Published in final edited form as:

IEEE Trans Autom Sci Eng. 2015 October ; 12(4): 1169–1180. doi:10.1109/TASE.2015.2469726.

6-DOF Pose Estimation of a Robotic Navigation Aid by Tracking Visual and Geometric Features

Cang Ye [Senior Member, IEEE], Soonhac Hong [Member, IEEE], and Amirhossein Tamjidi

Department of Systems Engineering, University of Arkansas at Little Rock, Little Rock, AR 72204, USA

Cang Ye: cxye@ualr.edu

Abstract

This paper presents a 6-DOF Pose Estimation (PE) method for a Robotic Navigation Aid (RNA) for the visually impaired. The RNA uses a single 3D camera for PE and object detection. The proposed method processes the camera's intensity and range data to estimate the camera's egomotion that is then used by an Extended Kalman Filter (EKF) as the motion model to track a set of visual features for PE. A RANSAC process is employed in the EKF to identify inliers from the visual feature correspondences between two image frames. Only the inliers are used to update the EKF's state. The EKF integrates the egomotion into the camera's pose in the world coordinate system. To retain the EKF's consistency, the distance between the camera and the floor plane (extracted from the range data) is used by the EKF as the observation of the camera's z coordinate. Experimental results demonstrate that the proposed method results in accurate pose estimates for positioning the RNA in indoor environments. Based on the PE method, a wayfinding system is developed for localization of the RNA in a home environment. The system uses the estimated pose and the floorplan to locate the RNA user in the home environment and announces the points of interest and navigational commands to the user through a speech interface.

Note to Practitioners—This work was motivated by the limitations of the existing navigation technology for the visually impaired. Most of the existing methods use a point/line measurement sensor for indoor object detection. Therefore, they lack capability in detecting 3D objects and positioning a blind traveler. Stereovision has been used in recent research. However, it cannot provide reliable depth data for object detection. Also, it tends to produce a lower localization accuracy because its depth measurement error quadratically increases with the true distance. This paper suggests a new approach for navigating a blind traveler. The method uses a single 3D time-of-flight camera for both 6-DOF PE and 3D object detection and thus results in a small-sized but powerful RNA. Due to the camera's constant depth accuracy, the proposed egomotion estimation method results in a smaller error than that of existing methods. A new EKF method is proposed to integrate the egomotion into the RNA's 6-DOF pose in the world coordinate system by tracking both visual and geometric features of the operating environment. The proposed method substantially reduces the pose error of a standard EKF method and thus supports a longer range navigation task. One limitation of the method is that it requires a feature-rich environment to work well.

Index Terms

robotic navigation aid; egomotion estimation; pose estimation; SLAM; extended Kalman filter; filter consistency; indoor localization; wayfinding

I. Introduction

Visual impairment reduces a person's independent mobility and severely deteriorates quality of life. In the US, 40% of people over 65 lose their vision as a result of age related diseases such as glaucoma, macular degeneration and diabetes [1]. As the baby boomers approach their 60s, more Americans will experience visual impairment. Therefore, there is a dire need to develop new technology to help the visually impaired live independent lives. The problem of independent mobility of a visually impaired person can be divided into two components, obstacle avoidance and wayfinding. Obstacle avoidance deals with a local problem in taking steps safely without bumping into anything, tripping, or falling. Wayfinding refers to a global problem of planning and following a route from one point to another while maintaining awareness of the current position and heading. The problem shares the common ground with mobile robot navigation. In spite of substantial advances in computer vision and robotics, independent mobility is still challenging to a visually impaired person. Up to date, white cane remains the most efficient mobility tool due to the powerful haptic feedback and low cost. However, a white cane is a point-contact device that cannot provide a "full picture" of its surroundings. Also, it has a limited range and is thus too nearsighted for path planning. In addition, it is unable to provide location information. To address these limitations, a number of Robotic Navigation Aids (RNAs) have been introduced as alternative mobility tools for the visually impaired. Up to date, there is no RNA that has effectively addressed both obstacle avoidance and wayfinding problems. The main technical challenge is that both problems must be addressed in a small platform with limited resources.

This paper considers independent mobility of the visually impaired in a home environment. Moving around safely in a home often poses great challenges to a blind traveler because: (1) the environment may be more densely populated with obstacles than an outdoor open space; (2) it may contain overhanging obstacles; and (3) it is GPS-denied. The first two conditions require an RNA to be capable of 3D perception while the third factor means that existing GPS-based blind navigation technology [2, 3, 4, 5] cannot be used for localization. However, a home environment may provide the RNA with off-board computing support that may minimize onboard computing need and help to create a system with higher autonomy and lower power dissipation. To address the challenges and take advantage of navigation in a home environment, we develop a new portable RNA [6] with a client-server architecture (described in V.C). The RNA, called Smart Cane (SC), is a computer vision enhanced white cane that uses a 3D camera—SwissRanger SR4000—to perceive its surroundings and provides navigational information to its user. The SC provides two functions—wayfinding and 3D object detection—to its user. The wayfinding function is achieved by using a new Pose Estimation (PE) method to locate the user in a floorplan. The SC can be regarded as a smart home system. On one hand, it provides navigational support to the SC user by the

home-based computer vision service. On the other hand, it maintains real-time location information of the user that may be used to provide other services such as automatically opening/closing a door in a nursing home for the user. This paper will be focused on the PE problem, aka Simultaneous Localization and Mapping (SLAM). An overview of the related work in RNAs and SLAM is given in the reminder of this Section.

A. Related wayfinding methods for RNAs

There have been some attempts to address wayfinding in GPS-denied environments. But only limited success has been achieved so far. In [7], a Robotic Guide Dog (RGD) is introduced to lead the way for a blind traveler. The RGD uses RFID tags deployed in the environment to guide itself from one point to another. A map containing the locations of the RFID tags and the Points Of Interest (POIs) is used by the RGD for navigation. The RGD uses an RFID reader to detect a tag and retrieve the stored information, based on which it determines its location and the desired movement towards the next tag until reaching the destination. The RGD lacks portability and the approach requires installation of RFID tags in the environment. Hesch *et al.* [8] propose a portable indoor localization aid that employs an Extended Kalman Filter (EKF) to fuse data from a 3-axis gyro, a 2D LIDAR (both installed on the device) and a user-worn pedometer for 6-DOF PE. The gyro and pedometer measure the device's rotation and translation, respectively. The motion is then used to predict the device pose and laser scan. The actual laser data is then acquired and used to update the device's pose. For laser scan prediction, the environment is assumed to be vertical and the map is assumed to be available. The method requires corner points as observations for state update. It may not work in a geometrically featureless environment. In addition, the assumptions on the environment and map availability may be unrealistic in some real-world scenarios.

In our SC project, we develop a 6-DOF PE method. The method first estimates the SC's Pose Change (PC) between two camera views by an egomotion estimation method, called Visual-Range Odometry (VRO) [9], by processing the intensity and range data of the camera. It then integrates the PCs over time by an EKF [10] to obtain the SC's pose in the world coordinate system. The PE method estimates the device pose without requiring any a priori information of the environment. The pose information is used for both localization and 3D map-building. Recently, the method is implemented in a wayfinding system that may provide real-time navigational information to guide a blind person in a home environment.

B. Related work in SLAM

In the literature of mobile robotics, a robot often obtains its pose by a SLAM method [11, 12] that builds and updates a map within an unknown environment while at the same time tracks its pose. Existing SLAM methods can be broadly classified into two categories: state filter based SLAM [13] and pose-graph SLAM [14, 15]. The former is preferred in this work because a graph SLAM method requires large amount of memory for graph storage. In a state filter based method, the pose estimate is a multi-dimensional probability distribution function represented by a mean (state vector) and covariance (state covariance). The filter maintains a posterior over the robot pose by tracking only a limited number of features from

the previous steps. Therefore, it is memory-efficient. SLAM algorithms have become mature in mobile robotics and they have recently been applied to wearable RNAs [16, 17, 18]. Technical factors to be considered in applying a SLAM algorithm to an RNA include the dimension of imaging sensor to meet portability requirement and the reliability of the sensor's range data for object detection. Among existing compact 3D imaging sensors, stereovision, RGB-D and 3D time-of-flight (TOF) cameras, a 3D TOF camera is adopted in this work because: (1) it has a smaller dimension; (2) it provides more reliable range data; and (3) its operating principle (phase shift measurement) allows the use of multiple cameras without interference.

Saez *et al.* [18] propose a Visual-SLAM (VSLAM) approach for 6-DOF PE of a wearable RNA that uses a stereo camera. The egomotion between two consecutive camera frames is obtained by a visual odometry algorithm [19] and the device's pose in the world coordinate system is determined by minimizing an entropy-based criterion. In [16], a metric-topological SLAM method is employed for PE of a stereovision-based wearable RNA. The method extracts and tracks features in the stereo camera's images step by step. A FastSLAM algorithm [12] is used to process the visual information and update the local metric maps and the global topological relations between them. Despite its suitability for VSLAM, a stereo camera has a limitation for indoor navigation—it cannot provide complete depth data for reliable obstacle detection [9]. To overcome this problem, an RGB-D camera is used in the wearable RNA in [17] because it may provide more reliable depth data in a low-textured environment. In all of these VSLAM solutions, the algorithms have to maintain and update a map with an increasing size for PE as the user moves around. This may result in an intractable memory requirement and computational cost.

To alleviate this problem, a VSLAM method, called 1-Point RANSAC EKF-SLAM (1-PRES), is introduced in [20]. The method maintains a dynamic local map consisting of recently observed visual features and uses an EKF to track the features and estimate the 6-DOF pose of a mobile robot. To obtain an accurate state estimate, the method employs a RANSAC procedure to generate hypothesized states and find the Most Supported State Hypothesis (MSSH). The features that support the MSSH are then used to update the state. However, the method incurs the scale problem due to the use of a monocular camera. Also, it requires odometry data. A fundamental issue with the 1-PRES method is that the linearization of the nonlinear state and measurement models accrues state estimation error [21] that may drive the filter into inconsistency. To overcome this problem, Hervier *et al.* [22] use a nonlinear filter—Invariant EKF—to fuse the egomotion estimated from a Kinect sensor and the motion data of an Inertial Measurement Unit (IMU) for SLAM. Their method uses the Iterative Closest Point (ICP) algorithm to compute the egomotion and covariance from the Kinect's depth data. The ICP method may fail in our case when the SC is in an open space with a flat scene. In addition, the method requires an additional sensor (the IMU) and the sensor fusion procedure adds computational complexity to the system. Ling *et al.* [23] propose a repetitive linearization method to alleviate the propagation of EKF linearization error. The method re-linearizes the measurement equation by iterating an approximate maximum a posteriori estimate around the updated state instead of the

predicated state. In their experiment with a Kinect sensor, an iteration number of 30 is used as a trade-off between estimation accuracy and computational cost.

In this paper, we propose a new approach that uses geometric feature extracted from range data to clip the camera's z coordinate error to a small value and thus reduce the EKF's linearization error propagation. Without adding computational complexity, the method retains the EKF consistency in a certain time windows to satisfy the wayfinding need in a home environment. Using the 1-PRES framework [20], the proposed method uses the VRO's output to predict the SC pose that is then updated by the EKF through tracking some visual features sampled from the previous images and the geometric feature extracted from the current range data. Based on the method, we develop a wayfinding system that is able to locate the SC user in an indoor environment (e.g., nursing home or residential home). The wayfinding system announces the POIs and navigational commands as the user moves around. It is noted that planes extracted from range data have been used for SLAM [24, 25, 26] recently. However, these SLAM methods rely on geometric features (multiple intersecting planes) only. They will not work in a geometrically featureless environment (e.g., a flat ground). The proposed method is a visual based one that uses simple geometric feature (distance) to assist VSLAM and improve VSLAM accuracy.

II. Smart Cane Prototype and SR4000

As shown in Fig. 1, the SC prototype uses a SwissRanger SR4000 for 3D perception. The SR4000 has a spatial resolution of 176×144 pixels and a field of view of $43.6^\circ \times 34.6^\circ$. It has a small physical dimension ($50 \times 48 \times 65$ mm³) that makes it suitable for the SC. The camera illuminates its environment with modulated infrared light. Based on phase shift measurement, it detects a range up to 5 meters (with ± 1 cm accuracy) for every point on the imaging plane. The SR4000 produces both range and intensity data simultaneously at a rate up to 50 frames per second. Its modulation frequency is programmable to allow simultaneous use of multiple cameras without interference. The SR4000 is mounted on the white cane with a tilt-up angle of 18° to ensure that the cane is out of the camera's field of view. The camera is configured at the software-trigger-mode. An image is captured whenever the camera receives an acquisition command from the user-worn computer. The camera, floor plane and world coordinate systems are defined and depicted in Fig. 1.

III. Visual-Range Odometry

The foundation of the proposed PE method is the VRO algorithm that estimates the egomotion (i.e., PC), between two camera views. The operating principle and the characterization study of the VRO are presented in this Section.

A. Operating principle

The VRO extract visual features from the current intensity image and match them to those from the previous intensity image based on the feature descriptors. In this paper, the SIFT [27] descriptors are used. As the matched features' 3D coordinates are known from the range data, the feature-matching process results in two 3D point sets, $\{a_i\}$ and $\{b_i\}$ for $i = 1$,

$\dots, N$. The rotation and translation matrices, $\mathbf{R}$ and $\mathbf{T}$, between the two images are determined by minimizing the error residual

$$e^2 = \sum_{i=1}^{i=N} \|a_i - \mathbf{R}b_i - \mathbf{T}\|. \quad (1)$$

This least-squares data fitting problem is solved by the Singular Value Decomposition (SVD) method [28]. As the feature matching process may result in incorrect matches (outliers), the following RANSAC process is used to reject the outliers:

1. Detect SIFT features in two consecutive intensity images, find the matched features and form the corresponding 3D data sets $\{a_i\}$ and $\{b_j\}$. Repeat steps 2 & 3 for K times.
2. Draw a sample by randomly selecting 4 associated point-pairs, $\{a_m\}$ and $\{b_m\}$ for $m = 1, \dots, 4$, from the two point sets and find the least-squares rotation and translation matrices ($\mathbf{R}_k$ and $\mathbf{T}_k$) for $\{a_m\}$ and $\{b_m\}$.
3. Project the entire point set $\{b_i\}$ onto $\{a_i\}$ by using $\mathbf{R}_k$ and $\mathbf{T}_k$ and compute error $e_i^2 = \|a_i - \mathbf{R}_k b_i - \mathbf{T}_k\|$, $i = 1, \dots, N$ for each point-pair. A threshold ξ is used to score the support S_k for this transformation: S_k is incremented by 1 if $e_i^2 < \xi$.
4. The transformation with $\max(S_k)$ is recorded. The corresponding point sets $\{a_j\}$ and $\{b_j\}$ for $j = 1, \dots, \max(S_k)$, called inliers, are used to compute the maximum likelihood estimate $\hat{\mathbf{R}}$ and $\hat{\mathbf{T}}$ of the camera by the SVD method. The camera's orientation and position changes is computed from $\hat{\mathbf{R}}$ and its translation is determined by $\hat{\mathbf{T}}$.

Given the true inlier ratio ε , the minimum K required to ensure, with confidence ζ , that point sets $\{a_j\}$ and $\{b_j\}$ are outlier-free, is computed by

$$K_{min} = \frac{\log(1 - \zeta)}{\log(1 - \varepsilon^m)}. \quad (2)$$

In this paper, $m=4$ and $\zeta=0.99$ are used. As ε is a priori unknown, it is estimated at each repetition using the sample with the largest support.

B. Accuracy and repeatability

Noises in the SR4000's intensity and range data may produce error in Pose Change Estimation (PCE). To analyze the noises, we took 540 data frames on the same scene from the camera in a typical office environment (with objects ranging 1–5 meters) and plot the distributions of the intensity and range values of each pixel. It is found that both intensity and range data follow Gaussian distributions, i.e., the noises are Gaussian. To reduce the noises, a 3×3 low-pass Gaussian filter ($\sigma=1$) was applied to the intensity and range data. The distribution plots of the raw and filtered data reveal that the filter substantially reduces the standard deviations with slight reductions in the mean values. To quantify the noises of the raw and filtered data, we computed the noise ratio (i.e., the ratio of the standard deviation to the mean) for each intensity/range data. Table I shows the maximum, mean and minimum noise ratios in a typical data frame. It can be seen that the Gaussian filter reduces the overall

noise levels (the mean noise ratio) of the intensity and range data by 57% and 63%, respectively. We have carried out experiments with different scenes and the results are similar.

In our earlier work [29], we have shown that the use of the Gaussian filter greatly reduced the VRO's PCE error. Some quantitative results of the PCE accuracy and repeatability of the VRO (using the Gaussian filter) with individual movement in φ , θ , ψ , X , and Y are tabulated in Tables II and III. From the tables, it can be seen that most of the mean and standard deviations of the PCE errors are within the SR4000's angular resolution (0.24°) and absolute accuracy (± 10 mm). In this characterization study, we used Euler angles with z - x - y rotation sequence (i.e., yaw-pitch-roll angles) to represent orientation. The data were collected with a series of roll/pitch/yaw movements (range: $3^\circ \sim 18^\circ$, step size: 3°) or X/Y movements (range: 100~400 mm, step size: 100 mm). 540 frames were captured from the camera in a typical office environment before and after each rotation/translation movement for computing the mean and standard deviation of the errors.

To further investigate the VRO's accuracy and repeatability, an in-house built motion table was used to produce ground truth rotation and translation simultaneously for the SR4000. The motion table has a ± 60 tilt and ± 180 pan ranges with an angular resolution of 0.018° . It has a 3-meter linear movement range with a ~ 0.01 mm resolution. Using the ground truth data, the VRO is further characterized with combinatory rotation and translation. The first combination consists of pitch and yaw rotations and Y translation and the second combination consists of pitch and yaw rotations and X translation. The quantitative results we obtained are similar to Tables II and III, meaning that the overall PE accuracy and repeatability of the VRO with combinatory motion are equally good. For simplicity, the results are omitted here. Reader are referred to [29] for details.

C. Gaussianity of pose change estimation error

We examined the Gaussianity of the VRO's PCE error by plotting the error histogram for each case with an individual or combinatory motion. The results show that each PCE error follows a Gaussian distribution. Therefore, we will use an Extended Kalman Filter (EKF) to estimate the SC's pose in this paper. The histogram plots are omitted for conciseness.

D. Effect of pitch angle

It can be seen from Table II (the 2nd group of data) that the pitch measurement incurs a relatively larger mean error and/or standard deviation when the SR4000 undergoes a pitch movement. This is an application specific issue with the SC where the SR4000 points forward and downward to the ground. In this case, the non-zero incident angle of the camera's light beam decreases the amplitude of the received signal. With a reduced signal amplitude, the range measurement noise increases. Therefore, the VRO's repeatability becomes worse. This phenomenon can be modeled theoretically. According to [30], the intensity of the received signal and the standard deviation of range error can be expressed by

$$I \propto \frac{\rho \cos \alpha}{r^2} \quad \text{and} \quad \sigma_r \propto \frac{r}{\sqrt{\rho \cos \alpha}}, \quad (3)$$

where r , α and ρ are the range, incident angle and the reflectance of target surface, respectively. A non-zero incident angle results in a smaller I and a larger σ_r .

Fig. 2 shows the experimental results of the SR4000 with varying pitch angles. In the experiment, the SR4000 is initially pointed perpendicularly towards the ground. It then pitched up by 10° and we started taking 540 data frames for each 2° pitch movement. The error bar of the range measurement and the mean intensity of the center pixel are plotted against the pitch angle. The result validates that the intensity decreases while the range measurement noise increases with an increasing incident angle.

For the SC application case, the major rotation movement is yaw. However, a yaw movement does not affect r and α in a flat region. The SC may experience small roll and/or pitch rotation, which has minor impact on r and α . In fact, a roll rotation has opposite effect on the right and left light beams of the camera and thus its impact on the VRO is small. In other words, the VRO's PCE errors are not affected by the SC's movement. The determining factor is then ρ . This means that the VRO may produce reliable PCE (robust to camera motion) as long as the environment is visually feature-rich.

E. Minimum number of required matched Points

For noise-free range data, the VRO requires three matched data points to compute a PC. In case of noise, more points are needed to attain a sufficiently small PCE error. In this paper, the Minimum Number of Matched Points (MNMP) required by the VRO is determined by simulation that computes the PCE errors from two matched point sets (with a known transformation) corrupted by noises. The noises are generated using the noise ratios of the filtered data in Table I. By running the simulation with varying number of data points and using the SR4000's angular resolution (0.24°) and absolute accuracy (± 10 mm) as the thresholds, the MNMP is found to be 12 [29]. In this paper, a failure of the VRO is detected if the number of matched features or the number of inliers is smaller than the MNMP. In this case, the VRO skips to the next frame.

In summary, the characterization study demonstrates that the VRO's PCE error can be approximated as a Gaussian noise. Therefore, an EKF may be used to estimate the camera's pose in the world coordinate system by using the VRO result as the motion model. However, the relatively large mean error and standard deviation in pitch may accrue a large error in state prediction and result in filter inconsistency. Therefore, it is essential to devise a method to alleviate the problem.

IV. Pose Estimation by EKF

Based on the VRO, we may obtain the SC's pose in the world coordinate system by dead reckoning. However, it may accrue a large pose error over time. In this paper, we propose an EKF to estimate the SC's pose by tracking some visual features and the geometric feature of the operating environment. By tracking these features, the EKF finds an optimized pose that satisfies the PC constraints between the current and a number of previous camera views and thus results in a smaller pose error than that of dead reckoning. The visual features are some SIFT features sampled from previous intensity images while the geometric feature is the

floor plane extracted from camera's range data at each step. To be more specific, the geometric feature is the distance between the camera and the extracted floor plane. It can be seen from Fig. 1 that the distance is a measurement of the camera's z coordinate in the world coordinate system. This means that the z coordinate of the EKF's state vector can be observed at each step. We recall that the VRO has relatively large mean and standard deviation in measuring a pitch change. This may accrue a large pose error in z quickly, resulting in a large EKF linearization error and eventually causing filter inconsistency. The inclusion of z coordinate observation may significantly reduce z error so as to retain the EKF consistency.

To measure the distance between the camera and the floor plane, a RANSAC plane fitting method [32] is used to extract planes from the SR4000's range data at time step k . The largest plane, denoted Ξ_k , is detected as a Potential Floor Plane (PFP). A PFP detected at $t=0$ is treated as a floor plane while each subsequent PFP is classified as a floor plane only if it passes the floor plane detection criteria (described IV.B). A floor plane is expressed by $ax + by + cz + d = 0$. The distance from camera to the floor plane is then calculated as

$l_k = d / \sqrt{a^2 + b^2 + c^2}$. In a real-world application, gating factors based on the camera's height and the plane orientation may be used to enhance the reliability of floor plane detection. For example, l_k should be between 0.6 m and 1 m given the length of the SC and the height of a SC user. In the remainder of this section, the technical details of the proposed EKF PE method will be described.

A. State Vector

The state vector $\mathbf{x}_k$ of the EKF at step k is the concatenation of the camera's pose $\mathbf{x}_k^s$ and the stacked vector of the SIFT feature points that are maintained in a local feature map

$\text{Map}_k = [\mathbf{f}_k^1, \mathbf{f}_k^2, \dots, \mathbf{f}_k^m]^T$, where $\mathbf{f}_k^j = [x_k^j, y_k^j, z_k^j]^T$ is the j^{th} SIFT feature's 3D point (x , y and z coordinates in the world coordinate system) at the time it is initialized into the map. Each feature's SIFT descriptor is also recorded for feature matching in the future. The camera's state $\mathbf{x}_k^s$ is represented by a 7-dimensional vector consisting of the position $\mathbf{r}_k^s = [x_k^s, y_k^s, z_k^s]^T$ and the orientation quaternion $\mathbf{q}_k^s = [q_1, q_2, q_3, q_4]^T$, of the camera in the world coordinate system. The state covariance matrix $\mathbf{P}_k$ contains the correlation information between the elements of the state vector. The dimensions of $\mathbf{x}_k$ and $\mathbf{P}_k$ are of $(3m + 7) \times 1$ and $(3m + 7) \times (3m + 7)$, respectively.

At the beginning ($k=0$), the world coordinate, state vector and state covariance must be initialized. The world coordinate system is fixed using the extracted floor plane Ξ_{k0} . The origin of the world coordinate O_W is the projection of the camera coordinate's origin O_C on Ξ_0 . The world coordinate system's Z axis, Z_W , is aligned with the normal of Ξ_0 ; and its Y axis, Y_W , is aligned with the projection of Y_C on Ξ_0 . X_W is then determined to form a right-handed coordinate system. Fig. 1 depicts the coordinate systems. After the world coordinate system is fixed, the rotation of the camera coordinate system with reference to the world coordinate system, denoted ${}^w_c \mathbf{R}$, is calculated. The state of the camera is then initialized with

$$\mathbf{x}_0^s = [0 \quad l_0 \quad 0 \quad \mathbf{q}_c^w \mathbf{R}] \quad (4)$$

where $\mathbf{q}_c^w \mathbf{R}$ is the quaternion representation of ${}^w_c \mathbf{R}$, and l_0 is calculated by Equation (4) for $k = 0$. The floor plane is represented by $z=0$ in the world coordinate system.

The motion model of the SC is given by:

$$\mathbf{x}_k = f_s(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}) + \mathbf{w}_{k-1} \quad (5)$$

$$\text{or} \quad \begin{pmatrix} \mathbf{r}_k^s \\ \mathbf{q}_k^s \\ \text{Map}_k \end{pmatrix} = \begin{pmatrix} \mathbf{r}_{k-1}^s + \mathbf{R}(\mathbf{q}_{k-1}^s) \Delta \mathbf{r}_{k-1}^s \\ \mathbf{q}_{k-1}^s \otimes \Delta \mathbf{q}_{k-1}^s \\ \text{Map}_{k-1} \end{pmatrix} + \mathbf{w}_{k-1} \quad (6)$$

where $\otimes$, $\mathbf{R}(\mathbf{q}_{k-1}^s)$ and $\mathbf{w}_{k-1}$ are quaternion product, the rotation matrix of quaternion vector $\mathbf{q}_{k-1}^s$ and the process noise, respectively. The input to the motion model,

$\mathbf{u}_{k-1} = [\Delta \mathbf{r}_{k-1}^s, \Delta \mathbf{q}_{k-1}^s]^T$, represents the PC between step $k-1$ and k as estimated by the VRO. The Jacobians of the model are $\mathbf{F}_{k-1} = \mathbf{x}_k / \mathbf{x}_{k-1}$ and

$\mathbf{G}_{k-1} = \partial \mathbf{x}_k / \partial (\Delta \mathbf{r}_{k-1}^s, \Delta \mathbf{q}_{k-1}^s)$. The state vector and its covariance are predicted as follows:

$$\hat{\mathbf{x}}_{k|k-1} = f_s(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{u}_{k-1}) \quad (7)$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_{k-1} \mathbf{P}_{k-1|k-1} \mathbf{F}_{k-1}^T + \mathbf{G}_{k-1} \mathbf{Q}_{k-1} \mathbf{G}_{k-1}^T,$$

where $\mathbf{Q}_{k-1}$ is the covariance matrix of $\mathbf{u}_{k-1}$.

B. Observation Vector

The observation vector $\mathbf{v}_k^{\text{obs}}$ consists of two parts, $\mathbf{v}_k^f$ (visual feature observation) and $\mathbf{v}_k^z = l_k$ (observation of the z coordinate if a floor plane is detected). If no floor plane is detected, then $\mathbf{v}_k^z$ is removed from $\mathbf{v}_k^{\text{obs}}$. In order to detect the presence of a floor plane, innovation value $\gamma_k = l_k - z_{k|k-1}^s$ is used. The predicted variance for γ_k is computed by $s_\gamma = P_{zz}^s + \eta$, where η is the observation noise for l_k and P_{zz}^s is the variance of $z_{k|k-1}^s$. P_{zz}^s is obtained from state covariance $\mathbf{P}_{k|k-1}$. If $\gamma_k < 2s_\gamma$, l_k is included in $\mathbf{v}_k^{\text{obs}}$; otherwise, it is dropped and $\mathbf{v}_k^{\text{obs}}$ only contains visual feature observations.

In this paper, we use SIFT features as observations in the EKF process. When a new feature point is added to the map, it is initialized with its 3D coordinates in the world coordinate system. Its subsequent observation on the image plane is calculated using the state and the camera's model $\mathbf{h}^c$. A pin hole camera model is used. To account for the SR4000's optical distortion, an optical calibration process was performed to obtain the camera's intrinsic and extrinsic parameters. The observation of the i^{th} feature at step k is predicted by

$\mathbf{h}^c(\hat{\mathbf{x}}_{k|k-1}^s, \mathbf{f}_{k-1}^i)$. The prediction of the z coordinate observation is $\hat{z}_{k|k-1}^s$, i. e., the z element

of the predicted state vector. The observation model, including the SIFT features and the z coordinate, is denoted by $\mathbf{h}(\hat{\mathbf{x}}_{k|k-1})$. The predicted observation is then given by

$$\hat{\mathbf{h}}_{k|k-1} = \mathbf{h}(\hat{\mathbf{x}}_{k|k-1}) \quad (8)$$

$$\mathbf{S}_{k|k-1} = \mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}_k^T + \mathbf{R}_k,$$

where $\mathbf{H}_k = \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \big|_{\hat{\mathbf{x}}_{k|k-1}}$ is the jacobian of the observation model and $\mathbf{R}_k$ is the covariance of the observation noise. $\mathbf{R}_k$ is a $n \times n$ diagonal matrix with $\mathbf{R}_k(n, n) = 4.9 \times 10^{-5}$ and the other diagonal elements are 1. These are computed based on that the SR4000's measurement repeatability σ is 7 mm and the camera model has a 1-pixel uncertainty. Here, $n = 2m + 1$.

C. Data Association—SIFT Feature Correspondences

Data association between the SIFT features in the local feature map and that of the current image is determined by the following procedure: (1) The state vector is predicted for step k by Equation (7) and the observation is predicted by Equation (8); (2) The predicted covariance of each SIFT feature observation is used to determine a region on the image, within which a matched SIFT feature should locate; (3) A matched SIFT feature within the region is identified as an Individually Compatible Match (ICM). This process is performed on all predicted feature observations to produce a set of ICMs.

D. RANSAC-based State Update

The 6-DOF PE problem requires three feature correspondences. However, using more correspondences may reduce the effect of the camera's data noise and improve the estimation accuracy. The process of finding the maximum number of feature correspondences out of the ICMs is achieved by a 3-point RANSAC procedure as follows: (1) Three ICMs are randomly drawn to generate a hypothesized state update $\hat{\mathbf{x}}_i$, based on which we calculate the predicted observation, $\hat{\mathbf{h}}_i$, for all features. (2) A threshold τ is used to find a set of features $\mathbf{v}_i^T$ whose innovations are below τ under hypothesis $\hat{\mathbf{x}}_i$. These features are the supports to the state hypothesis $\hat{\mathbf{x}}_i$. The 2-step procedure is repeated for a number of times to find the hypothesis with the maximum number of supports. The corresponding features are identified as low innovation inliers (LIIs) and denoted by $\mathbf{v}_k^{\text{li}}$. They are then used to perform a partial state update to the EKF.

Partial State Update Using LIIs—In this stage, $\mathbf{v}_k^{\text{li}}$ is augmented with $\mathbf{v}_k^z$ (if available) to form the observation vector $\mathbf{v}_k^{\text{obs}}$, which is then used to perform a partial state update by:

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{H}_k^T \mathbf{S}_{k|k-1}^{-1}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k (\mathbf{v}_k^{\text{obs}} - \mathbf{h}(\hat{\mathbf{x}}_{k|k-1})) \quad (9)$$

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_{k|k-1}$$

Partial Update Using High Innovation Inliers—After the partial state update, some feature correspondences that were originally classified as outliers may become inliers [16]. These features are called high innovation inliers (HIIs). In this stage, we first apply a χ^2 test to the outlier points resulted from the previous stage. Points passing the test are identified as HIIs and their corresponding observation $\mathbf{v}_k^{\text{hi}}$ is augmented with $\mathbf{v}_k^z$ (if available) to form another observation, based on which another partial update is then performed to the state by using (9).

E. EKF State Estimation Loop

The proposed EKF PE method is diagramed in Fig. 3. It is a repetitive process. Each repetition consists of four stages. They are described as follows:

1) Preprocessing—First, the range data at step k is first processed to extract the floor plane, from which l_k is computed. Then, the PC between images at steps $k-1$ and k and the corresponding covariance are calculated. The VRO method is used to determine the PC with a modification to eliminate unreliable range data for PC computation. The modification is to use the confidence values of the SR4000 to discard unreliable SIFT features. The camera produces a confidence value for each range data representing how reliable the range measurement is. In this work, we discard those SIFT features whose confidence values are smaller than half of the maximum confidence value. This treatment improves the results of the VRO as well as the EKF state estimation. In this paper, we used the bootstrap approach [33] to calculate $\mathbf{Q}_{k-1}$ (the covariance of the PC). Since we use quaternion to describe the camera's orientation, the PC computed by the VRO is $(\Delta \mathbf{r}_{k-1}^s, \Delta \mathbf{q}_{k-1}^s)$.

To retain a computationally affordable EKF state update, a map with N SIFT feature points ($40 \leq N \leq 160$) is used in this paper. These features are sampled from previous images. A map management scheme is used to ensure that the SIFT feature points in the map are active and their number is large enough for data association and state update. The map management process discards a SIFT feature point from the map if it becomes obsolete and add new feature points into the map if the number of feature points falls below 40. For obsolete feature detection, the ratio of the number of times a feature is observed to the number of times it is predicted to be visible on the image is computed. If the ratio falls below 0.5, the feature point is detected as an obsolete one and deleted from the map. When the total number of SIFT feature points falls below 40, new feature points are initialized into the map. It is noted that each feature point contains the SIFT descriptor and the feature's x , y and z coordinates at the time of being added. To retain a longer feature life span, we use a Gaussian kernel weighted sampling scheme to randomly draw features from the inliers (identified by the VRO in the PC computation process). The Gaussian kernel has its mean

located at the image center and a standard deviation σ that allows all image pixels reside within the 2σ range (i.e., 95% confidence ellipse). New features are randomly selected from the inliers using the roulette wheel selection scheme based on the Gaussian kernel. The feature sampling scheme utilizes the fact that a feature closer to the image center is more likely to be visible in the next a few time steps (i.e., has a longer life span) and is thus added to the map with a higher probability. The sampling scheme reduces the chance of adding to the map short-life feature points that are inefficient for the EKF tracking.

2) SIFT Feature Correspondences—In this stage, the data association scheme as described in IV.C is performed to create a set of ICMs.

3) Low Innovation Inlier Detection and the 1st Update—In this stage, the 3-point RANSAC procedure (described in IV.D) is used to determine LIIs out of the ICMs. A state update is then performed using the LLIs and v_k^z (if available) by (9).

4) High Innovation Inlier Detection and the 2nd Update—In this stage, the outliers from the previous stage are re-examined (as describe in IV.D) to find HIIs. The HIIs along with v_k^z (if available) are used to perform a further state update by (9).

V. Experimental Results

We carried out a number of experiments with the SC prototype in indoor environments to validate the proposed method. To demonstrate the efficacy of the use of z coordinate observation, we ran three methods—the VRO-based dead reckoning (DR), the EKF without z coordinate observation (EKF) and the EKF with z coordinate observation (EKF-Z)—on each of the experiment and compare their performances. The DR results are used as the baselines to showcase the PE improvement of the proposed method.

We define two metrics—Endpoint Position Error Norm (EPEN) and Average Absolute Z -coordinate Error (AAZE)—for performance evaluation. The EPEN is the error norm of the final camera position represented as the percentage of the camera's trajectory length. The AAZE is the average value of the absolute z error ($|e_z| = |z_e - z_t|$) over all points along the trajectory. Here, z_e and z_t are the estimated and true z coordinates of the camera, respectively.

A. Experiments in laboratory Environment

We carried out five experiments in our lab (Fig. 4(a)) equipped with an OptiTrack motion capture system (MCS), which provides the SC's ground truth pose. The MCS consists of 10 Prime 17W motion capture cameras that use 850 nm IR illuminators (the same wavelength as that of the SR4000). To avoid interference between the MCS and the SR4000, we set the MCS to its passive mode (with the IR illuminator off) in the experiments. We built an active target with five IR LEDs arranged in a T-shape and attached the target on top of the SR4000 (Fig. 4(b)). The LEDs were powered by the SC's battery. In each experiment, the MCS detected the five LEDs and computed the SR4000's pose at each capture time point. The poses were used as the ground truth in the experiments. The time stamps of the MCS and the

SR4000 were used to synchronize the data of the MCS and the SR4000. In each of the five experiments, we collected one dataset from the SR4000 and MCS.

We ran the three methods on each of the five datasets and computed their EPENs and AAZEs by using the ground truth data. Due to the RANSAC and random feature selection processes, poses estimated by the methods are suboptimal. This means that the results may be different for each run on the same dataset. For this reason, we ran each method 10 times on each dataset and compare the methods' performances by using the means and standard deviations as the performance metrics. The results are tabulated in Tables IV and V. To demonstrate the performance improvements of the EKF and EKF-Z over the DR, the two methods' EPEN Reductions (EPENRs) and AAZE Reductions (AAZER) for each experiment are computed and tabulated. The average values of these performance metrics over the five experiments are also computed to show the methods' overall performances. From the tables, we can see that the use of z coordinate observation in the EKF substantially reduces the EPENs and AAZEs and results in an overall much better localization performance of EKF-Z. We can also observe that the three methods' EPEN repeatabilities is similar. But the EKF-Z's AAZE repeatability is much better than that of the other two.

Fig. 5 shows the ground truth trajectory and the estimated trajectories of the three methods for experiment 3. It can be seen that the EKF-Z's trajectory is the closest to the ground truth trajectory and has the smallest EPEN. Therefore, the EKF-Z's trajectory may be used to approximate the ground truth trajectory of an experiment. Fig. 6 depicts the absolute values of the pitch estimation errors of the three methods along the path. It can be observed that the EKF-Z bounds the error under 1.7° while the DR and the EKF methods' errors go up to 3.0° and 2.6° , respectively.

B. Experiments in real-world environments

We carried out numerous experiments in various real-world indoor environments in the EIT, ETAS and RBUS buildings on campus. The results show that the use of z coordinate observation greatly improves the PE accuracy. Five of the experiments are reported here. In all experiments, the SC user walked in a looped trajectory and returned to the start point except for experiment 1 where the endpoint position was measured manually. We collected a dataset from the SR4000 for each experiment and ran the three methods on each dataset. In order to inspect the proposed method's activity at each step, we developed a visualization software to visualize the state estimation process by displaying the location and trajectory of the SC, feature map, unmatched features, inliers and outliers. The software created a video clip for the PE process at the end of each run. Fig. 7 shows a screenshot of the video clip (for experiment 2) at the time when the SC was returning to the start point. This visualization software has been helpful in both algorithm development and validation phases in this work. Brief descriptions of the five experiments are given in the following paragraphs.

Experiment 1 was carried out on the 1st floor of the RBUS building. The path-length is ~ 38 meters and the floor is carpeted and tiled. The user of the SC walked with a changing speed and stopped at several points along the path during the experiment.

Experiment 2 was performed on the 2nd floor of the ETAS building. The path-length is ~24 meters and the floor is covered with patterned carpet. The user swung the SC while walking. This is a case mimicking the typical use of a white cane.

Experiment 3 was conducted on another floor in the RBUS building. The path-length is ~26 meters and the floor is tiled. Similar to experiment 1, the user walked with varying speed and stopped at several points.

Experiment 4 was carried out in different area on the 2nd floor of the ETAS building. The path-length is ~23 meters and the floor is carpeted. The user walked with a slow speed (~0.20 m/s).

Experiment 5 was carried out on the 1st floor of the EIT building. The path-length is ~35 meters and the floor is tiled. The user walked with a slow speed (~0.20 m/s) and swung the SC while walking.

The numerical results (EPENs and AAZEs) of the experiments are tabulated in Tables VI and VII. It is noted that we used the poses estimated by the EKF-Z method to approximate the ground truth poses in computing the EPENs and AAZEs. In term of EPEN, the EKF method improves the results of the DR for experiments 1, 2, 4 and 5. However, it substantially degrades the DR result for experiment 3. As a result, it has a roughly same average EPEN as the DR and a larger average σ than the DR, meaning a same performance with a worse repeatability. But the average EPENR indicates that the EKF still has a better overall performance. The EKF-Z method substantially outperforms the DR and EKF methods in all experiments in term of EPEN and AAZE. This demonstrates that the use of z coordinate observation for EKF state estimation results in a much smaller EPEN and AAZE. In particular, the AAZE of the EKF-Z method is consistently below 2 cm, which is about the SR4000 camera's range measurement accuracy.

Based on the poses estimated by the EKF-Z method, we registered the SR4000's range data to build a point cloud map of the environment for each experiment and plot the trajectories generated by the three methods. Fig. 8 depicts the result of experiment 1. The SC's trajectories estimated by the DR and EKF methods are below the ground level at the end of the trajectory due to the large negative errors in Z axis. However, the EKF-Z method produces the most accurate poses and thus the most accurate SC trajectories.

C. A real-world wayfinding system based on EKF-Z

Based the EKF-Z method, we built a WayFinding System (WFS) that may be used to guide the SC user to move around in a residential or work environment using a speech interface. We adopted a client-server architecture and used a server computer for wayfinding computation. A handtop computer (the user-worn client computer) and a server (desktop) computer were connected through a wireless router. The server used an Nvidia GeForce GTX 770 GPU card to speed up SIFT feature extraction. Fig. 9 depicts the diagram of the software of the WFS. The floorplan of a home environment (Fig. 10) is used to locate the user and Points Of Interest (POIs) and determine the Navigational Commands (NCs). The user indicates the destination by speaking to the SC. The Speech Interface Module (SIM)

converts the speech into a file containing the name of the destination (e.g., home office) and sends the file to the server computer via Wi-Fi. The POI-Coordinate Mapping Module (PCMM) determines the coordinate of the destination for the EKF-Z SLAM Module (ESM). The ESM then acquires data from the SR4000 through the handtop computer using the SR4K Receiver and SR4K Acquisition & Transmission Modules (SATM) and computes the SC's pose by the EKF-Z method. Using the SC's location and the floorplan, the PCMM determines the NCs and/or nearby POIs and sends them as a text file to the SIM, which then converts the text file into speech and makes announcement to the user via the Bluetooth headset. Handshaking protocol is designed and used to ensure error-free data communication between the SR4k Receiver module and SATM and between the SIM and PCMM.

Fig. 10 shows the trajectory generated by the WFS for one of the navigation tasks performed in the home environment. In this task, the SC user walked from the laundry room to the home office. On the way to the destination, the system made announcements for the POIs and NCs, such as "kitchen ahead", "refrigerator on your right", "living room", "go straight ahead", "turn left/right", etc. The point cloud map produced by the EKF-Z method is depicted in Fig. 11 with the trajectories of the three methods. We can observe that: (1) the trajectories generated by the VRO and EKF methods ditch into the ground in the living room; and (2) both trajectories run into the bathroom's wall due to the large pose errors at the end of the navigation tasks. However, the trajectory of the EKF-Z passes through the doorway of the home office and is always above the ground. The result indicates that by observing the camera's z coordinate the EKF reduced not only the z error but also the x and y errors because the z observation reduces the filter's covariance matrix.

As the home environment is feature-rich, the EKF-Z method worked very well for all other tests. The results are omitted for conciseness. It is noted that the starting points are assumed to be known to the WFS in these experiments. The technique for obtaining the starting points is beyond the scope of this paper.

VI. Conclusion

A new 6-DOF PE method is presented for the localization of an RNA in an indoor environment for the visually impaired. The method consists of two components—egomotion estimation by the VRO and state estimation by an EKF. The VRO processes the SR4000 camera's intensity and range data to estimate the camera's egomotion. The characterization study of the VRO reveals that: (1) the PCE of the VRO approximately follows a Gaussian distribution; and (2) the SR4000's non-zero incident angle results in a relatively large error in pitch measurement. Therefore, an EKF is devised to track some SIFT features sampled from previous images and the z coordinate of the camera. Experimental results show that the proposed method results in much smaller pose errors than that of the dead reckoning or standard EKF method. In particular, the use of z coordinate observation in the EKF process greatly reduces the SC's final position error and the error in Z axis. As a result, the trajectory estimated by the proposed method is close to the ground truth path. Based on the 6-DOF PE method, we developed a wayfinding system and used the system to guide a person in a home environment. Using the floorplan and locations of the appliances, furniture, etc, the

wayfinding system sends both navigational commands and points of interest to the person all the way to the destination. The experimental results show that the entire system performed well in the home environment.

The proposed method can be applied to a mobile robot for PE and map-building. In term of future research, the geometric feature tracking approach can be employed to reduce the position errors in X and Y axes and the related rotation errors if a vertical plane can be extracted from the camera's range data. Also, a loop closure detection method may be used to further reduce the pose errors if the trajectory contains a loop.

It is noted that the SC prototype is in its first version. The SR4000 and battery account for the major weight of the device. A custom built SR4000 for the RNA (instead of robotic application) will largely reduce the weight. In the next version, we will consider to use a user-worn battery to power the system and reduce the device weight. Currently, some codes are in Matlab. Due to this limitation, the system has not yet achieved a real-time performance. We are translating the matlab codes into C codes. We expect that the system will have a real-time performance with a frame rate of ~12 FPS in the near future.

Acknowledgments

This work was supported in part by the NSF under Grant IIS-1017672, in part by the National Institute of Biomedical Imaging and Bioengineering and the National Eye Institute of the NIH under Award R01EB018117, and in part by NASA under Award NNX13AD32A. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

1. Estimated Statistics on Blindness and Vision Problems. New York: National Society to Prevent Blindness; 1979.
2. Tsukada K, Yasumura M. Activebelt: Belt-Type Wearable Tactile Display for Directional Navigation. *Proc. Ubiquitous Computing*. 2004:384–399.
3. Balachandran W, Cecelja F, Ptasiński P. A GPS Based Navigation Aid for the Blind. *Proc. International Conference on Applied Electromagnetics and Communications*. 2003:34–36.
4. Ivanov RS. Algorithm for GPS Navigation, Adapted for Visually Impaired People. *Proc. IASK International Conference E-Activity and Leading Technologies*. 2009
5. Sendero Group. Mobile Geo Navigation Software. 2009 [Online]. Available: <http://www.senderogroup.com>.
6. Ye C. Navigating a Portable Robotic Device by a 3D imaging sensor. *Proc. IEEE Sensors Conference*. 2010:1005–1010.
7. Kulyukin V, Gharpure C, Nicholson J, Osborne G. Robot-Assisted Wayfinding for the Visually Impaired in Structured Indoor Environments. *Autonomous Robots*. 2006; 21(1):29–41.
8. Hesch JA, Roumeliotis SI. Design and Analysis of a Portable Indoor Localization Aid for the Visually Impaired. *The International Journal of Robotics Research*. 2010; 29(11):1400–1415.
9. Ye, C.; Bruch, M. *Proc. Unmanned Systems Technology XII Conference at the SPIE Defense, Security, and Sensing Symposium*. Orlando, FL: 2010. A Visual Odometry Method Based On The Swissranger SR-4000.
10. Tamjidi A, Ye C, Hong S. An Extended Kalman Filter for Pose Estimation of a Portable Navigation Aid for the Visually Impaired. *Proc. IEEE International Symposium on Robotic and Sensors Environments*. 2013:178–183.
11. Durrant-Whyte H, Bailey T. Simultaneous Localization and Mapping: Part I. *IEEE Robotics & Automation Magazine*. 2006; 13(2):99–110.

12. Bailey T, Durrant-Whyte H. Simultaneous Localization and Mapping (SLAM): Part II. *IEEE Robotics & Automation Magazine*. 2006; 13(3):108–117.
13. Davison AJ, et al. MonoSLAM: Real-time Single Camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* 2007 Jun.29(6):1052–1067. [PubMed: 17431302]
14. Grisetti G, et al. A Tutorial on Graph-Based SLAM. *IEEE Intelligent Transportation Systems Magazine*. 2010; 2(4):31–34.
15. Huang, et al. Visual odometry and mapping for autonomous flight using an RGB-D camera. *Proc. International Symposium on Robotics Research*. 2011:1–16.
16. Pradeep V, Medioni G, Weiland J. Robot Vision for the Visually Impaired. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*. 2010:15–22.
17. Lee YH, Medioni G. RGB-D Camera Based Navigation for the Visually Impaired. *Proc. Workshop on RGB-D: Advanced Reasoning with Depth Camera*. 2011
18. Manuel Saez J, Escolano F, Penalver A. First Steps Towards Stereo-Based 6DOF SLAM for the Visually Impaired. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*. 2005:23–23.
19. Nister D, Naroditsky O, Bergen J. Visual Odometry. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*. 2004
20. Civera J, et al. 1-Point RANSAC for Extended Kalman Filtering: Application to Real-Time Structure from Motion and Visual Odometry. *Journal of Field Robotics*. 2010; 27(5):609–631.
21. Civera J, Davison AJ, Montiel J. Inverse Depth Parametrization for Monocular SLAM. *IEEE Transactions on Robotics*. 2008; 24(5):932–945.
22. Hervier T, et al. Accurate 3D maps from depth images and motion sensors via nonlinear Kalman filtering. *Proc. IEEE Intelligent Robots and Systems*. 2012:5291–5297.
23. Ling L, et al. An iterated extended Kalman filter for 3D mapping via Kinect camera. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013:1773–1777.
24. Weingarten J, Siegwart R. 3D SLAM using planar segments. *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2006:3062–3067.
25. Trevor A, et al. Planar surface SLAM with 3D and 2D sensors. *Proc. IEEE International Conference on Robotics and Automation*. 2012:3041–3048.
26. Salas-Moreno RF, et al. Dense planar SLAM. *Proc. IEEE International Symposium on Mixed and Augmented Reality*. 2014:157–164.
27. Lowe DG. Distinctive Image Features From Scale-Invariant Keypoints. *International Journal of Computer Vision*. 2004; 60(2):91–110.
28. Arun KS, et al. Least Square Fitting of Two 3-D Point Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1987; 9(5):698–700. [PubMed: 21869429]
29. Hong S, et al. Performance Evaluation of a Pose Estimation Method based on the SwissRanger SR4000. *Proc. IEEE International Conference on Mechatronics and Automation*. 2012:499–504.
30. Adam M. Lidar Design, Use, and Calibration Concepts for Correct Environmental Detection. *IEEE Transactions on Robotics and Automation*. 16(6):753–761.
31. Oggier, T., et al. *Proceedings of Opto-Conference*. Germany: Nürnberg; 2004. 3D-Imaging in Real-Time with Miniaturized Optical Range Camera.
32. Qian X, Ye C. NCC-RANSAC: A Fast Plane Extraction Method for Navigating a Smart Cane for the Visually Impaired. *Proc. IEEE International Conference on Automation Science and Engineering*. 2013:267–273.
33. Deans M, Kunz C, Sargent R, Park E, Pedersen L. Combined Feature based and Shape based Visual Tracker For Robot Navigation. *IEEE Aerospace Conference*. 2005:339–346.

Biographies

Cang Ye (S'97–M'00–SM'05) received the B. E. and M. E. degrees from the University of Science and Technology of China, Hefei, Anhui, in 1988 and 1991, respectively, and the Ph.D. degree from the University of Hong Kong, Hong Kong in 1999. He is currently a Professor with the Department of Systems Engineering, University of Arkansas at Little Rock. His research interests are in mobile robotics, intelligent system, and computer vision.

Soonhac Hong (M'15) received the B.S. and M.S. degrees in Mechanical Engineering from Hanyang University, Seoul, South Korea, in 1994 and 1996, the M.S. degree in Computer Science from Columbia University, New York City, NY, USA, in 2010, and the Ph.D. degree in Engineering Science and Systems from University of Arkansas at Little Rock, AR, USA, in 2014. He is currently a Senior Researcher of Systems, Algorithms & Services at Sharp Labs of America, Camas, WA, USA. His research interests include simultaneous localization and mapping, motion planning, computer vision, and machine learning.

Amirhossein Tamjidi received the B.S. degrees in Electrical Engineering from Tabriz University, Iran, in 2006, and the M.S. degree in Electrical Engineering from K.N. Toosi University of Technology, Iran, in 2009. At the time of doing this work, he was a Ph.D. student with the Department of Systems Engineering, University of Arkansas at Little Rock. His research interests include SLAM, 2D/3D computer vision, motion planning, and image processing.

Fig. 1.
The Smart Cane and its coordinate systems: the camera, floor plane and world coordinates are denoted by subscripts C , f , and W , respectively.

Fig. 2.

Range measurement error (and intensity) vs camera pitch angle: The mean distance error also increases with the pitch angle because the true distance happens to be in the region where the measurement error increases with the true distance. The error will decrease if the true distance becomes bigger [31].

Fig. 3.
Diagram of the proposed EKF pose estimation method

(a) Motion Capture System

(b) Active IR target

Fig. 4.
The MCS and active target on top of the SR4000 camera

Fig. 5.

The ground truth (GT) trajectory and the trajectories generated by the three methods

Fig. 6.
Pitch estimation errors of the three methods

$$[x \ y \ z \ q1 \ q2 \ q3 \ q4] = [0.86 \ -1.32 \ 0.77 \ -0.7734 \ 0.5819 \ -0.2391 \ 0.07]$$

Fig. 7.

A screenshot of the video clip of experiment 2 showing SIFT feature tracking. Left: '+' with an ellipse denotes the projection of a feature point on the intensity image. The ellipse depicts the position uncertainty within which a matched SIFT feature is anticipated. Blue indicates an unmatched feature point. Thick red indicates an LII and pink an HII. Right: The SC trajectory in the X-Y coordinate system. The feature points are depicted by red '+'. Their uncertainty ellipses are omitted for simplicity.

Fig. 8.

Experiment in a real-world indoor environment: Texture mapping is performed to the point cloud map by using the intensity images of the SR4000 camera. Green – DR, blue – EKF, red – EKF-Z; s – starting point, e – endpoint.

Fig. 9.
Diagram of the Software of the wayfinding system

Fig. 10.
Estimated trajectory of the SC for a trip from the laundry room to the home office.

(a) Snapshots of kitchen and living room

(b) Point cloud map with the trajectories of DR, EKF and EKF-Z

Fig. 11.

Validation of the wayfinding system in a home environment. The SC user walked from s to e . Green – DR, Blue – EKF, Red – EKF-Z; s – starting point, e – endpoint.

TABLE I

Noise Ratio of Raw and Filtered SR4000 Data

Noise Ratio (%)		Raw Data	Filtered Data	Noise Reduction
Intensity	Maximum	20.66	9.88	52%
	Mean	2.28	0.99	57%
	Minimum	0.47	0.42	11%
Range	Maximum	14.66	7.96	46%
	Mean	1.50	0.55	63%
	Minimum	0.31	0.12	61%

The maximum, mean and minimum are computed from 176×144 pixels.

TABLE II

PCE Errors in Rotation Measurement

MV: (μ , σ)	Roll ϕ ($^{\circ}$)	Pitch θ ($^{\circ}$)	Yaw ψ ($^{\circ}$)
TV: (ϕ , θ , ψ)			
(3, 0, 0)	(0.17, 0.11)	(0.06, 0.07)	(0.04, 0.06)
(6, 0, 0)	(0.16, 0.10)	(0.02, 0.06)	(0.03, 0.07)
(9, 0, 0)	(0.07, 0.10)	(0.07, 0.06)	(0.05, 0.06)
(12, 0, 0)	(0.02, 0.11)	(0.09, 0.07)	(0.01, 0.07)
(15, 0, 0)	(0.00, 0.10)	(0.05, 0.08)	(0.11, 0.09)
(0, 3, 0)	(0.05, 0.03)	(0.42, 0.06)	(0.08, 0.05)
(0, 6, 0)	(0.06, 0.04)	(0.40, 0.10)	(0.08, 0.06)
(0, 9, 0)	(0.08, 0.04)	(0.53, 0.16)	(0.10, 0.07)
(0, 12, 0)	(0.13, 0.06)	(0.76, 0.22)	(0.31, 0.13)
(0, 15 0)	(0.25, 0.06)	(0.91, 0.34)	(0.26, 0.15)
(0, 0, 3)	(0.02, 0.07)	(0.09, 0.13)	(0.17, 0.11)
(0, 0, 6)	(0.02, 0.08)	(0.09, 0.14)	(0.21, 0.11)
(0, 0, 9)	(0.01, 0.08)	(0.18, 0.16)	(0.14, 0.16)
(0, 0, 12)	(0.03, 0.09)	(0.18, 0.20)	(0.15, 0.22)
(0, 0, 15)	(0.01, 0.12)	(0.22, 0.22)	(0.23, 0.27)

MV: Measured Values, TV: True Values, μ : mean error, σ : standard deviation, ϕ , θ , and ψ mean ϕ , θ and ψ (is dropped for simplicity).

TABLE III

PCE Errors in Translation Measurement

MV: (μ , σ)	X (mm)	Y (mm)	Z (mm)
TV: (X, Y, Z)			
(100, 0, 0)	(9.8, 4.0)	(0.4, 1.4)	(3.3, 2.4)
(200, 0, 0)	(5.6, 5.5)	(2.7, 1.7)	(3.9, 2.9)
(300, 0, 0)	(10.5, 5.2)	(3.4, 1.6)	(7.7, 3.6)
(400, 0, 0)	(2.8, 8.9)	(4.7, 2.7)	(6.9, 6.8)
(0, 100, 0)	(1.4, 2.8)	(4.3, 1.7)	(3.4, 2.7)
(0, 200, 0)	(2.8, 2.8)	(6.2, 1.7)	(2.5, 3.1)
(0, 300, 0)	(0.9, 2.7)	(7.7, 1.8)	(0.3, 3.5)
(0, 400, 0)	(3.3, 3.1)	(9.5, 1.8)	(3.3, 3.7)

MV: Measured Values, TV: True Values, μ : mean error, σ : standard deviation, X, Y, Z: changes of position (is dropped for simplicity).

TABLE IV

EPENs of Experiments in the Lab

Experiment (path-length)	EPEN (%): (μ , σ)				EPENR (%)	
	DR	EKF	EKF-Z	EKF-Z	EKF	EKF-Z
1 (13.80 m)	(2.45, 0.35)	(2.11, 0.26)	(0.97, 0.27)		13.9	60.4
2 (12.27 m)	(2.05, 0.20)	(1.96, 0.52)	(1.73, 0.36)		4.4	15.6
3 (14.10 m)	(3.39, 0.21)	(3.80, 0.29)	(1.61, 0.45)		-12.1	52.5
4 (12.83 m)	(3.48, 0.41)	(3.80, 0.39)	(0.96, 0.22)		-9.2	72.4
5 (11.94 m)	(2.02, 0.40)	(1.56, 0.27)	(1.66, 0.49)		22.8	17.8
Average	(2.68, 0.31)	(2.65, 0.35)	(1.39, 0.36)		3.96	43.7

μ and σ are the mean and standard deviation over the 10 runs. A negative EPENR indicates a degraded performance.

TABLE V

AAZEs of Experiments in the Lab

Experiment (path-length)	AAZE (cm): (μ , σ)			AAZER (%)	
	DR	EKF	EKF-Z	EKF	EKF-Z
1 (13.80 m)	(8.1, 0.9)	(9.2, 2.1)	(1.3, 0.2)	-13.6	84.0
2 (12.27 m)	(4.1, 1.3)	(5.7, 3.3)	(1.8, 0.2)	-39.0	56.1
3 (14.10 m)	(13.5, 1.4)	(16.8, 2.5)	(1.8, 0.2)	-24.4	86.7
4 (12.83 m)	(13.2, 1.5)	(13.6, 1.6)	(1.4, 0.1)	-3.0	89.4
5 (11.94 m)	(6.6, 1.7)	(6.1, 2.1)	(1.1, 0.1)	7.6	83.3
Average	(9.1, 1.4)	(10.3, 2.3)	(1.5, 0.1)	-14.5	79.9

μ and σ are the mean and standard deviation over the 10 runs. A negative AAZER indicates a degraded performance.

TABLE VI

EPENs of Experiments outside the lab

Experiment (path-length)	EPEN (%): (μ , σ)				EPENR (%)	
	DR	EKF	EKF-Z	EKF	EKF	EKF-Z
1 (38.24 m)	(3.65, 0.31)	(3.32, 0.80)	(1.61, 0.46)	9.0	55.9	
2 (24.29 m)	(2.72, 0.93)	(2.21, 0.39)	(1.16, 0.59)	18.8	57.4	
3 (25.71 m)	(5.48, 0.47)	(7.64, 1.19)	(4.05, 0.97)	-39.4	26.1	
4 (22.48 m)	(2.60, 0.37)	(1.32, 0.53)	(0.77, 0.30)	49.2	70.4	
5 (34.91 m)	(1.86, 0.21)	(1.64, 0.58)	(1.16, 0.44)	11.8	37.6	
Average	(3.26, 0.46)	(3.23, 0.70)	(1.75, 0.55)	9.9	49.5	

TABLE VII

AAZEs of Experiments outside the lab

Experiment (path-length)	AAZE (cm): (μ , σ)			AAZER (%)	
	DR	EKF	EKF-Z	EKF	EKF-Z
1 (38.24 m)	(80.2, 4.7)	(44.7, 7.9)	(1.6, 0.1)	44.3	98.0
2 (24.29 m)	(20.3, 3.6)	(12.0, 2.8)	(1.2, 0.1)	40.9	94.1
3 (25.71 m)	(52.8, 8.1)	(53.6, 12.4)	(1.3, 0.2)	-1.5	97.5
4 (22.48 m)	(14.3, 2.4)	(11.6, 4.9)	(1.1, 0.1)	18.9	92.3
5 (34.91 m)	(52.7, 1.9)	(20.3, 3.6)	(1.0, 0.0)	61.5	98.1
Average	(40.1, 4.1)	(28.4, 6.3)	(1.2, 0.1)	32.8	96.0