects [489, 451]. The effect depends quadratically on the magnetic field and produces a change of ellipticity of circularly polarized light.

Apart from all methods described above, image intensifiers can also be used to construct an EOM. The use of an image intensifier as EOM in a ToF imaging system is reported in [154]. Previously, submillimetric depth resolution had been reported in [226], where an image intensifier was used as external EOM in combination with a CCD camera. Key points of the approach in [226] are the coaxial illumination and detection, the high modulation frequency (up to 225 MHz) and large number of phase steps (up to twelve). Image intensifiers are optoelectronic devices commonly used in medical imaging and low-light imaging applications. Several generations of image intensifiers have been developed in the last decades, driven by the military demand. A common factor among all of them is that they are based on the idea of the image intensifier tube, in which the photons are converted first into electrons, then the electrons are amplified and converted back into photons by means of a phosphorescent material. The image or beam first reaches the photocathode, where electrons are generated by photoelectric effect. A microchannel plate (MCP) at high voltage is used to accelerate the electrons, which, in turn, release multiple electrons when they hit the plate.

At the end of the tube, the electrons hit a phosphor screen, where a photon is released for every electron, generating an intensified image or beam. Clearly, decreasing the voltage of the MCP would slow down the electrons and fewer would arrive to the phosphor screen, decreasing the image intensity. That is, an image intensifier can be directly used as intensity modulator with amplification.

2.3. The Photonic Mixer Device (PMD)

This section describes the technique used to perform the ORFI demodulation or mixing of a modulated light signal with a reference electric signal—that is, electro-optical (de)modulation—at the detector. In the case of phase-shift-based ToF this implies that each pixel of the detector array has to be endowed with an apparatus to regulate the integration process according to the external reference signal. One can predict that, regardless of the specific implementation, this idea requires both *in pixel* control electronics and modifications in the structure of the active area of the pixel. Such pixels are often called *smart* or *intelligent* pixels and are introduced in Section 2.3.1. The initial concept of *intelligent* pixels aimed to integrate control functions and basic computational functions within the pixel [140], while keeping an acceptable fill factor. The first array of *smart* pixels that was made commercially available as a ToF imaging system implemented the so-called Photonic Mixer Device or PMD technology. The original idea of the PMD was conceived by Prof. Rudolf Schwarte in 1996 [402], at the *Institut für Nachrichtenverarbeitung* of the University of Siegen, today the Center for Sensor Systems (ZESS). The production of PMD chips was externalized to S-Tec Sensor GmbH, founded in 1996 and today integrated in the well-known PMD Technologies. Nowadays PMD technology remains being a reference technology worldwide and, for this reason, has been adopted as the reference phase-shift-based ToF technology in this work. A description of the PMD pixel structure and its principle of operation is given in Section 2.3.2.

2.3.1. *Smart* Pixels

A definition of *smart* pixel that fits the wide range of usages of the term found in the literature could be: a pixel whose measurement after integration is not directly the result of a photon count during the exposure time, but the result of a different process, still mainly influenced by the arrival of photons to it. According to its functionality, *smart* pixels used in ToF imaging can be

classified in those actually measuring the *time of flight* of the light and those intended to provide samples of the cross-correlation between the modulated light signal that arrives to the pixel and some reference signal.

The pulse counting ToF system of [19] is an example of a time-counting system, but recent research in this area points to the single-photon avalanche diodes (SPADs) [481] as the reference technology for pure ToF imaging. In very short terms, the diode is reverse biased above breakdown voltage for a short time, so that, when a photon generates a charge carrier, an avalanche may be triggered. Obviously, unless the avalanche is quenched, the diode will be destroyed. For that reason, once the avalanche is produced, the bias voltage is immediately lowered below breakdown. This operation mode is called Geiger mode and, in theory, implies a virtually infinite gain in the sensing process. SPADs can be constructed using conventional CMOS technology [102] and, consequently, dense arrays of SPAD pixels can be generated and used as a ToF image sensor. A 32×32 CMOS array of SPADs for ToF imaging was presented in [351], where the circular shape of the pixels prioritizes correct operation of the SPAD over fill factor. Millimetric depth standard deviations are reported in [351] and submillimetric in [352]. The shape evolved from circular to octagonal and the fabrication process [103] from $0.8\,\mu$m to 65 nm. SPAD array sizes have grown up to that of conventional correlating pixels (PMD arrays), e. g., the 160×120 single-photon image sensor in [443]. Conversely, scanning mirrors may be used to obtain an acceptable depth image size, at the cost of having to scan the scene, increasing motion sensitivity and decreasing the frame rate [10]. A disadvantage of SPAD arrays is the extremely low fill factor, e. g., 1.1% in [351] or 1% in [443], which has to be effectively increased by means of microlenses. Other technical issues are the complex readout schemes and low frame rates.

Smart pixels internally performing a cross-correlation are typically multitap pixels. A *multitap* pixel features two or more integration wells, also called integration channels or *taps*. The modulated light signal that arrives to the pixel surface is transferred from optical to electrical domain due to photoelectric effect. The generated charge carriers are then stored in one of the integration wells at a time, being the active tap selected by a control signal. Clearly, the level of each pixel channel at the end of the integration is equivalent to scalar product of the modulation signal and the reference binary signal used to control the integration in that channel. If the pixel features only two channels and the integration time is divided equally for integration in each of them, the result of each channel is a sample of the cross-correlation between the modulation signal and a square signal between

zero and some maximum level. The difference of both channels, provided that the reference signals are complementary, corresponds to the cross-correlation between the modulation signal and a square signal of double amplitude and no DC offset (i. e., pure AC). In case we have a large number of pixel channels, the process is equivalent to sampling the modulated light signal itself. If the number of channels tend to infinity, then each channel provides a sample of the cross-correlation between the modulation signal and a Dirac delta function, which is the definition of a conventional sampling process.

The first multitap pixels were presented in [423, 421]. The idea of a lock-in structure was immediately adopted to generate the first 2D arrays of demodulating pixels for ToF imaging [422, 278]. Note that this idea is closely related to the array of QE-modulated photodiodes presented in [20]. General introductions to lock-in pixels, together with the main challenges they face can be found in [71, 69]. The *smart* pixels can achieve very large dynamic ranges, over 100 dB [422], by means of an offset subtraction scheme, similar to that implemented in the active pixel sensor (APS) of [450, 449], able to achieve 150 dB while keeping a linear response. Other methods for increasing the dynamic range are the minimum charge transfer (MCT) and the pixelwise integration (PWI) [291]. The basis of MCT is the fact that only the differences between the charge level of the different integration gates is relevant, while the common mode part is due to unmodulated light (DC offset, background illumination) and can be discarded. Consequently only the relevant signal part is transmitted from the integration well to the sense node. To this end, the integration gate potential is lowered at a certain slew rate and the transfer is stopped when all taps become lower than a threshold. This way, the common mode stays underneath the integration gates and do not get transmitted to the sense node. This method may be seen as a precursor of the PMD suppression of background illumination (SBI), which is briefly introduced in Section 2.3.2. PWI adapts the integration time independently for each pixel. This is done by means of sequential integration steps and seems a prior hardware implementation of the *staggered integration* idea presented in [304]. Another hardware solution for increasing the dynamic range of these pixels [21] performs a reset of the common mode component before the charge reaches the saturation level, while preserving the difference between channels. This can be done as many times as the charge approaches that level during integration. Especial pixels have been designed to achieve outstanding inmunity to background light (BGL), such as that presented in [136], which uses a current-sample-and-hold (CS&H) circuit to sense the current generated by DC light. According to the authors, the pixel achieved the highest BGL-immunity reported at that time.

Multitap pixels with number of taps ranging from one to 16 are found in the literature. One-tap pixels are *sensu stricto* not multitap pixels, but a conventional pixel where the integration can be inhibited by an external signal. A one-tap pixel has been presented in [278], as an attempt of maximizing the fill factor. Note that in the one-tap configuration, charge carriers that are generated out of the integration intervals are not stored in any channel and do not contribute to any measurable signal, differently from, e. g., a two-tap configuration. If the signal controlling the integration is square of 50% duty cycle, then, supposing a random depth distribution, on average, 50% of the optical energy captured by the pixel is being wasted. Another drawback of one-tap pixels is the rise of the total acquisition time, since only one sample of the cross-correlation function is gathered per acquisition. A first design of a four-tap PMD pixel is presented in [406]. The main advantage of a four-tap structure is that only one acquisition is required, providing better robustness to motion. A clear drawback is the degradation of the fill-factor and the eventual need for longer exposure times per acquisition. Two pixel structures for a four-tap pixel are evaluated in [278]. The first one is a wheel-shaped CCD structure, where the integration channels or storage gates are located around the photogate, which is in the middle. The main disadvantage of the wheel-shaped four-tap pixel is the very low fill factor, direct consequence of the shape. Another design, with $1:2$ aspect ratio and two output diffusions instead one, is proposed. The fill factor is three times higher than that of the wheel-shaped pixel and the symmetry of this design seems to reduce inhomogeneities between taps.

Interestingly, the first solid-state ToF camera [279] implements one-tap pixels, in order to get rid of the inhomogeneities between pixels channels that affected the first arrays of *smart* pixels and maximize the fill factor, also very low in the first pixels. Note that increasing the fill factor implies decreasing the exposure times and, in consequence, partially compensating the need for more acquisitions derived from using only one channel per pixel. Nowadays, pixels with two-taps and four-taps are common and can be reliably produced. A dedicated diffusion node and amplifier per tap allows increasing the read-out speed by parallel operation. The last generation ToF cameras, e. g., PMD cameras, the SwissRanger ToF cameras [355, 354] or the new Kinect sensor implement two-tap pixels. The pixels in the SwissRanger SR3000, for instance, are similar to the two-tap PMD pixels, but featuring three photogates instead two for a smoother potential gradient [72]. The sensor provides QCIF resolution and the pixels can operate with frequencies up to 41 MHz. Fast carrier separation and storage is achieved using CCD structures, while CMOS technology is used to allow flexible and

fast readout. The first commercial ToF imaging sensor implementing these *smart* pixels was named Photonic Mixer Device (PMD), which became the reference technology for phase-shift ToF imaging and will be described in Section 2.3.2.

Especial pixels, designed for extreme operating conditions can be found in [66]. Two pixel architectures are presented, which enable modulation-demodulation frequencies of hundreds of megahertz. The first architecture is similar to that of photo-gate PMD pixels, but substituting the several conductive photogates by a single high-resistive photogate [70]. This allows a smoother, linearly-decreasing, potential in the semiconductor, instead of a staircase-shaped potential distribution. This modulated-drift-field pixel was said to allow for frequencies up to 1 GHz. Unfortunately, the high per-pixel power consumption makes it unsuitable for high-resolution sensors. The second design overcomes some issues of the first, such as high power consumption and low-pass filtering effect of the gate signal, by means of photogates with huge resistance and capacitance. The core idea of the pixel is to separate the detection from the demodulation, still happening both *in pixel*. A static drift field is applied to the detection area, allowing fast charge separation and transport to the demodulation area [68]. Instead a single photogate, many thin photogates connected in series, with intermediate resistors, are used, in order to achieve a high resistive and capacitive load. The demodulation process, which is just a switching between two storage capacitances, is carried out in a small region and can also be performed at high speed.

Recently, new pixel designs featuring three taps have been proposed, in an attempt to reduce the number of acquisitions per depth image. Recall that in a phase-shift-based ToF system with sinusoidal modulation there are three free parameters and, thus, three measurements suffice to estimate them. The LEFM proposed in [212] features four taps for three effective outputs and a drain. The three-tap structure allows automatic subtraction of the BGL, since one of the taps is integrating exclusively BGL an not the reflected light signal, while the other two integrate the sum of BGL and signal. The use of lateral drift-field for charge transport reminds the LDPD design in [420], which allows for on-chip SBI also by means of three pixel channels.

Some works exist presenting pixels that try to separate the correlation process from the photogeneration process, tightly coupled in the PMD pixels. In such pixels, a fast photodiode is used to sense the incoming light signal and the correlation process is carried out in a subsequent step, by means of additional *in pixel* circuitry. A 30×50 array of such pixels is presented in

[427]. The main drawbacks of the pixels in [427] are a large pixel pitch of $81.9 \times 81.7\,\mu$m and a low fill factor of 20% (one third less than, e. g., 19k or 41k PMD pixels). An improved *in pixel* circuitry is presented in [480], which offers outstanding immunity to ambient illumination (± 5 cm depth variation up to 100 klx) and low power consumption. They achieve a fill factor of 66%, but with a large pixel size of $125 \times 125\,\mu$m. A competitive 160×120 array of pixels of this kind is presented in [366]. Their 'Correlated Double Sampling' (CDS) emulates the natural CBS of two-tap PMD pixels. The prototype poses a serious competitor to the original PMD concept, not just because of the size of the array, equal to that of the PMD 19k, but also because of the combination of a similar fill factor (34%) and an outstandingly low pixel pitch ($29.1\,\mu$m). Another pixel design that is worth mentioning is that in [236], which features *in pixel* SBI and adaptive accumulation, while offering an outstanding fill factor of 77% (cf. 60% for the Xbox One sensor, with microlenses).

A review on different ToF 3D imaging pixel structures in CMOS is provided in [159]. In [277], a simulation framework is proposed to evaluate different ToF chip layouts, namely, single and dual readout configurations. Differently from existing simulations, where the focus is on the final depth image, a physically meaningful modeling allows obtaining raw images that are not a simple intermediate result, but a realistic simulation of the real ones.

2.3.2. Principle of Operation of the Photonic Mixer Device

The Photonic Mixer Device (PMD) [77] was born with the intention of bringing the demodulation process required in ORFI ToF imaging closer to the detection process. Differently from FLC-LCoS or MQW, which provide *on pixel* demodulation, the PMD principle of operation constituted a breakthrough in that it performs *in pixel* demodulation, that is, the processes of signal mixing and detection happen simultaneously, within the same physical support and are no longer independent. The first demodulation pixels, precursors of the actual PMD pixels, should fulfill the following tasks, according to [278]: photoelectric carrier generation, fast separation of the carriers, repeated addition of the separated carriers in consecutive cycles and *in pixel* storage.

PMD pixels, originally produced in CCD technology, are currently produced in standard CMOS technology. Different PMD pixel structures have been presented and different physical implementations have been tested. Examples of optical implementations of PMD technology are [428, 403] the photo-gate PMD (PG-PMD), which was the first PMD implementa-

tion in CMOS technology and the one used in commercial PMD chips, the metal-electrode PMD (ME-PMD), the metal-semiconductor-metal PMD (MSM-PMD) and the micro-channel-plate PMD. In the following we briefly present these PMD implementations, focusing on the PG-PMD. Additionally, a non-optical implementation of the PMD pixels working with microwave radiation has also been proposed, named Microwave Mixer Device (MMD), which contributes to bridge the gap between radar and ToF imaging.

Regardless of the physical implementation, there are some common characteristics of optical PMD pixels. In general, they present a silicon substrate where charge carriers are generated from the incoming photons by photoelectric effect. One or several pairs of electrodes, typically isolated from the silicon substrate by an oxide layer, are used to generate an electrical field that is utilized to move the carriers to one of two storage areas. Normally, PMD pixels have a *finger* structure, i. e., they feature several pairs of electrodes and storage areas, one after another, being each element long and narrow, like a finger. This structure is designed to minimize the distance the carriers have to travel from the generation area to the storage area. This decreases the transport time and increases the modulation bandwidth of the pixel.

Photo-Gate PMD The photo-gate PMD (PG-PMD) was the first implementation of the PMD principle, dating back to 1997/98. The structure of the pixel can be seen as two diodes whose anodes are grounded and whose cathodes are the readout points of the two pixel channels. Therefore, the diodes are reverse-biased. The photocarriers are generated in a photosensitive area between the diodes. Over the photosensitive area a layer of silicon oxide (or, alternatively, Si_3N_4) separates it from the two conductive and transparent MOS photogates. Fig. 2.6 provides an schematic representation of a PG-PMD pixel (a) and its principle of operation with DC light (b) and modulated light (c).

The *push-pull* voltages applied to the photogates ($U_0 + u_m$, $U_0 - u_m$ in Fig. 2.6-a) modify the potential distribution in the surface and lead to what was called a *dynamic seesaw* [269] for the generated carriers, pushing them to one side of the pixel or another. In order to clarify the mixing process, suppose that the reference signal applied to the photogates is a square signal and the light arriving to the pixel surface is not modulated. Then, at the end of the integration time, the output voltages of the two readout nodes (U_a, U_b in Fig. 2.6-a) is the same, since, for each period, the photocarriers were integrated in the A channel half of the time and in the B channel the other half (Fig. 2.6-b). If the light is modulated, say according to the same reference signal applied to the photogates, then the difference of

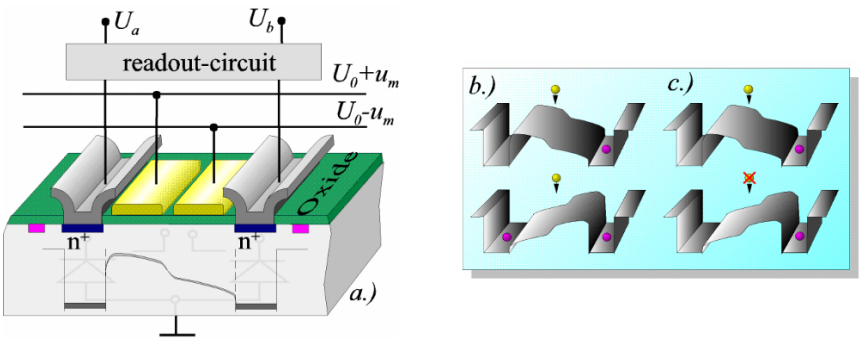


Figure 2.6.: Schematic representation of a photo-gate PMD pixel (a) and conceptual representation of the mixing process with DC (b) and modulated light (c). PMD pixels can be thought as two anode-grounded diodes, separated by a photosensitive region. A difference of potential between the transparent photogates (in yellow) induces an electric field that drives the photocarriers to one or another diode, where they are stored. The readout of each pixel channel is then performed at the cathodes of each diode. Images from [269].

potential between the two output nodes is directly dependent on the phase shift between the light signal and the reference signal applied to the pixel. It is clear that this difference provides a sample of the autocorrelation function of the reference signal (thus a triangular wave), at a phase given by the shift between light and reference signals. In the case of 0° or 180° phase shift, all the photocarriers should contribute to one channel and no one to the other (Fig. 2.6-c), i. e., a maximum or minimum of the autocorrelation function.

Depending on the voltage applied to the photogates, the diodes may operate in accumulation, depletion or inversion mode. In normal operation, they are in deep depletion mode, i. e., there is a deep space charge region or depletion region under the photogates [309] due to U_0. The modulation frequencies, typically ranging from tens to thousands of megahertz, are too fast to modify the deep depletion state, even for large values of the modulation voltage, $\pm u_m$. This region serves for fast charge transport.

Metal-Electrode PMD The metal-electrode PMD (ME-PMD) is similar to the PG-PMD, being the main difference the absence of the photogates. The transparent electrodes are substituted by the metal shields of the

readout electrodes, which are widened to that end. A diagram of an ME-PMD pixel is given in [403]. The main advantage of this implementation is a more homogeneous drift field, resulting in an increased modulation bandwidth. Additionally, the absence of photogates allows for narrowing the finger structure, yielding higher bandwidth. The reference signal is not only applied through the metallic shields of the readout electrodes, but also through the cathodes themselves, by means of a coupling capacitor that acts as a barrier for the DC component of the reference signal (U_0 in Fig. 2.6-a).

Metal-Semiconductor-Metal PMD The metal-semiconductor-metal PMD (MSM-PMD) exhibits a simpler structure than the ME-PMD. There are no dedicated modulation electrodes and both the modulation and the readout are carried out through contacts on metallic stripes with the characteristic finger structure. That is, the pixel is composed by metal stripes on negatively doped silicon, i. e., metal-semiconductor junctions or Schottky diodes, all over a grounded silicon substrate. Note the similarity with the ME-PMD pixel operation, in which the modulation was applied through the metallic shielding and the cathodes simultaneously. The simplified structure of the MSM-PMD requires a low pass filter per channel before the readout circuitry, in order to block the high-frequency modulation signal ($\pm u_m$). The modulation signal is applied through coupling capacitors like in the ME-PMD, which also serve to eliminate the photocurrent due to unmodulated BGL. Small-area Schottky diodes, such as those used in RF detectors, often operate at frequencies of tens of gigahertz. One of the main advantages of MSM-PMD pixels is, therefore, an outstanding bandwidth [269], yielding better depth resolution.

Micro-Channel-Plate PMD The micro-channel-plate PMD (MCP-PMD) [485] reminds the second generation (and posterior) image intensifiers, whose central component is an MCP at high voltage. The operation of EOM based on image intensifiers was already described at the end of of Section 2.2.2. The novelty of the MCP-PMD pixel is bringing the concept of the MCP-based image intensifier to a pixel architecture, instead of simply stacking the intensifier over a pixel array. As in conventional image intensifiers, electrons are generated when the light reaches the photocathode and travel to the MCP, due to the difference of portential between the photocathode and the metalized inner layer of the MCP, which acts as a first anode. The modulation signal is applied to this layer, causing the electrons to hit the MCP in a region destined to the readout A or another destined for the B. After the

electrons reach the MCP, they are conducted by finger-like microchannels where the electron multiplication takes place. Each microchannel is around 10 μm width and acts as a continuous-dynode electron multiplier. At the end of the microchannels the electrons from A-area microchannels and B-area microchannels are read in their corresponding readouts. In MCP-PMD pixels, as well as in conventional image intensifiers, voltages of more than a thousand volts between anode and cathode are required.

A related concept is that of the Photo-Multiplier-Tube PMD (PMT-PMD) [485], which is a special PMD pixel architecture including an individual PMT per pixel. The modulation signal is used to deviate the electron beam at the end of the tube to one of two anodes, which are the two channels of the pixel. The electrons are generated at the beginning of the tube, in a photocathode, as in conventional PMTs. Apart from clear concerns on the pixel size and the scalability of such a system, a problem that persists is the requirement of high differences of potential between the anodes and the photocathode.

An advantage of MCP-PMD and PMT-PMD is the large amplification of the light signal, which permits operation with very low optical power, low exposure or very large distances. The MCP-PMD could provide an amplification factor of 10^7 with 300 GHz [485] and maximum ranges of 1000 m [403].

Microwave Mixer Device (MMD) Silicon detectors are sensitive to radiation in the visible and NIR spectrum bands. Nevertheless, the PMD concept is general and can be applied to higher wavelength radiation, such as X-rays or microwaves, or even sound waves, as far as the conversion into electrical signals is possible. Once the carriers are generated, a mixing process can be carried out, as in PMD pixels. The so-called Microwave Mixer Device (MMD) is equivalent to the PMD, but using microwave radiation instead of light. There are clear advantages of using microwave radiation instead of optical radiation. One of them is the frequency of the radiation, from 300 MHz to 300 GHz, lower than that of optical radiation (wider unambiguous range in a pure interferometric setup), but high enough to achieve good depth resolution. Another advantage is the lower attenuation by meteorologic phenomena and the independence of daylight and background illumination. Unfortunately, a greater wavelength brings new issues, such as the lower lateral resolution and the requirement of large antenna apertures.

The high-frequency microwave carrier is modulated in amplitude by a lower-frequency signal, similarly to the optical case. An antenna is used to convert the received microwave signal into an electrical signal, substituting

the photoelectric conversion in the PMD. The charge separation, i. e., the core of the MMD pixel, is performed by means of a pair of Schottky diodes in antiparallel configuration, connected right after the antenna output. The modulation signal (the same used for amplitude modulation of the emitted microwave radiation, $\pm u_m$) is applied to the terminals of the diode pair. Clearly, this signal will block one of the diodes while driving the other to conduction. A block diagram of the MMD and a more detailed operation description is provided in [485], together with active and pasive pixel implementations. A sketch of a 3D-MMD camera using a Fresnel lens to map the scene onto a MMD pixel array is given in [403].

Dynamic Range and Background Illumination In general, PMD pixels are able to suppress the contribution of BGL by means of the so-called "correlated balanced sampling" (CBS), which is simply exploiting the fact that, for periodic reference signals of 50% duty cycle, the DC component of the light signal contributes equally to both pixel channels and, consequently, has no influence in the correlation value given by the difference $k(\theta) = D_\theta = I_\theta^{\mathrm{A}} - I_\theta^{\mathrm{B}}$. Nevertheless, this does not avoid saturation of the pixel channels if the light intensity is excessively high, e. g., under sunlight illumination. Consider the following classical definition of dynamic range of a pixel:

$$DR = 20 \log \left(\frac{I_{\mathrm{ceil}}}{I_{\mathrm{floor}}} \right) \tag{2.17}$$

where I_{floor} and I_{ceil} are the minimum and maximum values that the pixel (or pixel channel) can deliver. I_{floor} is given by the noise floor and I_{ceil} by the saturation level. The dynamic range of PMD pixels has been calculated in prior investigations and found to be of 73 dB for the one-tap PMD pixel in [278], 71 dB for the two-taps PMD pixel in [309] or 70 dB in [222].

An additional mechanism to suppress the disturbance of unmodulated light is the use of narrow bandpass optical filters, centered at the wavelength of the illumination source. Unfortunately, sunlight spectrum exhibits a considerable amplitude in the NIR region, often the band of choice for ToF systems. An intelligent selection of the illumination wavelength, e. g., making it coincide with the center of some atmospheric absorption band, together with a narrow bandpass filter at that wavelength can be of great help. In general, we seek a large ratio between signal amplitude and background level. The easiest solution is to increase the optical power of the illumination system, e. g., adding more emitters or increasing their power. This has a cost and increases the size of the ToF system. Alternatively, the emitters

may operate in burst mode. In burst operation, the light sources emit the same average power as in normal operation, but during a shorter time, i. e., with higher peak power. As a by-product of burst operation, fast-moving objects can be captured, due to the reduction of the exposure time.

Still, these methods do not increase the dynamic range of the PMD pixel itself. The amount of charge to be stored by some pixel channels might still be reaching the saturation threshold, while others are within the noise level. A typical case is saturation of few pixels due to background illumination. If the exposure time is reduced, saturation may be avoided, but at the cost of decreasing the SNR of the rest of the pixels. A mechanism is therefore necessary to artificially increase the dynamic range of the PMD pixels, in a local way and leaving the signal (difference between pixel channels) unaffected. The solution provided by PMD Technologies is the patented Suppression of Background Illumination system (SBI) [475, 74]. The system is a hardware solution, implemented as *in pixel* circuitry, which works independently for each pixel. The core idea of the SBI is the use of a compensation current to remove the charge due to uncorrelated light from the pixel. The system assures that exactly the same current is being removed from both pixel channels, keeping constant the difference. In principle, this means that the entire dynamic range of the pixel can be used to measure the correlated signal, leaving away the often-dominant background component. It has been shown that the SBI makes possible to operate in sunlight conditions up to 1.5×10^5 lux [387]. A model of the PMD pixels including SBI can be found in [398].

SBI operation presents several disadvantages. The intensity image cannot be recovered from the measurements, since the DC component is artificially decreased for those pixels where the SBI is active. If the current extracted from both pixel channels is not exactly the same, the SBI affects the measurements, producing a per-pixel bias, which translates into a noise-like pattern in the depth image.

The integration characteristic curves of PMD pixels are linear along the entire range of operation. The SBI introduces a severe non-linearity in the curve, changing from a positive slope to a zero slope in the pixel channel reaching saturation and a slightly-negative slope in the other. Other non-linear characteristic curves can be used to increase the dynamic range. In [11], linear, piecewise-linear, logarithmic and linear-logarithmic characteristic curves are compared in terms of dynamic range. A piecewise linear curve presents a point, for a certain charge level, from which the slope is lower than before. Logarithmic curves provide an increase of the dynamic range by a smoothly-decreasing slope. Linear-logarithmic curves combine the low

slope of a logarithmic ending with a constant high slope for signals with low optical power.

2.3.3. Depth Calculus and Performance Considerations

In this section we consider that the illumination signal exhibits a close-to-sinusoidal shape and, in consequence, the cross-correlation carried out in the PMD pixels is between a sinusoid and a periodic signal of equal fundamental frequency, which is equivalent to being between two sinusoidal signals of that frequency [304]. This is the case which is implicitly assumed in PMD cameras and it is supported by the fact that the LEDs, often used as illumination sources, show a low pass filter effect in the conversion of the electrical driving signal into optical at frequencies of tens of megahertz. This is due to the characteristic rising and falling times in the range of nanoseconds. If we neglect the asymmetry between rising and falling times, the effect the LEDs have over the illumination control signal (ICS), which is used to drive them, can be modeled as a convolution with a Gaussian kernel in time domain, $g(t)$, i. e., a smooth low pass filter. Note that modeling such asymmetry would require an asymmetric convolution kernel, thus, not Gaussian. This means that the final illumination signal can be defined as $i(t) = (ICS * g)(t)$ and follows Eq. 2.7. In consequence, the return optical signal is also sinusoidal (Eq. 2.9) and the measurements provided by the PMD pixels follow Eq. 2.12 in an ideal mixing process.

If we consider as measurements the difference of the two PMD pixel channels, we are correlating against a square signal of zero mean (no offset) and, in case of ideal mixing, there is no DC component in Eq. 2.12, since $A_0^q = 0$. In other words, measurements are insensitive to the DC component of the light received by the pixel, A_0^r. Thus, only two of these differential measurements suffice to estimate the remaining parameters, namely, the amplitude and the phase shift due to depth, θ_{depth}, in case of ideal modulation and demodulation. Let $D_{\theta_k} = I^{\text{A}}_{\theta_k} - I^{\text{B}}_{\theta_k}$ denote the differential measurement obtained as difference of the PMD pixel channels (A and B) for a sampling point $\tau_k = \theta_k/\omega$ of the cross-correlation given in Eq. 2.12. Despite only two sample points suffice, in PMD systems typically four measurements at four different phases are gathered. Gathering more measurements reduces the effect of harmonics and, indirectly, also the effect of noise of random nature due to the summations in both numerator and denominator in Eq. 2.5. In order to simplify the depth calculus, the four phases are selected to be uniformly distributed in phase domain, i. e., the shifts are $\theta \in \{0°, 90°, 180°, 270°\}$ of, equivalently, the samplings points are given by $\theta_k = \frac{k-1}{N}360,\ k \in [1, N]$, for

$N = 4$ samples. Note that this sampling scheme is, in the case of an ideal mixing process, actually equivalent to acquiring only two samples, since $D_\theta = -D_{\theta+180^\circ}$, $\forall\theta$, by definition (Eq. 2.12). The four phases schema allows further simplification of Eq. 2.6 [332] due to the trivial values of sines and cosines at these phases, yielding

$$\theta_{\text{depth}} = \arctan\left(\frac{D_{270^\circ} - D_{90^\circ}}{D_{180^\circ} - D_{0^\circ}}\right) \tag{2.18}$$

which, in combination with Eq. 2.1, yields a simple formula, known as the *four phases algorithm* [330, 388], which directly computes the depth from the raw measurements:

$$d = \frac{c}{4\pi f_{\text{mod}}} \arctan\left(\frac{D_{270^\circ} - D_{90^\circ}}{D_{180^\circ} - D_{0^\circ}}\right) \tag{2.19}$$

The amplitude and the DC offset can also be estimated from the measurements. The amplitude, which refers to that of the cosine in Eq. 2.12, can be estimated from measurements of one of the channels or from differential measurements. In the former case the reference signal has an offset of half of the amplitude, while in the latter case there is no offset. Note that in the latter case the amplitude is scaled by a factor of two, due to the corresponding double amplitude of the reference signal (A^q). The offset cannot be estimated from differential measurements, since it does not contribute to them, and has to be calculated from the measurements of one of the pixel channels. The formulas are derived from Eq. 2.12 and given by Eq. 2.20 and Eq. 2.21. The proofs of Eq. 2.18, Eq. 2.20 and Eq. 2.21 are given in Appendix A.3.

$$\begin{aligned} A &= \frac{\sqrt{(I_{270^\circ} - I_{90^\circ})^2 + (I_{180^\circ} - I_{0^\circ})^2}}{2} \\ &= \frac{1}{2}\frac{\sqrt{(D_{270^\circ} - D_{90^\circ})^2 + (D_{180^\circ} - D_{0^\circ})^2}}{2} \end{aligned} \tag{2.20}$$

$$A_0 = \frac{I_{0^\circ} + I_{90^\circ} + I_{180^\circ} + I_{270^\circ}}{4} \tag{2.21}$$

Demodulation Contrast A measurement of the performance of any kind of *smart* pixel performing an *in pixel* demodulation is the so-called demodulation contrast. This complements and is based on the concept of modulation depth (or contrast) of the light source, which measures the

quality of the modulation process. For completeness, consider the following classical definition of Michelson contrast [326]:

$$k_{\text{mod}} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{2.22}$$

where $I_{\min}$ and $I_{\max}$ are the minimum and maximum light intensities provided by the light source. Obviously, an ideal modulation provides a contrast of 1. Note that the numerator is the amplitude of the signal, while the denominator represents twice the average, i. e., twice the offset of a periodic illumination signal. Similarly, the demodulation contrast is defined as the ratio between measured amplitude and measured offset of the light signal. For any multitap pixel that works in a four-phases basis, the amplitude and offset can be obtained from Eq. 2.20 and Eq. 2.21, respectively. For the specific case of a PMD pixel, with two integration channels denoted by A and B, we can write [330]

$$k_{\text{demod}} = \frac{\text{measured amplitude}}{\text{measured offset}} = \frac{|I_{\text{A}} - I_{\text{B}}|}{I_{\text{A}} + I_{\text{B}}} \tag{2.23}$$

where I_{A} and I_{B} are the readouts of the corresponding pixel channels when the overall phase shift between the electrical signal controlling the integration and the optical signal as received at the pixel surface is 0° or 180°. The demodulation contrast quantifies the ability of the pixel to perform charge separation according to the reference signal. Typical values of this parameter are 50% for two-tap PMD pixels (at 20 MHz modulation frequency) or 60% for three-tap PMD pixels. Despite it is often omitted in the literature, this measurement is a meaningful descriptor of the performance of the pixel only when gathered using a light signal with modulation contrast of 1. Otherwise a composed measurement is obtained, which in most practical cases is dominated by the signal-to-background-light ratio. The total contrast of a ToF system is given by $k_{\text{total}} = k_{\text{mod}} k_{\text{demod}}$.

Other Operation Modes: Heterodyne Mode and Pseudonoise Modulation As already introduced in Section 2.2.2, homodyne operation is not the only possible modulation-demodulation framework in phase-shift-based ToF systems. PMD devices can be tuned to operate in heterodyne mode [117] or using pseudonoise (PN) modulation [75].

Operating in heterodyne mode brings the need for accurate frequency shifting and a perfect synchronization of the acquisitions with the low-frequency beat signal. This also means ensuring that an integer number

of acquisitions are gathered per period of the beat signal. The phase shift calculus in heterodyne mode can be performed in an equivalent way to homodyne mode, using a formula which is equivalent to Eq. 2.6, but actually sampling the beat signal in time, instead of sampling the cross-correlation function at different phase shifts of the reference signal, as in homodyne PMD. The heterodyne operation makes easier to acquire a custom number of equally-spaced measurements and reduce linearity errors, caused by the presence of unexpected high order harmonics in the modulated light.

The use of PN modulation in PMD sensors has been extensively studied in [75], motivated by the spread-spectrum techniques, which had been already applied in communications time before. Originally used for military communications in the 1950s, these techniques are nowadays used everywhere in civil wireless networks. In short terms, spread-spectrum methods deliberately spread a signal, which might be bandlimited in principle, along a wide frequency band. An immediate advantage is the improvement of the resistance to interference and intentional jamming. Additionally, these techniques allow using the same channel to transmit different information to different receivers simultaneously, that is, allow multiple access without the need of time multiplexing. This is known as the code-division multiple access (CDMA) or code-division multiplexing (CDM). It is clear that, if orthogonal codes (e. g., obtained from the columns of a Hadamard matrix) are used as carriers to send the simultaneous messages, no interference will occur, if each receiver decodes the message using the right code. Similarly, in ToF imaging, using orthogonal PN sequences for illumination modulation ensures interference-free simultaneous operation, which can not be achieved in single-frequency homodyne operation. If the codes are long enough, any two different random codes can be considered orthogonal. Additionally, the sharp autocorrelation function of PN codes ensures that accurate phase measurements are possible.

Physical Considerations Commercial PMD sensors feature a daylight blocking filter, whose mission is to block all light except from that of the wavelength used for the modulated light. The filter is a low-pass filter, with transmittance close to 100% for wavelengths higher than 830 nm. That is, PMD sensors are expected to operate with NIR light. This may profit from the spectral response of silicon, which is higher in the NIR than in the visible range. Additionally, working with relatively large wavelengths reduces the thermal diffusion in the silicon and, consequently, increases the accuracy of the measurements. Large wavelengths imply lower excess of energy in the photoelectric effect. Such excess of energy generates heat in the

silicon, increasing its temperature. For completeness, consider the following equations. For a certain wavelength of the radiation, λ, the excess of energy is given by Eq. 2.24,

$$\Delta E_{\text{th}} = \frac{hc}{\lambda} - E_g \tag{2.24}$$

where ΔE_{th} is the energy excess, destined to generate heat, and E_g is the band gap energy of the photosensitive material, silicon in this case. This excess should be as close to zero as possible. The wavelength for which the right side of Eq. 2.24 is zero is the lowest that will still produce photoelectric effect in silicon and it is around 1.1 μm. The equipartition theorem of the energy states that, once thermal equilibrium is achieved, energy is shared equally among all the particles in the system. This means that the total energy can be formulated as the sum of the kinetic energies of all particles moving at a certain average speed. Per particle, we obtain

$$\frac{1}{2} m_{\text{eff}} v_{\text{th}}^2 = \frac{3}{2} k_B T \tag{2.25}$$

where m_{eff} is the effective mass of the charge, v_{th} the thermal speed, k_B the Boltzmann constant and T the absolute temperature. An increase of the total energy in ΔE_{th} will add to the right side of Eq. 2.25, rising the temperature, and leading to an increase of v_{th}, $\Delta v_{\text{th}} = \sqrt{\frac{2\Delta E_{\text{th}}}{m_{\text{eff}}}}$. We are interested in the diffusion length, which is the average distance that the carriers travel until they recombine, i. e., the distance corresponding to its lifetime, τ_n, which is given by

$$L_{\text{diff}} = \sqrt{D\tau_n} \tag{2.26}$$

where D is the diffusivity of the material. Provided that D depends linearly on v_{th}^2, the energy excess, ΔE_{th}, produces an increase of the diffusion length which depends linearly on $\sqrt{\Delta E_{\text{th}}}$. Additionally, D and E_g strongly depend on the temperature, so that, the higher the temperature, the higher the resulting increase of the diffusion length. Now consider the following general definition of modulation transfer function (MTF):

$$MTF(f) = \left| \frac{T(f)}{T(0)} \right| \tag{2.27}$$

where $T(f)$ is the frequency transfer function of the system. In control theory transfer function refers to temporal frequency, but in this case it refers to spatial frequency, in the pixel domain. Eq. 2.27 describes the ability

to capture spatial information. Supposing that the pixel size is small enough to consider negligible the corresponding spatial quantization error, Eq. 2.27 yields an upper bound for the lateral resolution of the ToF imager.

A strong diffusion favors crosstalk and degrades the MTF [421]. The quantum efficiency (QE) of a photosensitive element is defined as the ratio between the number of photogenerated electrons effectively collected by the element and the number of impinging photons, under uniform illumination. This ratio strongly depends on the wavelength of the light, decreasing for high wavelengths, and on the diffusion length, approaching to one for large diffusion lengths. The existence of upper layers over the depleted silicon area produces interference between shifted wavefronts, since the order of magnitude of the layers and the radiation wavelength are approximately the same. This translates into a fringe pattern in the QE in wavelength domain. Depending on the relationship between the thickness of the upper layer and the wavelength, a local minimum or maximum of QE can be achieved [421]. Thermal diffusion might also increase, as a side-effect, the effective active area of the pixels. Electrons generated out of the space charge region are, in principle, not affected by the electrical field responsible for charge transport. Nevertheless, if they diffuse into the space charge region, they can be separated by the electrical field and pushed to the corresponding storage area.

The other key parameter in Eq. 2.24 is τ_n, which depends, in turn, on the feature size of the opto-electronic processes.

Depth Measurement Uncertainty Several sources of noise exist in PMD sensors and, in general, in any image sensor. Some of the most important noise sources [434] are the optical shot noise, the thermal noise, flicker noise and the quantization noise. A rule of thumb provided in [411] says that 1 pJ of accumulated NIR photon energy in a pixel allows the measurement of the distance to a precision of around 1 cm at 20 MHz modulation frequency.

Shot noise, or quantum noise, is due to the fact that the process of photon arrival and electron generation follows a Poisson distribution, which is a limit case of the binomial distribution, when the number of tries n tends to infinity and the probability of success tends accordingly to zero, yielding a finite number of successes. A discrete random variable following a Poisson distribution, $X \sim P(\lambda)$, is characterized by a single parameter, λ, which gives both the expected value and the variance, i. e.: $\mathrm{E}(X) = \mathrm{Var}(X) = \lambda$.

The thermal noise, also known as Johnson-Nyquist noise, is the electronic noise due to thermal agitation of the charge carriers in the silicon. Thermal noise approximately follows a zero-mean white Gaussian noise distribution,

with a power spectrum that is flat up to 1 THz [29] in resistive regions, such as MOS channels in strong inversion.

Flicker noise is also known as $1/f$ noise or pink noise due to its *pink* power density spectrum, i. e., decaying with frequency. This noise is found in many fields and is common in natural phenomena, where the systems tend to be more prone to low-frequency variations than violent changes. In electronics, it is related to a resistance fluctuation, which produces voltage or current variations. Due to the pink power density spectrum, at high frequencies it may be eclipsed by thermal noise.

Quantization noise is ubiquitous in electronics, since it appears there where an analog-to-digital conversion is performed. Quantization consists in mapping a large set of values into a countable smaller set. Such process also takes place in data compression, being the quantization noise the signal distortion due to the eventual information loss during compression. From a general point of view, if the discrete samples of the signal are considered measurements,—e. g., using an ideal Dirac delta function as sensing kernel,—the recent theory of Compressive Sensing (shortened CS, see Chapter 3) assures that exact recovery of the signal in a compressed domain is achievable from the measurements if the signal admits a sparse or compressible representation in that domain and its basis is incoherent with the sensing kernels. This holds even if the measurements are incomplete and affected by quantization noise, as far as the number of measurements is high enough for the given sparsity level.

Noise models for amplitude, phase and offset have already been provided in [335]. We are interested in determining how noise in the pixel readouts affects the final depth estimation. The relationship between the depth value and the measurements in a conventional PMD ToF camera is given by Eq. 2.19. Consequently, it suffices to apply error propagation to Eq. 2.19 to obtain the influence of measurement noise in the depth estimation. The result is provided in Eq. 2.28, where A is the amplitude, as defined in Eq. 2.20. The proof is given in Appendix A.4.

$$\Delta d = \frac{c}{4\pi f_{\text{mod}}} \frac{1}{(2A)^2} \sqrt{\begin{aligned}&(I_{270^\circ} - I_{90^\circ})^2(\Delta^2 I_{0^\circ} + \Delta^2 I_{180^\circ}) \\ &+(I_{180^\circ} - I_{0^\circ})^2(\Delta^2 I_{90^\circ} + \Delta^2 I_{270^\circ})\end{aligned}} \tag{2.28}$$

The depth uncertainty given by Eq. 2.28 is general and valid regardless of the value of the measurements, I_θ, and the nature of the measurement noise, ΔI_θ. A realistic hypothesis is that the measurement noise is dominated by

the shot noise, i. e., that follows a Poisson distribution. In that case we have that the variance is equal to the expected value and the standard deviation is $\Delta I_\theta = \sqrt{I_\theta}$. This, together with the fact that the sampling points are multiples of 90°, can be used to simplify Eq. 2.28, yielding Eq. 2.29, the proof of which is also given in Appendix A.4.

$$\Delta d = \frac{c}{4\pi f_{\text{mod}}} \frac{\sqrt{2A_0}}{2A} \tag{2.29}$$

Eq. 2.29 is equivalent to Eq. 4.11 of [278] and coincides with Eq. 10 of [411]. Other noise sources different from Poisson noise are signal-independent and can, therefore, be added to the DC component of the measurements, A_0.

Eq. 2.29 clearly shows that in cases when shot noise is dominant, e. g., due to poor illumination, the depth uncertainty is given by the ratio between the square root of the DC component and the amplitude of the cross-correlation function, which is, in turn, proportional to the amplitude of the illumination signal. If the background illumination is low, the DC offset is mainly due to pixel noise (e. g., thermal, flicker or reset noise). Therefore, it is clear that the pixel sensitivity plays a key role in the achievable depth accuracy. At the chip level, the sensitivity of a PMD sensor can be increased with larger pixel sizes or higher fill factor. If no further improvement can be achieved in this regard, the use of microlenses on the pixels to collect more light may be an alternative. The amount of photo-generated electrons also depends on the optical system. Lenses with large apertures allow collecting more light and getting a higher response in the pixels. It has been observed that, in most cases, the accuracy of the depth measurements improves linearly with the aperture, but in low-light cases, i. e., when the noise floor due to dark current and quantization is dominant, the accuracy increases quadratically with the aperture [330]. In [278, 309] an equation is provided that calculates the optical power received by a pixel in the case of Lambertian reflection, which we rewrite here as Eq. 2.30:

$$P_{\text{pixel}} = \frac{A_{\text{pixel}}}{A_{\text{image}}} P_{\text{source}} \rho \left(\frac{D_{\text{lens}}}{2r} \right)^2 k_{\text{lens}} \tag{2.30}$$

where A_{pixel} is the active area of the pixel, A_{image} is the area of the image in the sensor plane, P_{source} the optical power of the light source, ρ the reflectivity of the target, r the distance between the lens and the target, D_{lens} the aperture of the lens and k_{lens} a factor accounting for the optical losses of the lens system. Note that Eq. 2.30 relies on several assumptions that might not match the reality. First, it implicitly supposes that all the

optical power of the light source is somehow transferred to the target surface, being the power reflected by the target $P_{\text{object}} = \rho P_{\text{source}}$. Additionally, it assumes that the whole scene observed by the lens is a target whose points are all at a distance r from the lens and share the same reflectivity. This might be a valid assumption for small FOVs and large observation distances, so that the observed object area is small and the reflectivity can be considered uniform, while the relative depth variations are much lower than the average distance r. Nevertheless, in general, the power received by a pixel depends on the reflectivity and distance of the object point being observed by that pixel, which might be very different from those of points observed by other pixels of the array. For this reason, we derive here a more complete formula that accounts for the light propagation from the source to the scene and calculates the power received by the pixel directly in a per-pixel basis, avoiding the need for the ratio $\frac{A_{\text{pixel}}}{A_{\text{image}}}$ to obtain the per-pixel power from the total average power received at the image plane, as done in Eq. 2.30.

$$P_{\text{pixel}} = \frac{A_{\text{pixel}} \rho k_{\text{lens}}}{8\pi (rf\#)^2 \left(1 - \cos\left(\frac{FOV_{\text{source}}}{2}\right)\right)} P_{\text{source}} \tag{2.31}$$

where r and ρ denote now the average distance and reflectivity of the object area captured by the pixel, respectively. FOV refers here to the FOV of the illumination system, which might differ from the FOV of the camera. $f\#$ denotes the f-number of the lens. The derivation of Eq. 2.30 can be found in [278], while the proof of Eq. 2.31 is given in Appendix A.5.

The number of photogenerated electrons can then be calculated from the per pixel power obtained from Eq. 2.31 as:

$$N_{e^-} = \frac{E_{\text{pixel}}}{E_{\text{photon}}} QE(\lambda) = \frac{(P_{\text{pixel}} t_{\text{exp}}) QE(\lambda)}{\frac{hc}{\lambda}} \tag{2.32}$$

where t_{exp} denotes exposure time and $QE(\lambda)$ is the quantum efficiency for the light wavelength λ.

2.3.4. Typical and New Applications

Since the advent of the PMD technology, the targeted application was phase-shift-based ToF depth imaging. Depth images of simple objects obtained with early low-resolution PMD arrays are given in [310]. A milestone in this regard was the development of the first ToF camera based on *smart* pixels [279], which used one-tap pixels, similar to the posterior two-tap PMD

pixels. A real time 3D video camera featuring a PMD sensor is presented in [408]. The number of pixels of this first prototype was 16×16, but new designs up to the current 120×160 resolution were already contemplated in [408]. In that work it was suggested to combine the information from a 3D-PMD videocamera with that from a 2D-RGB videocamera with the same field-of-view (FOV). Similarly to [408], the PMD arrays have been presented as a multichannel lidar in [407], where the idea of a 3D camera with synchronized 2D-RGB image fusion is also present. The system presented in [216], and whose possibilities are thoroughly investigated in [304], fuses a PMD-based 3D camera and an RGB 2D camera into a single system. The camera, developed at ZESS, is called MultiCam [377] and allows both sensors to share the same optical path by means of a Bauernfeind prism with an integrated beam splitter, which lets the NIR light reach the PMD chip, while redirecting the visible light to the RGB sensor. The monocular setup allows achieving a perfect registration between the color and depth modalities, which is not possible in most commercial RGB-D sensors, such as the Kinect sensor. The latest model of the MultiCam implements a PMD 19k-S3 chip [374] (120×160 pixels) for ToF imaging and an Aptina three-megapixel color CMOS chip for RGB imaging. For further details the reader is referred to [280], where a MultiCam is used in a wide-area 2D/3D imaging system.

A linear 3D-ranging system based on the MSM-PMD pixels is presented in [76]. The large bandwidth of MSM-PMD pixels (up to 1 GHz) is exploited to generate PN modulation patterns. The same PN sequence is used for modulation and demodulation, leading to an autocorrelation function with a single sharp maximum. This allows accurate and unambiguous depth estimation, at the cost of increasing the number of acquisitions. If the method for finding the autocorrelation maximum is by regular sampling, the number of samples grows linearly with the depth resolution requirements. The system achieved depth standard deviation below 4 mm.

An appealing application, also considered from the origins of the PMD technology, is its use for optical communications [77, 75]. Arrays of PMD pixels can be used to implement CDMA in optical communications, allowing multiple users to use simultaneously the same physical medium without mutual interference by means of orthogonal PN codes. The idea of using the NIR for wireless communications has become of interest in recent times due to the saturation of the traditional radio bands and the comparatively low interference in the optical range. The development of solid state lighting technology, able to switch at high frequencies, is another factor that favors optical wireless for indoor communications [165]. In this regard, a system allowing the use of LED indoor lighting systems both for visible illumination

and wireless communications has been made commercially available under the name of *pureLiFi*. The core idea is to use conventional white LED lamps as a multiple access communication system [164] and is backed by the staggered prohibition of incandescence bulbs for household use, which is expected to lead to the omnipresence of LED illumination in the domestic domain. Either using orthogonal frequency division multiplexing (OFDM) modulation or CDMA, PMD arrays are an excellent candidate for multichannel reception and decoding of the light signal in a MIMO configuration.

In the last years, the development of transient imaging techniques [264] has opened a new field of application for PMD pixels. In principle, transient imaging aims to offer a unique combination of outstanding spatial, angular and temporal resolutions. In principle, conventional cameras offer a very strict tradeoff between temporal and spatial or angular resolution. In most cases, high spatial and angular resolutions are only achieved at rates below 100 Hz. Transient imaging aims keeping the spatial resolution of conventional cameras, angular resolution as high as that of phased arrays, eventually allowing to *look around corners*, while pushing the temporal resolution down to the femtosecond range. Consequently, these techniques require a laser source operating in pulsed mode in the femtosecond range and fast photodetectors. Alternatively to the fast photodetector, a CCD camera is used in [445], in combination with an external setup that reminds an MCP-based image intensifier, with additional electrodes at the sides of the tube to change the angle of the electrons generated in the photocathode. With this setup, which is close to a phase-shift-based ToF camera with external demodulation, the time profile of the returning light signal has been recovered, which allows *seeing* how the light propagates through the scene. In [248], a PMD chip is used as low-cost substitute of the CCD camera with external electron-sweeping system in a ToF system in order to deal with multipath interference. The use of a PMD sensor for transient imaging is proposed in [220], which allows sensing the surfaces of different translucent objects occluding each other. Further work in this direction has been recently presented in [221], where the sparsity of the vector of scatters in depth (i. e., propagation time) domain is exploited to achieve depth imaging in scattering media, such as non-transparent water.

2.4. Current Limits of the PMD-Based Time-of-Flight Imaging

2.4.1. Depth Accuracy

The depth accuracy of a PMD-based ToF imaging system is not directly given by Eq. 2.29. The uncertainty given by Eq. 2.29 derives from propagating the noise expected in the raw measurements to the depth measurement, according to a certain noise model. This provides a bound on the precision of the depth measurement, which mainly depends on the modulation frequency, the amplitude of the modulated signal and the DC offset. An experimental evaluation of the dependency of the depth standard deviation and accuracy on, e. g., modulation waveform, modulation frequency, exposure time, illumination geometry, etc. is given in [284], where an optimal modulation frequency of 45 MHz was found for LED-based illumination. The accuracy involves both the aleatory and the systematic error. If the real signals or the system do not follow their respective models, a systematic error will appear, which is not of random nature, but system- and often also signal-dependent. Indeed, in real PMD ToF cameras, depth errors in the centimeter range are often registered, even under good illumination conditions, low BGL and modulation frequencies of tens of megahertz. Nevertheless, in these cases the main component of the error is of systematic nature and, therefore, cannot be predicted with Eq. 2.29 or eliminated via temporal averaging. The standard deviation of the measurements is, in favorable cases, very low, in the millimeter range, close to the uncertainty predicted by Eq. 2.29. That is, PMD cameras exhibit good precision but often poor accuracy.

Wiggling Effect A method to improve the depth accuracy is to perform a calibration of the system. The easiest way is to perform extensive experiments acquiring the depth of reference objects at known distances and to generate a look-up table (LUT) establishing correspondences between measured distance and real distance [251]. This way, the true depth can be calculated from the measured depth by interpolation. This method requires an amount of memory that grows with the desired calibration quality and, conceptually, solves the problem by *brute force*. A more elegant solution is that provided by [300], which takes profit of the oscillating shape of the depth error curves to fit a cubic B-spline to it. Retrieving the real depth only requires evaluating a cubic polynomial.

If the systematic error is due to irregularities in the modulation/demodulation process, i. e., a non-sinusoidal modulation of the light plus a non-sinusoidal demodulation at the PMD pixel, then the well-known *wiggling* effect will appear in the depth measurements. This effect is due to the non-null contribution of the harmonic content of the light signal to the measurements and consists in a slightly non-linear relationship between real depth and measured depth, which shows oscillations, as observable in Fig. 2.3b for the ToF sensors analyzed (all except from the first Kinect sensor). The frequency of the wiggling phenomenon is given by the frequency of the harmonic of greatest amplitude, typically the second or the third, depending on the signal shape. This kind of depth error can be completely eliminated by calibration. Nevertheless, a calibration does not eliminate the causes of the problem, which are well-known, but it only palliates its effects on the depth estimation. The phase retrieval method used by PMD cameras (Eq. 2.19) is derived from the hypothesis of sinusoidal modulation of the illumination or, conversely, perfect sinusoidal reference signal at the PMD pixel. If this is not the case, the phase estimation method is to be adapted to the real situation, even when it comes at the cost of increasing complexity. Otherwise, part of the power of the illumination and the reference signals will be wasted in distorting the cross-correlation values, which will generate large systematic errors that have to be removed afterwards by calibration. This concerns the efficiency of the system. Part of the limited dynamic range of the PMD pixels is wasted in storing the undesired contribution of high-order harmonics.

Illumination Irregularities In addition to the wiggling effect, there is a number of other issues that translate into depth errors. If many light sources are used to illuminate the scene, at each point of the scene the light signal is given by the sum of the light signals coming from several, if not all, light sources. That is, the signal shape, apart from being non-sinusoidal, is 3D-space dependent. Consequently, if it is desired to eliminate the subsequent depth error by calibration, such calibration has to be performed independently per pixel for a given FOV. If the optics of the camera are changed, the per-pixel depth calibration is to be repeated. Alternatively, if the FOV is reduced, the new map of calibration parameters can be generated from the existent by interpolation, but if the FOV is increased, obtaining the new calibration map by extrapolating the original one might not lead to a correct calibration.

In medium and large range ToF systems, many high-power LEDs are often used to build the illumination system. In such systems, special care is put

in synchronizing the LED modules. The tracks that conduct the signals in the circuits for signal amplification and LED driving are of equal length, ensuring that all signals are equally delayed. The same applies to the wires that transport the signals between circuits. An extensive analysis of the medium-range illumination system presented in [283] has shown that, even equalizing the signal paths and using LEDs and auxiliary components of the same model, perfect synchronization is not achieved and relative phase shifts appear between the light signals emitted by different LEDs. At a modulation frequency of 20 MHz, the time delay induced by the illumination system is determined for each LED. The average of delay is around 6 ns, which means a depth offset of 90 cm. This kind of signal-propagation depth offset is always present in any ToF system and does not lead to depth errors because it is removed by a simple linear depth calibration. What generates depth inaccuracies is the slightly different delays registered for different LEDs, which is, on average, 0.66 ns. This asynchronism, which might be thought negligible at a first glance, means 9.9 cm in depth. The details of the study are omitted here for brevity, but a brief summary with the main results is provided in Appendix A.6. The asynchronism, together with slight variations in the waveform, lead to a unique depth error distribution in 3D-space, where the error does not only depend on the depth itself, but also on the 3D location of the scene point for which that depth was measured. Two points of the scene being at the same distance from the sensor can give rise to depth measurements that are, not only different from the real depth, but also different from each other, with differences lying in the centimeter range. The required calibration would have to be three-dimensional, adding two more dimensions to the depth itself, e. g., elevation and azimuth, in a polar coordinate system. Supposing that the complex illumination system is fixed and the camera can be displaced, if only the central pixel is considered, the system required to carry out the data collection for such calibration cannot be a 1D translation unit, such as that used in [251] and [282], but a 3D one.

The depth errors related to non-sinusoidality of the waveforms and inhomogeneous illumination, not just in terms of intensity, but also in terms of signal shape and delay, can be calibrated up to a certain extent. Other system-related error sources cannot be calibrated, unless additional sensors are included in the camera. For instance, the dark current, which can be included in Eq. 2.29 in the offset term, A_0, increases dramatically with the temperature, doubling its value each 8 °C, according to [330]. LEDs are also temperature-dependent and, in consequence, the shape and delay of the illumination signal with respect to the ICS change with the LED

temperature, which depends on the ambient temperature and the time the LED has been in operation.

Interference Between Optical Signals

Simultaneous operation Other external source of depth error is the interference between the reflected optical signal and other unexpected modulated light signals. An example of such situation is the simultaneous operation of two or more PMD cameras with overlapping FOVs. In this case, methods such as temporal division, wavelength division, frequency division or code division have been analyzed in [304] to avoid degradation of the depth accuracy.

Temporal division simply consists in acquiring at different times and, therefore, totally avoids interference. The main advantage is also a major inconvenient, since it makes simultaneous acquisition impossible. Additionally, a synchronization mechanism is required. This method also makes the frame rate inversely dependent on the number of cameras sharing FOV.

Wavelength division does not require any synchronization, since it consists in making each camera operate with light of different wavelengths. This requires a coherent modification of both illumination systems and camera filters. In order to avoid any overlapping, relatively well-spaced wavelengths in combination with narrow filters are to be used. If the number of cameras is large, the NIR region of the spectrum might not suffice and visible wavelengths may have to be used, which might be incompatible with the application. As already commented in Section 2.3.3, the spectral response of the silicon is not flat and using short visible wavelengths may compromise the sensitivity.

Frequency division consists in using orthogonal frequencies for the different cameras. The cross-correlation between periodic signals of orthogonal frequencies is zero, so the interference is eliminated. Unfortunately, in real systems the reference signal at pixel level is not sinusoidal, but a close-to-square (trapezoidal) signal, which contains more than a single frequency. For this reason, frequency multiplexing might still produce distortion in depth measurements unless the frequencies are largely different. Apart from that, if each ToF camera is operating in single-frequency mode, the use of different frequencies will lead to different unambiguous ranges and accuracies for the different cameras. Alternatively, one could use very closely-spaced frequencies, using integration times that are higher than the period of the corresponding beat signal, in order to get an amplitude that is independent

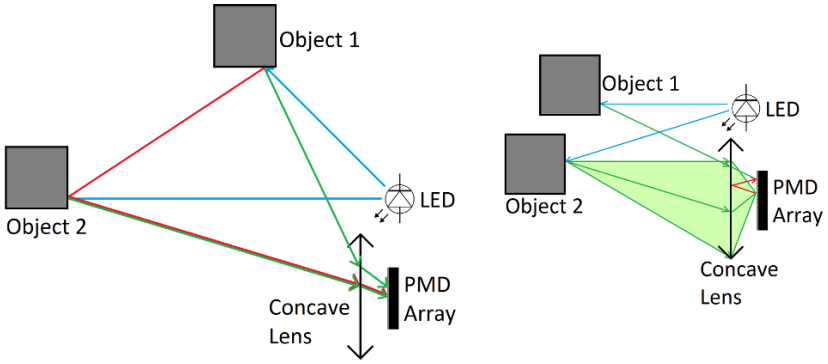
from the starting time. In [304, 302] frequency differences of several kilohertz are used to achieve simultaneous operation of several MultiCams.

Of special interest is the case of PN modulation [75], using different orthogonal codes for different cameras at a time [67]. This method does not impose any restriction on the acquisition, apart from ensuring that different codes are used for modulation in different cameras at the same time. As well as in the case of frequency division, the orthogonal illumination signals lead to a null contribution in the cross-correlation measurement, obtained as difference between pixel channels, but increase the BGL level measured by both pixel channels. This means wasting part of the dynamic range of the pixels.

Multipath Another example of unexpected interference between two or more optical signals is the multipath effect. Multipath interference (MPI) can easily cause depth errors ranging from tens of centimeters to meters in conventional ToF cameras. In short terms, the multipath effect occurs when a single pixel receives a combination of two or more reflected optical signals, which are not in phase. This can occur due to strong secondary reflections, caused by a shiny floor or walls, especially in corners [208]. A schematic representation of the effect is depicted in Fig. 2.7a. As a result, the depth measurement will be different from the real depth, the error being given by the relationship between the amplitude of the principal reflection and the amplitude(s) of the secondary reflection(s). This effect is scene-dependent and, therefore, could be seen as a *chicken and egg* problem, where the multipath cannot be corrected unless the scene structure is known and the scene structure cannot be retrieved correctly due to MPI. For this reason, most multipath-removal methods are optimization problems that jointly estimate the scene structure and the multiple paths followed by the illumination beam.

Another kind of multipath is light scattering, due to low quality optics in combination with objects that are too reflective or to close to the camera. As a consequence, the rest of the scene appears closer to the camera. In [192] a method was proposed to estimate the MPI in ToF cameras from the MPI-affected depth measurements and correct them. The method deals with diffuse multipath and was extended in an iterative fashion to deal with scenes with non-constant reflectance in [193]. Another iterative method for diffuse multipath is presented in [243], consisting in sequential scene reconstruction and ray tracing steps. In each iteration, the model of the scene is updated, until the error between the original depth map and that obtained from the scene model by multipath-affected ray tracing is lower

than a threshold. Despite the impressive error reduction (e. g., around 70% in [243]), these methods [192, 193, 243] have runtimes of one to ten minutes per frame and deal exclusively with diffuse multipath, arising from perfect Lambertian reflectors. Multipath due to scattering depends also on the pixel design. Pixels with large metal-shielded areas (exhibiting high reflectivity), in combination with a filter or lenses without anti-reflective coating, lead to internal reflections that are equivalent to secondary paths. This case, together with the general diffuse multipath due to Lambertian reflectors, is illustrated in Fig. 2.7b.



(a) Reflective Multipath **(b)** Diffuse Multipath

Figure 2.7.: Schematic representation of multipath interference in a ToF camera. (a): Two-path interference due to secondary reflection. The light rays coming from the light source (LED in the figure) are depicted in blue. The expected reflections, collected by the camera lens are in green. Object 2 receives light from the source but also from Object 1, generating a second reflection, delayed with respect to the first. This unexpected second path is depicted in red. (b): Diffuse multipath. Even in absence of clear secondary paths as in the left schema, diffuse multipath arises from objects whose surfaces behave as Lambertian reflectors. This means that the light collected by the lens for a single scene point undergoes slightly different paths, with slightly different lengths. Additionally, if Object 2 is very reflective and close to the camera, most pixels of the array will be receiving a certain contribution of the light reflected by it. The strength of the effect depends on the quality of the optical system. Optical elements without anti-reflective coating might generate internal reflections (in red).

In the case of highly-reflective or specular (non-Lambertian) surfaces, several modulation frequencies are often used in order to estimate more than one depth per pixel in a closed form, without the need of scene reconstruction and ray tracing. This is the case of the method in [261] and the first method presented in [203]. Alternatively, the multiple paths can be retrieved through an optimization process, as in [153]. The methods in [261, 203, 153] cope with only two paths per pixel. The method in [153] uses only two frequencies and, therefore, solves a determined system. The closed solutions of [203] require three or four different frequencies, while the method in [261] uses five frequencies.

If more than two reflective paths occur per pixel, none of the previous approaches offer an effective solution. The simple multifrequency approach for resolving the MPI presented in [45] deals with non-unique reflections per pixel, assuming sparsity of the reflectors in depth domain, i. e., only very few return paths per pixel. The extremely sparse vector of reflected amplitudes (zero elements for all depths where no object reflects the signal) is recovered from partial Fourier measurements in a classical compressive sensing (CS) framework. The sparse signal recovery is carried out via OMP, i. e., a pseudo-l_0-minimization, with full *a priori* knowledge of the signal sparsity. Appealing due to its simplicity, the feasibility of this approach is compromised by the large number of modulation frequencies required (77 for known sparsity of 3). Additionally, if multipath arising from specular surfaces occur in combination with diffuse multipath, the assumption of exact sparsity does not hold and the number of measurements required to achieve a similar reconstruction is expected to grow. A more feasible approach is provided in [43], where k interfering paths are estimated from $2k + 1$ frequency measurements in a closed-form manner. The performance of the method, which does not require phase sampling, is evaluated with real data from an Xbox One sensor. In the experiment, 21 measurements are used to separate only two paths. The frequency domain ToF framework proposed in [249] allows for easy separation of multiple paths, at the cost of requiring large modulation bandwidth of the illumination system.

The approach presented in [189], named Sparse Reflections Analysis (SRA), is a general framework that allows both diffuse and reflective multipath estimation, exploiting the underlaying sparsity of the reflectors in depth domain. This approach discretizes the depth space and supposes that the signal received at the pixel is a sparse combination of returns from the different discrete depths. Even for low dimensionalities, the vector containing the return intensities for the different discrete depths meets the sparsity assumption, since most depths correspond to free space. If

Lambertian reflectors are considered, the vector of backscattering intensities is compressible, instead of sparse. Even in that situation, according to CS theory (Chapter 3), the sparse vector might be recovered from few measurements by l_1-minimization. This method has been demonstrated with the Xbox One sensor [471], which uses three modulation frequencies. It allows multipath removal in real time. A similar approach has been simultaneously presented in [221], where the concept of multipath is extended to scattering media, where the medium produces a number of different paths per pixel. Since scattering media are a continuum, the sparsity hypothesis in the Dirac basis does not hold and an overcomplete dictionary of exponentially-modified Gaussian functions is adopted to adapt the model and recover the sparsity of the vector of scatters. The signal received at each PMD pixel is a sparse convolution of the emitted signal convolved with some of these functions centered at different depth locations. Depending on the width of the Gaussian and the tail of the exponential, these functions can equally model specular (close-to-Dirac function) and Lambertian (non-negligible exponential decay) reflections. Similarly to [189], the signal sparsity allows solving the problem by l_1-minimization.

In general, transient imaging can be understood as way of multipath estimation. Consider, for instance, a scene that returns multiple reflections (still few) for each pixel of the ToF camera. Recovering the sparse time profile means estimating the locations of the impulses that form the scene response, that is, at which depths the reflectors are located in the scene. Consequently, one could say that multipath depth estimation is a specific case of transient imaging. For this reason, works using a PMD sensor for transient imaging, such as [248] and [220], implicitly provide a tool for multipath ToF imaging. In [248], PR codes with broad frequency spectrum and an autocorrelation function that approaches a Dirac delta function are used as modulation signal. Different measurements are acquired at different shifts of the signal, as in a conventional PMD sensor, with the peculiarity that PR codes are used as ICS and reference signal, instead of a periodic signal with close-to-zero bandwidth. The scene responses are recovered exploiting their sparsity directly in time (or, equivalently, depth) domain. This assumption might not hold if the MPI is diffuse, in which case an appropriate dictionary is to be used to sparsify the scene response functions [221]. The approach in [220] sticks to the sinusoidal modulation, originally employed in PMD sensors, instead of using PR codes. Measurements are taken at different frequencies and phase shifts. Differently from [248], where only the sparsity prior in temporal domain was used to recover the scene responses, in [220], both temporal and spatial regularization terms are

included in a minimization formulation to recover the scene responses for the different pixels of the image. The spatial regularization terms are functions of the vertical and horizontal image gradients, motivated by the expectation of gradient sparsity not only in temporal domain, but also in spatial domain. One more prior related to how well the estimated time profile fits a model is also included in the cost function to minimize. The model considers Gaussian shapes for surface reflections and exponential decays for subsurface scattering. The problem is split in two simpler subproblems, namely, solving the minimization for the time profile to recover, supposing known model parameters, and estimating the model parameters supposing the time profile known. The coupled subproblems are then solved via the Chambolle and Pock primal-dual framework [98].

Differently from [248] and [220], which operate in time domain, the approach in [298] translates transient imaging to Fourier domain and obtains the time profiles by means of an inverse Fourier transform and subsequent post-processing. Similarly to [298], in [367] the measurements provided by a ToF camera are interpreted as the Fourier coefficients of the scene response. The reconstruction of the scene response in phase (or ambiguous depth or time) from the Fourier coefficients is interpreted as solving a trigonometric moment problem [252, 271]. Obviously, even for moderate time resolutions, the problem is underdetermined, unless an unfeasibly large number of frequencies are used. The authors solve this issue by looking for the solution that, fulfilling the measurements, exhibits minimal Burg entropy [64]. Using such prior allows for a closed-form solution, known as the maximum entropy spectral estimate. Alternatively, if the scene response is known to be exactly sparse, as in the case of specular multipath, the Pisarenko estimate optimally reconstructs k Dirac delta functions, i. e., $2k$ parameters, from k complex phasors (cf. [43], where $2k + 1$ measurements are required).

A single-shot transient imaging system able to capture non-repetitive events at frames rated up to 10^{11} frames per second is reported in [198]. The system operates with a single illumination pulse and, therefore, can be considered as a pulsed ToF camera. A CCD camera is used for sensing and resolution in time domain is attained by means of a DMD in combination with a streak tube. The system trades spatial resolution in the shearing dimension for temporal resolution. Such a system natively allows for resolving multiple bounces per pixel. The transient video is recovered from the compressed measurements in a CS framework.

Towards the end of the writing of this thesis, a new field has emerged that combines transient imaging with light transport analysis to achieve robust separation of the individual transport components (e. g., direct reflections,

inter-reflections, caustics, etc.). While approaches oriented to determine or approximate the light transport matrix make use of different 2D patterns that are projected onto the scene, transient imaging approaches project time-coded light using different patterns. Combining both approaches means projecting 3D patterns onto the scene, being the light modulated both in spatial (e. g., using an SLM) and temporal domains simultaneously. Obviously, this would imply adding one more dimension to the light transport matrix. The core motivation seems to be getting rid of strong assumptions about the scene backscattering (e. g., sparsity or compressibility of the vector of backscattering amplitudes in depth domain), required when operating exclusively in temporal frequency domain, which might be violated in real scenes. For instance, it is known that direct transport preserves high spatial frequencies, while sub-surface scattering and diffuse inter-reflections have a low-pass filtering effect in spatial domain. In [359], sinusoidal temporal modulation is adopted and the scene is *probed* with different modulation frequencies, yielding a different complex-valued *transient frequency transport matrix* per frequency. The main hardware components are a PMD 19k-S3 sensor and a DLP LightCrafter projector using 650 nm laser diodes as illumination source. Computing the depth from direct/retroreflective images allows getting rid of the effects of diffuse MPI. The pure direct/retro-reflective image is computed subtracting the indirect-only image from a conventional PMD acquisition, containing both components. The indirect-only image is obtained by means of complementary 2D binary patterns in emission and reception of the light, so that direct paths are blocked (see [360] for the underlying concept of primal-dual coding). Spatial modulation of the emitted light is carried out by the DLP projector, while the complementary masking at reception is performed computationally, avoiding the need for a second spatial modulator. A similar idea is exposed in [209], where a phasor formulation is also adopted for the light transport model when using sinusoidally modulated illumination. The radiance at the pixel surface is the sum of direct radiance, due to direct reflection from the correspondent scene point, and global radiance, due to global light transport. The phase of the direct component contains the true depth information, while the global component contains the contributions due to MPI. Differently from [359], they exploit the fact that, for a broad range of scenes, global radiance decreases with increasing frequency. Performing the separation by temporal frequency, instead of by spatial frequency, the method in [209] avoids the need for spatial modulation/demodulation hardware. On the other hand, profiting from the vanishing effect of high modulation frequencies in the global component of the radiance means prohibitively large modulation

frequencies (600 MHz to 6 GHz) when the scenes contain small geometrical features, which could be handled by MSM-PMD, but not by conventional PG-PMD technology. A direct consequence of using high modulation frequencies is the need for several frequencies to achieve an acceptable unambiguous range. In their experiments a PMD CamBoard Nano (featuring a 19k-S3 chip) with a bank of 650 nm laser diodes for illumination is used. The PMD chip limits the maximum frequency to the 100 MHz range, which, in turn, limits the effectiveness of the method to large scenes. A closed-form formula for computing MPI-compensated phase shift under single indirect bounce hypothesis is provided in [338]. The radiances due to direct and global illumination are separated using the fast method proposed in [341]. The performance of the method is demonstrated using an Xbox One sensor and a DLP Lightcommander for spatial pattern projection.

2.4.2. Lateral Resolution

The lateral resolution may be bounded by the number of pixels or by the quality of the optics. Professional photographic cameras feature sensors with often more than 20 Mpx, pushing the limitation of lateral resolution towards the optical system.

In a conventional camera setup, the limit in lateral resolution imposed by the lens can be characterized by its point spread function (PSF). If the optical system is free of imperfections, the resolution limit is given by diffraction and the system is said to be *diffraction limited*. In optics, a spot of light optimally focused by a perfect lens of circular aperture generates an Airy pattern on the surface where it is focused. The Airy pattern consists in a central bright disk, also known as Airy disk, surrounded by concentric bright rings of decaying intensity. The pattern exhibits axial symmetry and its intensity value is given by Eq. 2.33,

$$I(x) = I_0 \left(\frac{2J_1(x)}{x} \right)^2 \tag{2.33}$$

where I_0 is the maximum intensity, J_1 is the Bessel function of the first kind of order one and $x = kr_{\text{lens}} \sin\theta$, being k the wavenumber of the radiation, r_{lens} the radius of the circular aperture and θ the angle of observation. All points of the pattern at the same distance from the origin share the same observation angle and, therefore, the same intensity. For completeness, the normalized amplitude (recall that the intensity is given by the square of the amplitude) of the Airy pattern is represented in Fig. 2.8a. The resolution

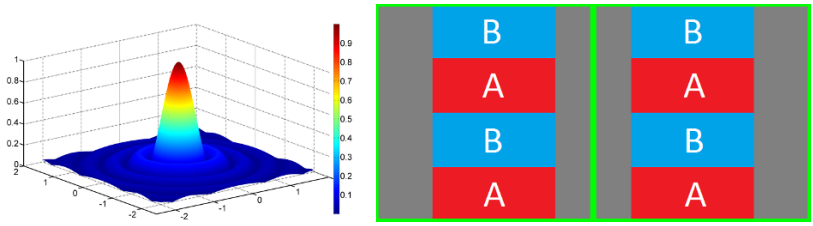
limit can be determined from the radius of the first Airy dark ring, given by Eq. 2.34 [465],

$$r_{\text{Airy}} = 1.22\frac{\lambda}{2NA} \tag{2.34}$$

where NA is the numerical aperture of the optical system and λ is the wavelength of the light. Clearly, larger wavelengths lead to larger Airy radius. In most cases, even using NIR light, the Airy radius lies in the micrometer range. Depending on the camera this may be higher or lower than the pixel size. In the case of a PMD camera, with a pixel size of $45 \times 45\,\mu$m (19k-S3), the lateral resolution is limited by the pixel size, unless the PSF of the optical system exhibit an abnormal width due to defocus or some system imperfection.

When the lateral resolution is given by the number of pixels (or, equivalently, by the pixel size, provided that the maximum sensor area is constant), two factors are to be taken into account: the pixel geometry and the FOV of the lens. Narrowing the FOV finer detail of the scene can be resolved with the same number of pixels. If the height and width of the pixels are different, the vertical and horizontal resolutions will also differ. The PMD pixel geometry for the specific case of the 19k-S3 chip is exposed in detail in Section 4.3.1. Here we only provide a general scheme of the two pixel geometries found in the chip, in Fig. 2.8b. The scheme has been generated from microscopic images of the pixel surface. Despite the regions do not perfectly correspond to physical boundaries, the pixel geometry and the relation between active and non-active areas is real.

Four regions per pixel are depicted in Fig. 2.8b, corresponding to the surroundings of four charge accumulation regions: two for the channel A and two for the B. In case of perfect demodulation capabilities, the location where a photon reaches the pixel should not determine in which integration channel the corresponding photogenerated charge carrier is stored. This should be only determined by the control signal. For now we adopt this hypothesis and leave further analysis for Section 4.3.1. Then, the active area of the pixel is the same for both pixel channels and equal to the colored regions in Fig. 2.8b (in red and blue in the color version, dark gray and light gray in the black and white version), while the rest of the pixel (in gray) is *blind*, i. e., non-active. The width of the active regions is calculated to be $25\,\mu$m, by simple proportionality, knowing that the size of the pixel is $45 \times 45\,\mu$m. This coarse model delivers an (optimistic) approximate fill factor of 56% and is enough to make clear that a PMD sensor will exhibit different behavior vertically and horizontally. While some scene structure that results



(a) Airy Pattern **(b)** PMD Pixels

Figure 2.8.: Normalized amplitude of an Airy diffraction pattern (a) and schematic representation of the active areas of PMD pixels in a 19k-S3 sensor (b). The axis units in (a) are arbitrary. The radius of the first Airy dark ring is, at maximum, of few micrometers, while the size of the square total pixel area (delimited in green in (b)) is $45 \times 45\,\mu$m. The gray areas in (b) are non-active areas, typically metal-shielded. For clarity, the active areas of each pixel are divided in A and B regions. A regions are centered at an A-gate readout wire, while B regions at a B-gate readout wire.

in a thin horizontal line in the image plane will be sensed—still, depending on the width, not fully resolved—, if it results in a vertical line, it might not be sensed at all, if it falls along the blind areas of the pixels. In terms of lateral resolution, if the pixel is considered as a whole, i. e., square, there is no change, but if only the active area is considered, the lateral resolution in horizontal direction is lower than in vertical direction. This means that one could either perfectly resolve or not sense at all finer structure in the horizontal direction. If the width of the non-active areas were constant and equal to the width of the active areas, splitting the image to two PMD sensors, horizontally displaced one width between each other, would be enough to obtain a PMD camera of double lateral resolution in horizontal direction. In Table 2.1 we provide values of the FOVs, angular resolution and corresponding lateral resolution at 1 m distance for some common focal lengths of the camera lens. The table columns are organized in pairs. The first pair contains the resolutions (horizontal or vertical) obtained using a pixel size of pixel is $45 \times 45\,\mu$m, while the second pair contains the horizontal resolution considering a pixel width equal to its active area ($25\,\mu$m).

In general, the lateral resolutions presented in Table 2.1 are poor, due to the low number of pixels in the chip (120×160). The lateral resolution degrades linearly with the distance to the scene. Consequently, unless

		Square Pixel Model		Rectangular Pixel Model	
f (mm)	FOV ($^\circ \times ^\circ$)	Ang. Res. ($^\circ$)	Lat. Res. 1 m (mm)	Ang. Res. ($^\circ$)	Lat. Res. 1 m (mm)
8.5	45.9×35.2	0.29	5.1	0.16	2.8
12	33.4×25.4	0.21	3.7	0.12	2.1
35	11.7×8.8	0.074	1.29	0.041	0.72

Table 2.1.: Field of View (FOV) and angular resolution of the PMD 19k-S3 sensor for different lenses of common focal lengths (f) and corresponding lateral resolution at 1 m distance. The resolutions obtained using the rectangular pixel model are theoretical horizontal resolutions, obtained using the width of the active area as pixel width.

extremely narrow FOVs are desired, the lateral resolution will easily be in the centimeter range at few meters distance. This also has an impact in the final depth accuracy, since it means that the pixel is measuring the average depth of a relatively large area of the scene. If the scene is, e. g., a plane at 45° with respect to the image plane, poor lateral resolution directly translates into an equivalent depth error, generating a staircase effect.

One could argue that the lateral resolution should be normalized by the size of the sensor, but even in that case, due to the pixel size, PMD sensors offer linear densities of 22.2 px/mm, while commercial photographic cameras range from 80 to 130 px/mm. Therefore, the low lateral resolution for the current chip size is an open issue in PMD ToF cameras.

2.4.3. Dynamic Range

The dynamic range of PMD pixels has shown to be around 70 dB (see corresponding paragraph in Section 2.3.2). Despite being relatively large, it is still below those of high-end digital cameras, close to 90 dB. The dynamic range is determined by the noise floor of the pixels and the maximum charge level they can store. Such parameters depend on the pixel design and fabrication technology and are expected to improve in new generations of PMD pixels.

We performed extensive experiments to determine the dynamic range of the pixels in a commercial PMD 19k-S3 sensor[5], which is the sensor included

[5] Note that PMD chips of newer generations, manufactured by Infineon, may exhibit significant improvements with respect to the PMD 19k-S3 we used for most experiments in this thesis, for instance, but not exclusively, in terms of dynamic range.

in the latest MultiCams and used in most experiments in this work. In our experiments, 100 different exposures were considered, ranging from 0.1 to 10 ms. For each exposure, 100 acquisitions were carried out. Since the PMD cameras operate with four phases, this means 4×100 pairs of A and B images. Two datasets with two different setups were acquired: one for determining the noise floor and another to determine the maximum value. In the first one, the camera has no optics and an opaque cap ensures that no light reaches the sensor. Additionally, the experiment was carried out in a dark room with no illumination. In the second setup, the camera is equipped with a lens and capturing a white wall, parallel to the image plane. The illumination is strong enough to drive most pixels to saturation in the cases of high exposure. It is important to note that the PMD pixels of the 19k-S3 are equipped with SBI system and, therefore, the saturation level is never achieved. Instead, the SBI does not allow the level of the pixel channels to exceed a close-to-saturation threshold (an exemplary value of $36,500$ is suggested in [284]).

All the images of all datasets are calibrated according to the per-pixel linear calibration described in Section 4.2.1. This allows averaging the results obtained for each pixel over the 120×160 pixels of the chip. The noise floor, I_{floor} in Eq. 2.17, is considered to be the standard deviation observed along the 100 acquisitions gathered per exposure time considered. The I_{floor} for a generic PMD pixel is obtained averaging the standard deviations for all pixels and exposures considered. The value of I_{ceil} in Eq. 2.17 is obtained by simply finding the maximum value for any of the pixel channels in the corresponding dataset.

Our experiments concluded that the dynamic range of a pixel channel of a PMD 19k-S3 array is 67.18 dB. This is close to the values presented in previous work, still lower, probably due to the fact that the SBI activation prevents pixels from reaching saturation level.

The SBI system can be understood as increasing the dynamic range of the PMD pixels, since it enables operation under stronger BGL conditions. In this sense, the SBI can provide an enormous improvement of the dynamic range, but it is essential to realize that the DC component of the light signals is being shrunken. If the goal is just to compute the depth using the four phases algorithm, this is irrelevant, but if we seek an *intelligent* processing of the signal that exploits eventual local correlations or data redundancy, we must be aware that information is being modified and the structure of the raw images might not meet our expectations anymore.

Another issue of capital importance is that the increment of dynamic range that the SBI can eventually provide is strongly scene-dependent, since it is

bounded by the discharge process of the less-charged channel. Depending on the depth being sensed by the pixel, the charge of the less-charged channel will range from zero (e. g., at 180° total phase shift for channel A or at 0° for B) to the same value as the other channel (at 90° or 270° total phase shift). Since the SBI removes the same current from both channels, the asymmetry of charge between channels limits its correct operation. Once the less-charged channel reaches zero (or some close-to-zero value) the difference between pixel channels is no longer preserved and the depth measurement will be erroneous. For clarity, we illustrate the SBI operation in Fig. 2.9, which plots data of a real experiment. Three operation areas are depicted in different colors, from left to right: the linear region, the area of correct SBI operation and the area of incorrect operation.

Channel A (in blue) charges at a higher rate than B (in red). When channel A reaches a threshold, the SBI starts to remove current from both channels, so that the level in channel A does not increase anymore. Consequently, the level in channel B starts to decrease with a negative slope, given by the difference between the slopes of the A and B response curves during the linear region. Clearly, if both channels had the same charge level when the SBI is activated, the slope of both curves in the central area of Fig. 2.9 (correct SBI operation) would be zero. In this situation the SBI would provide a theoretical infinite dynamic range, since, regardless of the light intensity and exposure time, the difference between channels would be preserved. Conversely, if the B channel had a close-to-zero level, which might occur for certain depths in cases of low background illumination, the SBI would provide no significant increment of the dynamic range, since the level of the B channel would go to zero immediately after SBI activation.

A close observation of the response curves in Fig. 2.9 reveals that the curves in the central region are not perfect straight lines. The difference between A and B channels is being distorted due to imperfect SBI operation. Additionally, the slope of the B channel curve starts to decrease before the level reaches zero, corrupting the differential measurement (A-B). The dotted lines extend the linear regions of both curves and should continue out of the graph. In case of ideal operation, if a vertical line is traced at any exposure time, the length of the segment between the dotted lines and between the real responses should be equal, which is clearly not true for exposures in the second half of the exposure range.

Due to its scene-dependency, the activation of the SBI produces a characteristic error pattern in the depth images, where some pixels exhibit much larger depth error than others, depending on the charge asymmetry between pixel channels in the measurements, i. e., depending on the depth.

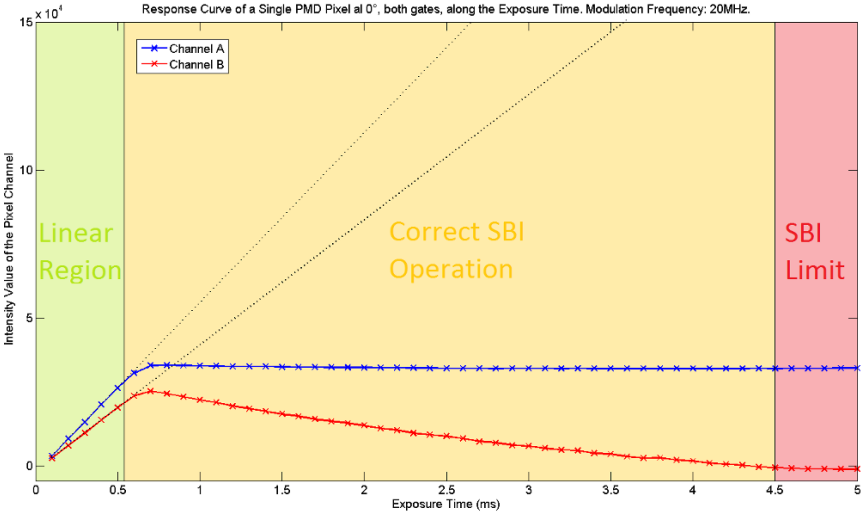


Figure 2.9.: Response curves of the A and B channels of a PMD pixel along the exposure time. The pixel corresponds to an object which is close to the camera, under strong 20 MHz-modulated IR illumination. Note the linear region of the curve, together with the good performance of the SBI, which subtracts the exact intensity to make the channel A charge remain at a constant level. The limit of the correct SBI operation is reached at almost 4.5 ms, where the charge in channel B reaches the zero level. This means that, if this close object is placed in a scene for which a higher exposure time is required, the depth measurements for the pixels of the object will be not just noisy, but wrong.

http://www.springer.com/978-3-658-18056-0

Compressive Sensing for the Photonic Mixer Device
Fundamentals, Methods and Results
Heredia Conde, M.
2017, XXXIII, 496 p. 97 illus., 10 illus. in color., Softcover
ISBN: 978-3-658-18056-0