Chapter 3
Color and Depth Sensing Sensor Technologies for Robotics and Machine Vision

Ali Shahnewaz and Ajay K. Pandey

Abbreviations

3D	Three dimension
AMCW	Amplitude-modulated continuous wave
APD	Avalanche photodiode
CAS	Computer-assisted surgery
CCD	Charge-coupled device
CMOS	Complementary metal oxide semiconductor
CNN	Convolutional neural network
CNN-CRF	Convolutional neural network-conditional random field
DoG	Difference of gradient
DSSC	Dye-sensitized solar cells
FIP	Focus-induced photoluminescence
FMCW	Frequency-modulated continuous wave
FOV	Field of view
FW	Fixed window
LiDAR	Light detection and ranging
MIS	Minimally invasive surgery
RANSAC	Random sample consensus
RGB	Red green blue
RGB-D	Red green blue depth
SAD	Sum of absolute differences
SfM	Structure from motion
SIFT	Scale-invariant feature transformation

A. Shahnewaz · A. K. Pandey (✉)
School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD, Australia
e-mail: shahnewaz.ali@hdr.qut.edu.au; a2.pandey@qut.edu.au

O. Sergiyenko et al. (eds.), *Machine Vision and Navigation*,
https://doi.org/10.1007/978-3-030-22587-2_3

SLAM	Simultaneous localization and mapping
SPAD	Single-photon avalanche diodes
SURF	Speeded-up robust feature
ToF	Time of flight

3.1 Introduction

Conventional vision technology projects 3D world information into 2D plane with information lacking in Z-axis, that is, the depth information of a scene. Access to depth information is paramount in capturing the real-world space; therefore, 3D vision systems form important research topic for robotic and autonomous systems. For instance, path planning and obstacle avoidance form the key aspects of autonomous vehicles and heavily rely on sensors providing situational awareness for system accuracy. A large body of research is being conducted on safety and obstacle avoidance, and 3D vision technologies remain an integral part of robotic systems [1–3]. It is not surprising that in addition to the usual red, green, and blue (RGB) color vision, most of the advanced robotic vision systems already deploy a form of active or passive depth information using the so-called RGB-D vision technology, where D stands for depth. In robotics, time-of-flight (ToF)-based sensors together with stereo vision systems are widely used to extract the depth information. ToF sensors are particularly suited to self-driving cars and autonomous aerial systems or drones. ToF-based depth sensor is the most promising form of long-range active depth sensing, and tech giants such as Texas Instruments, Sony, Panasonic, STMicroelectronics, AMS, etc. are currently developing micro-depth sensor for range imaging in a form that is compatible with portable device such as smartphones.

Object recognition in real time is yet another active research area in robotic vision, and use of RGB-D sensors for 3D object reconstruction is common. Information contained in voxels is used to compare and identify different objects and features contained within them [4–7]. The advantage of this approach benefits from the fact that a lot of salient features can be extracted from the 3D space to improve object recognition performance [4–7]. No wonder the demand for novel and high-resolution cameras that can provide depth is on the rise. Currently, many commercial 3D image sensors exist in the market, and imaging system providers are developing a new generation of 3D image sensors [8–10]. Surveillance system, vehicle identification, traffic control system, people counting system, activity and gesture identification etc. are the subdomains of this category where 3D information offers improved system efficiency [1, 2, 4, 12, 109]. Access to depth information has a big impact on computer graphics especially in games and in content and image retrieval as well as in archeology [13–15].

In medical robotics, depth information has a great influence on assigning perception. In computer-assisted surgery (CAS) or in robotic-assisted minimally invasive surgery (MIS), depth has an important role. In conventional MIS procedures, 3D surgical world is projected to 2D screen; hence, surgeons performing MIS face more challenges than open surgery. Surgeon has to operate 3D world in a 2D space without haptic sense that makes MIS system more complicated. Unintentional tissue damage is often reported that may later cause other difficulties such as arthritis or osteoarthritis. In MIS context, vision is the most crucial factor that improves surgical outcomes with respect to safety and unintentional injury [11]. Without the depth information, MIS faces difficulties to track surgical tools within the surgical space. Promising improvement has been reported when 3D vision is incorporated into the tracking system [16]. Recent studies show a significant amount of improvement in MIS procedure by presenting a comprehensive result of 3D MIS versus 2D MIS. According to their records, median error of MIS in 3D surgery versus 2D is 27 and 105, respectively, that reports 25.72% less median error [17]. Another study shows that 3D MIS reduces 71% performance time as well as 63% error rate [18, 19]. Therefore, 3D vision systems offer great advantage in countries where the number of skilled surgeons is limited.

In this chapter, we describe a diverse range of vision technology by reviewing the current scanning technologies specific to application areas of robotic and machine vision. We also aim to extend this discussion to capture the advantages and limitations of active and passive depth sensing technologies using in stereo vision, time of flight, and structured light with a particular focus on how to deal with constrained (indoor) or unconstrained (outdoor) environments.

3.2 3D Image Construction

Depth estimation technique mainly faces two big challenges: (1) depth accuracy and (2) computational cost in terms of time [20–23, 106, 107]. Two different branches, sensor technology and computer vision, actively involved in research to meet these constraints. Based on the imaging technology, current depth estimation technologies can be classified into two main categories, active or passive, as illustrated in Fig. 3.1. Passive estimation technology relies on machine learning algorithms and

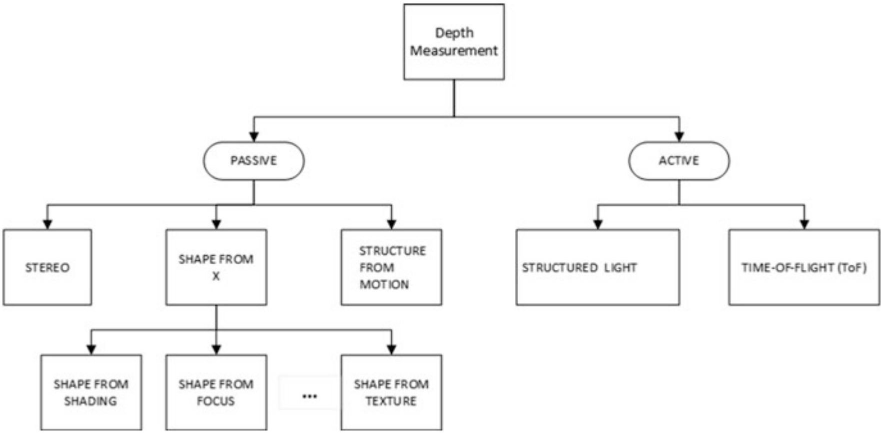


Fig. 3.1 Classification of depth measurement technology

mathematical approaches to extrapolate depth information from 2D image or images. Whereas, the other class is referred to as active depth sensing technology which relies on active controlled signal sources and sensor technology to estimate distance. The aim of this chapter is to provide a comprehensive review of the various depth estimation approaches along with their merits and demerits.

3.2.1 Image Sensor

Imaging technology uses natural or ambient illumination to capture the scene. Most of the image sensors are based on charge-coupled device (CCD) or complementary metal oxide semiconductor (CMOS). On the other hand, optical scanning sensors are used to estimate depth [24]. Wendy Flores-Fuentes et al. proposed a novel electronic sensor that consists of an electronic processing unit along with photodiode. To measure the distance, their proposed work infers the energy centre of an optical received signal. In the next section, we are going to introduce the most promising passive stereo technique to infer 3D structure.

3.2.2 Stereo Vision

Stereo vision is the most common approach to infer depth from a set of images. Computer vision algorithms are used to reconstruct depth from single or multiple images. Single-view 3D reconstruction methodology uses only one image. On the other hand, multi-view 3D construction considers two or more images to reconstruct depth information. It is also known as stereoscopic vision. When two images are used, the system is known as a binocular stereo vision system, and probably it is the most widely focused research area of computer vision.

Binocular stereo vision originally mimics the human vision system. In a binocular stereo vision, two images are taken from two different cameras at the same time [25]. The basic requirement is that two cameras are placed at a known distance. In this arrangement, the left camera is denoted as a reference camera where the right camera is called target camera. The distance between the optical center of these two cameras is referred as a baseline. Stereo vision system uses the concept of parallax and uses disparity as a vision cue. Figure 3.2 provides an overview of binocular stereo vision and how it is used to calculate depth.

As highlighted in Fig. 3.2, stereo matching is the core technique of the stereo vision. Stereo matching is the process that matches each pixel from the reference image to the target image to perceive the depth of each pixel [26]. The resulting output image is often referred as a depth map. An intensive comparison takes place to find the corresponding pixel in the target image. Offline camera calibration and pre-processing always take place before the actual stereo matching process [27–29]. Ideally, the reference and the target cameras capture the same scene point at the same time with a slightly different viewpoint, and this serves the basis for stereo vision algorithms [28]. Therefore, the term synchronization is always used to convey the

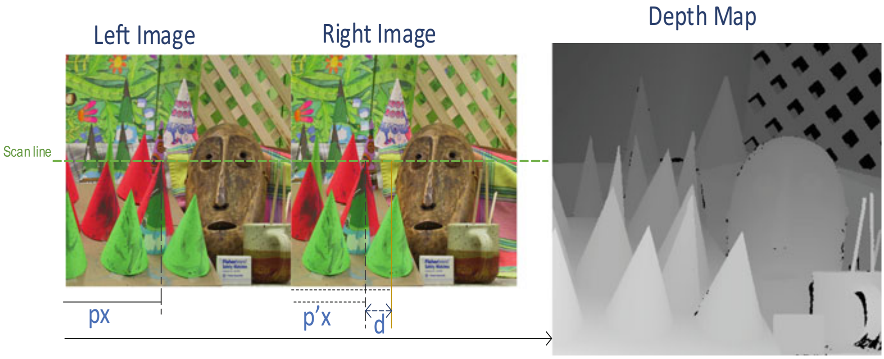


Fig. 3.2 Overview of binocular stereo vision system. Left and right cameras take the same scene image, then stereo matching is performed to find the corresponding points. The resulting corresponding points provide disparity information with respect to the left image. Finally, depth map is calculated from disparity

sense that the image acquisition system captures the same scene point at the same time with no time lag [30, 31]. When the object is in motion, this precondition plays a pivotal role to reduce reconstruction noises, and camera calibration process is used to eliminate image acquisition distortions [27, 32]. Basically, stereo rectification is a transformation process that aligns two images into the same plane, so that same horizontal line becomes parallel to both camera centers [33, 34].

Depth is calculated by finding disparity in a pair of images. Disparity refers the distance between two corresponding points in the left and right images of a stereo pair. It is inversely proportional to the depth and vice versa. In a stereo vision system, the relationship between depth and disparity can be expressed as follows [35–37]:

$$d = \frac{bf}{Z} \tag{3.1}$$

Here b is the baseline and f is the focal length. Z stands for depth, and disparity is expressed by the letter d. When the stereo matching process is completed, the difference of the pixel position in the right image with respect to the left image is referred to disparity of that pixel. The basic idea of the disparity calculation is to match each pixel from the left image to the right image. In some circumstances, parts of a scene may not be visible to one or both cameras. This part of a scene is known as a missing part. In other words, stereo matching process fails to find the best match. These outcomes are often reported as holes [38]. Hence, after calculating depth map, post processing algorithm are used to refine noises [35]. Depth is estimated from disparity by using the geometric principle of triangulation, and some of the common approaches are summarized in Fig. 3.3.

Generally, stereo matching algorithms are classified into two groups. Pixel-wise matching algorithm is categorized as a correlation-based approach. It is

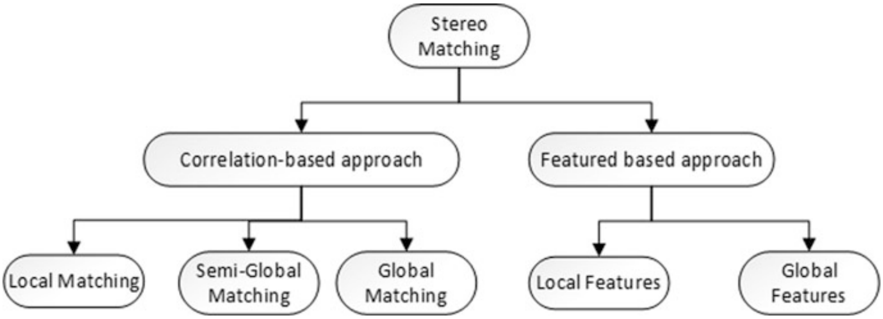


Fig. 3.3 Classification of stereo matching algorithm

further grouped into two main groups: (1) local matching and (2) global matching. Between the local matching and the global matching, semi-global matching resides that combines the upsides of both the algorithms. Stereo matching becomes highly ambiguous especially when a single pixel is considered. To alleviate this characteristic, often fixed or variable length of window is considered. Example of the local matching algorithm is the sum of absolute differences (SAD), fixed window (FW), etc. [39, 40].

Local methods can estimate disparity at high speed, but it compromises estimation accuracy to computational cost. The downside of local matching algorithms is that the disparity map often contains ambiguity. Though window-based approaches improve the overall accuracy, defining universal window size to balance both speed and accuracy is a challenging task. Probably, one of the major limitations of local matching algorithm is that it is incapable to handle occlusion due to lack of global information. Moreover, this group of algorithms is often limited to low texture images because, often, the local windows fail to capture smoothly varying texture features [13, 21] at low frequency. The rudimentary hypothesis of this group of the algorithm is that the corresponding pixel exists on the same horizontal scan line. For this reason, rectification is a crucial step to increase the accuracy of disparity estimation. However, accurate image rectification in practice is a hard task. Some algorithms also consider an additional path to estimate disparity [41]. But this additional path aggregation function again increases the disparity computational cost.

On the other hand, global matching algorithms provide improved and highly accurate depth map [102]. Instead of the local neighbor pixel, the global method takes into account all image pixels. Smoothness function is the most pivotal step of global method. The aim of this step is to minimizes the energy cost of the overall depth map. The objective is to reconstruct depth map with the lowest energy. Unlike local stereo matching algorithm, this set of algorithms requires very high computational cost. Generally, the energy minimization function is defined as [1, 42]:

$$E(d) = E_{\mathrm{D}}(d) + E_{\mathrm{s}}(d) \tag{3.2}$$

Here, $E_{\mathrm{s}}(d)$ is known as a smoothing function.

Local methods are not robust to noise, and accuracy gets compromised with respect to speed. On the other hand, global methods consume high computation cost, are robust to noises, and provide highly accurate results. Adaption of a global method in a real-time system is a challenging task. The semi-global method originally proposed by Hirschmuller [41] balances these two approaches. According to this method, the matching process is performed with a set of pixels that is basically a window-based approach. So, initially, the stereo matching approach starts with local stereo matching process. Census transformation and sum of absolute differences (SAD) are probably the most used algorithms to perform this task. Census transformation is more robust than SAD [43]. Here, window size or census kernel size plays an important role in identifying textureless or low texture properties. The drawback of larger window size is that it increases computational cost. In order to estimate the matching cost, generally hamming distances are used. The lowest hamming distance is preferred for each pixel over the total disparity level. This initial matching cost encounters the same problems of the local matching algorithm. Thus, it contains wrong correspondences due to limited or low textures. To alleviate these problems to some extent, the semi-global method introduces further cost aggregation function which is known as a path cost aggregation. Path cost is calculated from several directions, and in practice, 8–16 directions are used. Although the semi-global method improves local method matching accuracy, this method still falls short in fully overcoming the above-described limitations. Path cost aggregation of the semi-global method can be described as follows:

$$E(D)=\sum_{p}\Big(C\,(p,\mathrm{Dp})+\sum_{q\in N_{\mathrm{p}}}P_1T\left[|\mathrm{Dp}-\mathrm{Dq}\,|=1\right]+\sum_{q\in N_{\mathrm{p}}}P_2T\left[|\mathrm{Dp}-\mathrm{Dq}\,|>1\right] \tag{3.3}$$

Passive stereo matching technology encounters a set of challenges. Image may be contaminated by noises. Missing point due to occlusions or self-occlusion, the absence of texture, and the perfection of same horizontal scan line alignment are the rudimentary problems to reconstruct 3D structure using stereo images. Principally three pivotal metrics are used to describe a stereo matching algorithm as a whole. These are (1) robustness, (2) accuracy, and (3) computational cost. Feature matching-based algorithms are also widely used to estimate passive depth from images. In this approach, features are calculated to construct feature vectors. This process can be referred to as a feature descriptor process where features are extracted from the images. Then feature matching algorithms are used to find the correspondence feature, and disparity is calculated based on the matching outcomes. The most common image features are edges and corners. But these features are usually susceptible to noise but have less computational cost. Other widely used image features are scale-invariant feature transformation (SIFT), difference of gradient (DoG), and speeded-up robust feature (SURF) [42, 44–47]. Feature selection is a crucial process. Robust features are always preferable, but it increases computational cost. By definition, features are the most interesting points of an image that carries important image information. Therefore, feature-based methods create a sparse matrix. Only partially reconstructed depth can be achieved from feature-based stereo matching algorithm compared to dense depth map construction.

3.2.3 Shape from Shading

Binocular stereo is based on finding the corresponding problem. However, images also contain many visual cues such as shading, texture, etc. In computer vision, these visual cues are used to construct the shape of an object. These are classical approaches used in monocular stereo system. Among them, shape from shading and photometric stereo are the most prominent fields which are still viable and probably the most widely used research approaches to reconstruct 3D structure from a single image. This approach has many applications. Considering the growing interest of this approach, we focused on shape from shading and photometric stereo in more detail.

Originally, shape from shading approach was proposed by Horn [48] and later by his Ph.D. student Woodham. In his Ph.D. thesis, Woodham proposed a photometric approach which is the extension of shape from shading [49]. Though it seems a very old approach, it is still an active research area in computer vision to infer depth from monocular or single view.

Shading pattern of an object conveys information and visual cues of its surface. Under controlled lighting source, the reflected light intensity of an object surface has a sharp relation to its surface shape. It creates a bridge between the shading to surface slope. However, shape from shading is often marked as an ill-posed problem that refers to the same numerical solution representing two distinct surfaces, one is inversion of the other one. Photometric stereo, which is one step further from the approach of shape from shading, solved this problem using more light sources [112]. As shown in Fig. 3.4, the idea behind photometric stereo is to estimate surface reflectance coefficient, albedo, and surface normal. When these are estimated, depth of a surface is calculated by integrating surface normals or by solving nonlinear partial derivative equation. One important definition in this context is the surface albedo. It is the reflectance coefficient that tells the amount of light a surface can reflect. The value of albedo is between 1 and 0, and it is denoted by ρ. Limitations of the shape from shading is that this approach is based on some assumptions such as Lambertian surface, surface smoothness, and discontinuity. On the other hand, photometric stereo is often limited to complex lighting environment and specular nature of a surface. Using knowledge of radiometry, numerical solutions, and proper construction of the system especially known lighting source, photometric stereo is able to capture depth from textured, untextured, or textureless images.

Surface intensity or surface irradiance can be expressed as

$$I(x, y) = R(p, q) \tag{3.4}$$

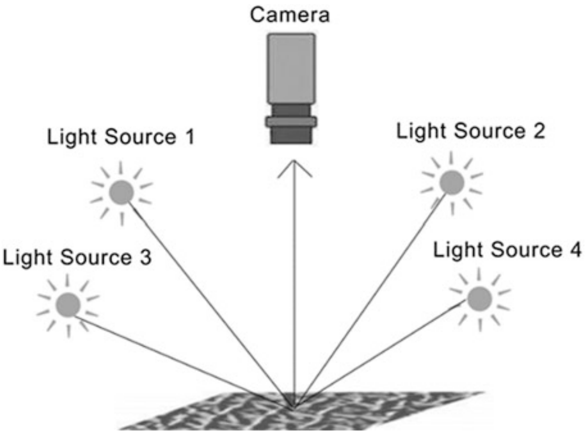


Fig. 3.4 Photometric stereo system. The figure is taken from S. Ali et al. [50]. Four images are taken from four different lighting sources at different angles. z is the depth of the point

From the radiometry and back to the literature, Horn used this relation to model shape from shading. Equation (3.4) tells that the image intensity is directly proportional to its reflectance (R) map which is also known as irradiance map. Reflectance map is a relational map which relates scene radiance, surface reflectance property, surface orientation, and observed brightness [51]. If surface reflectance property is estimated properly, then surface radiance depends on the surface shape. Horn approach p and q in Eq. (3.4) represent the surface gradient points and can be expressed as

$$p = \frac{dz}{dx} \tag{3.5}$$

$$q = \frac{dz}{dy} \tag{3.6}$$

Extension of this shape from shading is photometric stereo. The basic idea is to infer depth of a scene illuminated at different angle. In photometric stereo, a camera is placed in a fixed position. Usually, three or more lighting sources are used to construct photometric stereo. Images are captured one after another by changing lighting direction from one to another. The idea is to capture surface orientation from different illumination direction. Collected images are then processed to construct depth map.

One reflectance map corresponds to one light source. So, $\boldsymbol{l}$ number of light sources produce l number of reflectance maps. Unlike shape from shading, photometric stereo calculates surface property such as albedo. Photometric stereo is an overdetermined system where the number of unknowns is less than the number of equations. Hence, it eliminates the limitations of shape from shading. The surface normal can be defined as a vector on a surface in 3D space which is perpendicular to the surface. The basic principle is based on the radiance by calculating surface normal and the direction of light. Suppose **s** is the light source vector and **n** denotes surface normal, then image irradiance can be expressed as follows:

$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} = \rho \begin{bmatrix} \mathbf{s}_1^{\mathrm{T}} \\ \mathbf{s}_2^{\mathrm{T}} \\ \mathbf{s}_3^{\mathrm{T}} \end{bmatrix} \mathbf{n} \tag{3.7}$$

Finally, the depth map is calculated from the surface normal through numerical integration. There are some well-known numerical methods that already exist in literature mainly based on fast marching method and the integration methodology [52, 53]. According to the integration method, depth can be calculated from the following equation [54]:

$$z\,(x, y) = z\,(x_0, y_0) + \int p\,(x_0, y_0)\,dx + q\,(x_0, y_0)\,dy \tag{3.8}$$

Lambertian surface property is the preliminary assumption of the shape from shading or photometric stereo. Dynamically estimation of a surface property or photometric stereo for non-Lambertian surface is one of the active areas where a lot of contribution has been reported. In recent years, many contributions are reported where structured light, color image intensity, and fusion of photometric stereo with other approaches are used [55, 56]. The strong point of photometric stereo is that it provides fine surface shape with fine depth information. Recent patent has been reported in 2018 where photometric stereo process has been used to reconstruct 3D environment model [57]. The downside of this approach is that, unlike passive stereo system, it uses external lighting sources to estimate depth. Hence, it is limited to environmental lighting sources, or complex lighting sources make this approach hard to estimate depth.

3.2.4 Dynamic Vision

In contrary to other conventional imaging system or camera, the event camera meets high-speed vision sensor demand. The idea behind the event camera or dynamic vision sensor is to produce an image when an event has occurred. In other words, even if the brightness value of a single pixel changes, it produces an image. Event camera does not produce image at a fixed rate, but based on an event, it generates an image at high speed. An event can be translated into time series tuple of $\langle t_{\,k}, (X_{\,k}, Y_{\,k}), p_{\,k}\rangle$ [58, 59], where $t_{\,k}$ expresses time, $(X_{\,k}, Y_{\,k})$ is the coordinate of a pixel that raises an event, $p_{\,k}$ defines priority. Event camera can produce an event in some order of milliseconds [58, 59]. In robotic odometry, event camera provides the ability to solve many feature-based cutting-edge problems such as visual simultaneous localization and mapping (SLAM) [60, 61]. On the other hand, the event camera has great influence on passive depth estimation. Examples of event-based scene detection is shown in Fig. 3.5.

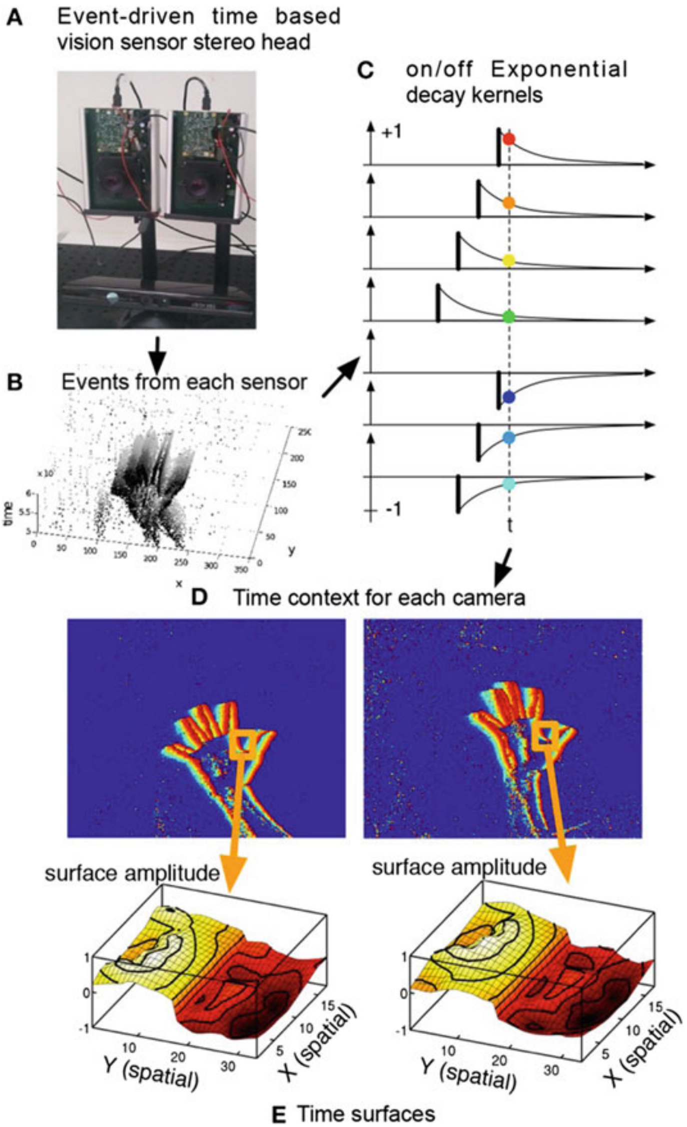

A Event-driven time based vision sensor stereo head
B Events from each sensor
time
x
y
C on/off Exponential decay kernels
+1
-1
t
D Time context for each camera
surface amplitude
surface amplitude
Y (spatial)
X (spatial)
E Time surfaces

Fig. 3.5 Dynamic vision based on events. Events in the scene are captured through dynamic vision cameras, and depth construction is obtained from the images. (**a**) Dynamic vision camera-based stereo system. (**b**) Output events and captured scene. (**c**) Extracted neighbourhood that allows to build the event context. (**d**) Time context of the most recent events. (**e**) Exponential decay kernel for spatial domain. Figure is taken from Ieng et al. [62]

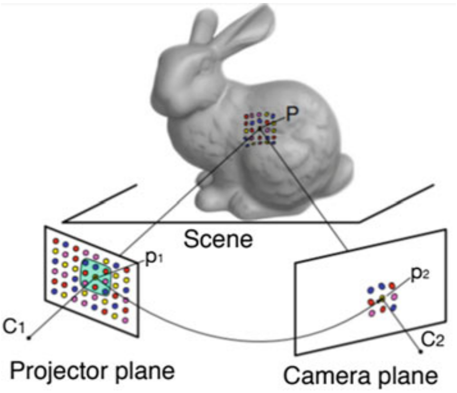


Fig. 3.6 Dynamic vision with structured light. Matching is performed between two view p1 and p2, and depth is recovered through triangulation method from known optical center C1 and C2. Figure is taken from T. Leroux et al. [64]

N. Julien et al. estimated depth using dynamic vision sensor using active approach [63]. In their work, they addressed passive stereo matching problem using event data. They generated events of an observed scene so-called light spots using lens and laser light, and scanning was performed by translating laser beam. The Fig. 3.5 shows the output of a stereo rig consisting of a dynamic vision sensor that produces overlapped stereo images. Stereo matching is performed over the sparse data at each event. It alleviates the stereo matching problem. An active pixel array is used to grab a visual scene. Though this work approaches to solve the stereo matching problem, scanning all the pixels of the field view area consumes time. Moreover, their approach is limited to a range in some meters.

T. Leroux et al. in their method used digitized structured light projection with an event camera to estimate the depth of a scene [64]. Their method as shown in Fig. 3.6 relies on the use of frequency-tagged light pattern. It generates a continuous event. Since structured light has a distinguishable property of pattern at a different frequency, it facilitates matching problem on event-based data.

The fundamental approach is based on the idea that unique projector pixel triggered a unique scene point that is captured by the image sensor. By knowing this two center points say C1 and C2, depth is recovered using the triangulation method.

3.3 Active 3D Imaging

The active 3D imaging system consists of an additional signal source known as a projector. The aim of the projector is to emit signals. Received reflected signals are analyzed to construct the 3D structure of the surrounding environments. The emitted signal can be laser light, ultrasound signal, near infrared light, etc. It is known as a projector, and its responsibility is to fire signals on the surrounding surface. Many terms are used to describe 3D active imaging technology such as a rangefinder and range imaging. Several methods are used to measure distance, but probably the most practiced methods are based on time of flight (ToF), triangulation and phase shift. This section provides a brief introduction of active sensing methods and technique. Dense depth map with less ambiguity and minimum depth error are the most reported advantages of active 3D imaging technology. However, the resolution of the depth map is limited. Miniaturized, high-resolution, and low-power active depth sensor has a potential demand in various fields like medical robotics.

3.3.1 Time of Flight

Time-of-flight (ToF) systems measure the distance from the scanner to surface points. The basic idea of the active sensing technology is to emit signal such as from a laser. When the signal is emitted by the projector, then the clocking system inside the active imaging system starts counting. This approach is known as direct time of flight. If the object exists within the range of the imaging system, then it reflects a potential amount of signal to the receiver. When the receiver part of the camera receives this signal, it then computes the round-trip time. Then the distance is estimated from the basic principle of the light or electromagnetic source as follows:

$$d = \frac{\Delta t * c}{2} \tag{3.9}$$

Here d expresses the distance of an object from the camera, Δt stands for total travel time, and c is the velocity of the light. Figure 3.7 captures the fundamental working procedure described above. Direct ToF imaging system consists of four basic components: (1) light source or transmitter, (2) optics, (3) light detector or receiver, and (4) electronic timer [65]. Light source or transmitter along with optics produce a signal transmitting unit for this system. Different light or signal source can be used such as near infrared or laser. The optical lens is used to diffuse the signal

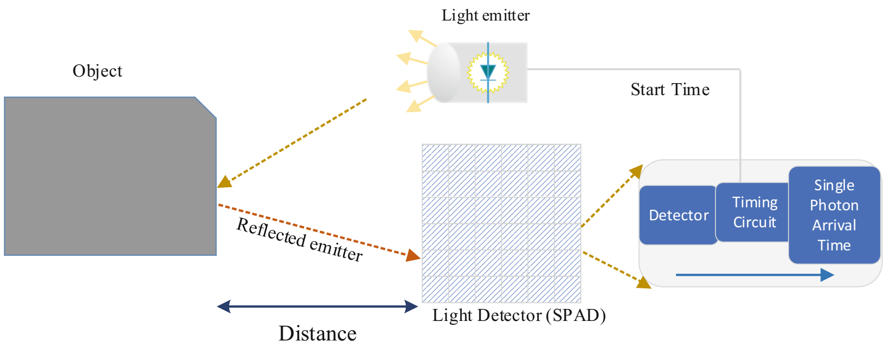


Fig. 3.7 Working principle of ToF-based sensor

over the surface. The optical lens is also used to collect light and project it to the receiver. It creates limited field view to avoid other outdoor lighting such as sunlight.

On the other hand, the receiver unit is generally composed with two system components: (1) photosensor and (2) electronic time counter. Within a defined range, ToF provides high-quality depth map. High-scale precision clock is the challenging part of this approach. When an object is placed very near to camera, for example, in millimeter distance, it is a challenging area to design a clock that can measure a time gap in nano or pico scale. That makes active direct ToF camera limited to very short distance.

Photosensor has to sense reflected light within a very short time. Some semiconductor components such as avalanche photodiode (APD) and single-photon avalanche diodes (SPADs) show their ability to sense signal within the picosecond range, and these components are used to fabricated ToF sensor [66]. To improve the resolution, efficiency, and scale down the dimension of the whole imaging system, currently leading manufacturers are involved in developing solid-state 3D active imaging system [9]. In the indirect approach, a continuous signal is sent by the transmitter or projector instead of the one-shot signal in order to avoid small-scale clock design. The transmitter contains an array of a signal emitter and generates the desired signal. Different kind of signals are used such as sine, square, etc. The received signal is compared to the original signal. Different signal characteristics are used such as signal phase to estimate distance. It is a continuous process and more flexible for silicon technology.

One-shot approach can measure both short and long distance with some range limitations. Long distance measurement requires stronger light source, in most of the cases coherent light sources, which can be hazardous. Moreover, strong and complex lighting source can contaminate the reflected signal. In practice, a multi-shot approach is adopted to overcome this problem. However, still high-power light source is the main drawback considered in this situation. The continuous pulsed signal is used to overcome this crucial problem. From the basic theory of signal processing, the target signal is wrapped into a carrier signal that has relatively low frequency. Often amplitude-modulated continuous wave (AMCW) or frequency-modulated continuous wave (FMCW) are used in this domain. In frequency-modulated continuous wave (FMCW), high-frequency signal is combined to a relatively low-frequency signal and then transmitted. This mechanism increases the system robustness. Suppose an emitted signal St_x is transmitted and a reflected back signal Sr_x is received. If a sine signal is transmitted, then they can be expressed as

$$St_x = \cos 2\pi\omega t \tag{3.10}$$

$$Sr_x = \cos\left(2\pi\omega t + \varphi\right) \tag{3.11}$$

where φ contains phase shift information that eventually expresses the amount of time and distance that the signal traversed after its emission. Basic electronics and filter approaches are used that estimate phase shift between transmitted and received signal. Multi-path propagation is one of the considered problems of ToF technique. When light hit a surface point, the scattered light may fall on the detector plane through different paths. Multiple detection of such event can generate noises.

In single-photon detection, especially in solid state where avalanche photodiodes are widely used to convert light energy to current energy, strong light source generates strong current. When light source traverses a long path, the signal becomes weaker and generates less current. Similarly, when the signal hit a diffuse surface or mat surface, it reflects weaker signal. The signal is scattered into different directions when it hits sharp edges and photodiode response becomes weaker. These are well-addressed problems of ToF technique especially for LiDAR (light detection and ranging). LiDAR-based system uses the fundamental principle of ToF [67]. Laser is a more preferable choice as a light source. When laser light is reflected back from a surface point, LiDAR can estimate the surface distance. Since light is projected on the surface, the surface property and other factors can also contaminate the reflected light as it is mentioned in the ToF section.

Spot scanner scans a single point at one time. This type of LiDAR projects laser light on the surface point. The back-propagated light is captured and projected to the light detection sensor. Single point distance is measured with this approach. To recover whole geometry covered by the field of view (FOV), conventional steering is used to scan all the points.

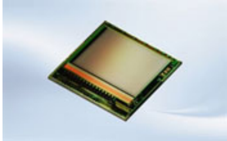
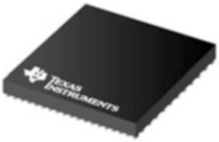
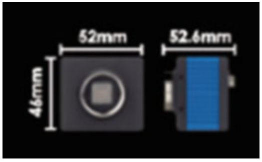
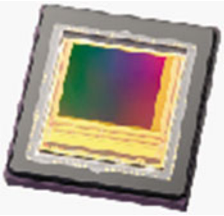
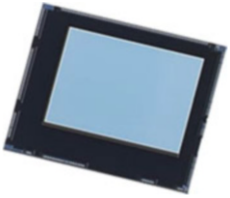
Apart from the pulsed shot approach, amplitude modulation continuous wave (AMCW), frequency modulation continuous wave (FMCW), and triangulation techniques are also adopted. In recent years, a big volume of literature currently exists which concentrates ToF camera especially LiDAR on solid state. Single photon on distance measurement technique is widely adopted. Some solid-state materials such as avalanche photodiode (APD) and single-photon avalanche photodiode (SPAD) are widely used in this research area to detect incoming light at very small time gap [68]. With the capability of detecting and discriminating incoming light in the range of picoseconds, these materials became the state-of-the-art choices to develop solid-state LiDAR.

Table 3.1 presents a comprehensive list of active depth measurement sensors and their characteristics

3.3.2 Structured Light

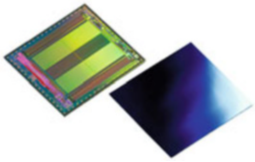
Structured light is used to estimate the depth of a surface. This technology is widely used to construct 3D image [69–71, 105]. Similar to time-of-flight (ToF) mechanism, it uses a projector that generates a pattern of light. Considering pattern generation procedure, structured light can be further categorized into basic two classes: single-shot structured light and multi-shot structured light. When a pattern is projected, the surface scene is captured by the image sensor. Based on the number of the image sensors used, a structured light depth estimation procedure has two well-studied directions: (1) monocular structured light and (2) binocular structured light [71, 72].

Table 3.1 List of active depth measurement sensors

Product name	Characteristics	Vendor
REAL3 	Dimensions 68 mm × 17 mm × 7.25 mm Measurement range 0.1–4 m Frame rate max 45 fps Resolution 224 × 172 pixel (38 k) Viewing angle (H × V) 62° × 45°	Infineon REAL3™
PMD PhotonICs® 19k-S3 	Time-of-flight 3D chip Dimensions 12 × 12 mm^2 Pixel array 160 × 120	PMD
OPT9221, OPT8241, OPT3101 OPT8320 	Time of flight Long-range distance Sensor resolution 80 × 60 to 320 × 240 Frame rate 1000–120	Texas Instruments
PX5 	Alternative spatial phase image 3D sensing Up to 5 MP resolution Fame rate 90	Photon-x
BORA 	Time of flight Resolution 1.3 Megapixel Distance various range Minimum range 0.5 m Maximum 500 m	Teledyne e2v
IMX456QL back-illuminated ToF 	ToF image sensor VGA resolution 10 µm square pixel Approx. 30 cm to 10 m distance measurement	Sony DepthSense™

(continued)

Table 3.1 (continued)

Product name	Characteristics	Vendor
Epc660 	Time of flight Resolution 320 × 240 pixels (QVGA) 1000 ToF images per second in advance operation mode Range millimeter to 100 m	ESPROS Photonics Corporation
MESA 4000 	Time of flight Distance 0.8–8 m Resolution 640 × 480	MESA Imaging
SR300 	Structured light (IR) Distance 0.2–1.5 m Resolution 640 × 480	Intel RealSense™
ASUS Xtion 	Structured light (IR) Distance 0.8–4 m Resolution 640 × 4800	Asus

Depth estimation procedure analyzes a captured structured light pattern to estimate depth. Different methods are already established so far and can be grouped into (1) spatial neighborhood pattern method, (2) time-multiplexing pattern method, and (3) direct coding pattern method [23]. The fundamental approach of structured light depth estimation is to calculate disparity and can be defined as $d = U_{\mathrm{a}} - U_{\mathrm{c}}$. U_{a} comes from the projector coordinate system and U_{c} comes from the camera coordinate system.

As shown in Fig. 3.8, in this arrangement, depth estimation can be defined as a pattern matching problem of the scene that is illuminated by a specific light pattern.

Some approaches are based on deformation of the received pattern. Considering the correspondence problem of passive stereo vision, it shows a significantly improved result on a textured and textureless region as well as it reduces ambiguity [19, 74]. In structured light triangulation method, camera calibration can be the first building block that estimates camera intrinsic matrix. It is also important to estimate the extrinsic parameter that maps projector coordinate to camera coordinate system known as stereo calibration. Encoded light patterns are projected on the surface, and reflected patterns are captured by the image sensor. Deformation depends on the surface planar characteristics. Matching is performed on the decoded pattern by using different approaches such as global optimization [75, 76]. Then depth is

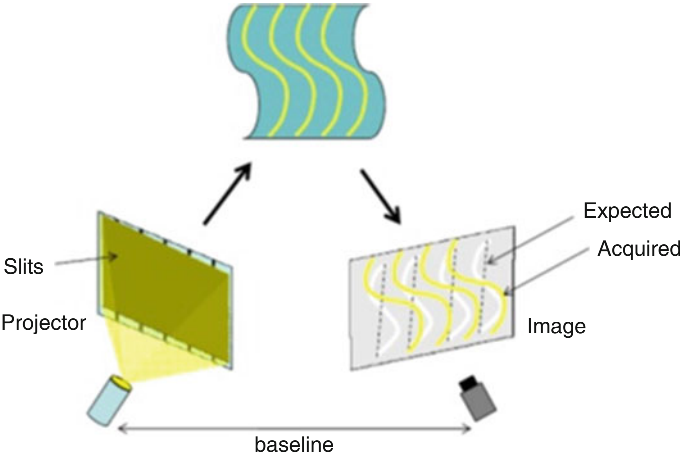
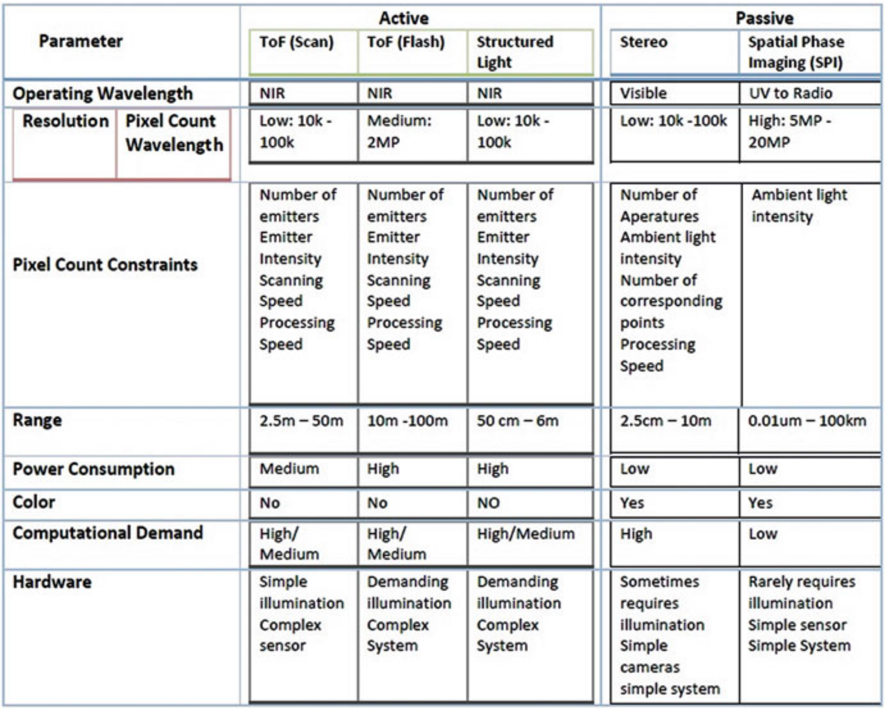


Fig. 3.8 Structured light system architecture [73]

inferred using triangulation [77–79]. Numerous pattern generation and structured light techniques exist in literature and are used in practice. A comprehensive list of structured light techniques has been stated by Jason Geng published in *Advances in Optics and Photonics* [80]. According to their research, structured light technique is categorized into five main categories: (1) sequential projections (multi-shots), (2) continuous varying pattern, (3) stripe indexing, (4) grid indexing, and (5) hybrid methods.

Structured light and photometric stereo both are active depth sensing technology. Through the advances of solid-state physics and micro lens technology, miniaturized depth sensor is now possible, and this technology is expected to improve further in terms of image resolution. A comprehensive comparison of depth sensing imaging technology is presented in Fig. 3.9.

Alternatively, the backbone of the image sensing technology beyond CMOS can also be modified to estimate depth. Recent research shows that new form of image sensor is capable of estimating depth directly from incident light [81]. Pixel aperture and depth from defocus are evaluated to construct a camera sensor. One of the merits of this sensor design is the control of pixel aperture-controlled pixel array design. Within a single die approach can capture blurred and sharp image. At a certain distance, a camera can produce a sharp image, which depends on camera focal setting. Otherwise, it produces a blurred image due to defocus. How much image blurring has occurred gives a cue to estimate depth [82, 83]. This approach is known as the depth from defocus. It requires two images, one sharp image and other one is blur image. Their constructed image sensor uses two different filters. A color filter is used to construct a blurred image, and a white filter is used to produce a sharp image. Both these images are then used to estimate depth. Pixel aperture controls the incident light. It can block, partially block, or pass whole amount of light to the image plane.



Parameter	Active			Passive	
	ToF (Scan)	ToF (Flash)	Structured Light	Stereo	Spatial Phase Imaging (SPI)
Operating Wavelength	NIR	NIR	NIR	Visible	UV to Radio
Resolution Pixel Count Wavelength	Low: 10k - 100k	Medium: 2MP	Low: 10k - 100k	Low: 10k -100k	High: 5MP - 20MP
Pixel Count Constraints	Number of emitters Emitter Intensity Scanning Speed Processing Speed	Number of emitters Emitter Intensity Scanning Speed Processing Speed	Number of emitters Emitter Intensity Scanning Speed Processing Speed	Number of Aperatures Ambient light intensity Number of corresponding points Processing Speed	Ambient light intensity
Range	2.5m – 50m	10m -100m	50 cm – 6m	2.5cm – 10m	0.01um – 100km
Power Consumption	Medium	High	High	Low	Low
Color	No	No	NO	Yes	Yes
Computational Demand	High/ Medium	High/ Medium	High/Medium	High	Low
Hardware	Simple illumination Complex sensor	Demanding illumination Complex System	Demanding illumination Complex System	Sometimes requires illumination Simple cameras simple system	Rarely requires illumination Simple sensor Simple System

Fig. 3.9 Comparison of 3D imaging techniques. Figure source is Photon-X [8]

Pekkola Oili et al. presented focus-induced photoresponse technique to measure distance [84]. Their approach is based on photoresponse materials such as dye-sensitized solar cells (DSSC) and optics. Photoresponse of a photosensor depends on the amount of incident photons and the surface area in which they fall [84]. The authors referred it as focus-induced photoluminescence (FIP) effect. In their technique, they use this property with the combination of lens, and they successfully derived the distance from the FIP effect. As shown in Fig. 3.10, they presented a single-pixel measurement technique, and to retrieve full geometry of an object, scan needs to be performed over the whole surface.

When lights fall on photodiode, it then generates photocurrent. FIP effect expresses the amount of light in terms of photocurrent. However, ambiguity arises when light radiant power is unknown. Their system arrangement consists of two photodiodes, and one lens is placed in front of the system that collects rays as it is mentioned in the figure. Instead of single photocurrent, their approach uses photocurrent ratio of two sensors to alleviate this situation. Moreover, their work shows the quotient changes with distance.

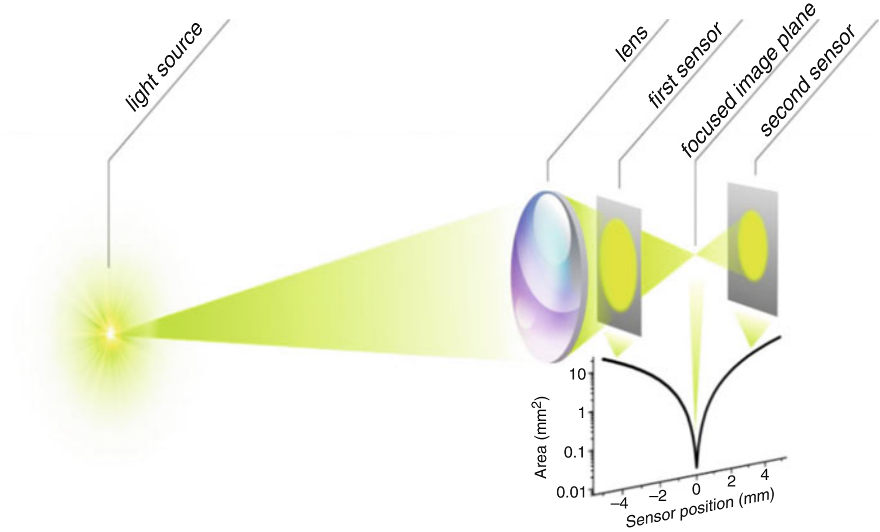


Fig. 3.10 FIP-based distance measurement technique [84]

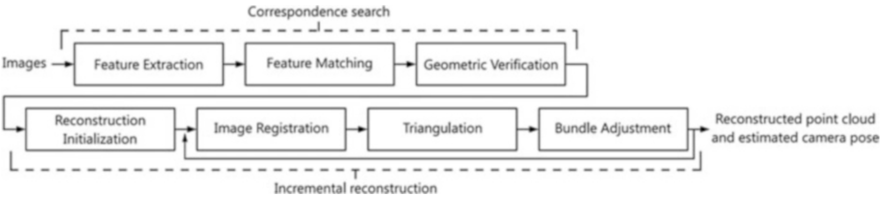


Fig. 3.11 Overview of structure from motion incremental pipeline. Input images are captured from different view angle. Figure is obtained from S. Bianco et al. [87]

3.3.3 Shape from Motion

Structure from motion (SfM) is one of the mature techniques to reconstruct a shape from a sequence of images. Some commercial 3D rendering softwares already adopt this approach to construct the 3D shape of an object [11, 85, 86, 103]. In this technique, motion is used to infer the depth of a scene. The concept behind SfM is shown in Fig. 3.11. Here motion means a scene is observed from a different angle of view. Generally, under orthographic camera model, at least three image sequences are used to estimate depth. Although various shape from motion algorithms exist, this chapter only focuses on the state of the art of shape from motion pipeline.

Using motion, multi-view images, structure from motion is the technique to reconstruct 3D shape of an object and simultaneous estimation of camera pose. SfM takes a series of input images from different camera view (motion). It is a sequential processing pipeline that iteratively estimates motion and shape. The first stage corresponds to the feature extraction process. Local features are extracted from every frame, and the extracted features matched within the overlapped image pair. Correspondence outliers are filtered out via random sample consensus (RANSAC) and bundle adjustment. Projective geometry is used to verify matched features. Different geometries and parameters such as homography, camera fundamental matrix, epipolar geometry, perspective n point, and triangulation are used to reconstruct shape. Highly overlapped images are a good candidate to improve its efficiency.

Feature extraction and matching action are then performed over the pair of overlapped images. Observed images are taken at different angles; hence, view angle and illumination condition affect the overlapped images. Thus, a feature that is observed from one view angle may not become visible from another view angle due to loss of illumination characteristics such as edge property. Moving from one view angle to another, same features have the probability to compromise its dimensional characteristics. Scaling factors may affect the matching process. Feature points are the key elements that describe scene context; hence, more feature points are desirable. In the last decade, scale-invariant feature transformation [88] approach has been widely adopted in this context because it is robust to noise.

3.4 Deep Learning Approaches to 3D Vision

Deep learning has gained much success in complex computer vision problems [86, 89–94]; recently it has been used to solve 3D reconstruction problem [104, 108, 110, 111]. Multilayer perception and its capacity to infer knowledge in 3D reconstruction domain have been deliberately used to solve different problems in different approaches. Considering shape from motion approach, features of an image sequence has great impact. Often low-textured and salient features are hard to extract. Convolutional neural network (CNN) has been used in this domain, and this approach significantly shows better performance compared to other feature extraction methodologies such as SIFT and DoG in a different environment [86]. Similar problem has been addressed by a deep learning context, and it shows a significant improvement [89, 93] in estimating pose. However, a more sophisticated approach has been developed, and a full network has been developed to solve structure from motion problems [90, 91]. Moreover, these approaches solve conventional structure from motion such as small camera translation problem [91, 95]. CNN is also used to infer depth that comes from the technique such as the depth from defocus [95]. Mainly it improves the depth uncertainty problem. Stereo matching problem also reffered as a finding correspondences. Deep learning has been well studied to solve both problems related to passive stereo vision: (1) finding feature and (2) finding correspondences [96–98].

Monocular or single depth image also has a great impact on computer vision as well as on robotics. SLAM is widely used to solve robot localization problem. SLAM depth is conventionally based on structure from motions that are limited to low texture and small translation [90, 99]. Improvement has been observed when deep learning is used to estimate depth especially when a low-textured region has been considered [99]. Convolutional neural network-conditional random field (CNN-CRF) framework for monocular endoscopy can estimate depth with relative error 0.152 and 0.242 on real endoscopic images [92].

Photometric stereo and structured light have been widely used in many areas. Moreover, object or camera in motion and surface property estimation such as Lambertian or non-Lambertian are the challenges and limitations of photometric stereo. Deep learning has been used to estimate surface normal vector which is the rudimentary step of photometric stereo before calculating depth [51, 90, 94]. Deep learning is also used to estimate depth in a supervised and unsupervised manner [51, 90, 94, 100]. In a supervised manner, the network needs to be trained with known data set and its ground truth depth map. In an unsupervised manner, depth can be estimated from both monocular and binocular views. It opens a freedom in such a way that, even if a stereo arrangement fails, network remains active and provides depth from any single image. The idea of this approach is to predict stereo images in a sense that for an input image say left image L, a network is trained in such a way that it can predict a disparity map [101]. Depth can be calculated from predicted disparity using triangulation method with a known baseline that is used to train the network. Several smoothing functions are used to reduce prediction error and noises.

Though deep network shows high accuracy [101] result, it cannot predict the depth of an unknown object shape which is not used during the training [90]. Also, the deep network needs a well-trained network to estimate depth in real-world environment.

3.5 Conclusion

Current sensors are able to achieve depth resolution from few centimeters to 100 m in real time, and sensor technologies like ToF, structured light, and stereo vision largely form the backbone of object detection and range finding applications in robotics and autonomous systems. Extraction of depth information from computational techniques is yet another growing area of research, and approaches like shape from shading and structure from motion offer some advantages in sensor design. Ambient light spectrum and light intensity planes play an important role in getting a dense depth map, and often lighting conditions experienced in complex environments contaminate depth estimation process. Demands on illumination pattern and computation limit the role of certain depth sensing mechanisms to static or less mobile platforms, and one sensor might not be a good fit. New sensing architectures and neuromorphic approaches to sensor design are already in progress to simplify some of these challenges. Ideally, miniature sensors with low power consumption and computational demands that can combine depth as well as accurate color information are preferred. The ability to add multi-spectral imaging on depth sensors is another area of interest and fusion of depth from different sensor technologies would solve some of the challenges in achieving robust vision for aerial, marine, and medical robotics.

References

1. Schauwecker, K., & Zell, A. (2014). On-board dual-stereo-vision for the navigation of an autonomous MAV. *Journal of Intelligent and Robotic Systems: Theory and Applications, 74*(1–2), 1–16.
2. Di Stefano, L., Clementini, E., & Stagnini, E. (2017). Reactive obstacle avoidance for multicopter UAVs via evaluation of depth maps. In *13th International Conference on Spatial Information Theory*.
3. Massimiliano, I., & Antonio, S. (2018). Path following and obstacle avoidance for an autonomous UAV using a depth camera. *Robotics and Autonomous Systems, 106*, 38–46.
4. Elaiwat, S., Bennamoun, M., Boussaid, F., & El-Sallam, A. (2014). 3-D face recognition using curvelet local features. *IEEE Signal Processing Letters, 21*, 172–175.
5. Maturana, D., & Scherer, S. (2015). VoxNet: A 3D convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 922–928).
6. Schwarz, M., Schulz, H., & Behnke, S. (2015). RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1329–1335).
7. Song, S., Lichtenberg, S. P., & Xiao, J. (2015). SUN RGB-D: A RGB-D scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 567–576).
8. Retrieved from http://www.photon-x.co/
9. ToF sensors. Retrieved from http://www.ti.com/sensors/specialty-sensors/time-of-flight/
10. NanEye Stereo web. Retrieved from https://ams.com/3d-sensing
11. Vélez, A. F. M., Marcinczak, J. M., & Grigat, R. R. (2012). Structure from motion based approaches to 3D reconstruction in minimal invasive laparoscopy. In A. Campilho & M. Kamel (Eds.), *Image analysis and recognition*. Berlin: Springer.
12. Xia, Y., Xu, W., Zhang, L., Shi, X., & Mao, K. (2015). Integrating 3d structure into traffic scene understanding with RGB-D data. *Neurocomputing, 151*, 700–709.
13. Wang, D., Wang, B., Zhao, S., & Yao, H. (2017). View-based 3D object retrieval with discriminative views. *Neurocomputing, 151*, 612–619.
14. Kokkonis, G., Psannis, K. E., Roumeliotis, M., et al. (2017). Real-time wireless multisensory smart surveillance with 3D-HEVC streams for internet-of-things (IoT). *The Journal of Supercomputing, 73*, 1044.
15. Santana, J. M., Wendel, J., Trujillo, A., Suárez, J. P., Simons, A., & Koch, A. (2017). Multimodal location based services—Semantic 3D city data as virtual and augmented reality. In G. Gartner & H. Huang (Eds.), *Progress in location-based services 2016*. Berlin: Springer.
16. Du, X., Allan, M., Dore, A., et al. (2016). Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery, 11*, 1109–1119.
17. Alaraimi, B., El Bakbak, W., Sarker, S., et al. (2014). A randomized prospective study comparing acquisition of laparoscopic skills in three-dimensional (3D) vs. two-dimensional (2D) laparoscopy. *World Journal of Surgery, 38*, 2746–2752.
18. Sørensen, S. M. D., Savran, M. M., Konge, L., et al. (2016). Three-dimensional versus two-dimensional vision in laparoscopy: A systematic review. *Surgical Endoscopy, 30*, 11–23.
19. Velayutham, V., Fuks, D., Nomi, T., et al. (2016). 3D visualization reduces operating time when compared to high-definition 2D in laparoscopic liver resection: A case-matched study. *Surgical Endoscopy, 30*, 147–153.
20. Hirschmuller, H., & Scharstein, D. (2007). Evaluation of cost functions for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis* (pp. 1–8).
21. Hosni, A., Bleyer, M., Rhemann, C., Gelautz, M., & Rother, C. (2011). Real-time local stereo matching using guided image filtering. In *IEEE International Conference on Multimedia and Expo* (pp. 1–6).

22. Domański, M., et al. (2015). Fast depth estimation on mobile platforms and FPGA devices. In *3DTV-Conference: The True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON)* (pp. 1–4).
23. Fan, Y., Huang, P., & Liu, H. (2015). VLSI design of a depth map estimation circuit based on structured light algorithm. *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems, 23*, 2281–2294.
24. Flores-Fuentes, W., Rivas-Lopez, M., Sergiyenko, O., Rodríguez-Quiñonez, J. C., Hernández-Balbuena, D., & Rivera-Castillo, J. (2014). Energy center detection in light scanning sensors for structural health monitoring accuracy enhancement. *IEEE Sensors Journal, 14*(7), 2355–2361.
25. Bleyer, M., & Breiteneder, C. (2013). Stereo matching—State-of-the-art and research challenges. In G. Farinella, S. Battiato, & R. Cipolla (Eds.), *Advanced topics in computer vision. Advances in computer vision and pattern recognition*. London: Springer.
26. Ding, J., Du, X., Wang, X., & Liu, J. (2010). Improved real-time correlation-based FPGA stereo vision system. In *IEEE International Conference on Mechatronics and Automation* (pp. 104–108).
27. Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(11), 1330–1334.
28. Liu, X., Li, D., Liu, X., & Wang, Q. (2010). A method of stereo images rectification and its application in stereo vision measurement. In *Second IITA International Conference on Geoscience and Remote Sensing* (pp. 169–172).
29. Santana-Cedrés, D., et al. (2017). Estimation of the lens distortion model by minimizing a line reprojection error. *IEEE Sensors Journal, 17*, 2848–2855.
30. Sousa, R. M., Wäny, M., Santos, P., & Morgado-Dias, F. (2017). NanEye—An endoscopy sensor with 3-D image synchronization. *IEEE Sensors Journal, 17*, 623–631.
31. Ascensão, B., Santos, P., & Dias, M. (2018). Distance measurement system for medical applications based on the NanEye stereo camera. In *International Conference on Biomedical Engineering and Applications (ICBEA)* (pp. 1–6).
32. Rodríguez-Quiñonez, J. C., Sergiyenko, O., Flores-Fuentes, W., Rivas-lopez, M., Hernandez-Balbuena, D., Rascón, R., & Mercorelli, P. (2017). Improve a 3D distance measurement accuracy in stereo vision systems using optimization methods' approach. *Opto-Electronics Review, 25*(1), 24–32.
33. Fusiello, A., & Trucco, E. (2000). Verri, a compact algorithm for rectification of stereo pairs. *Machine Vision and Applications, 12*, 16–22.
34. Kumar, S., Micheloni, C., Piciarelli, C., & Foresti, G. L. (2010). Stereo rectification of uncalibrated and heterogeneous images. *Pattern Recognition Letters, 31*, 1445–1452.
35. Hamzah, R. A., Ibrahim, H., & Hassan, A. H. A. (2016). Stereo matching algorithm for 3D surface reconstruction based on triangulation principle. In *1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (pp. 119–124).
36. Rivera-Castillo, J., Flores-Fuentes, W., Rivas-López, M., Sergiyenko, O., Gonzalez-Navarro, F. F., Rodríguez-Quiñonez, J. C., et al. (2017). Experimental image and range scanner datasets fusion in SHM for displacement detection. *Structural Control and Health Monitoring, 24*(10), e1967.
37. Real-Moreno, O., Rodriguez-Quiñonez, J. C., Sergiyenko, O., Basaca-Preciado, L. C., Hernandez-Balbuena, D., Rivas-Lopez, M., & Flores-Fuentes, W. (2017, June). Accuracy improvement in 3D laser scanner based on dynamic triangulation for autonomous navigation system. In *Industrial Electronics (ISIE), 2017 IEEE 26th International Symposium on* (pp. 1602–1608). IEEE.
38. Atapour-Abarghouei, A., & Breckon, T. P. (2018). A comparative review of plausible hole filling strategies in the context of scene depth image completion. *Computers & Graphics, 72*, 39–58.
39. Yoon, K. J., Member, S., & Kweon, I. S. (2006). Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*, 650–656.

40. Hamzah, R. A., Rahim, R. A., & Noh, Z. M. (2010). Sum of absolute differences algorithm in stereo correspondence problem for stereo matching in computer vision application. In *3rd International Conference on Computer Science and Information Technology* (pp. 652–657).
41. Hirschmuller, H. (2005). Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 807–814).
42. Joglekar, J., Gedam, S. S., & Mohan, B. K. (2014). Image matching using SIFT features and relaxation labeling technique—A constraint initializing method for dense stereo matching. *IEEE Transactions on Geoscience and Remote Sensing, 52*, 5643–5652.
43. Hafner, D., Demetz, O., & Weickert, J. (2013). Why is the census transform good for robust optic flow computation? In A. Kuijper, K. Bredies, T. Pock, & H. Bischof (Eds.), *Scale space and variational methods in computer vision*. Berlin: Springer.
44. Huang, F., Huang, S., Ker, J., & Chen, Y. (2012). High-performance SIFT hardware accelerator for Real-time image feature extraction. *IEEE Transactions on Circuits and Systems for Video Technology, 22*, 340–351.
45. Garstka, J., & Peters, G. (2015). Fast and robust keypoint detection in unstructured 3-D point clouds. In: *12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)* (pp. 131–140).
46. Kechagias-Stamatis, O., & Aouf, N. (2016). Histogram of distances for local surface description. In *IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2487–2493).
47. Prakhya, S. M., Lin, J., Chandrasekhar, V., Lin, W., & Liu, B. (2017). 3D HoPD: A fast low-dimensional 3-D descriptor. *IEEE Robotics and Automation Letters, 2*, 1472–1479.
48. Brooks, M. J., & Horn, B. K. P. (1985). Shape and source from shading. In *Proc. Int. Joint Conf. Artificial Intelligence* (pp. 932–936).
49. Woodham, R. J. (1980). Photometric method for determining surface orientation from multiple images. *Optical Engineering, 19*, 139–144.
50. Sohaib, A., Farooq, A. R., Atkinson, G. A., Smith, L. N., Smith, M. L., & Warr, R. (2013). In vivo measurement of skin microrelief using photometric stereo in the presence of interreflections. *Journal of the Optical Society of America. A, 30*, 278–286.
51. Woodham, R. J. (1978). Photometric stereo: A reflectance map technique for determining surface orientation from image intensity. In *Image Understanding Systems and Industrial Applications*.
52. Mostafa, M. G., Yamany, S. M., & Farag, A. A. (1999). Integrating stereo and shape from shading. In *Int. Conf. on Image Processing* (pp. 130–134).
53. Prados, E., & Soatto, S. (2005). Fast marching method for generic shape from shading. In N. Paragios, O. Faugeras, T. Chan, & C. Schnörr (Eds.), *Variational, geometric, and level set methods in computer vision. VLSM. Lecture notes in computer science* (Vol. 3752). Berlin: Springer.
54. Lu, S., & Yuanyuan, W. (2017). Three-dimensional reconstruction of macrotexture and microtexture morphology of pavement surface using six light sources–based photometric stereo with low-rank approximation. *Journal of Computing in Civil Engineering, 31*, I. 2.
55. Antensteiner, D., Štole, S., & Pock, T. (2018). Variational fusion of light field and photometric stereo for precise 3D sensing within a multi-line scan framework. In *24th International Conference on Pattern Recognition (ICPR)* (pp. 1036–1042).
56. Ju, Y., Qi, L., Zhou, H., Dong, J., & Lu, L. (2018). Demultiplexing colored images for multispectral photometric stereo via deep neural networks. *IEEE Access, 6*, 30804–30818.
57. Hilliges, O., Weiss, M. H., Izadi, S., & Kim, D. (2018). *Using photometric stereo for 3D environment modeling*. US Patent.
58. Piatkowska, E., Kogler, J., Belbachir, N., & Gelautz, M. (2017). Improved cooperative stereo matching for dynamic vision sensors with ground truth evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu*.
59. Zhu, A. Z., Chen, Y., & Daniilidis, K. (2018). *Realtime time synchronized event-based stereo, arxiv*. arXiv:1803.09025.

60. Censi, A., & Scaramuzza, D. (2014). Low-latency event-based visual odometry. In *IEEE International Conference on Robotics and Automation (ICRA)* (pp. 703–710).
61. Gallego, G., Lund, J. E. A., Mueggler, E., Rebecq, H., Delbruck, T., & Scaramuzza, D. (2018). Event-based, 6-DOF camera tracking from photometric depth maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*, 2402–2412.
62. Ieng, S. H., Carneiro, J., Osswald, M., & Benosman, R. (2018). Neuromorphic event-based generalized time-based stereovision. *Frontiers in Neuroscience, 12*, 442.
63. Martel, J. N. P., Müller, J., Conradt, J., & Sandamirskaya, Y. (2018). An active approach to solving the stereo matching problem using event-based sensors. In *IEEE International Symposium on Circuits and Systems (ISCAS)*.
64. Leroux, T., Ieng, S. H., & Benosman, R. (2018). *Event-based structured light for depth reconstruction using frequency tagged light patterns, arxiv*. arXiv:1811.10771.
65. Piatti, D., Remondino, F., & Stoppa, D. (2013). State-of-the-art of TOF range-imaging sensors. In F. Remondino & D. Stoppa (Eds.), *TOF range-imaging cameras*. Berlin: Springer.
66. Edoardo, C., Matt, F., Richard, W., Robert, K. H., & Cristiano, N. (2013). Spad-based sensors. In *TOF range-imaging cameras* (pp. 11–38). Berlin: Springer.
67. Behroozpour, B., Sandborn, P. A. M., Wu, M. C., & Boser, B. E. (2017). Lidar system architectures and circuits. *IEEE Communications Magazine, 55*, 135–142.
68. Beer, M., Schrey, O. M., Nitta, C., Brockherde, W., Hosticka, B. J., & Kokozinski, R. (2017). 1×80 pixel SPAD-based flash LIDAR sensor with background rejection based on photon coincidence. *IEEE Sensors*, 1–3.
69. Albitar, C., Graebling, P., & Doignon, C. (2007). Robust structured light coding for 3D reconstruction. In *IEEE 11th Int. Conf. on Computer Vision* (pp. 1–6).
70. Lee, D., & Krim, H. (2010). 3D surface reconstruction using structured circular light patterns. In J. Blanc-Talon, D. Bone, W. Philips, D. Popescu, & P. Scheunders (Eds.), *Advanced concepts for intelligent vision systems. ACIVS*. Heidelberg: Springer.
71. Ma, S., Shen, Y., Qian, J., Chen, H., Hao, Z., & Yang, L. (2011). Binocular structured light stereo matching approach for dense facial disparity map. In D. Wang & M. Reynolds (Eds.), *AI 2011: Advances in artificial intelligence. AI 2011. Lecture notes in computer science* (Vol. 7106). Berlin: Springer.
72. Zhao, L., Xu, H., Li, J., & Cai, Q. (2012). Binocular stereo vision measuring system based on structured light extraction algorithm. In *2012 International Conference on Industrial Control and Electronics Engineering, Xi'an* (pp. 644–647).
73. Retrieved from https://www.aniwaa.com/best-3d-scanner/
74. Choo, H., Ribera, R. B., Choi, J. S., & Kim, J. (2011). Depth and texture imaging using time-varying color structured lights. In *2011 International Conference on 3D Imaging (IC3D)* (pp. 1–5).
75. Pages, J., Salvi, J., Collewet, C., & Forest, J. (2005). Optimised De Bruijn patterns for one-shot shape acquisition. *Image and Vision Computing, 23*(8), 707–720.
76. Tyler, B., Beiwen, L., & Song Z. (2016). *Structured light techniques and applications*. Wiley Online Library.
77. Slysz, R., Moreau, L., & Borouchaki, H. (2013). On uniqueness in triangulation based pattern for structured light reconstruction. In *International Conference on 3D Imaging* (pp. 1–6).
78. Rodrigues, M., Kormann, M., Schuhler, C., & Tomek, P. (2013). Structured light techniques for 3D surface reconstruction in robotic tasks. In R. Burduk, K. Jackowski, M. Kurzynski, M. Wozniak, & A. Zolnierek (Eds.), *8th International Conference on Computer Recognition Systems CORES*.
79. Van, L. T., & Huei, Y. L. (2018). A structured light RGB-D camera system for accurate depth measurement. *Hindawi International Journal of Optics*.
80. Geng, J. (2011). Structured-light 3D surface imaging: A tutorial. *Advances in Optics and Photonics, 3*, 128–160.
81. Choi, B.-S., et al. (2017). Pixel aperture technique in CMOS image sensors for 3D imaging. *Sensors and Materials, 29*(3), 235–241.
82. Pentland, A. P. (1987). A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 9*, 523–531.

83. Rajagopalan, A. N., Chaudhuri, S., & Mudenagudi, U. (2004). Depth estimation and image restoration using defocused stereo pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*, 1521–1525.
84. Oili, P., Christoph, L., Peter, F., Anke, H., Wilfried, H., Stephan, I., Christian, L., Christian, S., Peter, S., Patrick, S., Robert, S., Sebastian, V., Erwin, T., & Ingmar, B. (2017). *Focus-induced photoresponse: A novel optoelectronic distance measurement technique, arXiv*. arXiv:1708.05000.
85. Xu, X., Che, R., Nian, R., He, B., Chen, M., & Lendasse, A. (2016). Underwater 3D object reconstruction with multiple views in video stream via structure from motion. *OCEANS*, 1–5.
86. Aji, R. W., Akihiko, T., & Masatoshi, O. (2018). *Structure-from-Motion using Dense CNN Features with Keypoint Relocalization, arXiv.*, arXiv:1805.03879.
87. Bianco, S., Ciocca, G., & Marelli, D. (2018). Evaluating the performance of structure from motion pipelines. *Journal of Imaging, 4*(8), 98.
88. Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE Int. Conf. on Computer Vision* (pp. 1150–1157).
89. Arjun, J., Jonathan, T., & Yann, L., & Christoph, B. (2014). *MoDeep: A deep learning framework using motion features for human pose estimation, arXiv*. arXiv:1409.7963.
90. Benjamin, U., Huizhong, Z., Jonas, U., Nikolaus, M., Eddy, I., Alexey, D., & Thomas, B. (2017). DeMoN: Depth and motion network for learning monocular stereo. *Benjamin Ummenhofer, arXiv*. arXiv:1612.02401.
91. Sudheendra, V., Susanna, R., Cordelia, S., Rahul, S., & Katerina, F. (2017). *SfM-Net: Learning of structure and motion from video, arXiv*. arXiv:1704.07804.
92. Faisal, M., & Nicholas, J. D. (2018). *Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy* (pp. 230–243).
93. Gyeongsik, M., Ju, Y. C., & Kyoung, M. L. (2018). *V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map, arXiv*. arXiv:1711.07399.
94. Liang, L., Lin, Q., Yisong, L., Hengchao, J., & Junyu, D. (2018). Three-dimensional reconstruction from single image base on combination of CNN and multi-spectral photometric stereo. *Sensors, 18*(3).
95. Marcela, C., Bertrand, L. S.,Pauline, T.-P, Andrés, A, & Frédéric, C. (2018). *Deep depth from defocus: How can defocus blur improve 3D estimation using dense neural networks? arXiv*. arXiv:1809.01567.
96. Jure, Ž., & Yann, L. (2016). *Stereo matching by training a convolutional neural network to compare image patches, arXiv*. arXiv:1510.05970.
97. Luo, W., Schwing, A. G., & Urtasun, R. (2016). Efficient deep learning for stereo matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5695–5703).
98. Sameh, K., Sean, F., & Christoph, R. (2018). *StereoNet: Guided hierarchical refinement for real-time edge-aware depth prediction, arXiv*. arXiv:1807.08865.
99. Tateno, K., Tombari, F., Laina, I., & Navab, N. (2017). CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6565–6574).
100. Zhan, H., Garg, R., & Weerasekera, C. S. (2018). *Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction*.
101. Godard, C., Aodha, O. M., & Brostow, G. J. (2017). *Unsupervised monocular depth estimation with left-right consistency clement godard, arXiv*. arXiv:1609.03677.
102. Srivastava, S., Ha, S. J., Lee, S. H., Cho, N. I., & Lee, S. U. (2009). Stereo matching using hierarchical belief propagation along ambiguity gradient. In *16th IEEE International Conference on Image Processing (ICIP)* (pp. 2085–2088).
103. Westoby, M., Brasington, J., Glasser, N. F., & Hambrey, M. J. (2012). Structure-from-Motion photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology, 179*, 300–314. Elsevier.
104. Yichuan, T., Ruslan, S., & Geoffrey, H. (2012). *Deep Lambertian networks, arXiv*. arXiv:1206.6445

105. Visentini-Scarzanella, M., et al. (2015). Tissue shape acquisition with a hybrid structured light and photometric stereo endoscopic system. In X. Luo, T. Reichl, A. Reiter, & G. L. Mariottini (Eds.), *Computer-assisted and robotic endoscopy*. CARE, Springer.
106. Wang, W., Yan, J., Xu, N., Wang, Y., & Hsu, F. (2015). Real-time high-quality stereo vision system in FPGA. *IEEE Transactions on Circuits and Systems for Video Technology, 25*, 1696–1708.
107. Ttofis, C., Kyrkou, C., & Theocharides, T. (2016). A low-cost Real-time embedded stereo vision system for accurate disparity estimation based on guided image filtering. *IEEE Transactions on Computers, 65*, 2678–2693.
108. Xie, J., Girshick, R., & Farhadi, A. (2016). Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision*. Berlin: Springer.
109. Jalal, A., Kim, Y. H., Kim, Y. J., Kamal, S., & Kim, D. (2017). Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognition, 61*, 295–308. Elsevier.
110. Ma, F., & Karaman, S. (2018). Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *IEEE International Conference on Robotics and Automation (ICRA), Brisbane* (pp. 1–8).
111. Li, R., Wang, S., Long, Z., & Gu, D. (2018). UnDeepVO: Monocular visual odometry through unsupervised deep learning. In *IEEE International Conference on Robotics and Automation (ICRA), Brisbane* (pp. 7286–7291).
112. Yuanhong, X., Pei, D., Junyu, D., & Lin, Q. (2018). *Combining SLAM with multi-spectral photometric stereo for real-time dense 3D reconstruction, arxiv*. arXiv:1807.02294.