A sequence-based visual place recognition method for aerial mobile robots

Minzhou Yang, Jun Mao*, Xiaofeng He, Lilian Zhang, Xiaoping Hu

College of Intelligence Science and Technology, National University of Defense Technology, Changsha, Hunan, 410073, China

*Corresponding author’s e-mail: maojun12@nudt.edu.cn

Abstract. Recent developments in sequence-based visual place recognition (VPR) have shown robust localization capability on ground vehicles; however, deploying existing sequence-based VPR techniques on aerial mobile robots is still challenging, as topologically ordered database is not available for aerial agents. In this paper, we develop a new sequence-based VPR algorithm, which is applicable to aerial mobile robots. The proposed method models the agent’s position as a probability function conditioned on all the visual observations and accumulates the position belief as the observation increases, which helps to filter out the inconsistent false recognitions. To testify the effectiveness of the proposed method, we have conducted series of experiments under different visual environments. The experimental results show that, by using a remote sensing image as reference and onboard downward-looking camera images as observations, the proposed method can robustly estimate the position of the agent and have a localization rate 30% higher than the single-image-based VPR algorithm.

1. Introduction

Visual place recognition (VPR) [1-2] determines whether a query image is taken at a known place in the database; it is widely used in mobile robotics for autonomous positioning. As place appearance changes, VPR algorithms must be capable of recognizing places under variations caused by viewpoint, weather and other factors. To achieve condition invariant VPR, recently techniques have used sequence images to filter out spatial inconsistent recognitions, and have shown great improvements over the single-image-based methods. However, most existing sequence-based VPR algorithms are designed for ground vehicles, where the reference database images have been ordered as a topological map according to historical trajectories’ topology [3-5] or road network topology [9-10,17]. For aerial mobile robots, their flight trajectories can be very different from their historical trajectories, which brings in difficulties for aligning the query image sequence with the reference database image sequence. Furthermore, current sequence-based VPR algorithms focus on using local view images as queries and references, but in aerial applications, using a remote sensing map, such as satellite map, as references is more efficient.

In this paper, we develop a novel sequence-based VPR algorithm for aerial robots. Unlike the ground sequential VPR algorithms, which need a ordered local view topological map as reference, the proposed method uses a remote sensing map as reference, which can be easily obtained from Google Earth or other map servers. The position of the agent is modelled as probability function conditioned on all the observations; by using Bayes estimation, we can numerically estimate the probability function and thus obtaining the agent’s position. We have conducted systematic experiments in
different visual environments, the results show the proposed sequence-based visual place recognition method significantly augmented the aerial mobile robot’s place recognition capability.

Figure 1. The query sequence images were collected in blue spots from the path shown in red. The reference sequence images were collected from the areas shown in black panes on the red path. The VPR system can match current query sequence images at topological map established through reference sequence images to locate the robot.

Figure 2. The reference images were collected beyond the red path. Note that the reference images cannot be ordered in sequence as a topological map according to historical trajectories’ topology or road network topology. It’s difficult for aerial robots to align the query images sequence with the reference database images.

2. Related work
Sequence-based VPR has been received great attentions in recent years. In 2008, Cummins proposed an appearance-only based VPR algorithm named as FAB-MAP, which has shown large scale self-localization capability; later in 2011, they included place topological connections to estimate location priors in FAB-MAP, and have shown improvements in algorithm efficiency and precision. However, FAB-MAP’s place recognition precision can reduce dramatically in severe appearance variation conditions.

One of the most famous sequence-based VPR algorithms is SeqSLAM [7-8], which has shown robust localization performance under severe appearance changes between sunny day and rainy night. However, the original SeqSLAM requires the constant velocity of vehicle and exhaustive sequence match searching through multiple locations and image scales. In the follow-up research, Milford
implements a VPR algorithm named SMART[10], which improves the general applicability of SeqSLAM by integrating self-motion information to form spatially consistent sequences, and new image matching techniques to handle greater perceptual change and variations in translational pose. However, the system has lower performance in areas with significant self-similarity, such as long stretches of road or track with common distal features. Furthermore, they present a condition-invariant, sequence-based algorithm [9] to solve two key road navigation challenges: intersections and multiple-lane roads, thus progressing towards a complete navigation system. In addition, they also make the algorithm automatically adjust the sequence length to adapt to the scene characteristics of different regions [11]. Nasser apply network flow to perform sequence-based VPR, which can extend the algorithm to adapt to different road network topology [12]. ABLE method [13] was proposed to deal with the long-term VPR across seasons, which can achieve more than million images match.

However, the conventional VPR based on sequences image, which is only applicable when the observed places have exactly the same order as the database reference places, cannot be applied to aerial environment without road constraint. Hence, we propose a VPR based on sequences image applied for aerial robots and describe how to create a suitable database without topological map. Hence, we propose a sequence-based VPR without using a topological map and applicable to aerial robotics.

3. Problem setting
We first describe our adopted setting for VPR for aerial robots. Let \( R = \{r_1, r_2, r_3, \ldots, r_n\} \) be a reference database of \( n \) places in the designated area. In the proposed VPR system, \( R \) is collected in distributed mode to cover all features of the specified area using aerial robots instrumented with cameras. The query video is represented as \( Q = \{q_1, q_2, q_3, \ldots, q_t\} \), which is a temporally-ordered sequence of \( t \) query scenes.

For each \( q_i \in Q(i = 1, 2, \ldots, t) \), the goal of VPR system is to retrieve an image from \( R \) that was taken from the similar location to \( q_i \), i.e., the field of view in the retrieved image overlaps to a large degree with \( q_i \). Note that the reference images in \( R \) for aerial vehicles aren’t ordered as sequence compared with that for ground vehicles, we design the probability function conditioned on all the observations instead of temporally-ordered sequence for position of agent at each \( q_i \). For each \( r_j \in R(j = 1, 2, \ldots, n) \), we do not need to build a topological map in \( R \). Using constraint of Bayes estimation, we can obtain the best match between \( q_i \) and \( r_j \).

4. Sequential visual place recognition based on Bayes estimation
As place appearance changes due to factors such as season, illumination and view point, only using appearance to perform VPR could include false positive matches, which brings in catastrophic localization information to autonomous systems. Note that the agent’s true positions are related to the observed visual features, therefore we can describe probability of being at a reference location as the following probability density function:

\[
\text{Bel}(L^k) = p(L_j \mid z_{t,k}) \quad (1)
\]

where \( L_i(x, y) \) represents the location label of the \( i^{th} \) reference landmark, \( z_{t,k} \) is all the camera images captured along the way. and \( \text{Bel}(L^k) \) represents the probability that agent is located in \( L_i \) at the \( k^{th} \) sample time. Through the Bayes estimation, Eq.(1) can be rewritten as:

\[
\text{Bel}(L^k_i) = \eta p(z_k \mid L_i, z_{t,k-1}) p(L_i \mid z_{t,k-1}) \quad (2)
\]
where $\eta$ is the normalization term. The first term at the right of Eq.(2) can be simplified by the first-order Markov model as:

$$p(z_k \mid L, z_{ik-1}) \approx p(z_k \mid L)$$

which means the current captured image $z_k$ is only determined by the current position $L$. Eq.(3) can be numerically calculated as the similarity of the visual features between the query image $z_k$ and the reference image $i^{th}$ as:

$$p(z_k \mid L) = \exp(-\|f_i - f_k\|)$$

where $f_i$ represents the feature descriptor of the $i^{th}$ reference image, and $f_k$ represents is the feature descriptor of the current query image.

The term $p(L_i \mid z_{ik-1})$ in Eq.(2) is the location prior, which uses all the historical observations to predict the probability of the agent locating at $L_i$. It can be written as

$$p(L_i \mid z_{ik-1}) = p(L_i \mid L_{j-i}) \sum_{N} p(L_{j-i} \mid z_{ik-1}) = p(L_i \mid L_{j-i}) \sum_{N} Bel(L_{j-i})$$

where $p(L_{j-i} \mid z_{ik-1})$ is the probability of being at $L_i$ at the previous sample time $k-1$; $p(L_i \mid L_{j-i})$ represents the state transition model, which characterizes the probability of agent travels from $L_j$ to $L_i$ during a sample time interval. As the motion range of the agent is small in a short period of time, we approximately calculate Eq.(5) as the maximum probability of the nearby reference images as:

$$p(L_i \mid z_{ik-1}) = \arg \max_{j \in h(i)} (Bel(L_{j-i}))$$

where $h(i)$ is the set of reference image indices adjacent to $L_i$, and can be expressed as:

$$h(i) = \{j_1, j_2, \cdots, st. \mid L_i - L_{j} \mid \leq r \}

(7)

$$

to select the best matched location, we apply a kernel density estimation on a neighbourhood around each reference image’s position as in [9,16] as:

$$p(L_i) = \sum_{j \in h(i)} Bel(L_j)$$

If $p(L_i)$ is greater than the set threshold, it is considered that the VPR is successful, and the current position of the agent is then calibrated to $L_i$.

Note that, the proposed method only uses image position labels to perform sequential VPR, while most current algorithms need to construct a topological map to align query and reference image sequence.

5. Experiments and results

This part presents the experimental setup and the experimental results. The experimental setup details the database construction and the flight simulation procedures, while the experimental results shows the VPR precision-recall curve and analyses to the results.
5.1. Experimental Setup

5.1.1. Database Construction.
Two datasets were constructed for the experiments, one is the Mountain dataset and the other is the Coast dataset. The Mountain dataset was captured at a hilly region as shown in Figure 3(a) and the Coast dataset was captured at a coastline as shown in Figure 3(b). Two remote sensing images downloaded from the NearMap [17] were used in each dataset, one was used to generate the database images and the other was used to generate the camera query images. The capture time of the two images in each dataset has more than 3 months’ time separation to stimulate the appearance variations in different time. The resolution of the remote sensing image of the Coast database is $\sigma = 4.77\text{m/px}$, and the Mountain database is $\sigma = 0.55\text{m/px}$.

To construct the reference image database, we cut reference remote sensing image with 32 $\text{px}$ stride and in 128$\text{px} \times 128\text{px}$ size to generate reference images database; all the reference images were ordered according to their spatial locations, other than a specified road topology or a travelled path as in the ground applications. There are 6882 images in the Coast database and 8658 images in Mountain database.

5.1.2. Flight Simulations.
In flight simulations experiment, we tested the proposed VPR algorithm on 20 navigation tasks in each dataset. As is shown in Table 2, the agent was set to fly along the planned path at a height of 64$\sigma$ m above the ground and at a speed of 20 $\sigma$ m/s along the heading direction. The onboard virtual camera was defined to have a view field of 90° and the output image was at the size of 128$\text{px} \times 128\text{px}$, with the optical axis at the middle of the image; the camera was set to take images at a frequency of 1 Hz.

The whole voyage was 68.323 $\sigma$ km and 3880 query images were collected on this process in each dataset in total, where $\sigma$ is the corresponding remote image’s resolution.

To simulate the real-world observation changes, we added viewpoint, position and scale variations to the reference image, as shown in Table 1. In detail, the Coast query images have a viewpoint variation of 10° (anticlockwise), and the Mountain query images have a viewpoint variation of 20° to the reference images. The observation position variation (represented in the remote map) along the east and south were set as a zero-mean Gaussian distribution. The Mountain query images also have a zero mean scale variation, to simulate the observation height difference. Figure 4 demonstrates the variations between the query and reference images.
Table 1. The details of variations added to the query image dataset

| Databases | Viewpoint Difference | Position Difference | Scale Difference |
|-----------|----------------------|---------------------|-----------------|
| Mountain  | 20°                  | $x \sim \mathcal{N}(0,10\text{pix})$; $y \sim \mathcal{N}(0,10\text{pix})$ | $s \sim \mathcal{N}(0,0.1)$ |
| Coast     | 10°                  | $x \sim \mathcal{N}(0.5\text{pix})$; $y \sim \mathcal{N}(0.5\text{pix})$ | \ |

Table 2. Parameter list

| Parameter | Value       | Description                       |
|-----------|-------------|-----------------------------------|
| $R_x$, $R_y$ | 128pix, 128pix | Image size                       |
| $\theta$   | 90°         | Camera view field                 |
| $H$        | 64σm        | Fly height                        |
| $V$        | 20σm/s      | Fly speed                         |
| $f$        | 1Hz         | Working frequency                 |
| $r$        | 64σ         | Radius of neighbourhood in Eq.(6) |

Figure 4. Example frame match from reference images and query images generated through above method. The details of variations in Table 1 are carried out to simulate the changes of query images, so that the algorithm has universal applicability.

5.1.3. Image Processing.

The VLAD feature [14] and the Gist feature [15] were used to describe the images. For VLAD descriptors, we used k-means algorithm to learn a dictionary of 512 visual words from 128-dimensional SIFT features; then we aggregated these visual words descriptors into a 65536-dimensional (512×128) vector representation for each image; finally, a dimensionality reduction of these vectors was made with the principal component analysis (PCA) to obtain 256-dimensional VLAD feature. For Gist descriptors, we first applied Gabor filter banks with 8 scales and 4 directions to process the database image and obtained 32 (4×8) feature images as size as the database images; then each feature image was divided into 4×4 regions; finally, 512-dimensional(32×4×4) Gist vector was obtained by combining the regions of all feature images.
5.2. Experimental results

We present the results in terms of precision-recall curve, which has been widely used to evaluate VPR algorithms [8-10]. Correct localizations are defined as the recognized places are within $64\sigma$ meters to the query image’s true positions. In order to analyse and verify the performance of our algorithms, a single-image-based VPR algorithm that uses nearest distance matching has been used for comparisons.

![Figure 5. Precision-recall curves](image)

Figure 5(a) shows the precision-recall curve for the Mountain dataset. For proposed sequence-based VPR with Gist feature, a precision rate of 90% is achieved at a recall rate of 75%, after which the precision drops steadily to 73% at the recall rate of 100%. The single-image-based VPR has much lower precision at the same recall rate, e.g. only has 52% precision at 100% recall. For proposed sequence-based VPR with VLAD feature, the precision has 45% at the recall rate of 100%, while the single-image-based VPR has much lower precision at the same recall rate, e.g. only has 22% precision at 100% recall. Figure 5(b) shows the precision-recall curve for the Coast dataset. The sequence-based VPR with Gist feature achieved a recall rate of 40% at 100% precision, after which precision dropped steadily to 68% at the recall rate of 100%. However, the single-image-based VPR match has 65% precision at 100% recall for Gist feature. For proposed sequence-based VPR with VLAD feature, the precision has 56% at the recall rate of 100%, while the single-image-based VPR has much lower precision at the same recall rate, e.g. only has 38% precision at 100% recall.

In conclusion, the proposed sequence-based VPR has shown significant visual place recognition capability improvements in aerial mobile robots. Though the vehicle’s navigation trajectories have never been explored, our method can align the observed images to the unordered reference places extracted from a remote sensing map and accumulates the localization belief sequentially.

Figure 6 shows the ground truth plots for the Coast databases at a recall rate of 100% by using the VLAD feature. As shown in the figure, the false positives of the sequence-based VPR are mostly located in the water regions, where salient visual features are not available; In comparison, the single-image-based VPR not only has a large number of false positives at the water region, but also has a significant number of false positives at the Mountain regions.
6. Discussion and future work

The paper presents a sequence-based VPR algorithm for aerial mobile robots. Unlike the ground applications, where a topologically ordered local view image database was required, the proposed method uses a remote sensing image as reference; this greatly improves the availability of the reference image database. Furthermore, the proposed method uses Bayes estimation to perform sequence visual place recognition and does not require the query image sequences have the same order as the database images. Experimental results in different visual environments have shown the effectiveness of the proposed method.

Currently, we only conducted experiments with small image scale variations. However, when aerial robots fly at different altitudes, the query image’s scales can be very different to the fixed scale reference images. In the future, we can construct a multiscale database, which can be act as reference to aerial vehicles fly at different altitudes.

References

[1] Garcia-Fidalgo, E., Ortiz, A. (2015) Vision-based topological mapping and localization methods: A survey. Robotics and Autonomous Systems, 64: 1-20.
[2] Lowry, S., Sunderhauf, N., Newman, P., etc. (2015) Visual Place Recognition: A Survey. IEEE Transactions on Robotics, 1-19.
[3] T. Naseer, W. Burgard, and C. Stachniss, (2018) Robust visual localization across seasons. IEEE T-RO, vol. 34, no. 2: 289–302.
[4] M. Milford and et al., (2015) Sequence searching with deep-learnt depth for condition-and viewpoint invariant route-based place recognition. In: CVPR Workshops. Boston. pp. 18–25.
[5] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, (2015) Place recognition with convnet land-marks: Viewpoint-robust, condition-robust, training-free. In: Proceedings of the 2010 Academy of Marketing Science (AMS) Annual Conference. Columbia.
[6] Cummins, M., Newman, P. (2011) Appearance-only SLAM at large scale with FAB-MAP 2.0. The International Journal of Robotics Research, 30 (9): 1100-1123.
[7] Sünderhauf, N., Neubert, P., Protzel, P. (2013) Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In: IEEE International Conference on Robotics and Automation (ICRA) on Proc. of Workshop on Long-Term Autonomy. Baden.
[8] Milford, M. J., Wyeth, G. F. (2012) SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In: IEEE International Conference on Robotics and Automation. Minnesota. pp. 1643-1649.
[9] Pepperell, E., Corke, P., Milford, M. (2016) Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints. International Journal of Robotics Research, 1–23.

[10] Pepperell, E., Corke, P. I., Milford, M. J. (2014) All-environment visual place recognition with SMART. In: IEEE International Conference on Robotics and Automation (ICRA). Hong Kong. pp: 1612-1618.

[11] Bruce, J., Jacobson, A., Milford, M. (2017) Look No Further: Adapting the Localization Sensory Window to the Temporal Characteristics of the Environment. IEEE Robotics and Automation Letters, 2 (4): 2209-2216.

[12] Naseer, T., Spinello, L., Burgard, W., etc. (2014) Robust visual robot localization across seasons using network flows. In: Twenty-Eighth AAAI Conference on Artificial Intelligence. Québec. pp: 289 – 302.

[13] Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., & Romera, E. (2017) Are you able to perform a life-long visual topological localization? Autonomous Robots.

[14] Jegou, H., Perronnin, F., Douze, M., etc. (2012) Aggregating local image descriptors into compact codes. IEEE transactions on pattern analysis and machine intelligence, 34 (9): 1704-1716.

[15] Oliva, A., Torralba, A. (2006) Building the gist of a scene: The role of global image features in recognition. Progress in brain research, 155: 23-36.

[16] Maddern, W., Milford, M., & Wyeth, G. (2012) Cat-slam: probabilistic localisation and mapping using a continuous appearance-based trajectory. International Journal of Robotics Research, 31(4): 429-451.

[17] Jun, M., Xiaoping, H., Michael, M. (2018) An adaptive localization system for image storage and localization latency requirements. Robotics & Autonomous Systems, 107: 246-261.