

LNCS 8927

Lourdes Agapito  
Michael M. Bronstein  
Carsten Rother (Eds.)

# Computer Vision – ECCV 2014 Workshops

Zurich, Switzerland, September 6–7 and 12, 2014  
Proceedings, Part III

3  
Part III

 Springer

**EXTRA**  
MATERIALS  
[springerlink.com](http://springerlink.com)

**VIDEOS**  
[springerimages.com](http://springerimages.com)

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

David Hutchison

*Lancaster University, Lancaster, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Friedemann Mattern

*ETH Zürich, Zürich, Switzerland*

John C. Mitchell

*Stanford University, Stanford, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

C. Pandu Rangan

*Indian Institute of Technology, Madras, India*

Bernhard Steffen

*TU Dortmund University, Dortmund, Germany*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max Planck Institute for Informatics, Saarbrücken, Germany*

Lourdes Agapito · Michael M. Bronstein  
Carsten Rother (Eds.)

# Computer Vision – ECCV 2014 Workshops

Zurich, Switzerland, September 6–7 and 12, 2014  
Proceedings, Part III

*Editors*

Lourdes Agapito  
University College London  
London  
UK

Carsten Rother  
Technische Universität Dresden  
Dresden  
Germany

Michael M. Bronstein  
University of Lugano  
Lugano  
Switzerland

Videos to this book can be accessed at

<http://www.springerimages.com/videos/978-3-319-16198-3>

ISSN 0302-9743  
Lecture Notes in Computer Science  
ISBN 978-3-319-16198-3  
DOI 10.1007/978-3-319-16199-0

ISSN 1611-3349 (electronic)

ISBN 978-3-319-16199-0 (eBook)

Library of Congress Control Number: 2015933663

LNCS Sublibrary: SL6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media  
(www.springer.com)



A Fast and Flexible Computer Vision System for Implanted Visual Prostheses . . . . .	686
<i>Wai Ho Li</i>	
An Intelligent Wheelchair to Enable Safe Mobility of the Disabled People with Motor and Cognitive Impairments . . . . .	702
<i>Yeounggwang Ji, Myeongjin Lee, and Eun Yi Kim</i>	
A Visual SLAM System on Mobile Robot Supporting Localization Services to Visually Impaired People . . . . .	716
<i>Quoc-Hung Nguyen, Hai Vu, Thanh-Hai Tran, David Van Hamme, Peter Veelaert, Wilfried Philips, and Quang-Hoan Nguyen</i>	
Scene-Dependent Intention Recognition for Task Communication with Reduced Human-Robot Interaction . . . . .	730
<i>Kester Duncan, Sudeep Sarkar, Redwan Alqasemi, and Rajiv Dubey</i>	
Egocentric Object Recognition Leveraging the 3D Shape of the Grasping Hand . . . . .	746
<i>Yizhou Lin, Gang Hua, and Philippos Mordohai</i>	
Face Recognition by 3D Registration for the Visually Impaired Using a RGB-D Sensor . . . . .	763
<i>Wei Li, Xudong Li, Martin Goldberg, and Zhigang Zhu</i>	
Learning Pain from Emotion: Transferred HoT Data Representation for Pain Intensity Estimation . . . . .	778
<i>Corneliu Florea, Laura Florea, and Constantin Vertan</i>	
Neural Network Fusion of Color, Depth and Location for Object Instance Recognition on a Mobile Robot . . . . .	791
<i>Louis-Charles Caron, David Filliat, and Alexander Gepperth</i>	
An Experimental Analysis of Saliency Detection with Respect to Three Saliency Levels . . . . .	806
<i>Antonino Furnari, Giovanni Maria Farinella, and Sebastiano Battiato</i>	
Recognizing Daily Activities in Realistic Environments Through Depth-Based User Tracking and Hidden Conditional Random Fields for MCI/AD Support. . . . .	822
<i>Dimitris Giakoumis, Georgios Stavropoulos, Dimitrios Kikidis, Manolis Vasileiadis, Konstantinos Votis, and Dimitrios Tzovaras</i>	
3D Layout Propagation to Improve Object Recognition in Egocentric Videos. . . . .	839
<i>Alejandro Rituerto, Ana C. Murillo, and José J. Guerrero</i>	
<b>Author Index . . . . .</b>	<b>853</b>

# A Visual SLAM System on Mobile Robot Supporting Localization Services to Visually Impaired People

Quoc-Hung Nguyen<sup>1</sup>(✉), Hai Vu<sup>1</sup>, Thanh-Hai Tran<sup>1</sup>, David Van Hamme<sup>2</sup>,  
Peter Veelaert<sup>2</sup>, Wilfried Philips<sup>2</sup>, and Quang-Hoan Nguyen<sup>3</sup>

<sup>1</sup> International Research Institute MICA, Hanoi University of Science and  
Technology, Hanoi, Vietnam  
{quoc-hung.nguyen, hai.vu, thanh-hai.tran}@mica.edu.vn

<sup>2</sup> Ghent University/iMinds (IPI), Gent, Belgium  
dvhamme@telin.ugent.be, peter.veelaert@hogent.be  
wilfried.philips@ugent.be

<sup>3</sup> Hung Yen University of Technology and Education, Hung Yen, Vietnam  
quanghoanptit@yahoo.com.vn

**Abstract.** This paper describes a Visual SLAM system developed on a mobile robot in order to support localization services to visually impaired people. The proposed system aims to provide services in small or mid-scale environments such as inside a building or campus of school where conventional positioning data such as GPS, WIFI signals are often not available. Toward this end, we adapt and improve existing vision-based techniques in order to handle issues in the indoor environments. We firstly design an image acquisition system to collect visual data. On one hand, a robust visual odometry method is adjusted to precisely create the routes in the environment. On the other hand, we utilize the Fast-Appearance Based Mapping algorithm that is may be the most successful for matching places in large scenarios. In order to better estimate robot's location, we utilize a Kalman Filter that combines the matching results of current observation and the estimation of robot states based on its kinematic model. The experimental results confirmed that the proposed system is feasible to navigate the visually impaired people in the indoor environments.

**Keywords:** Visual odometry · Place recognition · FAB-MAP algorithms · Kalman filter

## 1 Introduction

Autonomous localization and navigation are extreme desirable services for visually impaired people. Most commercial solutions are based on the Global Positioning System (GPS), WIFI, LIDAR, or fusion of them. iNavBelt uses ultrasonic sensors to procedure a 120-degree wide view ahead of the user [19]. GuideCane has an ultrasonic sensor head mounted on a long handle [3] The EyeRing developed by MIT's Media Lab, is a finger-won device that translates images into aural signals. Although such kind of devices are useful to blind/visually impaired people in some environments, major drawbacks are that they only give limited information, and require well-focused user control. Recent techniques in the computer vision and robotics community offer

substantial advantages to overcome those limitations. This paper proposes the use of visual sensors mounted on an intelligent system like a mobile robot to assist the visually impaired/blind people in indoor environments. The proposed system aims to solve two problems: 1. Understanding the current environments. 2. Robot self-location. Regarding the problem #1, the question is "What does the world look like?". Answering this question involves building a map of the environment. In contrast to this, self-location service relates to estimating a pose to a relative position on the created map, to answer the second question "Where am I?". A visual SLAM method relying on the visual appearance of distinct scenes is responsible for finding solutions to both problems: it builds and maintains a map of the robot's trajectory and the landmark positions. Recent approaches like FAB-MAP are aimed at reaching a high recall rate at 100% precision. In this work, we employ a robust FAB-MAP [4] that is reliable to recognize known places through autonomous operation in an intelligent system like a mobile robot. FAB-Map 2.0 has been applied to a 1000 km dataset and achieved a recall of 3.1% at 100% precision (14.3% at 90% precision respectively).

Although FAB-MAP approaches are reliable for recognizing places in large-scale environments, in indoor environments, repetitive structure and sensory ambiguity constitute severe challenges for any place recognition system. Our real experiments in indoor environments show that by setting a threshold on the probability for matching an observation, it is very difficult to obtain high recall rate (~ 14% at 100% precisions). Therefore, we focus on two improvements in this paper. We first clearly describe the visual dictionary of the discriminant scenes. In indoor environments, because many scenes has repetitive structure, the visual dictionary needs to include only representative scenes. Secondly, we deploy the Kalman filter to update vehicle position. This update incorporates prior knowledge of the vehicle (e.g. velocity of the mobile robot, pacing of the people).

The proposed system is implemented in two phases. The first phase is an off-line process including two main functions: building a map and learning (indexing) places in the environment. We simultaneously collect visual data for the off-line process by a self-designed image acquisition system. For building a map of the environment, we utilize a robust visual odometry proposed in [8]. This is an interesting method because it is able to build a trajectory using only one consumer-grade camera. Furthermore, in order to improve quality of the constructed map, we adapt the algorithms in [8] with contexts of the indoor environments. In order to learn places in the environment, we utilize so-called loop closure detection methods [4], [14]. The main idea for learning the visited places is that loop constraints can be found by evaluating visual similarity between the current observation and past images captured in one (or several) trials. The second phase is an online process. The current observation is matched to a place in the database. This matching procedure is similar to place recognition. A probabilistic model of FAB-MAP algorithms [4] is utilized to find the maximal likelihood. It is observed that the proposed system is not able to update new positions against the created map. We simply past new places using a simple motion model that is based on positions of the closest neighbor places.

We evaluate results of the proposed method through travels of a mobile robot which moves along corridors of a large building. The experimental results show succesful matching of places on the map with 74% precision and 88% recall. This demonstrates the possibility of guiding blind people with the mobile robot. The

remainder of the paper is organized as follows: in Section 2, we briefly survey the related works. In Section 3, we present our vision-based system for automatic map building and localization tasks. We report the experimental results in Section 4. Finally, we conclude and give some ideas for future work.

## 2 Related Works

Localization and navigation assistance tools for visually impaired people have received much attention in the autonomous robotics community [5]. Most of the works focus on finding efficient localization solutions based on positioning data from different sensors such as GPS, laser, Radio Frequency Identification (RFID), vision or the fusion of several of them. Loomis et al. [12] surveyed efficiency of GPS-based navigation systems 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 multi-path effect or line-of-sight restrictions due to the presence of buildings or even foliage. Kulyukin et al. [10] 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. [9] proposed a wireless pedestrian navigation system. They integrated several signals such as voiced, 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 offer not only safe navigation, but also provide a very rich and valuable description of the environment. For example, [2] develops an application named *LocateIt*, which helps blind people locate objects in indoor environments. In [22], *ShelfScanner* is a real-time grocery detector that allows online detection of items on a shopping list.

With respect to visual mapping and localization, Alcantarilla [6] 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. He then utilizes visual descriptors (such as Gauge-Speeded Up Robust Features, G-SURF) to mark local coordinates on the constructed 3-D map. Instead of building a prior 3-D map, Lui et al. [11] utilize 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 applications based on visual SLAM have also been proposed. Pradeep et al. [17] present a head-mounted stereo-vision platform for detecting obstacles in the path and warn subjects about their presence. They incorporate visual odometry and feature based metric-topological SLAM. Murali et al. [13] estimate the users location relative to the crosswalks in the current traffic intersection. They develop a vision-based smart-phone system for providing guidance to blind and visually impaired travelers at traffic intersections. The system of Murali et al. in [13] requires supplemental images from Google Map services, therefore its applicability is limited to outdoor travel.

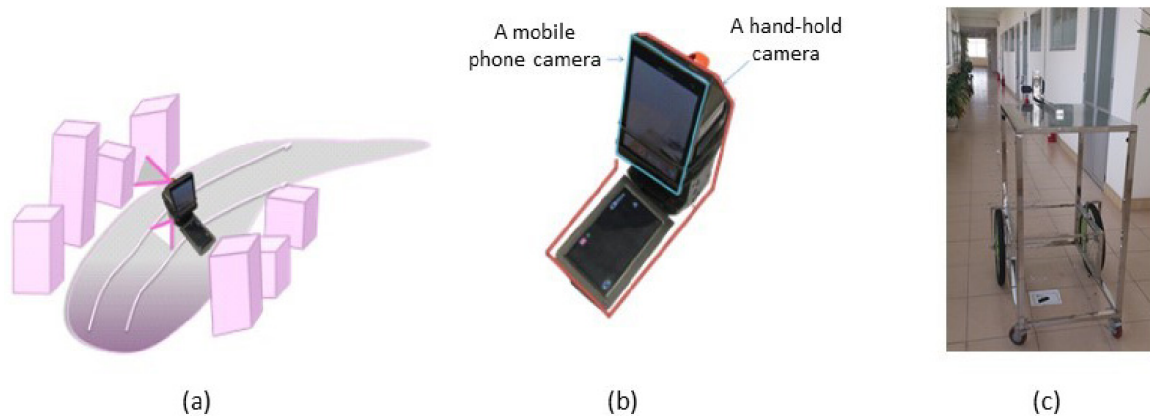
It is clear from these works that a SLAM-based approach is ideally suited to the task of guiding the visually impaired, because SLAM combines the two key elements required for a user-friendly and widely applicable system: map building and self-location. However, the complexity of the map building task varies in function of environment size. In some case, a map can be acquired from the visual sensor, but in other cases, the map is such that it must be constructed from other sensor modalities such as GPS, WIFI [4]. Furthermore, matching a current view to a position on the created map seems to be the hardest problem in many works [1], [7]. Important work towards appearance-based place recognition has been conducted in [20] which borrowed ideas from text retrieval systems and introduced the concept of the so called visual vocabulary. The idea was later extended to vocabulary trees by [15], allowing to efficiently use large vocabularies. [18] demonstrated city-scale place recognition using these tree structures.

Recently, Maddern et al. report an improvement to the robustness of FAB-Map by incorporating odometric information into the place recognition process. [21] propose BRIEF-Gist, a very simplistic appearance-based place recognition system based on the BRIEF descriptor. BRIEF-Gist is much easier to implement and its performance is comparable to FAB-MAP. In our point of view, an incremental map is able to support us in improving matching results. Therefore, different from the systems mentioned above, we attempt to create a rich map as good as possible through many trials. When new observations arrive, these new observations must be locally and globally consistent with the previously constructed map. To this end we employ the the loop closure algorithms from [4], [14]. Furthermore, we pay significant attention to the creation of the visual dictionary. We deploy the GIST features [16], a holistic representation of the natural scenes. Selection of the most representative frames helps to construct a robust visual dictionary of the environment.

### 3 The Proposed Approaches

#### 3.1 Imaging Acquisitions System

We design a compact imaging acquisition system to capture simultaneously scenes and routes in the indoor environments.



**Fig. 1.** (a) A schematic view of the visual data collection scheme. (b) The proposed imaging acquisition system in which a mobile phone camera is attached on rear of a hand-held camera. (c) The image acquisition system attached on a wheel vehicle.

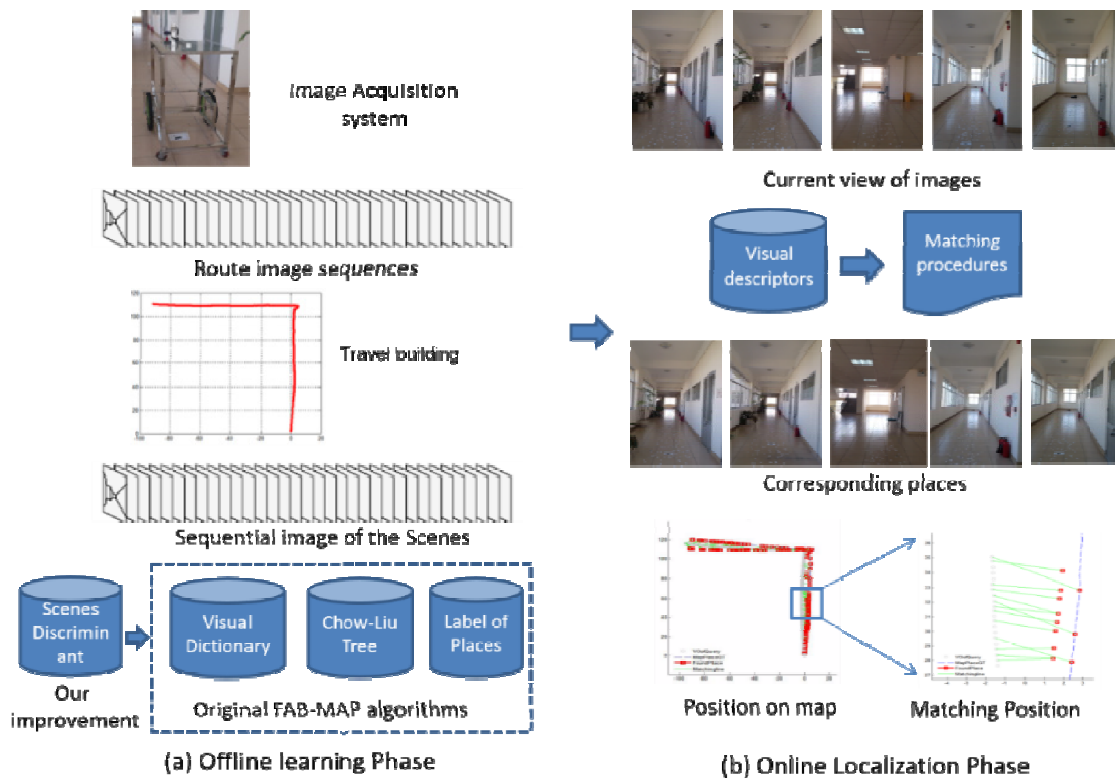


A schematic view of the data collection platform is shown in **Fig. 1(a)**. The proposed acquisition system has two cameras. One camera captures scenes around the environment. The second one aims at capturing road on the travels. The camera setting is shown in **Fig. 1 (b)**. These cameras are mounted on a vehicle, as shown in **Fig. 1 (c)**. The details of the collected images are described in the experiments. The vehicle will be only used during the offline phase to build the map of the environment and capture scene images. Using a vehicle in the offline phase has the advantage that it avoids the vibration of the camera system. As a consequence, it allows a more accurate reconstruction of the route.

### 3.2 The Proposed Framework

The proposed system is shown in **Fig. 2**. Its operation consists of two phases, as described below:

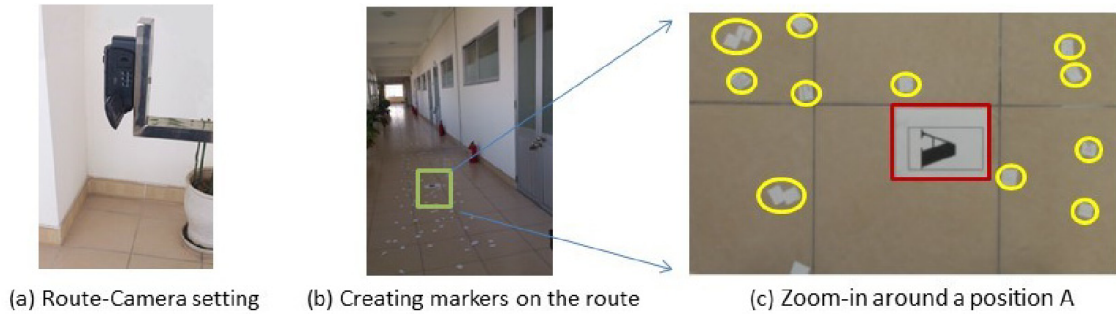
- **Off-line learning:** Using the collected visual data, this phase creates trajectories and learns the places along the travels. The techniques to construct the map and learning the places are described in Sec.3.4, Sec.3.6 respectively. Because scenes and route images are captured concurrently, the constructed map contains learnt places in corresponding positions of the travel.
- **Online localization:** The current view is described using a visual dictionary. A probabilistic function attempts to match this data to the database of labeled places obtained during the offline phase. The current observation can then be matched to a corresponding position on the constructed map.



**Fig. 2.** The framework of the proposed system

### 3.3 The Map Building Based on Visual Odometry Techniques

To build route of the travel, we utilize a visual odometry method proposed by Van Hamme et al. [8]. The method is based on the tracking of ground plane features. Particularly, it is designed to take into account the uncertainty on the vehicle motion as well as uncertainty on the extracted features.



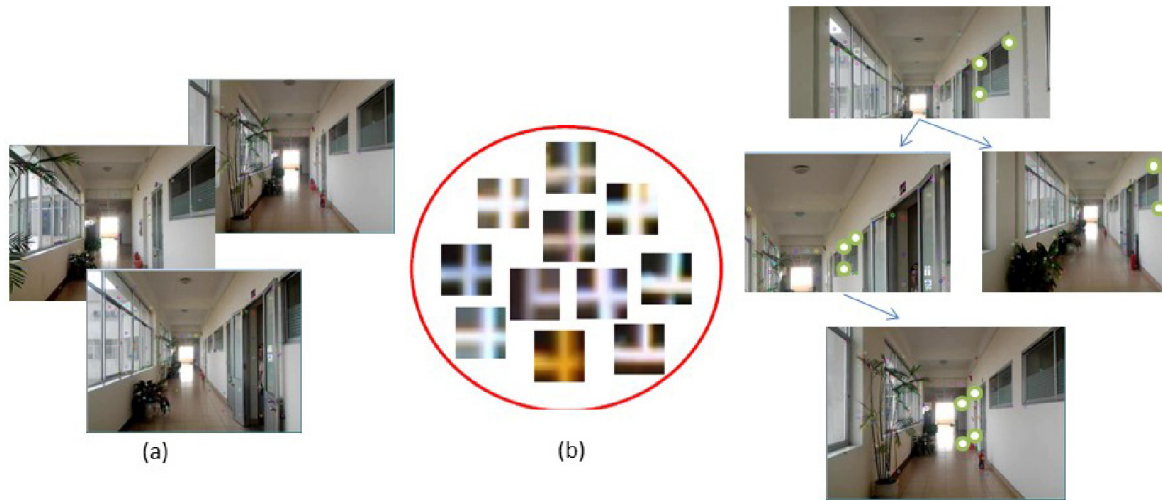
**Fig. 3.** The collection databases on Road

Our system configures the acquisition camera so that it is perpendicular to the ground plane, as shown in **Fig. 3(a)**. Well-known issues for visual odometry techniques are that they need to estimate precisely correspondences between the features of consecutive frames. Once the feature correspondences have been established, we can reconstruct the trajectory of the vehicle between the two frames. Due to the floor characteristic of the corridor environment, the number of feature points detected by the original work [8] is quite limited and leads to a very poor reconstruction of the travel. To solve this issue, we manually placed additional markers over the whole journey as shown in **Fig. 3 (b-c)**. In future work, the odometry method should be adapted to better work in case of sparse feature distribution.

### 3.4 Matching Image-to-map Procedure

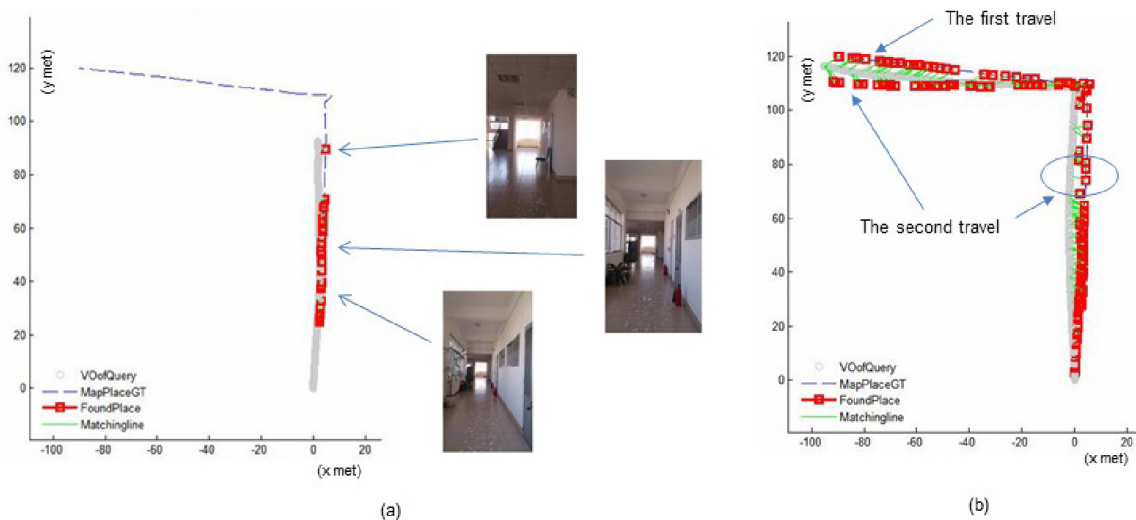
The places visited along the trajectory of interest will be stored in a condensed visual representation. This visual representation preferably needs to be easy to adapt to our specific indoor context and efficient at distinguishing scenes. To meet these goals, we involve the FAB-MAP technique [4] which was recently demonstrated to be successful at matching places in routes over a long period time. It is a probabilistic appearance-based approach to place recognition. Each time an image is taken, its visual descriptors are detected and extracted.

In our system, we utilize SURF extractors and descriptors for creating a visual vocabulary dictionary. A Chow Liu tree is used to approximate the probability distribution over these visual words and the correlations between them. **Fig. 4(a)-(b)** shows the extracted features and visual words to build the visual dictionary. Beyond the conventional place recognition approaches that simply compare image similarity between two visual descriptors. FAB-MAP examines co-occurrence of visual words for the same subject in the world. For example, **Fig. 4 (c)** shows that for several windows, some visual words co-appearances are present.



**Fig. 4.** FAB-MAP algorithm to learn places. (a) SURF features are extracted from image sequences. (b) Visual words defined from SURF extractors. (c). Co-occurrence of visual words corresponding to same object.

Consequently, the distinct scenes are learnt from visual training data. For updating new places, 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. **Fig. 5(a)** shows only few places are marked by the first travel, whereas various places that are updated after the second travel as shown in **Fig. 5 (b)**.



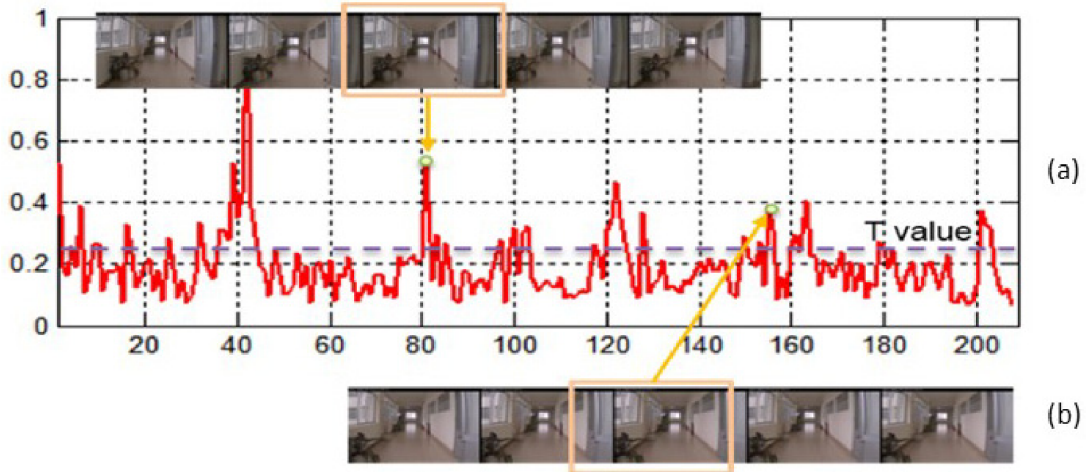
**Fig. 5.** (a) The places are learnt and their corresponding positions are shown in the constructed map data. (b) Many new places are updated after second trial.

### 3.5 Distinguishing Scenes for Improving FAB-MAP’s Performances

In related works [8], [6] report that FAB-MAP obtains reasonable results for place recognition over long travels in term of both precisions and recall measurements.



However, those experiments were implemented in outdoor environments which usually contain discriminate scenes. The original FAB-MAP [2] still has unresolved problems in discriminating scenes to define visual dictionary. This issue affects the results of FAB-MAP when we deploy it in indoor environments, where scenes are continuous and not clearly distinct.



**Fig. 6.** (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.

Therefore, a pre-processing step is proposed to handle these issues. Given a set of scene images  $S = \{I_1, I_2 \dots I_n\}$  we learn key frames from  $S$  by evaluating inter-frame similarity. A feature vector  $F_i$  is extracted for each image  $I_i$ . In this work, the GIST feature [2] is utilized to build  $F_i$ . GIST presents a brief observation or a report at the first glance of a scene that summarizes the quintessential characteristics of an image. Feature vector  $F_i$  contains 512 responses which are extracted from an equivalent of model of GIST proposed in [11]. A Euclidean distance  $D_i$  between two consecutive frames is calculated to measure dissimilarity. **Fig. 6(a)** shows distance  $D_i$  of a sequence including 200 frames. The key-frame then is selected by comparing  $D_i$  with a pre-determined threshold value  $T$ . Examples of selecting two key-frames are shown in **Fig. 6(b)**.

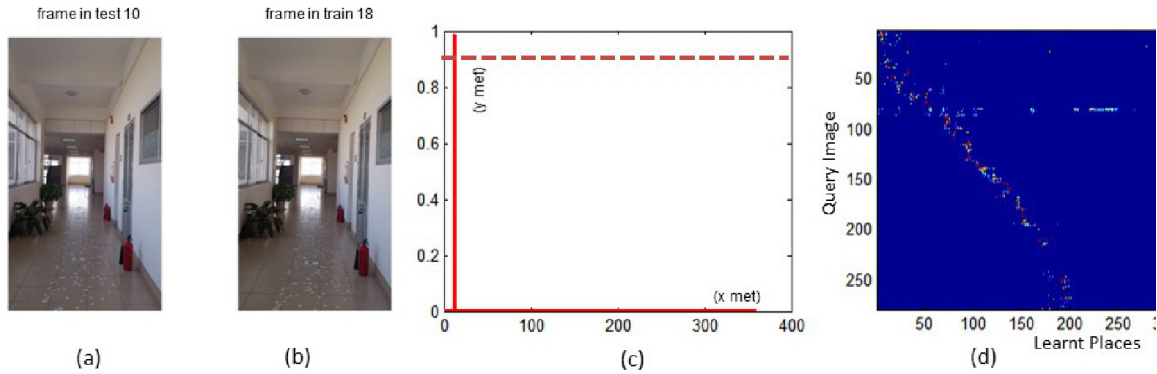
### 3.6 Localizing Places to Earlier Visited Ones in the Constructed Map

Given a current view, its position on the map is identified through a place recognition procedure. We evaluate the current observation at location  $L_i$  on the map by its probability when given all observations up to a location  $k$ :

$$\rho(L_i|Z^k) = \frac{\rho(Z_k|L_i)\rho(L_i|Z^{k-1})}{\rho(Z_k|Z^{k-1})} \tag{1}$$

Where  $Z_k$  contains visual words appearing in all observations up to  $k-1$ ; and  $Z^k$  presents visual words at current location  $k$ . These visual words are defined in the learning places phase. A probability  $p(Z_k|L_i)$  infers observation likelihood as learnt in the training data. In our system, a  $L_i$  is matched at a place  $k^*$  when  $argmax(p(Z_k|L_i))$  is large

enough (through a pre-determined threshold  $T = 0.9$ ). The **Fig. 7** shows an example of the matching procedure.



**Fig. 7.** (a) Given a current observation, (b) the best matching place. (c) The probability  $p(L_i|Z^k)$  calculated with each location  $k$  among  $K = 350$  learnt places. (d) Confusion matrix of the matching places with a sequence of collected images (290 frames).

Given an observation as shown in **Fig. 7(a)**, the best matching place is found at  $placeID = 12$ . The probability  $p(L_i|Z^k)$  is shown in **Fig. 7(c)** with a threshold  $value = 0.9$  whose the maximal probability is  $placeID = 12$ . A confusion matrix of the matching places for an image sequence is shown in **Fig. 7(d)**. This example shows that we can resolve most places in a testing phase.

### 3.7 The Kalman Filter (KF)

In our context, the observations of the robot are images captured over time, which are then converted to coordinates  $(x, y, z)$  in a predefined coordinate system using above matching procedure. However, in indoor environment, the scene does not always change significantly. Consecutive scenes could repeat when the robot moves. Therefore, the performance of image matching is not good. Sometimes, a current observation could be matched with a very far forward / backward image that makes incorrect localization of the robot. To overcome this problem, we propose to use a Kalman filter to correct the position of the robot from observation. A Kalman filter is one of the most popular techniques to improve SLAM results. In our context, we suppose that the robot moves in a flat plane, so the  $z$  coordinate of the robot is constant then we can ignore it. The *state vector* of the robot at a given time  $k$  is simply presented by its coordinates and velocity in two directions  $x$  and  $y$ . *Observation vector* is defined at each time where the image matching is found, the position of the robot could be estimated. We use this information as observation in Kalman filter. *State transition model*  $\mathbf{F}_k$  allows to predict the state vector at time  $k+1$  :

$$x_{k+1} = F_k * x_k + w_k \quad (2)$$

Where  $w_k$  is process noise, which is assumed to follow a normal distribution with covariance  $Q_k$ :  $w_k \sim N(0, Q_k)$ . *Observation model*  $\mathbf{H}_k$  maps the true state space into the observed space:

$$z_k = H_k * x_k + v_k \quad (3)$$

In our case:  $H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$  Where  $v_k$  is observation noise which is assumed to be zero mean Gaussian white noise with covariance  $R_k$ :  $v_k \sim N(0, R_k)$

## 4 Experimental Results

### 4.1 Evaluation Environments

- *Experimental environments*: We examine the proposed method in a corridor environment of a building. The evaluation environment is shown in **Fig. 8(c)**. The total length of the corridor is about 60 m.
- *Database*: Two camera devices are mounted onto a vehicle as shown in **Fig. 1(c)**. The vehicle moves at a speed of 1.25 feet/second along the corridor. The total length of the corridor is about 60 m. We collect data in four times (trials), as described in **Table 1**

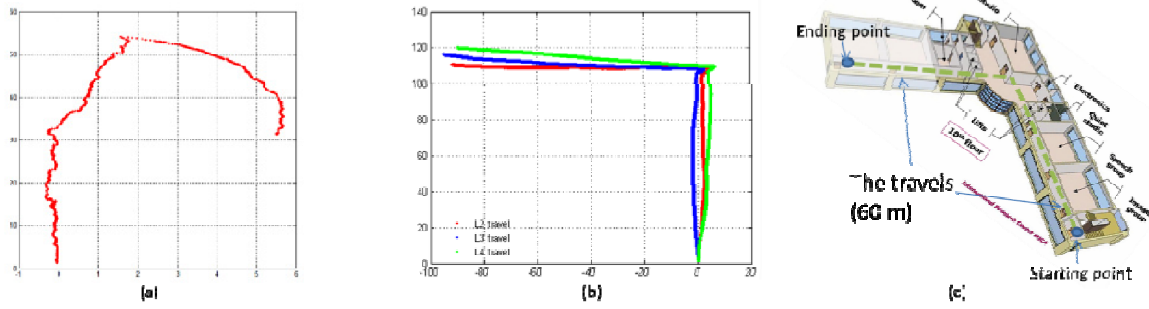
**Table 1.** Three rounds data results

<b>Trials</b>	<b>Total Scene images</b>	<b>Total road images</b>	<b>Duration</b>
L1	8930	2978	5:14
L2	10376	2978	5:30
L3	6349	2176	3:25
L4	10734	2430	4:29

### 4.2 Experimental Results

For map building, we use image acquisitions from L2, L3, and L4 trials. Results of the constructed map using original work of Van Hamme et al. [8] is shown in **Fig. 8(a)**, whereas the reconstructed travels using proposed method are shown in **Fig. 8 (b)**.

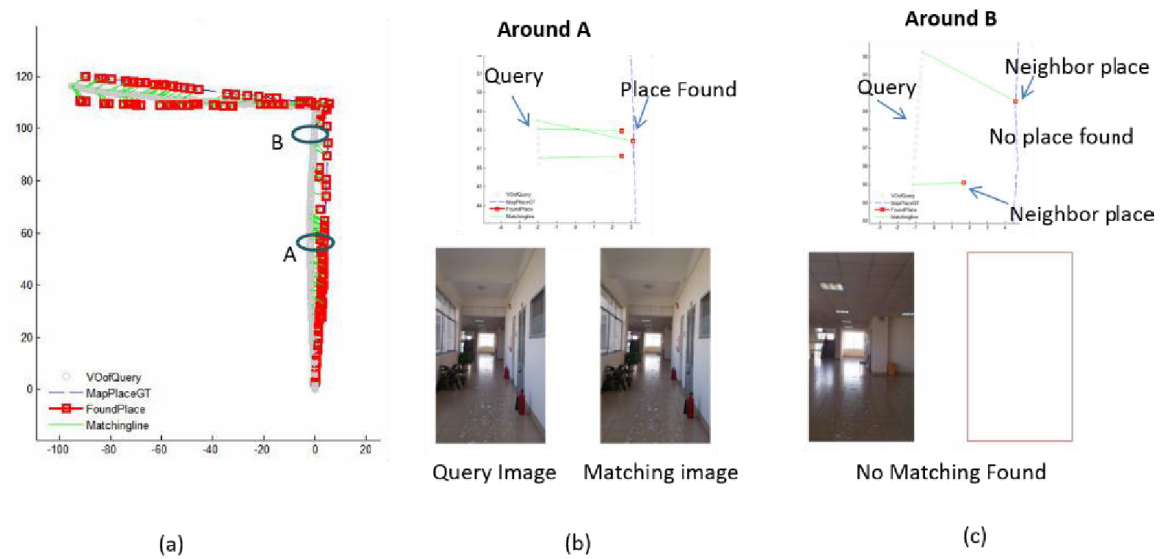
As shown, the results of map building from three travels are quite stable. All of them are matched to ground truth that are plotted in green dash-line in a model 3-D of the evaluation environments, as shown in **Fig. 8 (c)**. Our results are a substantial improvement on the ones using original method [8] without additional markers. We believe that creating highly textures on ground plane is more efficient for detecting and matching the features. The original algorithm in [8] is designed to be robust against high uncertainty of the detected features, but requires many features to create a high quality map.



**Fig. 8.** (a) The travel reconstructed using original works [8]. (b) Results of three time travels (L2, L3, and L4) using proposed method. (c) A 3-D map of the evaluation environment. The actual travels also plotted in green dashed line for comparing results between (a) and (b).

We continue evaluating the proposed system with aspects of the place recognition rate on the created map. To define the visual word dictionary as described in Sec.3.4, we use collected images from L1 trial. About 1300 words are defined in our evaluation environment. We then use dataset from L4 travel to learn places along the travel. In total,  $K = 140$  places are learnt. The visual dictionary and descriptors of these places are stored in XML files. The collected images in L2 and L3 travels are utilized for the evaluations.

Visually, some matching places results from L3 travel are shown in **Fig. 9**. Two demonstrations are shown in details in **Fig. 9** (around position A and position B). Case a shows a query image (from L3 travel) is matched to a learnt place. Therefore, its corresponding positions on the map is able to localize. A zoom-in version around position A is shown in the top panel. Case b show a “no place found” that query image was not found from learnt place database. For the qualitative measurement, we then evaluate the proposed system using two criteria: Precision is to measures total place detected from total query images, whereas Recall is to measure correct matching places from detected places. We setup a predetermined threshold for matching place ( $T = 0.9$ ).



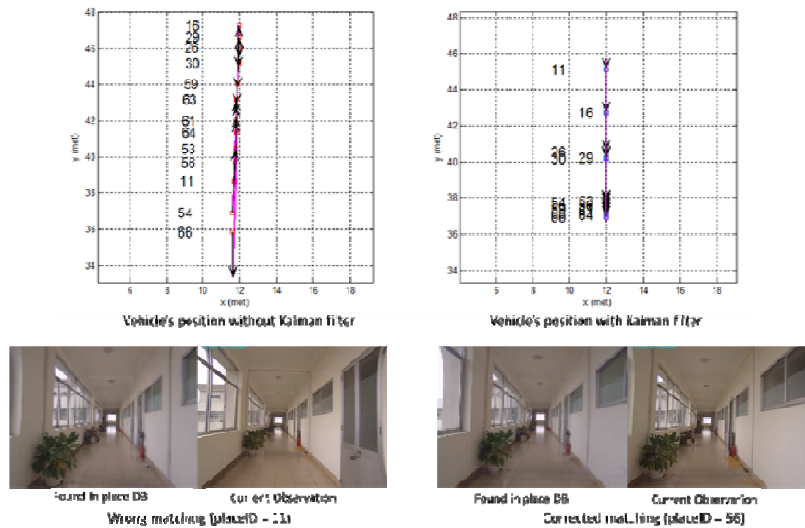
**Fig. 9.** (a) Results of the matching image-to-map with L3 trial. Two positions around A and B are given. (b)-(c): current view is on the left panel (query image); matching is on the right panel. Upper panel is a zoom-in around corresponding positions.

**Table 2.** Result of the matching places (FAB-MAP algorithms) without and with Scene discriminations

Travels	Without scene discrimination		With scene discrimination	
	Precision	Recall	Precision	Recall
L2	12%	90%	67%	82%
L3	36%	85%	<b>74%</b>	<b>88 %</b>

The **Table 2** shows precision and recall with L2 and L3 travels with/without scene discrimination step. For learning places (using original FAB-MAP, without scene discrimination), the recall of L3 travel is clearly higher than L2. The main reason is that some “new” places which were not learnt from L4 are able to update after L2 running. Therefore, more “found” places are ensured with L3 travel. **Table 2** also shows the efficiency of the scene discriminations step, the performances of image-to-map matching obviously increasing and stable for precisions measurement with scene discrimination step, whereas high confidence of the recalls is still consistent.

To show effectiveness of applying the Kalman filter, **Fig. 10** demonstrates navigation data without and with using Kalman filter. Using only the place recognition results (**Fig. 7**- left panel), the directions supporting navigation services are obviously uncontrolled. Some matching places (show in numbers) are misses and in the wrong order in this case. The main reason is the erroneous matching of some places (e.g., place ID = 11, shown in bottom panel). By using a Kalman Filter, directions supporting navigation services are correctly ordered. We can clearly observe the effectiveness on **Fig. 7**- right panel.



**Fig. 10.** Vehicle moving without/with Kalman Filter. Top row: Left panel: vehicle positions on the map using only results of the matching image-to-map procedures. The arrows show directions to guide vehicle. Numbers on left of each red box show *placeID* of the current observation. Right panel: positions of the vehicle are updated using Kalman filter. Bottom row: Left panel: This result shows wrong direction to vehicle. Right panel: is a good matching with Kalman filter.



## 5 Conclusions

In this paper, we presented a visual SLAM system with mobile robot supporting localization services to visually impaired people. We successfully created a map of the indoor environment using the visual odometry and learning places. The results of matching image-to-map are of high confidence for navigation service thanks to the application of a Kalman filter. The proposed system therefore is therefore feasible for deploying navigation services in indoor environments. The proposed system provides direction support for blind/visually impaired people. Further in-the-loop evaluations with the visually impaired/blind people will direct us to future work.

**Acknowledgment.** This work is supported by the project “*Visually impaired people assistance using multimodal technologies*” funded by the Vlaamse Interuniversitaire Raad (VLIR) in the framework of the VLIR’s Own Initiatives’ Program 2012 under the grant number ZEIN2012RIP19

## References

1. Bailey, T., Durrant-Whyte, H.: Simultaneous Localisation and Mapping (SLAM): Part II State of the Art. *IEEE Robotics & Automation Magazine* **13**(2), 99–110 (2006)
2. Bigham, J.P., Jayant, C., Miller, A., White, B., Yeh, T.: VizWiz::LocateIt - enabling blind people to locate objects in their environment. In: *CVPR Workshops 2010*, pp. 65–72 (2010)
3. Borenstein, J., Ulrich, I.: The guidecane-a computerized travel aid for the active guidance of blind pedestrians. In: *Proceeding of ICRA 1997*, pp. 1283–1288 (1997)
4. Cummins, M., Newman, P.: FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research* **27**, 647–665 (2008)
5. Dakopoulos, D., Bourbakis, N.G.: Wearable Obstacle Avoidance Electronic Travel Aids for Blind: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* **40**, 25–35 (2010)
6. Fernández Alcantarilla, P.: Vision based localization: from humanoid robots to visually impaired people. In: *Electronics, University of Alcalá, Ph.D. Thesis* (2011)
7. Fraundorfer, F., Scaramuzza, D.: Visual Odometry: Part II: Matching, Robustness, Optimization, and Applications. *IEEE Transaction on Robotics & Automation Magazine* **19**, 78–90 (2012)
8. Hamme, D.V., Veelaert, P., Philips, W.: Robust visual odometry using uncertainty models. In: *The Proceedings of 13th International Conference on Advanced Concepts for Intelligent Vision Systems, Belgium* (2011)
9. Helal, A., Moore, S.E., Ramachandran, B.: Drishti: an integrated navigation system for visually impaired and disabled. In: *Proceedings of the Fifth International Symposium on Wearable Computers 2001*, pp.149–156 (2001)
10. Kulyukin, V., Gharpure, C., Nicholson, J., Pavithran, S.: RFID in robot-assisted indoor navigation for the visually impaired. In: *The Proceeding of 2004 IEEE/RSJ IROS 2004*, pp. 1979–1984 (2004)
11. Liu, J.J., Phillips, C., Daniilidis, K.: Video-based localization without 3D mapping for the visually impaired. In: *The proceeding of CVPR Workshops 2010*, pp. 23–30 (2010)

12. Loomis, J.M., Golledge, R.D., Klatzky, R.L.: GPS-based navigation systems for the visually impaired. In: *Fundamental of Wearable Computers and Augmented Reality* (2001)
13. Murali, V.N., Coughlan, J.M.: Smartphone-based crosswalk detection and localization for visually impaired pedestrians. In: *Proceeding of IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–7 (2013)
14. Newman, P., Kin H.: SLAM-Loop Closing with Visually Salient Features. In: *The Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp. 635–642 (2005)
15. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: *The Proceeding of Computer Vision and Pattern Recognition 2006*, pp. 2161–2168 (2006)
16. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* **42**, 145–175 (2001)
17. Pradeep, V., Medioni, G., Weiland, J.: Robot vision for the visually impaired. In: *The Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.15–22 (2010)
18. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: *The Proceeding of the Computer Vision and Pattern Recognition*, pp. 1–7 (2007)
19. Shoal, S., Borenstein, J., Koren, Y.: Auditory guidance with the Navbelt-a computerized travel aid for the blind. *Trans. Sys. Man Cyber Part C* **28**, 459–467 (1998)
20. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: *Proceedings of the 9th IEEE CVPR 2003*, pp. 1470–1477 (2003)
21. Sunderhauf, N., Protzel, P.: Brief-gist-closing the loop by simple means. In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* pp. 1234–1241 (2011)
22. Winlock, T., Christiansen, E., Belongie, S.: Toward real-time grocery detection for the visually impaired. In: *The Proceeding of CVPRW 2010*, pp. 49–56 (2010)