sensors

MDPI

Article

A Comprehensive Survey of Indoor Localization Methods Based on Computer Vision

Anca Morar [1,*], Alin Moldoveanu [1], Irina Mocanu [1], Florica Moldoveanu [1], Ion Emilian Radoi [1], Victor Asavei [1], Alexandru Gradinaru [1] and Alex Butean [2]

[1] Faculty of Automatic Control and Computers, University POLITEHNICA of Bucharest, 060042 Bucharest, Romania; alin.moldoveanu@cs.pub.ro (A.M.); irina.mocanu@cs.pub.ro (I.M.); florica.moldoveanu@cs.pub.ro (F.M.); emilian.radoi@cs.pub.ro (I.E.R.); victor.asavei@cs.pub.ro (V.A.); alex.gradinaru@cs.pub.ro (A.G.)

[2] Faculty of Engineering, Lucian Blaga University of Sibiu, 550024 Sibiu, Romania; alex@butean.com

* Correspondence: anca.morar@cs.pub.ro

Received: 6 March 2020; Accepted: 29 April 2020 ; Published: 6 May 2020

Abstract: Computer vision based indoor localization methods use either an infrastructure of static cameras to track mobile entities (e.g., people, robots) or cameras attached to the mobile entities. Methods in the first category employ object tracking, while the others map images from mobile cameras with images acquired during a configuration stage or extracted from 3D reconstructed models of the space. This paper offers an overview of the computer vision based indoor localization domain, presenting application areas, commercial tools, existing benchmarks, and other reviews. It provides a survey of indoor localization research solutions, proposing a new classification based on the configuration stage (use of known environment data), sensing devices, type of detected elements, and localization method. It groups 70 of the most recent and relevant image based indoor localization methods according to the proposed classification and discusses their advantages and drawbacks. It highlights localization methods that also offer orientation information, as this is required by an increasing number of applications of indoor localization (e.g., augmented reality).

Keywords: indoor localization; computer vision; QR codes; fiducial markers; 3D reconstruction

1. Introduction

In recent years, the field of indoor localization has increased in popularity due to both the increasing number of applications [1] in domains such as surveillance [2], navigation (both assistive and general purpose) [3], robotics [4–6], and Augmented Reality (AR) [7] and the many proposed solutions that differ in terms of the devices used for tracking, the type of sensor data, and the localization algorithms.

This paper focuses on computer vision based localization methods; therefore, the solutions presented are based on input from cameras. Most navigation systems use cameras carried by the subject, which represents the mobile entity (e.g., person, robot) that requires positioning or tracking, as illustrated in the left-hand side of Figure 1. The other type of solutions uses an infrastructure of static cameras positioned at known locations throughout the building to track the subject, as shown in the right-hand side of Figure 1. The vision based localization systems use 2D or 3D cameras (e.g., stereo, depth, RGB-D cameras) and perform the localization by identifying artificial markers (such as Quick Response (QR) codes and fiducial markers like AprilTags, ARTags, and CALTags [8]) or objects that are part of the environment. In many cases, the cameras are used in combination with other sensors such as WiFi, beacon, or inertial sensors [1].

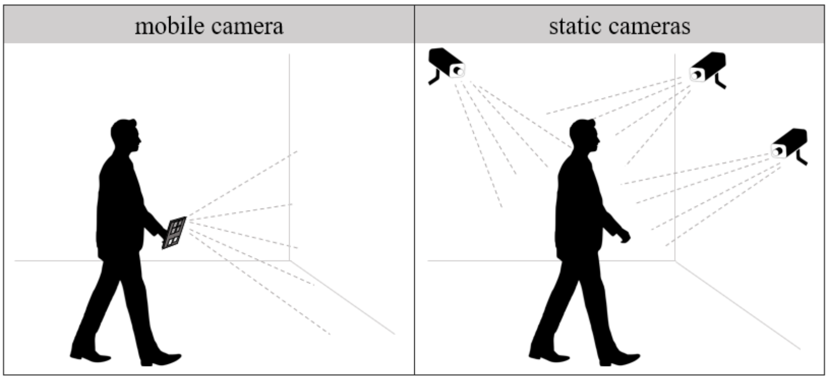

Figure 1. Indoor localization with a mobile camera (**left**) or with static cameras (**right**).

Depending on the application, the required level of location accuracy varies [9]. Navigation solutions for guiding people to find specific rooms in a building or when changing underground lines accept an accuracy of several meters. In the same accuracy range are assistive solutions for the elderly that monitor their approximate location to confirm their compliance with certain routines and to detect situations of emergency such as a person ceasing to move. Other tracking or surveillance applications require 1–2 m accuracy to assess risky situations such as a person getting too close to an exhibit item in a museum. However, some of these surveillance applications require a higher accuracy of 10–20 cm when aiming to detect whether restricted areas/perimeters are only entered by authorized people. Applications for indoor autonomous robots or assistive systems for the visually impaired that perform obstacle detection cannot rely on an approximate localization and need centimeter-level accuracy. Modern AR solutions take the accuracy requirements even further. To offer seamless integration of the multimedia content, superimposed over video flows on smartphones or over smart glasses' lenses, these applications require centimeter to millimeter accuracy of the position and orientation of the user's mobile device.

Even though a considerable number of surveys on indoor localization have been published [1,4–6,9–15], as this research space continuously developed and the types of localization solutions diversified, we find that for the area of vision based localization, the majority of the previous surveys could have better focus as they are too general (encompassing all kinds of sensing devices) or too specific (addressing only a segment of the vision based localization problem, such as Simultaneous Localization and Mapping (SLAM) [16–21], Structure from Motion (SfM) [22], or image matching [23]). Other surveys discuss indoor positioning solutions particular to certain application domains. For instance, Huang et al. [24] analyzed only localization solutions that combined visual and inertial information. Marchand et al. [7] provided a survey of pose estimation methods used only for AR. Silva and Wimalaratne [25] presented a survey of navigation and positioning aids for the visually impaired. In comparison, we offer a comprehensive survey of image based localization solutions regardless of the application domain and propose a new classification.

The paper is organized as follows: the next section describes the most impacted domains by indoor localization; Section 3 classifies 70 selected computer vision based indoor localization methods and details their main characteristics; Section 4 focuses on benchmarks used for evaluating image based indoor localization solutions; and Section 5 presents our conclusions.

2. Application Domains

2.1. Assistive Devices

With the advance in technology, researchers have focused on improving the lifestyle of people with various disabilities, including visual impairment [26]. Two of their main problems are navigating and perceiving unknown environments.

There are several solutions to this problem that map the environment with images containing different visual cues such as QR codes, bar codes, or other simple synthesized geometric shapes like circles and triangles. Idrees et al. [27] proposed an indoor navigation system that used QR codes that were placed on the floor at certain locations. The system guided the user to a selected destination, checking the user's location every time a QR code was scanned. Fusco and Coughlan [28] used sign detection and visual-inertial odometry to estimate the user's location inside a building, requiring only a digital map of that building that contained the locations of the signs.

Other solutions create a 3D reconstruction of the environment in a configuration stage or create a database with images of the indoor space, annotated with location information. Endo et al. [29] proposed a navigation system for visually impaired people, which applied Large-Scale Direct SLAM (LSD-SLAM) to estimate the user's position while constructing a 3D map of the environment. The 3D model of the environment allowed for the construction of an occupancy grid map, divided into quadrate cells, which stored information about the presence or absence of an obstacle in that location. The system created a cost map and conducted path finding through the navigation stack provided by the Robot Operating System (ROS) framework [30]. Li et al. [19] detected dynamic obstacles and applied path planning to improve navigation safety for people with visual disabilities. They built a 3D reconstruction of the environment with the visual positioning service provided by the Google Tango device. With a time-stamped map Kalman filter, they implemented an obstacle detection and avoidance algorithm that guided the user to a specified destination.

As previously mentioned, some localization solutions use an infrastructure of static cameras located at known positions in the environment. Heya et al. [31] employed color detection in an indoor localization system used for visually impaired people. A static camera located on the ceiling tracked the screen of a smartphone that was placed on the user's shoulder. Chaccour and Badr [32] proposed an ambient navigation system that was composed of static cameras attached to the ceiling. The system detected the users' location and orientation based on markers located on their heads. Static and dynamic obstacles were identified based on their shape or predefined images that were stored in a database. The proposed solution provided navigation assistance and obstacle avoidance and allowed the visually impaired users to locate missing objects.

Many navigation systems for visually impaired people assume the existence of a map for the current environment or create such a map in a configuration stage [25]. However, there are some solutions that allow the user to navigate in unknown environments by simply translating the visual information through audio or haptic signals, allowing the user to create a mental map of the surroundings. Sound of Vision [33,34] is such an example, which identifies the most important objects in the proximity of the user and sends information about their characteristics (weight, height, elevation, distance to the user, etc.) through headphones and a haptic belt. Sound of Vision is not only a navigation system, but also a solution for perceiving the environment. However, the understanding of the audio and haptic information that characterize the environment can be accomplished only through intensive training [35].

More information about indoor positioning systems for visually impaired people can be found in Siva's and Wimalaratne's survey [25].

2.2. Autonomous Robots

Designing autonomous robot applications can represent a challenge, since the localization methods cannot rely on external information. When navigating an environment, a human being can use the five senses, especially vision, touch, and hearing, to create a mental representation of the surroundings. This is not the case for robots, which navigate the environment only based on the information provided by the localization system. Therefore, localization solutions for autonomous robots require continuous computation of the robot's position and orientation relative to a digital representation of the environment, as well as obstacle detection and path planning. On the other hand, designing applications for specific robots can simplify the localization problem. Various characteristics of a robot, such as degrees of freedom, width, height, position of the sensors mounted on the robot, or wheel diameter can be used to make assumptions about the movement of the robot (dead reckoning), thus reducing the complexity of the localization algorithms.

Since most robots operate in controlled environments, a popular approach is to configure the space with artificial landmarks such as QR codes. Li and Huang [18] presented a system that assisted robots, as well as human beings, in navigating indoor environments. A Kinect device acquired color information that allowed the detection of QR codes attached to the walls at known locations in a room. The depth sensor measured the distance from the Kinect camera to the identified QR codes. Babu and Markose [36] proposed a navigation system for Internet of Things (IoT) enabled robots. A QR code based detection solution estimated the position of the robot, while a path optimization step based on Dijkstra's algorithm assisted the robot in reaching a destination node. Nazemzadeh and Macii [37] described a localization solution for unicycle-like wheeled robots. It computed the position of the robots by fusing information from QR codes, odometry based on dead reckoning, and a gyroscope platform. Cavanini et al. [38] proposed a low-cost QR code based localization system for robots operating indoors, experimentally validated on smart wheelchairs.

Other approaches used actual images of the environment, acquired with 2D or 3D cameras. Correa et al. [39] described a Kinect based reactive navigation system that guided robots while performing obstacle avoidance. It recognized different configurations of the indoor space with an artificial neural network. Xin et al. [40] introduced an RGB-D SLAM method that combined the Oriented FAST and Rotated BRIEF (ORB) and Random Sample Consensus (RANSAC) algorithms for feature extraction and matching. They created a 3D volumetric map of the environment that could be used for the navigation of a mobile robot. Kao and Huy [41] proposed an indoor navigation system that performed smartphone based visual SLAM with ORB features using a wheel-robot. They combined WiFi signals, information from inertial sensors, and monocular images for the computation of the robot's position.

Surveys dedicated to positioning solutions for autonomous robots [4,5] present in-depth information on this topic.

2.3. Augmented Reality

Currently, Augmented Reality (AR) has become very popular, due to the new technologies that are bringing it closer to the greater public. Smartphones are the most commonly used devices for displaying augmented content. However, smart glasses, such as Moverio [42] or Google Glass [43], are gaining ground [44]. For a seamless integration of the multimedia content within the real environment, the position and orientation of the display device must be estimated with high accuracy.

Gerstweiller [45] presented HyMoTrack , a tracking solution that generated a 3D model of the environment out of a vectorized 2D floor plan, and an AR path concept, called FOVPath, for guiding people. Using the FOVPath approach, the display of a trajectory depended on the user's position and orientation and also on the Field Of View (FOV) capabilities of the device. Wang et al. [46] described the development of a 3D augmented reality mobile navigation system that provided indoor localization based on Radio-Frequency Identification (RFID) readings and computer vision. They created 3D representations of internal and external structures of Oxford College, from the present and the past. Based on the position and orientation of the user's device, they displayed the 3D architectural appearance of the college during important time periods, as well as multimedia content such as texts, pictures, or 3D models related to various exhibits. Balint et al. [47] presented an AR multiplayer treasure hunt game, which combined GPS position information with localization based on image recognition. The treasures were virtual 3D objects that were displayed when the user reached a checkpoint, and the device was oriented towards the respective direction. Baek et al. [48] proposed an AR system for facility management, which computed the user's pose relative to the building, with a deep learning approach. The visualization module displayed location-specific information, holographic pipes in this case, which in reality were not visible because they were built within the walls.

Marchand et al. [7] offered more information on the topic of pose estimation for augmented reality, presenting the most important approaches on vision based positioning.

AR Commercial Solutions

This section presents some of the most popular commercial tools for developing augmented reality content, which have indoor localization capabilities.

Wikitude [49] is an augmented reality Software Development Kit (SDK) that uses the SLAM technology to reconstruct the environment. It also performs 2D and 3D image recognition and tracking, which can trigger the display of digital content, overlaid on the real world.

ARKit [50] is an iOS AR platform that provides scene understanding capabilities by combining inertial data with visual information to detect horizontal and vertical planes. It also recognizes images and 3D objects, determining the position and orientation of the camera relative to the target.

ARCore [51] is another SDK, launched by Google, which allows developers to build augmented reality experiences that seamlessly integrate the digital content into the real world. ARCore provides motion tracking capabilities, as well as environmental understanding based on plane detection. It performs SLAM, making use of inertial sensors and the data acquired with a smartphone camera, estimating the position and orientation of the user's device relative to a 3D coordinate system.

Vuforia [52] is a popular engine that provides detection and tracking of image targets and pose estimation of any tracked target or marker, allowing the rendering module to display the 3D virtual content naturally, depending on the position and orientation of the user. The pose computation is performed only relative to an image target or a marker, therefore not offering actual indoor positioning capabilities (relative to an entire room or another type of indoor space).

ARToolKit [53] is an open-source library intended for the development of augmented reality applications, which overlays 2D and 3D multimedia content on the real world. It is a tracking library that computes the camera position and orientation relative to square markers or to natural feature markers in real time. It works with both monocular and stereo cameras, providing calibration capabilities.

MAXST [54] is a cross-platform engine that provides a variety of tracking features for the development of augmented reality applications. It recognizes and tracks planar target images or planar surfaces, as well as markers with regular patterns or QR codes. Their implementation of visual SLAM can be used to create map files of the environment that are later loaded up by the Object Tracker module, which superimposes AR experiences on them. Another module, AR Fusion tracker, generates world representations and performs environment tracking by combining information from the other tracking modules.

Other popular commercial solutions for developing augmented reality applications are EasyAR [55], Kudan [56], Onirix [57], Pikkart [58], and DeepAR [59].

2.4. Surveillance and Monitoring

Indoor positioning can also be used for surveillance or monitoring purposes, detecting whether an unauthorized person has breached a perimeter, or tracking a certain person throughout an entire building or house. Generally, surveillance systems use an infrastructure of static cameras for the purpose of detecting and tracking the users. However, there are cases when information from other sensors, such as WiFi access points or beacons, is fused with the video frames.

Sun et al. [60] proposed a localization solution that used panoramic cameras and a map of the indoor environment. They applied a background subtraction method to detect human beings, matching their location to a corresponding position on the indoor map.

Desai and Rattan [61] used a pan/tilt camera and wireless sensor networks to track objects within an indoor space. The estimation of an object's position was performed with the time difference of arrival method. The camera, equipped with a laser pointer, followed the object continuously, by computing the pan and tilt angles based on a listener Cricket mote carried by the object. Grzechca et al. [62]

fused Received Signal Strength Indication (RSSI) information with data from a static video camera to track human beings in indoor environments. Zhang et al. [63] acquired video sequences with a surveillance camera and recognized a target person by matching the information provided by the inertial sensor of the person's smartphone with gait and heading azimuth features extracted from the videos. They applied a Convolutional Neural Networks (CNN) based object tracking technique in order to handle occlusion.

As far as we know, there is no recent survey on computer vision based indoor localization dedicated to surveillance and monitoring, but further information on this topic can be found in the work of Shit et al. [64], which presented localization solutions with static cameras, and in the survey of Jiao et al. [65], which discussed deep learning methods for object positioning.

3. Indoor Localization Solutions

3.1. Selection of Papers Included in the Survey

Image based indoor localization has been intensely researched. Among existing scientific publications, we chose 70 papers based on publication date and relevance to the domain, using only prestigious research databases (IEEE Explore, ACMDigital Library, SpringerLink, MDPI, and Elsevier). Figure 2 shows the distribution of selected papers over time, illustrating an increased interest in the image based localization domain in the last five years. Since it takes time for research papers to acquire visibility, the number of citations was not one of the selection criteria, as it disadvantaged the more recent research. The purpose of this survey was not to be exhaustive in terms of listing the work performed in the field of vision based indoor positioning, but to illustrate the main characteristics of the existing technologies and techniques. This allows the reader to attain an overview of the domain while understanding the advantages and drawbacks of the various methods. Choosing an appropriate solution does not boil down to just the application domain, but also to the particular requirements of the applications such as accuracy, computing time, equipment, and dynamic and static aspects (the properties of the objects contained in the environment).

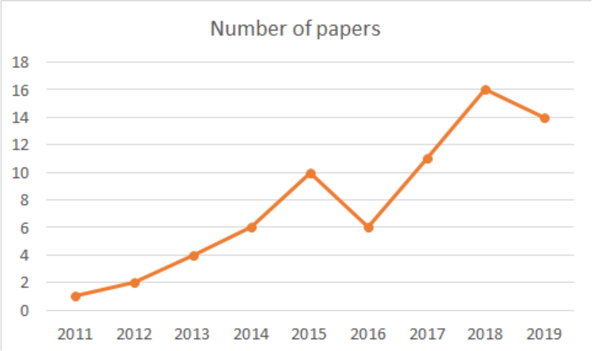

Figure 2. Distribution of selected papers over time. The horizontal axis shows the publication years. The vertical axis shows the number of papers published per year.

3.2. Classification

Recent surveys in the indoor positioning domain have proposed various classifications. For instance, the survey of Yassin et al. [1], which addressed the entire domain of indoor localization (not limited to vision based solutions), proposed a two-level classification. The first level grouped the solutions based on the positioning algorithms, which were divided into three classes: triangulation, scene analysis, and proximity detection. The second level classified the solutions within the first level classes based on the measurement techniques as follows: the triangulation class had two sub-classes: lateration and angulation; the scene analysis class had only one sub-class: fingerprinting based; and the proximity detection class had two sub-classes: cell-ID and RFID. Another general survey in indoor localization [10] classified existing research solutions into local infrastructure dependent techniques (ultra-wideband, wireless beacons), local infrastructure independent techniques (ultrasound, assisted global navigation satellite systems, magnetic localization, inertial navigation systems, visual localization, infrared localization), and visual/depth sensors (structured light technology, pulsed light technology, stereo cameras).

Mendoza-Silva et al. [9] presented a meta-review of indoor positioning systems, resulting from the analysis of 62 indoor localization-related surveys. They reviewed the most commonly used technologies for localization applications and proposed the following classes: light, computer vision, sound, magnetic fields, dead reckoning, ultra-wideband, WiFi, Bluetooth Low Energy, RFID, and Near-Field Communication (NFC). In the computer vision class, they discussed several positioning techniques, such as visual odometry and vision based SLAM, and mentioned different acquisition devices (monocular, stereo, omnidirectional). However, they did not propose any classification for this domain. They also observed the complete lack of recent surveys on computer vision based indoor localization solutions and claimed the necessity of such a work.

Analyzing the research papers mentioned in the previous section, several discriminating characteristics emerged. Therefore, we propose a new classification of computer vision based indoor localization solutions, as illustrated in Figure 3.

All indoor positioning methods have a configuration stage, in which the environment is filled with landmarks and sensors, images from the environment are saved into a database, or a 3D representation of the indoor space is created. Therefore, environment data could consist of information about the position of the markers (e.g., QR codes, geometric synthetic identifiers) or the location of the static cameras placed within the scene. Another type of environment data is represented by databases with images or features from images, annotated with position and orientation information. Lastly, environment data could consist of a 3D model of the environment, a point cloud, a 3D mesh, or a 3D map, obtained with various methods such as manual modeling, SLAM, or SfM.

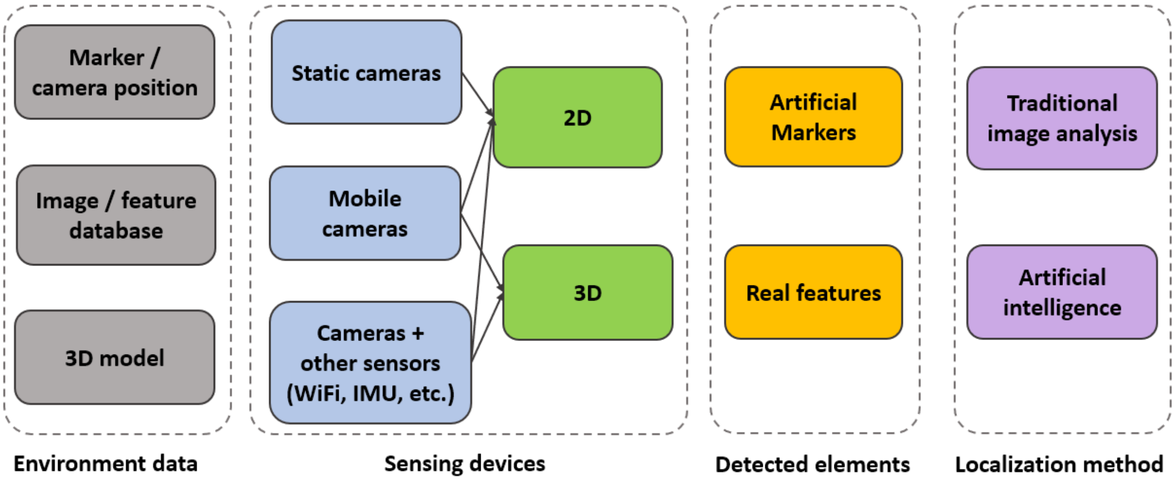

Figure 3. Proposed classification based on environment data, sensing devices, detected elements, and localization method.

Another element that helps discriminate between methods is the type of employed sensing devices. As previously mentioned in Section 1 and illustrated in Figure 1, the main acquisition devices are static and mobile cameras. Furthermore, the input information can be enriched with data from other sensors, such as WiFi access points or IMU devices. Another differentiating aspect of image based localization methods is the type of visual input, which can be either 2D or 3D.

The localization methods can search for artificial markers (e.g., QR codes and other fiducial markers such as AprilTags [66], ARTags, and CALTags [8]) or for features from the real environment. The latter category includes any type of element that can be extracted from the real environment (without the need to insert synthesized items into the scene), either features of interest such as

Speeded Up Robust Features (SURF) and Scale-Invariant Feature Transform (SIFT) or semantic objects. Therefore, we propose a new level of classification, namely the *detected elements*, which refers to the type of features (artificial markers or natural, real elements from the environment) that are tracked or matched within the images.

Indoor positioning solutions employ various localization methods, which range from low-level feature matching to complex scene understanding. We divided the techniques into traditional image analysis and artificial intelligence. The ones belonging to the second category include any type of artificial intelligence, such as Bayesian approaches, Support-Vector Machine (SVM), and neural networks.

We applied the proposed classification to the selected indoor localization solutions. Table 1 assigns each of the chosen research papers to a class, based on environment data, sensing devices, detected elements, and localization method.

Out of all the classes that could result from combining the differentiating elements from Figure 3, we chose only 17 of the more popular ones, which were represented by a large number of research papers.

Table 1. Classification of computer vision based localization research papers considering the environment data, the sensing devices, the detected elements, and the localization algorithm.

Classification				**Research Papers**
Environment Data	**Sensing Devices**	**Detected Elements**	**Localization Method**	
Marker/camera position	2D static cameras	Artificial	Image analysis	[31,32]
Marker/camera position	2D static cameras	Real	Image analysis	[60,67–69]
Marker/camera position	2D static cameras	Real	AI	[70–74]
Marker/camera position	2D mobile cameras	Artificial	Image analysis	[38,75–81]
Marker/camera position	3D mobile cameras	Artificial	Image analysis	[82,83]
Marker/camera position	2D cameras, sensors	Artificial	Image analysis	[36,37,84]
Image/feature database	2D mobile cameras	Real	Image analysis	[85–88]
Image/feature database	2D mobile cameras	Real	AI	[89–91]
Image/feature database	3D mobile cameras	Real	AI	[63,92]
Image/feature database	2D cameras, sensors	Real	Image analysis	[93–96]
Image/feature database	2D cameras, sensors	Real	AI	[97–101]
Image/feature database	3D cameras, sensors	Real	Image analysis	[102,103]
3D model	2D mobile cameras	Real	Image analysis	[29,104–111]
3D model	2D mobile cameras	Real	AI	[112,113]
3D model	3D mobile cameras	Real	Image analysis	[114–119]
3D model	3D mobile cameras	Real	AI	[120,121]
3D model	2D cameras, sensors	Real	Image analysis	[41,45,46,122–125]

The following sub-sections analyze each category, presenting representative indoor localization solutions and discussing their advantages and drawbacks. For each examined scientific paper, we include in Tables 2–18 information about the characteristics of the datasets used for evaluation, the computing time or refresh rate (related to a certain running platform), and the achieved accuracy. If there were papers that did not report information about a certain characteristic, the field corresponding to that characteristic is marked with “-”. Some papers evaluated their solutions only visually, while others applied various metrics, such as average and/or absolute errors for position and orientation, percentage of tested cases when accuracy was within certain intervals, Detection Success

Rate (DSR), Root Mean Squared Error (RMSE), Navigation Success Rate (NSR), Relative Pose Error (RPE), and Absolute Trajectory Error (ATE).

3.2.1. Indoor Localization Solutions with 2D Static Cameras, Markers, and Traditional Image Analysis

This class of indoor localization methods uses an infrastructure of 2D static cameras with known locations. The images from these cameras are processed with traditional computer vision algorithms in order to detect synthetic identifiers carried by people or robots.

Belonging to this class is the work of Heya et al. [31], where the screen of the user's smartphone was detected with a simple color tracking algorithm. Each user was assigned a color, which was displayed on the smartphone, and the system tracked the screen of the device, which was placed on the user's shoulder. Another example of indoor localization solution using static 2D cameras and traditional image processing was an ambient navigation system proposed by Chaccour and Badr [32], which detected the users' location and orientation based on markers located on their heads. The system was evaluated within a home composed of three rooms, kitchen, living room, and bedroom, each containing an IP camera placed on the ceiling. The tests performed with eight people, including dynamically added obstacles, proved the reliability of the system.

The methods in this class require the map of the building and a configuration step that consists of annotating the positions of the static cameras on the map. They can achieve good, centimeter-level, accuracy, as can be seen in Table 2, which makes them viable solutions for scenarios requiring high accuracy positioning in small spaces. However, maintaining this accuracy level in large indoor spaces comes with high costs in terms of both effort and infrastructure, due the cumbersome configurations and the high number of cameras required.

Table 2. Characteristics of indoor localization solutions with 2D static cameras (with known positions), markers, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[31]	own dataset: 1 static camera, covering 1.26 m × 1.67 m	avg. 0.2 s per frame on a server	err. between 0.0002 and 0.01 m (max. err.: 1 cm)
[32]	own dataset: 3 rooms, each with 1 IP camera	real-time	observational

3.2.2. Indoor Localization Solutions with 2D Static Cameras, Real Features, and Traditional Image Analysis

In this class of indoor localization solutions, the images from the static cameras are processed with traditional computer vision algorithms in order to track objects or people and compute their positions within a certain room. Localization solutions within this class identify people or robots without the need for the tracked entities to carry devices or artificial markers.

Bo et al. [67] recursively updated the position of multiple people based on the detected foreground and the previous known locations of each person. The foreground was identified by analyzing changes in image structure (edges) based on the computation of the normalized cross-correlation for each pixel. They applied a greedy algorithm to maximize the likelihood of observing the foreground for all people. The efficiency of their algorithm was evaluated on public datasets, using the Multiple Object Tracking Accuracy (MOTA), a metric computed based on object misses, false positives, and mismatches.

Shim and Cho [69] employed a homography technique to create a 2D map with accurate object position, using several surveillance cameras. Dias and Jorge [68] tracked people using multiple cameras and a two level processing strategy. Firstly they applied region extraction and matching to track people, and secondly, they fused the trajectories detected from multiple cameras in order to obtain the positions relative to a global coordinate system, using homography transformations between image planes.

Sun et al. [60] proposed a device-free human localization method using a panoramic camera. They employed pre-processing, human detection with background subtraction (with mean filtering and a Gaussian low pass filtering), and an association between the location of users in the image space and their location on a given map of the indoor environment.

Compared to the previous class of solutions that use artificial markers, the methods in this class have slightly higher localization errors, as can be observed in Table 3. However, this accuracy level (tens of centimeters) is still good for many types of applications, and these methods have a wider applicability, especially in the monitoring and surveillance domains, due to them not requiring the tracked entities to carry devices or markers.

Table 3. Characteristics of indoor localization solutions with 2D static cameras (with known positions), real features, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[67]	public datasets: PETS2009 [126], TUD-Stadtmitte [127]	approximately 140 ms on Intel Core2Quad 2.66 GHz with 8 GB RAM	Multiple Object Tracking Accuracy (MOTA): 87.8% on the PETS2009 and 64.2% on the TUD-Stadtmitte
[69]	own dataset: indoor space with 2.2 m × 6 m, images with 320 × 240 pixels from 2 cameras	-	less than 7.1 cm
[68]	own dataset: 12,690 frames acquired with 3 cameras; public dataset: PETS2001	-	95.7% hit rate and 96.5% precision
[60]	own dataset: office with 5.1 m × 8.5 m × 2.7 m	-	mean err. of 0.37 m

3.2.3. Indoor Localization Solutions with 2D Static Cameras, Real Features, and Artificial Intelligence

This class of indoor localization methods differs from the class described in Section 3.2.2 by the type of employed algorithms for determining the entities' positions. An alternative to traditional image processing algorithms is artificial intelligence, in the form of Bayesian approaches, SVM, or neural networks. For instance, Utasi and Benedek [70] proposed a Bayesian method for people localization in multi-camera systems. First, pixel-level features were extracted, providing information about the head and leg positions of pedestrians. Next, features from multiple camera views were fused to compute the location and the height of people with a 3D Marked Point Process (MPP) model, which followed a Bayesian approach. They evaluated their method on two public datasets and used the Ground Position Error (GPE) and Projected Position Error (PPE) metrics for accuracy computation. Cosma et al. [73] described a location estimation solution based on 2D images from static surveillance cameras, which used pose estimation from key body points' detection to extend the pedestrian skeleton in case of occlusion. It achieved a location estimation accuracy of approximately 45 cm, as can be observed in Table 4, in complex scenarios with a high level of occlusion, using a power efficient embedded computing device. See-your-room [74] represents another localization solution that uses cameras placed on the ceiling. It employs Mask R-CNN and OpenPose [128] to detect people and their pose (standing, sitting) and the perspective transformation to obtain the position of the users on a map. Hoyer et al. [71] presented a localization framework for robots based on Convolutional Neural Networks (CNN) using static cameras. In a first stage, they used a CNN object detection to estimate the type and the bounding box of a robot. In the second stage, they ran two more neural networks, one for computing the orientation of the robot and another one to provide identification (based on a code placed on the robot). An algorithm was also proposed for generating synthetic training data by placing contour-cropped images of robots on background images. The solution described by Jain et al. [72] was based on the assumption that, in an office, employees tend to keep their phones lying on the table and that the ceiling layout is unique throughout the building, containing different tiles. They used a combination of artificial intelligence and traditional image processing to detect landmarks such as ceiling tiles, heating or air conditioning vents, lights, sprinklers, audio speakers, or smoke detector sensors. First, they applied the Hough transform to extract tiles, then SURF for feature extraction, and SVM to classify the type of landmark with the ECOCframework [129].

Table 4 presents the characteristics of the localization methods that use 2D static cameras, real features, and artificial intelligence based algorithms. The computational challenge of using neural networks or other AI based implementations can be met with the use of GPUs, as can be observed for several methods [71,73], which achieve interactive or real-time performance. Although a higher complexity of the algorithms would lead to expecting a higher accuracy level compared to the previous class of solutions, relevant accuracy comparisons cannot be made due to the evaluations being performed on different datasets/scenarios.

Table 4. Characteristics of indoor localization solutions with 2D static cameras (with known positions), real features, and artificial intelligence.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[70]	public datasets: PETS 2009 (City Center), EPFLterrace dataset	-	Ground Position Error (GPE) metric with total err. rate 0.122/0.131, Projected Position Error (PPE) metric with total err. rate 0.107/0.140
[71]	training: 1542 images (own) + 25,608 images from MS COCO; evaluation: 1400 images/robot type + 110/pattern	50 Hz on a GPU and 10 Hz on a CPU	detection rate between 70% and 97.9%; orientation err. between 1.6 and 11.9 degrees
[72]	own dataset: 47 employees, 18 rooms and 6 cubicles, 960 ceiling images	2.8 s per image (offline computation)	88.2% accuracy for identifying locations
[73]	own dataset: over 2100 frames in 42 scenarios	6.25 fps on Jetson TX2	approximately 45 cm mean err.
[74]	own dataset: office room with 1 camera and supermarket with 6 cameras	5 fps on a server	detection success rate of 90% and avg. localization err. of 14.32 cm

3.2.4. Indoor Localization Solutions with 2D Mobile Cameras, Markers with Known Positions, and Traditional Image Analysis

This class of indoor localization solutions employs a configuration step, in which artificial landmarks, predominantly QR codes, are placed at known locations inside a building (generally on the ceiling, walls, or floor). These solutions make use of cameras attached to people or robots and apply traditional image processing during the localization stage. Each QR image codifies its position within the coordinate system of the building. Based on the appearance of the QR code in the acquired images during the localization stage, compared to the raw images of the QR codes, the orientation of the camera can also be estimated by computing the projective transform matrices.

QR codes allow for fast detection and decoding of stored information. However, in cases where the video camera is moving fast, the detection of these codes can be difficult. This led Lee et al. [75] and Goronzy et al. [76] to surround their codes with simple borders such as circles or rectangles, which can be detected faster than QR codes with Hough transform.

Ooi et al. [79] used QR codes to reposition mobile sensor networks, in the form of four wheeled robots. When QR codes were not in range, the system estimated the position of the robot using dead reckoning.

Lightbody et al. [78] proposed WhyCode, a new family of circular markers that enable faster detection and pose estimation, of up to two orders of magnitude compared to other popular fiducial marker based solutions. They extended the WhyCon algorithm [130], which localizes a large number of concentric black and white circles with adaptive thresholding, flood fill, and a circularity test. The position of a marker, along with the pitch and roll, was estimated based on eigenvalues with a method proposed by Yang et al. [131]. The yaw was computed by detecting the Necklace code contained in the WhyCode marker. Benligiray et al. [80] presented STag, a fiducial marker system that used geometric features to provide stable position estimation. The markers contained an inner circular border and an outer square border used for detection and homography estimation. They compared their detection capabilities against the ARToolkit, ArUco [132], and RUNE-Tag [133] fiducial markers. Khan et al. [81] proposed a generic approach for indoor navigation and pathfinding using simple markers (ARToolkit) printed on paper and placed on ceilings. The orientation of the smartphone relative to a marker enabled the computation of the user's direction along a certain path.

As can be observed in Table 5, the performance of these methods is quite impressive. The centimeter or even sub-centimeter level position accuracy is achieved due to the precise matching mechanism when dealing with synthesized images. The fast detection and decoding of QR codes and fiducial markers enables real-time applications.

Table 5. Characteristics of indoor localization solutions with 2D mobile cameras, markers with known positions, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[75]	own dataset: classroom with area 2.4 m × 1.8 m and 4 QR codes	Nexus 4 Google (fps not mentioned)	localization err. 6–8 cm, heading direction err. 1.2 angles
[76]	own dataset, simplified and complex scenarios	47 ms for QR code extraction on a Raspberry Pi 2	complex scenario: planar position err. 17.5 cm, 3D pose estimation self-localization err. 10.4 cm
[77]	public dataset proposed by Mikolajczyk and Schmid [134], 4 image pairs	0.11 s, 0.16 s, 0.27 s, 0.14 s; 1–2 iterations to reach the threshold similarity	threshold similarity 0.8
[38]	own dataset: hall with 6 QR codes and 2 possible trajectories (circular and 8-shape)	10 Hz on Linux Ubuntu 12.04 OS running ROS framework	err. for circular path 0.2 m, err. for 8-shape path 0.14 m, orientation err. 0.267 radians
[78]	own dataset acquired with an RGB camera fixed on an FLIR Pan Tilt Unit mounted on a mobile platform with an SICK s300 laser scanner (ground truth), markers placed on a wall	0.07 s avg processing time of a scene with 550 markers (up to 200 times faster than AprilTags)	avg. error of angle estimates: 0.02 rad.for pitch/roll
[79]	own dataset: space covered with 4 × 4 QR codes, placed 50 cm apart	-	the robot can travel more than 7 times on the same route
[80]	own dataset, images of resolution 1280 × 720	18.1 ms on an image with a cluttered scene and a single marker, using a single core 3.70 GHz Intel Xeon	less than 0.6 degrees std. dev. for rotation, less than 0.4 cm std. dev. for translation
[81]	own dataset in an academic building, four different paths, markers on the ceiling, guidance test with 10 blindfolded users	-	0 miss detections, 2 false detections out of 40 tests

Compared to the previously presented static camera based solutions, even though deploying such a system in a large built environment also comes with a considerable effort in the configuration stage, it is significantly less expensive (artificial markers are practically free in comparison to static cameras). However, the tracked entity is required to carry a mobile camera, which in certain scenarios can represent an inconvenience, and mapping a building with artificial images can have a negative impact on the building's appearance.

Another important aspect when choosing marker based localization solutions is their detection success when facing occlusion. This problem was addressed in the solution proposed by Garrido-Jurado et al. [132], which combined multiple markers with an occlusion mask computed by color segmentation. Sagitov et al. [8] compared three fiducial marker systems, ARTag, AprilTag, and CALTag, in the presence of occlusion, claiming that CALTags showed a significantly higher resistance for both systematic and arbitrary occlusions.

3.2.5. Indoor Localization Solutions with 3D Mobile Cameras, Markers with Known Positions, and Traditional Image Analysis

Localization based on fiducial markers can also be performed by analyzing RGB-D images with traditional image processing methods. Li et al. [82] used RGB-D images in order to detect and recognize QR landmarks with the Zbar [135] code reader. The distance to the QR code was computed based on the depth image. Dutta [83] proposed a real-time application for localization using QR codes from RGB-D images, based on the keystone effect in images from range cameras (the apparent distortion of an image caused by projecting it onto an angled surface).

Some solutions achieve centimeter accuracy when computing the distance from the camera to the artificial marker (see Table 6). These solutions are very practical, since 3D cameras already offer a depth map of the environment, allowing for a faster and less complex computation of the position in a 3D coordinate system. However, as can be observed in Section 3.2.4, detection and pose computation for markers is very fast for 2D cameras as well, due to the geometric properties of the synthetic images. Therefore, using 3D cameras could represent an unnecessary excess of resources. Furthermore, RGB-D cameras usually have a lower resolution than RGB cameras, both for the color and depth maps. Thus, their use is rarely justified for marker based solutions.

Table 6. Characteristics of indoor localization solutions with 3D mobile cameras, markers with known positions, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[82]	own dataset	-	distance from the camera to the QR code: 1 cm err.
[83]	own dataset	real time	maximum distance and angles from which the robot can see the QR code are: 270 cm and 51○.

3.2.6. Indoor Localization Solutions with 2D Cameras + Other Sensors, Markers with Known Positions, and Traditional Image Analysis

Synthetic identifiers represent a very powerful tool when estimating the subject's position and orientation in indoor scenarios. However, the use of other sensors, such as inertial sensors, WiFi, or beacons, could enrich the information, thus increasing the accuracy, or could help reduce the number of necessary synthetic landmarks. Nazemzadeh et al. [37] proposed a localization solution for unicycle-like wheeled robots, using Zbar and OpenCV to detect QR codes that were placed on the floor. They applied an Extended H-Infinity Filter (EHF) to compute the odometry based on dead reckoning and on a gyroscope platform. Babu and Markose [36] also invoked dead reckoning with accelerometer and gyroscope information, increasing the accuracy of their QR based localization solution.

Gang and Pyun [84] configured the indoor space, in an offline phase, by creating a fingerprint map with the RSSI of the beacon signals and the intensity of the geomagnetic field at each reference point. In the localization stage, they combined the information from the beacons and the inertial sensors with the coordinates extracted from QR codes, obtaining an accuracy of approximately 2 m, as can be observed in Table 7.

The use of other sensors besides cameras can add many benefits to a localization solution, especially if there is no need to acquire supplementary equipment. This is the case for WiFi access points, already installed in a building for other purposes. However, most of the WiFi localization solutions are based on the WiFi fingerprinting procedure, a manual and cumbersome configuration stage in which the signal strengths of the access points are recorded for known locations on the map of the building.

Since smartphones have become very popular and their cameras have reached impressive capabilities, they can be successfully used as acquisition devices in computer vision based localization solutions. Another advantage of using a smartphone is represented by the built-in inertial sensors. Thus, an application that combines input from the camera and the inertial sensors of a smartphone does not require equipment that is not already owned by the users.

Table 7. Characteristics of indoor localization solutions with 2D cameras + other sensors, markers with known positions, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[37]	own dataset	less than 4 ms per frame; convergence time 18 s	less than 0.2 m for position and less than 0.1 orientation for EHF
[36]	own dataset	computational load increases if dead reckoning is invoked with IMU sensors	visual (performance affected if dead reckoning is not used)
[84]	own dataset: corridor with 100 m × 2.25 m and hall with 14 m × 6.5 m	-	accuracy is within 2 m 80% of the time

3.2.7. Indoor Localization Solutions with Real Image/Feature Databases, 2D Mobile Cameras, and Traditional Image Analysis

Using a database of real images or features from real images of the environment in localization solutions represents an alternative to decorating the indoor space with QR codes or other synthesized images.

In a configuration stage, images or features, labeled with location and orientation information, are stored in a database. For instance, Hu et al. [85] obtained a panoramic video of the scene, which was processed with traditional computer vision algorithms for computing omni-projection curves. Bai et al. [86] constructed a landmark database by using a laser distance meter to measure the distance between the location of the camera and selected landmarks.

In the localization stage, the images acquired with the mobile camera were compared with the ones from the database using feature matching algorithms such as SIFT, SURF, or ORB. The processing time in this stage is highly affected by the number of images/features in the database, which must be compared against the images from the mobile camera's video flow. The first line of Table 8 is a good example, as it shows that running the localization algorithm with a database of 1000 frames was eight times faster than with a database of 8000 frames. To reduce the processing time, Elloumi et al. [87] limited the similarity search of two images to only a selection of areas within the images, thus reducing the number of features by 40%. These areas were considered to contain the most important characteristics and were selected based on a metric that combined orientation, color, intensity, flickering effects, and motion.

Compared to solutions that use artificial markers, the solutions in this class do not require decorating the indoor space with visual markers, thus not affecting the aesthetics of the indoor space. Although they have a higher localization error (few meters), this error level can still be acceptable for certain applications.

Table 8. Characteristics of indoor localization solutions with real image/feature databases, 2D mobile cameras, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[85]	own datasets: 862 frames, 3674 frames	single CPU/Kepler K 20 chip: 20 ms/1.13 ms for a database of 1000 frames, 160 ms/8.82 ms for 8000 frames	within 2 m in most cases
[86]	own datasets: hallways 15 m long	-	estimated moving speed compared to ground truth: max absolute err. 0.0643 m/s, RMSE for speed 0.24–0.37 m/s, RMSE for distance 0.16–0.23 m
[87]	own dataset: 1866 images, 40 key frames	-	-
[88]	own dataset: over 90,000 annotated frames out of 60 videos from six corridors (approximately 3.5 km of data)	-	4 m avg. absolute err. for HOG3D and 1.3 m for SF GABOR, over a 50 m traveling distance

3.2.8. Indoor Localization Solutions with Real Image/Feature Databases, 2D Mobile Cameras, and Artificial Intelligence

Artificial intelligence includes a plethora of localization algorithms for systems that use mobile cameras. For instance, Lu et al. [89] proposed a multi-view regression model to determine the location and orientation of the user accurately. Xiao et al. [90] determined the location of a smartphone, based on the detection of static objects within images acquired with the smartphone's cameras. Faster-RCNN was used for static object detection and identification. Another deep CNN, Convnet, was used in the localization system proposed by Akal et al. [91]. This network uses compound images from four non-overlapping monocular images placed on a ground robot, achieving centimeter accuracy, but requiring a sizeable dataset of compound images for training. As can be observed in Table 9, the machine learning based solutions achieved interactive computing times or even real-time performance and a localization accuracy of under one meter to tens of centimeters. These solutions seemed to have better accuracy performance compared to the solutions in the previous class, while benefiting from the same advantages of not requiring deploying visual markers in the indoor space.

Table 9. Characteristics of indoor localization solutions with real image/feature databases, 2D mobile cameras, and artificial intelligence.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[89]	own dataset: 1800 images from 30 different locations, 480 indoor videos of buildings (each lasting around 2–3 s), public dataset: Dubrovnik [136]	0.00092 s for image based localization and 0.0012 s for the video based method	95.56%/94.44% accuracy for location/orientation with image based localization and 98% with the video based method
[90]	own dataset: 302 training images, resolution 3024 × 4032	object detection phase takes 0.3 s	location accuracy is within 1 m
[91]	own dataset: 112,919 compound images (composed of 4 images taken by 4 Google Nexus phones) of resolution 224 × 224	close to real-time	avg. median err. after a 20 step moving for compound images is 12.1 cm

3.2.9. Indoor Localization Solutions with Real Image/Feature Databases, 3D Mobile Cameras, and Artificial Intelligence

Another class of indoor localization methods uses RGB-D images acquired with mobile cameras that are processed with the help of CNN. Guo et al. [92] used a CNN (PoseNet network) for exploiting the vision information and the long short-term memory network for incorporating the temporal information. Zhang et al. [63] applied visual semantic information for performing indoor localization. A database with object information was constructed using Mask-RCNN, extracting the category and position for each object. Then, using the SURF descriptor, keypoints of the recognized objects were detected. Furthermore, CNN features were obtained using a pre-trained ResNet50 network. The visual localization was performed in two steps: the most similar key frames were obtained using the selected CNN features; the bundle adjustment method [137] was used to estimate the matrix between the current image and candidate frames. Both methods were tested on public datasets. Localization results were within 0.3 m and 0.51 m (as shown in Table 10).

3D cameras give access to a depth map of the environment, either through built-in algorithms, as in the case of structured light or time-of-flight devices, or through stereo matching algorithms that have multiple implementations, available to the public. However, these cameras come with various limitations. For instance, the estimation of the depth map with stereo cameras in the case of untextured surfaces (such as white walls) is very inaccurate. Furthermore, structured light and time-of-flight depth cameras cannot estimate the distance to reflective surfaces or in case of sunlit environments. Moreover, although 3D cameras have gained popularity, they are not as common as 2D cameras, and therefore, their applicability is reduced. While localization solutions with 2D mobile cameras can be easily deployed, using generally available smartphones, 3D cameras are more appropriate for specialized applications, in areas like assistive devices or autonomous robots.

Table 10. Characteristics of indoor localization solutions with real image/feature databases, 3D mobile cameras, and artificial intelligence.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[92]	ICL-NUIMdataset [138] and TUM dataset [139]	296 ms to find the most similar frame and 277 ms to estimate the final pose on Intel Xeon E5-1650 v3 CPU 3.5 GHz, NVidia TITAN GPU	more than 80% of the images are localized within 2.5 degrees and more than 90% are localized within 0.3 m
[63]	ICL-NUIM dataset [138]	-	0.51 m living room, 0.41 m office

3.2.10. Indoor Localization Solutions with Real Image/Feature Databases, 2D Cameras + Other Sensors, and Traditional Image Analysis

If WiFi signals, inertial sensors, beacons, or other sensors can increase the accuracy of marker based localization solutions or can help reduce the number of synthesized images that should be placed on the ceiling/floor/walls of the building (as discussed in Section 3.2.6), a hybrid approach can be even more useful when dealing with natural features from the environment. Acquiring additional information from various sensors can help reduce the search space in the image matching stages.

Yan et al. [94] also used WiFi information to increase the accuracy and improve the processing time of a natural feature extraction algorithm, which combined Features from Accelerated Segment Test (FAST) with SURF.

Marouane et al. [93] used accelerometer data for step counting and gyroscope information for orientation and transformation of images into histograms for more efficient image matching. Rotation invariance was achieved by adding the perspective transformation of two planes. Another solution that used inertial sensors was the one proposed by Huang et al. [95]. They applied the vanishing points method and indoor geometric reasoning, taking advantage of rules for 3D features, such as the ratio between width and height, the orientation, and the distribution on the 2D floor map. Arvai and Dobos [96] applied the perspective-n-point algorithm to estimate the user's position inside the 2D floor-plan of a building, relative to a series of landmarks that were placed in the configuration stage. They used an extended Kalman filter to estimate the position by combining visual and inertial information.

Table 11 presents the characteristics of indoor localization solutions that combine data from 2D cameras and other sensors, estimating the position and orientation of the subject with traditional image processing. Several such solutions achieved centimeter location accuracy, due to this fusion between images and information from inertial sensors, WiFi signals, RFID devices, or beacons. However, this fusion of data from several sensors brings a computational load.

Table 11. Characteristics of indoor localization solutions with real image/feature databases, 2D cameras + other sensors, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[93]	own dataset: 75 location images from 1.5–2 m distance	query time 40–230 ms	mean distance err. rate is 2.5/2.21 m for extended distance estimation method/hybrid approach
[94]	own dataset	90 ms FAST-SURF, 100–130 ms indoor positioning, 3–7 ms character detection, 30–45 ms tracking and registration on Honor 3C smartphone	-
[95]	own dataset (offices and hallways)	0.5 s per frame	90% of location and orientation errors are within 25 cm and 2 degrees
[96]	-	iPhone5s, iPhone X, LG Nexus 5X, Samsung Galaxy S7, S9, Huawei Mate tablet	best results, on Samsung S9: 4.5 deg. avg. rotation and 250 mm position err. from 1 m in front of the marker

3.2.11. Indoor Localization Solutions with Real Image/Feature Databases, 2D Cameras + Other Sensors, and Artificial Intelligence

The solutions based on the detection of objects or markers from RGB images offer a relative position and orientation estimation, but are unreliable when markers or objects are not visible. Furthermore, detection is influenced by camera exposure time. Thus, images combined with data from other sensors can increase the precision of the localization.

Rituerto et al. [97] estimated the user's location using values acquired from inertial sensors combined with computer vision methods applied on RGB images. The particle filtering method was used for combining all these data. A map with walls, corridors, and rooms and some important signs (such as exit signs and fiducial markers) was also considered.

Neges et al. [98] combined an IMU step based counter with video images for performing indoor localization. IMU data were used to estimate the position and orientation of the mobile device, and different semantic objects were extracted from the video (e.g., exit signs, fire extinguishers, etc.) for validation of the obtained position. The recognition of different markers was achieved using Metaio SDK [140], a machine learning based development tool. In Sun et al. [99], RSS samples, surveillance images, and room map information were used for performing indoor localization. People were detected using background subtraction from images acquired with a camera placed on the ceiling of the room. The foreground pixel that was the nearest to the location of the camera would approximate the person position in the image. Then, this position was mapped to a localization coordinate using a multi-layer neural network (with three layers). The iStart system [100] combines WiFi fingerprints and RGB images for indoor localization. The system proposed by Zhao et al. [101] was based on a combination of CNN with a dual-factor enhanced variational Bayes adaptive Kalman filter. Channel State Information (CSI) was extracted from an MIMO-OFDM PHY layer as a fingerprint image to express the spatial and temporal features of the WiFi signal. CSI features were learned with a CNN inspired by the AlexNet network obtaining the mapping relationship between the CSI and the 2D coordinates. Results were processed with the Bayes adaptive Kalman filter in order to achieve noise attenuation. These methods were evaluated on their own datasets with good results (position accuracy of approximately 1 m), as shown in Table 12.

Table 12. Characteristics of indoor localization solutions with real image/feature databases, 2D cameras + other sensors, and artificial intelligence.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[97]	own dataset, 3 blind volunteers	2 fps on a laptop	visual inspection
[98]	own dataset, 5 people with different weight and height, walking at 3 different speeds, on two tracks (straight or zig-zag)	real time	93% accuracy in case of normal speed
[99]	own dataset: office floor 51 m × 20 m × 2.7 m and 7 WiFi routers	-	panoramic camera based method: mean err. for localization 0.84 m, cumulative probability within localization err. of 1 m/2 m is 70%/86%
[100]	own dataset: room-level environment and open large environment	4 s per image (0.8 s fingerprint location on server, 2.9 s image location on smartphone, 1 s data transmission)	less than 0.6 m avg. location err. and less than 6 degrees avg. direction err., 90% location deviations are less than 1 m
[101]	own dataset: 50 m^2 office and 2 cases (with and without line-of-sight); ground divided into 42 reference points	real time on Intel5300 NIC laptop with 3 antennas as signal receiver and Ubuntu server with Intel Xeon e5-2609 CPU, GeForce GTX TITAN X GPU and 256 GB RAM	avg. position err. is 0.98/1.46 m for line-of-sight/none line-of-sight

Even though artificial intelligence and especially deep convolutional networks have become very popular, they still come with certain limitations. First, they require a large amount of training data, usually manually annotated. Second, the training stage is both time consuming and hardware demanding. Even though in the online stage, the already trained network requires less resources, adding the complexity of fusing the visual data with information from other sensors can have a negative impact on the runtime, as can be observed for several selected papers [97,100].

3.2.12. Indoor Localization Solutions with Real Image/Feature Databases, 3D Mobile Cameras + Other Sensors, and Traditional Image Analysis

Localization precision can be increased by matching of RGB-D images using traditional feature descriptors combined with information obtained from an IMU sensor. In Gao et al. [102], key points were extracted from the RGB-D images using an improved SIFT descriptor. Then, the RANSAC algorithm [141] eliminated mismatched points from the matching pairs. Their corresponding depth coordinates were obtained from the depth images. Using this information, the rotation matrix and translation vector were computed from two consecutive frames. Furthermore, IMU data were used to eliminate the noise, improving the stability and positioning accuracy. Adaptive fading extended

Kalman filter fused the position information of Kinect and IMU outputs. Furthermore, this fusion eliminated the noise and improved the stability and accuracy of the system. A similar idea was proposed by Kim et al. [103]. Their solution generated 3D feature points using the SURF descriptor, which were next rotated using IMU data to have the same rigid body rotation component between two consecutive images. The RANSAC algorithm [141] was used for computing the rigid body transformation matrix. Table 13 shows the dataset characteristics and obtained accuracy for the localization methods based on RGB-D images processed with traditional image analysis algorithms and sensor fusion. Since robots can be equipped with many sensors, including 3D cameras and inertial units, the solutions in this class have been successfully applied to the autonomous robots domain.

Table 13. Characteristics of indoor localization solutions with real image/feature databases, 3D mobile cameras + other sensors, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[102]	own dataset: small number of experiments and short tested trajectories	real time	avg. err. in the X-axis direction is 0.06 m with IMU
[103]	own dataset	real time	translation error: 0.1043 m, rotation err.: 6.6571 degrees for static environments; translation err.: 0.0431 m ± 0.0080 m, rotation err.: 2.3239 degrees ± 0.4241 degrees for dynamic environments

3.2.13. Indoor Localization Solutions with a 3D Model of the Environment, 2D Mobile Cameras, Real Features, and Traditional Image Analysis

Simultaneous Localization and Mapping is a very popular algorithm in several domains, such as autonomous robots or Augmented Reality. During recent years, various solutions to the problem of localization and mapping have been proposed. For instance, Endo et al. [29] used LSD-SLAM for map construction, localization, and detection of obstacles in real time. Teixeira et al. [104] used the pattern recognition SURF method to locate natural markers and reinitialize Davison's Visual SLAM [142].

Several SLAM based solution use 3D cameras in the configuration stage, to create a 3D reconstruction of the environment, and then change the acquisition device to a monocular camera in the localization stage. Sinha et al. [105] applied RGBD-SLAM on images acquired with Microsoft Kinect to reconstruct 3D maps of indoor scenes. In the localization stage, they used monocular images acquired with a smartphone camera and estimated the transformation matrix between frames using RANSAC on the feature correspondences. They applied SIFT or SURF for feature extraction, in order to detect landmarks, which were cataloged as sets of distinguished features regularly observed in the mapping environment, being stationary, distinctive, repeatable, and robust against noise and lighting conditions. Deretey et al. [106] also applied RGBD-SLAM in an offline, configuration stage, to create 3D point clouds that contained intensity information. 2D features were extracted with a matching algorithm (SIFT, SURF or ORB), and then, a projection matrix of matched features between 2D images and 3D points was computed. A comparison with RGBD-SLAM was offered by Zhao et al. [109], which used Kinect to collect the 3D environment information in a configuration stage. They also built a 2D map of the indoor scene with Gmapping, an ROS package that used Rao–Blackwellized Particle Filters (RBPF) [143] to learn grid maps. In the online phase, they applied Monte Carlo localization based on the previously created 2D map.

Ruotsalainen et al. [107] performed Visual SLAM for tactical situational awareness by applying a Kalman filter to combine a visual gyroscope and a visual odometer. The visual gyroscope estimated the position and orientation of the camera by detecting straight lines in three orthogonal directions. The visual odometer computed the transformation of the camera from the motion of image points matched using SIFT in adjacent images. A similar approach, which took into account the structural regularity of man-made building environments and detected structure lines along dominant directions, was the solution proposed by Zhou et al. [108]. They also applied an extended Kalman filter to solve the SLAM problem. Ramesh et al. [110] combined imaging geometry, visual odometry, object detection with aggregate channel features, and distance-depth estimation algorithms into a Visual SLAM based navigation system for the visually impaired.

A different approach was the one proposed by Dong et al. [111], which reused a previous traveler's (leader) trace experience to navigate future users or followers. They used ORB features for the mobile Visual SLAM. To combat environmental changes, they culled non-rigid contexts and kept only the static contents in use.

SLAM based approaches can attain centimeter or even millimeter location accuracy, but at a high computational cost. They also require significant memory resources to store the 3D representation of the scene. Table 14 presents the characteristics of some solutions that create a 3D reconstruction of the environment in an offline stage, acquire images with a monocular camera in the localization stage, and perform low-level image processing to estimate the position and orientation of the user/robot.

Table 14. Characteristics of indoor localization solutions with an existing/generated 3D model of the environment, 2D mobile cameras, real features, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[29]	own dataset: simple experiment with obstacles along a route	real time on a single CPU	visual inspection
[104]	own dataset: images of resolution 320 × 240	3 fps for SURF on 2.20 GHz dual-core computer	90% detection success rate and 14.32 cm avg. localization err.
[105]	3 own datasets: images of resolution 640 × 480 (from Galaxy S4 camera); public dataset: feature set of Liang et al. [144]	avg. search time for Dataset 1 (176 frames) is 10 ms and for Dataset 4 (285 frames) is 28 ms	80–100% accuracy, depending on dataset; 0.173–0.232 m localization error
[106]	own dataset: reconstruction with RGBD-SLAM, Dataset 1 (50 frames, 139 mp), dataset 2 (33 frames, 37 mp)	avg. localization time is 0.72 s per frame on an Intel Core i7 with 8 GB RAM	avg. localization error is less than 10 mm: translation err. for Dataset 1 is 0.9–35 mm and for Dataset 2 is 0.3–17 mm
[107]	own dataset: office environment, with 154 m long route	images captured at 0.8 Hz using a smartphone camera, computing time not mentioned	1.5 degrees mean accuracy for visual gyroscope, 0.3 m/s mean accuracy for visual odometer, 1.8 m localization err.
[108]	own dataset: synthetic scenes (20 m × 20 m scene with 88 lines and 160 points, 794 generated images); public dataset: Biccoca_2009 [145]	0.5–1 s for MATLAB version, 25.8 ms avg. running time for C++ version	0.79 accuracy err. in position on a 967 m path; 0.2 m accuracy err. for synthetic scenes
[109]	own dataset: Guangdong Key Laboratory, Shantou University	-	only visual inspection, in comparison with the RGB-D SLAM method [146]
[110]	own dataset: indoor environment; public dataset: Karlsruhe outdoor datasets [147]	real time on an i7 processor	94–98% distance and depth measurements accuracy; absolute err. of 5.72/9.63 m for 2 outdoor datasets and 4.07/1.35 cm for 2 indoor datasets
[111]	own datasets: office building (400 m^2), gymnasium (1000 m^2), and a shopping mall (6000 m^2), 21 navigation paths, 274 checkpoints	approximately 0.1 s for relocalization and navigation on Huawei P10, Nexus 6, Nexus 7, Lenovo Phab2 pro	98.6% immediate NSR, 93.1% NSR after 1 week, 83.4% NSR after 2 weeks

3.2.14. Indoor Localization Solutions with a 3D Model of the Environment, 2D Mobile Cameras, Real Features, and Artificial Intelligence

Artificial intelligence based 2D localization methods can also be applied on 3D representations of the space. Han et al. [112] removed obstacles detected with the Mask-RCNN network to enhance the performance of the localization. It detected persons as potential obstacles and split these obstacles from the background. Then, ORB-SLAM2 [148] was used for localization. Xiao et al. [113] proposed Dynamic-SLAM for solving SLAM in dynamic environments. It was based on ORB-SLAM. First, a CNN was used for static or dynamic object detection. Then, applying a missed detection compensation algorithm based on the speed invariance from adjacent frames, the detection recall rate was improved. Finally, tracking was performed using ORB features extracted from each keyframe image for performing feature based visual SLAM by processing feature points of dynamic objects. The pose estimation was obtained by solving the perspective-n-point problem with the bundle adjustment method.

Table 15 presents the characteristics of some solutions belonging to the current class. The neural networks introduce a high computational load, but can help not only with the localization, but also with the scene understanding problem.

Table 15. Characteristics of indoor localization solutions with an existing/generated 3D model of the environment, 2D mobile cameras, real features, and artificial intelligence.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[112]	public dataset: TUM Dynamic Object dataset (RGB images, depth information, ground truth trajectory)	5fps for Mask-RCNN on NVidia Tesla M40 GPU [149]	RMSE between 0.006134 and 0.036156
[113]	own dataset: 370 m from route; public datasets: TUM dynamic dataset, KITTI dataset (outdoor large scenarios)	real time	the trajectory RMSE err. is 2.29 m, the accuracy is 7.48–62.33% higher than ORB-SLAM2 [148]

3.2.15. Indoor Localization Solutions with a 3D Model of the Environment, 3D Mobile Cameras, Real Features, and Traditional Image Analysis

Several localization solutions use 3D cameras in the configuration step, as well as in the actual localization stage. For instance, Du et al. [114] created an interactive mapping system that partitioned the registration of RGB-D frames into local alignment, based on visual odometry, and global alignments, using loop closure information to produce globally consistent camera poses and maps. They combined RANSAC inlier count with visibility conflict in the three point matching algorithm to compute 6D transformations between pairs of frames. Paton and Kosecka [115] applied feature extraction and mapping on RGB-D data with SIFT, motion estimation and outlier rejection with RANSAC, and estimation refinement to compute the position and orientation of a camera. Correspondences established between SIFT features could initialize a generalized Iterative Closest Point (ICP) algorithm.

Salas-Moreno et al. [116] proposed a GPGPUparallel 3D object detection algorithm and a pose refinement based on ICP. Their real-time incremental SLAM was designed to work even in large cluttered environments. Prior to SLAM, they created a database of 3D objects with KinectFusion. The scene was represented by a graph, where each node stored the pose of an object with a correspondent entry in the database. Their object level scene description offered a huge representation compression in comparison with the usual reconstruction of the environment into point clouds.

A robust key-frame selection from RGB-D image streams, combined with pose tracking and global optimization based on the depth camera model, vertex-weighted pose estimation, and edge-weighted global optimization, was described by Tang et al. [118].

Most solutions acquire images with structured light or time-of-flight cameras, but stereo cameras can also provide 3D information. For instance, Albrecht and Heide [117] acquired images with a stereo camera and applied ORB-SLAM2 for poses of the keyframes, creating a 3D reconstruction of the environment with OpenCV's Semi-Global Block Matching (SGBM) algorithm. Then, they condensed the point cloud into a blueprint-like map of the reconstructed building, based on ground and wall segmentation.

Martin et al. [119] applied Monte Carlo based probabilistic self-localization on a map of colored 3D points, organized in an octree. They demonstrated that their algorithm recovered quickly from cases of unknown initial position or kidnappings (the robot was manually displaced from one place of the environment to another).

Table 16 presents the computing capabilities and obtained accuracy for several SLAM based localization solutions that apply low-level image processing on data that contain both color and depth information. It can be observed that some of the researchers evaluated their algorithms only through visual inspection. Even so, inspection of the obtained 3D reconstruction and especially loop closure can demonstrate the performance in the case of SLAM based solutions. This class is reduced to a 3D to 3D matching problem, much less complex than the 3D to 2D matching problem described in Sections 3.2.13 and 3.2.14. However, the requirement to have a 3D camera both in the configuration stage and in the online phase greatly reduces the applicability of this kind of solution.

Table 16. Characteristics of indoor localization solutions with an existing/generated 3D model of the environment, 3D mobile cameras, real features, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[114]	own dataset: a room, check if the virtual objects have the same size as the real ones	3–4 fps for map building on a laptop with i7-720qm CPU	difference in dimensions between 3D reconstructed and real objects: dm/cm level accuracy
[115]	own dataset: barren office hallway; public dataset: TUM dataset [139]	-	0.1–1.5 m translation RPE, 2–18 degrees rotational RPE, 0.02–1.1 m ATE
[116]	own dataset: room of size 15 × 10 × 3 m^3	20 fps on a gaming laptop	visual inspection, checking loop closure
[117]	own dataset: a path of 70 m through a building	25 Hz	visual inspection
[118]	own datasets: taken with a handheld structure sensor; public datasets: Freiburg Benchmark, TUM dataset [139]	5 fps	0.011–0.062 RMSE of ATE for public datasets; 1.4–4.1 cm closing distance and 1.08–3.32 degrees closure angle for own datasets
[119]	own dataset collected with 2 wheeled robots (RB-1 and Kobuki) with Asus Xtion RGB-D sensors	less than 30 s	tracking mode (robot starts from a known position): x-mean: 0.082 m, y-mean 0.078 m; global mode (robot starts from unknown position): x-mean 0.27 m, y-mean 0.43 m

3.2.16. Indoor Localization Solutions with a 3D Model of the Environment, 3D Mobile Cameras, Real Features, and Artificial Intelligence

Another class of indoor localization solutions is the one that uses 3D cameras in the configuration stage, to create a reconstruction of the scene with SLAM or other algorithms, but also in the localization stage, applying high level computer vision techniques for computing the position and orientation of the user.

Guclu et al. [121] proposed an SLAM method applied on RGB-D images using a graph based approach. The keyframe autocorrelogram database estimated motion between frames. Keyframes were indexed based on their image autocorrelograms [150], using a priority search k-means tree. Adaptive thresholding was used to increase the robustness of loop closure detection.

Kuang et al. [120] improved ORB-SLAM. A combination between quasi-physical sampling algorithm (based on BING features [151], obtained by SVM training) with depth information was used to pre-process an image for decreasing the computing time of the ORB algorithm. Then, improved KD-trees were used to increase the matching speed of the ORB algorithm. Furthermore, using RGB-D images, a 3D dense point cloud map system was constructed, instead of a sparse map from ORB-SLAM.

As can be observed in Table 17, the use of 3D cameras can improve the accuracy of known localization methods such as ORB-SLAM or ORB-SLAM2. Still, the dimensionality of the data introduces a high computational cost. Furthermore, the lack of 3D training data could represent a limitation of this class.

Table 17. Characteristics of indoor localization solutions with an existing/generated 3D model of the environment, 3D mobile cameras, real features, and artificial intelligence.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[120]	public dataset: TUM dataset [139]	37.753 ms per frame on Intel i5 2.0 GHz CPU with 3 GB RAM	0.015 and 0.103 RMSE for the error size of the posture, better than ORB-SLAM [152]
[121]	public dataset: ICL-NUIM [138] and TUM dataset [139]	119.0 ms (average) on a desktop PC running Ubuntu 12.04 with an Intel Core i7-2600 CPU at 3.40 GHz and 8 GBRAM	Absolute Trajectory Error ATE: 1 cm–5 cm, mostly competing with ORB-SLAM2 [152]

3.2.17. Indoor Localization Solutions with a 3D Model of the Environment, 2D Mobile Cameras + Other Sensors, Real Features, and Traditional Image Analysis

A hybrid approach that fuses information from 2D cameras and other sensors can be applied on 3D models of the environment as well.

For instance, Wang et al. [46] used RFID readers for an approximate estimation of the location and calculation of 3D image coordinates with low-level image matching.

Kao and Huy [41] combined information from WiFi access points with the K-nearest neighbor method, inertial sensors (accelerometer and gyroscope), and a CMOS camera. They chose ORB features in their SLAM implementation to navigate Bluetooth connected wheeled robots in indoor environments.

Yun et al. [122] saved the WiFi access point information in a configuration stage and assembled the images acquired with an Xtion PRO LIVE depth camera, building a 3D indoor map of the indoor location. In the localization stage, they reduced the per-frame computation by splitting a video frame region into multiple sub-blocks and processing only a sub-block in a rotating sequence at each frame. They applied SIFT based keypoint detection and optical flow for tracking.

Huang et al. [95] applied an extended Kalman filter to fuse data from LSD-SLAM computed on RGB images, ZigBee localization, and IMU sensors (accelerometer, gyroscope, and magnetometer). Ullah et al. [125] combined data from a monocular visual SLAM and an IMU with an unscented Kalman filter. Gerstweiler [45] also fused IMU information with SLAM, using the HyMoTrack framework [153], a hybrid tracking solution that uses multiple clusters of SLAM maps and image markers, anchored in the 3D model.

Chan et al. [124] computed a laser based SLAM and a RBPF based visual SLAM. Perspective trajectories obtained from the laser SLAM were mapped into images, and the essential matrix between two sets of trajectories was combined with the monocular camera based SLAM.

Even if the fusion between sensor data and visual information introduces a high computational load, several solutions achieve real-time frame rates on commodity computers, as can be observed in Table 18.

Table 18. Characteristics of indoor localization solutions with an existing/generated 3D model of the environment, 2D mobile cameras + other sensors, real features, and traditional image analysis.

Research Paper	Dataset Characteristics	Computing Time and Platform	Accuracy
[46]	own dataset	-	visual evaluation
[41]	own dataset: robot moves along a specific path in a lab room, QVGA resolution	200 ms for basic image processing on LG P970, whole pipeline processed offline (manual extraction of features)	position err. converges from 35–50 cm to less than 3 m
[122]	own dataset: 120 m indoor hallway with 5200 video frames of size 640 × 480	from 3.2 fps to 23.3 fps on commodity laptop (2.6 GHz quad-core CPU and 4 GB RAM)	0.17 m position err.
[95]	own dataset	-	visual inspection
[124]	own dataset	map building and fusion process: real time on Intel Core i7-8550U CPU	relative error of ORB-SLAM2 [148] calibrated with proposed mapping matrix is less than 5%
[125]	own dataset and public dataset: EuRoCdataset [154]	real time on an embedded board (1.92 GHz processor and 2 GB DDR3L RAM)	0.01–0.15 m position err. for own dataset; 0.234 m max. err. for EuRoC dataset
[45]	own dataset: Vienna airport, path of 200 m	23 s for the proposed method to complete a guiding task	visual inspection

3.3. Discussion

This section draws conclusions from our analysis of the proposed classes of vision based localization solutions, enabling readers to make better informed choices in terms of indoor positioning technologies to accommodate their specific requirements or particularities. While positioning technologies are numerous and do not limit themselves to image processing, vision based solutions have become popular due to the increasing affordability of cameras and their integration in pervasive devices such as smartphones.

Localization methods that use static cameras can benefit from the camera surveillance infrastructure already available in most modern large office and public buildings. Furthermore, since most robotic platforms have RGB or RGB-D cameras, it makes it easier to port visual positioning solutions on the different platforms, enabling their use in assisted living scenarios. Other applications in the autonomous robots domain can take advantage of 2D/3D cameras already integrated in the robots. Smart glasses with cameras can enable a more seamless user experience for indoor localization applications; however, until they reach a wide market adoption, most applications that localize people (especially in the domains of assistive devices and augmented reality) use smartphones.

Even though 2D cameras have larger applicability due to their ubiquity and the dimensionality of the acquired data, 3D cameras have several advantages. 3D cameras offer a depth map of the environment, either obtained from a disparity map computed with stereo matching algorithms or estimated with time-of-flight and structured light technologies. Stereo cameras require optimal lighting conditions and are affected by lens distortion, similar to 2D cameras. Furthermore, depth cannot be estimated in untextured environments through stereo matching. On the other hand, structured light and time-of-flight cameras work even in unlit environments and can estimate the depth regardless of texture properties. Although these cameras are affected by bright light and reflective surfaces, typical indoor environments contain untextured surfaces (especially uniformly painted walls) and are rarely characterized by bright sunlight. Therefore, we considered that among 3D cameras, structured light, and time-of-flight devices are the most suited for indoor applications.

While cameras have many advantages, they are affected by lighting conditions, occlusion, and position changes of objects from the environment. In order to increase localization accuracy or to decrease the computational load of the computer-vision algorithms, visual data can be combined with data from other sensors. Other popular indoor localization solutions are those based on sensors such as WiFi, beacons, and RFID. WiFi based solutions use the received signal strength and the media access control address of access points to determine the position. WiFi based methods also enjoy the advantage of using existing infrastructure in buildings, as WiFi access points are even more widely available in buildings than camera surveillance systems. While beacon based positioning technologies can reach higher accuracy than WiFi based solutions, they require deploying additional hardware. The RFID technology poses even more limitations in terms of range. Although the positioning algorithms that use sensors such as WiFi, beacons, or RFID have a lower accuracy compared to vision based methods, they also have a lower complexity. Thus, possible localization solutions can benefit from a two-step positioning algorithm: firstly obtaining a quick, approximate location using beacons or WiFi, which tightens the search space of computer vision algorithms; secondly achieving an accurate and also quick location and orientation estimation of the tracked entity.

Vision based indoor localization solutions can detect fiducial markers or features from real images of the environment. The use of artificial markers enables extremely fast detection and position estimation. Due to the geometric properties of the fiducial markers and their accurate localization with 2D cameras, the use of 3D cameras is unjustified. The biggest disadvantage of using markers is the requirement of covering the space with synthetic images, which can have a negative visual impact on the environment. Therefore, the applicability of such solutions is reduced. Features or semantic objects detected from real images of the environment do not visually influence the environment. However, setting up a database of features/images, annotated with position and orientation information, or creating a 3D model of the environment represent cumbersome processes. Furthermore, changes in the environment, such as rearranging furniture or paintings and posters, would require another configuration stage for rebuilding the feature/image database or the 3D model of the scene.

Objects or features from images can be detected using traditional image processing or artificial intelligence methods. Traditional image processing methods perform detection by comparing different features that are extracted from the images, and the recognition success depends on the selected features. On the other hand, the artificial intelligence methods used for object recognition are mainly based on convolutional neural networks, thus not needing to select the features for recognizing objects, as convolutional neural networks learn specific objects directly from images. One disadvantage of these networks is the high number of images required for training the network. Training data can be obtained either by manual acquisition and annotation or from publicly available datasets. Public datasets are very helpful; however, they are only a few available (especially containing 3D data), and they are limited to several semantic classes.

4. Benchmarks

The evaluation methods used for the indoor localization solutions presented in this paper differ. Some are based on visual inspection, some on testing the solutions in certain scenarios or testbeds, and some on using public datasets. This section presents benchmarks created for evaluating localization methods. They can differ based on the input information, which consists of monocular or RGB-D images and WiFi, along with other sensor readings. Some testbeds are designed with the purpose of evaluating only the location and orientation accuracy, while others can also evaluate the correctness of 3D reconstructions in SLAM based methods. Several research papers released to the public a series of datasets and evaluation tools [139,155,156], while others proposed reference systems that can be used for testing the accuracy of localization solutions [157], the latter enabling fair comparisons of existing localization systems in similar conditions [158–160].

Sturm et al. [139] proposed a benchmark for the evaluation of RGB-D SLAM based localization solutions. The TUM dataset and this benchmark represent a popular testing tool. This is noticeable in the tables in Section 3. The database consists of images acquired with a Kinect sensor, containing both color and depth information, at a resolution of 640 × 480. They provide ground truth trajectories that are computed with a motion-capture system composed of eight cameras that acquire images at 100 Hz. These sequences cover a variety of cases, from short to long trajectories, with or without loop closure. Their benchmark offers automatic evaluation tools to assess the drift of visual odometry solutions, as well as the global pose error of SLAM based methods. Another popular benchmark, which contains the ICL-NUIM dataset, was provided by Handa et al. [138]. The database consists of RGB-D frames within synthetically generated scenes with the point of view of handheld cameras. It contains ground truth camera poses and surface models, which enable not only the evaluation of localization solutions, but also of the surface reconstruction accuracy of SLAM based methods. Sun et al. [155] proposed a dataset for evaluating computer vision based localization solutions that compute the pose of a 2D camera with respect to a 3D representation of the scene. Their database contained training data acquired with cameras and a LiDAR scanner, which measured the distance to a target by illuminating it with a laser light and computing the difference in return times for the reflected light. The LiDAR point clouds were used as a reference in a semi-automatic localization workflow that estimated the camera pose with six degrees of freedom. They compared this dataset with several image based localization datasets produced with the SfM algorithm [136,161,162], claiming the creation of a point cloud with much higher density and precision. EgoCart [156] is another benchmark dataset, comprising of almost 20,000 RGB-D images, annotated with information of the camera position and orientation. The authors made the dataset public, along with the evaluation (in terms of accuracy, computing time, and memory requirements) of various machine learning based localization solutions [163–166].

Schmitt et al. [157] presented an indoor localization system that relied on visual information provided by two Microsoft Kinect devices and on wheel-odometry data acquired with a Roomba robot. Within the ROS framework, they enhanced a pre-drawn floor plan with SLAM, achieving an average error of 6.7 cm for the position estimation. The authors claimed that the accuracy was sufficient to use the system as a reference, when testing the performance of other systems. Their robot could carry the components of the system under test and collect data, without interfering with the localization process.

Ibragimov and Afanasyev [158] analyzed the feasibility of using different visual SLAM based localization methods for robot systems in homogeneous indoor spaces. Their evaluation testbed was built with a monocular camera, a LiDAR sensor, a ZEDstereo camera, and a Kinect device. LIDAR based HECTORSLAM and a tape measure are considered ground truth for comparing trajectories obtained with ORB-SLAM [152], Dense Piecewise Parallel Tracking and Mapping (DPPTAM) [167], Stereolabs' ZedFu [168] 3D mapping tool, and Real-Time Appearance Based Mapping (RTAB-MAP) [169]. Filipenko and Afanasyev [159] compared various SLAM based methods integrated in the ROS framework: GMapping [170], Parallel Tracking and Mapping (PTAM) [171], HectorSLAM [172], Semi-direct Visual Odometry (SVO) [173], LSD-SLAM [174], RTAB-MAP [169], ORB-SLAM [152], DPPTAM [167], Direct Sparse Odometry (DSO) [175], Cartographer [176], and Stereo Parallel Tracking and Mapping (S-PTAM) [177]. They built a robot equipped with a 2D LiDAR, a monocular camera, and a ZED stereo camera. ATE was the chosen means of evaluation, represented through statistical metrics such as RMSE or the standard deviation. Ragot et al. [160] performed an evaluation of two Visual SLAM algorithms, ORB-SLAM2 [148] and RTAB-MAP [169,178,179]. During the comparison of the two solutions, they used a VICON motion capture system as the ground truth. They performed various experiments, with straight-line, straight-line and back, circular paths, or trajectories containing loop closure.

5. Conclusions

Indoor localization has an increasingly vast applicability in domains such as AR, navigation systems, assistive devices (especially for the visually impaired), autonomous robots, surveillance, and monitoring. Since surveillance cameras and smartphone cameras represent a commodity currently, researchers have developed a plethora of indoor localization solutions based on visual input.

This paper offers an overview of the computer vision based indoor localization domain, discussing applications areas, commercial solutions, and benchmarks and presenting some of the most relevant contributions in the area. It also provides a survey of selected positioning solutions, proposing a new classification that organizes the solutions according to the use of known environment data, the sensing devices, the type of detected elements (artificial markers or real features), and the employed localization methods. The research papers selected in the 17 classes of the proposed taxonomy were chosen from prestigious research databases based on their relevance to the domain and publication date, and their purpose was to be illustrative for the reader in terms of the indoor positioning technologies. Since many relevant papers were too recent to have a considerable number of citations, we decided to not use this criterion for selection. The focus was on providing short descriptions of the solutions, highlighting the advantages and disadvantages and presenting the achieved performances (in terms of running time and location estimation accuracy), along with the properties of the datasets used for testing.

Tables 2–18 show that many papers did not report computing times and the specifics of their testbeds, and few papers discussed the financial implications of implementing such solutions in real life. The evaluation methodologies for the presented solutions differed. While some used visual observations to test their solutions, others chose to use private datasets. Although public datasets and benchmarks exist, they come with limitations in terms of required environment data and input from cameras, limiting their use to only certain localization solutions.

We considered this paper a good guide to the field of computer vision based indoor localization. Our proposed classification and the selection of localization solutions for each category aimed to allow the reader to easily grasp the advantages and applicability of each class of solutions.

Author Contributions: Conceptualization, A.M. (Alin Moldoveanu), A.M. (Anca Morar) and I.M.; writing—original draft preparation, A.M. (Anca Morar), I.M. and I.E.R.; writing—review and editing, all authors; funding acquisition, A.M. (Alin Moldoveanu), A.M. (Anca Morar) and A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the grant "Ecosistem de cercetare, inovare si dezvoltare de produse si servicii TIC pentru o societate conectata la Internet of Things–NETIO" (MySMIS Code = 105976), subsidiary contract "Platformă imersivă de realitate virtuală și augmentată pentru spații culturale–CultReal", Number 4192/18.03.2019.

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

References

1. Yassin, A.; Nasser, Y.; Awad, M.; Al-Dubai, A.; Liu, R.; Yuen, C.; Raulefs, R. Recent Advances in Indoor Localization: A Survey on Theoretical Approaches and Applications. *IEEE Commun. Surv. Tutor.* **2016**, *19*, 1327–1346. [CrossRef]
2. Ferdous, S.; Vyas, K.; Makedon, F. A Survey on Multi Person Identification and Localization. In Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA'12), Heraklion, Greece, 6–8 June 2012.
3. Wei, W.; Tan, L.; Jin, G.; Lu, L.; Sun, C. A Survey of UAV Visual Navigation Based on Monocular SLAM. In Proceedings of the 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 14–16 December 2018.
4. Panchpor, A.A.; Shue, S.; Conrad, J.M. A survey of methods for mobile robot localization and mapping in dynamic indoor environments. In Proceedings of the 2018 Conference on Signal Processing and Communication Engineering Systems (SPACES), Vijayawada, India, 4–5 January 2018.
5. O'Mahony, N.; Campbell, S.; Carvalho, A.; Harapanahalli, S.; Velasco-Hernández, G.A.; Riordan, D.; Walsh, J. Adaptive Multimodal Localisation Techniques for Mobile Robots in Unstructured Environments: A Review. In Proceedings of the 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), Limerick, Ireland, 15–18 April 2019.

6. Desai, A.; Ghagare, N.; Donde, S. Optimal Robot Localisation Techniques for Real World Scenarios. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018.
7. Marchand, E.; Uchiyama, H.; Spindler, F. Pose Estimation for Augmented Reality: A Hands-On Survey. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 2633–2651. [CrossRef] [PubMed]
8. Sagitov, A.; Shabalina, K.; Lavrenov, R.; Magid, E. Comparing fiducial marker systems in the presence of occlusion. In Proceedings of the 2017 International Conference on Mechanical, System and Control Engineering (ICMSC), St. Petersburg, Russian, 19–21 May 2017.
9. Mendoza-Silva, G.M.; Torres-Sospedra, J.; Huerta, J. A Meta-Review of Indoor Positioning Systems. *Sensors* **2019**, *19*, 4507. [CrossRef] [PubMed]
10. Alkhawaja, F.; Jaradat, M.; Romdhane, L. Techniques of Indoor Positioning Systems (IPS): A Survey. In Proceedings of the 2019 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, UAE, 26 March–10 April 2019.
11. Zafari, F.; Gkelias, A.; Leung, K.K. A Survey of Indoor Localization Systems and Technologies. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 2568–2599. [CrossRef]
12. Lashkari, B.; Rezazadeh, J.; Farahbakhsh, R.; Sandrasegaran, K. Crowdsourcing and Sensing for Indoor Localization in IoT: A Review. *IEEE Sens. J.* **2019**, *19*, 2408–2434. [CrossRef]
13. Jang, B.; Kim, H. Indoor Positioning Technologies Without Offline Fingerprinting Map: A Survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 508–525. [CrossRef]
14. Gu, F.; Hu, X.; Ramezani, M.; Acharya, D.; Khoshelham, K.; Valaee, S.; Shang, J. Indoor Localization Improved by Spatial Context—A Survey. *ACM Comput. Surv.* **2019**, *52*. [CrossRef]
15. Birsan, J.C.R.; Moldoveanu, F.; Moldoveanu, A.; Dascalu, M.; Morar, A. Key Technologies for Indoor Positioning Systems. In Proceedings of the 2019 18th RoEduNet Conference: Networking in Education and Research (RoEduNet), Galati, Romania, 10–12 October 2019.
16. Cremers, D. Direct methods for 3D reconstruction and visual SLAM. In Proceedings of the 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), Nagoya, Japan, 8–12 May 2017.
17. Chen, Y.; Zhou, Y.; Lv, Q.; Deveerasetty, K.K. A Review of V-SLAM*. In Proceedings of the 2018 IEEE International Conference on Information and Automation (ICIA), Fujian, China, 11–13 August 2018.
18. Li, J.; Liu, Y.; Wang, J.; Yan, M.; Yao, Y. 3D Semantic Mapping Based on Convolutional Neural Networks. In Proceedings of the 2018 37th Chinese Control Conference (CCC), Wuhan, China, 25–27 July 2018.
19. Li, A.; Ruan, X.; Huang, J.; Zhu, X.; Wang, F. Review of vision based Simultaneous Localization and Mapping. In Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 15–17 March 2019.
20. Chahine, G.; Pradalier, C. Survey of Monocular SLAM Algorithms in Natural Environments. In Proceedings of the 2018 15th Conference on Computer and Robot Vision (CRV), Toronto, ON, Canada, 9–11 May 2018.
21. Zhao, B.; Hu, T.; Shen, L. Visual odometry—A review of approaches. In Proceedings of the 2015 IEEE International Conference on Information and Automation, Lijiang, China, 8–10 August 2015.
22. Saputra, M.R.U.; Markham, A.; Trigoni, N. Visual SLAM and Structure from Motion in Dynamic Environments: A Survey. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
23. Li, X.; Wang, J. Image matching techniques for vision based indoor navigation systems: Performance analysis for 3D map based approach. In Proceedings of the 2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Sydney, Australia, 13–15 November 2012.
24. Huang, G. Visual-Inertial Navigation: A Concise Review. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019.
25. Silva, C.S.; Wimalaratne, P. State-of-art-in-indoor navigation and positioning of visually impaired and blind. In Proceedings of the 2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 7–8 September 2017.
26. Singh, B.; Kapoor, M. A Survey of Current Aids for Visually Impaired Persons. In Proceedings of the 2018 3rd International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU), Bhimtal, India, 23–24 February 2018.
27. Idrees, A.; Iqbal, Z.; Ishfaq, M. An efficient indoor navigation technique to find optimal route for blinds using QR codes. In Proceedings of the 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), Auckland, New Zealand, 15–17 June 2015.

28. Fusco, G.; Coughlan, J.M. Indoor Localization Using Computer Vision and Visual-Inertial Odometry. In Proceedings of the International Conference on Computers Helping People with Special Needs, Linz, Austria, 11–13 July 2018.
29. Endo, Y.; Sato, K.; Yamashita, A.; Matsubayashi, K. Indoor positioning and obstacle detection for visually impaired navigation system based on LSD-SLAM. In Proceedings of the 2017 International Conference on Biometrics and Kansei Engineering (ICBAKE), Kyoto, Japan, 15–17 September 2017.
30. ROS. Robot Operating System. Available online: https://www.ros.org/(accessed on 3 March 2020).
31. Heya, T.A.; Arefin, S.E.; Chakrabarty, A.; Alam, M. Image Processing Based Indoor Localization System for Assisting Visually Impaired People. In Proceedings of the 2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS), Wuhan, China, 22–23 March 2018.
32. Chaccour, K.; Badr, G. Computer vision guidance system for indoor navigation of visually impaired people. In Proceedings of the 2016 IEEE 8th International Conference on Intelligent Systems (IS), Sofia, Bulgaria, 4–6 September 2016.
33. Caraiman, S.; Morar, A.; Owczarek, M.; Burlacu, A.; Rzeszotarski, D.; Botezatu, N.; Herghelegiu, P.; Moldoveanu, F.; Strumillo, P.; Moldoveanu, A. Computer Vision for the Visually Impaired: The Sound of Vision System. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017.
34. Morar, A.; Moldoveanu, F.; Petrescu, L.; Moldoveanu, A. Real Time Indoor 3D Pipeline for an Advanced Sensory Substitution Device. In Proceedings of the Image Analysis and Processing (ICIAP), Catania, Italy, 11–15 September 2017.
35. Moldoveanu, A.D.B.; Ivascu, S.; Stanica, I.; Dascalu, M.; Lupu, R.; Ivanica, G.; Balan, O.; Caraiman, S.; Ungureanu, F.; Moldoveanu, F.; et al. Mastering an advanced sensory substitution device for visually impaired through innovative virtual training. In Proceedings of the 2017 IEEE 7th International Conference on Consumer Electronics–Berlin (ICCE-Berlin), Berlin, Germany, 3–6 September 2017.
36. Babu, S.; Markose, S. IoT Enabled Robots with QR Code Based Localization. In Proceedings of the 2018 International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR), Cochin, India, 11–13 July 2018.
37. Nazemzadeh, P.; Fontanelli, D.; Macii, D.; Palopoli, L. Indoor Localization of Mobile Robots Through QR Code Detection and Dead Reckoning Data Fusion. *IEEE/ASME Trans. Mechatron.* **2017**, *22*, 2588–2599. [CrossRef]
38. Cavanini, L.; Cimini, G.; Ferracuti, F.; Freddi, A.; Ippoliti, G.; Monteriù, A.; Verdini, F. A QR-code localization system for mobile robots: Application to smart wheelchairs. In Proceedings of the 2017 European Conference on Mobile Robots (ECMR), Paris, France, 6–8 September 2017.
39. Correa, D.S.O.; Sciotti, D.F.; Prado, M.G.; Sales, D.O.; Wolf, D.F.; Osorio, F.S. Mobile Robots Navigation in Indoor Environments Using Kinect Sensor. In Proceedings of the 2012 Second Brazilian Conference on Critical Embedded Systems, Campinas, Brazil, 20–25 May 2012.
40. Xin, G.X.; Zhang, X.T.; Wang, X.; Song, J.M. A RGBD SLAM algorithm combining ORB with PROSAC for indoor mobile robot. In Proceedings of the 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), Harbin, China, 19–20 December 2015.
41. Kao, W.; Huy, B.Q. Indoor navigation with smartphone based visual SLAM and Bluetooth-connected wheel-robot. In Proceedings of the 2013 CACS International Automatic Control Conference (CACS), Nantou, Taiwan, 2–4 December 2013.
42. Moverio Website. Available online: https://moverio.epson.com (accessed on 3 March 2020).
43. Google Glass Website. Available online: https://www.google.com/glass/start (accessed on 3 March 2020).
44. Hedili, M.K.; Ulusoy, E.; Kazempour, S.; Soomro, S.; Urey, H. Next Generation Augmented Reality Displays. In Proceedings of the 2018 IEEE SENSORS, New Delhi, India, 28–31 October 2018.
45. Gerstweiler, G. Guiding People in Complex Indoor Environments Using Augmented Reality. In Proceedings of the 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Reutlingen, Germany, 18–22 March 2018.
46. Wang, C.; Chiang, D.J.; Ho, Y.Y. 3D augmented reality mobile navigation system supporting indoor positioning function. In Proceedings of the 2012 IEEE International Conference on Computational Intelligence and Cybernetics (CyberneticsCom), Bali, Indonesia, 12–14 July 2012.

47. Bálint, Z.; Kiss, B.; Magyari, B.; Simon, K. Augmented reality and image recognition based framework for treasure hunt games. In Proceedings of the 2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics, Subotica, Serbia, 20–22 September 2012.
48. Baek, F.; Ha, I.; Kim, H. Augmented reality system for facility management using image based indoor localization. *Autom. Constr.* **2019**, *99*, 18–26.10.1016/j.autcon.2018.11.034. [CrossRef]
49. Wikitude Website. Available online: https://www.wikitude.com (accessed on 3 March 2020).
50. ARKit Website. Available online: https://developer.apple.com/augmented-reality (accessed on 3 March 2020).
51. ARCore Website. Available online: https://developers.google.com/ar (accessed on 3 March 2020).
52. Vuforia Website. Available online: https://developer.vuforia.com (accessed on 3 March 2020).
53. ARToolKit Website. Available online: https://github.com/artoolkit (accessed on 3 March 2020).
54. MAXST Website. Available online: http://maxst.com (accessed on 3 March 2020).
55. EasyAR Website. Available online: https://www.easyar.com (accessed on 3 March 2020).
56. Kudan Website. Available online: https://www.kudan.io (accessed on 3 March 2020).
57. Onirix Website. Available online: https://www.onirix.com (accessed on 3 March 2020).
58. Pikkart Website. Available online: https://developer.pikkart.com/augmented-reality/sdk (accessed on 3 March 2020).
59. DeepAR Website. Available online: https://www.deepar.ai/augmented-reality-sdk (accessed on 3 March 2020).
60. Sun, Y.; Zhao, K.; Wang, J.; Li, W.; Bai, G.; Zhang, N. Device-free human localization using panoramic camera and indoor map. In Proceedings of the 2016 IEEE International Conference on Consumer Electronics-China (ICCE-China), Guangzhou, China, 19–21 December 2016.
61. Desai, P.; Rattan, K.S. Indoor localization and surveillance usingwireless sensor network and Pan/Tilt camera. In Proceedings of the IEEE 2009 National Aerospace & Electronics Conference (NAECON), Dayton, OH, USA, 21–23 July 2009.
62. Grzechca, D.; Wróbel, T.; Bielecki, P. Indoor Location and Idetification of Objects with Video Survillance System and WiFi Module. In Proceedings of the 2014 International Conference on Mathematics and Computers in Sciences and in Industry, Varna, Bulgaria, 13–15 September 2014.
63. Zhang, W.; Liu, G.; Tian, G. A Coarse to Fine Indoor Visual Localization Method Using Environmental Semantic Information. *IEEE Access* **2019**, *7*, 21963–21970. [CrossRef]
64. Shit, R.C.; Sharma, S.; Puthal, D.; James, P.; Pradhan, B.; Moorsel, A.v.; Zomaya, A.Y.; Ranjan, R. Ubiquitous Localization (UbiLoc): A Survey and Taxonomy on Device Free Localization for Smart World. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3532–3564. [CrossRef]
65. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A Survey of Deep Learning-Based Object Detection. *IEEE Access* **2019**, *7*, 128837–128868. [CrossRef]
66. Wang, J.; Olson, E. AprilTag 2: Efficient and robust fiducial detection. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016.
67. Bo Bo, N.; Deboeverie, F.; Veelaert, P.; Philips, W. Real-Time Multi-People Tracking by Greedy Likelihood Maximization. In Proceedings of the 9th International Conference on Distributed Smart Cameras (ICDSC'15), Seville, Spain, 8–11 September 2015.
68. Dias, J.; Jorge, P.M. People Tracking with Multi-Camera System. In Proceedings of the 9th International Conference on Distributed Smart Cameras (ICDSC'15), Seville, Spain, 8–11 September 2015.
69. Shim, J.H.; Cho, Y.I. A Mobile Robot Localization using External Surveillance Cameras at Indoor. *Procedia Comput. Sci.* **2015**, *56*, 502–507. [CrossRef]
70. Utasi, A.; Benedek, C. A Bayesian Approach on People Localization in Multicamera Systems. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 105–115. [CrossRef]
71. Hoyer, L.; Steup, C.; Mostaghim, S. A Robot Localization Framework Using CNNs for Object Detection and Pose Estimation. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Bengaluru, India, 18–21 November 2018.
72. Jain, M.; Nawhal, M.; Duppati, S.; Dechu, S. Mobiceil: Cost-free Indoor Localizer for Office Buildings. In Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '18), Oldenburg, Germany, 5–8 October 2018.

73. Cosma, A.; Radoi, I.E.; Radu, V. CamLoc: Pedestrian Location Estimation through Body Pose Estimation on Smart Cameras. In Proceedings of the 2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Pisa, Italy, 30 Septmeber–3 October 2019.

74. Sun, M.; Zhang, L.; Liu, Y.; Miao, X.; Ding, X. See-your-room: Indoor Localization with Camera Vision. In Proceedings of the ACM Turing Celebration Conference–China (ACM TURC '19), Chengdu, China, 17–19 May 2019.

75. Lee, S.; Tewolde, G.; Lim, J.; Kwon, J. QR-code based Localization for Indoor Mobile Robot with validation using a 3D optical tracking instrument. In Proceedings of the 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), Busan, Korea, 7–11 July 2015.

76. Goronzy, G.; Pelka, M.; Hellbrück, H. QRPos: Indoor positioning system for self-balancing robots based on QR codes. In Proceedings of the 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Madrid, Spain, 4–7 October 2016.

77. Wan, K.; Ma, L.; Tan, X. An improvement algorithm on RANSAC for image based indoor localization. In Proceedings of the 2016 International Wireless Communications and Mobile Computing Conference (IWCMC), Cyprus, Paphos, 5–9 September 2016.

78. Lightbody, P.; Krajnik, T.; Hanheide, M. An Efficient Visual Fiducial Localisation System. *SIGAPP Appl. Comput. Rev.* **2017**, *17*, 28–37. [CrossRef]

79. Ooi, Y.; Lee, W.K.; Chea, K.C. Localization of Mobile Sensor Nodes Using QR Codes and Dead Reckoning with Error Correction. In Proceedings of the 2018 21st International Conference on Electrical Machines and Systems (ICEMS), Jeju, Korea, 7–10 October 2018.

80. Benligiray, B.; Topal, C.; Akinlar, C. STag: A stable fiducial marker system. *Image Vis. Comput.* **2019**, *89*, 158–169. [CrossRef]

81. Khan, D.; Ullah, S.; Nabi, S. A Generic Approach toward Indoor Navigation and Pathfinding with Robust Marker Tracking. *Remote. Sens.* **2019**, *11*, 3052. [CrossRef]

82. Li, Z.; Huang, J. Study on the use of Q-R codes as landmarks for indoor positioning: Preliminary results. In Proceedings of the 2018 IEEE/ION Position, Location and Navigation Symposium (PLANS), Monterey, CA, USA, 23–26 April 2018.

83. Dutta, V. Mobile Robot Applied to QR Landmark Localization Based on the Keystone Effect. In *Mechatronics and Robotics Engineering for Advanced and Intelligent Manufacturing*; Springer: Cham, Switzerland, 2017; pp. 45–60.

84. Gang, H.S.; Pyun, J.Y. A Smartphone Indoor Positioning System Using Hybrid Localization Technology. *Energies* **2019**, *12*, 3702. [CrossRef]

85. Hu, F.; Zhu, Z.; Zhang, J. Mobile Panoramic Vision for Assisting the Blind via Indexing and Localization. In Proceedings of the ECCV Workshops, Zurich, Switzerland, 6–7 September 2014.

86. Bai, Y.; Jia, W.; Zhang, H.; Mao, Z.; Sun, M. Landmark based indoor positioning for visually impaired individuals. In Proceedings of the 2014 12th International Conference on Signal Processing (ICSP), Hangzhou, China, 19–23 October 2014.

87. Elloumi, W.; Guissous, K.; Chetouani, A.; Treuillet, S. Improving a vision indoor localization system by a saliency-guided detection. In Proceedings of the 2014 IEEE Visual Communications and Image Processing Conference, Valletta, Malta, 7–10 December 2014.

88. Rivera-Rubio, J.; Alexiou, I.; Bharath, A.A. Appearance based indoor localization: A comparison of patch descriptor performance. *Pattern Recognit. Lett.* **2015**, *66*, 109–117. [CrossRef]

89. Lu, G.; Yan, Y.; Sebe, N.; Kambhamettu, C. Indoor Localization via Multi-view Images and Videos. *Comput. Vis. Image Underst.* **2017**, *161*, 145–160. [CrossRef]

90. Xiao, A.; Chen, R.; Li, D.; Chen, Y.; Wu, D. An Indoor Positioning System Based on Static Objects in Large Indoor Scenes by Using Smartphone Cameras. *Sensors* **2018**, *18*, 2229. [CrossRef] [PubMed]

91. Akal, O.; Mukherjee, T.; Barbu, A.; Paquet, J.; George, K.; Pasiliao, E. A Distributed Sensing Approach for Single Platform Image-Based Localization. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018.

92. Guo, F.; He, Y.; Guan, L. RGB-D camera pose estimation using deep neural network. In Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, 14–16 November 2017.

93. Marouane, C.; Maier, M.; Feld, S.; Werner, M. Visual positioning systems—An extension to MoVIPS. In Proceedings of the 2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Busan, Korea, 27–30 October 2014.
94. Yan, X.; Liu, W.; Cui, X. Research and Application of Indoor Guide Based on Mobile Augmented Reality System. In Proceedings of the 2015 International Conference on Virtual Reality and Visualization (ICVRV), Xiamen, China, 17–18 October 2015.
95. Huang, Z.; Gu, N.; Hao, J.; Shen, J. 3DLoc: 3D Features for Accurate Indoor Positioning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *1*, 141:1–141:26. [CrossRef]
96. Árvai, L.; Dobos, G. On demand vison based indoor localization. In Proceedings of the 2019 20th International Carpathian Control Conference (ICCC), Krakow-Wieliczka, Poland, 26–29 May 2019.
97. Rituerto, A.; Fusco, G.; Coughlan, J.M. Towards a Sign-Based Indoor Navigation System for People with Visual Impairments. In Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '16), Reno, NV, USA, 24–26 October 2016.
98. Neges, M.; Koch, C.; König, M.; Abramovici, M. Combining visual natural markers and IMU for improved AR based indoor navigation. *Adv. Eng. Inform.* **2017**, *31*, 18–31. [CrossRef]
99. Sun, Y.; Meng, W.; Li, C.; Zhao, N.; Zhao, K.; Zhang, N. Human Localization Using Multi-Source Heterogeneous Data in Indoor Environments. *IEEE Access* **2017**, *5*, 812–822. [CrossRef]
100. Guo, J.; Zhang, S.; Zhao, W.; Peng, J. Fusion of Wifi and Vision Based on Smart Devices for Indoor Localization. In Proceedings of the 16th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry (VRCAI '18), Hachioji, Japan, 2–3 December 2018.
101. Zhao, B.; Zhu, D.; Xi, T.; Jia, C.; Jiang, S.; Wang, S. Convolutional neural network and dual-factor enhanced variational Bayes adaptive Kalman filter based indoor localization with Wi-Fi. *Comput. Netw.* **2019**, *162*, 106864. [CrossRef]
102. Gao, M.; Yu, M.; Guo, H.; Xu, Y. Mobile Robot Indoor Positioning Based on a Combination of Visual and Inertial Sensors. *Sensors* **2019**, *19*, 1773. [CrossRef]
103. Kim, D.H.; Han, S.B.; Kim, J.H. Visual Odometry Algorithm Using an RGB-D Sensor and IMU in a Highly Dynamic Environment. In *Robot Intelligence Technology and Applications 3*; Kim, J.H., Yang, W., Jo, J., Sincak, P., Myung, H., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 11–26.
104. Teixeira, L.; Raposo, A.B.; Gattass, M. Indoor Localization Using SLAM in Parallel with a Natural Marker Detector. In Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC '13), Coimbra, Portugal, 18–22 March 2013.
105. Sinha, D.; Ahmed, M.T.; Greenspan, M. Image Retrieval Using Landmark Indexing for Indoor Navigation. In Proceedings of the 2014 Canadian Conference on Computer and Robot Vision, Montreal, QC, Canada, 6–9 May 2014.
106. Deretey, E.; Ahmed, M.T.; Marshall, J.A.; Greenspan, M. Visual indoor positioning with a single camera using PnP. In Proceedings of the 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Calgary, AB, Canada, 13–16 October 2015.
107. Ruotsalainen, L.; Gröhn, S.; Kirkko-Jaakkola, M.; Chen, L.; Guinness, R.; Kuusniemi, H. Monocular visual SLAM for tactical situational awareness. In Proceedings of the 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Calgary, AB, Canada, 13–16 October 2015.
108. Zhou, H.; Zou, D.; Pei, L.; Ying, R.; Liu, P.; Yu, W. StructSLAM: Visual SLAM With Building Structure Lines. *IEEE Trans. Veh. Technol.* **2015**, *64*, 1364–1375. [CrossRef]
109. Zhao, L.; Fan, Z.; Li, W.; Xie, H.; Xiao, Y. 3D Indoor Map Building with Monte Carlo Localization in 2D Map. In Proceedings of the 2016 International Conference on Industrial Informatics–Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), Wuhan, China, 3–4 December 2016.
110. Ramesh, K.; Nagananda, S.N.; Ramasangu, H.; Deshpande, R. Real-time localization and navigation in an indoor environment using monocular camera for visually impaired. In Proceedings of the 2018 5th International Conference on Industrial Engineering and Applications (ICIEA), Singapore, 26–28 April 2018.
111. Dong, E.; Xu, J.; Wu, C.; Liu, Y.; Yang, Z. Pair-Navi: Peer-to-Peer Indoor Navigation with Mobile Visual SLAM. In Proceedings of the IEEE INFOCOM 2019–IEEE Conference on Computer Communications, Paris, France, 29 April–2 May 2019.

112. Han, S.; Ahmed, M.U.; Rhee, P.K. Monocular SLAM and Obstacle Removal for Indoor Navigation. In Proceedings of the 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, Australia, 3–7 December 2018.
113. Xiao, L.; Wang, J.; Qiu, X.; Rong, Z.; Zou, X. Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment. *Robot. Auton. Syst.* **2019**, *117*, 1–16. [CrossRef]
114. 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 (UbiComp '11), Beijing, China, 17–21 September 2011.
115. Paton, M.; Kosecka, J. Adaptive RGB-D Localization. In Proceedings of the 2012 Ninth Conference on Computer and Robot Vision, Toronto, ON, Canada, 28–30 May 2012.
116. Salas-Moreno, R.F.; Newcombe, R.A.; Strasdat, H.; Kelly, P.H.J.; Davison, A.J. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
117. Albrecht, A.; Heide, N. Mapping and Automatic Post-Processing of Indoor Environments by Extending Visual SLAM. In Proceedings of the International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, China, 16–17 July 2018.
118. Tang, S.; Li, Y.; Yuan, Z.; Li, X.; Guo, R.; Zhang, Y.; Wang, W. A Vertex-to-Edge Weighted Closed-Form Method for Dense RGB-D Indoor SLAM. *IEEE Access* **2019**, *7*, 32019–32029. [CrossRef]
119. Martín, F.; Matellán, V.; Rodríguez, F.J.; Ginés, J. Octree based localization using RGB-D data for indoor robots. *Eng. Appl. Artif. Intell.* **2019**, *77*, 177–185. [CrossRef]
120. Kuang, H.; Wang, X.; Liu, X.; Ma, X.; Li, R. An Improved Robot's Localization and Mapping Method Based on ORB-SLAM. In Proceedings of the 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 9–10 December 2017.
121. Guclu, O.; Can, A.B. k-SLAM: A fast RGB-D SLAM approach for large indoor environments. *Comput. Vis. Image Underst.* **2019**, *184*, 31–44. [CrossRef]
122. Yun, D.; Chang, H.; Lakshman, T.V. Accelerating Vision based 3D Indoor Localization by Distributing Image Processing over Space and Time. In Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology (VRST '14), Edinburgh, UK, 11–13 November 2014.
123. Huang, C.; Lin, C.; Shih, S.; Chang, P.; Lin, Y.; Huang, C. Indoor environmental data collection, localization and fusion. In Proceedings of the 2017 IEEE International Conference on Consumer Electronics–Taiwan (ICCE-TW), Taibei, Taiwan, 12–14 June 2017.
124. Chan, S.; Wu, P.; Fu, L. Robust 2D Indoor Localization Through Laser SLAM and Visual SLAM Fusion. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 7–8 October 2018.
125. Ullah, S.; Song, B.; Chen, W. EMoVI-SLAM: Embedded Monocular Visual Inertial SLAM with Scale Update for Large Scale Mapping and Localization. In Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Nanjing, China, 27–31 August 2018.
126. Ferryman, J.; Shahrokni, A. PETS2009: Dataset and challenge. In Proceedings of the 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Snowbird, UT, USA, 7–12 December 2009.
127. Andriluka, M.; Roth, S.; Schiele, B. Monocular 3D pose estimation and tracking by detection. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 July 2010.
128. OpenPose Website. Available online: https://github.com/CMU-Perceptual-Computing-Lab/openpose (accessed on 3 March 2020).
129. Dietterich, T.G.; Bakiri, G. Error-Correcting Output Codes: A General Method for Improving Multiclass Inductive Learning Programs. In Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI'91), Anaheim, CA, USA, 14–19 July 1991.
130. Krajník, T.; Nitsche, M.; Faigl, J.; Vaněk, P.; Saska, M.; Přeučil, L.; Duckett, T.; Mejail, M. A Practical Multirobot Localization System. *J. Intell. Robot. Syst.* **2014**, *76*, 539–562. [CrossRef]
131. Yang, S.; Scherer, S.A.; Zell, A. An Onboard Monocular Vision System for Autonomous Takeoff, Hovering and Landing of a Micro Aerial Vehicle. *J. Intell. Robot. Syst.* **2013**, *69*, 499–515. [CrossRef]

132. Garrido-Jurado, S.; Muñoz-Salinas, R.; Madrid-Cuevas, F.J.; Marín-Jiménez, M.J. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognit.* **2014**, *47*, 2280–2292. [CrossRef]
133. Bergamasco, F.; Albarelli, A.; Cosmo, L.; Rodolà, E.; Torsello, A. An Accurate and Robust Artificial Marker Based on Cyclic Codes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2359–2373. [CrossRef]
134. Mikolajczyk, K.; Schmid, C. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1615–1630. [CrossRef]
135. Zbar Website. Available online: http://zbar.sourceforge.net/about.html (accessed on 3 March 2020).
136. Li, Y.; Snavely, N.; Huttenlocher, D.P. Location Recognition Using Prioritized Feature Matching. In Proceedings of the ECCV, Heraklion, Greece, 5–11 September 2010.
137. Triggs, B.; McLauchlan, P.F.; Hartley, R.I.; Fitzgibbon, A.W. Bundle Adjustment—A Modern Synthesis. In *Vision Algorithms: Theory and Practice*; Triggs, B., Zisserman, A., Szeliski, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2000. pp. 298–372.
138. Handa, A.; Whelan, T.; McDonald, J.; Davison, A.J. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–5 June 2014.
139. 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, Vilamoura, Portugal, 7–12 October 2012.
140. Metaio SDK Website. Available online: https://metaio-sdk.software.informer.com/5.5 (accessed on 3 March 2020).
141. Raguram, R.; Frahm, J.M.; Pollefeys, M. A Comparative Analysis of RANSAC Techniques Leading to Adaptive Real-Time Random Sample Consensus. In Proceedings of the Computer Vision (ECCV 2008), Marseille, France, 12–18 October 2008.
142. Davison, A. Real-time simultaneous localisation and mapping with a single camera. *Proc. Int. Conf. Comput. Vis.* **2003**, *2*, 1403–1410. [CrossRef]
143. Doucet, A.; de Freitas, N.; Murphy, K.; Russell, S. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In *Uncertainty in Artificial Intelligence*; Springer: New York, NY, USA, 2000.
144. Liang, J.Z.; Corso, N.; Turner, E.; Zakhor, A. Reduced-complexity data acquisition system for image based localization in indoor environments. In Proceedings of the International Conference on Indoor Positioning and Indoor Navigation, Montbeliard-Belfort, France, 28–31 October 2013.
145. Bonarini, A.; Burgard, W.; Fontana, G.; Matteucci, M.; Sorrenti, D.; Tardos, J. RAWSEEDS: Robotics Advancement through Web-publishing of Sensorial and Elaborated Extensive Data Sets. In Proceedings of the IROS Workshop Benchmarks Robot, Beijing, China, 9–15 October 2006.
146. Endres, F.; Hess, J.; Sturm, J.; Cremers, D.; Burgard, W. 3-D Mapping With an RGB-D Camera. *IEEE Trans. Robot.* **2014**, *30*, 177–187. [CrossRef]
147. Geiger, A. Karlsruhe Dataset: Stereo Video Sequences + rough GPS Poses. Available online: http://www.cvlibs.net/datasets/karlsruhe_sequences/ (accessed on 3 March 2020).
148. Mur-Artal, R.; Tardós, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]
149. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
150. Huang, J.; Kumar, S.R.; Mitra, M.; Zhu, W.J.; Zabih, R. Image indexing using color correlograms. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 17–19 June 1997.
151. Cheng, M.; Zhang, Z.; Lin, W.; Torr, P. BING: Binarized Normed Gradients for Objectness Estimation at 300fps. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June, 2014.
152. Mur-Artal, R.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [CrossRef]
153. Gerstweiler, G.; Vonach, E.; Kaufmann, H. HyMoTrack: A Mobile AR Navigation System for Complex Indoor Environments. *Sensors* **2016**, *16*, 17. [CrossRef] [PubMed]

154. Burri, M.; Nikolic, J.; Gohl, P.; Schneider, T.; Rehder, J.; Omari, S.; Achtelik, M.W.; Siegwart, R. The EuRoC micro aerial vehicle datasets. *Int. J. Robot. Res.* **2016**. [CrossRef]
155. Sun, X.; Xie, Y.; Luo, P.; Wang, L. A Dataset for Benchmarking Image-Based Localization. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
156. Spera, E.; Furnari, A.; Battiato, S.; Farinella, G.M. EgoCart: A Benchmark Dataset for Large-Scale Indoor Image-Based Localization in Retail Stores. *IEEE Trans. Circuits Syst. Video Technol.* **2019**. [CrossRef]
157. Schmitt, S.; Will, H.; Aschenbrenner, B.; Hillebrandt, T.; Kyas, M. A reference system for indoor localization testbeds. In Proceedings of the 2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Sydney, Australia, 13–15 November 2012.
158. Ibragimov, I.Z.; Afanasyev, I.M. Comparison of ROS based visual SLAM methods in homogeneous indoor environment. In Proceedings of the 2017 14th Workshop on Positioning, Navigation and Communications (WPNC), Bremen, Germany, 25–26 October 2017.
159. Filipenko, M.; Afanasyev, I. Comparison of Various SLAM Systems for Mobile Robot in an Indoor Environment. In Proceedings of the 2018 International Conference on Intelligent Systems (IS), Madeira, Portugal, 25–27 September 2018.
160. Ragot, N.; Khemmar, R.; Pokala, A.; Rossi, R.; Ertaud, J. Benchmark of Visual SLAM Algorithms: ORB-SLAM2 vs RTAB-Map*. In Proceedings of the 2019 Eighth International Conference on Emerging Security Technologies (EST), Essex, UK, 22–24 July 2019.
161. Irschara, A.; Zach, C.; Frahm, J.; Bischof, H. From structure-from-motion point clouds to fast location recognition. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
162. Sattler, T.; Weyand, T.; Leibe, B.; Kobbelt, L. Image Retrieval for Image-Based Localization Revisited. In Proceedings of the BMVC, Surrey, UK, 3–7 September 2012.
163. Sánchez, J.; Perronnin, F.; de Campos, T.E. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognit. Lett.* **2012**, *33*, 2216–2223. [CrossRef]
164. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
165. Kendall, A.; Grimes, M.; Cipolla, R. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
166. Kendall, A.; Cipolla, R. Geometric Loss Functions for Camera Pose Regression with Deep Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
167. Concha, A.; Civera, J. DPPTAM: Dense piecewise planar tracking and mapping from a monocular sequence. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015.
168. ZedFu Website. Available online: https://www.stereolabs.com/blog/positional-tracking-3d-reconstruction-and-more-with-zed-camera (accessed on 3 March 2020).
169. Labbé, M.; Michaud, F. Online global loop closure detection for large-scale multi-session graph based SLAM. In Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014.
170. Grisetti, G.; Stachniss, C.; Burgard, W. Improved Techniques for Grid Mapping With Rao-Blackwellized Particle Filters. *IEEE Trans. Robot.* **2007**, *23*, 34–46. [CrossRef]
171. Klein, G.; Murray, D. Parallel Tracking and Mapping for Small AR Workspaces. In Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Nara, Japan, 13–16 November 2007.
172. Kohlbrecher, S.; Stryk, O.v.; Meyer, J.; Klingauf, U. A flexible and scalable SLAM system with full 3D motion estimation. In Proceedings of the 2011 IEEE International Symposium on Safety, Security, and Rescue Robotics, Kyoto, Japan, 31 October–5 November 2011.

173. Forster, C.; Pizzoli, M.; Scaramuzza, D. SVO: Fast semi-direct monocular visual odometry. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–5 June 2014.

174. Engel, J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-Scale Direct Monocular SLAM. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.

175. Engel, J.; Koltun, V.; Cremers, D. Direct Sparse Odometry. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 611–625. [CrossRef]

176. Hess, W.; Kohler, D.; Rapp, H.; Andor, D. Real-time loop closure in 2D LIDAR SLAM. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016.

177. Pire, T.; Fischer, T.; Castro, G.; De Cristóforis, P.; Civera, J.; Jacobo Berlles, J. S-PTAM: Stereo Parallel Tracking and Mapping. *Robot. Auton. Syst.* **2017**, *93*, 27–42. [CrossRef]

178. Labbé, M.; Michaud, F. Memory management for real-time appearance based loop closure detection. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 25–30 September 2011.

179. Labbé, M.; Michaud, F. Appearance-Based Loop Closure Detection for Online Large-Scale and Long-Term Operation. *IEEE Trans. Robot.* **2013**, *29*, 734–745. [CrossRef]