An Indoor Positioning Framework Based on Panoramic Visual Odometry for Visually Impaired People

Weijian Hu, Kaiwei Wang, Hao Chen, Ruiqi Cheng, Kailun Yang

Zhejiang University, Zheda Road, Hangzhou, China

E-mail: wangkaiwei@zju.edu.cn

May 2019

Abstract. Positioning is one of the most urgent problems for assisting visually impaired people, especially in indoor environments where GPS signals are weak. In this paper, we present a positioning framework based on panoramic visual odometry for assisting visually impaired people. We introduce panoramic annular lens to visual odometry and use the Taylor camera model to describe its projection rules. Some critical techniques in visual odometry, including direct image alignment and stereo matching, are extended to fit this camera model. Besides, an easy-to-maintain coordinate alignment method based on multi-pinhole image rectification and marker detection is also proposed to unify the results of visual odometry in the world coordinate system. We evaluate our system on both synthetic and real-world datasets in a comparative set of experiments and compare against state-of-the-art algorithms. The experiment results show that the robustness of positioning has been significantly improved by the proposed visual odometry algorithm with panoramic annular lens and the system has the ability to provide reliable positioning results in indoor environments.

Keywords: Visual Odometry, Panoramic Annular Lens, Blind Assistance

Submitted to: *Meas. Sci. Technol.*

1. Introduction

In the world, there are estimated 253 million people suffering from severe vision impairments and blindness (Bourne et al.; 2017). One of the most urgent problems faced by Visually Impaired People (VIP) lies in how to determine their position correctly, especially in unfamiliar environments. Thanks to the development of the Global Positioning System (GPS) and the Geographic Information System (GIS), VIP can locate themselves in the open air and go anywhere with the help of a smart phone. But in indoor environments, GPS signals are usually weak, and how to help VIP locate themselves indoors is still a challenging problem.

Visual Odometry (VO) is one of the promising positioning methods to solve this problem. It has been widely used in the field of robotics, autonomous driving and etc. (Huang et al.; 2017; An et al.; 2017; Dimas et al.; 2017). VO determines the position and orientation of a camera by tracking scene features (e.g. point features and line features) in a sequence of images. Thus, a challenging issue of VO in indoor environments lies in the lack of texture in images. For example, if a camera shoots a white wall, most part of the image will be occupied by the wall, and there are not enough textures that can be utilized to estimate the motion of the camera. Another problem is related to the dynamic objects in real-world scenarios, such as pedestrians, which introduces unstable features into images and finally affects the positioning accuracy.

In this paper, we present a panoramic ceiling-view positioning framework based on Panoramic Annular Lens (PAL) to help VIP locate themselves in indoor environments. PAL is a kind of panoramic optical system, as shown in figure 1(a)-(b). It utilizes two reflective surfaces to achieve wide Field of View (FoV), typically near 180 degrees. Thus the commonly used pinhole camera model cannot describe its projection rules correctly. To solve this problem, we make use of the Taylor camera model (Scaramuzza; 2007), a kind of unified omnidirectional camera model, to describe the projection rules between 3D world and 2D images. For the VO algorithm, we extend DSO (Engel et al.; 2017), a state-of-the-art monocular VO algorithm, to fit the Taylor camera model. Moreover, a coordinate transformation method based on ArUco markers (Romero-Ramirez et al.; 2018; Garrido-Jurado et al.; 2016) and multi-pinhole rectification is also proposed to transform VO results to the world coordinate system. A proof of concept prototype is also designed with a PAL system mounted on a helmet, as shown in figure 1(c).

The benefits of using PAL as the sensor of a ceiling-view VO in indoor positioning are manifold. First, the camera is mounted on the head and the FoV of PAL is about 180 degrees. This means the camera can capture all information above the user, including the entire ceiling, as well as part of walls and doors, which can provide enough stable textures for positioning. At the same time, most dynamic objects on the ground will not be captured by the camera, which makes our VO algorithm more robust. Second, compared with conventional fisheye lenses and catadioptric imaging systems, PAL has compact size, small f-theta distortion and is easier to design and manufacture (Huang et al.; 2013), which makes our wearable devices more compact and cost-effective.

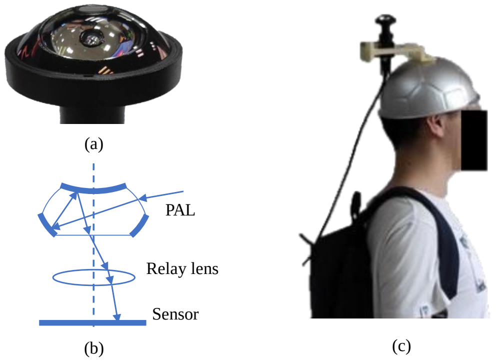
[redacted]

Figure 1. Panoramic annular lens and the proposed prototype. (a) The PAL used in our system. (b) PAL imaging principle. A ray is reflected twice in the PAL block before reaching the relay lens. (c) The prototype of a head-mounted PAL.

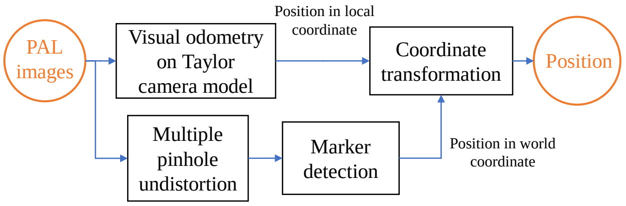
PAL images
Visual odometry on Taylor camera model
Position in local coordinate
Coordinate transformation
Position
Multiple pinhole undistortion
Marker detection
Position in world coordinate

Figure 2. The schematic diagram of the positioning framework.

The schematic of our algorithm is shown in figure 2. Input PAL images are sent to two modules and processed simultaneously. Visual odometry analyses the image sequence and outputs camera positions in the local coordinate system where the origin is the first image frame and the unit is uncertain due to the observability of monocular camera. Direct image alignment, stereo matching and sliding window optimization are utilized to calculate and refine 6-DoF camera poses as well as the depth of sparse points. Meanwhile, PAL images are also rectified by the multi-pinhole model to generate four narrow FoV pinhole images. ArUco marker is detected on these undistorted pinhole images and the 6-DoF camera pose is estimated in the world coordinate system, where the origin is the detected marker and the unit is meter. The marker can be seen as the anchor of the building. Once enough camera poses are calculated both in the world coordinate system and the local coordinate system, we can estimate the 3D similarity transformation (rotation, translation and scale change) from local coordinate system to world coordinate system through a least squared method. After that, all positions produced by VO can be transformed to the world coordinate system for positioning and assisting VIP. The contribution of this paper are summarized as follows:

- In order to increase the robustness of VO, we introduce PAL to VO and perform direct image alignment and stereo matching based on Taylor camera model.
- To address the issue that VO results do not match the world coordinate system, an easy-to-maintain coordinate alignment method based on multi-pinhole rectification and marker detection is proposed.
- Our implementation of wearable prototype is evaluated in real world environments and the experiments demonstrate the effectiveness of our positioning framework.

The rest of this paper is organized as follows. Section 2 reviews the relevant literature on indoor pedestrian positioning. Section 3 introduces the proposed VO method for PAL. Section 4 describes the multi-pinhole rectification and the transformation of local coordinate system and world coordinate system. In section 5, a comprehensive set of experiments on both synthetic and real-world datasets are performed to evaluate the performance of our positioning framework. Finally, the conclusions and future work are presented in section 6.

2. Related Works

In recent years, with the development of sensors and mobile computing, a wide variety of portable navigation systems have been proposed to assist VIP to avoid obstacles (Dimas et al.; 2019; Yang et al.; 2016), navigate (Jayakody et al.; 2020; Donati et al.; 2020), and perceive the environment (Iakovidis et al.; 2020; Yang et al.; 2018). Positioning plays an important role in assisting VIP. Accurate positioning is the premise of navigation and other scene perception functions. In the literature, many works have been proposed to assist VIP in positioning in indoor environments. According to the type of sensors, the technique of indoor positioning falls into two categories: inertial positioning and visual positioning.

Inertial sensors-based indoor positioning systems usually take use of inertial sensors (gyroscope and accelerometer) to estimate the movement and use radio beacons, such as WiFi and bluetooth, to correct drifts and recognize key points, such as entrances and elevators (Harle; 2013). In 2018, M. Murata et al. presented an indoor positioning algorithm based on inertial sensors in a mobile phone and Bluetooth Low Energy (BLE) beacons (Murata et al.; 2018). They used particle filter to build a probabilistic representation of positions and proposed some techniques to make their algorithm perform better in multi-storey building-scale environments. An experiment was conducted in a shopping mall with four floors and 218 pre-installed BLE beacons to evaluate the system.

Another clusters of research use camera as main sensor to realize indoor positioning for VIP. In 2016, Lee and Medioni presented an RGB-D camera based wearable indoor navigation system for VIP (Lee and Medioni; 2016). They adapted FOVIS (Huang et al.; 2017), a feature-based 6-DoF motion estimation algorithm, to estimate camera poses and used point clouds alignment to refine poses. Two dimensional occupancy grid map and dynamic path planning were implemented to achieve point-to-point indoor navigation. A smart phone application and a tactile feedback vest were also developed as user interface. Their system included multiple functions, including navigation, obstacle avoidance, etc. and can provide comprehensive assistance to VIP without a prior map. However, their positioning method did not perform well in the experiment, with an error of 2.6 meters after 49 meters walking, which limits its usability. ISANA is another types of VO-based indoor navigation system that requires building map input first (Li et al.; 2016). It was developed on Google Project Tango device and used built-in visual inertial odometry to estimate camera poses. A map alignment algorithm based on semantic landmark (e.g. room num) was proposed to bridge the position among building map and camera poses. Obstacle detection and avoidance function was also designed and a field test was conducted to demonstrate the effectiveness of the system. However, their map alignment algorithm (can be regarded as the initialization of the entire system) needed several successful room number recognitions, which may take a long time in actual use.

Some studies also take advantage of panoramic cameras to enhance the robustness of indoor positioning. In 2012, A. C. Murillo et al. proposed a personal positioning system using a head-mounted omnidirectional camera (Murillo et al.; 2012). It used extended Kalman filter to achieve a VO. Limited by the development of VO algorithm at that time, their positioning accuracy suffers from large scale and rotation drifts. F. Hu et al. presented a localization system based on panoramic images indexing technique to help VIP navigate in indoor environments (Hu et al.; 2015). They extracted features of current images and searched it in a pre-created image feature database to find the location and the orientation. The time-consuming search process was done on a cloud server with GPU parallel acceleration to obtain real-time performance. However, this method needs to build a dataset of images-position pairs, which usually takes a long time. In scenes with repeating textures, such as corridors, this appearance-based method is more likely to fail.

3. Direct Odometry for Panoramic Annular Lens

In this section, we will introduce the VO framework briefly and elaborate how it runs on Taylor camera model. Our algorithm is an extension of DSO (Engel et al.; 2017) and the pipeline is shown in figure 3. Initialization is the first step of the proposed system. In this step, points with high gradient on the first frame are selected and the initial depth of these points are estimated by direct image alignment with following frames. After a successful initialization, a fix-size keyframe sliding window is created immediately and the first frame is pushed into the sliding window as the first keyframe. Every keyframe in the sliding window contains a 6-DoF camera pose in the local coordinate system and maintains some points with depth. The sliding window is the core of the system and all following steps are based on it. When a new PAL frame is captured, its 6-DoF pose relative to the newest keyframe is estimated by using the direct image alignment. Then stereo matching is performed to refine the depth on newest keyframe. If the movement between the current frame and the newest keyframe is larger than a threshold, the current frame is converted to a keyframe and pushed into the sliding window. All poses and points of keyframes in the sliding window are optimized jointly. A marginalization step is also executed to keep the number of keyframes less than the size of the sliding window.

Notations. We denote vectors ($\mathbf{t}$) and matrices ($\mathbf{R}$) with bold letters and scalars (a) with light letters. Camera poses are represented in two equivalent ways. For coordinate transformation, 4×4 matrices of the 3D Special Euclidean group $\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \in \mathrm{SE}(3)$ is used, where $\mathbf{R}$ is the 3×3 rotation matrix and $\mathbf{t}$ is the 3×1 translation vector respectively. For optimization, a minimal representation is required and we use 6×1 vectors of the Lie algebra corresponding to the Special Euclidean group $\boldsymbol{\xi} \in \mathfrak{se}(3)$ to represent camera poses. These two representations are connected through the exponential map and the logarithmic map (Engel et al.; 2014).

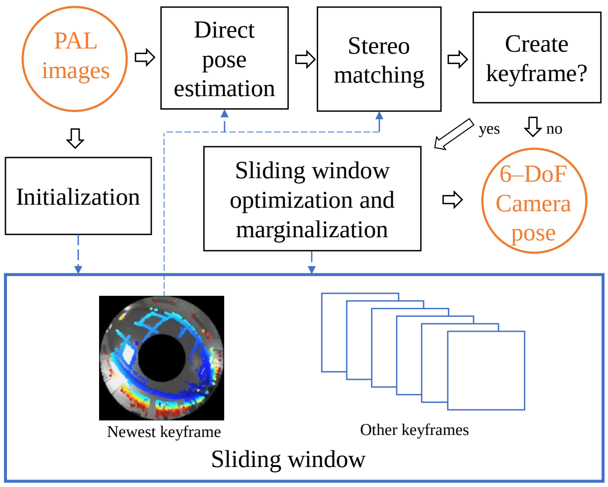

Figure 3. The schematic diagram of our visual odometry.

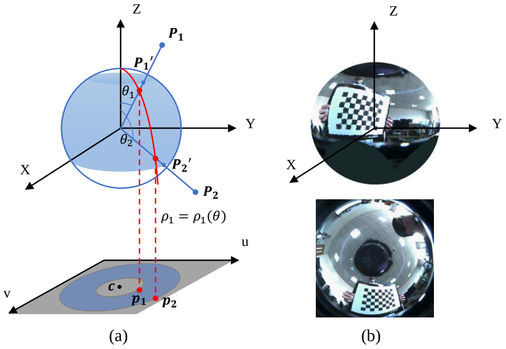

Figure 4. Taylor camera model. (a) The projection flow. The blue area indicates the effective FoV of PAL and the red curve is the visualization of the polynomial function. (b) Mapping an PAL image onto the unit sphere based on the camera calibration results.

3.1. Taylor camera model

Camera models describe the mathematical relationship between a 3D point in the world and its projection on the image. The pinhole camera model is the most widely used camera model for ordinary cameras. It projects a point in 3D space onto the image plane through an ideal pinhole. However, if the FoV is larger than 180 degrees, the pinhole model is no longer suitable. Thus we use the Taylor model, a kind of unified sphere projection model proposed by Scaramuzza (Scaramuzza; 2007) to describe the imaging formula.

As shown in figure 4(a), a 3D point $\mathbf{P_1} = [x, y, z]^T$ in the camera coordinate system is projected onto a pixel $\mathbf{p_1} = [u, v]^T$ in the image coordinate by three steps. First, $\mathbf{P_1}$ is normalized onto a unit sphere, denoted as $\mathbf{P_1'}$. Then, a polynomial factor $\rho(\theta)$ is multiplied to change the norm of $\mathbf{P_1'}$, and finally it is directly projected on the image plane with a bias $\mathbf{c} = [c_x, c_y]^T$ added. The complete mathematical expression to project a 3D point can be written as:

$$\mathbf{p} = \Pi(\mathbf{P}) = \begin{bmatrix} \frac{x}{\sqrt{x^2+y^2+z^2}}\rho_1(\theta) \\ \frac{y}{\sqrt{x^2+y^2+z^2}}\rho_1(\theta) \end{bmatrix} + \begin{bmatrix} c_x \\ c_y \end{bmatrix} \tag{1}$$

where

$$\theta = atan(\frac{\sqrt{x^2+y^2}}{z}) \tag{2}$$

$$\rho_1(\theta) = a_0 + a_2\theta^2 + a_3\theta^3 + ... + a_N\theta^N \tag{3}$$

Note that the a_1 is constant equal to zero due to the symmetry of lens. The unprojection function is

$$\mathbf{P} = \Pi^{-1}(\mathbf{p}, d_p) = d_p \begin{bmatrix} (u - c_x)/\rho_2(r) \\ (v - c_y)/\rho_2(r) \\ 1 \end{bmatrix} \tag{4}$$

where

$$r = \sqrt{u^2 + v^2} \tag{5}$$

$$\rho_2(r) = b_0 + b_2r^2 + b_3r^3 + ... + b_Nr^N \tag{6}$$

It is noteworthy that the polynomial functions ρ_1, ρ_2 used in projection and unprojection are different, but the meaning of ρ is the same, i.e. the relation between camera coordinates and pixel coordinates. All parameters that need to calibrate are c_x, c_y and coefficients of ρ_1 and ρ_2. According to Scaramuzza's experiment (Scaramuzza; 2007), fourth-order polynomial is an appropriate choice to balance the computational complexity and precision.

We use OCamCalib toolbox ‡, an omnidirectional camera calibration toolbox for Matlab, to calibrate the parameters of the Taylor camera model (Scaramuzza et al.; 2006). Ten PAL images with checkerboard are captured for calibration and the calibration process include four steps:

- Extract corner points of the checkerboard.
- Calibrate the coefficients of ρ by a least-square method.
- Estimate the optical center c_x and c_y through a iterative searching method.
- Optimize all parameters globally to decrease the reprojection error.

Figure 4(b) shows a PAL calibration image and visualizes the performance of projecting an PAL image onto a unit sphere according to the calibration results.

3.2. Direct image alignment

The direct image alignment is used in initialization and direct pose estimation stages to estimate the pose between two images. We use the newest keyframe in the sliding window as a reference frame $\mathbf{I}_i$, which includes a pose $\mathbf{T}_i$ and N_p points with depth. The pose $\mathbf{T}_j$ of a following frame $\mathbf{I}_j$ can be estimated by directly minimizing the photometric error of two frames, formulated as

$$\hat{\mathbf{T}}_j = \underset{\hat{\mathbf{T}}_j}{\operatorname{argmin}} \sum_{\mathbf{p}\in\mathbf{I}_i} w_p \left\| \mathbf{I}_j[\mathbf{p}'] - \frac{t_j}{t_i}\mathbf{I}_i[\mathbf{p}] \right\|_\gamma \tag{7}$$

where $\mathbf{p}$ and $\mathbf{p}'$ are pixels on $\mathbf{I_i}$ and $\mathbf{I_j}$. Before calculating the photometric error between $\mathbf{p}$ and $\mathbf{p}'$, the exposure time of frames are also considered, denoted as t_i and t_j. Besides, w_p is a gradient dependent weight to reduce the influence of low gradient points and $\|\cdot\|_\gamma$ denotes the Huber norm. $\mathbf{p}'$ is the reprojection point of $\mathbf{p}$ based on the its depth d_p and the camera model that described in section 3.1, given by

$$\mathbf{p}' = \Pi(\mathbf{R}\Pi^{-1}(\mathbf{p}, d_p) + \mathbf{t}) \tag{8}$$

where

$$\begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} = \mathbf{T}_j\mathbf{T}_i^{-1} \tag{9}$$

We use Levenberg-Marquad method to solve this weighted non-linear least-squares problem. The $\mathfrak{se}(3)$ increment $\delta\xi$ is computed by the following equation:

$$(\mathbf{J^TWJ} + \lambda\mathbf{I})\delta\xi = \mathbf{J^TWr} \tag{10}$$

where $\mathbf{r}$ is the stacked residual vector of N_p tracked points and $\mathbf{W}$ is a diagonal matrix of weights w_p. $\mathbf{J}$ is a $N_p \times 6$ jacobian matrix of stacked residuals. For each row of $\mathbf{J}$, it can be decomposed as

$$\mathbf{J} = \left[\frac{\partial \mathbf{I}_j}{\partial \mathbf{p}'}\right]_{1\times2} \cdot \left[\frac{\partial \mathbf{p}'}{\partial \mathbf{P}'}\right]_{2\times3} \cdot \left[\frac{\partial \mathbf{P}'}{\partial \xi}\right]_{3\times6} \tag{11}$$

where $[\cdot]_{m\times n}$ denotes the dimension of matrix, and

- $\frac{\partial \mathbf{I}_j}{\partial \mathbf{p}'}$ is the gradient of image $\mathbf{I}_j$ at point $\mathbf{p}'$.
- $\frac{\partial \mathbf{p}'}{\partial \mathbf{P}'}$ is the jacobian of the projection function $\Pi(P)$.
- $\frac{\partial \mathbf{P}'}{\partial \xi}$ is the jacobian of left-compositional derivative.

Some tricks are employed to make the direct image alignment run in real time. First, we use a coarse-to-fine method in optimization: creating an image pyramid, and then optimizing variables from low resolution images to high resolution Images gradually (Engel et al.; 2014). This trick can reduce the number of iterations and the optimization will converge quickly. Second, some key parameters are selected carefully to make the computation of error and jacobian matrix

‡ https://sites.google.com/site/scarabotix/ocamcalib-toolbox

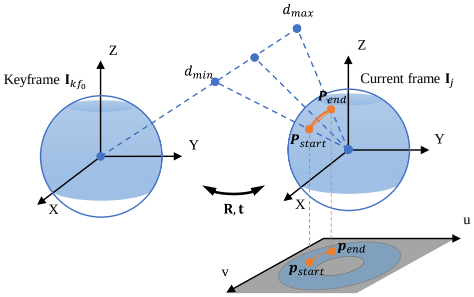

Figure 5. Stereo matching. The left sphere indicates the newest keyframe and the right sphere indicates the current frame. The epipolar curve segments on units sphere and image are shown in the orange curves.

not too slow. For example, N_p is set to 2000, so up to 2000 points will be initialized and included in the calculation.

Note that the initial depth of points are also jointly optimized in the initialization. Moreover, a brightness affine transformation is performed before calculating the photometric error. These details can be further referenced to DSO (Engel et al.; 2017).

3.3. Stereo matching

After estimating the pose of the current image $\mathbf{I}_j$, we can improve the accuracy of depth on the newest keyframe $\mathbf{I}_{kf_0}$.The problem can be regarded as a stereo matching problem: $\mathbf{I}_{kf_0}$ and $\mathbf{I}_j$ are the images of the left and right camera respectively and the rotation $\mathbf{R}$ and translation $\mathbf{t}$ between them is known. We need to match points on $\mathbf{I}_{kf_0}$ to points on $\mathbf{I}_j$ and reconstruct the depth by triangulation.

For pinhole camera model, the best match point can be found through searching along the epipolar line segment on $\mathbf{I}_j$ (Vogiatzis and Hernández; 2011). But for Taylor camera model, the epipolar line is a curve segment, as shown in figure 5. As the epipolar curve segment is usually short, we split it into many short segments to solve this issue.

The depth of a point on $\mathbf{I}_{kf_0}$ is parameterized by min inverse depth d_{min} and max inverse depth d_{max}. We first calculate the start and end points of epipolar curve segment on the unit sphere, denoted by $\mathbf{P}_{min}$ and $\mathbf{P}_{max}$. The mathematical description of epipolar curve on the unit sphere can be expressed by the interpolation of $\mathbf{P}_{min}$ and $\mathbf{P}_{max}$, given by

$$\mathbf{P}_L(\alpha) = \alpha \mathbf{P}_{max} + (1-\alpha)\mathbf{P}_{min} \qquad (12)$$

with $\alpha \in [0,1]$. and the epipolar curve segment on the image is

$$\mathbf{p}_L(\alpha) = \Pi(\mathbf{P}_L(\alpha)) \qquad (13)$$

We start searching the best match pixel at $\mathbf{p}_L(0)$ and increase α gradually. Sum of Squared Differences (SSD) over eight equidistant pixels is used as the matching cost. To search the epipolar curve segment pixel by pixel, the increment $\delta\alpha$ is approximately set to $1/|\mathbf{p}_L(1) - \mathbf{p}_L(0)|$, and this approximation is good in practice.

After searching the best matching point, we use triangulation process to calculate the depth of point on $\mathbf{I}_{kf_0}$. Let $\mathbf{p}_{kf_0}$ and $\mathbf{p}_j$ are the matched points on $\mathbf{I}_{kf_0}$ and $\mathbf{I}_j$, the inverse depth of point d_{kf_0} and d_j can be calculated by solving the following equation:

$$\frac{1}{d_{kf_0}} \cdot \Pi^{-1}(\mathbf{p_{kf_0}}) = \frac{1}{d_j} \cdot \mathbf{R}\Pi^{-1}(\mathbf{p_j}) + \mathbf{t} \qquad (14)$$

The uncertainty of depth $\sigma_{d_{kf_0}}$ is also estimated by assuming we have 1 pixel error in point matching (Engel et al.; 2013). We only keep $d_{min} = d_{kf_0} - \sigma_{d_{kf_0}}$ and $d_{max} = d_{kf_0} + \sigma_{d_{kf_0}}$ to parameterize the depth and use them for the stereo matching of the next frame.

3.4. Sliding window optimization

Sliding window optimization is an effective technique to reduce error and keep scale consistency. As we follow the optimization method in (Engel et al.; 2017), so here we briefly introduce the basic flow.

If the change of movement, rotation and exposure time is larger than a threshold, or the photometric error of direct pose estimation is too large, the current frame will be selected as a keyframe and be inserted into the sliding window. Then some pixels on the new keyframe with high gradient will be selected as tracked points. A joint optimization is performed to minimize the photometric error over all keyframes in the sliding window and all camera poses as well as the depth of points will be adjusted. If the number of keyframes in the window exceeds N_f (e.g. $N_f = 7$ in our method), a marginalization process will be done to remove a keyframe from the window to balance the computational complexity and precision.

4. Positioning in world coordinate

The results of monocular VO are in the local coordinate system and the unit is also indeterminable. Thereby, the real-time output of VO cannot be used for navigation directly. We use the ArUco marker (Garrido-Jurado et al.; 2016) to estimate the rotation, translation, as well as the scale change between the local coordinate system and the world coordinate system in the building. First, a PAL image is rectified to four virtual pinhole images. Then, ArUco marker detection is performed on four pinhole images respectively. If a marker is successfully detected, the pose in the marker coordinate system can be estimated by calculating and decomposing the homography matrix. Once enough camera poses in both the local

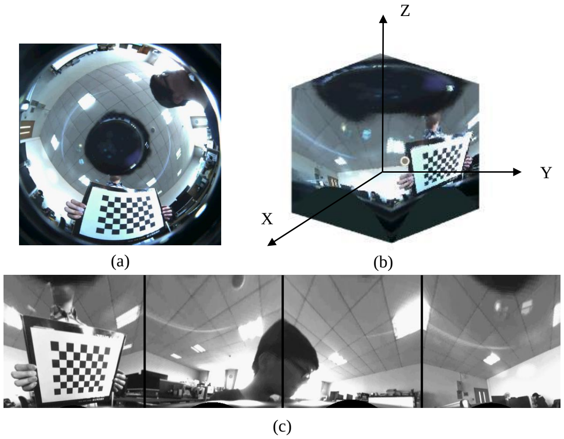

Figure 6. Multi-pinhole PAL image rectification. (a) Raw PAL image. (b) Mapping the PAL image on the unit cube. (c) Complete undistorted multi-pinhole image.

coordinate system and the marker coordinate system are estimated, we can calculate the 3D similarity transformation between them. The marker can be regarded as an anchor of the building. Through a marker, we can unify multiple trajectories in a common world coordinate system, where its origin is the marker and the unit is meter.

4.1. Multiple-pinhole rectification

In order to detect the ArUco marker correctly, the marker on the image must satisfy the perspective projection rules, i.e. the border of a marker on the image must keep straight. Raw PAL images obviously do not satisfy this condition, as shown in figure 6(a).

Based on the taylor camera model, we can map a PAL image onto a unit sphere, as demonstrated in section 3.1. It is also feasible to extend the unit sphere to a cube and then we use four virtual pinhole cameras to capture the side surfaces of the cube, as shown in figure 6(b) and 6(c). The virtual pinhole camera parameters $\mathbf{K}^{mp}$ are set carefully to make the most of valid sensing areas of a PAL image. The projection rules of the undistorted pinhole image is given by

$$\mathbf{p} = \Pi^{mp}(\mathbf{P}) = \mathbf{K}^{mp}\mathbf{R}_i\mathbf{P} \tag{15}$$

where $\mathbf{R}_i$ is the rotation matrix between the PAL camera coordinate system and the coordinate system of the ith virtual pinhole camera.

4.2. Coordinate alignment

The 3D similarity transformation matrix $\mathbf{S}$ from the VO local coordinate system to the marker coordinate system can be decomposed as

$$\mathbf{S} = \begin{bmatrix} s\mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \tag{16}$$

We calculate $\mathbf{R}$, and then estimate s and $\mathbf{t}$ through a least squared method.

Assuming we have n pose pairs, denoted by $\mathbf{R}_i^{mk}, \mathbf{t}_i^{mk}$ and $\mathbf{R}_i^{VO}, \mathbf{t}_i^{VO}$, where $i = 1...n$. The $\mathbf{R}$ can be calculated by

$$\mathbf{R} = \frac{1}{n}\sum \mathbf{R}_i^{mk}(\mathbf{R}_i^{VO})^{-1} \tag{17}$$

Note that rotation matrices do not have summation operation, so we perform it on the Lie algebra $\mathfrak{so}(3)$.

Then we stack s and $\mathbf{t}$ into a vector to construct the following linear equations:

$$\begin{bmatrix} \mathbf{R}\mathbf{t}_1^{VO} & \mathbf{I} \\ \vdots & \vdots \\ \mathbf{R}\mathbf{t}_n^{VO} & \mathbf{I} \end{bmatrix} \cdot \begin{bmatrix} s \\ \mathbf{t} \end{bmatrix} = \begin{bmatrix} \mathbf{t}_1^{mk} \\ \vdots \\ \mathbf{t}_n^{mk} \end{bmatrix} \tag{18}$$

The least squared solution of s and $\mathbf{t}$ can be obtained by solving this overdetermined equations.

5. Evaluation

In this section, we will evaluate the performance of our panoramic VO and the accuracy of positioning in a building in detail. The camera we used is a PAL with 33 to 96 degrees half FoV and a CMOS sensor with 480 × 480 resolution and global shutter. Our entire positioning system runs at 10 Hz on average on a laptop with Intel(R) Core(TM) i7-8550U CPU and 16G RAM.

5.1. Evaluation of the visual odometry

Accuracy and robustness of our method are evaluated in this subsection. We compared our algorithm with three state-of-the-art monocular VO/SLAM algorithms, DSO (Engel et al.; 2017), SVO (Forster et al.; 2014) and ORB-SLAM2 (Mur-Artal and Tardós; 2017), on both synthetic datasets and real-world datasets. ORB-SLAM2 is a representation of feature-based SLAM algorithms. It utilizes ORB features (Rublee et al.; 2012) to match the corresponding points in two images and then estimates the camera motion. we disabled the loop-closing function and only kept the VO function of ORB-SLAM2 for fair comparison. SVO is a hybrid VO method. It takes use of sparse optical flow and direct image alignment to estimate the camera motion first and minimizes the reprojection error of points in the back-end, unlike our method and DSO that minimize the photometric error of pixels. Note that these three methods can only run on narrow FoV image sequences which can be described by the pinhole model.

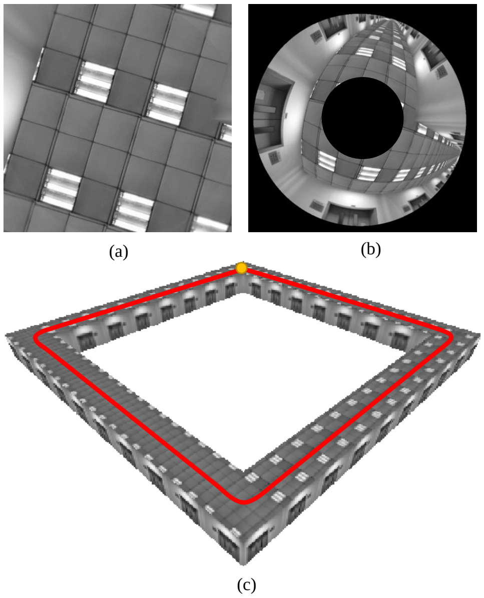

Figure 7. Synthetic datasets. (a) A synthetic pinhole image. (b) A synthetic PAL image. (c) Top view of the virtual corridor.

5.1.1. Synthetic datasets Synthetic datasets can provide ground truth to evaluate the accuracy of pose estimation comprehensively. Considering the application of our system, we created a virtual circular corridor by mapping real images of ceiling and wall to a 3D model, as shown in Figure 7. Image sequences were synthesized by simulating a pedestrian walking in the corridor. We synthesize both PAL images and pinhole camera images for our algorithm and other algorithms respectively. The parameters of synthetic PAL images are consistent with the real PAL images. The resolution of synthetic pinhole images is set to be the same as PAL images (480×480) for fair comparison, and the FoV is set to 90 degrees, a common FoV for pinhole cameras.

In order to simulate different scenarios, we synthesized six datasets in total, with two different ceilings and three different walking speed. Each sequence contains 2500 images, and some noise was also added to poses to simulate the camera shaking caused by walking. Note that the noise only affects poses and do not cause motion blur on the image. The accuracy was evaluated by the alignment error proposed in (Engel et al.; 2016). Let $\mathbf{p}_{1\ldots n} \in \mathbb{R}^3$ and $\mathbf{g}_{1\ldots n} \in \mathbb{R}$ are the generated trajectory and the ground truth with n frames. First we align both the start segment $\mathbf{p}_{1\ldots n/2}$ and end segment $\mathbf{p}_{n/2+1\ldots n}$ to the corresponding ground truth trajectory $\mathbf{g}$ independently and get two similarity transformation matrices $\mathbf{S}_{start}$ and $\mathbf{S}_{end}$. Then the alignment error is defined by the translational Root Mean Square Error (RMSE) between two trajectories that transformed by $\mathbf{S}_{start}$ and $\mathbf{S}_{end}$ respectively:

$$e_{align} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\mathbf{S}_{start}\mathbf{p}_i - \mathbf{S}_{end}\mathbf{p}_i|^2} \tag{19}$$

The scale drift e_{scale} and rotation error $e_{rotation}$ can also be evaluated quantitatively by decomposing the rotation part and scale part of the difference between two similarity transformation $\mathbf{S}_{start}$ and $\mathbf{S}_{end}$:

$$e_{rotation} = rotation(\mathbf{S}_{start}\mathbf{S}_{end}^{-1}) \tag{20}$$

$$e_{scale} = scale(\mathbf{S}_{start}\mathbf{S}_{end}^{-1}) \tag{21}$$

Each sequences was run five times for each algorithm, and the mean alignment error of five times was recorded, as shown in table 1.

From table 1, we can see that in synthetic datasets with block ceiling, almost all methods can run successfully. DSO and our method are more accurate than SVO and ORB-SLAM2 (without loop-closing) in block ceiling datasets. The main reason is that both SVO and ORB-SLAM2 need to extract FAST corner features. But for the block ceiling, only the cross point of lines can be regarded as stable corner features. DSO and our method rely on points with high gradient, so the point on lines can also be tracked. Although our method and DSO follow a similar flow, there is still a difference of accuracy between. This difference can be explained from the perspective of angular resolution. PAL images and pinhole images have the same resolution (480 × 480), but the FoV of PAL images is nearly twice as large as pinhole images, which means pinhole images have higher angular resolution. In white ceiling datasets, only our method can run successfully. Benefit from large FoV, our method can track points on walls and keep not failing.

5.1.2. Real-world datasets Two datasets of different scales are used to evaluate the performance of VO qualitatively. We captured both PAL images and pinhole camera images simultaneously for comparison as we did on synthesized datasets. The pinhole camera was the left infrared camera of Intel RealSense D435 (*Intel(R) Realsense(TM) Depth Camera D400-Series*; 2018) with resolution 1280 × 720, horizontal FoV 87 degrees and global shutter. Two cameras were rigidly fixed to a 3D printed shelf to ensure that two cameras have the same motion.

In the first dataset, we traversed a looped route in a hall. The size of the hall was about 9.3m × 9.3m. Both SVO and ORB-SLAM2 cannot be initialized due to the lack of detected corner features. As map creation and camera pose estimation complement each other, a consistent map is equal to an accurate trajectory. Thereby, We qualitatively evaluate the camera poses

Table 1. Results on synthetic datasets. BC and WC indicate the block ceiling and the white ceiling used in synthetic datatsets. LS, MS, and HS indicate low speed, medium speed and high speed at witch the virtual camera moves. The unit of $e_{rotation}$ is degree.

	Ours			DSO			SVO			ORB-SLAM2		
	e_{align}	$e_{rotation}$	e_{scale}	e_{align}	$e_{rotation}$	e_{scale}	e_{align}	$e_{rotation}$	e_{scale}	e_{align}	$e_{rotation}$	e_{scale}
s1(BC+LS)	96.6(15.7)	0.25(0.11)	1.0%(0.1%)	62.2(6.0)	0.15(0.01)	0.4%(0.1%)	175.0(59.6)	0.22(0.08)	1.7%(0.6%)	171.8(36.1)	0.64(0.18)	0.3%(0.6%)
s2(BC+MS)	95.6(11.2)	0.28(0.06)	0.3%(0.6%)	67.0(3.8)	0.21(0.04)	0.5%(0.1%)	207.0(95.7)	0.24(0.01)	2.3%(1.2%)	250.4(31.1)	1.21(0.11)	1.3%(0.6%)
s3(BC+HS)	133.2(58.2)	0.37(0.05)	1.3%(0.6%)	102.3(3.2)	0.32(0.01)	0.7%(0.1%)	280.4(77.4)	0.30(0.08)	2.4%(1.2%)	-	-	-
s4(WC+LS)	228.0(136.5)	0.34(0.01)	1.7%(1.5%)	-	-	-	-	-	-	-	-	-
s5(WC+MS)	314.4(40.3)	0.42(0.05)	2.7%(0.6%)	-	-	-	-	-	-	-	-	-
s6(WC+HS)	504.0(114.5)	0.45(0.25)	5.0%(1.0%)	-	-	-	-	-	-	-	-	-

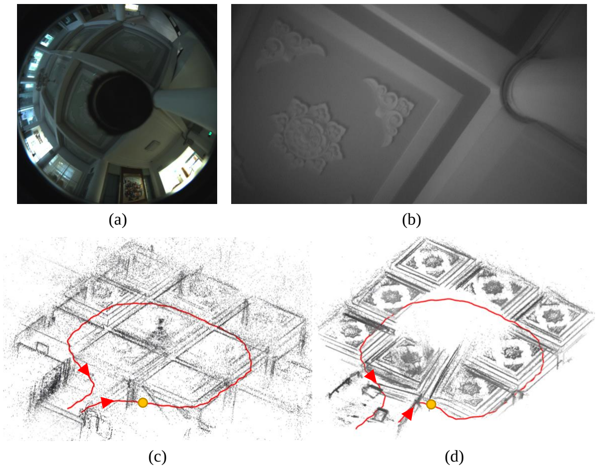

Figure 8. (a, b) Images captured by PAL and the left infrared camera of Intel RealSense D435. PAL can capture entire ceiling and part of walls, while the conventional pinhole camera can only capture a small portion of the ceiling. (c, d) The map created by our method and DSO respectively. The map created by our method is more consistent than DSO.

estimation by observing the consistency of the map. The results of our algorithm and DSO are shown in figure 8. The map created by DSO is clean, but at the start and end of the trajectory, the map has obvious inconsistencies. In contrast, the map created by our algorithm is sparser, but more consistent. Benefit from large FoV, tracked points can stay in the FoV for a long time, and are gradually refined by stereo matching and windowed optimization. On the other hand, expanding FoV but not changing the resolution results in the loss of image details, causing the map to become sparse.

The second dataset was captured in a two-story building and the full trajectory contained long distance corridors, up and down stairs, as well as loop in a hall. All three pinhole camera VO methods failed in this dataset due to the directing illumination of lights in the corridor as well as the white ceiling at the stairs. The results of our algorithm and some image samples are shown in figure 9. Although there are some accumulative errors, the map is complete. From the comparison of raw PAL images and pinhole camera images, we can see that PAL images have more textures when going upstairs and downstairs. In the corridor, lights only occupy a small part of the image, and our algorithm can still use other stable textures to achieve positioning. The result of these two datasets demonstrate the robustness of our algorithm, which is critical for VIP indoor positioning.

5.2. *Evaluation of the global positioning*

We also quantitatively evaluated the positioning accuracy of the entire system in the world coordinate system. The experiment field was in a hall. We defined a start point and an end point in the field, and an ArUco marker was attached to the wall at the start point. We walked from the start point to the end point along a similar route for five times and the full route is about 50 meters. The positioning error from the end of the five trajectories to the real end point indicates the positioning error of the entire system.

The experiment field and five routes are shown in figure 10. All five trajectories are unified correctly in the world coordinate system. Route 1 has the largest positioning error of 1.625m and the error of other routes are all less than 1m. The results show that our system has the ability to provide reliable positioning results in indoor environments

5.3. *Discussion*

In this subsection, we discuss the advantages of our positioning framework in assisting VIP.

Compared with narrow FoV camera, PAL is more suitable for assisting VIP to positioning. PAL with wide FoV can capture more information in a less textured indoor environment, which improves the robustness of our VO algorithm, as demonstrated in Section 5.1. Besides, the ceiling-view configuration avoids the influence of dynamic pedestrians on positioning. Moreover, the compact size of PAL also makes our system more portable.

To unify positioning results into a world coordinate system, we propose an ArUco marker-based coordinate alignment algorithm. PAL has the ability to observe 360-degree scenes around the user. The marker can be attached above the door, and once the user approaches to the marker, the coordinate alignment is triggered. For narrow FoV cameras, this method is

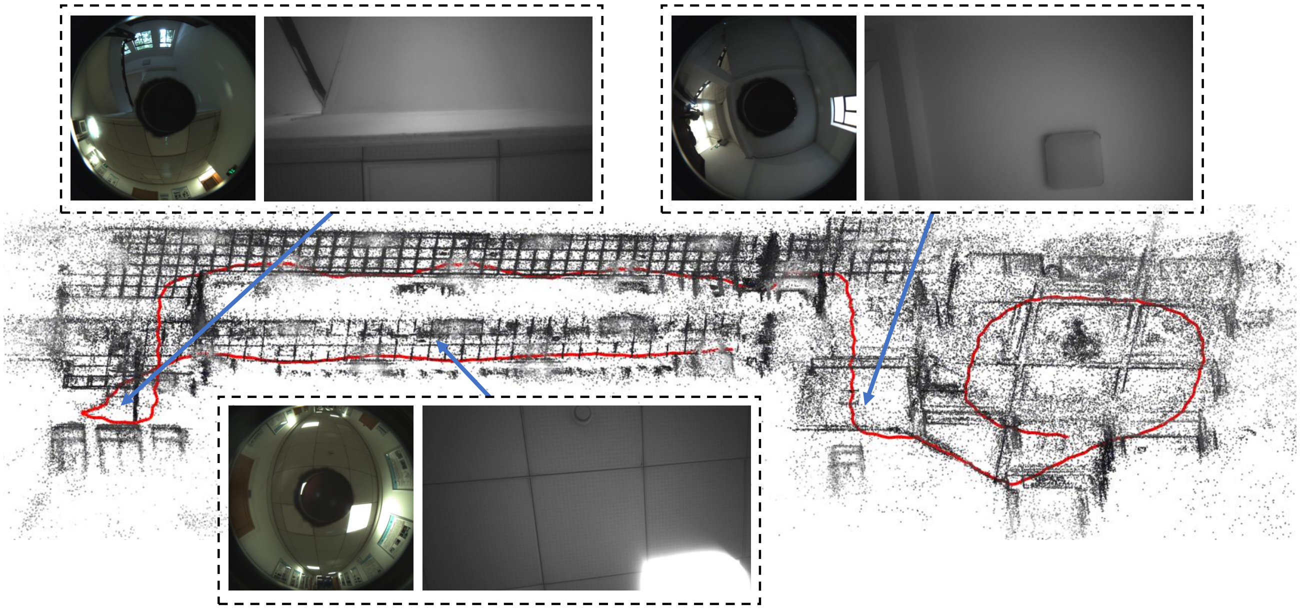

Figure 9. The generated point cloud and tracked trajectory of the large scale real-world dataset. The dataset was captured in a two-story building, containing long corridors, up and down stairs, as well as loop in a hall. The comparison of PAL images and common pinhole images at three places are also shown besides the trajectory. We can observe that PAL images contain much more information compared with the common pinhole images, which makes our algorithm more robust, especially in the scene with low textures.

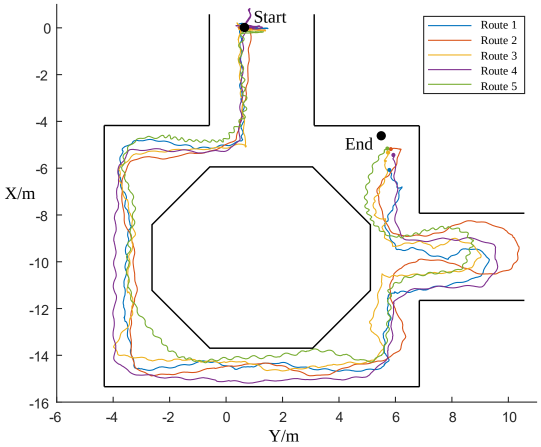

Figure 10. Trajectories of five routes.

also feasible. But limited by the narrow FoV, the area that can observe the marker and trigger the coordinate alignment becomes very small. And it is hard for blind user to move to that area and shoot the marker to trigger coordinate alignment.

At last, we discuss the positioning precision. As an odometry, our method estimates the camera translation and rotation frame-by-frame, so positioning error will accumulate inevitably. If an ArUco marker is detected, the positioning result will be aligned to the world coordinate system and the accumulative drifts will be eliminated immediately. From the global positioning experiment in Section 5.2, we can observe that our method produced a positioning error of up to 1.625m in a 50m route. This error is acceptable for normal indoor navigation tasks. Moreover, the positioning error can be further reduced by adding more ArUco markers according to the requirement of applications.

6. Conclusions

In this paper, we proposed indoor positioning framework based on panoramic visual odometry for visually impaired people. Based on panoramic annular lens, our method achieve robust indoor positioning compared with conventional pinhole cameras. An easy-to-maintain coordinate alignment method is also proposed to transform positioning results to the world coordinate system. In the feature, we plan to add an auditory-based human-machine interface to make our system more practical. In addition, we would also like to research how to reuse map to enhance the positioning accuracy.

7. Acknowledgments

This work was supported by the State Key Laboratory of Modern Optical Instrumentation and Hangzhou KrVision Technology Co., Ltd (krvision.cn).

References

An, L., Zhang, X., Gao, H. and Liu, Y. (2017). Semantic segmentation–aided visual odometry for urban autonomous driving, *International Journal of Advanced Robotic Systems* **14**(5): 1729881417735667.

Bourne, R. R. A., Flaxman, S. R., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., Keeffe, J., Kempen, J. H., Leasher, J., Limburg, H., Naidoo, K., Pesudovs, K., Resnikoff, S., Silvester, A., Stevens, G. A., Tahhan, N., Wong, T. Y., Taylor, H. R., Bourne, R., Ackland, P., Arditi, A., Barkana, Y., Bozkurt, B., BRAITHWAITE, T., Bron, A., Budenz, D., Cai, F., Casson, R., Chakravarthy, U., Choi, J., Cicinelli, M. V., Congdon, N., Dana, R., Dandona, R., Dandona, L., Das, A., Dekaris, I., Monte, M. D., Deva, J., Dreer, L., Ellwein, L., Frazier, M., Frick, K., Friedman, D., Furtado, J., Gao, H., Gazzard, G., George, R., Gichuhi, S., Gonzalez, V., Hammond, B., Hartnett, M. E., He, M., Hejtmancik, J., Hirai, F., Huang, J., Ingram, A., Javitt, J., Jonas, J., Joslin, C., Keeffe, J., Kempen, J., Khairallah, M., Khanna, R., Kim, J., Lambrou, G., Lansingh, V. C., Lanzetta, P., Leasher, J., Lim, J., LIMBURG, H., Mansouri, K., Mathew, A., Morse, A., Munoz, B., Musch, D., Naidoo, K., Nangia, V., PALAIOU, M., Parodi, M. B., Pena, F. Y., Pesudovs, K., Peto, T., Quigley, H., Raju, M., Ramulu, P., Resnikoff, S., Robin, A., Rossetti, L., Saaddine, J., SANDAR, M., Serle, J., Shen, T., Shetty, R., Sieving, P., Silva, J. C., Silvester, A., Sitorus, R. S., Stambolian, D., Stevens, G., Taylor, H., Tejedor, J., Tielsch, J., Tsilimbaris, M., van Meurs, J., Varma, R., Virgili, G., Volmink, J., Wang, Y. X., Wang, N.-L., West, S., Wiedemann, P., Wong, T., Wormald, R. and Zheng, Y. (2017). Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis, *The Lancet Global Health* **5**(9): e888 – e897.

Dimas, G., Iakovidis, D. K., Karargyris, A., Ciuti, G. and Koulaouzidis, A. (2017). An artificial neural network architecture for non-parametric visual odometry in wireless capsule endoscopy, *Measurement Science and Technology* **28**(9): 094005.

Dimas, G., Ntakolia, C. and Iakovidis, D. K. (2019). Obstacle detection based on generative adversarial networks and fuzzy sets for computer-assisted navigation, *International Conference on Engineering Applications of Neural Networks*, Springer, pp. 533–544.

Donati, M., Iacopetti, F., Celli, A., Roncella, R. and Fanucci, L. (2020). An aid system for autonomous mobility of visually impaired people on the historical city walls in lucca, italy, *Technological Trends in Improved Mobility of the Visually Impaired*, Springer, pp. 379–411.

Engel, J., Koltun, V. and Cremers, D. (2017). Direct sparse odometry, *IEEE transactions on pattern analysis and machine intelligence* **40**(3): 611–625.

Engel, J., Schöps, T. and Cremers, D. (2014). Lsd-slam: Large-scale direct monocular slam, *European conference on computer vision*, Springer, pp. 834–849.

Engel, J., Sturm, J. and Cremers, D. (2013). Semi-dense visual odometry for a monocular camera, *Proceedings of the IEEE international conference on computer vision*, pp. 1449–1456.

Engel, J., Usenko, V. and Cremers, D. (2016). A photometrically calibrated benchmark for monocular visual odometry, *arXiv:1607.02555*.

Forster, C., Pizzoli, M. and Scaramuzza, D. (2014). Svo: Fast semi-direct monocular visual odometry, *2014 IEEE international conference on robotics and automation (ICRA)*, IEEE, pp. 15–22.

Garrido-Jurado, S., Munoz-Salinas, R., Madrid-Cuevas, F. J. and Medina-Carnicer, R. (2016). Generation of fiducial marker dictionaries using mixed integer linear programming, *Pattern Recognition* **51**: 481–491.

Harle, R. (2013). A survey of indoor inertial positioning systems for pedestrians, *IEEE Communications Surveys & Tutorials* **15**(3): 1281–1293.

Hu, F., Zhu, Z. and Zhang, J. (2015). Mobile panoramic vision for assisting the blind via indexing and localization, *in* L. Agapito, M. M. Bronstein and C. Rother (eds), *Computer Vision - ECCV 2014 Workshops*, Springer International Publishing, Cham, pp. 600–614.

Huang, A. S., Bachrach, A., Henry, P., Krainin, M., Maturana, D., Fox, D. and Roy, N. (2017). Visual odometry and mapping for autonomous flight using an rgb-d camera, *Robotics Research*, Springer, pp. 235–252.

Huang, Z., Bai, J., Lu, T. X. and Hou, X. Y. (2013). Stray light analysis and suppression of panoramic annular lens, *Optics express* **21**(9): 10810–10820.

Iakovidis, D. K., Diamantis, D., Dimas, G., Ntakolia, C. and Spyrou, E. (2020). Digital enhancement of cultural experience and accessibility for the visually impaired, *Technological Trends in Improved Mobility of the Visually Impaired*, Springer, pp. 237–271.

Intel(R) Realsense(TM) Depth Camera D400-Series (2018). Accessed: 2019-06-22.
URL: *https://software.intel.com/en-us/realsense/d400*

Jayakody, J. A. D. C. A., Murray, I., Hermann, J., Lokuliyana, S. and Dunuwila, V. (2020). Intelligent vision impaired indoor navigation using visible light communication, *Technological Trends in Improved Mobility of the Visually Impaired*, Springer, pp. 181–206.

Lee, Y. H. and Medioni, G. (2016). Rgb-d camera based wearable navigation system for the visually impaired, *Computer Vision and Image Understanding* **149**: 3–20.

Li, B., Muñoz, J. P., Rong, X., Xiao, J., Tian, Y. and Arditi, A. (2016). Isana: Wearable context-aware indoor assistive navigation with obstacle avoidance for the blind, *in* G. Hua and H. Jégou (eds), *Computer Vision – ECCV 2016 Workshops*, Springer International Publishing, Cham, pp. 448–462.

Mur-Artal, R. and Tardós, J. D. (2017). Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras, *IEEE Transactions on Robotics* **33**(5): 1255–1262.

Murata, M., Ahmetovic, D., Sato, D., Takagi, H., Kitani, K. M. and Asakawa, C. (2018). Smartphone-based indoor localization for blind navigation across building complexes, *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, IEEE, pp. 1–10.

Murillo, A. C., Gutiérrez-Gómez, D., Rituerto, A., Puig, L. and Guerrero, J. J. (2012). Wearable omnidirectional vision system for personal localization and guidance, *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, pp. 8–14.

Romero-Ramirez, F. J., Muñoz-Salinas, R. and Medina-Carnicer, R. (2018). Speeded up detection of squared fiducial markers, *Image and vision Computing* **76**: 38–47.

Rublee, E., Rabaud, V., Konolige, K. and Bradski, G. R. (2012). Orb: an efficient alternative to sift or surf, *International Conference on Computer Vision*.

Scaramuzza, D. (2007). *Omnidirectional Vision: From calibration to root motion estimation*, PhD thesis, ETH Zurich.

Scaramuzza, D., Martinelli, A. and Siegwart, R. (2006). A toolbox for easily calibrating omnidirectional cameras, *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, pp. 5695–5701.

Vogiatzis, G. and Hernández, C. (2011). Video-based, real-time multi-view stereo, *Image and Vision Computing* **29**(7): 434–441.

Yang, K., Wang, K., Bergasa, L., Romera, E., Hu, W., Sun, D., Sun, J., Cheng, R., Chen, T. and López, E. (2018). Unifying terrain awareness for the visually impaired through real-time semantic segmentation, *Sensors* **18**(5): 1506.

Yang, K., Wang, K., Hu, W. and Bai, J. (2016). Expanding the detection of traversable area with realsense for the visually impaired, *Sensors* **16**(11): 1954.