IMAGE-BASED LOCALIZATION FOR INDOOR ENVIRONMENT USING MOBILE PHONE

Yewei Huang, Haiting Wang, Kefei Zhan, Junqiao Zhao, Popo Gui, Tiantian Feng

KEY WORDS: Indoor Localization, Image Matching, SIFT, HoG, Mobile phone

ABSTRACT:

Real-time indoor localization based on supporting infrastructures like wireless devices and QR codes are usually costly and labor intensive to implement. In this study, we explored a cheap alternative approach based on images for indoor localization. A user can localize him/herself by just shooting a photo of the surrounding indoor environment using the mobile phone. No any other equipment is required. This is achieved by employing image-matching and searching techniques with a dataset of pre-captured indoor images. In the beginning, a database of structured images of the indoor environment is constructed by using image matching and the bundle adjustment algorithm. Then each image's relative pose (its position and orientation) is estimated and the semantic locations of images are tagged. A user's location can then be determined by comparing a photo taken by the mobile phone to the database. This is done by combining quick image searching, matching and the relative orientation. This study also try to explore image acquisition plans and the processing capacity of off-the-shell mobile phones. During the whole pipeline, a collection of indoor images with both rich and poor textures are examined. Several feature detectors are used and compared. Pre-processing of complex indoor photo is also implemented on the mobile phone. The preliminary experimental results prove the feasibility of this method. In the future, we are trying to raise the efficiency of matching between indoor images and explore the fast 4G wireless communication to ensure the speed and accuracy of the localization based on a client-server framework.

1. INTRODUCTION

In recent years, the increasingly matured global positioning technology has breed various location-based applications, such as the navigation system of mobile devices and even unmanned vehicles and aircrafts. However, the signals of the global navigation system usually fails to penetrate into indoor environment, which leads to the unreliability of the indoor localization. One of the solutions to this problem is by establishing an indoor referencing framework based on wireless communication, i.e. Wi-Fi, Zigbee, Bluetooth etc. [1]. Nevertheless various issues exist in current approaches. They are vulnerable to signal interference and multiple reflections, and the layout scheme of hotspots dramatically affects the accuracy of localization [1]. In addition, this technique can only be applied in coarse indoor positioning with meter level accuracy. Other localization methods based on QR code or RFID [2] are simpler to implement. However these landmarks which provide location reference can only be embedded in limited locations therefore can hardly provide continuous localization in the indoor space.

In this study, we explored a simple, yet accurate image-based approach for indoor localization. A structured image database of an indoor environment is constructed, and used as the indoor location referencing, since each image's relative pose can be estimated. A user's location is determined based on a new captured image by first quickly searching the similar images from the database and then precisely matching them and conduct relative orientation. This idea is not new and has been already explored in many fields such as robotics [3]. However, this study tries to examine the practicability and limitations of such methods for human users in specified environment such as exhibitions and museums. We also try to implement the whole pipeline on off-the-shell mobile phones and plan to explore the fast LTE wireless communication to ensure the speed and accuracy of the localization by using a client-server framework.

The remainder of this paper is organized as follows. In section 2, we discuss related work on image matching and indoor localization. In section 3, we introduce how to build a structured image database and how to localize by image searching and matching. The result of the experiment is presented in section 4. At last, we conclude the paper and describe future improvements in section 5.

2. RELATED WORK

Vision-based localization has drawn intensive attention because of its passive nature and is analogous to human localization [1]. Many methods have been proposed, such as the visual vocabulary tree [4, 5], which packs the SIFT features extracted from images into vectors of visual words. Although the searching is speeded up but constructing the visual words is complex and costly. [3, 6] use landmarks to implement indoor localization. The landmarks are features or group of features detected from the images. They must be stationary, distinctive in the map, repeatable and robust against noise [6]. During the searching period, features which are detected from the query image are matched to the landmarks. Comparing to the previous method, this method is more efficient, but is vulnerable to similar features presented in various locations [6]. In topology-based method [8], a topological map is built from a series of images or video sequences, and then is refined by learning vector quantization (LVQ). During the searching stage, nearest neighbor rule is used to detect the similar region in the query image. This method assumes that the navigation path is unique in the topological map. Therefore, the query image can be misclassified in some case.

There are also methods using stereo images [1, 3, 7]. The stereo images can provide depth information which is helpful for 3D reconstruction. However, they are more or less the same as monocular images in localization applications.

3. INDOOR LOCALIZATION BASED ON STRUCTURED IMAGE DATABASE

The whole pipeline of our image-based localization is shown in Fig. 1. It can be divided into two main stages. The first is the building of structured image database. Our goal is to estimate all the camera poses and position information of pre-

doi:10.5194/isprsarchives-XL-4-W5-211-2015

captured indoor images. All these information compose a structured image database. The second stage is the localization. Once the user takes a photo, the image features are extracted on the fly. These features are then matched with the feature descriptions extracted in the former step, from which the image with the richest matches can be detected. Then the parameters of the relative orientation of the newly taken photo can be estimated. As a result, the location of the user can be identified.

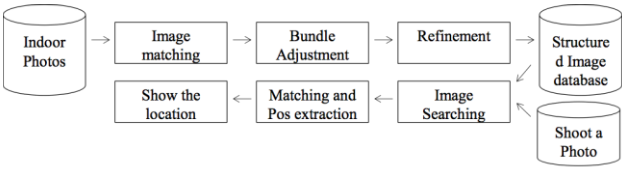

Fig. 1 The pipeline of image-based indoor localization

3.1 Collection of indoor environment images

It is significant to collect proper indoor images because its quality has a great effect on the construction of structured image database and the localization. Ideally, any images of an indoor environment can be incorporated in the database. However, due to the robustness problem of feature descriptors and the richness of image texture, we find the following aspects are important.

First, the collected indoor images should have at least 70% overlap and the camera should not be too close to the target. Otherwise, it is difficult to capture enough textures and would cause difficulties in image matching.

Second, we should avoid view directions containing surfaces of strong reflection, transparency, or vegetation and light sources. All these situations will introduce ambiguous to matching.

Finally, the orientation of camera should be roughly perpendicular to the target and the moving path, so that to relieve the influence of rotation and the problem of reflection for the target objects such as glasses. Fig. 2 show several indoor images used for building the structured image database.

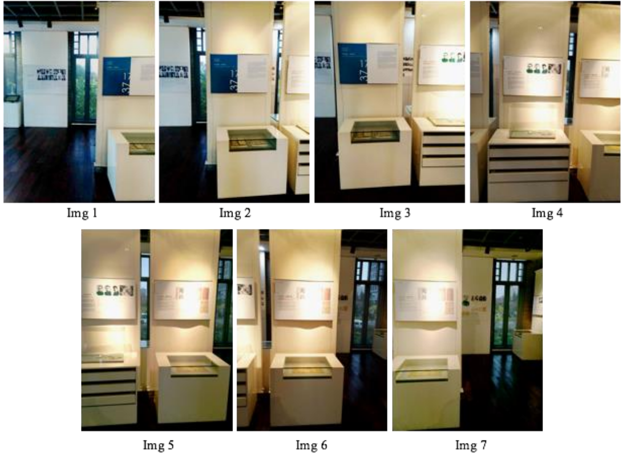

Fig. 2 Examples of the pre-captured indoor image sets

3.2 Construction of structured image database

In this period, the first step is to detect features from the images. We compared the SIFT descriptor and two SIFT-like descriptors: Speeded Up Robust Features (SURF) [8] and Affine-SIFT (ASIFT) [9]. These methods are known to be robust in feature extraction and description. The captured photos are resized to 800x600 to balance the computational cost and the richness of image details. Then the images are matched to their neighbors according to the Flann-based matcher [10]. Fig. 3 show the results of matching using different feature descriptors, from which the results of ASIFT are more accurate and 649 matched points are extracted. The implementation of SIFT and SURF extract more feature points but also introduce distractions to the matching stages, since the RANSAC estimation is influenced by the outliers. As a result, the ASIFT is chosen in our method.

In the second step, all the indoor images are aligned in a queue and we match each image with its neighbors. After that the relative orientation can be conducted on each image pair, hence the relative image and camera poses are calculated. To globally optimize the poses, the bundle adjustment is used [11]. Then we get sparse point cloud and the poses of the images as well.

To construct the structured image database, the poses are described in a unified coordinate frame. We define the frame based on the pose of the first image and all the consequent images are transformed into this frame. The global coordinates are not needed at this moment because the relative location conjugated with the semantic description of the image, e.g. the 2nd floor, room 3, can already provide enough information for localization. However, if the accurate distance is demanded for routing, we have to introduce the global scale factor.

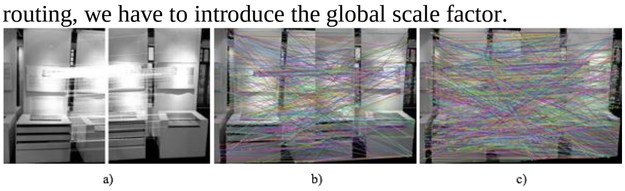

Fig. 3 The comparison of image matching based on ASIFT a), SIFT b) and SURF c)

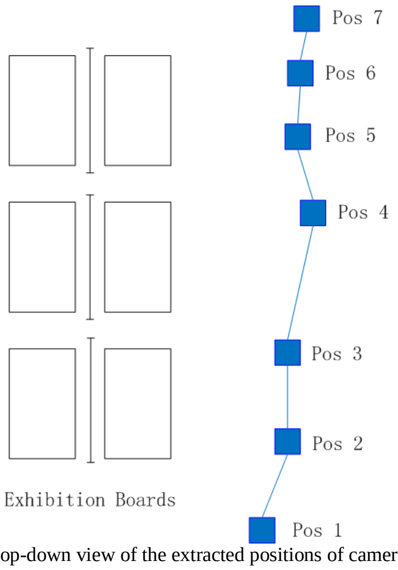

Fig. 4 The top-down view of the extracted positions of cameras from the input image set.

3.3 Image searching based on HoG features

In a localization scenario, we provide the possibility that a user can be aware of his/her ego location simply by just shooting one photo of the indoor scene. This is realized by the technique of fast image searching and deducing the camera's pose using image matching and orientation.

Image searching is known to be nontrivial since the description of an image can be complex which leads to high computation cost. In this study, a two stage searching mechanism is developed. Firstly, a simple statistic-based image description, i.e. the histogram, is used to filter out the non-similar images from the database. Fig. 5 is the newly captured photo by the user and Fig. 6 is the histograms of the previously shown indoor images and the query image. It can be seen that Img5, Img6 and Img7 are most unlike to be the similar scene. Before comparing, the captured image is enhanced by histogram equalization on mobile phone.

The histogram is only used to roughly rank the images by similarity. We should further use more precise features such as SIFT features to directly match the similar images. The number of matched points then can be used as an evaluation of the similarity of images [8]. However, this approach is too burdensome in practice because matching N image pairs are needed (where N is equal to the number of images in the structured image set). Building the image pyramid could hardly help in this case because SIFT features extracted from multi-scaled image set can hardly be compared directly. An alternative is to store and match SIFT features based on a binary-tree index. However, this can be overwhelming when the structure image database is built from too many photos. As a result, the Histogram of Oriented Gradients (HoG) [12] is introduced to detect images containing similar scene (as shown in Fig. 7).

At first, we decompose the query image into 3 levels using quadtree and calculation the HoG features for each image block (Fig. 8). The reason is that all these image patches may contain part of overlapping scene in the database. During localization, the 3-level HoG features of the shoot photo is compared with all the images stored in the image database using a moving window. Therefore once a similar image patch is found, we can deduce that the image may overlap part with the query image. To solve the problem of scale variance, we constructed image pyramid for all the images in the database. After the "similar image" is detected, the image matching using ASIFT is then conducted. And the pose of the query camera can be extracted by the same method as described in section 1.2. However, once the matching is failed. The next image in the ranked list is selected to match with the query image.

The full pipeline of this approach is still in construction but the preliminary results show the effectiveness of this solution. The computational cost of HoG feature extraction and patch detecting is much less then direct SIFT-based image matching. Another problem is that HoG is not rotation invariable. Nevertheless we make an assumption in our application that the photos are always taken either vertically or horizontally. A detection of the aspects of the photo would solve the most of the problem instead of rotating images in multiple degrees.

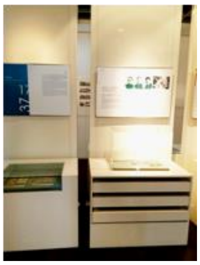

Fig. 5 The newly captured image to be used for localization

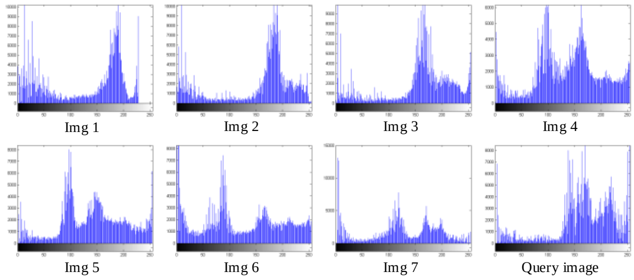

Fig. 6 The histograms of the image set

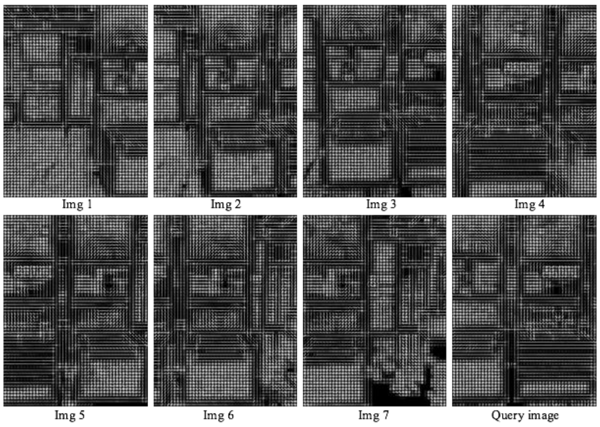

Fig. 7 The HoG feature extracted from the indoor images

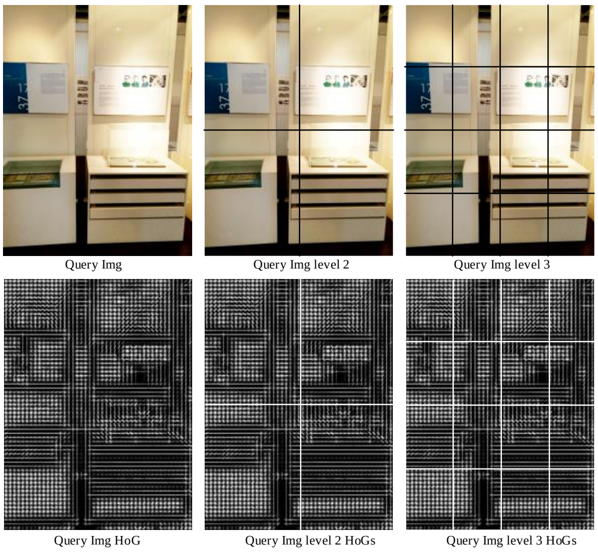

Fig. 8 The 3 level HoG features extracted from the query image

4. RESULTS

We take photos of the second and the third floor of the History Museum of Tongji University with iPhone 5s at a fixed focal length. The camera is control and calibrated by own developed program. The size of the every image is 3264×2448 pixels. Taking the speed of calculation into account, the images were down sampled into the size of 800 ×600 pixels.

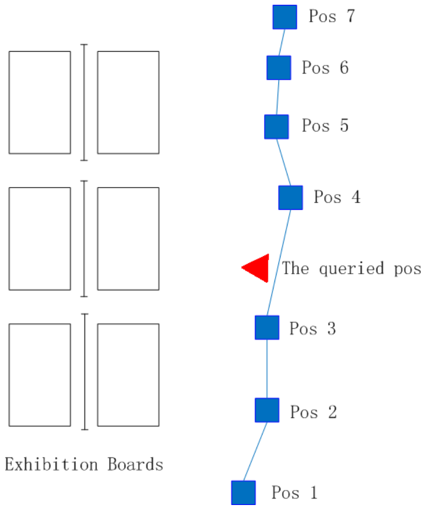

Fig. 9 The calculated camera position from the newly captured image

SIFT and SURF we deployed is based on OpenCV [13], while ASIFT is provide by the author[1]. The original version cost around 20 seconds to matching an image pair. There remains space to lessen the computing time.

The query image can be located with the proposed method and Fig. 9 shows its correct location.. The precision of the result still needs evaluation, however comparing to the wireless-based method this location yields much higher accuracy.

The localization result is shown in Fig. 10, which is an app developed on Android. The red spot displays the location of a user in the floor plan of the museum, and the semantic information is shown on the side.

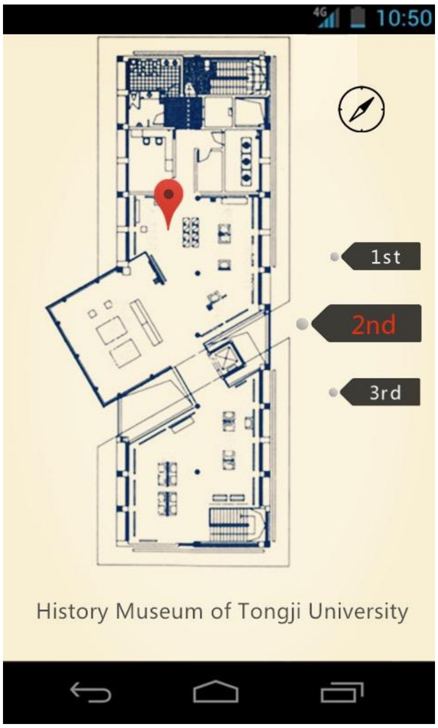

Fig. 10 The localization result shown in mobile app

5. CONCLUSION

In this paper, we present an image-based indoor localization system implemented on mobile phone, which is free from specified devices thus are easy to implement and flexible to use in practice. The key adopted techniques are the image matching and searching. Although the preliminary experiments has shown its feasibility, the accuracy of the results has to be examined in detail and the performance remains space to be improved. To optimize the construction of the database, the geometric constrains of indoor image set can be incorporated. Our HoG-based image searching still needs further refinements. At present, only the photo capture and enhancement is fully implemented on mobile phone using OpenCV. The image searching and matching are migrating from desktop to the mobile platform. We planned to use the client-server framework to relieve the task load on the mobile phone by uploading the image to the server-side and return results on the mobile phone. This can be done by using the fast 4G LTE communication.

Acknowledgement

This research is supported by the NSFC young scholar funding (No.41201379), the Program for Young Excellent Talents in Tongji University (No.2014KJ027) and the National Student Innovation Training Program (No.0250107010).

6. REFERENCES

[1]Mautz, R. (2012). *Indoor positioning technologies* (Doctoral dissertation, Habilitationsschrift ETH Zürich, 2012).

[2]Alghamdi, S., Van Schyndel, R., & Alahmadi, A. (2013, April). Indoor navigational aid using active RFID and QR-code for sighted and blind people. In *2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, (pp. 18-22).

[3]Lategahn, H., & Stiller, C. (2012, July). City gps using stereo vision. In *2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES),* (pp. 1-6).

[4]Sattler, T., Leibe, B., & Kobbelt, L. (2011, November). Fast image-based localization using direct 2D-to-3D matching. In *2011 IEEE International Conference on Computer Vision (ICCV)*, (pp. 667-674). IEEE.

[5]Li, Y., Snavely, N., & Huttenlocher, D. P. (2010). Location recognition using prioritized feature matching. In *Computer Vision–ECCV 2010* (pp. 791-804). Springer Berlin Heidelberg.

[6]Sinha, D., Ahmed, M. T., & Greenspan, M. (2014, May). Image Retrieval using Landmark Indexing for Indoor Navigation. In *2014 Canadian Conference on Computer and Robot Vision (CRV)*, (pp. 63-70).

[7]Lategahn, H., & Stiller, C. (2014). Vision-only localization. *Intelligent Transportation Systems IEEE Transactions on, 15, 3*, 1246 - 1257.

[8]Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. In *Computer vision–ECCV 2006* (pp. 404-417). Springer Berlin Heidelberg.

[9]Morel, J. M., & Yu, G. (2009). ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, *2*(2), 438-469.

[10]Muja, M., & Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. *Visapp International Conference on Computer Vision Theory & Applications*, 331--340.

[11]B. Triggs, P. McLauchlan, & R. Hartley, & A. Fitzgibbon (1999). "Bundle Adjustment — A Modern Synthesis".

[1] http://www.cmap.polytechnique.fr/~yu/research/ASIFT/

ICCV '99: Proceedings of the International Workshop on Vision Algorithms. Springer-Verlag. (pp. 298–372).

[12]Suard, F., Rakotomamonjy, A., Bensrhair, A., & Broggi, A. (2006, June). Pedestrian detection using infrared images and histograms of oriented gradients. In *Intelligent Vehicles Symposium, 2006 IEEE* (pp. 206-212).

[13]Lowe, D. G., “Distinctive Image Features from Scale-Invariant Keypoints”, *International Journal of Computer Vision, 60, 2* (pp. 91-110, 2004).