Solving Monocular Visual Odometry Scale Factor with Adaptive Step Length Estimates for Pedestrians Using Handheld Devices

Nicolas Antigny [1,2,3,*], Hideaki Uchiyama [4], Myriam Servières [2,3,5], Valérie Renaudin [1,3], Diego Thomas [4] and Rin-ichiro Taniguchi [4]

[1] Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux (IFSTTAR) AME GEOLOC, 44340 Bouguenais, France; valerie.renaudin@ifsttar.fr
[2] Centrale Nantes, 44300 Nantes, France; myriam.servieres@ec-nantes.fr
[3] Institut de Recherche en Sciences et Techniques de la Ville (IRSTV), 44300 Nantes, France
[4] Laboratory for Image and Media Understanding (LIMU), Kyushu University, Fukuoka 819-0395, Japan; uchiyama@limu.ait.kyushu-u.ac.jp (H.U.); thomas@ait.kyushu-u.ac.jp (D.T.); rin@kyudai.jp (R.-i.T.)
[5] Centre de Recherche Nantais Architectures Urbanités (CRENAU) AAU, 44262 Nantes, France
[*] Correspondence: nicolas.antigny@ifsttar.fr

Abstract: The urban environments represent challenging areas for handheld device pose estimation (i.e., 3D position and 3D orientation) in large displacements. It is even more challenging with low-cost sensors and computational resources that are available in pedestrian mobile devices (i.e., monocular camera and Inertial Measurement Unit). To address these challenges, we propose a continuous pose estimation based on monocular Visual Odometry. To solve the scale ambiguity and suppress the scale drift, an adaptive pedestrian step lengths estimation is used for the displacements on the horizontal plane. To complete the estimation, a handheld equipment height model, with respect to the Digital Terrain Model contained in Geographical Information Systems, is used for the displacement on the vertical axis. In addition, an accurate pose estimation based on the recognition of known objects is punctually used to correct the pose estimate and reset the monocular Visual Odometry. To validate the benefit of our framework, experimental data have been collected on a 0.7 km pedestrian path in an urban environment for various people. Thus, the proposed solution allows to achieve a positioning error of 1.6–7.5% of the walked distance, and confirms the benefit of the use of an adaptive step length compared to the use of a fixed-step length.

Keywords: pose estimation; localization; handheld device; pedestrian navigation; urban mobility; augmented reality

1. Introduction

In the context of pedestrian navigation and handheld device pose estimation for urban mobility, localization and orientation would gain from an accurate global pose estimation using Global Navigation Satellite Systems (GNSS), Inertial Measurement Unit (IMU) and cameras [1–3]. Such systems generally operate well in outdoor open environments. However, urban environments comprising closely spaced buildings (i.e., urban canyon) still represent challenging areas for GNSS, which suffer from attenuation, reflection, and blockage effects [4–6]. It is even more challenging using low-cost Micro Electro Mechanical System (MEMS) sensors and low computational resources, which are typically embedded in pedestrian mobile devices [7]. Moreover, the hand can move freely even when the pedestrian's centre of mass is static, making it difficult to estimate an accurate pose.

In order to assist the pedestrian navigation during large displacements in urban environments, e.g., through on-site augmented reality applications on handheld mobile device, our aim is to provide a continuous and accurate pose estimate. In such context, we propose to integrate human motion estimation, i.e., step length estimation, using a handheld IMU proposed in the field of Pedestrian Dead Reckoning (PDR), to scale the monocular Visual Odometry (VO). To complete the pose estimation, we propose to use an estimation of the holding in hand of mobile devices coupled with data content in Geographical Information System (GIS). Although embedded technologies in general public devices become more efficient, the choice was made to propose a solution that can be fully embedded with a minimal hardware setup and a low memory requirement, and without any connection with networks, any deployment of new infrastructures, to be totally autonomous.

Thus, our first contribution is to achieve a continuous pose estimation with a mobile device held in hand during large displacements in urban environments. Our second contribution is to present an adaptive method, based on the use of an IMU and a monocular camera, that takes advantage of human motion estimation and data content in Geographic Information Systems to dynamically solve the scale ambiguity and suppress the scale drift in the monocular Visual Odometry. Finally, our third contribution is to propose a solution that does not require pedestrians to make specific and unnatural movements or to revisit the same place, as is the case with Simultaneous Localization And Mapping (SLAM) techniques [8–10].

The Section 2 conducts a state of the art on pose estimation and Visual Odometry scale estimation. The Section 3 presents the coupling process between PDR and VO, i.e., the scale estimation based on step length estimation, and the position estimation on the vertical axis using a hand-held model of the equipment and GIS data. The Section 4 details the hardware setup used in our experiments, the acquisition scenario and the establishment of reference waypoints and tracks to evaluate our method. The assessment by comparison between a foot-mounted INS aided by GNSS phase measurements and the proposed method using a handheld device is presented in the Section 5. Finally, we conclude that human motion analysis can be informative as an additional clue and constraint to scale the monocular Visual Odometry and to improve the pose estimation with general public handheld devices.

2. Related Work

In this section, we briefly summarize the solutions for handheld device pose estimation in urban environments using a camera or/and an IMU. In the literature on urban localization, one approach is to use a camera and a database comprising images associated with their global poses, namely appearance-based localization [3,11,12]. The pose of the camera is determined by searching an image containing a part of the input image in the database. This is typically performed by using Content-Based Image Retrieval techniques. However, the accuracy of this approach strongly degrades under appearance changes. In addition, such a huge database is not always available, not easy to prepare, and also may not be applicable to mobile devices for pedestrians due to limited computational resources.

In the field of pedestrian navigation, Pedestrian Dead-Reckoning approaches have been investigated for the urban localization [13,14]. The pose is incrementally determined by computing the displacement between the poses after the initial global pose is acquired with other approaches such as appearance-based localization. Sensor fusion is an alternative approach and has also been investigated to compensate for the drawback of each sensor [15–18]. The fusion was initially introduced to compensate for weaknesses in appearance-based approaches due to failure under fast rotational motions (i.e., motion blur and large displacement in the images) [15,19,20]. A gyroscope in the IMU has an advantage that it can accurately track sensor angular rates at high frequency for short time intervals. An accelerometer in the IMU is useful during static phases to compensate for global orientation estimation by using the gravity measurement.

Monocular Visual Odometry is a frame-to-frame relative pose tracking approach using a camera for the incremental pose estimation [21–23]. In unknown environments, the camera pose

is incrementally determined by, first, computing feature correspondences between two consecutive images, and then, computing the relative pose from the correspondences. However, one problem is to solve for scale ambiguity in the estimated poses [23]. The relative poses can fundamentally be estimated up-to-scale if known landmarks are used into the VO [8] or an initial translation of a camera is assumed to be known [24]. These approaches solve the initial scale estimation but still face the problem of scale drift that occurs in large displacements and degrades the localization accuracy [25]. As reported in [26], both direct and indirect VO approaches cannot avoid the scale drift when only a monocular camera is available, and the tracking errors are quickly accumulated over time. The drift can be punctually corrected by using appearance-based localization, using some known landmarks [3,27] or knowledge of existing 3D city models [28]. Ref. [29] presents a method for resolving the scale ambiguity and drift by using the slant distance obtained from a skyline matching between the camera and images synthesized using a 3D building model. The correction of the scale factor shows a 90% improvement of the positioning solution compared to a solution that does not correct the scale drift. These methods require to be located in an area where the 3D model is available and known with precision, whereas it is generally not. Loop closure, when revisiting the same place, is also used to correct the drift [25]. However, this approach is not applicable in the context of pedestrians navigating from point A to point B, without passing through the same place again.

Since the unit of the accelerometer in the IMU is metric, it is also useful to estimate the scale in the VO [30–32]. To estimate the scale, one approach is to compute the double integration of the acceleration after removing the gravity elements. However, the double integration is sensitive to noise because the error is quickly accumulated, even though the error is small during a short time interval. To further improve the accuracy and stability of metric scale estimation, additional methodologies need to be investigated. An alternative approach is to estimate the scale with an Extended Kalman Filter, that includes the scale factor in the state vector. As reported in [31], the method is also sensitive to a dynamic bias which is difficult to estimate.

Others scale estimation approaches exist but there is a hard constraint that the sensors need to be rigidly fixed to the pedestrian. For example, Ref. [33,34] have the sensors attached to a helmet. This constraint is not valid in the context of handheld AR.

As a recent approach dedicated to pedestrians, the pedestrian face is used as a known object to estimate the scale [32]. On a mobile device (e.g., a smartphone), two cameras are usually installed: a user-facing camera and a world-facing camera. In the context of handheld AR, the user-facing camera captures the user's face, while performing the VO by using the world-facing camera. When a relative pose of the two cameras is calibrated, the metric scale for the VO can be computed from the face. This setup is reasonable for handheld devices, however, this approach requires that the pedestrian be static while the hand is moving. It should also be noted that VO using a stereo camera does not have these scale issues because the scale can be uniquely determined from the disparity between two cameras [26].

Other approaches, close to ours, use a pedometer and an average step length for scale estimation in monocular VO [35,36]. As the pedestrian has to avoid others pedestrians, bikers, cars, elements in the scene and to wait at pedestrian crossings, the step lengths are not constant during the walk and a average step length does not correspond to pedestrian movement in urban environments. This is why we propose to use a dynamic step length estimation to scale monocular VO.

3. Scaled Monocular Visual Odometry

In this paper, we propose a novel approach for pose estimation with sensors held in hand based on monocular Visual Odometry and Pedestrian Dead-Reckoning. A human motion analysis from inertial data, i.e., a step length estimation, is used to dynamically solve the scale ambiguity and suppress the scale drift. The overview of the proposed method is illustrated in Figure 1. The proposed method can be divided into three main processes: "PDR" part, "VO" part and "Coupling" part. Since the "PDR" part and "VO" part are independent, our approach can be referred as a loosely coupled approach. The variables needed to develop the proposed method are all detailed in their corresponding following parts.
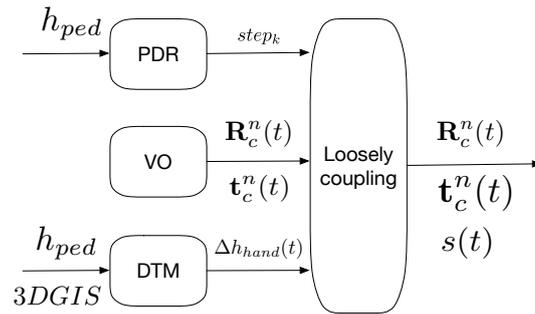
Figure 1. Block diagram of the proposed approach.

The "PDR" part is dedicated to the processing of IMU measurements. The same origin is defined for the IMU sensors and the IMU measurements are given in the *Body* frame, labeled by b. We estimate pedestrian step lengths with sensors held in hand, while the pedestrian walks naturally [37,38]. The process outputs a step length $step_k$ in meter at every instant of step t_{step_k}. The details are given in Section 3.1.

The "VO" part is dedicated to monocular VO for relative pose estimation based tracking. The origin of the camera frame is the optical center of the camera and the visual measurements are given in the *Camera* frame, labeled by c. In this part, because the scale estimation is independent of the image processing, any VO or vSLAM method can be applied if it outputs a complete pose estimate [23,39]. The implemented Visual Odometry framework is detailed in Section 3.2.

The "Coupling" part is dedicated to the scale determination by fitting trajectories from the "PDR" part and the "VO" part. This part finally solves the scale ambiguity in monocular VO. The scaled pose in monocular VO is the final output of the proposed solution. The final pose estimate is given in the *Navigation* frame (i.e., the North-East-Down (NED) frame), labeled by n. In the proposed solution, we need to consider the frequency difference between sensors and the timing of the scale estimation. The details are given in Section 3.4.

3.1. Step Length Estimation Process

Classically, inertial signals are integrated according to a strap-down mechanization to compute the recursive positions of an IMU. This is possible only if the accumulated error caused by low-cost inertial sensors is frequently calibrated or reset, such as using Zero velocity UPdaTes (ZUPT) [14]. As a moment of zero velocity does not always occur in the context of hand-held device-based pedestrian navigation, ZUPT is not easily achievable. Also, the location of the pedestrian is normally represented as a pedestrian's center of mass on a 2D map. This is obviously different from the location of the device held in hand, because the hand performs free motions even when the pedestrian's center of mass is static. Therefore, instead of double-integrating the measured accelerations, a step length model is adopted to derive step lengths as pedestrian's displacements for handheld devices. We summarize the procedure of pedestrian step length estimation proposed in [37,38]. It is claimed that this step length estimation is performed with a 2.5% up to 5% error on the walked distance. The different steps of the procedure are presented in Figure 2 and the notations are detailed hereafter.
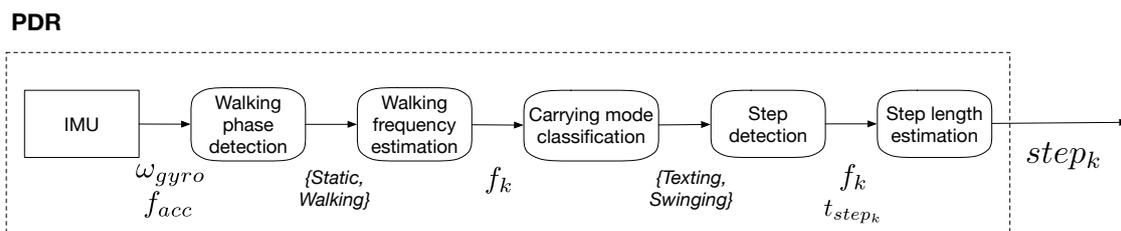

Figure 2. Detailed steps of "PDR" block.

To estimate step lengths, IMU signals acquired with sensors held in hand are analyzed as follows. A first motion classification is operated to determine the walking phases or the static phases of the pedestrian. A peak detection and a thresholding on the energy are applied on gyroscope signal ω_{gyro} and accelerometer signal f_{acc} to determine the events t_{step_k} when the pedestrian's foot comes in contact with the ground at k^{th} step. The step frequency is also computed as f_k. A second motion classification is operated by analyzing the variance of the IMU measurements to determine the device's carrying mode (*Static* and *Walking* in *Texting* or *Swinging* mode) according to [38]. Then, a generic model is used to compute the step lengths $step_k$ according to [37]. It is based on the user's height h_{ped} and on a set of three parameters $\{a, b, c\}$ trained on 12 subjects.

$$step_k = h_{ped} \times (a \times f_k + b) + c \tag{1}$$

3.2. Monocular Visual Odometry

Because step length estimation is independent of monocular Visual Odometry, any existing methods can be used. Here, we introduce our implementation based on the following standard procedure [22]. As a pre-processing, a standard camera calibration, using a checkerboard [40] and modeling the camera as a pinhole camera [41], is performed for a fixed resolution to express the camera's coordinates in a normalized space [42]. This enables to correct image distortions and to determine the intrinsic parameters matrix **K** of the camera.

In the VO, to determine the unknown pose at time (t), the known poses at $(t-1)$ and $(t-2)$ are used. The different steps of the vision procedure are detailed in Figure 3 and the notations are detailed hereafter.
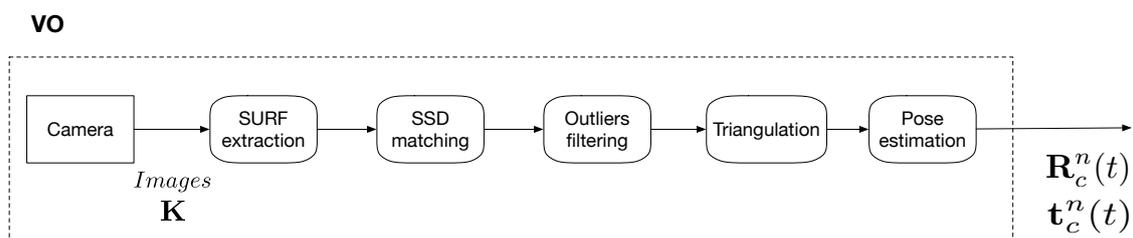

Figure 3. Detailed steps of "VO" block.

Knowing the intrinsic parameters matrix **K** of the camera and the correspondences between $\mathbf{x}_i$ and $\mathbf{X}_n$, expressed in homogeneous coordinates symbolized with a tilde (˜), the pose of the camera's optical center can be computed. It is expressed as a rotation matrix $\mathbf{R}_n^c$, giving the rotation from the global coordinate system to the camera one, and a translation vector $\mathbf{t}_n^c$ giving the translation from the origin of the global coordinate system to the one in the camera coordinate system **C**. The output of the "Vision" part in Figure 3 are $\mathbf{R}_c^n(t)$, the rotation from the *Camera* frame to the *Navigation* frame at time (t) and $\mathbf{t}_c^n(t)$, the translation from *Camera* frame to the *Navigation* frame $\mathbf{t}_c^n(t)$ with a dimensionless scale factor $s(t)$ at time (t).

For the relative pose estimation between images, feature points $\mathbf{x}_i$ are extracted by SURF detector [43]. A sparse feature points extraction is operated to improve the reliability by suppressing redundantly extracted points. Then, extracted feature points in consecutive images are matched using the Sum Square Difference (SSD) distance to select unique correspondences. A filtering stage, using a geometric constraint such as epipolar constraint, is then applied with the M-estimator SAmple Consensus (MSAC) algorithm [44] to exclude outliers. A triangulation stage is operated only on inliers that are visible in the two images at times $(t-1)$ and $(t-2)$, and in the current image at time (t) to estimate the 3D points $\mathbf{X}_n$ in the *Navigation* frame. To perform pose estimation using a

calibrated camera, PnP algorithms constitute one of the most suitable solutions [45]. The camera pose is computed using the following formulation:

$$\tilde{x}_i = K \begin{bmatrix} R_n^c & | & t_n^c \end{bmatrix} \tilde{X}_n \tag{2}$$

$$t_n^c = -R_n^c C^n \tag{3}$$

where C^n is the position of the camera's optical center in the *Navigation* frame. The successive positions of the camera's optical center are used for the scale estimation. It should be noted that we did not use any map optimization, such as pose graph optimization and bundle adjustment typically used in visual SLAM because they are computationally expensive. We simply implemented monocular Visual Odometry and did not keep the map in the memory as Dead-Reckoning.

3.3. Digital Terrain Model and Handheld Height

To complete the position estimation on the horizontal plane, the position on the vertical axis is determined in urban environments using data contained in 3D Geographical Information System (GIS), i.e., the Digital Terrain Model (DTM) and an estimation of the hand height holding a mobile device with respect to the pedestrian height. The DTM is a set of points referenced in planimetry (X,Y) and altimetry (Z). With an interpolation method, this provides the elevation of the ground level relief in digital form [46]. In the context of the use of an augmented reality applications for mobility assistance, the screen of the equipment held in the hand is considered maintained at a height h_{hand} to see information. In order to estimate the handheld device position on the vertical axis, the height of several pedestrians h_{ped}, as well as the screen center height h_{hand} on which an augmented reality display would be proposed, were measured. These measurements are presented in Table 1.

Table 1. Measurements of pedestrian height and handheld equipment height.

Pedestrians	*M1*	*M2*	*W1*	*W2*	*W3*
Gender	Male	Male	Female	Female	Female
h_{ped}	1.87 m	1.80 m	1.69 m	1.69 m	1.60 m
h_{hand}	1.67 m	1.60 m	1.55 m	1.50 m	1.40 m
$h_{ped} - h_{hand}$	0.20 m	0.20 m	0.14 m	0.19 m	0.20 m
h_{hand}/h_{ped}	89%	88%	91%	88%	87%

Thus, knowing the distance between the center of the screen and the optical center of the camera, the height of the equipment held in the hand can be experimentally defined as a ratio of the pedestrian height:

$$h_{hand} \sim 0.9 \times h_{ped} \tag{4}$$

It should be noted that the height variations $\Delta h_{hand}(t)$, i.e., the variations of the DTM, is used to estimate the scale factor of the displacement along the vertical axis in the following section. In addition, since the variations in the DTM are very small, displacements on the vertical axis could be ignored to simplify the scale-factor calculation.

3.4. Scale Determination

Because steps and images are not sampled at the same times, an interpolation is needed. Then using the knowledge of pedestrian step lengths and the frequencies of both the visual measurement and the pedestrian's step, a linear interpolation is operated to determine the pedestrian displacement between the instants of two consecutive images, as illustrated in Figure 4. The scale of the VO during the step $k + 1^{th}$ is computed by using the kth step length based on the interpolation.

The step length estimation $step_k$ provides the magnitude of the pedestrian displacements on the horizontal plane. With the assumption that the displacements of the handheld device are mainly on the horizontal plane, displacements on the vertical axis are not taken into account.
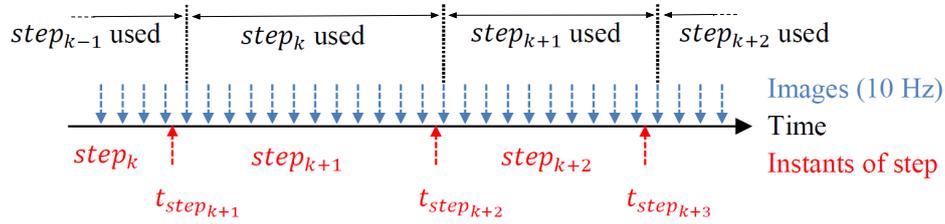

Figure 4. Step length interpolation.

Therefore, to estimate the scale $s(t)$, a comparison is made between $D_{VO}(t)$, i.e., the magnitude of the displacement of the camera's optical center $\mathbf{C}^n$ on the horizontal plane estimated by the VO at times $(t-1)$ and (t), and $D_{step}(t)$, i.e., an interpolation at times $(t-1)$ and (t) of the estimated pedestrian step length on the horizontal plane $step_k$ between the instants of step $(k-1)$ and (k).

$$D_{VO}(t) = \|\mathbf{C}^n(t) - \mathbf{C}^n(t-1)\| \tag{5}$$

$$D_{step}(t) = step_k \times \frac{\Delta t_{image}}{\Delta t_{step_k}} \tag{6}$$

The scale factor $s(t)$ is then defined as:

$$s(t) = \frac{D_{step}(t)}{D_{VO}(t)} \tag{7}$$

Thus, at each relative pose estimated by the VO, the scale is computed and used to correct the displacement of the handheld device on the horizontal plane in the *Navigation* frame. The final outputs are $\mathbf{R}_c^n(t)$, the rotation matrix from the *Camera* frame to the *Navigation* frame at time (t) and $\mathbf{t}_c^n(t)$, the translation vector from the *Camera* frame to the *Navigation* frame at time (t).

It should be noted that during the whole process, several steps can fail. In the "PDR" part, there might be some miss or false step detections. In the "VO" part, there might be false correspondences between features even if the matching process should limit it.

3.5. Known Object Recognition-Based Pose Estimation

To punctually suppress the drift and correct the relative pose estimate based on the scaled monocular Visual Odometry, the detailed knowledge of existing sparse known objects contained in 3D GIS is used. A known object allows to estimate absolute pose when one is detected in video frames [47]. The mean positioning accuracy is claimed to be 25 cm on the horizontal plane when a known object was detected in video frames. It should be noted that the track estimated with the monocular VO is highly dependent on initialization and reinitializations. Therefore, an inaccuracy in the known object-based pose estimation results in a bias in the orientation estimate of the trajectory. As the last view before a known object detection loss is a grazing view, which is the most degraded and the less accurate case for the known object-based pose estimation, the absolute pose estimated when the pedestrian is in a static phase in front of a known object since a few seconds is preferred to re-initialize the monocular VO. That corresponds to the most accurate case for the known object-based pose estimation. Figure 5 presents an illustration of our global approach for continuous pose estimation with a handheld device in urban environments. It comprises three main stages:

1. The Known Object Detection;
2. The pose estimation;

(a)   The Scaled Monocular Visual Odometry (relative pose estimate).

(b)   The Known Object-Based Pose Estimation (absolute pose estimate).

3.   The AR visualization.
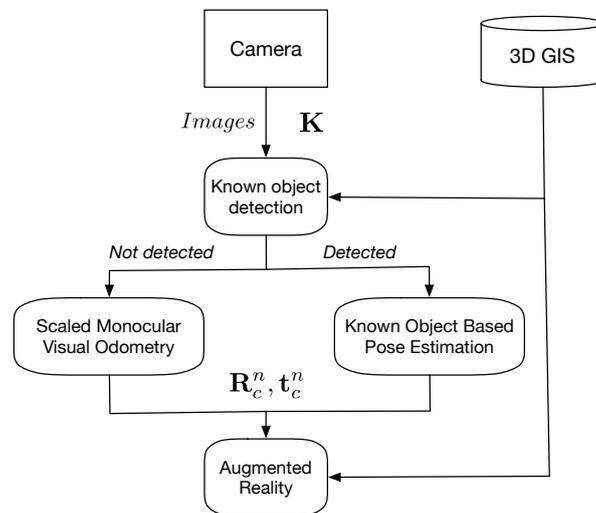
Figure 5. Global approach for continuous pose estimation in urban environments.

4. Experiments

4.1. Hardware Setup

According to the hardware setup needed to develop the proposed approach, the handheld device used in the experiments is composed of a monocular camera and an IMU rigidly attached together, as illustrated in Figure 6.

Figure 6. Hardware setup: the ULISS platform (**below**) and the Garmin VIRB camera (**top**).

This hardware configuration gives access to raw data without the filters commonly applied to mobile device signals. To synchronize the IMU signals and the monocular camera recordings, timestamps from the GPS receivers embedded in both devices were used. The details of the devices are as follows:

A Garmin camera "VIRB 30 Ultra" (https://virb.garmin.com), set up with a fixed focal length, was used for image acquisition. The resolution of the camera was 1920×1080 pixels and was chosen to correspond to a standard resolution of smartphone's acquisition. The video was acquired at 60 Hz frame rate. According to the computational resources in the device, the video was down-sampled at

10 Hz to reduce the computation time. The image resolution could also be resized, but this has not been done for the implementation of our solution.

A dedicated platform named ULISS [48] was used for measurements. It comprises a tri-axis Inertial Measurement Unit and a tri-axis magnetometer sampled at 200 Hz, a barometer, a High Sensitivity GPS receiver and an antenna. They are all low-cost sensors similar to those embedded in mobile devices owned by pedestrians.

4.2. Digital Terrain Model

The DTM, used in the experiments, was computed by the French National Geographical Institute (IGN) (professionnels.ign.fr/mnt). The resolution of the mesh is 1 meter with a decimetric accuracy for the altitude [49]. Other DTM, provided by public data (e.g., Open Street Map, Google Earth, etc.) could also be used. In our implementation, DTM data are processed using the OBJ format.

4.3. Scenario

A 0.7 km walk, which includes passages with an open view and buildings with important specular reflections, was performed by three different people in urban environments with the acquisition device held in hand. Details of the different pedestrians are given in Table 2. The pedestrian activity was in *Texting* mode at all times, i.e., the pedestrian walked while looking at the screen of the handheld device.

Table 2. Acquisition measurements.

Pedestrians	*W1*	*M1*	*W2*
Gender	Female	Male	Male
h_{ped}	1.69 m	1.87 m	1.80 m
Acquisition duration	550 s	486 s	574 s

4.4. Estimation of Reference Waypoints and Tracks

During acquisitions, pedestrians walked on absolute reference waypoints marked on the ground (i.e., with a crossed-out white circle) and made a stop of several seconds over them, as illustrated in Figure 7. Their locations were measured with a centimetric accuracy using a geodetic dual-frequency GNSS receiver (http://www.septentrio.com/) in differential mode. The starting and finishing positions of the acquisition were also determined using a differential GNSS solution.

To assess pedestrian track estimates, reference tracks need to be established during acquisitions. As described in Table 3, in some urban spaces, GNSS-based solutions were not accurate and not available all the time. When the pedestrian walks in an open environment, the standard deviation of the positioning error was less than one metre. When the pedestrian walks in urban canyons, the standard deviation of the positioning error was up to 21 m for differential GNSS and up to 30 m for standalone GPS. These results represent the difficulty of urban location using such sensors, and is far from the accuracy required for an assessment.

Table 3. GNSS measurements.

| Pedestrians	*W1*	*M1*	*W2*
Differential GNSS positioning availability	90.7 %	94.5 %	81.2 %
Standalone GPS positioning availability	86.3 %	78.9 %	54.8 %

Thus, to estimate an accurate and continuous reference track, the PERSY (PEdestrian Reference SYstem) platform [50] was mounted on the foot, on the same side as the handheld device was used during the data acquisition, as illustrated in Figure 7. The PERSY platform outputs location relative to a starting position and claims to have a 0.3% positioning mean error of the distance traveled. Quasi-static

phases of the acceleration and the magnetic field are used to mitigate inertial sensor errors. GNSS phase measurements are added to the strap-down EKF to improve the positioning accuracy.

Figure 7. Handheld device (ULISS + Garmin), foot-mounted PERSY for reference track and reference waypoint.

Figure 8 presents the relative PERSY reference track (in green) and the absolute reference waypoints (in white). Environmental specificities that may introduce difficulties in the monocular VO are also identified (in red).
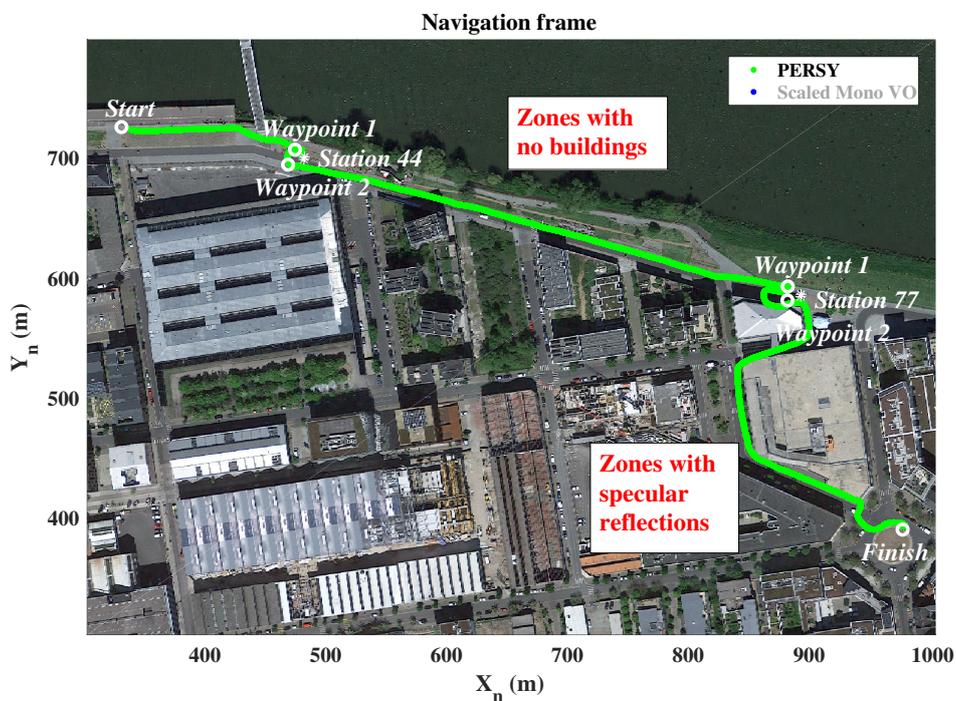

Figure 8. Absolute reference waypoints and relative PERSY reference track of the 0.7 km walk in urban environments.

5. Evaluation

5.1. Activities Classification and Step Length Estimation

Figure 9a presents the results of the activity classification for a walk in urban environments. It can be observed that the walking phases with the device held in hand in *Texting* mode (in green), and the moments when the pedestrian is static on each of the different reference waypoints (in pink) are correctly classified. Figure 9b presents the step lengths estimated, thanks to the walking frequency analysis and the knowledge of the pedestrian's height. It also can be observed that the step frequency and the step lengths are not constant during the walk. This is due to the fact that the pedestrian has to avoid others pedestrians, bikers or cars, and to wait at pedestrian crossings. When the pedestrian walks normally, an amplitude of 40 cm is observed for the step lengths estimation, i.e., mainly between 1 m and 1.4 m. This validates the fact that the use of an average step length is not appropriate for an accurate scale estimation.
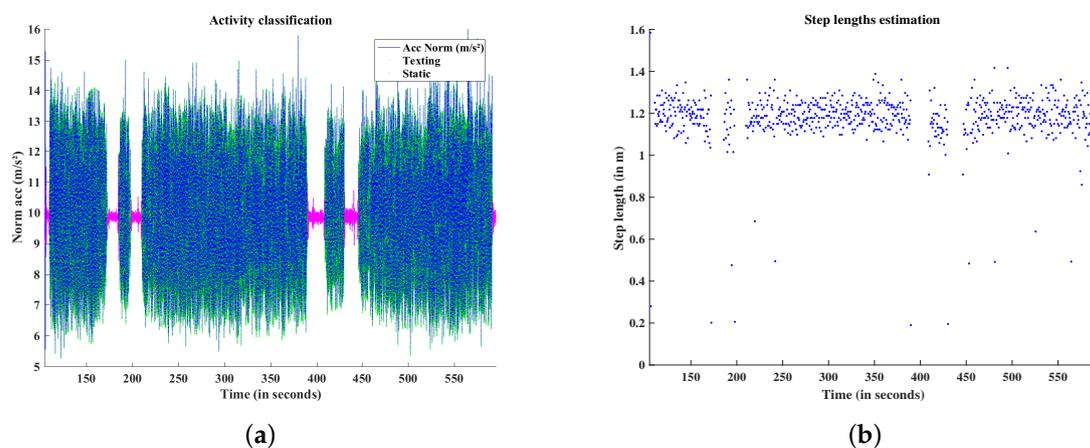
Figure 9. (**a**) Activities classification: *Static* (pink), *Texting* (green); (**b**) Step length estimation.

Table 4 presents the median step length of each pedestrian, as well as the standard deviation of their step lengths.

Table 4. Median step lengths and step lengths standard deviation.

| Pedestrians	*W1*	*M1*	*W2*
Median step length	1.06 m	1.18 m	0.99 m
Step lengths standard deviation	0.20 m	0.14 m	0.29 m

5.2. Estimated Trajectory

To perform the proposed scaled monocular VO, the initial pose in the *Navigation* frame must be estimated. For this process, position from GNSS and orientation from the IMU [51] were used. Then, to punctually suppress the drift and reinitialize the pose in the *Navigation* frame, the known object-based pose estimation proposed in [52] was used.

Figure 10 presents the relative PERSY reference track (in green) and the track estimated with the scaled monocular VO (in blue). From these results, it is observed that the estimated track is close to the relative PERSY reference track. This means that the proposed approach, using adaptive pedestrian step lengths estimates, allows to correctly scaled the monocular VO and to estimate the localization of a device held in hand by pedestrian while walking in urban environments. A visualization of the position estimation is also available on Youtube (https://youtu.be/3bSFrtF2lwU).
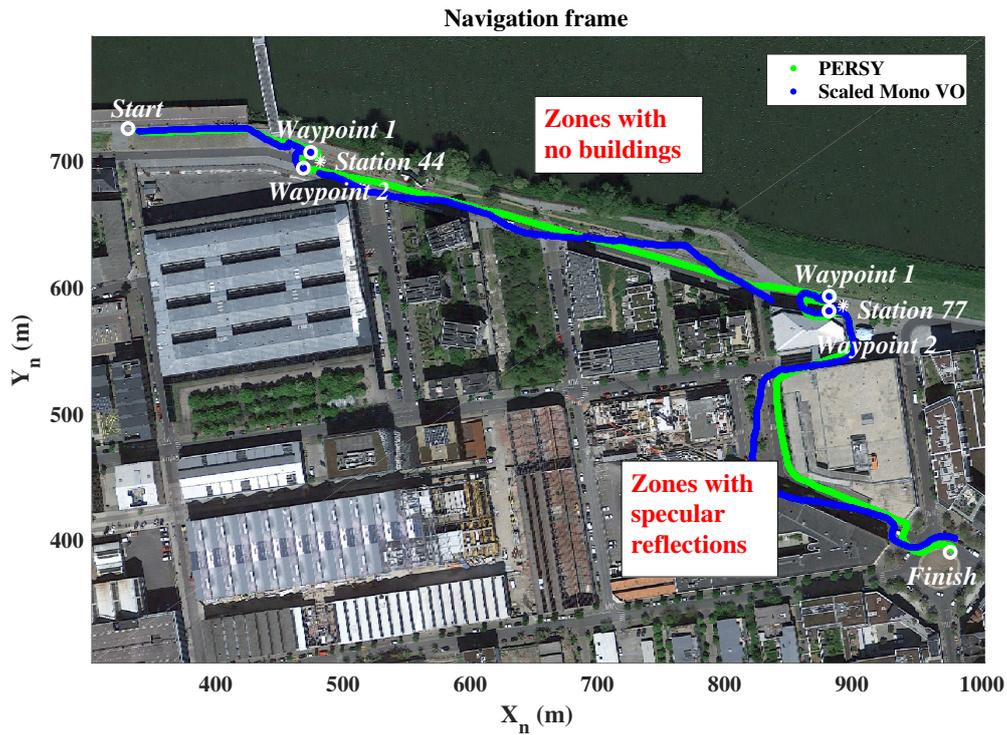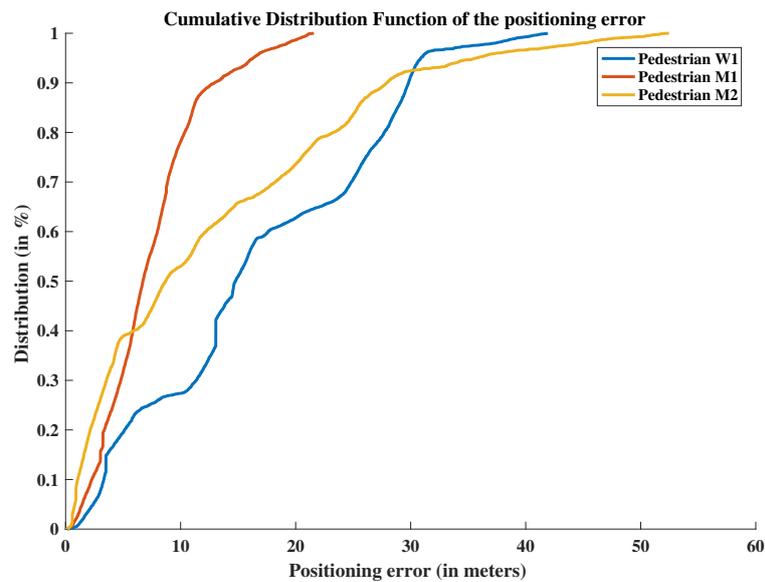
Figure 10. Scaled monocular Visual Odometry track, absolute reference waypoints and relative PERSY reference track of the 0.7 km walk in urban environments.

To assess the performance of the proposed approach, comparisons are made between the positions estimated with the scaled monocular VO and the relative PERSY reference tracks. Figure 11 presents the "Horizontal Positioning Error" and Figure 12 presents the "Cumulative Distribution Function" of the positioning errors for the three different pedestrians.
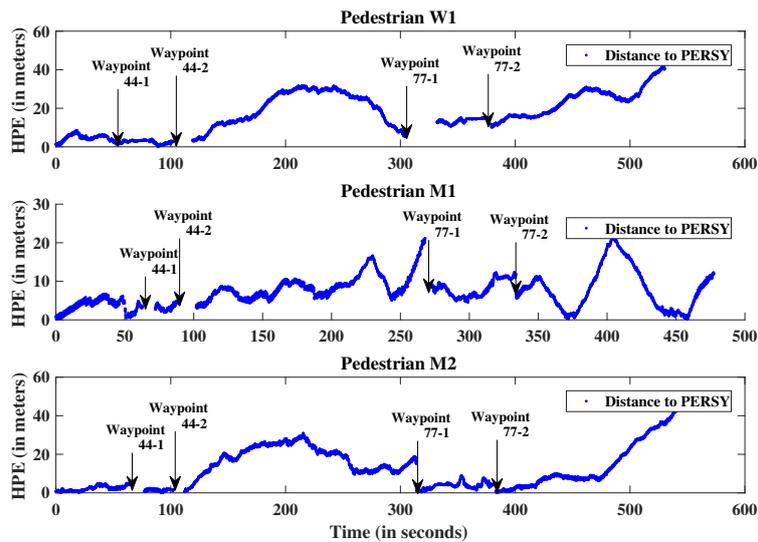
Figure 11. Scaled monocular Visual Odometry Horizontal Positioning Error compared to relative the PERSY reference track.

Figure 12. Scaled monocular Visual Odometry Cumulative Distribution function of the positioning errors compared to relative the PERSY reference track.

Because the relative PERSY reference track is not perfect due to slight inaccuracies that accumulate during the walk, an assessment of the positioning accuracy is also proposed in comparison to absolute reference waypoints. Thus, Table 5 presents the positioning errors between the scaled monocular VO and the absolute reference waypoints before a known object-based reinitialization.

A comparison with a scaling of monocular visual odometry by fixed-step length is also proposed in Table 5. A fixed-step length is defined in relation to the size of each pedestrian. The walking speed is also defined as fixed for all phases where pedestrians walk, and as zero for phases where pedestrians are static. It is observed that the proposed solution allows a positioning error of 1.6–7.5% of the walked distance, where the use of a fixed-step length only allows a positioning error of 7.1–12.7% of the walked distance. This comparison highlights the benefit of the proposed solution compared to the use of a fixed-step length.

Table 5. Scaled monocular Visual Odometry positioning errors and percentage of positioning error with respect to the walked distance compared to absolute reference waypoints and mean positioning error compared to relative the PERSY reference track. Results are given for the use of an adaptive step length (left) and for the use of a fixed-step length (right).

Pedestrians	*W1*	*M1*	*W2*
On waypoint 44-1 (after 120 m)	2.84 m | 27.97 m (2.3% | 23.3%)	3.58 m | 13.17 m (2.9% | 10.9%)	4.46 m | 35.16 m (3.7% | 29.3%)
On waypoint 77-1 (after 550 m)	6.60 m | 101.06 m (1.2% | 18.3%)	21.10 m | 78.74 m (3.8% | 14.3%)	15.11 m | 167.7 m (2.7% | 30.5%)
On finish (after 700 m)	22.43 m | 54.96 m (3.2% | 7.8%)	11.77 m | 49.80 m (1.6% | 7.1%)	52.56 m | 89.19 m (7.5% | 12.7%)
Mean positioning error	16.59 m | 26.58 m	7.33 m | 20.59 m	12.58 m | 45.09 m

As a general assessment of the proposed approach, it can be observed that there are positioning errors during the walk. There are mainly due to the fact that the estimation of the 3D world points coordinates is not perfect in the monocular VO, particularly in urban canyons and open areas, where extracted feature points correspond to elements that are several hundred meters away from the camera. However, these errors can be corrected if a known object is detected. In addition, due to important specular reflections, the implemented monocular VO does not allow to accurately estimate the changes

in the orientation. This results in an error in the orientation estimate, which increases the positioning error at the end of the walk. This could be solved by using more sophisticated VO frameworks such as [53].

It should be noted that when the pedestrian makes small displacements, the proposed approach estimates a smooth trajectory, whereas the use of an average step length would have distorted the scale and the position estimate. Figure 13 presents a focus on an area where the pedestrian makes small displacements. A visualization is proposed on the horizontal plane and in a 3D environment containing 3D models of the known object and buildings around it.

(**a**) (**b**)
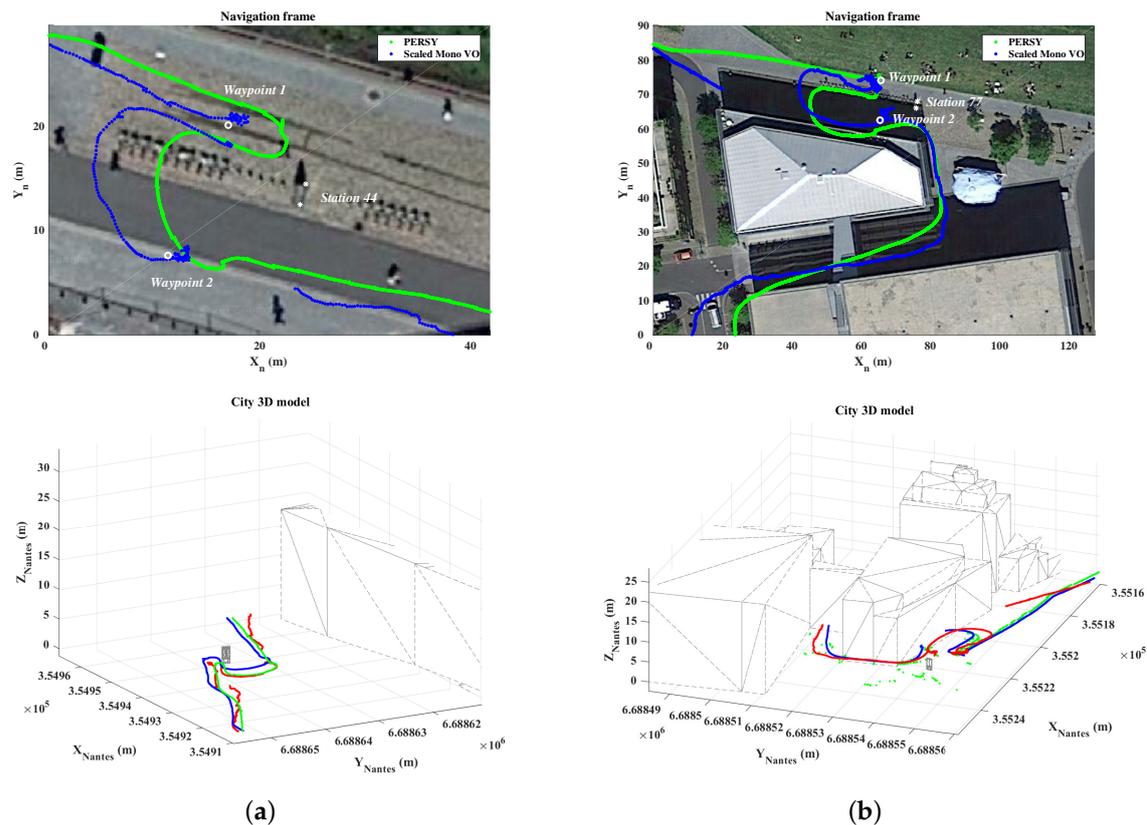
Figure 13. Example of a situation where the pedestrian makes small displacements: (**a**) vicinity of the station 44, (**b**) vicinity of the station 77.

6. Conclusions

In the context of pedestrian navigation, urban environments constitute challenging areas for both localization and handheld device pose estimation. Accurate position and orientation estimation are even more challenging, using only low-cost sensors available in general public devices, i.e., monocular camera and Inertial Measurement Unit.

To address these challenges, we propose a general approach, based on monocular Visual Odometry, to continuously estimate the pose of a handheld device. Our approach does not require any connection or any deployment of new infrastructures. To solve the scale ambiguity and suppress the scale drift in monocular Visual Odometry, an adaptive pedestrian step lengths estimation is used for the displacement on the horizontal plane. To complete the estimation, a handheld equipment height model, with respect to the Digital Terrain Model contained in Geographical Information Systems, is used for the displacement on the vertical axis. In addition, known objects allow to correct the pose estimate and reset the monocular Visual Odometry when one is detected in video frames.

A long walk of about 0.7 km with an IMU, that integrates low-cost sensors, combined with a camera held in hand has been conducted by three different pedestrians in urban environments

with sparse known objects. An assessment is conducted using absolute reference waypoints, whose coordinates have been precisely determined with a Differential GNSS solution in an off-line phase. The assessment is completed by comparing results to a relative reference track obtained with a foot-mounted INS aided by GNSS phase measurements. The proposed approach enables to estimate the pose of a handheld device in urban environments, which is needed for augmented reality applications. Furthermore, this also allows to accurately estimate the pedestrian displacements without any use of GNSS positioning, which strongly deteriorates in urban and indoor environments. A comparison is also proposed between the use of an adaptive step length and the use of a fixed-step length to scale the monocular visual odometry. The proposed solution allows to achieve a positioning error between 1.6% and 7.5% of the walked distance, and confirms the benefit of the proposed solution compared to the use of a fixed-step length.

We plan that part of future works will be dedicated to more accurately estimating the pose, by fusing the presented global approach with PDR and barometric height in a tightly coupling process, mainly to improve the pose estimate in case of handheld device's fast rotational motions. The proposed global approach will also be extended to other known objects in order to reduce the distance between two absolute pose estimates.

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

References

1. Schall, G.; Wagner, D.; Reitmayr, G.; Taichmann, E.; Wieser, M.; Schmalstieg, D.; Hofmann-Wellenhof, B. Global pose estimation using multi-sensor fusion for outdoor Augmented Reality. In Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality, Orlando, FL, USA, 19–22 October 2009; pp. 153–162. [CrossRef]
2. Arth, C.; Klopschitz, M.; Reitmayr, G.; Schmalstieg, D. Real-time self-localization from panoramic images on mobile devices. In Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, Switzerland, 26–29 October 2011; pp. 37–46.
3. Ventura, J.; Arth, C.; Reitmayr, G.; Schmalstieg, D. Global localization from monocular slam on a mobile phone. *IEEE Trans. Visual Comput. Graphics* **2014**, *20*, 531–539. [CrossRef] [PubMed]
4. Zandbergen, P.A.; Barbeau, S.J. Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-enabled Mobile Phones. *J. Navig.* **2011**, *64*, 381–399. [CrossRef]
5. Lategahn, H.; Schreiber, M.; Ziegler, J.; Stiller, C. Urban localization with camera and inertial measurement unit. In Proceedings of the 2013 IEEE Intelligent Vehicles Symposium (IV), Gold Coast, Australia, 23–26 June 2013; pp. 719–724.
6. Yu, L.; Joly, C.; Bresson, G.; Moutarde, F. Improving robustness of monocular urban localization using augmented Street View. In Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016; pp. 513–519.
7. Khan, W.Z.; Xiang, Y.; Aalsalem, M.Y.; Arshad, Q. Mobile Phone Sensing Systems: A Survey. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 402–427. [CrossRef]
8. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [CrossRef] [PubMed]
9. Engel, J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-scale direct monocular SLAM. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2014; pp. 834–849.
10. Mur-Artal, R.; Tardós, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE Trans. Rob.* **2017**, *33*, 1255–1262. [CrossRef]

11. Cummins, M.; Newman, P. Appearance-only SLAM at large scale with FAB-MAP 2.0. *Int. J. Rob. Res.* **2011**, *30*, 1100–1123. [CrossRef]

12. Lynen, S.; Sattler, T.; Bosse, M.; Hesch, J.; Pollefeys, M.; Siegwart, R. Get Out of My Lab: Large-scale, Real-Time Visual-Inertial Localization. In Proceedings of the Robotics: Science and Systems XI, Rome, Italy, 13–17 July 2015. [CrossRef]

13. Jimenez, A.R.; Seco, F.; Prieto, C.; Guevara, J. A comparison of pedestrian dead-reckoning algorithms using a low-cost MEMS IMU. In Proceedings of the 2009 IEEE International Symposium on Intelligent Signal Processing, Budapest, Hungary, 26–28 August 2009; pp. 37–42.

14. Fischer, C.; Sukumar, P.T.; Hazas, M. Tutorial: Implementing a pedestrian tracker using inertial sensors. *IEEE Pervasive Comput.* **2013**, *12*, 17–27. [CrossRef]

15. Jiang, B.; Neumann, U.; You, S. A robust hybrid tracking system for outdoor augmented reality. In Proceedings of the IEEE Virtual Reality 2004, Chicago, IL, USA, 27–31 March 2004; pp. 3–11. [CrossRef]

16. Beich, M.J.; Veth, C.M. Tightly coupled image-aided inertial relative navigation using Statistical Predictive Rendering (SPR) techniques and a priori world Models. In Proceedings of the IEEE/ION Position, Location and Navigation Symposium, Indian Wells, CA, USA, 4–6 May 2010; pp. 552–560. [CrossRef]

17. Oskiper, T.; Samarasekera, S.; Kumar, R. Tightly coupled robust vision aided inertial navigation algorithm for augmented reality using monocular camera and IMU. In Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, Switzerland, 26–29 October 2011. [CrossRef]

18. Oskiper, T.; Samarasekera, S.; Kumar, R. [POSTER] CamSLAM: Vision Aided Inertial Tracking and Mapping Framework for Large Scale AR Applications. In Proceedings of the 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct), Nantes, France, 9–13 October 2017; pp. 216–217. [CrossRef]

19. You, S.; Neumann, U.; Azuma, R. Orientation tracking for outdoor augmented reality registration. *IEEE Comput. Graph. Appl.* **1999**, *19*, 36–42. [CrossRef]

20. Satoh, K.; Anabuki, M.; Yamamoto, H.; Tamura, H. A hybrid registration method for outdoor augmented reality. In Proceedings of the IEEE and ACM International Symposium on Augmented Reality, New York, NY, USA, 29–30 October 2001; pp. 67–76. [CrossRef]

21. Nistér, D.; Naroditsky, O.; Bergen, J. Visual odometry. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004.

22. Scaramuzza, D.; Fraundorfer, F. Visual odometry [tutorial]. *IEEE Rob. Autom Mag.* **2011**, *18*, 80–92. [CrossRef]

23. Aqel, M.O.A.; Marhaban, M.H.; Saripan, M.I.; Ismail, N. Review of visual odometry: Types, approaches, challenges, and applications. In *SpringerPlus*; Springer: Cham, Switzerland, 2016.

24. 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; pp. 225–234. [CrossRef]

25. Strasdat, H.; Montiel, J.; Davison, A.J. Scale drift-aware large scale monocular SLAM. In *Robotics: Science and Systems VI*; MIT Press: Cambridge, MA, USA, 2010.

26. Wang, R.; Schwörer, M.; Cremers, D. Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

27. Botterill, T.; Mills, S.; Green, R. Correcting Scale Drift by Object Recognition in Single-Camera SLAM. *IEEE Trans. Cybern.* **2013**, *43*, 1767–1780. [CrossRef] [PubMed]

28. Lothe, P.; Bourgeois, S.; Royer, E.; Dhome, M.; Naudet-Collette, S. Real-time vehicle global localisation with a single camera in dense urban areas: Exploitation of coarse 3D city models. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 863–870. [CrossRef]

29. Gakne, P.V.; O'Keefe, K. Tackling the Scale Factor Issue in A Monocular Visual Odometry Using A 3D City Model. In Proceedings of the ITSNT 2018, International Technical Symposium on Navigation and Timing, Toulouse, France, 13–16 November 2018. [CrossRef]

30. Weiss, S.; Siegwart, R. Real-time metric state estimation for modular vision-inertial systems. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 4531–4537.

31. Nützi, G.; Weiss, S.; Scaramuzza, D.; Siegwart, R. Fusion of IMU and vision for absolute scale estimation in monocular SLAM. *J. Intell. Rob. Syst.* **2011**, *61*, 287–299. [CrossRef]

32. Knorr, S.B.; Kurz, D. Leveraging the User's Face for Absolute Scale Estimation in Handheld Monocular SLAM. In Proceedings of the 2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Merida, Mexico, 19–23 September 2016; pp. 11–17.

33. Lupton, T.; Sukkarieh, S. Removing scale biases and ambiguity from 6DoF monocular SLAM using inertial. In Proceedings of the 2008 IEEE International Conference on Robotics and Automation, Pasadena, CA, USA, 19–23 May 2008; pp. 3698–3703. [CrossRef]

34. Gutiérrez-Gómez, D.; Guerrero, J.J. Scaled monocular SLAM for walking people. In Proceedings of the 2013 International Symposium on Wearable Computers, Zurich, Switzerland, 8–12 September 2013; pp. 9–12.

35. Leppäkoski, H.; Collin, J.; Takala, J. Pedestrian navigation based on inertial sensors, indoor map, and WLAN signals. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 1569–1572. [CrossRef]

36. Tomazic, S.; Skrjanc, I. Fusion of visual odometry and inertial navigation system on a smartphone. *Comput. Ind.* **2015**, *74*, 119–134. [CrossRef]

37. Renaudin, V.; Susi, M.; Lachapelle, G. Step Length Estimation Using Handheld Inertial Sensors. *Sensors* **2012**, *12*, 8507–8525. [CrossRef] [PubMed]

38. Susi, M.; Renaudin, V.; Lachapelle, G. Motion Mode Recognition and Step Detection Algorithms for Mobile Phone Users. *Sensors* **2013**, *13*, 1539–1562. [CrossRef] [PubMed]

39. Taketomi, T.; Uchiyama, H.; Ikeda, S. Visual SLAM algorithms: A survey from 2010 to 2016. *IPSJ Trans. Comput. Vis. Appl.* **2017**, *9*, 16. [CrossRef]

40. Bouguet, J.Y. Camera Calibration Toolbox for Matlab, 2008. Available online: http://www.vision.caltech.edu/bouguetj/calib_doc/ (accessed on 22 February 2019).

41. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*; Cambridge University Press: Cambridge, UK, 2003.

42. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [CrossRef]

43. Bay, H.; Ess, A.; Tuytelaars, T.; Gool, L.V. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]

44. Torr, P.; Zisserman, A. MLESAC. *Comput. Vis. Image Underst.* **2000**, *78*, 138–156. [CrossRef]

45. Wu, Y. Image Based Camera Localization: An Overview. *arXiv* **2016**, arXiv:1610.03660.

46. Cai, J.; Miklavcic, S. The generation of digital terrain models from LiDAR data using seeding and filtering and its application to flood modelling. In Proceedings of the 2011 International Conference on Multimedia Technology, Hangzhou, China, 26–28 July 2011; pp. 5622–5625. [CrossRef]

47. Antigny, N.; Servières, M.; Renaudin, V. Fusion of 3D GIS, Vision, Inertial and Magnetic Data for Improved Urban Pedestrian Navigation and Augmented Reality Applications. *Navigation* **2018**, *65*, 431–447. [CrossRef]

48. Ortiz, M.; De Sousa, M.; Renaudin, V. A New PDR Navigation Device for Challenging Urban Environments. *J. Sens.* **2017**, *2017*, 4080479. [CrossRef]

49. Signorelli, V.; Leduc, T. *Utilisation du Socle 3D CityGML Nantes-Secteur Centre-Ville*; Technical Report; UMR CNRS 1563 AAU/CRENAU: Nantes, France, 2015.

50. Scornec, J.L.; Ortiz, M.; Renaudin, V. Foot-mounted pedestrian navigation reference with tightly coupled GNSS carrier phases, inertial and magnetic data. In Proceedings of the 2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Sapporo, Japan, 18–21 September 2017; pp. 1–8. [CrossRef]

51. Renaudin, V.; Combettes, C. Magnetic, Acceleration Fields and Gyroscope Quaternion (MAGYQ)-Based Attitude Estimation with Smartphone Sensors for Indoor Pedestrian Navigation. *Sensors* **2014**, *14*, 22864–22890. [CrossRef] [PubMed]

52. Antigny, N.; Servieres, M.; Renaudin, V. Pedestrian Track Estimation with Handheld Monocular Camera and Inertial-Magnetic Sensor for Urban Augmented Reality. In Proceedings of the 2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Sapporo, Japan, 18–21 September 2017.

53. Engel, J.; Sturm, J.; Cremers, D. Semi-dense visual odometry for a monocular camera. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 1449–1456.