digital

MDPI

Review

On 3D Reconstruction Using RGB-D Cameras

Kyriaki A. Tychola *, Ioannis Tsimperidis and George A. Papakostas *

MLV Research Group, Department of Computer Science, International Hellenic University, 65404 Kavala, Greece
* Correspondence: kytzcho@cs.ihu.gr (K.A.T.); gpapak@cs.ihu.gr (G.A.P.)

Abstract: The representation of the physical world is an issue that concerns the scientific community studying computer vision, more and more. Recently, research has focused on modern techniques and methods of photogrammetry and stereoscopy with the aim of reconstructing three-dimensional realistic models with high accuracy and metric information in a short time. In order to obtain data at a relatively low cost, various tools have been developed, such as depth cameras. RGB-D cameras are novel sensing systems that capture RGB images along with per-pixel depth information. This survey aims to describe RGB-D camera technology. We discuss the hardware and data acquisition process, in both static and dynamic environments. Depth map sensing techniques are described, focusing on their features, pros, cons, and limitations; emerging challenges and open issues to investigate are analyzed; and some countermeasures are described. In addition, the advantages, disadvantages, and limitations of RGB-D cameras in all aspects are also described critically. This survey will be useful for researchers who want to acquire, process, and analyze the data collected.

Keywords: RGB-D camera; 3D reconstruction; depth image processing; camera pose estimation; stereo matching; point cloud mapping

Citation: Tychola, K.A.; Tsimperidis, I.; Papakostas, G.A. On 3D Reconstruction Using RGB-D Cameras. *Digital* **2022**, *2*, 401–421. https://doi.org/10.3390/digital2030022

Academic Editor: Mikael Sjödahl

Received: 11 July 2022
Accepted: 9 August 2022
Published: 13 August 2022

1. Introduction

One of the main tasks that computer vision is dealing with is the 3D reconstruction of the real world [1]. In computer science, three-dimensional reconstruction means the process of capturing the shape and appearance of real objects. This process is accomplished by active and passive methods, and for this purpose, several tools have been developed and applied. In the last decade, an innovative technology has emerged: depth cameras (RGB-D). The elementary issue is "object detection," which refers to recognizing objects in a scene and is divided into instance recognition and category level recognition. Object recognition depends highly on RGB cameras and instance-specific information [2], whereas the quality of the recognized category depends on the generalization of the properties or functionalities of the object and the unseen instances of the same category. Although depth cameras provide 3D reconstruction models in real time, one of the main issues for researchers is robustness and accuracy [3]. In addition, the representation of the object can undergo changes such as scaling, translation, occlusion, or other deformations, which make category level recognition a difficult topic. In addition, object detection has weaknesses due to illumination, camera viewpoint, and texture [4]. The recovered information is expressed in the form of a depth map, that is, an image or image channel that contains information relating to the distance of objects' surfaces from the capturing camera. Depth maps are invariant to texture and illumination changes [5]. The modeling of objects can be categorized into *geometric* and *semantic*. Geometric modeling provides an accurate model related to geometry, whereas semantic analyzes objects in order to be understood by humans. A typical example of semantic information is the integration of RGB-D in people's daily life for space estimation (odometry), object detection, and classification (doors, window, walls, etc.). RGB-D camera have continuous detailed feedback for the existing configuration area, and this is especially helpful for visually impaired people. The RGB-D camera is a navigational aid both indoors and outdoors, in addition to the classic

aids they use, such as walking stick. In particular, in recent years, with the emergence of the COVID-19 pandemic, people have been spending more time in the household, and this is an issue mostly for people in this category. In this context, algorithms for combining RGB cameras and depth cameras have been developed that allow audio messages about the distance of objects from humans [6]. In addition, to ensure their real-time performance, the systems are accelerated by parallel offloading functions to the GPU [7].

Moreover, objects are represented either *dynamically* [8] or *statically* [9], where scenes with rapid movements or complex topology changes and mapping area are captured, respectively. Three-dimensional reconstruction is applied in various fields, such as natural disasters [10], healthcare [11], heritage [12], and so on. Using various camera techniques and methods, scenes are captured in both dynamic and static environments. For a dynamic scene, there are two categories with their own advantages and limitations. One is related to tracking the motion of the object (fusion-based methods), and the other to modelling a reconstruction without taking into account the number of photos taken simultaneously (frame-based methods) [13]. Detailed and complete analysis of various scenes is of major importance, especially for robotic applications where different environments exist. In these cases, semantic segmentation methods are applied, which enhance various tasks, such as semantically assisted person perception, (semantic) free space detection, (semantic) mapping, and (semantic) navigation [14]. Depth images provide complementary geometric information to RGB images, which can improve segmentation [15]. However, the integration of depth information becomes difficult, because depth introduces deviating statistics and characteristics of a different kind modality [16]. For the semantic recording, understanding, and interpretation of a scene, several methods were developed that were not particularly successful due to the increased memory consumption required [17–22]. To solve this issue, new model architectures have been proposed, such as a shape-aware convolutional layer (ShapeConv) [23], separation-and-aggregation gate (SA-Gate) [24], attention complementary network (ACNet) [25], RFne [26,27], and CMX, [28], as well as methods that focus on extracting modality-specific features in order to ensure that the best features are extracted without errors [29–32]. The result of objects' 3D reconstruction with depth cameras is a depth map that suffers from some limitations of this technology, such as sensors' hardware, errors of the pose, and low-quality or low-resolution capture. In short, RGB-D cameras are a new technology and different from classical RGB cameras in the sensors' integration. Although their advantages are many, there are still some limitations regarding geometric errors, camera trajectory, texture, and so on [33]. Moreover, deep-learning-based reconstruction methods and systems are used in various applications, such as human action and medical images, and they directly learn from image data to extract features [34–37]. ANNs are the foundation of deep learning techniques.

To achieve a variety of applications, databases are open to anyone interested in experimenting. Databases contain small and large datasets depending on the application desired [38].

In addition, nowadays, there is a discussion about fusion sensors, that is, the use of different sensors, such as RGB-D and thermal cameras, LiDAR and IMU sensors, and radar, with the aim of improving results with the minimum number of sensors and minimum system complexity for the lowest cost. Sensors operate independently but can be combined to produce a more complete, accurate, and dependable picture. Moreover, they provide increased data quality and reliability, cover larger areas, and estimate unmeasured states [39].

The rest of the paper is organized as follows: Section 1 provides an overview of the search problem, highlights the scope and contribution of the paper, and reports and compares similar surveys. It also describes the methodology of the research strategy. Section 2 gives an overview of the RGB-D technology timeline. Section 3 discusses the hardware and technology of RGB. Section 4 describes and discusses the concept and approaches of 3D reconstruction techniques, on both static and dynamic scenes, and RGB-D SLAM methods for 3D reconstruction. Section 5 presents data acquisition methods, describes some common databases, and focuses on the depth map as the final "product" and its limitations, and pro-

poses countermeasures. Section 6 presents pros, cons, and limitations of RGB-D technology, in various aspects. Finally, Section 7 evaluates RGB-D cameras and presents the conclusions.

1.1. Motivation

The representation of the physical world through realistic three-dimensional models has gained much interest in recent years and is therefore one of the subjects of research in computer science. The models are reconstructed with various modern and automated technologies. Relatively recently, depth cameras have been developed that provide real-time 3D models.

1.2. Scope and Contribution

The research carried out to date refers to various techniques and methods of 3D object reconstruction, while others discuss the differences between depth cameras and optimization to texture mapping algorithms for static scenes. This paper, through a bibliographic review, clarifies terms such as depth cameras and depth maps, and describes the construction and technology under which the RGB-D camera operates, as well as its applications. In addition, emphasis is placed on the research conducted, and the advantages and limitations of this technology are analyzed. The contribution of this paper is to discuss the state of the art and current status of RGB-D cameras. Hence, the novelty of this paper, compared to all other reviews, lies in the holistic approach to depth camera technology. In particular, in this research, the concepts (terms) of the RGB-D camera are clarified, its technology and operation are described in detail, the functions of the incorporated sensors are categorized, the process of 3D reconstruction models is explained in detail, the weaknesses and limitations that they present (especially from a technological point of view but also due to their principle of operation) are highlighted, some important datasets are listed, all of the above are evaluated, and the cons and limitations created due to a variety of factors are highlighted. In addition, appropriate countermeasures to solve any problem are proposed, and finally, challenges and open issues that require further study and research are identified.

The research questions are summarized as follows:

- What is RGB-D camera technology?
- What kind of data is acquired from the RGB-D camera?
- What algorithms are applied for applications?
- What are the benefits and limitations of the RGB-D camera?
- Why is a depth map important?

1.3. Related Work

Previous work on RGB-D camera technology has several aspects; however, there is no work that fully describes depth camera technology from all points of view. Existing research focuses on various RGB-D applications and applies experiments with various techniques and methods to address and resolve the respective constraints arising from this technology. In particular, the research done to date can be categorized into experiments involving data acquisition [40,41], noisy data [42–44], high quality of depth maps [43,45], object recognition in both static and dynamic scenes [46–51], pose estimation by robots [52–58], human–machine interaction [59], 3D reconstruction models in real time [60–65], and solving problems created by the environment (mainly outdoor environments), such as sharp lighting changes [66,67] and occlusion (repetitive textures and structures) [68–70].

Hence, the novelty of this paper lies in the holistic approach to depth camera technology. In particular, in this research, the concepts (terms) of the RGB-D camera are clarified, its technology and operation are described in detail, the functions of the incorporated sensors are categorized, the process of 3D reconstruction models is explained in detail, the weaknesses and limitations that they present (especially from a technological point of view but also due to their principle of operation) are highlighted, all of the above are evaluated, and the cons and limitations created due to a variety of factors are highlighted. In addition,

appropriate countermeasures to solve any problem are proposed, and finally, challenges and open issues that require further study and research are identified.

1.4. Methodology of Research Strategy

The literature search was achieved by identifying the appropriate criteria used by other papers published in peer-reviewed journals or peer-reviewed conference proceedings, with keywords according to the title in English.

Data Sources and Search

The desired information extraction from existing surveys was done in the online bibliographic database Scopus, which includes the most important digital libraries (Elsevier, Springer, IEEE), provides a refined search, and facilitates the export of files. Below, the query that was performed in Scopus for the period 2011–2022 is described:

TITLE-ABS-KEY (rgb-d AND cameras) AND TITLE-ABS-KEY (depth AND cameras).

The query results were saved in a CSV file. The process revealed 1869 documents. Of these, however, only journal articles, conference papers, and book chapters were selected, thus reducing the number to 1827. Of the remaining, only 1758 were written in English, and some of these were not ultimately relevant to the subject, while for others, only the abstract was available. Finally, 124 publications were selected as the most representative and influential and discussed in this study. Figure 1 shows the process by which information was obtained from the literature database, while Figure 2 shows the percentage of each type of publication in the papers selected.

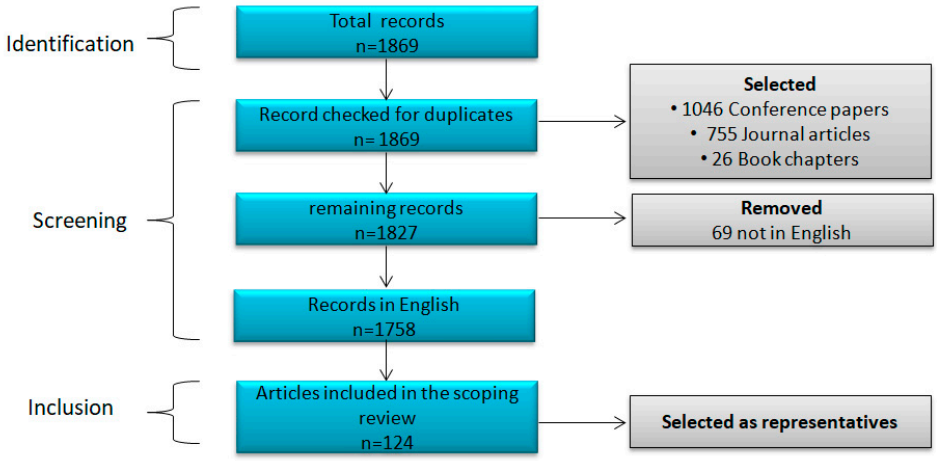

Figure 1. Process of obtaining information from the literature databases.

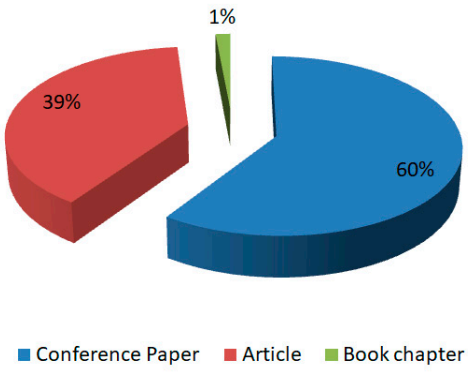

Figure 2. The percentage of each type of publication in the papers selected.

The chart on Figure 3 shows the number of published papers per year. The papers start from 2011 and continue until 2017, with the researchers' interest showing a slight

upward trend. After a small decline in the number of papers per year, in 2016, there was a continuous increase again, and 2019 recorded the largest number of publications.

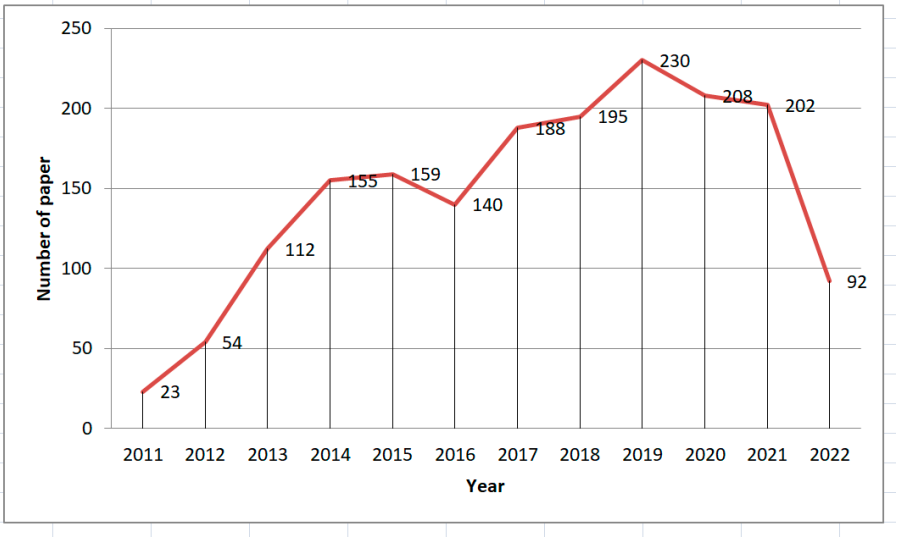

Figure 3. RGB-D camera papers per year.

Figure 4 lists the top 10 countries that have published on 3D reconstruction with RGB-D depth cameras, with China standing out by publishing about a third of the total documents.

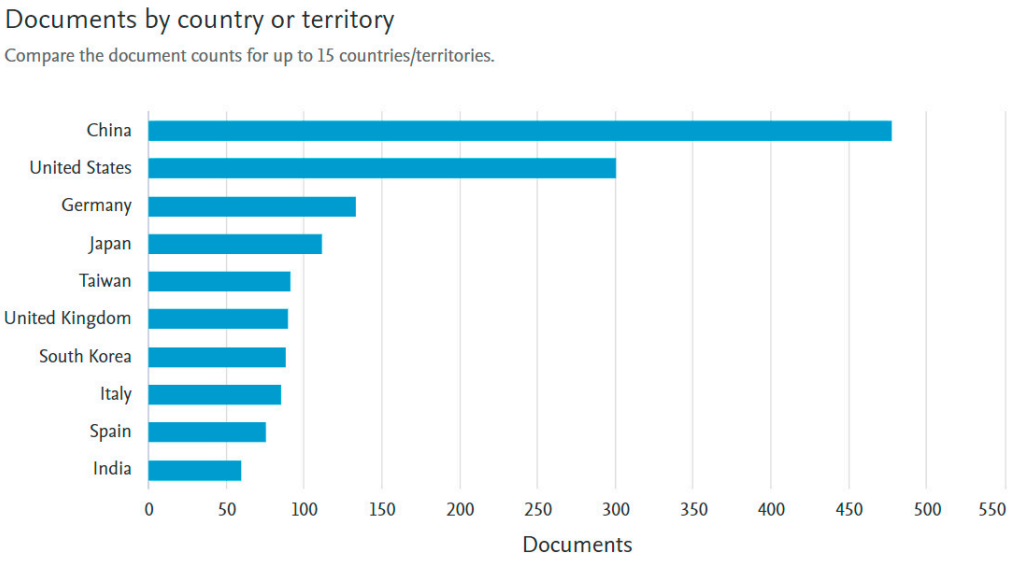

Figure 4. Countries that have published (Scopus source) papers related to RGB-D depth cameras.

2. History of RGB and 3D Scene Reconstruction

The technology of time-of-flight (ToF) cameras with 3D imaging sensors that provide a depth image and an amplitude image with a high frame rate has developed rapidly in recent years. Figure 5 shows the historical course of the development and evolution of depth cameras.

Depth cameras were developed in the last decade; however, the foundations were laid in 1989. Some milestones in the history of 3D Reconstruction and RGB-D technology are as follows: In the 1970s, the idea of 3D modeling and the significance of object shape were introduced. In the 1980s, researchers focused on the geometry of objects. From the 2000s, various techniques and methods related to the features and textures of objects

and scenes were developed. In the 2010s, appropriate algorithms were developed and implemented in applications, which mainly include dynamic environments and robotics. In this framework, deep learning methods are used that provide satisfactory models with high accuracy. Nowadays, research is focused on ways of solving the existing limitations associated with quality fusion of the scenes.

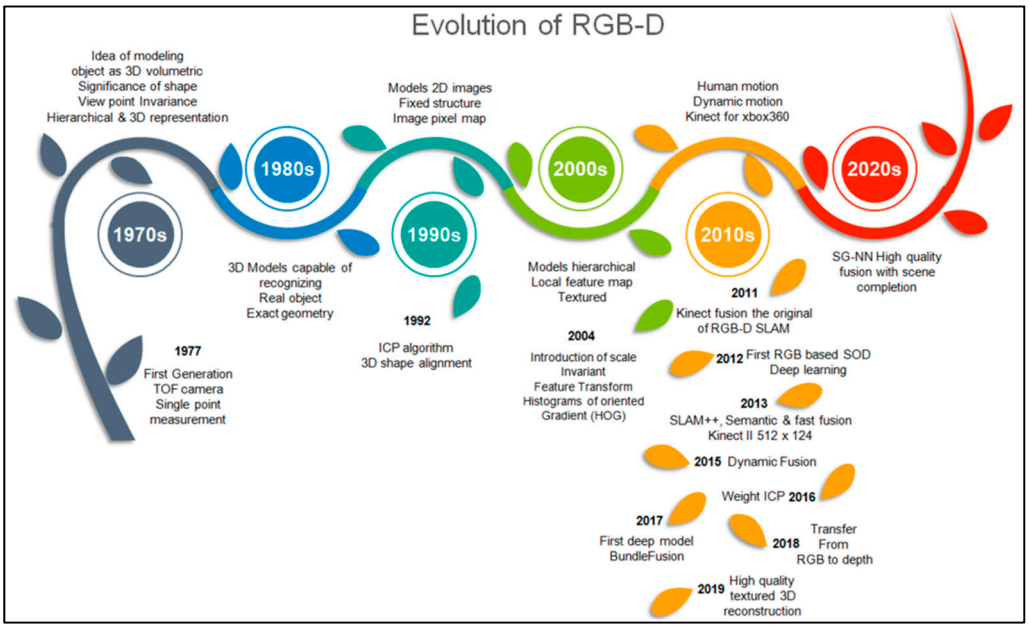

Figure 5. Timeline progression diagram of RGB-D cameras.

3. Hardware and Basic Technology of RGB-D

Depth cameras are vision systems that alter the characteristics of their environment, mainly visual, in order to capture 3D scene data from their field of view. These systems have structured lighting, which projects a known pattern on the scene and measures its distortion when viewed from a different angle. This source may use a wavelength belonging to the visible field, but more commonly, will be selected from the infrared field. The more sophisticated systems implement the time of flight (ToF) technique, in which the return time of a light pulse after reflection by an object in the scene captures depth information (typically over short distances) from a scene of interest [71].

A typical depth camera (RGB-D) incorporates an RGB camera, a microphone, and a USB port for connection to the computer. In addition, it includes a depth sensor, which uses infrared structured light to calculate the distance of each object from the camera's horizontal optical axis (depth). In order to achieve this, it takes into account each point of the object being studied. In addition, some cameras have an infrared emitter (IR emitter) that consists of an IR laser diode to beam modulated IR light to the field of view. The reflected light is collected by the depth sensor and an infrared absorber (IR sensor) mounted anti-diametrically. RGB-D sensors are a specific type of depth-sensing device that work in association with an RGB (red, green, blue color) sensor camera. They are able to augment the conventional image with depth information (related with the distance to the sensor) on a per-pixel basis. The depth information obtained from infrared measurements is combined with the RGB image to yield an RGB-D image. The IR sensor is combined with an IR camera and an IR projector. This sensor system is highly mobile and can be attached to a mobile instrument such as a laptop [72] (see Figure 6).

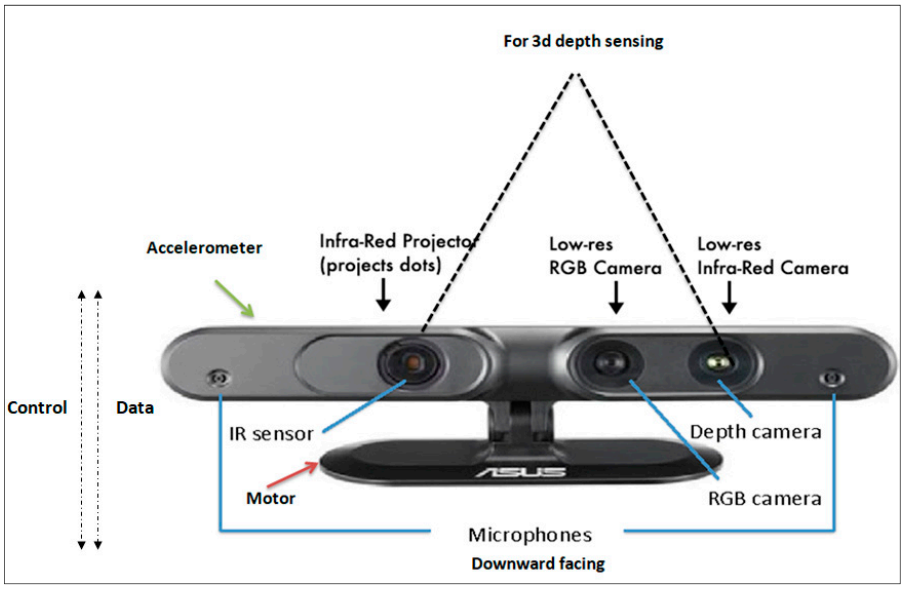

Figure 6. Example of RGB-D camera and its components. Asus Xtion PRO LIVE. (Source: https://www.asus.com/supportonly/Xtion%20PRO/HelpDesk_Manual/ (accessed on 12 July 2022)).

As for how it works, it emits a pre-defined pattern of infrared light rays. The light is absorbed by existing objects, and the depth sensors measure it. Since the distance between the emitter and sensor is known, from the difference between the observed and expected position, the depth measurement, with respect to the RGB sensor, is taken for each pixel. Trigonometric relationships are used for this purpose (see Figure 7).

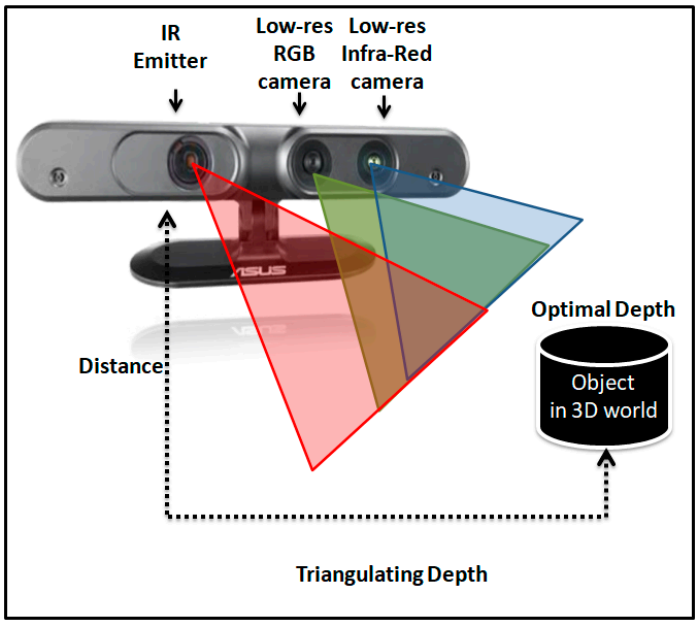

Figure 7. Diagram capture object from RGB-D camera.

Figure 7 illustrates the processes of RGB-D camera operation. RGB and infrared cameras capture one scene at the same time. Through this process, the information is visualized, and a 2D monochrome digital image is created, in which the color of each pixel indicates the distance of the homologous point (key point) from the camera. Dark shades indicate near-camera objects while light ones indicate otherwise.

Like all technologies, RGB-D camera technology has certain limitations. To address these, various techniques and methods have been devised and developed. For example,

it often suffers from specific noise characteristics and data distortions [73,74]. In general, although depth cameras operate with the same technology, they show differences related to the camera's resilience against background and structured light [75], but almost all of them have a low cost.

4. Conceptual Framework of 3D Reconstruction

The 3D representation of the physical world is the core and basic purpose of computer vision. The term 3D representation refers to the mapping and three-dimensional reconstruction of a region (scene), including its objects. This can be achieved through various techniques and methods such as *active*, *passive*, or *hybrid* (i.e., combination of active and passive), *monocular* or *stereoscopic* (stereo vision), and techniques based either on the *depth* or *content* of the image. The use of depth cameras for this purpose is a valuable tool, as they capture the scene of indoor and outdoor environments in real time. How is the reconstruction achieved, though? Firstly, the information is collected by the depth camera, and then, with the help of appropriate algorithms, it is processed to obtain the final result, which is a reconstructed object that combines disparate information into a textured 3D triangle mesh. For this purpose, many algorithms have been developed, among which the more frequently used are bundle fusion [76], voxel hashing [77], SIFT, SURF [78], FAST [79], ORB [80], RANCSANC [81,82], MVS [83], ICP [84], and signed distance function (SDF) [85,86]. Each algorithm contributes to each stage of the scene reconstruction. Table 1 presents the functions of these algorithms.

Table 1. Main algorithms of computer vision for 3D reconstruction model.

Algorithms	Features/Operation
Bundle Fusion	Optimization pose algorithm
Voxel Hashing	Real-time reconstruction and easy-to detect scene changes
SIFT	Provides uniform scaling and orientation, illumination changes and rotation, measures key points DoG, image translation, uses affine transformation, descriptor type integer vector
SURF	Allows a rapid differentiation of light characteristic points in dark background and inverse, key points Hessian, descriptor type real vector, unchanged in various scaling and rotations
ORB	Performance in noise scenes, descriptor type binary string
FAST	Checks corner points in image block, faster than SIFT/SURF, does not detects orientation of feature points
RANCSANC	Copes with outliers in the input image, uses the minimum number of data points
ICP	Based on combining images for the dynamic environment and an image with the 3D position information of the feature
SDF	Describes geometrical shapes, gives a distance of point X from the boundary of a surface, and determines if a point lies inside or outside the boundary

A standard procedure for assessing the structure of a three-dimensional object with depth cameras begins with the combination of its sensors. RGB-D sensors simultaneously capture color (RGB) and depth (D) data, and then the features of the considered scene are detected and extracted. In the next step, homologous (key) points are sought between the camera frames and are matched. Initially, a sparse point cloud is created, and then its local coordinates are transformed into the global coordinate system using the co-linearity equations. The sparse point cloud has a low resolution. From the sparse cloud emerges the dense cloud, which has metric value, and its density depends on the number of frames. In the next step, a triangle grid is created among points of the dense cloud, and the resolution of the depth map is determined by the number of triangular surfaces. The triangular grid creation process is called triangulated irregular network (TIN) spatial interpolation; a TIN is represented as a continuous surface consisting of triangular faces. A texture is given at each triangular surface. Finally, the position of the camera is evaluated, and the

three-dimensional reconstructed model is extracted. In 3D reconstruction scenes in real time, the same procedure is followed, except that the calibration procedure is required to calculate the position and orientation of the camera for the desired reference system. In addition, errors are identified and corrected, in order to operate the camera accurately (see Figure 8).

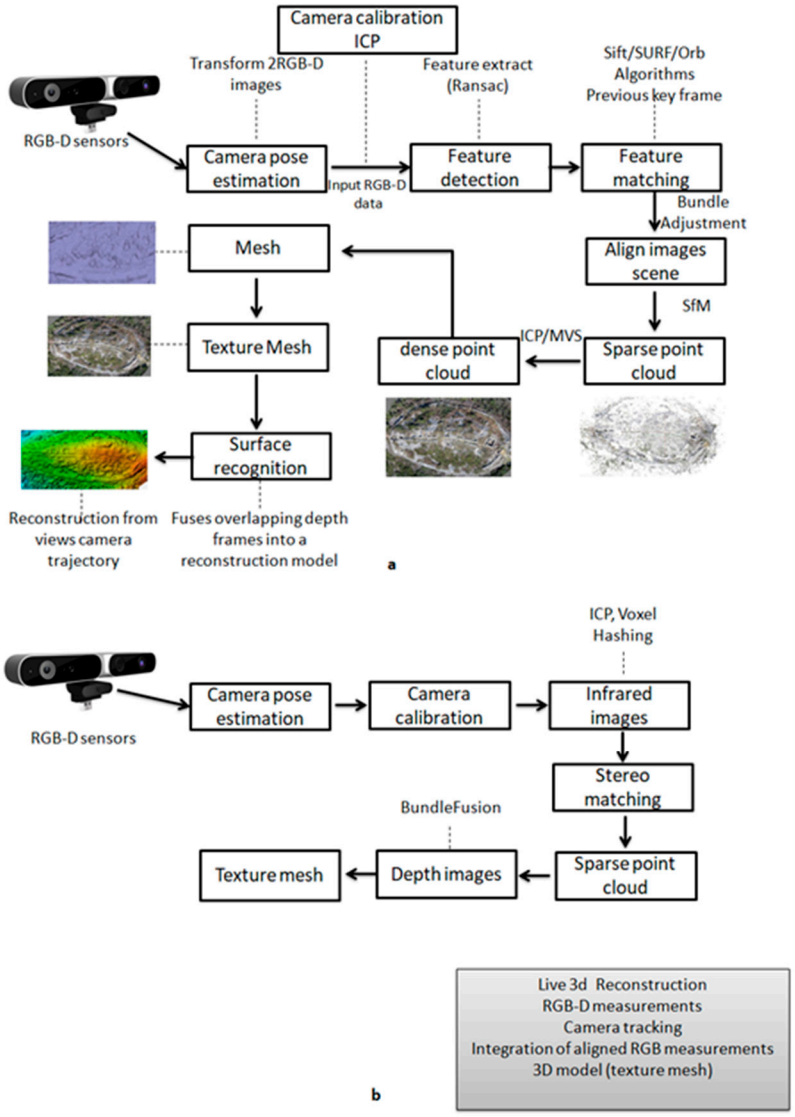

Figure 8. Pipeline of 3D scene reconstruction with an RGB-D camera, (**a**) a general 3D reconstruction workflow and (**b**) a typical live 3D reconstruction.

Figure 8a shows a general 3D reconstruction object workflow from RGB-D cameras, while Figure 8b shows a typical live 3D reconstruction. Sometimes the algorithms may be different, but the core technique is the same.

4.1. Approaches to 3D Reconstruction (RGB Mapping)

Researchers have approached the subject of 3D reconstruction with various techniques and methods. Table 2 presents some of these.

Table 2. Approaches and characteristics of 3D reconstruction.

Approaches of 3D Reconstruction	
Techniques and Methods	**Characteristics**
Align the current frame to the previous frame with the ICP algorithm [47]	For large-scale scenes, creates error propagation [33]
Weighted average of multi-image blending [78]	Motion blur and sensitive to light change
Sub-mapping-based BA [79]	High reconstruction accuracies and low computational complexity
Design a global 3D model, which is updated and combined with live depth measurements for the volumetric representation of the scene reconstructed [87]	High memory consumption
Visual and geometry features, combines SFM without camera motion and depth [88]	Accuracy is satisfactory, cannot be applied to real-time applications
Design system that provides feedback, is tolerate in human errors and alignment failures [89]	Scans large area (50 m) and preserves details about accuracy
Design system that aligns and maps large indoor environments in near real time and handles featureless corridors and dark rooms [47]	Estimates the appropriate color, implementation of RGB-D mapping is not real time

According to the literature, for the 3D reconstruction of various scenes, depth cameras are used in combination with various techniques and methods to extract more accurate, qualitative, and realistic models. However, when dealing with footage in mostly dynamic environments, there are some limitations that require solutions. Table 3 describes these limitations and suggests possible solutions.

Table 3. Limitations and proposed solutions for 3D reconstruction in dynamic scenes.

Limitation of RGB-D in Dynamic Scenes	Proposed Solutions
High-quality surface modeling	Surface modeling, no points
Global model consistency	When the scale changes, errors and distortions are corrected at the same time
Robust camera tracking	If the camera does not fail in areas with lack of features, then incremental errors will not occur. Does not consider preceding frames exclusively
On-the-fly model updates	Updates model with new poses each time
Real-time rates	Camera pose feedback in new spontaneous data
Scalability	Scanning in small- and large-scale areas, especially in the robotics sector and virtual reality applications, that are unexpectedly changing. Additionally, maintains local accuracy

4.2. Multi-View RGB-D Reconstruction Systems That Use Multiple RGB-D Cameras

Three-dimensional reconstruction of a scene from a single RGB-D camera is a risk and should be taken seriously because there are certain limitations. For example, in complex large scenes, it has low performance and requires high memory capacity or estimation of the relative poses of all cameras in the system. To address these issues, the multiple RGB-D camera was developed. Using this method, we can acquire data independently from each camera, which are then put in a single reference frame to form a holistic 3D reconstruction of the scene. Therefore, in these systems, calibration is necessary [90].

4.3. RGB-D SLAM Methods for 3D Reconstruction

3D reconstruction can be used as a platform to monitor the performance of activities on a construction site [91]. The development of navigation systems is one of the major issues in robotic engineering. A robot needs information about the environment, objects in space, and its position; therefore, various methods of navigation have been developed based on odometry [92], inertial navigation, magnetometer, active labels (GPS) [93], and label and map matching. The simultaneous localization and mapping (SLAM) approach is one of the most promising methods of navigation. Recent progress in visual simultaneous localization

and mapping (SLAM) makes it possible to reconstruct a 3D map of a construction site in real-time. Simultaneous localization and mapping (SLAM) is an advanced technique in the robotics community and was originally designed for a mobile robot to consistently build a map of an unknown environment and simultaneously estimate its location in this map [94]. When a camera is used as the only exteroceptive sensor, this technique is called visual SLAM or VSLAM [95]. Modern SLAM solutions provide mapping and localization in an unknown environment [96]. Some of them can be used for updating a map that has been made before. SLAM is the general methodology for solving two problems [97,98]: (1) environment mapping and 3D model construction, and (2) localization using a generated map and trajectory processing [99].

5. Data Acquisition and Processing

5.1. RGB-D Sensors and Evolution

The data acquisition from depth cameras plays an important role in further processing of data in order to produce a qualitative and accurate 3D reconstructed model of the physical world. Therefore, the contribution of depth cameras' incorporated sensors is of major importance. Nowadays, the sensors have many capabilities and continue to evolve. The rapid evolution is due to the parallel development of technologies, and this is to be expected considering that depth cameras work with other devices or software. In short, there are two main types of sensors, active and passive, which complement each other in various implementations [100]. Although sensors provide many benefits, they also present errors [101] and inaccurate measurements [102]. In general, to achieve a high degree of detail, depth cameras should be calibrated. RGB-D cameras were developed in the last decade, but the foundations were laid in 1989. Figure 9 illustrates the evolutionary history of RGB-D sensors.

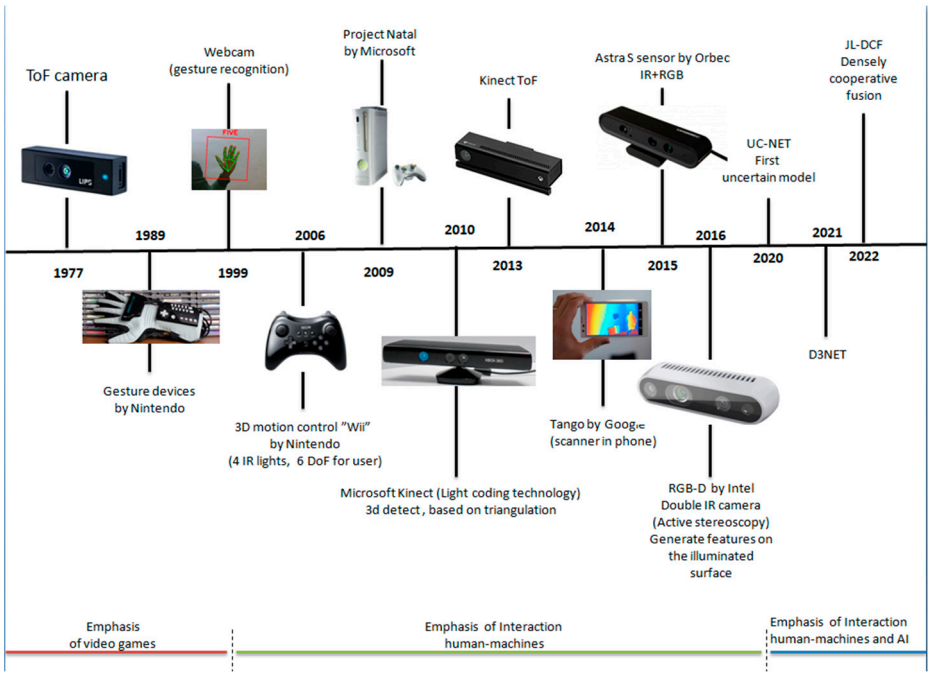

Figure 9. Evolution timeline of RGB-D sensors.

5.2. Sensing Techniques of RGB-D Cameras

There are different techniques to acquire data from depth cameras. These techniques fall into two categories, active and passive sensing, as well as the recently developed monocular depth estimation. The techniques of the first category uses a structured energy emission to capture an object in a static environment [103], as well as capture the whole scene at the same time. With active techniques, the 3D reconstruction becomes simpler.

In this case, there are also two subcategories, time of flight (ToF) and structured light (SL) cameras [104]. The second category is based on the triangulation principle [105,106], and through epipolar geometry, correspondence of the key points. In the third category, the depth estimation for the 3D reconstruction of an object is done by two-dimensional images [107]. Figure 10 shows the categories of techniques for data acquisition from RGB-D cameras.

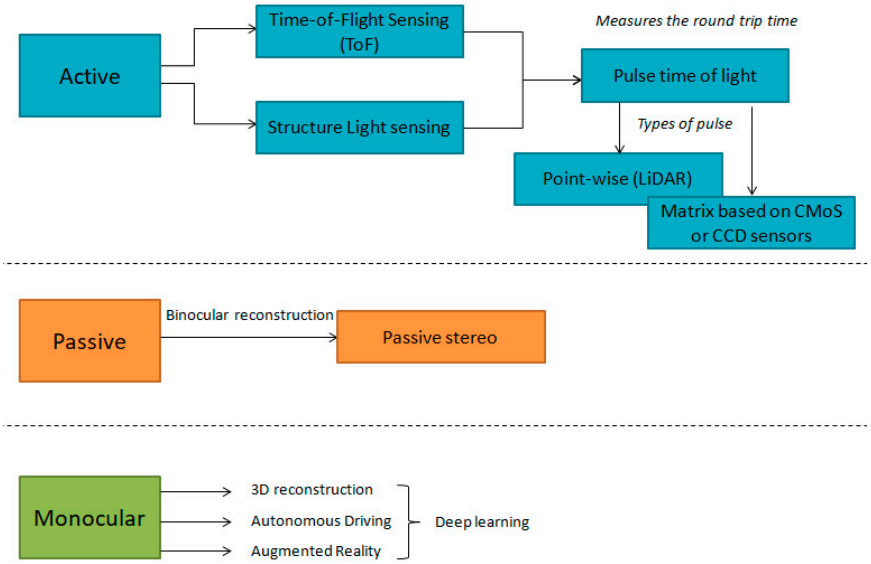

Figure 10. Techniques for RGB-D to acquire data and information.

5.3. Depth Image Processing (Depth Map)

The depth of the scene, combined with the color information, will compose the RGB-D data, and the result will be a depth map. A depth map is a metric value image that provides information relating to the distance of the surfaces of the scene objects. In fact, it is through depth estimation that the geometric relationships of objects within a scene are understood [108]. This process is achieved by epipolar geometry (i.e., the geometry of stereoscopic vision), which expresses a scene viewed by two cameras placed at different angles, or simply by the same camera shifted to different viewing angles (See Figure 11).

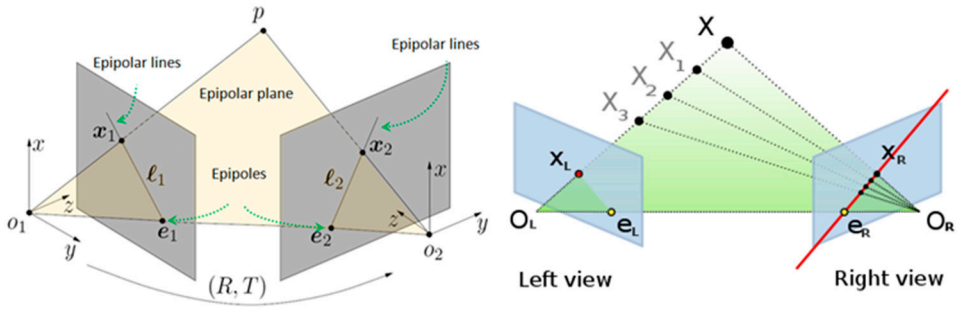

Figure 11. Epipolar geometry (source: https://en.wikipedia.org/wiki/Epipolar_geometry (accessed on 12 July 2022)).

According to Figure 11, a point P in world coordinates (X,Y,Z) is projected on the camera sensor at point x = K [R,T]X, with x = (u,v,1) and X = (X,Y,Z,1), where K is the camera calibration matrix, and R,T the rotation and translation matrices of 3 × 3 and 3 × 1 size, respectively. The only information that can be obtained is the calculation of the half-line on which this point is located, a half-line starting from the center of the camera projection and extending away from the camera. Therefore, if there is a second camera in a different part of space, covering the same scene, it is possible, through trigonometry, to calculate the

exact 3D coordinates of the point, as long as the mapping of the points of one camera to the points of the other can be achieved [109]. Solving the problem is simple, as it requires solving a three-equation system with three unknown values. The pixel position in the frame of each camera, as well as the affinity transformation between the coordinate systems of the two cameras, is available as data. The mapping of pixels to a frame plane is done through the algorithms discussed above.

Depth maps are produced using the methods described in Section 5.2, and they are directly related to environment lighting, object reflection, and spatial analysis. For example, bright lighting is responsible for creating outliers [110]. In addition, depth maps suffer from view angle reflective surfaces, occlusion boundaries [111], levels of quantization, and random noise (mainly indoor scene distances) [112], which are related to the distance of the object and the pixel position. To a certain extent, some of the above disadvantages are addressed. For example, the fusion of frames from different viewpoints, shape from shading (SfS), shape from polarization (SfP) techniques, or bilateral filtering help to repair the noise and smooth the depth map [113]. Qualitative depth maps have been an important concern for researchers, and they have devised techniques to solve the problems created during the 3D reconstruction process. Table 4 illustrates some problems with depth images and technical countermeasures tested.

Table 4. Problems of the depth maps and countermeasures.

Cons of Depth Maps	Countermeasures
Low accuracy	Apply bilateral filter [106]
Noise	Convolutional deep autoencoder denoising [107]
(HR) RGB but (LR) depth images	Super-resolution techniques, high-resolution color images [83]. CNN to downsample an HR image sampling and LR depth image [87]
Featureless region	Polarization-based methods (reveal surface normal information) [102]
Shiny surfaces, bright, transparency	TSDF to voxelize the space [105], ray-voxel pairs [106]

Depth maps are of great importance for extracting 3D reconstruction models; however, there are still limitations that pose challenges to the scientific community. Moreover, there are still some issues that remain open and need to be explored in the future. The main limitations are as follows:

- Recording the first surface seen cannot obtain information for refracted surfaces;
- Noise from the reflective surface viewing angle. Occlusion boundaries blur the edges of objects;
- Single-channel depth maps cannot convey multiple distances when multiple objects are in the location of the same pixel (grass, hair);
- May represent the perpendicular distance between an object and the plane of the scene camera and the actual distances from the camera to the plane surface seen in the corners of the image as being greater than the distances to the central area;
- In the case of missing depth data, many holes are created. To address this issue, a median filter is used, but sharp depth edges are corrupted;
- Cluttered spatial configuration of objects can create occlusions and shadows.

From the above limitations emerge some challenges, such as occlusions, camera calibration errors, low resolution, and high levels of ambient light (ToF), which are unsuitable for outdoor operation (structured light). In addition, depth noise increases with distance (SL) quadratically. Moreover, issues such as the correspondence between stereo or multiview images, multiple depth cues, computational complexity, spatial resolution, angle of projection, and multiple camera interference for dynamic scenarios remain open to investigation.

5.4. RGB-D Datasets

RGB-D data is essential for solving certain problems in computer vision. Nowadays, there are open databases containing large datasets of both indoor and outdoor scenes collected by RGB-D cameras and different sensors. The data are related to scenes and objects, human activities, gestures, and the medical field, and are used for applications such as simultaneous localization and mapping (SLAM) [114], representation [115], object segmentation [116], and human activity recognition [117]. Table 5 lists some of the most well-known datasets and the applications they used, such as semantic segmentation (SS), object detection (OD), pose (P), normal maps (NM), 3D semantic-voxel segmentation (3D SvS), instance segmentation (IS), and diffuse reflectance (DR).

Table 5. RGB-D datasets for SLAM, odometry, reconstruction, segmentation.

Dataset	Year	Sensor Type	Apps	Images/Scenes
NYU Depth	(V1) 2011	Structured light	SS	64 scenes (108,617 frames) with 2347 labeled RGB-D frames
	(V2) 2012	Structured light	SS	464 scenes (407,024 frames) with 1449 labeled aligned RGB-D images
SUN RGB-D	2015	Structured light and TOF	SS, OD, P	10335 images
Stanford2D3D	2016	Structured light	SS, NM	6 large-scale indoor areas (70,496 images)
ScanNet	2017	Structured light	3D SvS	1513 sequences (over 2.5 million frames)
Hypersim	2021	Synthetic	NM, IS, DR	461 scenes (77,400 images)

The NYU Depth dataset is the most popular for RGB-D indoor segmentation. It was created using a Microsoft Kinect v1 sensor, is composed of aligned RGB and depth images, and consists of labeled data containing semantic segmentation as well as raw data [118]. There are two versions: NYUv1 and NYUv2 (464 scenes (407,024 frames) with 1449 labeled aligned RGB-D images with 640 $\times$ 480 resolution). The dataset is split into a training set of 795 images and a testing set of 654 images. The difference is that the first type has fewer scenes and total frames (64 Scenes (108,617 Frames) with 2347 labeled RGB-D frames) [119]. NYUv2 originally had 13 different categories. However, recent models mostly evaluate their performance at the more challenging 40-classes settings [120].

The SUN RGB-D dataset [110] is the same category as NYU. Data was acquired with structured light and ToF sensors and used for semantic segmentation, object detection, and pose. This dataset provides 10,335 RGB-D images with the corresponding semantic labels. It contains images captured by different depth cameras (Intel RealSense, Asus Xtion, Kinect v1/2) since they are collected from previous datasets. Therefore, the image resolutions vary depending on the sensor used. SUN-RGBD has 37 classes of objects. The training set consists of 5285 images, and the testing set consists of 5050 images [121].

The Stanford2D3D dataset consists of indoor scene images, taken with a structured light sensor, which are used for semantic segmentation. It is a large-scale dataset that consists of 70,496 RGB images with the associated depth maps. The images are in 1080 $\times$ 1080 resolution and are collected in a 360° scan fashion. The usual class setting employed is 13 classes [122].

The ScanNet dataset is an indoor dataset collected by a structured light and contains over 2.5 million frames from 1513 different scenes. It is used for 3D semantic-voxel segmentation [115].

The Hypersim dataset consists of indoor scenes that are captured synthetically and used for normal maps, instance segmentation, and diffuse reflectance [123].

6. Advantages and Limitations of RGB-D

RGB-D camera technology, as mentioned above, is increasingly being used in a variety of applications. However, as with any technology, apart from the advantages it provides, it also has some limitations, especially in terms of data collection. For instance, the real-time performance, dense model, no drifting with local optimization, and robustness to scene changes are camera innovations that do not work for large areas (voxel-grid), far away from objects (active ranging), or outdoors (IR). Moreover, it requires a powerful graphics cards and uses lots of battery (active ranging) resources. Table 6 shows the advantages and limitations of RGB-D cameras based on their sensors. In particular, the advantages and limitations of active and passive sensors in general are listed, and then specified in the subcategories of active sensors. In addition, there is a focus on sensor errors and the inaccuracies that arise from the measurement procedure.

Table 6. Advantages and limitations of RGB-D cameras and sensors.

Advantages Active & Passive techniques	• Cheap construction • Energy efficient • Low power consumption • High frame rate and autonomy • High resolution (passive) • Creates color point cloud • Accuracy classification among similar objects • Color information improves scene details • Reliable scene segmentation • Use of deep learning • Outdoor and indoor implementations • Easy to find features • Calibrated IMU with accuracy	
Limitations of RGB-D cameras	***Active*** • Needs sufficient local intensity and textured scenes • Lack of features in the scene • Low resolution and spatial resolution • Robust non-textured regions • Low-precision distance measurement, thermal stability, and repeatability	***Passive*** • Limited depth accuracy • Non-robust textured regions • No correspondence in different views of camera • Limited with smooth surfaces
Limitations of Active sensors	***ToF*** • Reflections from materials • No reflection light in dark materials • Depth accuracy mm to cm • Motion blurred • Noisy characteristics • Limited scanning speed	***Structure Light Sensing*** • Sun's infrared radiation saturates the sensor • Missing depth information • Does not work under strong light • Limited range (15 m) • Affected by metallic surfaces • High energy consumption

Table 6. *Cont.*

Systematic & Random Errors of Sensors	• Ambient background light • Multi-device interference • Temperature drift • Systematic distance error • Depth in homogeneity at object boundaries (flying pixel) • Multi-path effects • Intensity-related distance error (ToF) • Semi-transparent and scattering media • Dynamic scenery	
Measurements inaccuracies	***Pose estimation of RGB-D*** • Inaccuracies of pose estimation device • Relative transformation (camera-pose estimation device) and alignment of world and model coordinate system • Temporal offset (pose and depth image acquisition)	***RGB-D camera itself*** • Random measurement errors • Systematic measurement errors • Motion blur effect

7. Conclusions

In this report, through a literature review, the main aspects of the 3D reconstruction of scenes and objects, in both static and dynamic environments, using RGB-D cameras, are gathered, compared, discussed, and critically analyzed. In addition, approaches, methodologies, and techniques applied to date are summarized. Depth cameras are powerful tools for researchers, as this technology provides real-time stereoscopic models. On the other hand, this technology presents serious limitations, which need to be solved. For example, their infrared operation prevents the reconstruction of the model outdoors, and this is an issue that is still being studied. To eliminate the resulting problems, in both the process of 3D reconstruction of the objects and the final products, it may be necessary to devise stronger techniques and algorithms that cover all the weak points, resulting in more reliable objects in terms of their geometry. In addition, a very important issue that needs to be resolved concerns the amount of memory this process requires. Hence, technology should take into account both the integrity and reliability of 3D models and the performance of the systems concerned.

Some future research goals of this scientific field are to perform experiments with new CNN network architectures, to create methods with smaller memory requirements, to use machine learning, and to combine UAVs with depth cameras so that it is possible to capture scenes throughout the day and night.

Author Contributions: Conceptualization, K.A.T.; methodology, K.A.T., I.T. and G.A.P.; formal analysis, K.A.T. and I.T.; investigation, K.A.T.; resources, K.A.T.; data curation, K.A.T.; writing—original draft preparation, K.A.T.; writing—review and editing, I.T. and G.A.P.; visualization, K.A.T. and I.T.; supervision, G.A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by the MPhil program "Advanced Technologies in Informatics and Computers", hosted by the Department of Computer Science, International Hellenic University, Kavala, Greece.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Orts-Escolano, S.; Rhemann, C.; Fanello, S.; Chang, W.; Kowdle, A.; Degtyarev, Y.; Kim, D.; Davidson, P.L.; Khamis, S.; Dou, M.; et al. Holoportation: Virtual 3D teleportation in real-time. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, Tokyo, Japan, 16–19 October 2016; pp. 741–754. [CrossRef]
2. Lai, K.; Bo, L.; Ren, X.; Fox, D. A large-scale hierarchical multi-view RGB-D object dataset. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 1817–1824. [CrossRef]
3. Seitz, S.M.; Curless, B.; Diebel, J.; Scharstein, D.; Szeliski, R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 1 (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 1, pp. 519–528. [CrossRef]
4. Finlayson, G.; Fredembach, C.; Drew, M.S. Detecting illumination in images. In Proceedings of the IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8. [CrossRef]
5. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2004; pp. 239–241. [CrossRef]
6. Martinez, M.; Yang, K.; Constantinescu, A.; Stiefelhagen, R. Helping the Blind to Get through COVID-19: Social Distancing Assistant Using Real-Time Semantic Segmentation on RGB-D Video. *Sensors* **2020**, *20*, 5202. [CrossRef]
7. Vlaminck, M.; Quang, L.H.; van Nam, H.; Vu, H.; Veelaert, P.; Philips, W. Indoor assistance for visually impaired people using a RGB-D camera. In Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), Santa Fe, NM, USA, 6–8 March 2016; pp. 161–164. [CrossRef]
8. Palazzolo, E.; Behley, J.; Lottes, P.; Giguere, P.; Stachniss, C. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 7855–7862. [CrossRef]
9. Zollhöfer, M.; Stotko, P.; Görlitz, A.; Theobalt, C.; Nießner, M.; Klein, R.; Kolb, A. State of the Art on 3D reconstruction with RGB-D Cameras. *Comput. Graph. Forum* **2018**, *37*, 625–652. [CrossRef]
10. Verykokou, S.; Ioannidis, C.; Athanasiou, G.; Doulamis, N.; Amditis, A. 3D Reconstruction of Disaster Scenes for Urban Search and Rescue. *Multimed Tools Appl.* **2018**, *77*, 9691–9717. [CrossRef]
11. He, Y.B.; Bai, L.; Aji, T.; Jiang, Y.; Zhao, J.M.; Zhang, J.H.; Shao, Y.M.; Liu, W.Y.; Wen, H. Application of 3D Reconstruction for Surgical Treatment of Hepatic Alveolar Ehinococcosis. *World J. Gastroenterol. WJG* **2015**, *21*, 10200–10207. [CrossRef] [PubMed]
12. Gomes, L.; Regina Pereira Bellon, O.; Silva, L. 3D Reconstruction Methods for Digital Preservation of Cultural Heritage: A Survey. *Pattern Recognit. Lett.* **2014**, *50*, 3–14. [CrossRef]
13. Newcombe, R.A.; Fox, D.; Seitz, S.M. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 12 June 2015; pp. 343–352. [CrossRef]
14. Seichter, D.; Köhler, M.; Lewandowski, B.; Wengefeld, T.; Gross, H.-M. Efficient RGB-D semantic segmentation for indoor scene analysis. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13525–13531. [CrossRef]
15. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In *Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 213–228.
16. Jiang, J.; Zheng, L.; Luo, F.; Zhang, Z. RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation. *arXiv* **2018**, arXiv:1806.01054.
17. Zhong, Y.; Dai, Y.; Li, H. 3D geometry-aware semantic labeling of outdoor street scenes. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20 August 2018; pp. 2343–2349.
18. Xing, Y.; Wang, J.; Chen, X.; Zeng, G. 2.5D Convolution for RGB-D semantic segmentation. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22 September 2019; pp. 1410–1414.
19. Xing, Y.; Wang, J.; Zeng, G. Malleable 2.5D convolution: Learning receptive fields along the depth-axis for RGB-D scene parsing. In Proceedings of the European Conference on Computer Vision 2(ECCV), Glasgow, UK, 23 August 2020; pp. 555–571.
20. Wang, W.; Neumann, U. Depth-aware CNN for RGB-D segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 144–161.
21. Chen, L.Z.; Lin, Z.; Wang, Z.; Yang, Y.L.; Cheng, M.M. Spatial Information Guided Convolution for Real-Time RGBD Semantic Segmentation. *arXiv* **2020**, arXiv:2004.04534. [CrossRef]
22. Chen, Y.; Mensink, T.; Gavves, E. 3D Neighborhood convolution: Learning DepthAware features for RGB-D and RGB semantic segmentation. In Proceedings of the International Conference on 3D Vision (3DV), Quebeck City, QC, Canada, 16–19 September 2019; pp. 173–182.
23. Cao, J.; Leng, H.; Lischinski, D.; Cohen-Or, D.; Tu, C.; Li, Y. ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021. [CrossRef]
24. Chen, X.; Lin, K.Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 561–577.

25. Hu, X.; Yang, K.; Fei, L.; Wang, K. ACNet: Attention based network to exploit complementary features for RGBD semantic segmentation. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019.
26. Hu, J.; Zhao, G.; You, S.; Kuo, C.C.J. Evaluation of multimodal semantic segmentation using RGB-D data. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*; SPIE: Budapest, Hungary, 2021. [CrossRef]
27. Hu, Y.; Chen, Z.; Lin, W. RGB-D Semantic segmentation: A review. In Proceedings of the 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), San Diego, CA, USA, 23–27 July 2018; pp. 1–6. [CrossRef]
28. Liu, H.; Zhang, J.; Yang, K.; Hu, X.; Stiefelhagen, R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers. *arXiv* **2022**, arXiv:2203.04838.
29. Xing, Y.; Wang, J.; Chen, X.; Zeng, G. Coupling two-stream RGB-D semantic segmentation network by idempotent mappings. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 16–19 October 2019; pp. 1850–1854.
30. Park, S.J.; Hong, K.S.; Lee, S. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In Proceedings of the IEEE International Conference On Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4990–4999.
31. Valada, A.; Mohan, R.; Burgard, W. Self-supervised model adaptation for multimodal semantic segmentation. *Int. J. Comput. Vis. (IJCV)* **2019**, *128*, 1239–1285. [CrossRef]
32. Fooladgar, F.; Kasaei, S. Multi-Modal Attention-based Fusion Model for Semantic Segmentation of RGB-Depth Images. *arXiv* **2019**, arXiv:1912.11691.
33. Penelle, B.; Schenkel, A.; Warzée, N. Geometrical 3D reconstruction using real-time RGB-D cameras. In Proceedings of the 2011 International Conference on 3D Imaging (IC3D), Liege, Belgium, 7–8 December 2011; pp. 1–8. [CrossRef]
34. Bakator, M.; Radosav, D.J.M.T. Deep learning and medical diagnosis: A review of literature. *Multimodal Technol. Interact.* **2018**, *2*, 47. [CrossRef]
35. Lundervold, A.S.; Lundervold, A. An Overview of Deep Learning in Medical Imaging Focusing on MRI. *Z. Für Med. Phys.* **2019**, *29*, 102–127. [CrossRef]
36. Yaqub, M.; Jinchao, F.; Arshid, K.; Ahmed, S.; Zhang, W.; Nawaz, M.Z.; Mahmood, T. Deep Learning-Based Image Reconstruction for Different Medical Imaging Modalities. *Comput. Math. Methods Med.* **2022**, 8750648. [CrossRef]
37. Pain, C.D.; Egan, G.F.; Chen, Z. Deep Learning-Based Image Reconstruction and Post-Processing Methods in Positron Emission Tomography for Low-Dose Imaging and Resolution Enhancement. *Eur. J. Nucl. Med. Mol. Imaging* **2022**, *49*, 3098–3118. [CrossRef] [PubMed]
38. Lopes, A.; Souza, R.; Pedrini, H. A Survey on RGB-D Datasets. *Comput. Vis. Image Underst.* **2022**, *103489*, 222. [CrossRef]
39. Elmenreich, W. An Introduction to Sensor Fusion. Research Report 47/2001. Available online: https://www.researchgate.net/profile/Wilfried_Elmenreich/publication/267771481_An_Introduction_to_Sensor_Fusion/links/55d2e45908ae0a3417222dd9/AnIntroduction-to-Sensor-Fusion.pdf (accessed on 1 August 2022).
40. Nguyen, C.V.; Izadi, S.; Lovell, D. Modeling kinect sensor noise for improved 3D reconstruction and tracking. In Proceedings of the 2nd International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission, Zurich, Switzerland, 13–15 October 2012; pp. 524–530. [CrossRef]
41. Morell-Gimenez, V.; Saval-Calvo, M.; Azorin-Lopez, J.; Garcia-Rodriguez, J.; Cazorla, M.; Orts-Escolano, S.; Fuster-Guillo, A. A comparative study of Registration Methods for RGB-D Video of Static Scenes. *Sensors* **2014**, *14*, 8547–8576. [CrossRef] [PubMed]
42. Tombari, F.; Salti, S.; Di Stefano, L. Unique signatures of histograms for local surface description. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 356–369. [CrossRef]
43. Newcombe, R.A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A.J.; Kohi, P.; Shotton, J.; Hodges, S.; Fitzgibbon, A. KinectFusion: Real-time dense surface mapping and tracking. In Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, Switzerland, 26–29 October 2011; pp. 127–136. [CrossRef]
44. Chen, Y.; Medioni, G. Object modeling by registration of multiple range images. *Image Vis. Comput.* **1992**, *10*, 145–155. [CrossRef]
45. Wang, R.; Wei, L.; Vouga, E.; Huang, Q.; Ceylan, D.; Medioni, G.; Li, H. Capturing dynamic textured surfaces of moving targets. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–14 September 2016; Springer: Cham, Switzerland, 2016; pp. 271–288. [CrossRef]
46. Li, J.; Gao, W.; Wu, Y.; Liu, Y.; Shen, Y. High-quality indoor scene 3D reconstruction with RGB-D cameras: A brief review. *Comput. Vis. Media* **2022**, *8*, 369–393. [CrossRef]
47. Henry, P.; Krainin, M.; Herbst, E.; Ren, X.; Fox, D. RGB-D Mapping: Using Kinect-Style Depth Cameras for Dense 3D Modeling of Indoor Environments. *Int. J. Robot. Res.* **2012**, *31*, 647–663. [CrossRef]
48. Zaharescu, A.; Boyer, E.; Varanasi, K.; Horaud, R. Surface feature detection and description with applications to mesh matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 373–380. [CrossRef]
49. Knopp, J.; Prasad, M.; Willems, G.; Timofte, R.; Gool, L.V. Hough transform and 3D SURF for robust three dimensional classification. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; Springer: Berlin, Heidelberg, 2010; Volume 6316, pp. 589–602. [CrossRef]
50. Salti, S.; Petrelli, A.; Tombari, F.; di Stefano, L. On the affinity between 3D detectors and descriptors. In Proceedings of the 2nd International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission, Zurich, Switzerland, 13–15 October 2012; pp. 424–431. [CrossRef]

51. Steinbrücker, F.; Sturm, J.; Cremers, D. Real-time visual odometry from dense RGB-D images. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 719–722. [CrossRef]
52. Gao, F.; Sun, Q.; Li, S.; Li, W.; Li, Y.; Yu, J.; Shuang, F. Efficient 6D object pose estimation based on attentive multi-scale contextual information. *IET Comput. Vis.* **2022**, *2022*, cvi2.12102. [CrossRef]
53. Rodriguez, J.S. A comparison of an RGB-D cameras performance and a stereo camera in relation to object recognition and spatial position determination. ELCVIA. *Electron. Lett. Comput. Vis. Image Anal.* **2021**, *20*, 16–27. [CrossRef]
54. Huang, A.S.; Bachrach, A.; Henry, P.; Krainin, M.; Maturana, D.; Fox, D.; Roy, N. Visual odometry and mapping for autonomous flight using an RGB-D camera. In *Robotics Research*; Springer: Cham, Switzerland, 2016; pp. 235–252; ISBN 978-3-319-29362-2. [CrossRef]
55. Jaimez, M.; Kerl, C.; Gonzalez-Jimenez, J.; Cremers, D. Fast odometry and scene flow from RGB-D cameras based on geometric clustering. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 3992–3999. [CrossRef]
56. Yigong, Z.; Zhixing, H.; Jian, Y.; Hui, K. Maximum clique based RGB-D visual odometry. In Proceedings of the 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2764–2769. [CrossRef]
57. Yang, J.; Gan, Z.; Gui, X.; Li, K.; Hou, C. 3-D geometry enhanced superpixels for RGB-D data. In Proceedings of the Pacific-Rim Conference on Multimedia Springer, Nanjing, China, 13–16 December 2013; Springer: Cham, Switzerland; Volume 8294, pp. 35–46. [CrossRef]
58. Hu, G.; Huang, S.; Zhao, L.; Alempijevic, A.; Dissanayake, G. A robust RGB-D slam algorithm. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Algarve, Portugal, 7–12 October 2012; pp. 1714–1719. [CrossRef]
59. Maisto, M.; Panella, M.; Liparulo, L.; Proietti, A. An accurate algorithm for the identification of fingertips using an RGB-D camera. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2013**, *3*, 272–283. [CrossRef]
60. Berger, M.; Tagliasacchi, A.; Seversky, L.M.; Alliez, P.; Levine, J.A.; Sharf, A.; Silva, C.T. State of the art in surface reconstruction from point clouds. *Eurographics State Art Rep.* **2014**, *1*, 161–185. [CrossRef]
61. Weinmann, M.; Klein, R. Exploring material recognition for estimating reflectance and illumination from a single image. In Proceedings of the Eurographics Workshop on Material Appearance Modeling, Dublin, Ireland, 22 June 2016; pp. 27–34. [CrossRef]
62. Curless, B.; Levoy, M. A volumetric method for building complex models from range images. In Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 4–9 August 1996; pp. 303–312. [CrossRef]
63. Chen, J.; Bautembach, D.; Izadi, S. Scalable real-time volumetric surface reconstruction. *ACM Trans. Graph. (ToG)* **2013**, *32*, 113. [CrossRef]
64. Zollhöfer, M.; Nießner, M.; Izadi, S.; Rehmann, C.; Zach, C.; Fisher, M.; Wu, C.; Fitzgibbon, A.; Loop, C.; Theobalt, C.; et al. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Trans. Graph. (ToG)* **2014**, *33*, 156. [CrossRef]
65. Jia, Q.; Chang, L.; Qiang, B.; Zhang, S.; Xie, W.; Yang, X.; Sun, Y.; Yang, M. Real-time 3D reconstruction method based on monocular vision. *Sensors* **2021**, *21*, 5909. [CrossRef] [PubMed]
66. Cui, Y.; Schuon, S.; Chan, D.; Thrun, S.; Theobalt, C. 3D shape scanning with a time-of-flight camera. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1173–1180. [CrossRef]
67. Kim, P.; Lim, H.; Kim, H.J. Robust visual odometry to irregular illumination changes with RGB-D camera. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 3688–3694. [CrossRef]
68. Camplani, M.; Hannuna, S.; Mirmehdi, M.; Damen, D.; Paiement, A.; Tao, L.; Burghardt, T. Real-time RGB-D tracking with depth scaling kernelised correlation filters and occlusion handling. In Proceedings of the BMVC, Swansea, UK, 7–10 September 2015; pp. 145.1–145.11. [CrossRef]
69. Liu, Y.; Jing, X.-Y.; Nie, J.; Gao, H.; Liu, J.; Jiang, G.-P. Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in RGB-D videos. *IEEE Trans. Multimed.* **2019**, *21*, 664–677. [CrossRef]
70. Sarkar, S.; Venugopalan, V.; Reddy, K.; Ryde, J.; Jaitly, N.; Giering, M. Deep learning for automated occlusion edge detection in RGB-D frames. *J. Signal Process. Syst.* **2017**, *88*, 205–217. [CrossRef]
71. Scharstein, D.; Szeliski, R. High-accuracy stereo depth maps using structured light. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 16–22 June 2003; pp. 195–202. [CrossRef]
72. Tang, S.; Zhu, Q.; Chen, W.; Darwish, W.; Wu, B.; Hu, H.; Chen, M. Enhanced RGB-D Mapping Method for Detailed 3D Indoor and Outdoor Modeling. *Sensors* **2016**, *16*, 1589. [CrossRef]
73. Kočevar, T.N.; Tomc, H.G. Modelling and visualisation of the optical properties of cloth. In *Computer Simulation*; Cvetkovic, D., Ed.; InTech: London, UK, 2017. [CrossRef]
74. Langmann, B.; Hartmann, K.; Loffeld, O. Depth camera technology comparison and performance evaluation. In Proceedings of the 1st International Conference on Pattern Recognition Applications and Method, Vilamura, Algarve, Portugal, 6–8 February 2012; pp. 438–444. [CrossRef]
75. Tran, V.-L.; Lin, H.-Y. Accurate RGB-D camera based on structured light techniques. In Proceedings of the International Conference on System Science and Engineering (ICSSE), Ho Chi Minh City, Vietnam, 21–23 July 2017; pp. 235–238. [CrossRef]

76. Dai, A.; Nießner, M.; Zollhöfer, M.; Izadi, S.; Theobalt, C. BundleFusion: Real-Time Globally Consistent 3D Reconstruction Using On-the-Fly Surface Reintegration. *ACM Trans. Graph.* **2017**, *36*, 76a. [CrossRef]
77. Nießner, M.; Zollhöfer, M.; Izadi, S.; Stamminger, M. Real-time 3D reconstruction at scale using voxel hashing. *ACM Trans. Graph. (ToG)* **2013**, *32*, 169. [CrossRef]
78. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, pp. 511–518. [CrossRef]
79. Maier, R.; Sturm, J.; Cremers, D. Submap-based bundle adjustment for 3D reconstruction from RGB-D data. In Proceedings of the Conference on Pattern Recognition, Münster, Germany, 2–5 September 2014; Springer: Cham, Switzerland; pp. 54–65. [CrossRef]
80. Rosten, E.; Drummond, T. Fusing points and lines for high performance tracking. In Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–21 October 2005; Volume 2, pp. 1508–1515. [CrossRef]
81. Wang, X.; Zou, J.; Shi, D. An improved ORB image feature matching algorithm based on SURF. In Proceedings of the 3rd International Conference on Robotics and Automation Engineering (ICRAE), Guangzhou, China, 17–19 November 2018; pp. 218–222. [CrossRef]
82. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
83. Derpanis, K.G. Overview of the RANSAC Algorithm. *Image Rochester NY* **2010**, *4*, 2–3.
84. Kim, D.-H.; Kim, J.-H. Image-based ICP algorithm for visual odometry using a RGB-D sensor in a dynamic environment. In *Robot Intelligence Technology and Applications 2012*; Advances in Intelligent Systems and Computing; Kim, J.-H., Matson, E.T., Myung, H., Xu, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 208, pp. 423–430. [CrossRef]
85. Yariv, L.; Gu, J.; Kasten, Y.; Lipman, Y. Volume Rendering of Neural Implicit Surfaces. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 4805–4815. [CrossRef]
86. Minh, P.; Hoai, V.; Quoc, L. WSDF: Weighting of Signed Distance Function for Camera Motion Estimation in RGB-D Data. *Int. J. Adv. Res. Artif. Intell.* **2016**, *5*, 27–32. [CrossRef]
87. Huang, P.-H.; Matzen, K.; Kopf, J.; Ahuja, N.; Huang, J.-B. Deepmvs: Learning multi-view stereopsis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2821–2830. [CrossRef]
88. Wang, K.; Zhang, G.; Bao, H. Robust 3D reconstruction with an RGB-D camera. *IEEE Trans. Image Process.* **2014**, *23*, 4893–4906. [CrossRef]
89. Du, H.; Henry, P.; Ren, X.; Cheng, M.; Goldman, D.B.; Seitz, S.M.; Fox, D. Interactive 3D modeling of indoor environments with a consumer depth camera. In Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing, China, 17–21 September 2011; pp. 75–84. [CrossRef]
90. Xu, H.; Hou, J.; Yu, L.; Fei, S. 3D Reconstruction System for Collaborative Scanning Based on Multiple RGB-D Cameras. *Pattern Recognit. Lett.* **2019**, *128*, 505–512. [CrossRef]
91. Shang, Z.; Shen, Z. Real-time 3D reconstruction on construction site using visual SLAM and UAV. In *Construction Research Congress*; American Society of Civil Engineers: New Orleans, LA, USA, 2018; pp. 305–315. [CrossRef]
92. Shekhar, S.; Xiong, H.; Zhou, H.; Eds, X. Visual odometry. In *Encyclopedia of GIS*; Springer International Publishing: Cham, Switzerland, 2017; p. 2425. [CrossRef]
93. Bailey, T.; Durrant-Whyte, H. Simultaneous Localization and Mapping (SLAM): Part II. *IEEE Robot. Automat. Mag* **2016**, *13*, 108–117. [CrossRef]
94. Durrant-Whyte, H.; Bailey, T. Simultaneous Localization and Mapping: Part I. *IEEE Robot. Automat. Mag.* **2006**, *13*, 99–110. [CrossRef]
95. Artieda, J.; Sebastian, J.M.; Campoy, P.; Correa, J.F.; Mondragón, I.F.; Martínez, C.; Olivares, M. Visual 3-D SLAM from UAVs. *J. Intell. Robot. Syst.* **2009**, *55*, 299–321. [CrossRef]
96. Cao, F.; Zhuang, Y.; Zhang, H.; Wang, W. Robust Place Recognition and Loop Closing in Laser-Based SLAM for UGVs in Urban Environments. *IEEE Sens. J.* **2018**, *18*, 4242–4252. [CrossRef]
97. Hahnel, D.; Triebel, R.; Burgard, W.; Thrun, S. Map building with mobile robots in dynamic environments. In Proceedings of the IEEE International Conference on Robotics and Automation (Cat. No.03CH37422), Taipei, Taiwan, 14–19 September 2003; pp. 1557–1563. [CrossRef]
98. Wang, Y.; Huang, S.; Xiong, R.; Wu, J. A Framework for Multi-Session RGBD SLAM in Low Dynamic Workspace Environment. *CAAI Trans. Intell. Technol.* **2016**, *1*, 90–103. [CrossRef]

99. Alliez, P.; Bonardi, F.; Bouchafa, S.; Didier, J.-Y.; Hadj-Abdelkader, H.; Munoz, F.I.; Kachurka, V.; Rault, B.; Robin, M.; Roussel, D. Real-time multi-SLAM system for agent localization and 3D mapping in dynamic scenarios. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020; pp. 4894–4900. [CrossRef]
100. Yang, Q.; Tan, K.-H.; Culbertson, B.; Apostolopoulos, J. Fusion of active and passive sensors for fast 3d capture. In Proceedings of the IEEE International Workshop on Multimedia Signal Processing, Saint-Malo, France, 4–6 October 2010; pp. 69–74. [CrossRef]
101. Kooij, J.F.P.; Liem, M.C.; Krijnders, J.D.; Andringa, T.C.; Gavrila, D.M. Multi-modal human aggression detection. *Comput. Vis. Image Underst.* **2016**, *144*, 106–120. [CrossRef]
102. Kahn, S.; Bockholt, U.; Kuijper, A.; Fellner, D.W. Towards precise real-time 3D difference detection for industrial applications. *Comput. Ind.* **2013**, *64*, 1115–1128. [CrossRef]
103. Salvi, J.; Pages, J.; Batlle, J. Pattern Codification Strategies in Structured Light Systems. *Pattern Recognit.* **2004**, *37*, 827–849. [CrossRef]
104. Alexa, M. Differential coordinates for local mesh morphing and deformation. *Vis. Comput.* **2003**, *19*, 105–114. [CrossRef]
105. Beltran, D.; Basañez, L. A comparison between active and passive 3d vision sensors: Bumblebeexb3 and Microsoft Kinect. In Proceedings of the Robot 2013: First Iberian Robotics Conference, Madrid, Spain, 28 November 2013; pp. 725–734. [CrossRef]
106. Lee, J.-H.; Kim, C.-S. Monocular depth estimation using relative depth maps. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9721–9730. [CrossRef]
107. Ibrahim, M.M.; Liu, Q.; Khan, R.; Yang, J.; Adeli, E.; Yang, Y. Depth map artefacts reduction: A review. *IET Image* **2020**, *14*, 2630–2644. [CrossRef]
108. Kadambi, A.; Bhandari, A.; Raskar, R. 3D depth cameras in vision: Benefits and limitations of the hardware. In *Computer Vision and Machine Learning with RGB-D Sensors*; Advances in Computer Vision and Pattern Recognition; Shao, L., Han, J., Kohli, P., Zhang, Z., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 3–26. [CrossRef]
109. Patias, P. *Introduction to Photogrammetry*; Ziti Publications: Thessaloniki, Greece, 1991; ISBN 960-431-021-6.691991. (In Greek)
110. Song, S.; Lichtenberg, S.P.; Xiao, J. Sun RGB-D: A RGB-D scene understanding benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 567–576. [CrossRef]
111. Khoshelham, K.; Elberink, S.O. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors* **2012**, *12*, 1437–1454. [CrossRef]
112. Malleson, C.; Hilton, A.; Guillemaut, J.-Y. Evaluation of kinect fusion for set modelling. In Proceedings of the European Conference on Visual Media Production (CVMP 2012), London, UK, 5–6 December 2012.
113. Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. In Proceedings of the 1998 Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 7 January 1998; pp. 839–846. [CrossRef]
114. Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; Cremers, D. A Benchmark for the evaluation of RGB-D SLAM systems. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vilamoura-Algarve, Portugal, 7–12 October 2012; pp. 573–580. [CrossRef]
115. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. ScanNet: Richly-Annotated 3D reconstructions of indoor scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
116. McCormac, J.; Handa, A.; Leutenegger, S.; Davison, A.J. SceneNet RGB-D: Can 5M synthetic images beat generic ImageNet pre-training on indoor segmentation? In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2697–2706. [CrossRef]
117. Zhang, J.; Li, W.; Wang, P.; Ogunbona, P.; Liu, S.; Tang, C. A large scale RGB-D dataset for action recognition. In *Understanding Human Activities Through 3D Sensors*; Lecture Notes in Computer, Science; Wannous, H., Pala, P., Daoudi, M., Flórez-Revuelta, F., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 10188, pp. 101–114. [CrossRef]
118. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from RGBD images. In *Computer Vision—ECCV 2012*; Lecture Notes in Computer, Science; Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7576, pp. 746–760. [CrossRef]
119. Silberman, N.; Fergus, R. Indoor scene segmentation using a structured light sensor. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 601–608. [CrossRef]
120. Gupta, S.; Arbelaez, P.; Malik, J. Perceptual organization and recognition of indoor scenes from RGB-D images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 564–571.
121. Janoch, A.; Karayev, S.; Jia, Y.; Barron, J.T.; Fritz, M.; Saenko, K.; Darrell, T. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*; Springer: London, UK, 2013; pp. 141–165.
122. Armeni, I.; Sax, S.; Zamir, A.R.; Savarese, S. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv* **2017**, arXiv:1702.01105.
123. Roberts, M.; Ramapuram, J.; Ranjan, A.; Kumar, A.; Bautista, M.A.; Paczan, N.; Webb, R.; Susskind, J.M. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. *arXiv* **2020**, arXiv:2011.02523.