sensors

MDPI

Article

An Indoor Positioning System Based on Static Objects in Large Indoor Scenes by Using Smartphone Cameras

Aoran Xiao [1], Ruizhi Chen [1,2,*], Deren Li [1,2,*], Yujin Chen [3] and Dewen Wu [1,2]

1 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; xiaoaoran@whu.edu.cn (A.X.); wudewen@whu.edu.cn (D.W.)
2 Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China
3 School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China; yujin.chen@whu.edu.cn
* Correspondence: ruizhi.chen@whu.edu.cn (R.C.); drli@whu.edu.cn (D.L.)

Received: 4 June 2018; Accepted: 9 July 2018; Published: 11 July 2018

Abstract: The demand for location-based services (LBS) in large indoor spaces, such as airports, shopping malls, museums and libraries, has been increasing in recent years. However, there is still no fully applicable solution for indoor positioning and navigation like Global Navigation Satellite System (GNSS) solutions in outdoor environments. Positioning in indoor scenes by using smartphone cameras has its own advantages: no additional needed infrastructure, low cost and a large potential market due to the popularity of smartphones, etc. However, existing methods or systems based on smartphone cameras and visual algorithms have their own limitations when implemented in relatively large indoor spaces. To deal with this problem, we designed an indoor positioning system to locate users in large indoor scenes. The system uses common static objects as references, e.g., doors and windows, to locate users. By using smartphone cameras, our proposed system is able to detect static objects in large indoor spaces and then calculate the smartphones' position to locate users. The system integrates algorithms of deep learning and computer vision. Its cost is low because it does not require additional infrastructure. Experiments in an art museum with a complicated visual environment suggest that this method is able to achieve positioning accuracy within 1 m.

Keywords: indoor positioning; smartphone; large indoor scene; computer vision; deep learning

1. Introduction

It seems obvious for us to conclude that human beings as well as most of animals locate themselves by visual perception. According to our own experiences, people observe the environment surrounding them intentionally or unintentionally, and "draw" a rough map in their mind. The winners of The Nobel Prize in Physiology or Medicine 2014 proved this idea: O'Keefe et al. [1,2] discovered that some "place cells", which are a type of nerve cells in particular area of the brain, were always activated if a rat was located at a particular place in a room. Other place cells were also activated if the rat moved to different places. O'Keefe concluded that the room map in the mind is formed by these place cells. In 2005, May-Britt and Edvard Moser [3] made a further discovery. They found another important component of location system in brain. Another type of nerve cell, which they named "grid cells", that can create a coordinate system to realize precise positioning and find accurate paths. Their subsequent research [4] indicated how place cells and grid cells are able to determine position and allow one to navigate.

Inspired by this interesting discovery, we propose an indoor positioning system. We used smartphone cameras as data acquisition devices to capture visual information for location. Considering

that specific indoor surroundings, including objects and layouts, activate place cells in the rat's brain, we decided to use particular objects in the rooms for positioning. In an application, objects used as positioning references should be common and immovable, such as doors and windows. These objects are widely distributed (many of which are necessary for indoor scenes) and remain still and well-preserved for a long time. As their locations are known in plan maps, they become ideal reference sources for positioning. In this paper, we called them "static objects". Besides, large indoor scenes such as libraries, airports and museums, provide a wide and stable vision condition for visual algorithms of the system. Traditional vision-based indoor positioning methods have their own limits when implementing in this kind of relatively large indoor environments (See Section 5.2. Evaluation). Thus, we decided to use the system to locate people in large indoor scenes.

In our previous research, Wu. et al. [5] compared the positioning accuracy between human brains and a visual algorithm was proposed, which proved that the indoor visual positioning method via smartphone outperforms human brains. They experimented with the positioning algorithm in several places including libraries and an office to prove the robustness. In this paper, we improved the method to a whole system, realizing positioning in large indoor spaces in our application.

The system proposed in this paper aims to locate smartphones (users) via static objects. Static objects in image are detected and identified firstly. Then the positions of users are calculated as output. The main contribution of our work can be summarized as follows:

(1) We propose an indoor positioning system by using smartphone cameras, which is designed for large indoor scenes. Previous studies of indoor positioning based on smartphone cameras have their own shortcomings in such large indoor scenes. The system integrates computer vision (CV) and deep learning (DL) algorithms. Common static objects (such as doors and windows) in the indoor scene are used as references for locating purposes, making our method general, and easy to replicate.
(2) We tested our system in a large indoor space with a complicated field of vision—an art museum. Experiments indicated that our method is able to achieve a positioning accuracy within 1 m in such circumstances.
(3) Our method is low-cost, as developers only need to take several photos to the static objects as a sample collection, without any additional infrastructure. It is also easily operated using monocular photography, which means users don't have to photograph scenes from multiple angles or take a video.

The rest of the paper is organized as follows: Section 2 reviews related works. Details about the method are demonstrated in Section 3. Experiments with performance evaluation are presented in Section 4. Section 5 is the discussion and Section 6 is the conclusions.

2. Related Works

Despite of the increasing demand for indoor location-based services, there is still no persuasive solution for indoor location and navigation like Global Navigation Satellite System (GNSS) for outdoor environments, as GNSS signals are too weak to penetrate into walls. Also, the complex spatial topology and RF transmission make localization in indoor environments very complicated.

Recent years have witnessed many studies on indoor positioning, especially by means of smartphones for their widespread use and development. Equipped with various sensors and supporting rich RF signals, smartphones hence can be located by various means, which can be divided into three categories [6]: (1) GNSS signal receiver, including GPS, BDS, GLONASS and Galileo; (2) Built-in sensors in smartphones, such as accelerometers, gyroscopes, magnetometers, barometers, lighter sensors, microphones, loudspeakers and cameras, etc.; (3) RF signals like Wi-Fi, Bluetooth and cellular wireless communication signal, etc. Except for the GNSS signal receiver, all of other sensors and RF signals are not designed for positioning purposes, but they can be used to calculate indoor locations via different principles and algorithms. Among them, the establishment of

fingerprinting of Wi-Fi [7–11], geomagnetism [12] or Bluetooth [13–16] are popular approaches due to their effectiveness and independence from infrastructure. These methods can reach positioning accuracies of 2~5 m but are easily interfered by changing environments and nearby human bodies. Besides, the fingerprinting database has to be updated every few months, which is inconvenient for developers. He et al. [17] combined the strengths of trilateration and fingerprinting to form an accurate and effective indoor localization framework. The Bluetooth antenna array [18] can achieve a higher location accuracy but has limitations such as high cost and short working distance. Cellular technology [19,20] has great potential in indoor positioning because cellular signals are widespread, but the positioning error of these kind of methods is relatively large. Infrared technology [21] and ultrasonic waves [22,23] can achieve higher indoor positioning accuracy, with the limitation of needing additional infrastructure. Since all of these methods have their own advantages and disadvantages, multi-sensor fusion approaches have been researched by many people to take advantage of different methods. Kim et al. [24] combined magnetic signals, Wi-Fi signals and cellular signals to realize indoor location. Jeon et al. [25] proposed a method integrating Bluetooth RSSI with an accelerometer and a barometer on smartphones to reach higher positioning accuracy compared with the approach without Bluetooth RSSI. Chen et al. [26] integrated a typical Wi-Fi indoor positioning system with a PDR system and achieved a better positioning performance than the PDR system or Wi-Fi positioning system alone. Li et al. [27] proposed a dead-reckoning (DR)/Wi-Fi fingerprinting/magnetic matching (MM) integration structure. The structure uses consumers' portable devices with off-the-shelf sensors and existing Wi-Fi infrastructures to raise positioning accuracy. Liu et al. [28] fused multiple sensors, including cameras, Wi-Fi and inertial sensors, and used a deep learning method to realize indoor localization. Becker et al. [29] used vision images for classification to recognize corridors, with existing wireless LAN access points in corresponding corridor to realize positioning. Gao et al. [30,31] designed a method which combined camera and gyroscope in smartphone to realize indoor positioning with accuracy of 2 m–8 m. In their method, at least three smartphone photos are photographed in each test place to capture references points (such as store logos) that provide positioning information, and gyroscope records angle information at the same time. After that, the position can be calculated via trilateration.

In addition, the vision-based indoor positioning problem has always been a hotspot issue of research in the last decades. In 1998, the vision technology group of Microsoft discussed an easy-living life in the future [32], and one of the core technologies was to locate people in house by using several video surveillance techniques. After that, methods of indoor positioning based on vision kept developing, and they can be roughly divided into three categories [33]. The first category uses references from building models. These kinds of methods detect objects in images and match them with those in the building database. Hile and Borriello [34] compared the images with the floor plan of the building to locate smartphones. Kohoutek et al. [35] detected specific objects in cloud points obtained by a range imaging camera, and compared them with the digital spatial-semantic interior building model CityGML to determine location and orientation. The second category of visual positioning methods are based on images. These approaches mainly compare similarities among testing images and reference images captured in offline phase and output the location of the reference image with the highest score. Kim and Jun [36] matched the current view of a camera with image sequences stored in a database. Their method is designed for augmented reality applications for extra information. Werner et al. [37] estimated position and orientation by using reference images and location information acquired from the pre-built database. They designed an algorithm to estimate distance through the ratio of matched pixel distance to measure viewpoint-to-image distance. Möller et al. [38] designed an mobile indoor navigation system combined interfaces of Virtual Reality (VR) and Augmented Reality (AR) elements. The last category is utilizing deployed coded targets, including concentric rings, barcodes or patterns consisting of colored dots, etc. Mulloni et al. [39] pasted the barcode in different places, so cameras can capture these marks to get location as well as other information. ByteLight company created a special LED light with specific frequency (represent different position information), which can be

captured by cameras instead of human eyes [40]. In addition to these systems and methods, visual gyro [41] and visual odometer technology [42] are also used as visual positioning. The algorithm of vision positioning is more complex, larger computation and higher power consumption than other methods. However, with further improvement of smartphone performance, this kind of methods is expected to further popularize in the future.

3. System and Methodology

In this section, a system overview is presented at first. Then key modules are illustrated with process of the system.

3.1. System Overview

The main idea of our system is to use smartphone images to locate users via static objects in large indoor scenes, and the system flow diagram is shown as Figure 1. The proposed indoor positioning system consists of two parts: static objects recognition and position calculation. The static objects recognition aims to detect and identify static objects in images, and then determine coordinates of control points for calculating users' location (Section 3.1). The position calculation includes position estimation, distance estimation and a filter screening gross points and output users' position (Section 3.2).

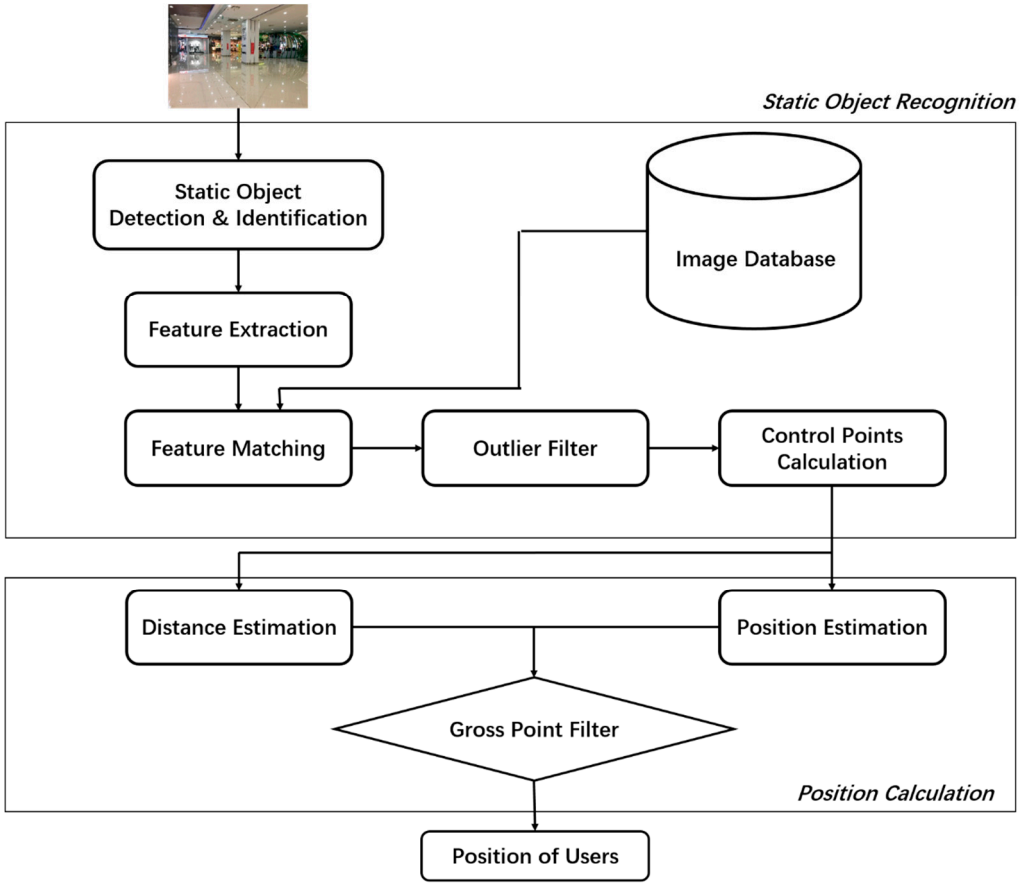

Figure 1. System flow diagram.

The current version of our system is web-based. After the smartphone photographs test images, a desktop as server will implement the rest of algorithms and return the results.

3.2. Static Objects Recognition

3.2.1. Static Object Detection & Identification

When users take a photo as input, the first task of the system is to detect static objects in images and recognize their unique identities. This is a key module of the system, outputting boundaries and identities of static objects in images. The boundaries of static objects in image influence the performance of feature extraction and matching in the following procedures, and identities of static objects are the key to find corresponding attributes in database, such as room number, pixel coordinates and objects coordinates of control points, etc.

In this paper, we implement Faster-RCNN algorithm [43] for this task. Faster-RCNN integrates region proposal, feature extraction, classification and rectangle-refine into one end-to-end network, which greatly reduce the amount of calculation and speed up the detection process. Smartphone images are firstly zoomed into a fixed size, then those fixed-size images are send to the network. Just as Figure 2 illustrates, at the beginning of the network there are 13 conv layers, 13 relu layers and four pooling layers. This combination of different layers is actually a part of VGG16 network [44], which is a famous network in image classification, realizing feature extraction for smartphone images; Then a Region Proposal Network (RPN) generates foreground anchors as well as bounding box regression bias to calculate proposals from these features; ROI pooling layers use proposals to extract proposal feature for subsequent fully convolutional network and softmax network to classify proposals. The whole network is trained on the basis of the convergence of loss function as below:

$$\text{Loss} = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (1)$$

Here, i is the index of anchor; p represents prediction probability for classification of foreground anchor (i.e., static object) and t is the outer rectangle of predicted target. p^* and t^* represent the corresponding ground truth of p and t respectively. N_{cls} and N_{reg} are numbers of anchors and outer rectangles. $L_{cls}(\cdot)$ and $L_{reg}(\cdot)$ run subtraction.

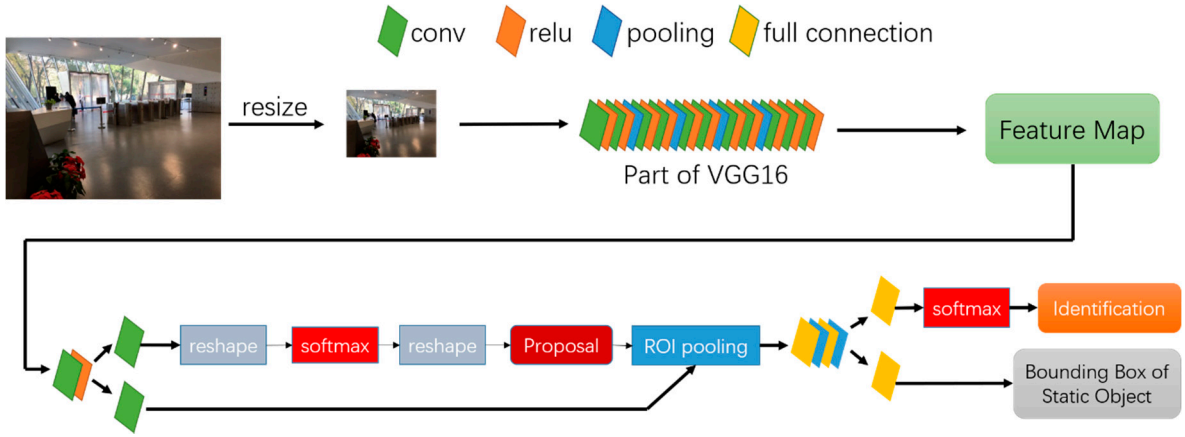

Figure 2. Faster-RCNN network for static objects detection and identification.

In order to improve performance and robustness of the system, the whole network shall be retrained in offline phase. Photos of static objects photographed from various angles at different distances are used for training images. After training customized network, the system outputs outer rectangles and identities of static objects appeared in images (as Figure 3).

Figure 3. An example of output (right) in process of Static Object Detection & identification.

3.2.2. Obtaining Control Points Coordinates

We define "control points" as those physical feature points on static objects with accurately surveyed coordinate location and can be identified relatively easy. By building relationship between pixel coordinates in image and corresponding space coordinates of control points (Collinear Equation Model [45]), the position of the smartphone can then be obtained. Thus, the key problem is to find the corresponding pixel coordinate of control points in test images. Here is our strategy: In the offline phase, images of static objects are photographed and stored in dataset, called the "reference images". Pixel coordinates of control points in these images are measured and recorded. In the online phase, after detecting and identifying static objects in test images, feature points in testing image and corresponding reference images are extracted. Then the feature matching algorithm is implemented to get enough homonymy feature points, which is used to calculate homographic matrix in next step. The homographic matrix represents the mapping relationship between pixels of testing image and reference image. Finally, the pixel coordinates of control points in testing image can be calculated from the homographic matrix and reference images coordinates of control points. The details of the algorithm are showed in Algorithm 1.

Algorithm 1. Obtaining Pixel Coordinates of Control Points in Test Images

Input: image block of static objects from test image

Procedure:

(1) Get reference image through identity of static object from database;

(2) Extract feature points for both test image block and reference image by SIFT operator [46];

(3) Perform feature matching to get homonymy feature point pairs;

(4) Employed RANSAC [47] to remove false matching points; the remaining matching points marked as P_{test} for test image and P_{ref} for reference image;

(5) Calculate homographic matrix H_{homo} by solving formula below:

$$P_{Test} = H_{homo} \times P_{ref}$$

(6) Estimate pixel coordinates of control points in test images CPT as following formula, CPT_{ref} is the set of pixel coordinates of control points in reference images:

$$CPT = H_{homo} \times CPT_{ref}$$

Output: CPT

Figure 4 shows an example of output by Algorithm 1. The pixel coordinates of control points in reference images are measured in the offline phase. The reason why we do not directly choose feature points from test image as output is that the specific control points may not belong to the set of feature points by SIFT when the texture of images are too complicated. Also, it is hard to design a robust and effective filter to screen the specific point from plenty of feature points. Algorithm 1 is a fast and effective approach instead.

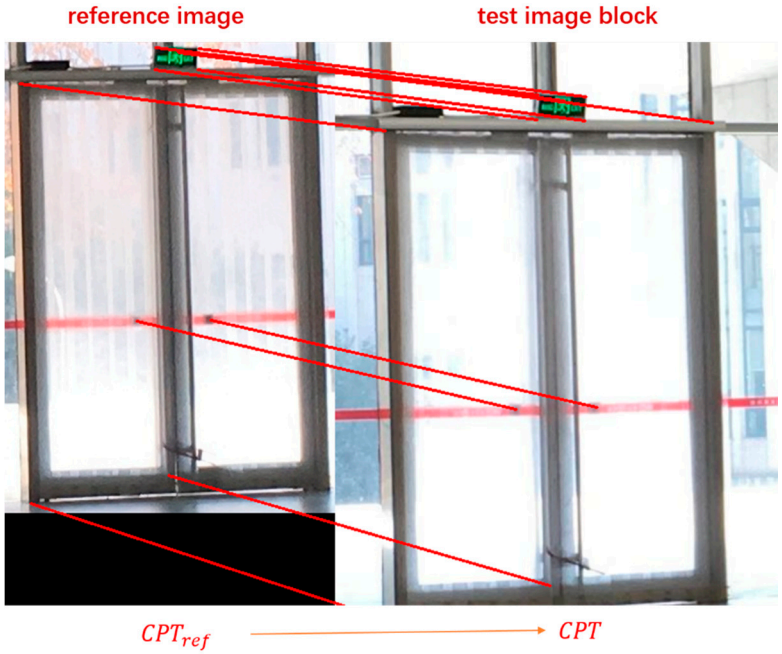

Figure 4. An example of output by Algorithm 1. The pixel coordinates of control points in test image are obtained from reference image.

3.3. Position Calculation

3.3.1. Position Estimation

The geometric relation between control points in image and object space can be illustrated via collinear equation model as Equation (2), for pixel coordinate (x, y) and space coordinate (X, Y, Z) of the same control point, the geometric relation can be illustrated as below:

$$\begin{cases} x - x_0 = -f\dfrac{t_{11}(X - X_O) + t_{12}(Y - Y_O) + t_{13}(Z - Z_O)}{t_{31}(X - X_O) + t_{32}(Y - Y_O) + t_{33}(Z - Z_O)} \\ y - y_0 = -f\dfrac{t_{21}(X - X_O) + t_{22}(Y - Y_O) + t_{23}(Z - Z_O)}{t_{31}(X - X_O) + t_{32}(Y - Y_O) + t_{33}(Z - Z_O)} \end{cases} \tag{2}$$

In this formula, (x_0, y_0, f) are the intrinsic parameters of the camera, which can be measured by camera calibration offline [48]. (X_O, Y_O, Z_O) are the space coordinates of the smartphone camera, i.e., the position of user. $t_{ij}(i, j = 1, 2, 3)$ are nine directional cosines related to the exterior orientation of smartphone. Hence, as long as more than three control points are offered (including pixel coordinates and space coordinates, which are the result of CPT from last step, Section 3.2.2), such as (A, a), (B, b) and (C, c) in Figure 5, the position can be calculated through this model. The system outputs estimated position $\left(X_O^*, Y_O^*, Z_O^*\right)$ by an iterative process.

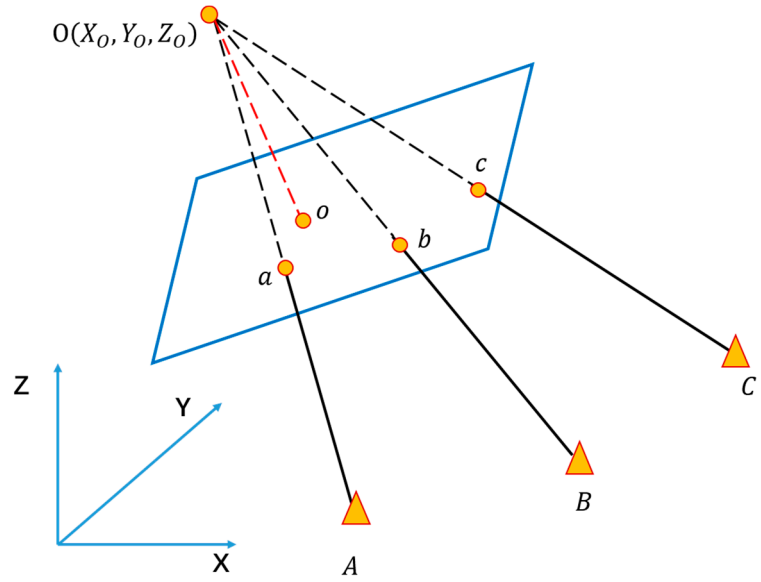

Figure 5. Principal of Position Estimation. o(x_0, y_0) is the principal point of camera image, and the length of Oo is the focal length f. By calculating collinear equation models of control point pairs, A, B, C and a, b, c, the positioning of smartphone $O(X_O, Y_O, Z_O)$ can be obtained.

3.3.2. Distance Estimation

In order to avoid gross error for the final position, distance estimation is then implemented to check the output of collinear equation model.

The principle of distance estimation can be illustrated as Figure 6a: A and B are two control points. The parallelogram in blue represents image plane; a and b are corresponding pixels of control points in image. G and g are midpoints of line AB and ab respectively. O is the focal point, which is also the position of the smartphone. The distance between the smartphone and a static object can be simplified as the length of line OG, which can be estimated by following formula:

$$OG = \frac{Og}{ab} \times AB = d_r \tag{3}$$

The length of Og can be calculated as Figure 6b: o is principal point of camera image, line Oo is the focal length f. Thus:

$$Og = \sqrt{(Oo)^2 + (og)^2} \tag{4}$$

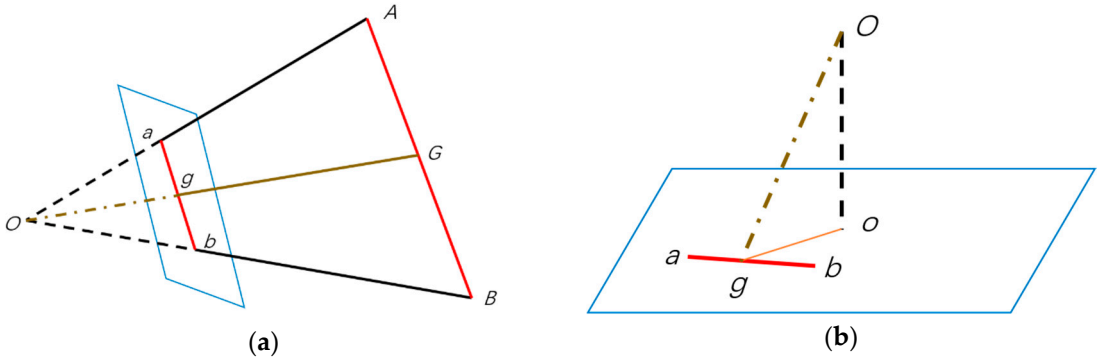

Figure 6. Principal of Distance Estimation. (**a**) Is geometric relation between smartphone camera and control points on static object; (**b**) is interior geometric relation of smartphone camera.

In addition, since the position of smartphone $O = (X_O^*, Y_O^*, Z_O^*)$ has been estimated in previous, we can calculate the distance OG directly and marked it as d_e. Then the controlled error γ as following are used to screen out gross error. If γ is less than the threshold, the estimated position is acceptable as final system output:

$$\gamma = \|d_r - d_e\| \tag{5}$$

4. Experiments

In this section, details of experiments are represented. We tested our system in a large indoor space with a relatively complicated environment and compared the positioning results with the ground truth.

4.1. Experiment Setup

The experiment was conducted on the first floor of Wanlin Art Museum in Wuhan University. The museum has about 8400 m^2 building area, and its first floor has more than 1000 m^2 with an open field and stable illumination, which is a typical large indoor scene (Figure 7). In order to verify the effectiveness, static objects shall be common and easy-to-catch in image. There are three glass doors in the experimental site, all of which can be seen at any place of the room (Figure 8). Two doors (identified as "door1" and "door2") are on the south of the museum, and "door3" locates in the north. We chose these three glass doors as static objects and used them for locating.

Figure 7. Test environment of the proposed system, i.e., Wanlin Art Museum. (**a**) is outside look of the art museum. (**b**,**c**) are photos of inside look on the first floor.

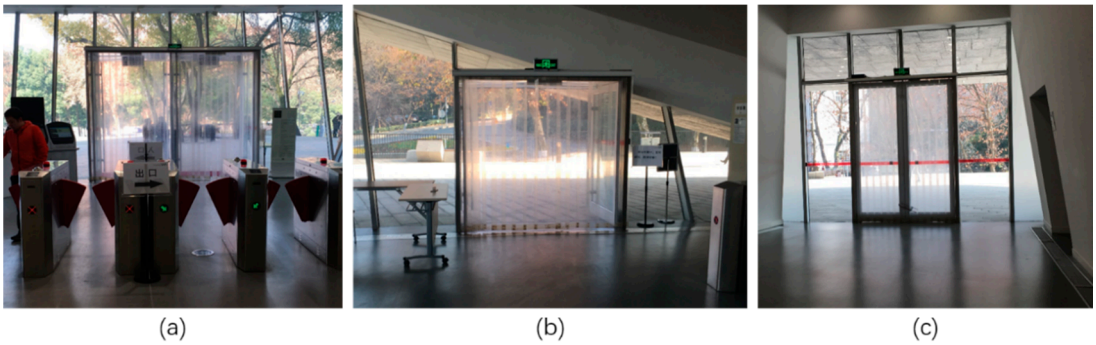

Figure 8. Three glass doors as static objects in the experiment. (**a**–**c**) are "door1", "door2" and "door3" respectively.

We used an iPhone 6S smartphone to take images in both offline and online phases, including training images for network, reference images and test images. Other procedures of the system were conducted in a computer with Titan Xp graphics card, for a purely web-based solution. In this case, the batteries of smartphones do not consume much power. The ground truth of space coordinates for all control points are measured by a Hi-Target ZTS-420R Total Station (Hi-Target Surveying Instrument Co. Ltd, Guangzhou, China) (Figure 9), with 2 mm positioning error for every 1000 m distance.

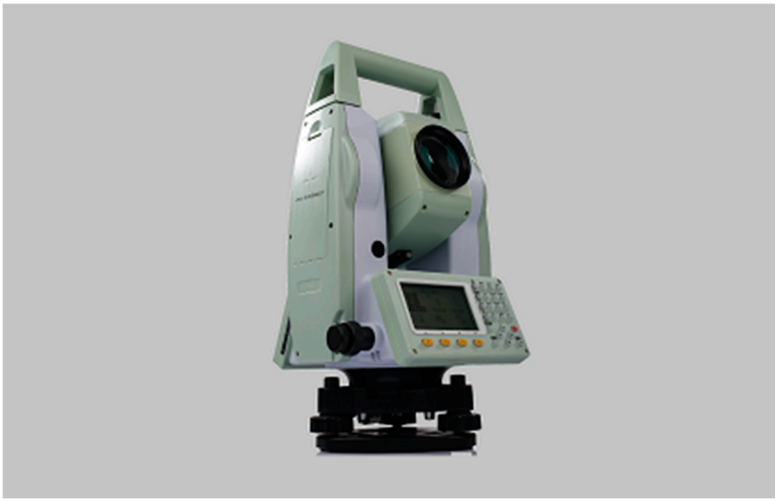

Figure 9. Hi-Target ZTS-420R Total Station is used for measuring space coordinates of control points and ground truth of test points.

We randomly selected twelve places as test points and photographed plenty of static objects images. All the test points are distributed throughout the room evenly (Figure 12).

4.2. Performance of Static Object Recognition

In the phase of static object detection and identification, we did data augmentation for training images in order to prevent the network from overfitting and improve the success rate of static object detection. We randomly blocked 30% area of the static target area in the training image to simulate the actual situation that the static objects may be blocked by pedestrians or other things. There were 302 training images in total. We adopted the strategy of transfer learning, and retrained the networked by using training images on the basis of a model trained by ImageNet [49]. Using the cross-validation method, we randomly selected 50% of the images for training, 25% for testing, and 25% for validation. The Accurate Precision (AP) of the detection and identification is shown in Table 1.

Table 1. Performance of Faster-RCNN network to detect and identify static objects. The ground truth of test images is offered by human eye judgement.

Phase	Static Object	Accurate Precision (AP)
Training	door1	100%
	door2	100%
	door3	90.9%
	mean	97.0%
Testing	door1/door2/door3	100%

The accuracy of coordinates of control points in test image is determined by the homographic matrix H_{homo}, which determines the final positioning accuracy to a large extend. We manually measured and recorded the ground truth of control points' pixels in test images (with 2~3 pixels error)

and compared them with the calculation results CPT. The errors of pixel coordinates for matching, as showed in Figure 10, mostly fall within ten pixels, which we considered as acceptable.

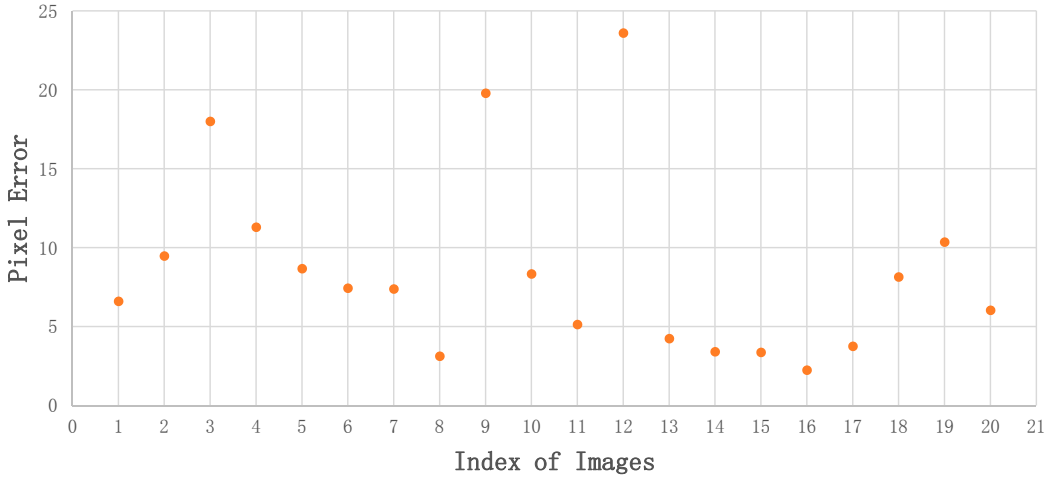

Figure 10. Accuracy of control points coordinates in test images. The horizontal axis is the index of test images; The vertical axis is pixel error between obtained pixel coordinates of control points and ground truth.

The size of images photographed by iPhone 6S is 3024 × 4032 pixels. For all of test images with such size, time cost in static object recognition phase are about 0.3 s.

4.3. Positioning Results and Analysis

This part demonstrates positioning results. We evaluated the accuracy through Euclidean distance of calculated position and ground truth position. Figure 11 illustrates the relationship of distance (user-static object) and positioning error. Within the range of 40 m, there is no significant correlation between distance and positioning precision. All of the test points achieved an accuracy of 1.5 m, and most of them are within 1 m.

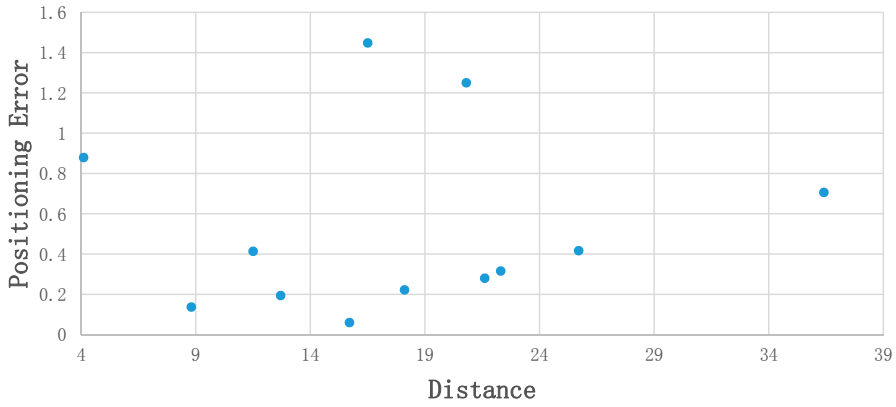

Figure 11. Relation between position error and distance.

Figure 12 is the plan map of the experimental site. Green circles on the map are error boundaries, and centers of these circles represent test points. The words near circles such as "0.14 m" means that the positioning accuracy error of this test point is 0.14 m. Just as the figure shows, nearly all test points achieve accuracy within 1 m, except for two test points, which is caused by unreasonable distribution of control points (see Section 5. Discussion). From the plan map we can see that our system has ability to locate smartphone within accuracy of 1 m in such a large indoor scene.

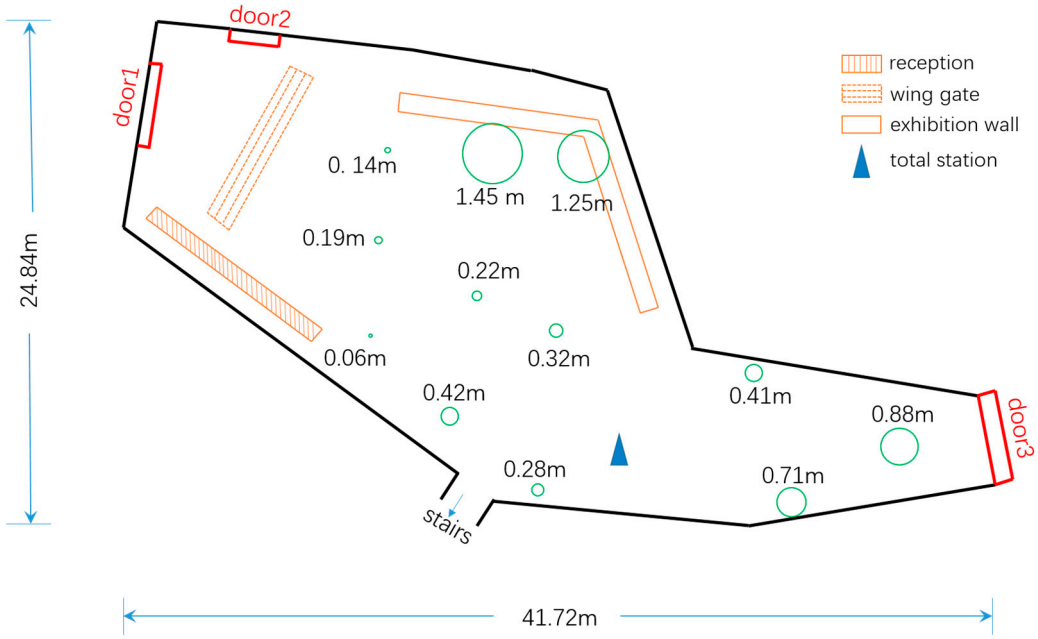

Figure 12. Plan map of experimental site and positioning results.

5. Discussion

Large indoor spaces with wide field of vision and stable illumination (such as museums, malls and airports) provide applicable environments for visual positioning. Hence, we use common static objects in these spaces as references to locate smartphone camera via visual algorithms. Our experiment in an art museum suggests that our system is able to achieve positioning accuracy within 1 m even that the experimental environment is complicated in vision.

5.1. Experimental Difficulties and Criteria for Choosing Static Objects

The static objects (doors) we chose in experimental site are actually complicated to process for a visual algorithm, because they are made of glasses. Since glass doors are transparent, their texture in image is depends on the outside environment, which can change due to many factors, like weather, time, season, illuminance and shooting angles, etc. This characteristic limits performance of feature extraction and feature matching. In this case, the strategy that designing a feature filter to get homonymy points of feature points from reference images directly, as the final control points in test image, is not robust: At first, pixels of control points in test images may not be extracted as feature points. In such condition, choosing the nearest neighbor feature points in test image increases pixel coordinate error. Secondly, it is hard to design a robust feature filter to select the correct one from huge amount of feature points when facing changeable image texture. Thus, we designed the strategy to get control points in test images from both homographic matrix H_{homo} and control points in reference images instead. By this strategy, the practicality of the system increases. However, in order to generate more feature points to increase accuracy of homographic matrix H_{homo}, static objects with non-transparent material will be better. Although we used doors as static objects in our experiments, anything can be set as static objects in indoor scenes, which increases the practicability of the method.

Besides, control points on static objects shall be chosen properly. Two test points with higher positioning accuracy error (1.45 m and 1.25 m respectively) resulted from improper distribution of control points. Only "door1" can be seen on these two places, and the control points in images were roughly distributed on a straight line. Other test points that have proper 'observation condition' reached ideal accuracy. Generally, control points shall be easy-to-capture in image, with characters different from neighboring regions, such as those with unique color or edge points, etc., and they

shall be distributed evenly throughout the static objects. Further, static objects chosen as references cannot be too small, otherwise all control points will be distributed too close to output accurate positioning results.

5.2. Evaluation

Vision-based indoor positioning problem has always been a hot research issue in recent years. However, methods that are completely based on vision algorithms and conducted on smartphones are much less frequent, and each method has its own scope of application. Due to the low cost or even without the need for any infrastructures as well as the popularity of smartphones, we believe that this kind of method has a promising future. However, the methods or systems designed so far may have their own shortcomings when implemented in large indoor scenes. In the following, we will discuss and evaluate these state-of-art of purely visual indoor positioning methods or systems when performing in large indoor spaces, most of them are based on smartphone cameras as our system. We excluded methods that integrated smartphone camera with RF signals, such as Wi-Fi, Bluetooth, wireless LAN and so on, because our system uses vision algorithm only, and thus without any infrastructures.

The *Signpost* system designed by Mulloni et al. [39] detects unobtrusive fiduciary markers that contain position information in different places as the location of users. This strategy is easy to transplant due to the low cost and low computation power of cellphones, but users have to find the nearest marker as these tiny markers cannot be made large in size. This is inconvenient for people in large indoor spaces since they may have to move a long distance to find and scan a marker. Hile et al. [34] built a novel system to find precise locations and orient users by matching features extracted from smartphone images with relevant sections of a floorplan map. However, as this creative system is designed for environments like hallways, it may perform poorly in large indoor scenes, because the available features such as edge or corners are much less abundant in large spaces. The distance estimation algorithm by Werner et al. [37] is adopted by our system as Distance Estimation module (Section 3.3.2). This algorithm is able to calculate accurate distances between cameras and reference targets. However, since the trilateral location method requires at least three reference target positions as well as the corresponding distances to the camera (one distance for each image), this algorithm performs better in environments such as corridors with narrow width (requiring one target/distance only) than wide open indoor districts (require at least three targets/distances). Van Opdenbosch et al. [50] realized indoor positioning by image recognition. In order to achieve meter-accurate localization, they photographed images in every 1 m $\times$ 1 m grid cell, using 16 viewing angles for each spot. However, this strategy has low efficiency in larger rooms because the vast number of images in the dataset will result in a huge computation cost and increase the burden of smartphones and web servers. Kawaji et al. [51] realized indoor positioning in a large indoor space—a railway museum. They used omnidirectional panoramic images captured by an omnidirectional camera and supplemental images captured by digital cameras to build a dataset and matched the test digital camera images with reference images in the dataset to locate users. Since omnidirectional panoramic images are suitable for large scenes, this method is effective for localization in large indoor spaces. However, the output location of test images is the same as that of the omnidirectional panoramic images, which cannot achieve a relatively accurate position. Deretey et al. [52] proposed a method by matching camera images with 3D models of indoor scenes by a Simultaneous Localization and Mapping (SLAM) system to realize positioning. Although it requires a 3D model building process in an offline phase, we think it is a promising method due to the development of SLAM technology. Xu el al. [53] proposed a novel monocular visual method to locate positions in offices based on ceilings. This strategy may not function in large indoor scenes like museums or airports because these places usually do not have a planar ceiling floor with a series of chessboard-like blocks like offices. Our solution to locate users is more workable in large indoor environments. Static objects such as doors and windows are not

only common in indoor scenes, but also relatively large enough to be captured from a long distance. The longest distance in our experiment is nearly 40 m, with a positioning accuracy of 0.7 m.

There are some factors that may influence the performance of our system: illumination of the indoor scenes as well as shooting angles may change images and have an impact on feature matching. In addition, the performance of the smartphones may also affect the final result. The system does not consume much battery power of smartphones as our system is web-based and the smartphone only take images as input. The clarity of images taken by smartphone cameras is high enough for our task. The distortion of images by different smartphones may change the final position output, but a camera calibration process can fix this problem (details can be found in our previous research [5]). In the future, we will try to improve the robustness of our system and experiment in other large indoor spaces with more rooms and more complex topologies.

6. Conclusions

In this paper, a positioning system in large indoor spaces by using smartphone cameras based on static objects is proposed. Our system uses smartphone images to detect specific static objects indoors and calculate users' position. The system imitates the human brain's cognitive mode and integrates algorithms of deep learning and computer vision. We experimented in an art museum with a large indoor area and a complex visual environment. Experimental results show that this method has the ability to achieve the positioning accuracy within 1 m in a distance range of 40 m indoors. We believe that it has potential for wide application in large indoor scenes.

Author Contributions: This paper is a collaborative work by all authors. Proposed the idea, A.X. implemented the system, performed the experiments, analyzed the data and wrote the manuscript. R.C. and D.L. helped to propose the idea, give suggestions and revise the rough draft. Y.C. helped with all of experiments, especially data acquisition. D.W. helped revised code and helped to do some of the experiments.

Funding: This study is supported by the National Key Research and Development Program of China (2016YFB0502201 and 2016YFB0502202), the NSFC (91638203), the State Key Laboratory Research Expenses of LIESMARS.

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

References

1. O'Keefe, J. Place units in the hippocampus of the freely moving rat. *Exp. Neurol.* **1976**, *51*, 78–109. [CrossRef]
2. O'Keefe, J.; Dostrovsky, J. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **1971**, *34*, 171–175. [CrossRef]
3. Fyhn, M.; Molden, S.; Witter, M.P.; Moser, E.I.; Moser, M.-B. Spatial representation in the entorhinal cortex. *Science* **2004**, *305*, 1258–1264. [CrossRef] [PubMed]
4. Sargolini, F.; Fyhn, M.; Hafting, T.; McNaughton, B.L.; Witter, M.P.; Moser, M.-B.; Moser, E.I. Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science* **2006**, *312*, 758–762. [CrossRef] [PubMed]
5. Wu, D.; Chen, R.; Chen, L. Visual positioning indoors: Human eyes vs. smartphone cameras. *Sensors* **2017**, *17*, 2645. [CrossRef] [PubMed]
6. Ruizhi, C.; Liang, C. Indoor Positioning with Smartphones: The State-of-the-art and the Challenges. *Acta Geod. Cartogr. Sin.* **2017**, *46*, 1316–1326. [CrossRef]
7. Youssef, M.; Agrawala, A. The Horus WLAN location determination system. In Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services, Seattle, WA, USA, 6–8 June 2005; ACM: New York, NY, USA, 2005; pp. 205–218.
8. Bahl, P.; Padmanabhan, V.N. RADAR: An in-building RF-based user location and tracking system. In Proceedings of the IEEE INFOCOM 2000 Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064), Tel Aviv, Israel, 26–30 March 2000; Volume 2, pp. 775–784.

9. Vaupel, T.; Seitz, J.; Kiefer, F.; Haimerl, S.; Thielecke, J. Wi-Fi positioning: System considerations and device calibration. In Proceedings of the 2010 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Zurich, Switzerland, 15–17 September 2010; pp. 1–7.
10. Hansen, R.; Wind, R.; Jensen, C.S.; Thomsen, B. Algorithmic strategies for adapting to environmental changes in 802.11 location fingerprinting. In Proceedings of the 2010 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Zurich, Switzerland, 15–17 September 2010; pp. 1–10.
11. Teuber, A.; Eissfeller, B. WLAN indoor positioning based on Euclidean distances and fuzzy logic. In Proceedings of the 3rd Workshop on Positioning, Navigation and Communication, Lower Saxony, Germany, 16 March 2006; pp. 159–168.
12. Haverinen, J.; Kemppainen, A. Global indoor self-localization based on the ambient magnetic field. *Rob. Auton. Syst.* **2009**, *57*, 1028–1035. [CrossRef]
13. Chen, L.; Kuusniemi, H.; Chen, Y.; Pei, L.; Kröger, T.; Chen, R. Information filter with speed detection for indoor Bluetooth positioning. In Proceedings of the 2011 International Conference on Localization and GNSS (ICL-GNSS), Tampere, Finland, 29–30 June 2011; pp. 47–52.
14. Chen, L.; Kuusniemi, H.; Chen, Y.; Liu, J.; Pei, L.; Ruotsalainen, L.; Chen, R. Constraint Kalman filter for indoor bluetooth localization. In Proceedings of the 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 1915–1919.
15. Chen, L.; Pei, L.; Kuusniemi, H.; Chen, Y.; Kröger, T.; Chen, R. Bayesian fusion for indoor positioning using bluetooth fingerprints. *Wirel. Pers. Commun.* **2013**, *70*, 1735–1745. [CrossRef]
16. Bargh, M.S.; de Groote, R. Indoor localization based on response rate of bluetooth inquiries. In Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-Less Environments, San Francisco, CA, USA, 19 September 2008; ACM: New York, NY, USA, 2008; pp. 49–54.
17. He, S.; Chan, S.H.G. INTRI: Contour-Based Trilateration for Indoor Fingerprint-Based Localization. *IEEE Trans. Mob. Comput.* **2017**, *16*, 1676–1690. [CrossRef]
18. Quuppa Company. Available online: http://quuppa.com/company/ (accessed on 10 July 2018).
19. Lakmali, B.D.S. Database Correlation for GSM Locationin Outdoor & Indoor Environments. In Proceedings of the 4th International Conference on Information and Automation for Sustainability (ICIAFS), Colombo, Sri Lanka, 12–14 December 2008.
20. Zhao, Y. Standardization of mobile phone positioning for 3G systems. *IEEE Commun. Mag.* **2002**, *40*, 108–116. [CrossRef]
21. Want, R.; Hopper, A.; Falcao, V.; Gibbons, J. The active badge location system. *ACM Trans. Inf. Syst.* **1992**, *10*, 91–102. [CrossRef]
22. Ward, A.; Jones, A.; Hopper, A. A new location technique for the active office. *IEEE Pers. Commun.* **1997**, *4*, 42–47. [CrossRef]
23. Priyantha, N.B.; Chakraborty, A.; Balakrishnan, H. The cricket location-support system. In Proceedings of the 6th nnual International Conference on Mobile Computing and Networking, Boston, MA, USA, 6–11 August 2011; ACM: New York, NY, USA, 2000; pp. 32–43.
24. Kim, B.; Kwak, M.; Lee, J.; Kwon, T.T. A multi-pronged approach for indoor positioning with WiFi, magnetic and cellular signals. In Proceedings of the 2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Busan, Korea, 27–30 October 2014; pp. 723–726.
25. Jeon, J.S.; Kong, Y.; Nam, Y.; Yim, K. An Indoor Positioning System Using Bluetooth RSSI with an Accelerometer and a Barometer on a Smartphone. In Proceedings of the 2015 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA), Krakow, Poland, 4–6 November 2015; pp. 528–531.
26. Chen, L.H.; Wu, H.K.; Jin, M.H.; Chen, G.H. Intelligent Fusion of Wi-Fi and Inertial Sensor-Based Positioning Systems for Indoor Pedestrian Navigation. *IEEE Sens. J.* **2014**, *14*, 4034–4042. [CrossRef]
27. Li, Y.; Zhuang, Y.; Zhang, P.; Lan, H.; Niu, X.; El-Sheimy, N. An improved inertial/wifi/magnetic fusion structure for indoor navigation. *Inf. Fusion* **2017**, *34*, 101–119. [CrossRef]
28. Liu, M.; Chen, R.; Li, D.; Chen, Y.; Guo, G.; Cao, Z.; Pan, Y. Scene Recognition for Indoor Localization Using a Multi-Sensor Fusion Approach. *Sensors* **2017**, *17*, 2847. [CrossRef] [PubMed]
29. Becker, M.; Ahuja, B. Implementing real-life indoor positioning systems using machine learning approaches. In Proceedings of the 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), Larnaca, Cyprus, 27–30 August 2017; pp. 1–6.

30. Gao, R.; Ye, F.; Wang, T. Smartphone indoor localization by photo-taking of the environment. In Proceedings of the 2014 IEEE International Conference on Communications (ICC), Sydney, Australia, 10–14 June 2014; pp. 2599–2604.
31. Tian, Y.; Gao, R.; Bian, K.; Ye, F.; Wang, T.; Wang, Y.; Li, X. Towards ubiquitous indoor localization service leveraging environmental physical features. In Proceedings of the IEEE INFOCOM 2014—IEEE Conference on Computer Communications, Toronto, ON, Canada, 27 April–2 May 2014; pp. 55–63.
32. Shafer, S.; Krumm, J.; Brumitt, B.; Meyers, B.; Czerwinski, M.; Robbins, D. The New EasyLiving Project at Microsoft Research. In Proceedings of the 1998 Joint DARPA/NIST Smart Spaces Workshop, Gaithersburg, MD, USA, 30–31 July 1998; pp. 30–31.
33. Mautz, R. Indoor Positioning Technologies. Habilitation Thesis, Institute of Geodesy and Photogrammetry, Department of Civil, Environmental and Geomatic Engineering, ETH Zurich, Zurich, Switzerland, 2012.
34. Hile, H.; Borriello, G. Positioning and orientation in indoor environments using camera phones. *IEEE Comput. Graph. Appl.* **2008**, *28*, 32–39. [CrossRef]
35. Kohoutek, T.K.; Mautz, R.; Donaubauer, A. Real-time indoor positioning using range imaging sensors. In Proceedings of the Real-Time Image and Video Processing 2010, Brussels, Belgium, 4 May 2010; Volume 7724, pp. 1–8.
36. Kim, J.; Jun, H. Vision-based location positioning using augmented reality for indoor navigation. *IEEE Trans. Consum. Electron.* **2008**, *54*, 954–962. [CrossRef]
37. Werner, M.; Kessel, M.; Marouane, C. Indoor positioning using smartphone camera. In Proceedings of the 2011 International Conference on Indoor Positioning and Indoor Navigation, Guimaraes, Portugal, 21–23 September 2011. [CrossRef]
38. Möller, A.; Kranz, M.; Huitl, R.; Diewald, S.; Roalter, L. A mobile indoor navigation system interface adapted to vision-based localization. In Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, MUM 2012, Ulm, Germany, 4–6 December 2012; pp. 4:1–4:10.
39. Mulloni, A.; Wagner, D.; Barakonyi, I.; Schmalstieg, D. Indoor positioning and navigation with camera phones. *IEEE Pervasive Comput.* **2009**, *8*, 22–31. [CrossRef]
40. Ganick, A.; Ryan, D. Light positioning system using digital pulse recognition. U.S. Patent 824,846,7B1, 26 July 2011.
41. Ruotsalainen, L.; Kuusniemi, H.; Bhuiyan, M.Z.H.; Chen, L.; Chen, R. A two-dimensional pedestrian navigation solution aided with a visual gyroscope and a visual odometer. *GPS Solut.* **2013**, *17*, 575–586. [CrossRef]
42. Ruotsalainen, L. Visual Gyroscope and Odometer for Pedestrian Indoor Navigation with a Smartphone. In Proceedings of the 25th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2012), Nashville, TN, USA, 17–21 September 2012; pp. 2422–2431.
43. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
44. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
45. Zuxun, Z.; Jianqing, Z. *Digital Photogrametry*, 2nd ed.; Wuhan Univerty Press: Wuhan, China, 2002.
46. Keypoints, S.; Lowe, D.G. Distinctive Image Features from. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
47. Fischler, M.A.; Bolles, R.C. Random Sample Consensus: A Paradigm for Model Fitting with. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
48. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [CrossRef]
49. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
50. Van Opdenbosch, D.; Schroth, G.; Huitl, R.; Hilsenbeck, S.; Garcea, A.; Steinbach, E. Camera-based indoor positioning using scalable streaming of compressed binary image signatures. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 2804–2808.
51. Kawaji, H.; Hatada, K.; Yamasaki, T.; Aizawa, K. Image-based indoor positioning system: Fast image matching using omnidirectional panoramic images. In Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis, Firenze, Italy, 29 October 2010; ACM: Firenze, Italy, 2010; pp. 1–4.

52. 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), Banff, AB, Canada, 13–16 October 2015; pp. 1–9.
53. Xu, D.; Han, L.; Tan, M.; Li, Y.F. Ceiling-based visual positioning for an indoor mobile robot with monocular vision. *IEEE Trans. Ind. Electron.* **2009**, *56*, 1617–1628.