Vision-based indoor localization via a visual SLAM approach

Minglei Li[1,2], Franz Rottensteiner[2]

[1] College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, China – minglei_li@nuaa.edu.cn
[2] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany – rottensteiner @ipi.uni-hannover.de

Commission IV, WG IV/5

KEY WORDS: Indoor localization, Image retrieval, Geometric constraint, Bag-of-visual-word, SLAM

ABSTRACT:

With an increasing interest in indoor location based services, vision-based indoor localization techniques have attracted many attentions from both academia and industry. Inspired by the development of simultaneous localization and mapping technique (SLAM), we present a visual SLAM-based approach to achieve a 6 degrees of freedom (DoF) pose in indoor environment. Firstly, the indoor scene is explored by a keyframe-based global mapping technique, which generates a database from a sequence of images covering the entire scene. After the exploration, a feature vocabulary tree is trained for accelerating feature matching in the image retrieval phase, and the spatial structures obtained from the keyframes are stored. Instead of querying by a single image, a short sequence of images in the query site are used to extract both features and their relative poses, which is a local visual SLAM procedure. The relative poses of query images provide a pose graph-based geometric constraint which is used to assess the validity of image retrieval results. The final positioning result is obtained by selecting the pose of the first correct corresponding image.

1. INTRODUCTION

Indoor localization systems have a wide range of applications, such as automated industry, health care, augmented reality, and robot-based disaster relief. Since GPS signals are usually unavailable inside buildings, many different techniques have emerged to make up for the lack of positioning system in indoor environment (Möller et al., 2012; Farid et al., 2013; Sánchez-Rodríguez et al., 2017). Most **Vision-based localization (VBL)** techniques do not require complex facilities or devices in buildings, so they are more affordable for some scenes and shows great potential to be deployed in reality.

For a typical VBL system, there is a remote server maintaining a photo database, where each photo is bound with a location tag and other possible auxiliary information. Implementing VBL systems for indoor scenes faces some severe difficulties. First of all, there is universally no global coordinate frame. This means for every building one needs to specify a unique frame to present its spatial character to end-users. A good approach should design an efficient method to guide map makers to record the entire environment. Secondly, the image retrieval process needs to detect invalid correspondences between query images and database images. This task is challenged by different imaging devices and changing imaging conditions. Current VBL systems mainly focus on various aspects of feature matching and position calculation. However, it would still be required to come up with a robust method for solving the general problem of localisation in such environments.

It is common that when a person enters an unfamiliar scene, he/she looks around to determine his/her position. A visual system that uses only one image for localization is error-prone. The indoor environment is complicated due to the repeated occurrence of similar patterns, occlusions and varying lighting conditions, which lead to confusion in image retrieval. Consequently, our idea is that the original query data from the end-user is not a single image but a short sequence of images. We assume more image correspondences can provide additional geometric constraints to validate and strengthen the image retrieval results.

In specific, inspired by the development of visual simultaneous localization and mapping techniques (SLAM), we propose a visual SLAM-based multi-image retrieval strategy, which recovers the geometric relationships between query images to improve the robustness of the VBL system. The accuracy of vision based methods highly depends on the quality of feature correspondences. In our method, the query information includes not only 2D features but also raw relative poses of the sequence of images.

We use a bag-of-visual-words (BoVW) model to find initial candidate corresponding images that are likely to have similar appearance to query images (Nister and Stewenius 2006). The six parameters of exterior orientation are determined by spatial resection. Our approach is closely related to the re-localization step used in some traditional SLAM methods (Eade and Drummond, 2008; Kendall et al., 2015; Mur-Artal and Tardós, 2017). However, the approach differs in two aspects:

(1) Instead of using one image to retrieve corresponding images in the database, we use a sequence of images to retrieve images at the querying stage and the related poses of these images are recovered by a local SLAM method. The related poses are local poses which integrate a pose graph-based geometry constraint further used to detect and handle inaccurate image retrieval results.

(2) We assume the interior orientation parameters of the camera used to generate the query images to be un-calibrated. The focus length is estimated from the image EXIF file and the principal point is assumed to be at the centre of the frame, so the global poses of query images, calculated by spatial resection based on correspondences between map points and image points, are some coarse estimations. After image retrieval, a set of pose graphs of candidate global poses in a global coordinate frame can be built. We design a pose graph-based alignment function,

which calculates the transformation parameters between global poses and local poses by least squares fitting. After minimizing the alignment error, the impact of interior parameters is weakened. Basically, because querying images only span a short distance in the global scene, the alignment accuracy is enough for end users.

2. RELATED WORK

With the emergence of location-based services, several techniques have been proposed to provide indoor locations. One of them is the fingerprinting localization technique, which means to use the signals, such as Wi-Fi or Bluetooth, to measure the distances from some known base stations and calculate the query position using trilateration (Farid et al., 2013; He and Chen, 2016). Fingerprinting technologies suffer from signal attenuation during diffusion, so sometimes the accuracy is a big problem. The localization error can be several meters.

VBL systems have drawn intensive attention in recent years. The first type of VBL systems (*indirect methods*) use an **appearance-based localization** strategy, which merely provide coarse poses by inheriting location tags from the corresponding images in the database. The database contains images, e.g. street view panoramas, images of buildings or landmarks. These methods are mainly applicable to localization tasks at city scale, e.g. for landmark identification and tourism guidance. Robertson and Cipolla (2004) built a database of views of building facades. The pose of a query image is determined by the corresponding facade images using a wide-baseline matching algorithm. Zamir and Shah (2010) constructed a SIFT descriptor (Lowe, 2004) based tree for the database images. In order to retrieve corresponding images of query image, they use a GPS-tag-based pruning method to remove less reliable descriptors and use a voting strategy to determine how reliable the localization of a particular image is.

Another type of VBL systems (*direct methods*) is based on **fine pose localization**, which uses 3D structure information to retrieve images and estimate the poses. In these cases, the spatial information of scenes are recorded through structure from motion (SfM) techniques, and each image in the database is associated with a 3D coordinate and orientation information (Xiao et al., 2008; Irschara et al., 2009; Lim et al., 2012; Li et al., 2012). Then, the localization result is calculated by querying database images and corresponding 3D points to determine the exterior orientation of the query image by spatial resection.

Both for indirect and direct methods, a critical task is to retrieve the corresponding images from the database. The first type of techniques is based on point features, such as the Hessian-affine detector (Lindeberg, 1998) combined with SIFT (Lowe, 2004) or SURF (Bay et al., 2008) descriptors. Some recent visual search systems are based on a BoVW model. BoVW approaches generally extract feature descriptors from an image, quantise the descriptors to a vocabulary of visual words, and use the histogram of observed words as an image descriptor (Nister and Stewenius, 2006; Yang et al., 2007; Galvez-López and Tardós, 2012; Radenovic et al., 2016). The advantage of BoVW is that it is unaffected by positions and orientations of objects in an image. BoVW-based search can quickly return the possible locations of the query image, avoiding searching through tens of millions of local feature descriptors.

3. METHODOLOGY

3.1 Overview

The proposed indoor VBL approach manages to address the following three problems: (1) how to record the spatial information with sufficient accuracy and limited storage; (2) how to find the image correspondences between the query images and the database images under different conditions with respect to resolution and illumination; and (3) how to calculate poses of query images and handle outliers. The workflow of the approach consists of **a keyframe-based global mapping phase** and **a visual SLAM-based local positioning phase**. An illustration of the approach is shown in Fig. 1.

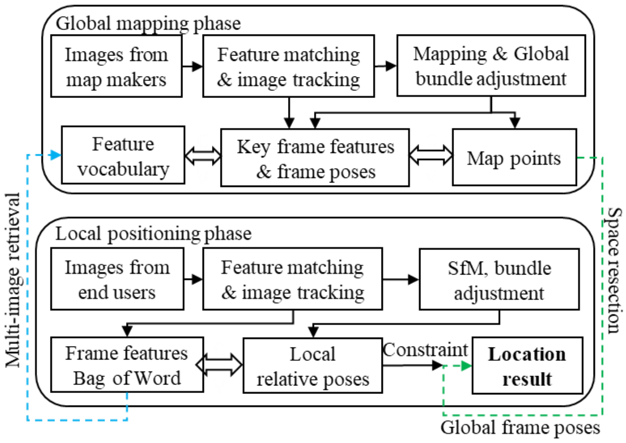


Figure 1. Workflow of our indoor localization approach.

3.2 Global mapping

In order to prepare the database, a manual operation is carried out first to map the interior of the building using a camera. The indoor scene is explored by a keyframe-based global mapping technique. Input data are a sequence of images provided by the operator who traverses the entire building. The 3D map representation is built based on a SfM mapping module. We use the ORB (oriented FAST and rotated BRIEF) feature as the feature tracking strategy (Michael et al., 2010; Rublee et al., 2011), which extracts FAST corners in 4 image resolution levels of each frame, and describe the corners by BRIEF descriptors. It is powerful to establish 2D-3D matches between image features and map points.

The selection of keyframes is a self-adaptive process, which culls the redundant images and at the same time ensures that the adjacent keyframes have enough overlapping regions and each keyframe contains sufficient feature correspondences. This corresponds to a self-adaptive recording process which stores the keyframes based on a dynamic selection strategy instead of using a fixed internal distance. There are three criteria for selecting keyframes: (1) the frame shares at least 40 points with other images; (2) the current frame shares less than 80% of its features with previous key frames; and (3) the frame interval between the current frame and previous keyframes is larger than a threshold, which we set equal to frame frequency. Compared with the manual selection of images in some typical SfM workflows, this dynamic strategy has an advantage of automatically generating a dataset that is less constrained by the experience of the person taking the images. Another advantage is that keyframe-based databases have compact image representations, which decreases the memory requirements and increases the search efficiency compared to databases containing an unfiltered set of images.

The initial scale of map is determined by the first two keyframes which consist a normalized baseline and generate an initial map by stereo matching and triangulation. The map is incrementally built by selecting new keyframes and calculating new map points. Finally, a global bundle adjustment is used to refine the structure of the map. In order to obtain the absolute map positions corresponding to the physical world, a coordinate transformation might be conducted using transformation parameters estimated by interactive aligning some predefined control points in the world coordinate system and the map points.

After the exploration, each keyframe in the database is associated with its exterior orientation parameters and some map points corresponding to the features in the frame. Each map point is associated with the ORB descriptors found in the images for feature matching. A feature vocabulary tree is trained using the feature vectors of all frames. The nodes in the vocabulary tree are visual feature words, which can speed up the following image retrieval process. In return, we cast ORB features of keyframes into a BoVW based on the vocabulary tree. In sum, the spatial structures, i.e. 3D map points, of the scene and the poses of keyframes are stored in the database.

3.3 Local Positioning

In the querying phase, instead of querying by a single image, a short sequence of images of the site is used to extract both features and the relative image poses, which is a local SfM procedure.

Following a workflow similar to the Parallel Tracking and Mapping (PTAM) (Klein and Murray, 2007) and ORB-SLAM (Mur-Artal and Tardós, 2017) methods, we initialize the local 3D model by a stereo matching method which takes two viewpoints different enough to guarantee a certain baseline and similar enough to share enough feature correspondences. In our implementation of feature matching, we use ORB (Rublee et al., 2011) for detection and matching. We use a simple constant velocity motion model combined with feature tracking method to predict the camera poses. After that, every subsequent tracking image $\boldsymbol{I}$ is associated with a local relative pose $\boldsymbol{P}_{\boldsymbol{I}}^{\boldsymbol{L}}$.

In the query phase, we predict orientation parameters for every frame. We cast the features of every query frame into a BoVW based on the vocabulary generated in the mapping phase. Based on the BoVW, for every query image we find corresponding candidate keyframes in the image database via searching through the vocabulary tree. For instance, given an image sequence containing k frames $\{\boldsymbol{I}_1, \boldsymbol{I}_2, ..., \boldsymbol{I}_k\}$, we conduct image retrieval for every frame, which outputs the following ranking correspondence list:

$$\begin{array}{l} \boldsymbol{Query}(\boldsymbol{I}_1) = \{\boldsymbol{J}_1^1, \boldsymbol{J}_2^1, ..., \boldsymbol{J}_m^1\} \\ ... \qquad\qquad\qquad\qquad\qquad , \\ \boldsymbol{Query}(\boldsymbol{I}_k) = \{\boldsymbol{J}_1^k, ..., \boldsymbol{J}_n^k\} \end{array}$$

where $\boldsymbol{J}_{\boldsymbol{n}}^{\boldsymbol{k}}$ is the image at the n^{th} rank returned for the k^{th} query frame $\boldsymbol{I}_{\boldsymbol{k}}$. The order in the returning set is based on the BoVW histogram matching scores.

This search is a 1-to-n problem where n may be 0, 1 or more than 1. If $\boldsymbol{n} \neq 0$, the search has successfully found at least one candidate image for the query frame in the database. Then, we match ORB features between each query image and its candidate correspondences in the database. As the features of the database keyframes are linked to map points, after matching we obtain some correspondences between map points and feature points of the query images. If there are more than 15 correspondences between ORB features and map points, we continue to calculate the exterior orientation of a query image by RANSAC, using the direct method of (Gao et al., 2003) in each iteration. The focus length is estimated from the image EXIF file, and the principal point is assumed to be at the centre of the frame. Subsequently, we refine the pose by least squares estimation using all inliers. The localization method returns an ordered list of candidate poses for each successful query image; we present these candidate global poses as

$$\left\{\left\{\boldsymbol{P}_{1,1}^{\boldsymbol{G}}, \boldsymbol{P}_{1,2}^{\boldsymbol{G}}, ..., \boldsymbol{P}_{1,k_1}^{\boldsymbol{G}}\right\}, \left\{\boldsymbol{P}_{2,1}^{\boldsymbol{G}}, ..., \boldsymbol{P}_{2,k_2}^{\boldsymbol{G}}\right\}, ..., \left\{\boldsymbol{P}_{m,1}^{\boldsymbol{G}}, ..., \boldsymbol{P}_{m,k_m}^{\boldsymbol{G}}\right\}\right\}, \qquad (1)$$

where $\boldsymbol{P}_{\boldsymbol{m},\boldsymbol{k_m}}^{\boldsymbol{G}}$ means the $k_m{}^{th}$ global pose of query image $\boldsymbol{I}_{\boldsymbol{m}}$.

The local relative poses of images with candidate poses are used to generate a local pose graph, whose nodes are defined at the coordinates of the frames and the weight of edge is related to the Euclidean distance. These images take a similar role as keyframes, but they are selected according to the image retrieval results and not by applying rules related to extracted features or intersection geometry. As there is only a short sequence of query images, we assume the related poses can correctly reflect the geometric relationship of frames. We use $\boldsymbol{P}_1^{\boldsymbol{L}}, \boldsymbol{P}_2^{\boldsymbol{L}}, ..., \boldsymbol{P}_{\boldsymbol{m}}^{\boldsymbol{L}}$ to denote the local poses.

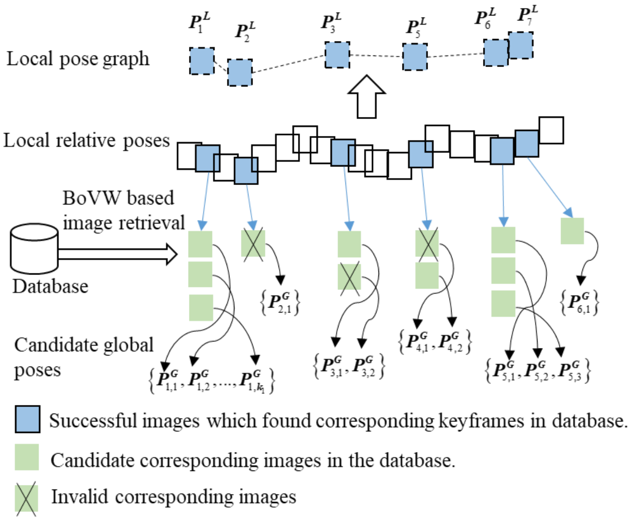

P_1^L
P_2^L
P_3^L
P_5^L
P_6^L
P_7^L
Local pose graph
Local relative poses
BoVW based image retrieval
Database
Candidate global poses
$\{P_{1,1}^G, P_{1,2}^G, ..., P_{1,k_1}^G\}$
$\{P_{2,1}^G\}$
$\{P_{3,1}^G, P_{3,2}^G\}$
$\{P_{4,1}^G, P_{4,2}^G\}$
$\{P_{5,1}^G, P_{5,2}^G, P_{5,3}^G\}$
$\{P_{6,1}^G\}$
Successful images which found corresponding keyframes in database.
Candidate corresponding images in the database.
Invalid corresponding images

Figure 2. An illustration of local pose graph and global candidate poses after image retrieval process. After local SLAM, "local relative poses" are recovered, and then we retrieve images from database to get corresponding keyframes. The local poses of successful query images integrate "local pose graph". At the same time, we calculate the "candidate global poses"

based on the connection between features in keyframes and map points.

For the successful query images with correspondences we have two sets of poses, i.e. the local relative poses obtained from SfM and global poses obtained via spatial resection. An illustration of the successful query images (blue boxes) and their corresponding image retrieval results (green boxes) is shown in Fig. 2. $\boldsymbol{P}_m^L$ and $\{\boldsymbol{P}_{m,k_m}^G\}$ denote the local poses and candidate global poses respectively. $\boldsymbol{P}_1^L$-$\boldsymbol{P}_2^L$-...-$\boldsymbol{P}_m^L$ represent a local pose graph. This information is used to eliminate wrong correspondences and, consequently, wrong global poses based on a geometric criterion.

3.4 Removal of outliers based on geometrical constraints

Using BoVW for image retrieval can result in some false correspondences (green boxes with forks shown in Fig. 2), which may lead to incorrect global candidate poses. We will use the information about the local poses to derive geometric constraints to eliminate these outliers.

The local and global poses of a frame are defined in the local and global coordinate systems, respectively. Therefore, we estimate the seven parameters of a spatial similarity transformation (three shifts, three rotations and a scale) using the projection centres of the keyframes as identical points in a least squares estimation:

$$\min \sum_{i} \sum_{j} \left\| \delta_{i,j} \left(\boldsymbol{P}_{i,j}^G - \lambda \boldsymbol{R}^{LG} (\boldsymbol{P}_i^L - \boldsymbol{t}^{LG}) \right) \right\|, \tag{2}$$

where λ is the scale factor between the local and the global coordinate systems; $\boldsymbol{R}^{LG}$ and $\boldsymbol{t}^{LG}$ are the rotation and translation parameters between two set of frame positions. $\delta_{i,j} \in \{0,1\}$ is an indicator term used to determine whether the global pose is valid in the function. The unknowns in this estimation procedure are λ, $\boldsymbol{t}^{LG}$ and the three angles parameterizing $\boldsymbol{R}^{LG}$.

A crucial task is to determine the value of $\delta_{i,j}$, i.e. to find out which global poses are outliers. We design a **distance ratio code (DRC)** that records a set of distance ratios between poses. Firstly, the local pose graph provides a **reference DRC**. If the pose graph consists of $\boldsymbol{P}_1^L, \boldsymbol{P}_2^L, ..., \boldsymbol{P}_m^L$, its DRC is $\left[\boldsymbol{r}_{1,2}^L(=1), \boldsymbol{r}_{2,3}^L, ..., \boldsymbol{r}_{m-1,m}^L\right]$, where each value *r* is the ratio of the relative distances of a pose to the first two positions:

$$\boldsymbol{r}_{p,q}^L = \left| \frac{\boldsymbol{P}_p^L - \boldsymbol{P}_q^L}{\boldsymbol{P}_2^L - \boldsymbol{P}_1^L} \right|. \tag{3}$$

As every query frame may have multiple candidate poses, there exists a corresponding set of DRCs. The first process to detect errors in the global localization is based on a global pose distance constraint. We assume that our query images are captured within a small range not exceeding 2 meters. Hence, if the Euclidean distance between two neighbouring global candidate poses $\boldsymbol{P}_p^G$ and $\boldsymbol{P}_q^G$ is larger than 2 meters, we think this edge length between $\boldsymbol{P}_p^G$ and $\boldsymbol{P}_q^G$ is exceed the scope of the query images. So, their relative distance ratio is omitted, i.e. $\delta_{i,j}=0$. Secondly, we calculate the correlation coefficient between the reference DRC and global DRCs. Only if the correlation score is higher than a threshold, $\delta_{i,j}$ of the global pose is set to 1. By solving (2), we can transform the local poses of successful query images to the global mapping coordinate frame and get the absolute poses for successful query images.

Thus, by constraining the pose distribution using the information about local poses, we can detect and remove outliers in the global poses, making the strategy robust to erroneous image retrieval results. Because our interior parameters of query images are roughly estimated, the accuracy of positioning results is lower than decimetre level. Errors in the range of some decimetres in the local SfM procedure can be tolerated because the correct result of image retrieval is the key to the localization task. After detecting and eliminating wrong image retrieval results, we use all the retrieved and query images that are left to get the correct feature correspondences and corresponding map points. Using these correspondences, we estimate the exact camera location and orientation information for query images by bundle adjustment.

4. EXPERIMENTS

4.1 Mapping

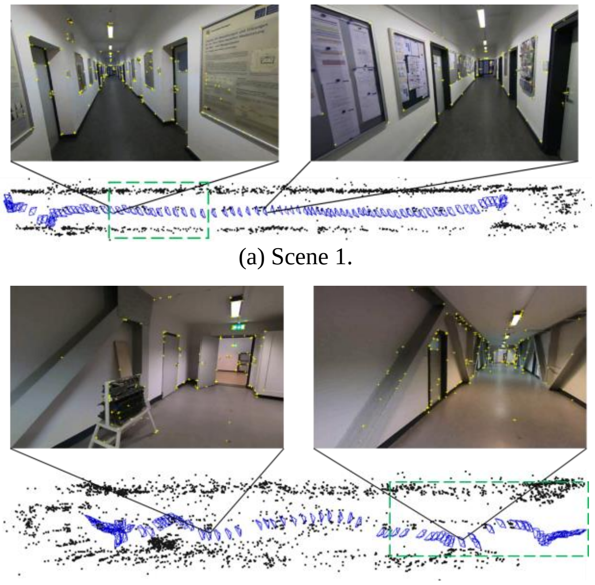
To test the proposed indoor VBL system, we scanned two corridors of a building using a Gropro sport camera with a fixed frame size of 1920×1440 pixels. The camera was calibrated. After the global mapping process, all keyframes and 3D map points are stored in memory as the database. In this experiment, the coordinate framework of a map is defined to coincide with the image coordinate system of the first keyframe, so we did not use ground control points to transform the coordinates. We only adjust the map scale. The physical scale of the scenes was derived by measuring the widths of the corridors, and the map was scaled accordingly.

Table 1 gives the statistics for these two scenes. These two videos contain more than 2000 frames each and have a frame rate of 24 frame/sec. Two scenes span approximately 25 m and 23 m respectively and there are many keyframes and map points evenly distributed. A visualization of the database is shown in Fig. 3. In each scene, the reconstructed database, including map points (**black points**) and poses of keyframes (**blue wireframes**), are shown at the bottom, and two arbitrary keyframes are displayed at the top for intuitive visualization. In the keyframes in Fig. 3, the yellow dots indicate the image features been used to calculte the map points. SfM techniques provide a relative pose of a sequence of data, then the VBL system tackles the problem of retrieving the absolute pose of a query data according to the known representation of scenes.

	Video frames	Rate (FPS)	No. of Keyframe	Map points	Span distance
Scene 1	2952	24	101	3967	25 m
Scene 2	2496	24	69	3271	23 m

Table 1. Statistics of the test scenes.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W13-827-2019



(a) Scene 1.

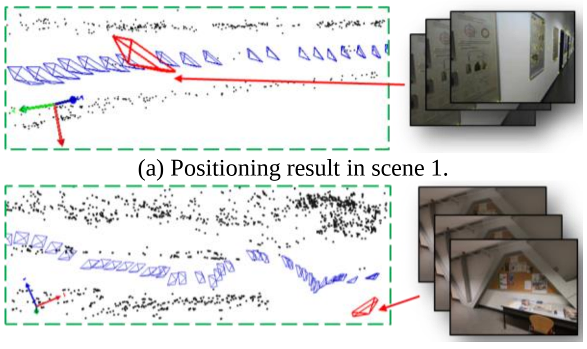
(b) Scene 2.

Figure 3. Global mapping results for the two test scenes (bottom) with sample images from the sequences (top).

4.2 Positioning

Using the proposed localization framework, we tried to determine the location of query images in each scene. The query images are obtained by a camera of a smartphone with a frame size of 1920 × 1080 pixels. The internal parameters of the camera are estimated from image EXIF file. Taking advantage of efficiency of the BoVW feature matching strategy, for query image sequences we found their corresponding keyframes in the database. The localization results corresponding to Fig. 4 is shown in Fig. 4, where the red wireframe indicates the position of the querying site in the whole scene. The region of Fig. 3 corresponds to the green dotted wireframe region in Fig. 3.

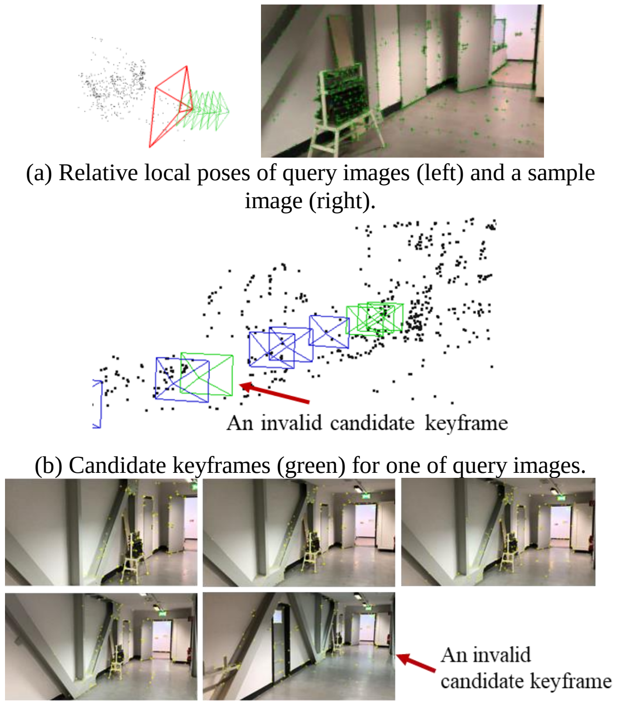
We can convert the local relative poses into absolute poses by determining the scale between two kinds of distances between adjacent frames. The poses recovered from local SLAM provide a geometric constraint, because when there are enough successful query frames (at least 3), the outliers in global poses can be detected as described in Section 3.3. In practice, the first verified global image pose in the query site is chosen as the output result.



(a) Positioning result in scene 1.

(b) Positioning result in scene 2.

Figure 4. Examples of positioning results.



(a) Relative local poses of query images (left) and a sample image (right).

(b) Candidate keyframes (green) for one of query images.

(c) Images of the keyframes shown (b).

Figure 5. The effect of relative poses on detecting invalid image retrieval result.

In Fig. 5, we shown an illustration of the method that relative local poses can detect invalid extracted candidate keyframes. In Fig. 5 (a), the red frame represents a current query image, and the green images are a few successful query images. The green dots in the right image denote the feature points detected in a query image. Fig. 5 (b) depicts an example of extracted candidate keyframes (green frames) from the database corresponding to one of query images. Fig. 5 (c) shows the images corresponding to these keyframes in Fig. 5 (b). The red arrows in (b) and (c) point out a false candidate keyframe which is extracted because it share a small part of scene with the query image. Using the method described in Sec. 3.4 to deal with the poses in (a) and (b), we can detect the invalid keyframe and yield correct positioning result.

4.3 Limitation

Image based localization has a common drawback: the confusion caused by similar decorations inside a building. A lack of features or a highly occluded environment can also reduce the success rate of the approach. Besides, the usage of BoVW still requires further testing for large changes in scale and viewpoint. In the future, a compensation solution could be to use Wi-Fi based finger printing techniques to reduce the search range.

5. CONCLUSION

We propose a vision-based localization system in an indoor environment by exploiting a SfM algorithm. The proposed method takes advantage of a visual vocabulary, under BoVW framework to exploit connections between physical locations and feature clusters. As the query phase only needs a small part of the scene, we assume the local pose graph provides a reliable shape, so the image retrieval results can explicitly handle the structure of global poses. The method is a very suitable option for low cost camera-based indoor positioning and navigation. Experimental results show that this method has a high potential for applications.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant: 41801342 and the Natural Science Foundation of Jiangsu Province, China, under Grant: BK20170781.

REFERENCES

Bay H., Ess A., Tuytelaars T., Luc Van Gool, 2008. SURF: Speeded up robust features, *Computer Vision and Image Understanding*, 110(3), 346–359.

Calonder M., Lepetit V., Strecha C., and Fua P., 2010, BRIEF: Binary Robust Independent Elementary Features, In: *11th European Conference on Computer Vision*, 778–792.

Chen D.M., Baatz G., Koser K., Tsai S.S., Vedantham R., Pylvanainen T., Roimela K., Chen X., Bach J., Pollefeys M., Girod B., Grzeszczuk R., 2011. City-scale landmark identication on mobile devices. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 737–744.

D. Gálvez-López and J. D. Tard´os, Bags of binary words for fast place recognition in image sequences, *IEEE Trans. Robot.*, vol. 28, no. 5, pp.1188–1197, 2012.

Eade E., Drummond T., 2008. Unified loop closing and recovery for real time monocular SLAM. In: *Proceedings British Machine Vision Conference*. 1–10

Farid Z., Nordin R., Ismail M., 2013. Recent advances in wireless indoor localization techniques and system, *Journal of Computer Networks and Communications*, Vol. 2013, Article ID 185138.

Galvez-López D., Tardós J.D. 2012. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5), 1188–1197.

Gao X.S., Hou X.R., Tang J., Cheng H.F., 2003. Complete solution classification for the perspective-three-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8), 930–943.

Huang Y., Wang H., Zhan K., Zhao J., Gui P., Feng T., 2015. Image-based localization for indoor environment using mobile phone. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume XL-4/W5, 211–215.

Klein G., Murray D., 2007. Parallel Tracking and Mapping for Small AR Workspaces. In: *6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, 225–234.

Kendall, A., Grimes, M., Cipolla, R., 2015. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In: *IEEE International Conference on Computer Vision*, 2938–2946.

Li Y., Snavely N., Huttenlocher D., and Fua P., 2012. Worldwide pose estimation using 3d point clouds. In: *European Conference on Computer Vision*, 2012, 15–29.

Lim H., Sinha S.N., Cohen M.F., Uyttendaele M., 2012. Real-time image-based 6-DOF localization in large-scale environments. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 1043–1050.

Lindeberg T., 1998. Feature detection with automatic scale selection, *International Journal of Computer Vision*, 30(2), 77–116.

Lowe D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, 60(2), 91–110.

Möller A., Kranz M., Huitl R., Diewald S., Roalter L., 2012. A mobile indoor navigation system interface adapted to vision-based localization, In: *ACM Proceedings of the 11th International Conferenceon Mobile and Ubiquitous Multimedia*, p. 4.

Mur-Artal R. and J. Tardós, 2017. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5), 1255–1262.

Nister D., and Stewenius H., 2006. Scalable recognition with a vocabulary tree. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2161–2168.

Radenovic F., Tolias G., Chum O., 2016. CNN Image Retrieval Learns from BoW: Unsupervised fine-tuning with hard examples. In: *Proceedings of the IEEE European Conference on Computer Vision*. Vol. 9905. pp. 3–20.

Ravi N., Shankar P., Frankel A., Elgammal A., and Iftode L., 2005. Indoor localization using camera phones, In: *7th IEEE Workshop on Mobile Computing Systems & Applications*, 1-7.

Robertson D.P. and Cipolla R., 2004. An image-based system for urban navigation. In: *Proceedings British Machine Vision Conference*, 819–828.

Rosten E. and Drummond T., 2006. Machine learning for high speed corner detection. In: *European Conference on Computer Vision*, volume 1, 430–443.

Rosten E., Porter R., and Drummond T., 2010. Faster and better: A machine learning approach to corner detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32, 105–119.

Rublee E., Rabaud V., Konolige K., Bradski G., 2011. ORB: an efficient alternative to SIFT or SURF. In: *IEEE International Conference on Computer Vision*, 2564–2571.

Sánchez-Rodríguez, D., Alonso-González, I., Sánchez-Medina, J., Ley-Bosch, C., and Díaz-Vilariño, L., 2017. Performance analysis of classification methods for indoor localization in VLC networks, *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, IV-2/W4, 385–391.

Wan, X., Liu, J., Yan, H., Morgan, G. L. K., 2016. Illumination-invariant image matching for autonomous UAV localisation based on optical sensing. ISPRS Journal of Photogrammetry and Remote Sensing 119, 198–213.

Xiao J.X., Chen, J.N., Yeung, D.Y., Quan L., 2008. Structuring visual words in 3D for arbitrary-view object localization. In: *European Conference on Computer Vision*, 725–737.

Yang J., Jiang Y.G., Hauptmann A.G., Ngo C.W., 2007. Evaluating bag-of-visual-words representations in scene classification, In: P*roceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval*, MIR 2007, 197–206.

Zamir A.R. and Shah M.， 2010. Accurate image localization based on google maps street view. In: *European Conference on Computer Vision*, 2010, 255–268.

https://doi.org/10.5194/isprs-archives-XLII-2-W13-827-2019