Multimed Tools Appl
DOI 10.1007/s11042-015-3204-2

Developing a way-finding system on mobile robot assisting visually impaired people in an indoor environment

Quoc-Hung Nguyen[1,2] · Hai Vu[1] · Thanh-Hai Tran[1] · Quang-Hoan Nguyen[3]

Received: 12 September 2014 / Revised: 27 November 2015 / Accepted: 29 December 2015

Abstract A way-finding system in an indoor environment consists of several components: localization, representation, path planning, and interaction. For each component, numerous relevant techniques have been proposed. However, deploying feasible techniques, particularly in real scenarios, remains challenging. In this paper, we describe a functional way-finding system deployed on a mobile robot to assist visual impairments (VI). The proposed system deploys state-of-the-art techniques that are adapted to the practical issues at hand. First, we adapt an outdoor visual odometry technique to indoor use by covering manual markers or stickers on ground-planes. The main purpose is to build reliable travel routes in the environment. Second, we propose a procedure to define and optimize the landmark/representative scenes of the environment. This technique handles the repetitive and ambiguous structures of the environment. In order to interact with VI people, we deploy a convenient interface on a smart phone. Three different indoor scenarios and thirteen subjects are conducted in our evaluations. Our experimental results show that VI people, particularly VI pupils, can find the right way to requested targets.

Keywords Visual odometry · Environmental map · Vision-based localization · Assistive technology for visually impaired · Indoor navigation

✉ Quoc-Hung Nguyen
quoc-hung.nguyen@mica.edu.vn

[1] International Research Institute MICA, HUST - CNRS/UMI - 2954 - INP Grenoble, Hanoi University of Science and Technology, Hanoi, Vietnam

[2] Thai Nguyen Medical College, Thai Nguyen, Vietnam

[3] Hung Yen University of Technology and Education, Hung Yen, Vietnam

Published online: 19 January 2016

1 Introduction

1.1 Motivation and system introduction

People with vision disability meet many difficulties in perceiving and understanding environments. Nowadays, developing assistive systems, particularly, way-finding (navigational) systems supporting visually impaired (VI) people, has been an attractive topic for both industry and research communities [22]. Several assistive technologies have been intensively proposed to address critical issues in both indoor and outdoor environments. While outdoor navigation systems conveniently rely upon Global Positioning System (GPS), indoor systems face many challenges because GPS signals can not be received. As surveyed in [8, 22], although many efforts have been proposed to develop indoor navigation systems, none of them have been deployed at a large scale because of the challenges in cost, accuracy and usability. Therefore, deploying a way-finding system in indoor environments is still considered an open problem. The system proposed here, as shown in Fig. 1, helps VI people navigate in indoor environments by holding onto a mobile robot. Importantly, the system successfully handled practical issues in real scenarios. The proposed system has been evaluated through a series of pilot experiments in three indoor scenarios with the participation of VI pupils.

Key issues consist of: (1) How to represent indoor environments; (2) How to handle accurate localization and path planning; (3) How to interact with end-users. The first and second issues are typically based on modern SLAM (Simultaneous Localization and Mapping) algorithms [2]. While SLAM methods have now reached a state of considerable maturity [2], they still face many difficulties with unstructured environments, especially in situations that GPS-like solutions are unavailable or unreliable. An extensive research on SLAM is not in the scope of this paper. In this research, we simply consider the first and second as issues of a place recognition system. The fact is that reliable place recognition is a difficult

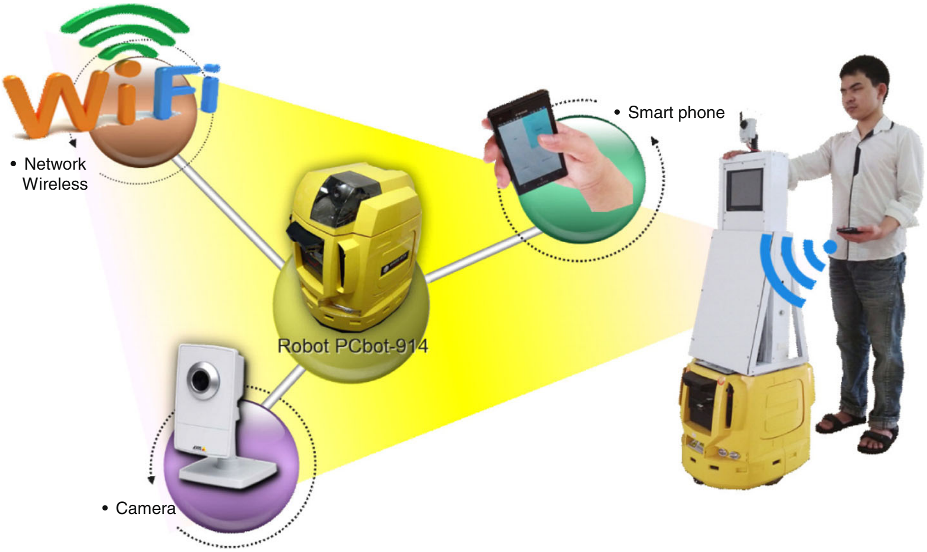


Fig. 1 A snapshot of the proposed system. One camera is rigidly attached to the moving robot. A smartphone helps VI user to communicate with robot

problem, especially in environments with repetitive and ambiguous structures. To solve this, we divide the proposed system into two parts: off-line phase and on-line phase. The off-line phase builds and represents an environmental map by the robot travel routes and landmarks given by representative scenes. The on-line phase associates visual data, matching current observation to places defined previously in the off-line phase. The proposed system interacts with an end-user through a smart phone. This device sends requested targets and receives navigational feedback via vibration signals. The proposed system is wrapped on a mobile robot. A monocular consumer-graded camera rigidly attaches to the moving platform. Consequently, VI people will move following the robot. Prototype of the proposed system is shown in the Fig. 1.

The proposed system has been tested with VI people in various indoor environments, such as the hallway of a public building, a library, and a dormitory of school for VI pupils. Such environments offer many challenges which commonly appear in daily practices such as narrow passage-ways, low-lighting conditions, and suddenly moving objects. Other aspects of complex environments, such as elevators and staircases, are outside of the scope of this work. The evaluation results show that the proposed system is usable for VI people, particularly when they move in unfavorable environments.

1.2 Solution requirements

Before describing the proposed system, we define two requirements that the proposed solution should meet.

- Accuracy: In order to support the blind user safely and efficiently, the system must be accurate in terms of robot movement and localization. Through a practical survey, most VI people accept that a navigation system may have, in the worst cases, a positioning error of about 50cm. This is an acceptable margin of error for the safe movement of the VI people.
- Usability and Usefulness: The use of the system should be as natural as possible for end-users. This means it is easy to learn and robust to small changes in the environment. The information delivered by the system must be useful to allow the users to navigate, even if it is their first time in the place.

This paper is organized as follows. The next section presents related works. Section 3 shows the overall proposed system, then describes in detail its components. Section 4 describes an experimental setup, and evaluation results in three different environments. Finally, we discuss and conclude the work, as well as providing possible research directions in the future.

2 Related works

Systematic reports on assistive technologies for VI people are found in several textbooks such as [1, 22]. In the interest of space, we limit this review to works which are close to our research. They are divided into following categories.

General assistive technologies for VI people Developing localization and navigation tools for visually impaired people has received much attention in the autonomous robotics community [6]. Most of the works to date focus on finding efficient localization solutions based on positioning data from different sensors such as GPS, laser, Radio Frequency

Identification (RFID), vision, or a fusion of them. Loomis et al. [21] surveyed the efficiency of GPS-based navigation systems in supporting visually impaired people. The GPS-based systems share similar problems: low accuracy in urban environments (localization accuracy is limited to approximately 20 m), signal loss due to a multi-path effect or line-of-sight restrictions due to the presence of buildings or even foliage. Kulyukin et al. [16] proposed a system based on Radio Frequency Identification (RFID) for aiding the navigation of visually impaired people in indoor environments. The system requires the design of a dense network of location identifiers. Helal et al. [12] proposed a wireless pedestrian navigation system. They integrated several types of signals such as voice, wireless networks, Geographic Information System (GIS) and GPS to provide the visually impaired people with an optimized route. Recent advanced techniques in computer vision offer substantial improvements with respect to localization and navigation services in known or unknown environments. The vision-based approaches not only offer safe navigation, but also provide a very rich and valuable description of the environment. For example in [3], Bigham et al. develop an application named LocateIt, which helps blind people locate objects in indoor environments. In [31], ShelfScanner is a real-time grocery detector that allows online detection of items in a shopping list.

Robot assisted way-finding for the visually impaired people Many autonomous service robots have been deployed to assist and serve in human environments, such as janitorial robots [7], hospital aides [13], museum tour guides, and robots that aid the blind [17] and the elderly [19]. The idea of using a robot for guiding VI people appeared in several systems. In [15], Kulyukin et al. utilized RFID sensors to deploy such a robot for navigation services in indoor environments. Two different indoor environments have been examined in [15]. The authors in [18] deployed a wireless sensor network on a mobile robot. This work offers new ways to monitor the environment and do so continuously and invisibly. Some approaches use a stereo camera to build vision-based navigation systems. As opposed to the above works, in this research we only use a monocular sensor to build a vision-based way-finding system.

Vision-based SLAM and Place recognition techniques In terms of visual mapping and localization, Alcantarilla [1] utilizes well-known techniques, such as Simultaneous Localization and Mapping (SLAM) and Structure from Motion (SfM) to create a 3-D Map of an indoor environment. Visual descriptors, such as Gauge-Speeded Up Robust Features and G-SURF, are utilized to mark local coordinates on the constructed 3-D map. Instead of building a prior 3-D map, Liu et al. [20] use a pre-captured reference sequence of the environment. Given a new query sequence, their system attempts to find the corresponding set of indices in the reference video. Some wearable-cameras based on visual SLAM techniques have also been proposed. Pradeep et al. [27] present a head-mounted stereo-vision platform for detecting obstacles on the path and warn subjects about their presence. They incorporate visual odometry and feature-based metric-topological SLAM. Murali et al. [23] estimate the user's location relative to the crosswalks in the intersections of roads. The system of Murali et al. in [23] requires supplemental images from Google Map services, therefore its applicability is limited to the outdoors. These works presume that a SLAM-based approach is ideally suited to the task of guiding the visually impaired people, because SLAM combines two key elements required for friendly and a widely user-applicable system: map building and self-location. However, the complexity of the map building task varies as a function of the environment's size. In some cases, a map can be acquired from the visual sensor, but in other cases, the map must be constructed from other sensor modalities, such as GPS, WIFI [5].

Matching a current view to a position on the created map seems to be the hardest problem [2, 9]. Work towards appearance-based place recognition has been conducted in [29]. A common place recognition system (e.g., [29]) is adopted from text retrieval systems. Such a system introduces the concept of so-called visual vocabulary. This idea was later extended to vocabulary trees by [25], allowing for the efficient use of a large number of vocabularies [28] for city-scale place recognition. Recently, Maddern et al. reported an improvement to the robustness by incorporating odo-metric information into the place recognition process. In [30], the authors proposed BRIEF-GIST combinations, a highly simplified appearance-based place recognition system based on the BRIEF descriptors and GIST features [26]. In our view, an incremental map is able to support us on improving matching results. When new observations arrive, these new observations must be locally and globally consistented with the previously constructed map. To this end, we employ the loop closure algorithms based on a Fast Appearance-Based Map (FAB-MAP) proposed by [5, 24].

3 Proposed way-finding system

3.1 Main functionalities of the proposed system

We refer to categories for the navigation systems as defined in a survey [8]. According to this survey, relevant works employ either *path integration* or *a landmark-based* navigation system. Many other related works may also use a combination of both *path integration* and *landmark-based* navigation. Whereas *path integration* does not need deployment or to pre-build a cognitive map (or a map of the environment), *landmark-based* navigation approaches rely on perceptual cues together with an external cognitive map. Matching these definitions, our proposed way-finding system could be a *landmark-based* navigation. The cognitive mapping is concerned with an off-line process, in which a map of the environment is created. In an online navigating system, many decisions need to be made based on knowledge of the environment stored in the cognitive map. Note that we use only a vision sensor in the proposed system. This makes the system simpler than any combination/fusion of sensors. It is the most effective way to create a cognitive map, because details about the environment can be acquired in a relatively short time. We describe relevant components of the proposed system in following sections:

- **Environment representation** (Section 3.2): In order to locate the user and find a path to go to a desired location, the map of environment must be computed and stored.
- **User localization** (Section 3.3): To assist the user moving to the desired location, the system has to determine their current position on the pre-built map.
- **Path planning and navigation** (Section 3.4): Once the user is localized, the path from their current position to the desired destination is computed and indications of navigation sent to the user.
- **System user interaction** (Section 3.5): The user sends a target destination to the system using a touch screen. He/she receives feedback via vibration signals encoding the navigation information to safely travel in the environment.

3.2 Environment representation

Unlike SLAM techniques that build the map and locate vehicles simultaneously, in our case, the map of the environment will be pre-built for way-finding and navigation guidance. The proposed map consists of a set of a representative view/scene along with robot's trajectory route. To obtain this, we utilize two different techniques: one is a visual odometry technique to build a 2-D map of robot routes; another is inspired by the idea of a loop-closure detection algorithm, namely FAB-MAP [5]. This method has been shown to be very efficient for Visual SLAM problem in the literature. It aims to represent the map of environment by a set of locations $\mathcal{L}^N = \{L_1, L_2, ..., L_N\}$. However, unlike the original work that builds the map incrementally at runtime, we build this map off-line. Moreover, the physical position (x_i, y_i) of each location and its corresponding appearance model L_i are added in the pre-built routes. The problem now is to determine (x_i, y_i) and L_i for each $i \in [1, N]$.

To resolve this problem, we design a compact imaging acquisition system to visually capture simultaneously scene and the travel routes in the environment. A schematic view of the system is shown in Fig. 2a. The proposed image acquisition has two cameras. One captures scenes surrounding the environment. The second one captures the travel road. Setting of two cameras is shown in Fig. 2b. These cameras are mounted on a wheel-vehicle, as shown in Fig. 2c, and time-synchronized with pre-selected frame rates. The travel route observation is used to rebuild the vehicle travel, while scene observation is used for computing appearance model of each location. The vehicle will be only used during the off-line phase to build a map of the environment and describe the appearance of locations. Using a vehicle in the off-line phase has the advantage of avoidance of the camera system's vibration. As a consequence, it is a more accurate reconstruction of the travel.

3.2.1 Vehicle travel reconstruction using visual odometry technique

To reconstruct the trajectories of the robot, we rely on a visual odometry technique. A well-known issue for visual odometry techniques is that they need to estimate precisely the correspondences between features of consecutive frames. Once feature correspondences have been established, we can reconstruct the trajectory of the vehicle.

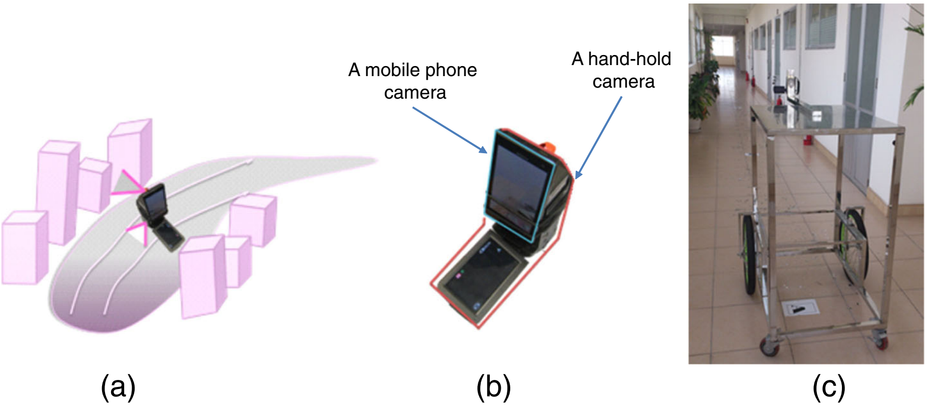


Fig. 2 **a** A schematic view of the visual data collection scheme. **b** The proposed image acquisition system in which a mobile phone camera is attached on rear of a hand-hold camera. **c** The image acquisition system is attached on a wheel vehicle

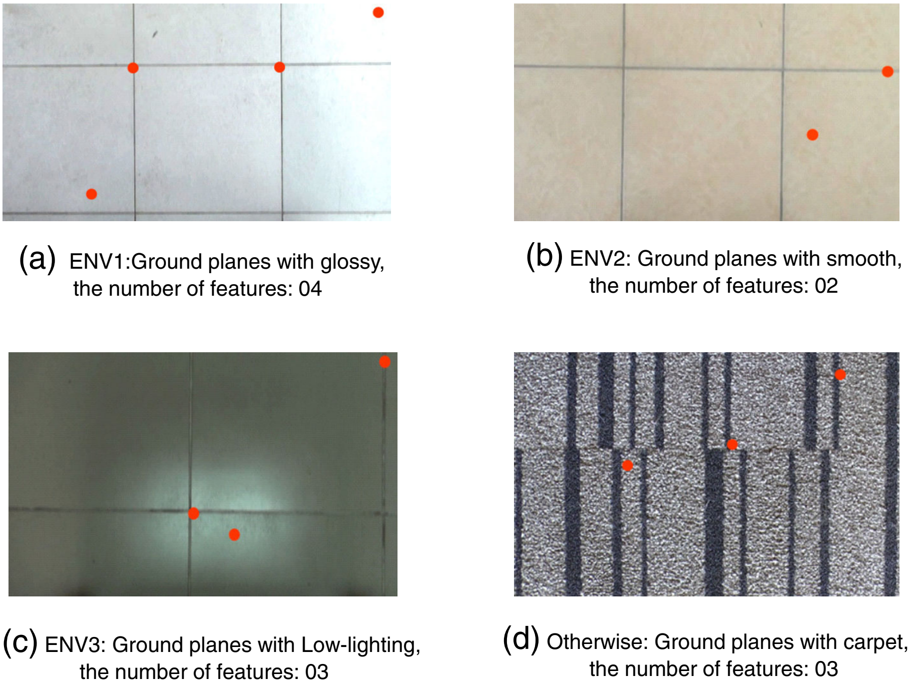

(a) ENV1:Ground planes with glossy, the number of features: 04
(b) ENV2: Ground planes with smooth, the number of features: 02
(c) ENV3: Ground planes with Low-lighting, the number of features: 03
(d) Otherwise: Ground planes with carpet, the number of features: 03

Fig. 3 Number of detected feature points is limited in indoor environments

Among many existing visual odometry techniques, we utilize the method presented in [10] due to its suitability to our context (using a single monocular camera, plane-ground based tracking), and its performance advantages over traditional fundamental matrix estimation. This method is based upon the tracking of ground plane features. It was designed to take into account the uncertainty as to vehicle motion and as to the extracted features. The method has been shown to be efficient in an outdoor application with a camera mounted on a moving car, but its precision is decreased when it is applied indoor environments.

We propose to adapt this algorithm as follows. In an indoor environment, ground planes are often covered by carpet or brick tiling. These materials make the ground-plane as homogenous regions with low texture and low brightness. The appearance of common ground planes are shown in Fig. 3. As a consequence, the number of detected feature points

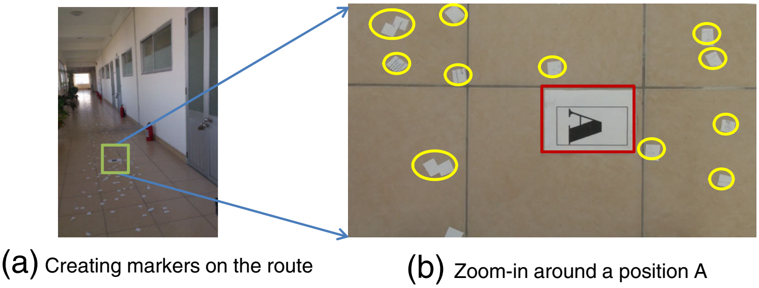

(a) Creating markers on the route
(b) Zoom-in around a position A

Fig. 4 We scatter markers along the travel

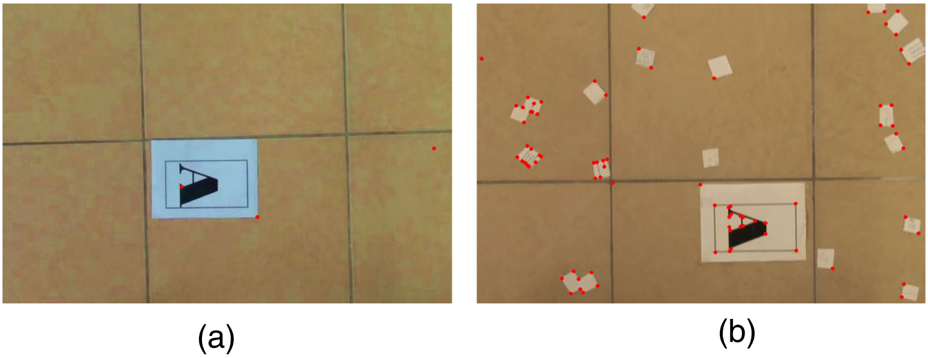


Fig. 5 Results of feature point detection using the same parameters: a) without markers and b) with markers

is not sufficient (Fig. 3) for a precise estimation of the vehicle's travel. To handle this issue, we scatter man markers (or stickers) along the whole journey as shown in Fig. 4.

This simple and feasible solution improves significantly the number of detected key points (Fig. 5) and gives better results of feature detection and matching (see the experimental results section).

3.2.2 Computation of appearance model of locations on the map

Image representation Similar to [5], we represent the appearance of an observation (image) using a "Bag of Word" (BoW) model. Specifically, at time k, an image I_k captured from the scene camera is represented as a collection of visual words $Z_k = \{z_1, z_2, ..., z_{|v|}\}$ where $|v|$ is the size of vocabulary and z_i is a binary variable indicating the presence (or absence) of the ith word of the vocabulary. $Z^k = \{Z_1, Z_2, ..., Z_k\}$ denotes a set of all observations up to time k.

To build vocabulary model, a set of M consecutive images captured along the vehicle travel $I = \{I_1, I_2, ..., I_M\}$ is considered. SURF features are extracted from every image and classified into v clusters using K-Means. In [5], all images captured along the travel are taken into account to build the BoW model.

The related work [1] reports that [5] obtains reasonable results for place recognition over long travel in term of both precision and recall measurements. However, those experiments were implemented in outdoor environments which usually contain discriminative scenes. The original work [5] still has unresolved problems in discriminating scenes to define a visual dictionary. This issue affects the results when we deploy it in indoor environments, where scenes are continuous and not clearly distinctive.

To overcome this problem, we propose a solution to remove similar samples and take only discriminative frames for vocabulary building. Specifically, we extract a subset $I_d \in I$, $I_d = \{I_{i1}, I_{i2}, ..., I_{id}\}$ where each I_{ij} is different from other (see Fig. 6).

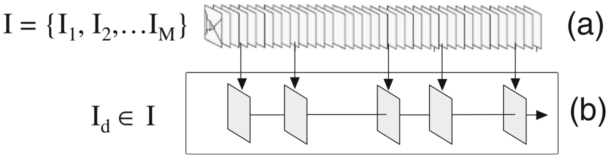


Fig. 6 Selecting a subset of distinctive frames for building bag of word model: **a** Image Sequence. **b** Representative scene

We define a distance function $D(I_{ij}, I_{ik})$ to measure the difference between two images I_{ij} and I_{ik}. Firstly, we describe each image by a GIST descriptor. This descriptor was firstly introduced by [26] and has been shown to be very efficient in scene classification. It is computed by filtering the input image by a set of Gabor filters of different scales and orientations. Then $D(I_{ij}, I_{ik})$ is Euclidean distance between two corresponding GIST descriptors $G(I_{ij}), G(I_{ik})$ (Fig. 7):

$$D(I_{ij}, I_{ik}) = D(G(I_{ij}), G(I_{ik})) \tag{1}$$

Given a frame sequence I, for every frame I_{ij} we compute the distance from this frame to its previous frame I_{ij-1}. I_{ij} will be added in the subset I_d if $D(I_{ij-1}, I_{ij})$ is bigger than a threshold T.

Once the subset I_d is determined, the same procedure done for computing the vocabulary model. As clarified in the original work [5], to capture the co-occurrence of visual words in the scene, a Chow Liu tree is utilized to approximate the probability distribution over these visual words and the correlations among them [4].

Appearance representation of location Each location L_i in the map model $\mathcal{L}^N$ has an associated appearance. We follow the same representation of location presented in [5].

$$L_i : \{p(e_1 = 1|L_i), ..., p(e_{|v|} = 1|L_i)\} \tag{2}$$

where e_q is a hidden variable. The variable e_q is the event that an object which generates observations of type z_i exists.

Key locations We could consider every point on the vehicle travel as locations composing the map. However, if locations are too dense, image matching will be very time consuming, making the localization intractable and unfeasible in a large environment. Uniform sampling is a solution to this problem. However, this will not take environmental characteristics into account (complex sections (turning) requires more samples than simple sections). Therefore image matching and localization will be inaccurate.

We then run FAB-MAP [5] in an off-line phase to determine key locations of the map. Given a scene sequence $I_s = \{I_{s_1}, I_{s_2}, ..., I_{s_n}\}$, at time k, we suppose that the set of locations

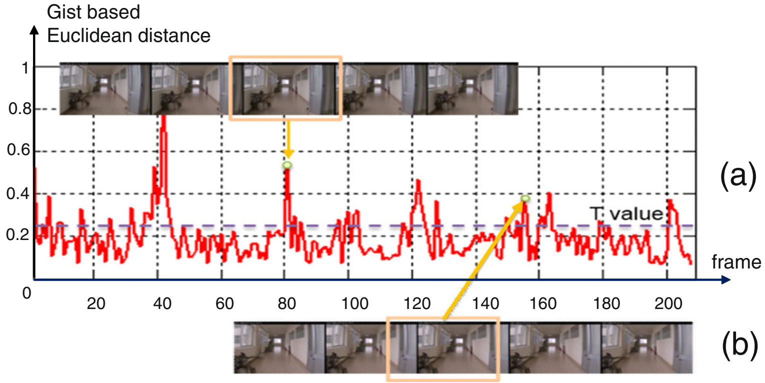


Fig. 7 Determining distinctive frames using GIST descriptor: **a** Dissimilarity between two consecutive frames. A threshold value $T = 0.25$ is pre-selected. **b** Two examples shows the selected key frames and their neighbor frames

on the map is $L^k = \{L_1, L_2, ..., L_{n_k}\}$. We compute $P(L^i|Z^k)$ with Z^k is the observation up to time k using recursive Bayes technique for each $i \in [1, n_k]$.

$$p\left(L_i|Z^k\right) = \frac{p\left(Z_k|L_i\right) p\left(L_i|Z^{k-1}\right)}{p\left(Z_k|Z^{k-1}\right)} \tag{3}$$

We check if $argmax(p(L_i|Z^k))$ is large enough, this means the current location is revisited, we then update the L^k. Otherwise, we create a new location L_{n_k+1} and update the map. Evidently, as scene image and road image are captured simultaneously, we could imply the physical position (x_{n_k+1}, y_{n_k+1}) of the location L_{n_k+1} thank to pre-built vehicle travel using visual odometry technique (Section 3.2.1).

Consequently, the distinctive locations are determined from visual training data. To fill the travel, we incorporate captured images through several trials. For each new trial, we compare the images with the previously visited places which are already indexed in a place database. This procedure calls a loop closure detection, these detections are essential for building an incremental map. Figure 8a shows only few places are marked by the first trial, whereas various places that are updated after the second trial as shown in Fig. 8b.

3.3 User localization

The entire map of environment has been built off-line $\mathcal{L}^N = \{L_1, L_2, ..., L_N\}$. At runtime, given a current observation I_k, the corresponding position on the map is identified through a place recognition procedure. We evaluate $p(L_i|Z^k)$ for all $i \in [1, N]$:

$$p\left(L_i|Z^k\right) = \frac{p\left(Z_k|L_i\right) p\left(L_i|Z^{k-1}\right)}{p\left(Z_k|Z^{k-1}\right)} \tag{4}$$

Where Z^k contains visual words appearing in all observations up to k; and Z_k presents visual words at current time k. A probability $p(Z_k|L_i)$ infers observation likelihood as learnt in the training data. In our system, the current position is assigned to the location L_k* with $k*$ satisfying $argmax(p(L_i|Z^k))$ is bigger than a threshold (through a pre-determined threshold $T = 0.9$).

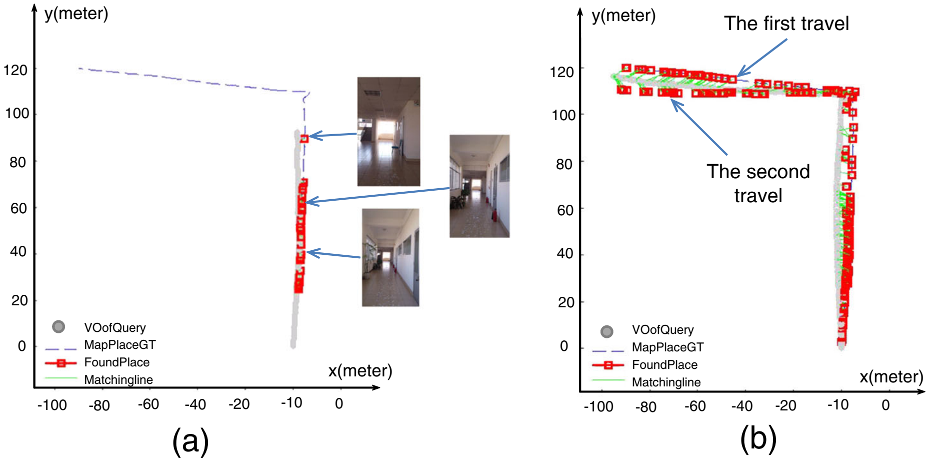


Fig. 8 **a** The locations are created in the first trial. **b** Many new locations are updated after second trial

The Fig. 9 shows an example of the localization procedure. Given an observation as shown in Fig. 9a, the probability $p(L_i|Z^k)$ is shown in Fig. 9c with a threshold $T = 0.9$ whose the maximal probability is $placeID = 18$ (Fig. 9b). A confusion matrix of the matching places for an image sequence is shown in Fig. 9d. This example shows that we can resolve most places in a testing phase.

3.4 Path planning and navigation

3.4.1 Path planning

Path-planning is a popular research topic which could be solved using the well-known A* [11] algorithm or based on some variants, such as Iterative-deepening-A (IDA) [14]. Using the well-known A* algorithm for finding a path, we need to define uniformly shaped grids in the environment and calculate geometric computation through visibility graphs in order to prevent collision. This requires high computational time. Furthermore, in real scenarios, a narrow way-path is common situation. A* and its variant are not effective in those cases. In our work, the map of the environment has been determined: $\mathcal{L}^N = \{L_1, L_2, ..., L_N\}$ where $\mathcal{L}^N$ is a set of connected locations. Given two locations on the map L_{k_s}, L_{k_d}, the path going from L_{k_s} to L_{k_d} is simply determined by a subset $L^{sd} = \{L_{k_s}, ..., L_{k_d}\}$, with $L^{sd} \in \mathcal{L}^N$.

3.5 System-user interaction

System-user interaction is one of the main issues for a navigation system. The system has to be designed so that the interaction between human and robot is as natural as possible. To exploit user behavior and their requirements, we have made a survey to the participated VI pupils.

Statistical results show that 70.59 % participated pupils want to cling to the robot to move as walking with a friend (29.41 % want to walk separately). 94.12 % want to use mobile phone to communicate with robot (5.88 % don't want to use mobile phone). 41.18 % want to use vibration for navigation indication and only give indications at important locations (35.29 % want synthetic speech, 23.53 % want both).

Through these analyses, we use a smart phone with a WIFI connection to send control commands to the robot and receive feedback from the robot. The communication for a working section is shown in Fig. 10, and is described below:

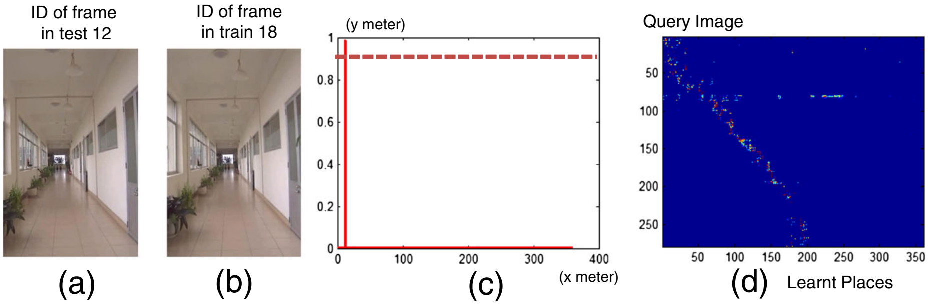


Fig. 9 **a** Given a current observation, **b** the best matching place. **c** The probability $p(L_i|Z^k)$ calculated with each location k among $N = 350$ learnt places. **d** Confusion matrix of the matching places with a sequence of collected images $(290 frames)$

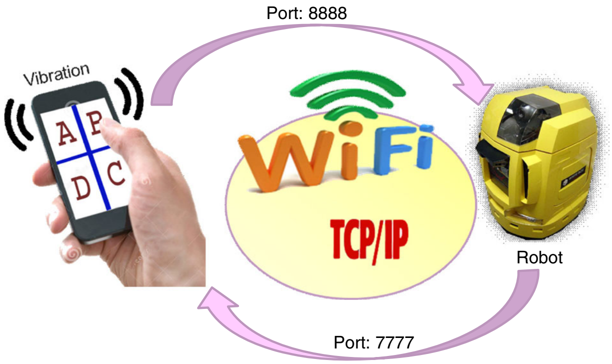


Fig. 10 Communication model using TCP/IP between human and robot

- **Human**: The user turns on the application on the mobile phone. An interface with M sub-windows appears. A screenshot of the application is shown in Fig. 11. In this screen, each sub-window corresponds to a destination which is a pre-defined location in the indoor environment (e.g., a department, a sharing-room, cafeteria, rest-room). If user wants to go to a specific location A, he touches on the corresponding sub-window. The application will send a command like ”going to A” to the robot.
- **Robot:** The robot scans every 100 ms for requests from the user. If it detects a command ”going to A”, it starts the localization module and determines the path going from its current location to A. Then the robot moves following the intermediate locations. Robot’s location is corrected by a Kalman filter. The directional movements of the robot are sent to the application on the mobile phone.
- **Human:** The user follows the robot. The feedbacks of the robot are encoded to vibration signals to alarm the user. There are four vibrations signals meaning ”turn left direction”, ”turn right”, ”go straight”, and ”stop”.

We have made many tests on vibration frequency to choose the best ones that the VI people could recognize the four indicators. The tests have been done with thirteen VI pupils.

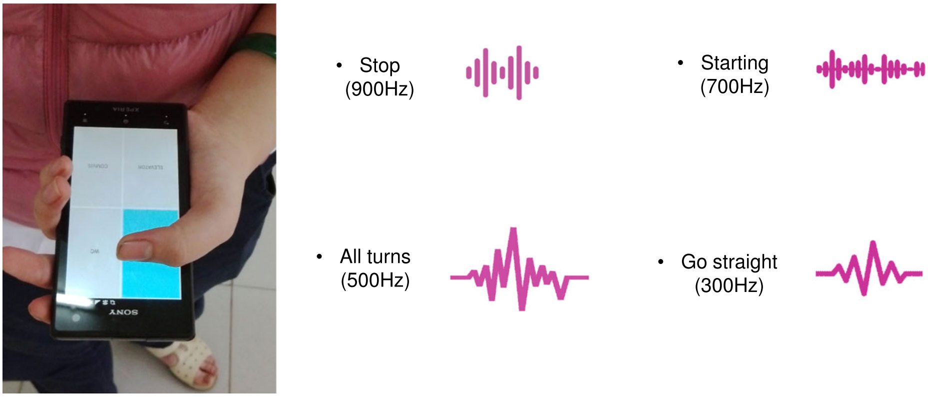


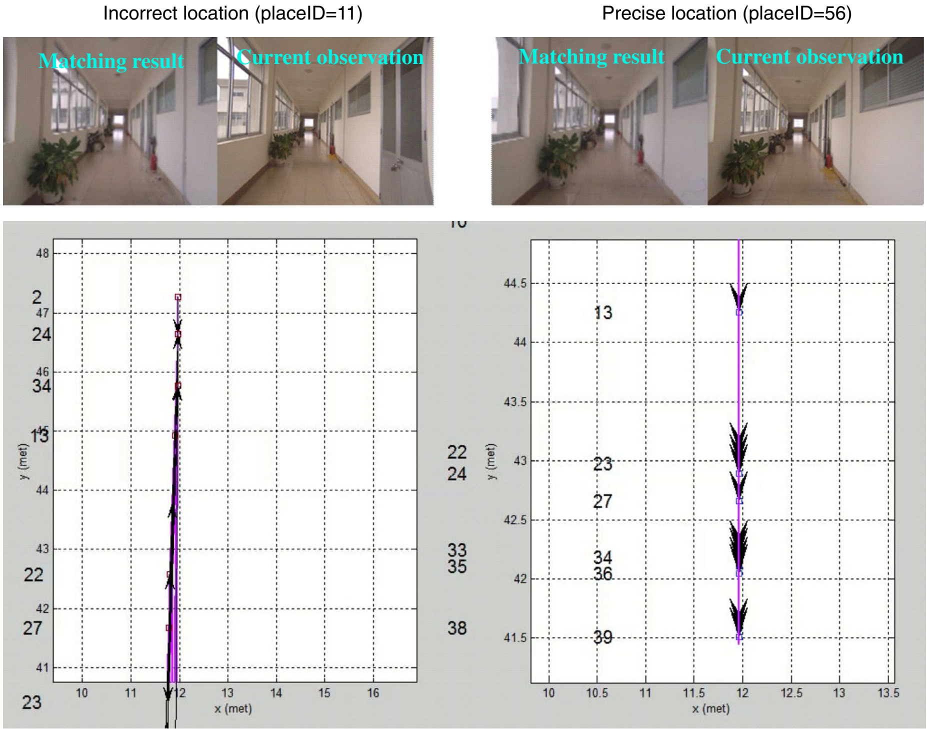
Fig. 11 Vibration frequencies corresponding to four navigation indications

After three trials for each person, we obtain the highest recognition result at frequencies of 300Hz, 500Hz, 700Hz, and 900Hz, respectively.

3.5.1 Navigation

Given a path $L^{sd} = \{L_{k_s}, ..., L_{k_d}\}$, robot will be controlled to move linearly from L_{k_i} to L_{k_j} at a constant speed. However, due to the hardware structures, the robot can not come exactly to L_{k_j} but to an unobservable L'_{k_j}. At this new position, the vision based localization is activated and predicts robot position at $L^*_{k_j}$. The robot will be controlled to move from $L^*_{k_j}$ to L_{k_j} until the distance between $L^*_{k_j}$ and L_{k_j} is smaller than a threshold.

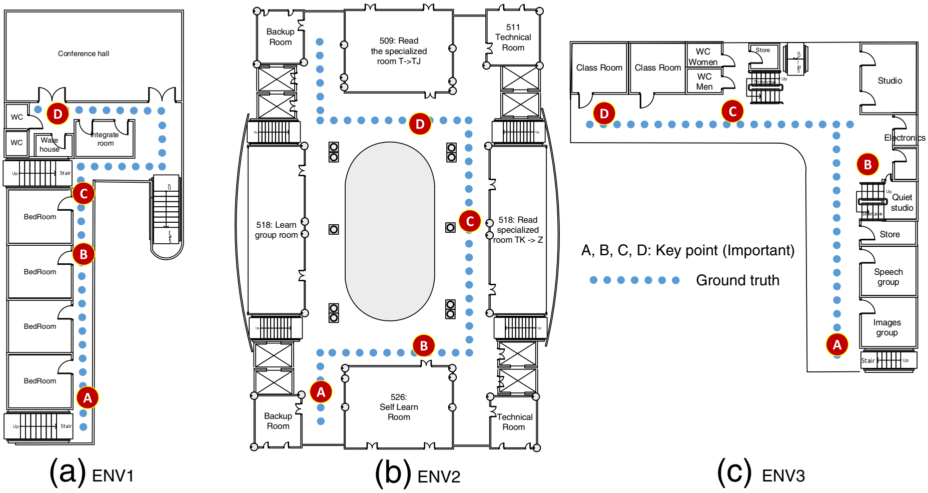
Due to the error of the localization, the predicted position $L^*_{k_j}$ would be very far forward or backward from the true location. As a consequence, the robot could move backward/forward many times. This makes the system inefficient. To overcome this problem, we use a Kalman filter to correct the position of the robot from the observation. Kalman filter helps to estimate the position of the robot near to the real unobservable given an error observation $L^*_{k_j}$. To evaluate the role of Kalman filter, we control the robot to move along the corridor. As shown in Fig. 12, with a pre-selected velocity of the robot, Kalman filter helps to make the robot's movement more smooth and linear. In practice, the localization error has on average reduced from 0.86m to 0.41m.



(a) Vehicle's position without Kalman filter

(b) Vehicle's position with Kalman filter

Fig. 12 Correted robot locations thanks to Kalman filter



(a) ENV1 (b) ENV2 (c) ENV3

Fig. 13 Three environments to be considered for evaluation

4 Experimental results

The proposed system is built on a 914 PC-BOT[1] with an M3-controller. 914 PC-BOT's configurations include Intel Core 2 Duo 2 GHz, 1 GB RAM, 80 GB HDD, on a platform of Windows 7 OS. The algorithms are wrapped in a package using Visual Studio C++. The propose system deploys the FAB-MAP algorithms based on openFABMAP package[2]. A wide-angle IP camera (Axis 270MW[3]) is attached on the PC-BOT. We use lens distortion function of OpenCV library to find intrinsic parameters. The captured images are undistorted before any processing. The image is collected at 1 *fps* and the robot is setup at a constant speed of 300 mms.

Our experiments are conducted with three real scenarios and evaluated by participants of 13 VI pupils. Aims of the evaluations are to answer to following questions: i) how man markers help to improve the precision reconstruction of vehicle travel; ii) how discriminative frame selection helps to improve performance of localization; iv) how the whole system satisfies the main predefined requirements.

4.1 Experimental environments and data collection

4.1.1 Experimental environments

Three scenarios of the indoor environments are conducted. Figure 13 shows maps of such environments. Blue dotted lines show the trajectories that are data-collected. We could see that each trajectory composes straight and turning sections. Each environment has it own characteristics, as the below explanations describe. The Table. 1 summarizes relevant information of experimental environments.

[1] http://www.whiteboxrobotics.com/Product/spec.html

[2] https://code.google.com/p/openfabmap/

[3] http://www.axis.com/global/sv/products/axis-207

Table 1 Summarization of three environments

$N°$	Environment	Length(m)	Width (m)	Turning (90°)
01	ENV1	76.8	1	3
02	ENV2	154	3.2	4
03	ENV3	60	2.5	1

- $ENV1$: This is the 2^{nd} floor of dormitory of a School for blind pupils[4]. This environment is familiar for blind pupils because they have been living here for several years. The environment is composed of four straight sections. The two first sections have about 36m in length. One side of the hallway is comprised of the walls and doors of the dormitory's rooms. Another side is retaining walls and columns, therefore these sections are highly illuminated by sunlight. The two remaining sections are covered by walls. The area is in low lighting condition, mostly the observation at the fourth section is very dark. The floor plane is composed of quite smooth tiles. As this is living environment for blind pupils, no objects are put in the hallway. On the first section, the scene structures are very repetitive from one room to another.
- $ENV2$: This is a floor of a Library[5]. This environment is completely new for blind people. It is composed of of five sections. The lighting condition is low and quite similar because all sections are covered by walls and not exposed to sunlight. The repeatability of environment structures is lower than in ENV1.
- $ENV3$: This is a floor of a Office building[6]. It is a new environment for blind people. It is composed of only of two main sections. It has high lighting condition. One side of the hallway is exposed to sunlight. The repeatability of environment structures is very strong.

4.1.2 Data collection

To be able to evaluate the performance of each module of the system, we move the self designed vehicle following to predefined trajectories to collect data for training and testing our method in each environment 3 times: (L_1, L_2) in the morning and L_3 in the afternoon. The resolution of frame is 640x480. Tables 2, 3 and 4 show details of the datasets.

The number of road frames and scene frames are different because the frame rate of the road and scene camera is different. To match one road frame to a scene frame, we have to scale to the frame rate of video recording.

4.2 Evaluate the precision of travel reconstruction

4.2.1 Evaluation measurement

Suppose that $\{(x_i, y_i), i \in [1, P]\}$ are P ground truth points manually marked on the trajectory of the vehicle. (x_i^*, y_i^*) are the corresponding reconstructed points using visual

[4] Nguyen Dinh Chieu Blind School, Hanoi

[5] 5^{nd} floor of Ta Quang Buu Library, HUST

[6] 10^{th} International Research Institute MICA, HUST

Table 2 Data collected at ENV1

Trial	#Scene frames	#Road Frames	Duration (s)	Used for
L1	12090	20150	06:43	Training
L2	11700	19500	06:30	Training
L3	10380	17300	05:46	Testing

odometry technique. We measure the precision of reconstructed travel by Root Mean Squared Error (RMSE). As (x_i, y_i) and (x_i^*, y_i^*) are measured in real coordinate system, the RMSE has meter unit.

$$RMSE = \sqrt{\frac{1}{P}\sum_{i=1}^{P}(x_i - x_i^*)^2 + (y_i - y_i^*)^2} \tag{5}$$

4.2.2 Results of travel reconstruction

We evaluate the precision of travel reconstruction on three environments. We use $L1$ sequence to build the travel. Figure 14 shows the reconstructed travels with and without man markers overlapped on the ground-truth trajectory.

As we can see in the Fig. 14 we sampled uniformly $P = 32$ points on the ground-truth trajectory to compute RMSE. Without man markers, the result of reconstruction is poor due to cumulative error. Using man markers, the RMSE is improved approximatively linearly according to the number of features detected on each frame. We notice also that RMSE at ENV3 environment is better than RMSE at ENV1 and ENV2. The reason is that ENV3 is highly illuminated so the number of detected key points is good enough for travel reconstruction.

4.3 Evaluate the performance of localization

4.3.1 Localization evaluation measurements

To evaluate the performance of localization, we follow the same measures used in [5] that are Precision and Recall.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Table 3 Data collected at ENV2

Trial	#Scene frames	#Road frames	Duration(s)	Used for
L1	12354	10650	07:06	Training
L2	11803	10175	06:47	Training
L3	10295	8875	05:55	Testing

Table 4 Data collected at ENV3

Trial	#Scene frames	#Road frames	Duration(s)	Used for
L1	7801	13450	04:29	Training
L2	10005	17250	05:45	Training
L3	5945	10250	03:25	Testing

Where TP, FP, FN stand for True Positive, False Positive, False Negative. They are defined as follows. Suppose that at time k, the observation of the robot up to time k is Z^k. The localization procedure computes $L_j = argmax P(L_i|Z^k)$.

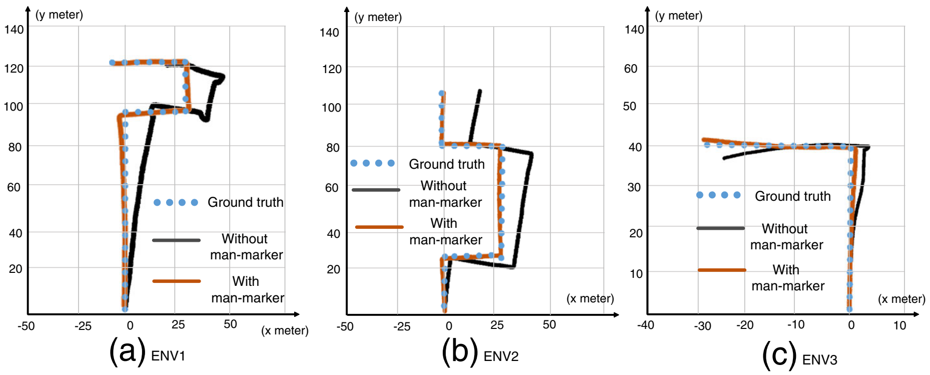
- If $P(L_j|Z^k) < \theta_1$ this is a false negative.
- If $P(L_j|Z^k) \geqslant \theta_1$ & $|L_j - L_{groundtruth}| \leqslant \theta_2$ this is a true positive.
- If $P(L_j|Z^k) \geqslant \theta_1$ & $|L_j - L_{groundtruth}| > \theta_2$ this is a false positive.

In our experiments, we set θ_2 to 0.4m. The value of θ_1 is varied from 0.1 to 0.9 with spacing 0.1 to build to Precision-Recall curve.

4.3.2 Localization results

We use L_2 sequence containing about 10000 frames to build a Bag of Word model and Chow Liu tree. Two procedures are tested and compared: (1) using all frames of the sequences and (2) using selective frames with a GIST descriptor. In both cases, the size of vocabulary is set to 1000.

We test our localization module with the sequence L_3. Precision-Recall curves are shown in Fig. 15 at three environments. The curves were generated by varying the probability at which a loop closure was accepted, specifically varying the value of threshold θ_1 from 0.1, .., 0.9. Ground truth was labeled by hand.



VO based (Visual Odometry)	**RMSE(m)**		
Travel reconstruction (man-marker)	*ENV1*	*ENV2*	*ENV3*
Without	1.23	1.78	0.68
With	0.11	0.14	0.12

Fig. 14 Result of travel reconstruction with / without man-markers at environments using the same algorithm parameters

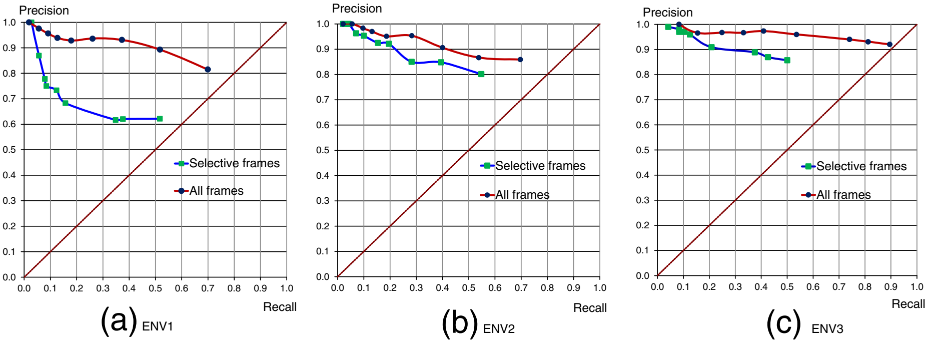


Fig. 15 Recall/Precision curves at environments

It could be seen that using selective frames to build a Bag of Word model gives better results than using all frames. In [5] the authors were interested in the recall rate at 100 % precision because they tuned the system to outdoor environment with high speed car moving. In our context with the slow robot movement in the indoors, we accept a small portion of false alarms and correct locations using the Kalman filter. Therefore we are interested in a recall rate at a high level of precision such as when $\theta_1 = 0.4$, at ENV3 environment, at 95.98 % precision, the system achieves a 53.59 % recall rate.

The Figs. 16, 17 and 18 illustrate the results of localization at three environments. We observe that the lighting conditions at the training and testing phase are quite different.

4.4 Evaluate the whole system with real users

All separated modules (i.e. environment representation, localization, navigation, interaction) are integrated on a mobile Robot for a real application of navigation. 13 VI pupils (6 Female and 7 Male aging from 13 to 21) of a School for Blind are invited to participate into

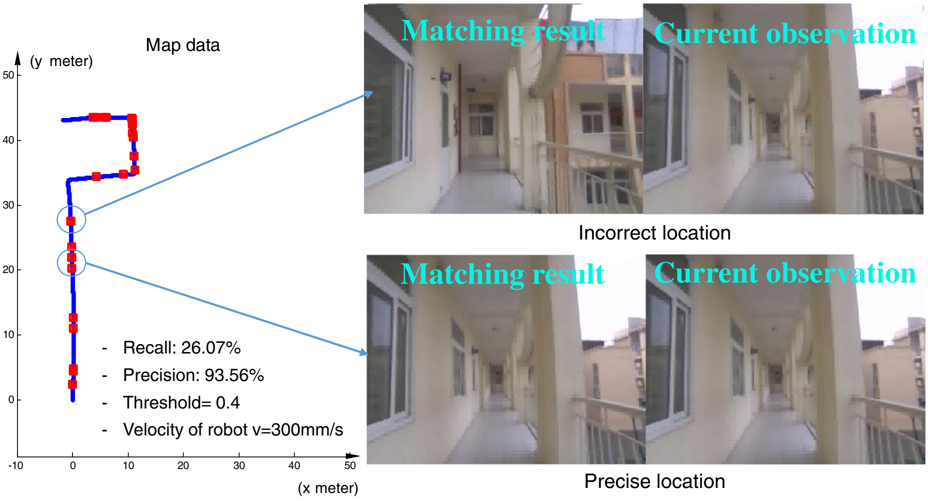


Fig. 16 Example of localization at ENV1 environment

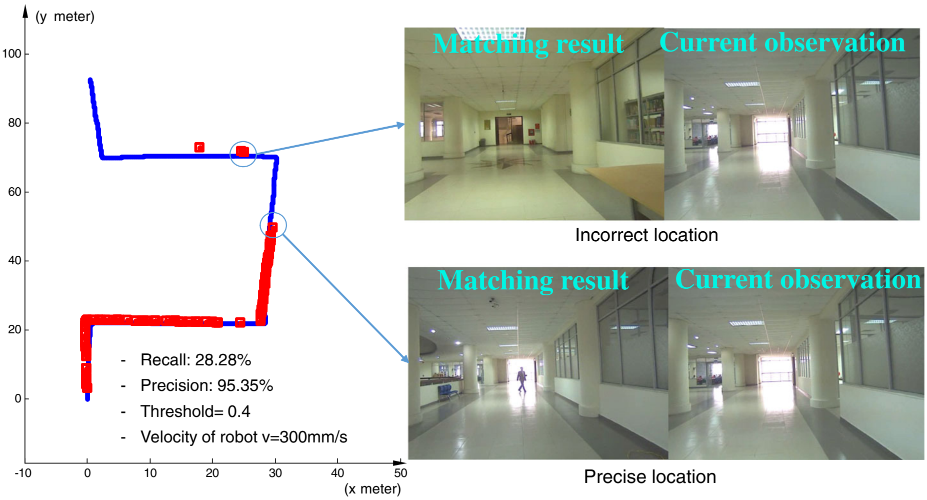


Fig. 17 Example of localization at ENV2 environment

experiments. We described the aid system and trained the users to use the mobile phone for communicating with robot. We then asked each pupil to use robot to go from a location A to B in a certain environment. For ENV2 and ENV3, a blind pupil did not know where she/he is in the environment to ensure that they would follow the guidance of the robot and not rely on their casual habits.

4.4.1 Navigation performance

The localization errors are presented in the Table 5. Averagely, the errors range from 50cm to 60cm in the Fig. 19. It equal a footstep and as mentioned in the solution requirement, accuracy of the proposed localization is favorable.

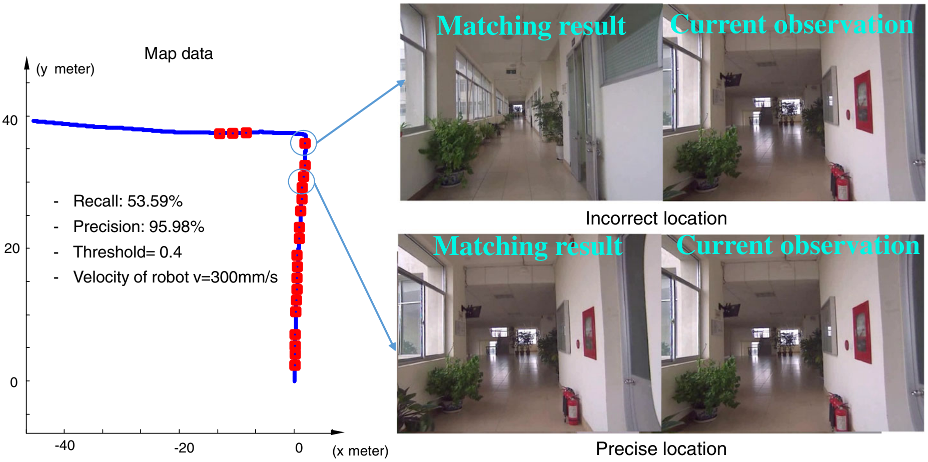


Fig. 18 Example of localization at ENV3 environment

Table 5 Localization errors (RMSE) at three environments

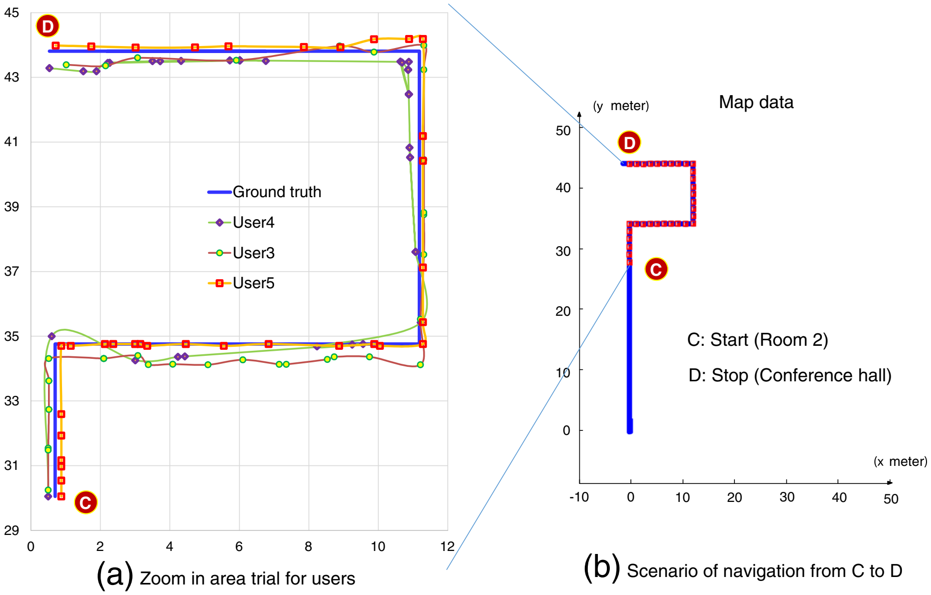
Environments	RMSE(m)
ENV1	0.58
ENV2	0.55
ENV3	0.45

4.4.2 Feedback information from participants

The system was tested in different indoor environments with simple and complex trajectories. The lighting condition slightly affects on the performance of the localization, then the navigation efficiency. After conducting the test, we asked the participants to answer a questionnaire including four main questions about their experience with the system.

- Q1: How do you rate the system usability? (scaling from 1 to 5 meaning from hard to use to easy to use)
- Q2: How do you understand the indication of the system (scaling from 1 to 5 meaning from incomprehensible to clear)?
- Q3: How fast is the movement of the system? (scaling from 1 to 5 meaning from slow to fast)
- Q4: How useful is the system? (scaling from 1 to 5 meaning from useless to very useful)

Table 6 shows system score according to each question. It appears that almost all participant VI pupils found the system very useful, mostly in unfamiliar environments. The system is highly usable. The blind pupils could understand the indication of the system through vibration with little confusion. However, as the user holds on and follows the robot, some



(a) Zoom in area trial for users

(b) Scenario of navigation from C to D

Fig. 19 Illustration of robot movement guiding blind user to go from C to D location at ENV1 School with three among 13 VI pupils

Table 6 Results of interviews with participated pupils after using the system

Question	Score
Q1	4.5
Q2	4.7
Q3	3.5
Q4	5

misunderstanding of indication is not very important. Currently, the blind pupils found that the robot movement (at 300mm/s) is quite slow and needs to be improved in the future.

4.5 Discussions

Currently, our system is a prototype that has been tested in real environments with visually impaired users that validate the feasibility of the approach. The localization error varies from 40cm to 60cm with a response time of 100ms. As mentioned in the previous section, it satisfies pre-defined requirements. In our evaluations, results of matching a current observation to the defined places in the corresponding environment is in a range from 26 % to 53 % for the recall rate, and averagely 94 % for precision. It is noticed that robot velocity is setup at 300 mm/s and processing time is 1 fps. For example, to go a distance of 1m, requires 3.3 seconds. There are 3 captured frames. When the proposed system archives the minimal recall rate (26 %), it still has at least one among three captured frames are matched (because of high precision); two remaining frames are not able to be matched. These results infer that input to path finding modules is convenient within 1 m. Consequently, performance of the proposed system is acceptable. However, the current system has still some main limitations:

- The idea of enriching the number of detected feature points for travel reconstruction is a simple solution. However, this method has still a drawback of the accumulated error when the vehicle moves along a complex trajectory. In addition, actually, the explored road is one-way only. The appearance model of location is therefore one-way. Consequently, the robot can not support the blind user going in the return direction.
- Currently, as map is quite simple, the way-finding is replied mainly on connected locations on the map. In a more complex environment, the map could compose of many nodes and links. In that case, it needs to have more sophisticated path finding algorithm.
- The current system reacts only on the knowledge that he has learnt before (the pre-built map). It has not ability to incrementally update the map of environment as well as learn a lesson to avoid the wasted trajectory thanks to Kalman filter.
- The system is designed to guide only one user at a time in indoor environment. So it is more suitable for public environment (public administration) and the first time the user comes.

5 Conclusions and future work

In this paper, we proposed a vision-based way-finding system supporting blind people in indoor environments. We adapt relevant techniques to handle issues to make a feasible system. The core solution of the proposed system is designed a vision-based place recognition system on a mobile robot. In which, we represented an indoor environment by representative scenes along robot's routes. Through extensive evaluations, our enhancements show their effectiveness. The man markers were very efficient for travel reconstruction using visual odometry with uncertainty model. Selecting distinctive frames before building bag of word model helped to improve significantly recall and precision of localization. A Kalman filter showed its important role to increase localization accuracy. Finally, we have integrated all separated modules into a mobile robot that communicates with users through mobile phone. Experiments with different environments and participated by blind people shows that the system is useful and usable in real situations. The main advantage of the system is to help people to go to a desired position in a completely unfamiliar environment.

In the future, we shall try to overcome some limitations of the system. Firstly, we plan to improve visual odometry algorithm for more precise travel reconstruction. We will make two way road map and test the system with more complex road. Second, we would like to update the map incrementally in order to enrich map at runtime for further processing. We could also memorize the real movement of the robot to go from one specific point to another to optimize the computation. Other extension of this work is to develop new functionality such as informing the user about environments. Currently, we are working on obstacle recognition. However, we would like to create encoded pictures of surrounding environment to the blind to give much more information about the world.

Acknowledgments This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number FWO.102.2013.08.

References

1. Alcantarilla FP (2011) Vision based localization: from humanoid robots to visually impaired people. Ph.D Thesis
2. Bailey T, Durrant-Whyte H (2006) Simultaneous localization and mapping (slam): part ii. IEEE Robot Autom Mag 13(3):108–117
3. Bigham J, Jayant C, Miller A (2010) White: Vizwiz::locateit - enabling blind people to locate objects in their environment. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 65–72
4. Chow C, Liu C (1968) Approximating discrete probability distributions with dependence trees. IEEE Trans Inf Theory 14(3):462–467
5. Cummins M, Newman P (2008) Fab-map: Probabilistic localization and mapping in the space of apperance. Int J Robot Res 27(6):647–665
6. Dakopoulos D, Bourbakis NG (2010) Wearable obstacle avoidance electronic travel aids for blind: a survey. IEEE Trans Syst, Man, Cybernet Part C: Appl Rev 40(1):25–35
7. Endres H, Feiten W, Lawitzky G (1998) Field test of a navigation system: Autonomous cleaning in supermarkets. In: the Proceeding of International Conference on Robotics and Automation. IEEE
8. Fallah N, Apostolopoulos I, Bekris K, Folmer E (2013) Indoor human navigation systems - a survey. Interact Comput 25(1):21–33
9. Fraundorfer F, Scaramuzza D (2012) Visual odometry : Part ii: Matching, robustness, optimization, and applications. IEEE Robot Autom Mag 19(2):78–90
10. Hamme D, Veelaert P (2011) Robust visual odometry using uncertainty models. Proc Adv Concepts Intell Vis Syst 6915:1–12
11. Hart PE, Nilsson NJ, Raphael B (1968) A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cybern 4(2):100–107
12. Helal A, Moore SE, Ramachandran B (2001) Drishti: An integrated navigation system for visually impaired and disabled. In: Proceedings. Fifth International Symposium on Wearable Computers, 2001. IEEE, pp 149–156
13. King S, Weiman C (1990) Helpmate autonomous mobile robot navigation system. In: the Proceeding of the SPIE Conference on Mobile Robots, pp 190–198
14. Korf RE (1985) Iterative-deepening-a: an optimal admissible tree search. In: Proceedings of the 9th international joint conference on Artificial intelligence-Volume 2. Morgan Kaufmann Publishers Inc, pp 1034–1036

15. Kulyukin V, Gharpure C (2006) Nicholson: Robot assisted way-finding for the visually impaired in structured indoor environments. Auton Robot 21:29–41
16. Kulyukin V, Gharpure C, Nicholson J, Pavithran S (2004) Rfid in robot-assisted indoor navigation for the visually impaired. In: Proceedings. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004), vol 2. IEEE, pp 1979–1984
17. Lacey G, Dawson-Howe K (1998) The application of robotics to a mobility aid for the elderly blind. Robot Auton Syst 23:245–252
18. LaMarca A, Brunette W, Koizumi D (2002) Making sensor networks practical with robots. In: the Proceeding of International Conference on Pervasive Computing. IEEE
19. Lehel P, Hemayed E, Farag A (1999) Robot assisted way-finding for the visually impaired in structured indoor environments. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 2
20. Liu JJ, Phillips C, Daniilidis K (2010) Video-based localization without 3d mapping for the visually impaired. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp 23–30
21. Loomis JM, Golledge RD (2001) Klatzky: Gps-based navigation systems for the visually impaired. Fundamentals of wearable computers and augmented reality, pp 429–446
22. Marion AH, Micheal AJ (2008) Assistive Technology for Visually Impaired and Blind People. Springer
23. Murali VN, Coughlan JM (2013) Smartphone-based crosswalk detection and localization for visually impaired pedestrians. In: 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). IEEE, pp 1–7
24. Newman P, Ho K (2005) Slam-loop closing with visually salient features. In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, 2005. ICRA 2005. IEEE, pp 635–642
25. Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 2. IEEE, pp 2161–2168
26. Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
27. Pradeep V, Medioni G, Weiland J (2010) Robot vision for the visually impaired. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, pp 15–22
28. Schindler G, Brown M, Szeliski R (2007) City-scale location recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07, pp 1–7
29. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: Proceedings. Ninth IEEE International Conference on Computer Vision, 2003, vol 2, pp 1470–1477
30. Sunderhauf N, Protzel P (2011) Brief-gist-closing the loop by simple means. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp 1234–1241
31. Winlock T, Christiansen E, Belongie S (2010) Toward real-time grocery detection for the visually impaired. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp 49–56

Quoc-Hung Nguyen received the Master degrees in computer science from Thai Nguyen University of information and communication technology in 2010. He is currently a PhD student of Hanoi University of Science and Technology. His research interests includes methods for acquiring, processing, analyzing in order to understand images and human-robot interaction.

Hai Vu received B.E. degree in Electronic and Telecommunications in 1999 and M.E. in Information Processing and Communication in 2002, both from Hanoi University of Technology. He received Ph.D. in Computer Science from Graduate School of Information Science and Technology, Osaka University, 2009. He join MICA International Research Institute from 2012. He am interested in computer vision, medical imaging techniques, mainly video capsule endoscopy analysis for diagnostic assistance.

Thanh-Hai Tran graduated in Information Technology from Hanoi University of Science and Technology. She is currently lecturer/researcher at Computer Vision group, International centre MICA, Hanoi University of Science and Technology. Her main research interests are visual object recognition, video understanding, human-robot interaction and text detection for applications in Computer Vision.

Quang-Hoan Nguyen received the University degree the former Soviet Union (1973) Former head of the IT department and TT Internet and Library Academy of Telecommunications Technology (1998-2010). His research interests includes methods for Artificial intelligence (especially artificial neural networks), Computer Hardware, Modern control and system stability.