ISSN 0868 - 3980

Tạp chí

KHOA HỌC & CÔNG NGHỆ

CÁC TRƯỜNG ĐẠI HỌC KỸ THUẬT

JOURNAL OF

SCIENCE & TECHNOLOGY

TECHNICAL UNIVERSITIES

C

No. 101
2014

CONTENTS

A Vision-Based System for Autonomous Map Building and Localization

Quoc- Hung Nguyen[1], Hai Vu[1], Thanh- Hai Tran[1], Quang- Hoan Nguyen[2]*
[(1)] International Research Institute MICA, HUST - CNRS/UMI 2954 - Grenoble INP
Hanoi University of Science and Technology, No. 1, Dai Co Viet Str., Hai Ba Trung, Ha Noi, Viet Nam
[(2)] Hung Yen University of Technology and Education
Received: March 3, 2014; accepted: April 22, 2014

Abstract

This paper describes techniques for developing a visual-based system that autonomously operators for both building a map and localization tasks. The proposed system tends to assist navigation services in small or mid-scale environments such as inside a building where conventional positioning data such as GPS, WIFI signals are often not available. We firstly design an image acquisition system to collect visual data. On one hand, a robust visual odometry method is adjusted to precisely create a map of the indoor environment. On the other hand, we utilize the FAB-MAP (Fast Appearance-Based Mapping) algorithms. We propose a scene discrimination procedure using GIST feature to dealing with issues of the FAB-MAP. In our experiments, these enhancements give better results comparing with original techniques for both map-building and localization tasks. Such results confirmed that the proposed system is feasible to support visually impaired people navigating in the indoor environments.

Keywords: Visual Odometry, Place Recognition, FAB-MAP algorithms

1. Introduction

Understanding and representing environments have been research topics over long period of time in field of the autonomous mobile robots. These research works aim to answer two questions. The first question is that given a presentation, “What does the world look like?” (Or build map of the environment).

In contrast to this, localization service is to estimate the pose of object of interest relative to a position on the created map. It is to answer the second question “Where am I?” (Positioning the object of interest on the created map). To solve these questions, the positioning data come from various types of sensors such as GPS, WIFI, LIDAR, and Vision (with single or multiple camera systems). However, these source data are not always available or convenient acquisitions, particularly, in a small or moderately environment. To overcome these issues, this paper presents a vision-based system that solely relies on the visual sensor. Our proposed system therefore is flexible and easily setup in indoor environments.

The proposed system aims to automatize map building and localization services. Main purpose of the map building task is to create available trajectories and learn scene elements from the evaluated environments. We simultaneously collect visual data using an own-designed imaging acquisition system. We utilize a robust visual odometry technique to build trajectory using only one consumer-grade camera. In order to learn places in the environment, we utilize so-called loop closure detections method [1] [2]. For localizing task, an agent (such as vehicle, robot, and human) is required to wear a consumer-grade camera. The current observation is matched to the place database which was learnt. This matching procedure is similar to place recognition.

A probabilistic model in FAB-MAP algorithms (Fast Appearance Based Mapping) is utilized to find the maximal likelihood. It is notice that our proposed system is not able to update new positions against the created map. We simply past new places using a simple motion that is based on positions of the closest neighbor places. We evaluate results of the matching image-to-map through travels along corridor in a large building. Experimental results show that we successfully create a map of the evaluated environments. The results of matching places on the map are successful with 74% precision and 88% recall. The main contributions of the proposed method are:

- Improving quality of the map building algorithms that usually make incremental errors after long travels.

- Exploiting discriminate scenes in order to create an efficient visual dictionary of the FAB-MAP algorithms.

- The proposed system is validated through experiments. It confirms that the visual solution is feasible to navigate the visually impaired people.

[*] Corresponding author: Tel: (+84) 912251253,
Email: Quoc-Hung.Nguyen@mica.edu.vn

The next sections of paper are organized as follows: In Section II, we briefly survey related works. In Section III, we present our vision-based system for autonomous map building and localization. We report the experimental results in Section IV. Finally, we conclude and give some directions for future works.

2. Related Works

The vision-based mapping and localizing services are fundamental topics in field of mobile robotics and computer visions. There are uncountable publications of these topics. Readers can refer a good survey in [1]. In this section, we focus on recent advanced techniques in computer vision that offer substantial solutions with respect to localization and navigation services in known or unknown environments.

Alcantarilla [3] utilizes well-known techniques such as Simultaneous Localization and Mapping (SLAM) and Structure from Motion (SfM) to create 3-D Map of indoor environments. He then utilized means of visual descriptors (such as Gauge Speeded Up Robust Features, G-SURF) to mark local coordinate on the constructed 3-D map. Instead of building a prior map, Lui et al. [4] utilized a pre-captured reference sequence of the environment. Given a new query, their system desired to find the corresponding set of indices in the reference video.

By means of visual SLAM techniques, some wearable applications are proposed. Pradeep et al. [5] presented a head-mounted, stereo-vision for detecting obstacles in the path and warn subjects about their presence. They incorporated the visual odometry and feature based metric-topological SLAM. Murali et al. in [6] estimated the users location relative to the crosswalks in the current traffic intersection. They developed 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 [6] required supplemental images from Google Map services, therefore it was suitable with travels at outdoor environments only.

Complexity of the map building task varies in function of the environment size. For example, indoor environment is more complex than outdoor environment because of the office supplies as chairs, tables, etc. Furthermore, matching a current view to a position on the created map seems to be the hardest problem in many works [1,7]. In our point of view, an incremental map is able to increase accuracy of the matching procedures. Therefore, different from above systems, our approach tends to create an incremental map through many trials. When new observations appear, they can be locally and globally consistent with the previous construction. These problems are solved through the loop closure algorithms [2,8]. However, major differences from those work are that our proposed system builds up using sole visual data for both tasks: building a map and localization services. Whereas the works in [2],[8] used GPS data for localizing on the map. Furthermore, the localization service in this paper suffers from limitations of a consumer camera or smart-phone device. The matching image-to-map's performance is acceptable.

3. The Proposed Approach

3.1. Imaging Acquisitions System

We design a compact imaging acquisition system to capture simultaneously scene and route in the experimental environments.

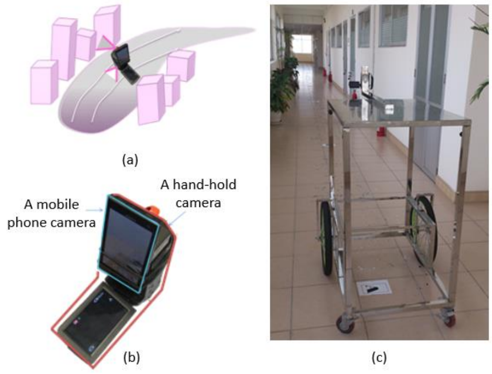

Fig. 1. (a) A schematic view of the visual data collection scheme. (b) The proposed imaging system. Mobile phone camera is attached on rear of a hand-hold camera. (c) The cameras are mounted on wheel-vehicle

A schematic view the system is shown in ***Fig. 1(a).*** It has two cameras. One captures scenes around the environment. The second one captures road on the travel. The camera setting is shown in Fig. 1(b). These cameras are mounted on a wheel-vehicle, as shown in Fig. 1(c). The details of the collected data in the evaluated environment are described in Section IV.

3.2. The proposed framework

General framework of the proposed system is shown in Fig. 2. It has two phases, as described below.

- **Offline 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. III.3, Sec. III.4, 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 observation is described using a visual dictionary. These data are associated matching images to the places where are labeled in the database. The current pose thus is localized on the travel.

3.3 Route building based on visual odometry techniques

To build route of the travel , we utilize a visual odometry method proposed by Van Hamme et al [9]. 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. Our system setups 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. To solve these issues, we utilize the man-markers in the whole journey as shown in Fig. 3(b-c). Because the detected features are projected on a ground plane, it is more accuracy for detecting and matching features.

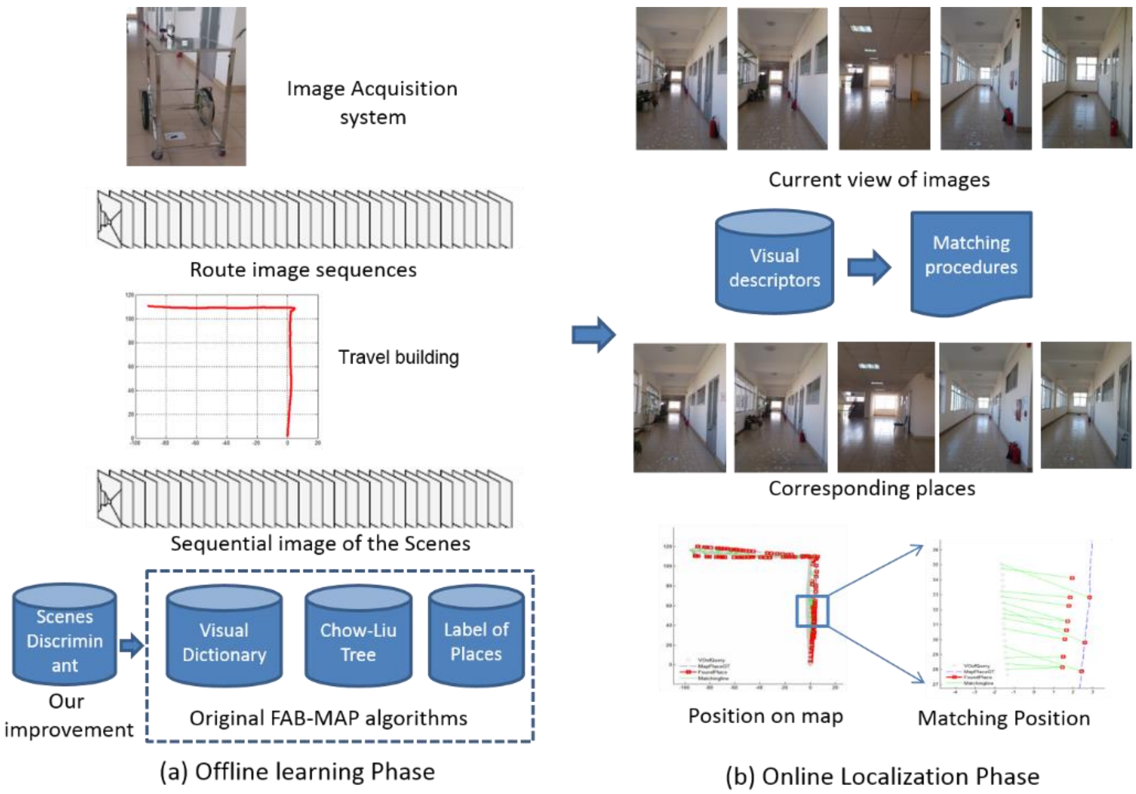

Fig. 2. The framework of the proposed system

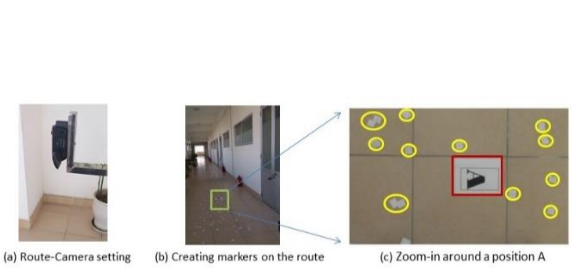

Fig. 3. The collection databases on Road

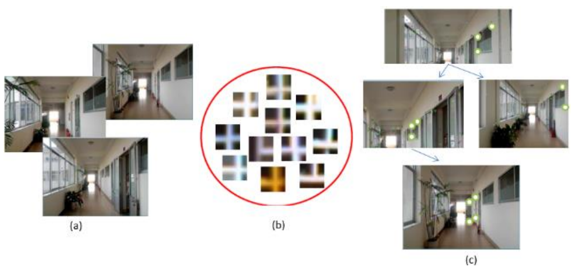

Fig. 4. FAB-MAP algorithm to learn places. (a) SURF features are extracted from image sequences. (b) Visual words defined from SURF extractors. (c). Co-occur of visual words by same object

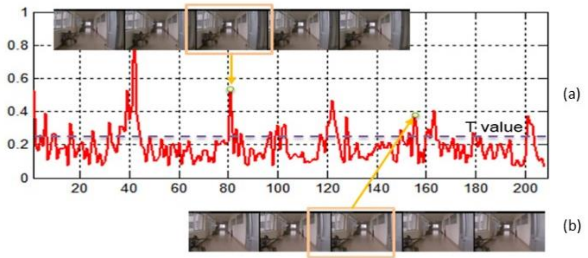

Fig. 5. (a) Dissimilarity between two consecutive frames. A threshold value T = 0.25 is pre-selected. (b) Two examples shows the selected key frames and their neighbor frames

3.4 Original FAB-MAP algorithms for leaning places on the travel

The learning places aim to visually present scenes along the travel. These visual presentations need to be easy implementation and efficient distinguishing scenes. To adapt with these issues, we utilize the FAB-MAP algorithms [2] which is recently successful for matching places in routes over long period time. It is a probabilistic appearance-based approach to place recognition. Each time the image taken, its visual descriptors are detected and extracted. In our system, we utilize SURF extractors and descriptors for creating on 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 visual dictionary. Furthermore, FAB-MAP involves co-occur visual word of same subject in the worlds. For example, Fig. 4(c) shows window subject in various contexts.

3.5. Distinguishing scenes for improving FAB-MAP's performances

Although related works [2,8] 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. Original FAB-MAP [2] is still unresolved problems of discriminating scenes to define visual dictionary. This issue affects to results of FAB-MAP when we deploy it in indoor environments, where scenes are continuous and not clearly distinct. Therefore, a pre-processing step is proposed to handle these issues. Given a set of scene images S={I_1, $I_{2,\ldots,}$ I_n} we learn key frames from S by evaluating similarity of intra-frames. A feature vector F_i is extracted for each image I_i. In this work, the GIST feature [10] 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

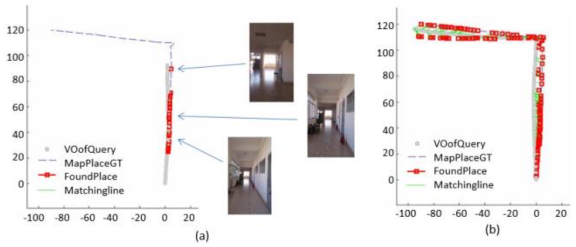

Fig. 6. (a) The places are learnt and their corresponding positions are shown in the constructed map data. (b) Many new places are updated after second trial.

512 responses which are extracted from an equivalent of model of GIST proposed in [10]. A Euclidean distance D_i between two consecutive frames is calculated to measure dissimilarity. Fig. 5(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. 5 (b)

3.6 Localizing a place to visited one in the constructed map

For updating new places, we implement captured images through several trials. For each new trial, we compare the image with the previous visited places which are already learnt. This procedure calls a loop closure detection. These detections are essential for building an incremental map. Fig. 6(b) shows various places that are updated after the second travel, whereas only few places are marked by the first travel (Fig. 6(a)).

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:

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

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 that 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). ***Fig. 7*** shows an example of the matching procedure.

Given a current observation as shown in Fig. 7(a), the most 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 a sequence of the collected images is shown in Fig. 7 (d) infers that we can resolve almost places in a testing phase.

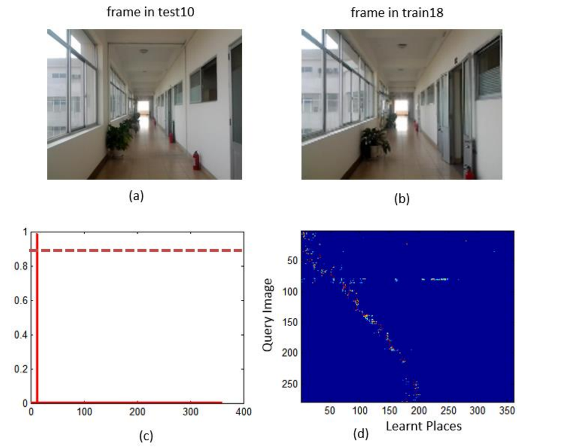

Fig. 7. (a) Given a current observation, (b) the most matching place. (c) The probability $p(L_i|Z^k)$ calculated with each location k among K = 350 learnt places. (d) Confusion matrix of the matching places with a sequential collected images (290 frames).

4. Experimental results

4.1 Evaluation Environments

- **Setting up environments**: We examine the proposed method in an indoor building, where is 10th floor of International Research Institute MICA Hanoi University of Science and Technology (HUST). A 3-D model of the evaluation environment is shown in Fig. 8(c).
- **Database collection:** Two camera devices are mount into a vehicle as shown in Fig. 1(c). A person moves at a speed of 0.4 m/second along the corridor. The total length of the corridor is about 60m. We collect data in four times (trials), as described in

Table. 1. Three rounds data results

Trials	Total Scene images	Total road images	Duration
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 [9] is shown in Fig. 8(a), whereas the reconstructed travels using proposed method are shown in Fig. 7(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 substantial comparing with the ones using original method [9]. We believe that creating highly textures on ground plane is more efficient for detecting and matching the features. Event original algorithm in [9] is designed to be robust with uncertainty of the detected features, more precisely the features matching more higher quality creating the map.

We continue evaluating the proposed system with aspects of the place recognition rate on the created map. To define visual word dictionary as described in Sec. III.4, we use collected images from L1 trial. About 1300 words are defined in our evaluation environments. We then use dataset from L4 travel to learn place along the travel. Totally, 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*.

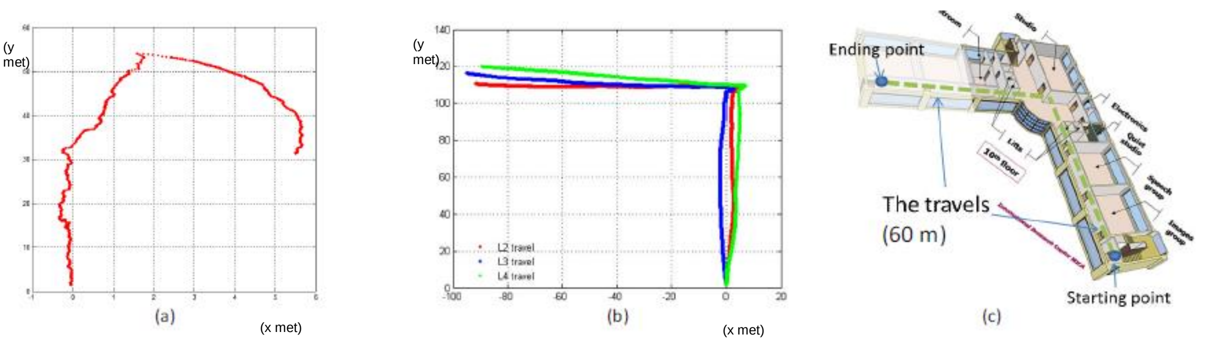

Fig. 8. (a) The travel reconstructed using original works [1] (b) Results of three time travels (L2, L3, and L4) using proposed method. (c) A 3-D map of the evaluation environment. The actual travels also plotted in green dashed line for comparing results between (a) and (b).

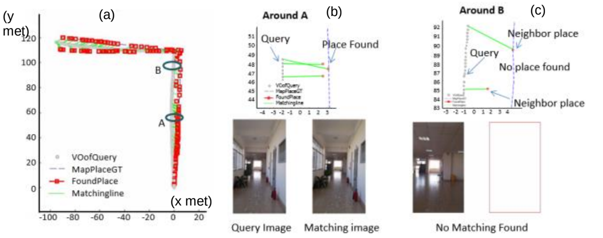

Fig. 9. (a) Results of the matching image-to-map with L3 trial. Two positions around A and B are given. (b)-(c): current view is on the left panel (query image); matching is on the right panel. Upper panel is a zoom-in around corresponding positions.

Two demonstrations are shown in details in 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). Table. 2 shows precision and recall with L2 and L3 travels with/without scene discriminant step. For learning place (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 where were not learnt from L4 are able to update after L2 running. Therefore, more "found" places is ensured with L3 travel. Table 2 also shows efficient of scene discriminations step (Sec. III.5). 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.

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%	**74%**	**88 %**

It is notice that although the evaluations report moderately precisions (Table 2, with scene discrimination), it is acceptable for localizing services. An agent (robots, wheel-vehicle) moves approximately at speed of 10 cm/sec and computational time at 1 frame/sec. Therefore, the precision of 65% (averagely) means that by moving a distance of 100 cm (10 frames captured), 6~7 positions (in a distance of 1 m) are confidence located.

5. Conclusions

In this paper, we presented a feasible vision-based system for both tasks: building map and localizing services. We successfully created the map using the visual odometry and learning places techniques. The matching image-to-map procedure gives high confidence results. The experimental results confirmed that the system is able to deploy mapping services in an indoor environment. Further evaluations with complex environments are implemented in future works. The proposed system also needs to fuse with other services to support navigating services to visually impaired people.

Acknowledgements

This work was supported by the Research Grant from Vietnam's National Foundation for Science and Technology Development (NAFOSTED), No. FWO.102.2013.08

References

[1] Bailey, T. and H. Durrant-Whyte, Simultaneous localization and mapping (SLAM): part II state of the art. Robotics & Automation Magazine, IEEE, 13 (2006) 108-117.

[2] Cummins, M. and P. Newman, FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. Int. J. Rob. Res., 27 (2008) 647-665.

[3] Fernández Alcantarilla, P., Vision based localization: from humanoid robots to visually impaired people, in Electronics2011, University of Alcala: Ph.D. Thesis.

[4] Liu, J.J., C. Phillips, and K. Daniilidis. Video-based localization without 3D mapping for the visually impaired. in Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. 2010.

[5] Pradeep, V., G. Medioni, and J. Weiland. Robot vision for the visually impaired. in Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. 2010.

[6] Murali, V.N. and J.M. Coughlan. Smartphone-based crosswalk detection and localization for visually impaired pedestrians. in Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on. 2013.

[7] Fraundorfer, F. and D. Scaramuzza, Visual Odometry : Part II: Matching, Robustness, Optimization, and Applications. Robotics & Automation Magazine, IEEE, 19 (2012) 78-90.

[8] Newman, P. and H. Kin. SLAM-Loop Closing with Visually Salient Features. in Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on. 2005.

[9] Hamme, D.V., P. Veelaert, and W. Philips, Robust visual odometry using uncertainty models, in Proceedings of the 13th international conference on Advanced concepts for intelligent vision systems2011, Springer-Verlag: Ghent, Belgium. p. 1-12.

[10] Oliva, A. and A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. Int. J. Comput. Vision, 42 (2001) 145-175.