REVIEW ON LOOP CLOSURE DETECTION OF VISUAL SLAM

Kavitha M L[1], Deepambika V A[2]
[1]PG Scholar, [2]Assistant Professor,
Dept. of ECE, LBS Institute of Technology for Women, Kerala, India

Abstract

Loop closure detection is an important issue of visual SLAM (Simultaneous Localization and mapping).It reduces the drift of localization and is essential for building a consistent SLAM map. Here in this review paper a comparative study of various image descriptors used for loop closure detection is discussed. Some handcrafted feature based descriptors BoVW, FV, VLAD, GIST and a CNN based image descriptors are taken for the study. Handcrafted feature descriptors share the weakness of lack of robustness with respect to illumination changes and high computational cost which can be overcome by a CNN based image descriptor.

Index Terms: image descriptor, loop closure, SLAM.

I. INTRODUCTION

Loop closure detection is considered one of the most important problems in SLAM. A SLAM algorithm aims to simultaneously localizing the position of some sensor with respect to the unknown environment at the same time mapping the structure of the environment. In case of visual SLAM the sensor used is one or more camera. Loop closure detection is the problem of determining whether a mobile robot has returned to a previously visited location, and it is critical for building a consistent map of the environment by correcting errors that accumulate overtime.

One class of the popular algorithm for loop closure detection is image matching. Image matching typically proceeds in two steps they are image description and similarity measurement. Image description is the most critical in visual loop closure detection. In the paper a comparative study of various image descriptors BoVW, FV, VLAD, GIST which was handcrafted features and a CNN based image descriptor was done.

The rest of this paper is organized as follows. Section II gives a brief description of SLAM problem. In section III we present various solutions to the SLAM problem. Section IV is about the basic SLAM system. The section V is about literature review on various image descriptors and in section VI similarity measurement techniques were discussed. Finally, we conclude the paper in section VII.

II. THE SLAM PROBLEM

Simultaneous localization and mapping (SLAM) is the computational problem of constructing or updating a map of an unknown environment while simultaneously keeping track of an agent's location within it. While this initially appears to be a chicken-and-egg problem that is the challenge in SLAM is to recover both the camera pose and map structure while initially knowing neither. A solution to the SLAM problem has been seen as a holy grail for the mobile robotics community as it provides the means to make a robot truly autonomous. The SLAM problem can be abstracted as follows. As a robot explores an environment, it moves through a series of discrete poses $x_1, \ldots\ldots, x_k$. The movement of the robot from one pose to the other is modelled by a control input, u_k. At each pose, a measurement of the environment is captured, resulting in a series of measurements $z_1, \ldots\ldots, z_k$. The end goal is to estimate the map, m, and robot pose x_k. In probabilistic form, the SLAM problem requires the probability distribution $p(x_k, m|z^k, u^k, x_0)$ to be calculated. A recursive estimator which is central to virtually all SLAM algorithms is as follows:

$$p(x_k, m|z^k, u^k, x_0) = \eta_k \quad p(z_k|x_k ,m) \int p(x_k|x_{k-1}, u_k)\, p(x_{k-1}, m|z^{k-1}, u^{k-1}, x_0)\, dx_{k-1}$$

where the recursive estimator is a function of a motion model $p(x_k|x_{k-1},u_k)$and a measurement model $p(z_k|x_k, m)$.

III. SOLUTION OF SLAM PROBLEM

Solution to the probabilistic SLAM problem involve finding an appropriate representation for both the observation model and motion model . The most common representation is in the form of a state space model with additive Gaussian noise, leading to the use of the Extended Kalman filter (EKF) to solve the SLAM problem. Another alternative representation to describe the vehicle motion model is the Rao-Blackwellized particle filter, or Fast SLAM algorithm

A. EKF-SLAM

It concurrently provides an estimate of uncertainty of the robot pose and the landmark positions based on predictive model of the robot's motion and the relative measurement of landmarks. The state vector, X, consisting of robot pose, x_v, and *n* landmarks, $x_{f1},……,x_{fn}$, can be expressed as $X = [x_v, x_{f1}, x_{f2},….., x_{fn}]$. Associated with the state vector is a map covariance matrix, P(x), where the o®-diagonal sub-matrices encode the correlations between landmark location estimates and provide the mechanism for updating all the relational estimates. One limitation of the EKF approach is computational in nature. Computational complexity increases quadratically with *n* number of landmarks stored in the state vector. This limits the number of landmarks this approach can handle. The EKF can only maintain one hypothesis with its unimodal Gaussian distribution model. To resolve this, multiple Kalman filters were used to maintain multiple hypotheses at the expense of computational complexity. Landmark-based EKF SLAM systems generally employ data association techniques.

B. FAST SLAM

Fast SLAM which is based on Monte Carlo sampling, or particle filtering was the first to directly represent the nonlinear process model and non-Gaussian pose distribution. Fast SLAM decomposes the SLAM problem into a robot localization problem and a collection of landmark estimation problems conditioned on robot pose estimate. Taking advantage of an insight that the posterior can be factored, the problem of determining *N* landmark locations is decoupled into *N* independent estimation problems when robot path is known. The path estimator is implemented using a Rao-Blackwellised particle filter while the landmark location estimator is implemented using EKFs. Each particle consists of an estimate of the present robot pose, s_k, all the previous poses of the robot, s^{k-1}, and a set of *N* independent EKFs that estimate the *n* landmark locations conditioned on the path estimate. The particle set is resampled based on consideration of control input, u_k and measurement, z_k A tree-based data structure was developed to reduce the running time of Fast SLAM to $O(mlog(n))$, where *m* is the number of particles and *n* is the number of landmark. Fast SLAM is capable of mapping larger environments than EKF SLAM due to less computational complexity. Fast SLAM does not face the limitation of unimodal Gaussian distribution as EKF-SLAM. Fast SLAM is capable of maintaining multiple data association hypotheses for landmarks simultaneously because each landmark location estimators can make different data association decisions and are independent of each other.

IV. THE SLAM SYSTEM

Most of the state of art system consists of

- A feature detector : That finds the point of interest within the image
- A feature descriptor: That matches tracks features from one image to the next.
- An optimization backend: that uses said correspondences to build geometry of the scene and find the position of the robot.
- A loop closure detection algorithm: That recognizes previously visited areas and adds constraints to the map.

Loop closure detection is considered as one of the most important problem in SLAM .Loop closure detection is the problem of determining whether a mobile robot has returned to the previously visited areas. A correct loop closure guarantees the consistency of the SLAM map and improves all round accuracy. Thus LCD is critical for creating a topologically correct map and for improving the metric information of the map. Also a successful LCD improves the computational efficiency and robustness to false positives. Most popular LCD is based on image matching that is matching the current view of the image with already captured image. Mainly image matching consists of two steps- image description and similarity measurement.

- Image description: Compress an image into a one dimensional vector that is more compact and discriminating is the most critical step in LCD.
- Similarity measurement: Here the distance with dimensions representing features of the image is obtained. If the distance is small, it will show high degree of similarity and if the distance is large, it exhibits lower degree of similarity.

In this paper a comparative study of the various image descriptors for LCD was done. Most of the descriptors chosen were based on hand crafted features. Hand crafted features are designed through a process of where human knowledge dominates for obtaining the desired characteristics. The hand crafted feature based image descriptors chosen here for comparative study are BoVW, FV, GIST, and VLAD. They perform well only on homogeneous dataset. In case of heterogeneous and unfamiliar dataset their performance is poor. For comparative study CNN based image descriptors is also chosen. Due to the high levels of abstraction CNN is able to extract semantic information which is difficult to obtain with hand crafted features.

V. LITERATURE REVIEW

In [3] BoVW characterizes an image as a histogram of visual words. To build BoVW the local key points of an image is first detected and described online. Then cluster this local key point descriptors whose cluster centre forms the visual words, which are simply the vector quantized version of local key point descriptors such as SIFT, SURF etc. This offline clustering thus creates the visual dictionary. BoVW then measures the similarity between the observations in word space. Due to the invariance properties of this key point descriptors BoVW has become one of the popular techniques for LCD. The creation of this visual dictionary is one of the major weaknesses of BoVW as this adopts an appearance assumption which makes it inadequate in situations such as when different cameras are used or when using it on a totally different environment. Also the vector quantization in BoVW leads to perceptual aliasing that means images from totally different location are considered as correct matches. In addition to this considerable amount of time is consumed for feature extraction and matching process. BoVW method relies on Bayesian filtering to obtain the loop closure probability. For each new coming image it calculate the probability that the current image comes from an already captured scene. Loop closure hypothesis whose probability is above some threshold are confirmed when the epipolar geometry constraint is satisfied. This ultimate validation step is achieved by multiple view geometry algorithms.

[4] Is based on fisher vector (FV) image descriptor .In FV method images are characterized by first detecting and describing key point descriptors in an online method an then extracting it and computing their deviation from a universal generative mode which is a probabilistic visual vocabulary learned offline from a large set of samples. The deviation is measured by computing the gradient of the sample log-likelihood with respect to the model parameters. This leads to a vectorial representation which called as FV. To construct the visual word dictionary it uses a Gaussian mixture model (GMM), where the mean of the Gaussian components are cluster centres as in BoVW and the covariance captures the distribution of the key point descriptors within the clusters. Due to the computation of deviation of the key point descriptors it uses a second order statistics. Thus it encodes richer information than BoVW .Since richer information is encoded, it gives a better representation of the image. FV is impractical for large scale application as it being high dimensional and dense, requires large storage requirements. After feature description using suitable similarity score can be obtained which can then be used for loop closure detection applications.

[5] Was about VLAD based image descriptors. It is a compact descriptor which is used for large scale image retrieval. Compared to FV, VLAD descriptor is preferable when the trade-off between the performance and memory footprint of an image descriptor is important. VLAD vector is a simplified version of fisher vector which can be considered as a first order statistics of the non-probability fisher vector. It represents an image as a VLAD vector, which is obtained by training a codebook of k visual words using k

means or HKM and the similarity is estimated by measuring the distance of related vectors. For LCD there are two steps that is offline learning and online detection. Offline learning learns surf feature visual codebook which is then used for computing VLAD feature vector of an image while the online detecting part consists of VLAD computing, image database query and geometric check. It requires only less execution time and memory consumption than BoVW and FV.

[6] Is about GIST image descriptor. Most of the image descriptor so far discussed were based on local key point descriptors where we observed that they require large computation time for online extraction and for offline processing. This Gabor-gist method is a single efficient image descriptor of low dimension to describe and measure similarities among images. It is a global image descriptor that provides a compact representation of the image. Using this compact image descriptor appearance based loop closure detection is

obtained here. To track the matching candidates for avoiding unnecessary matching, this method exploits the temporal coherence in the image sequence with the particle filter framework. Also it uses an efficient likelihood function in a probabilistic framework and maintains a fraction of the entire hypothesis during the LCD process. For improving the computational efficiency and discriminative power of the image discriminator PCA is employed for dimension reduction of Gabor Gist descriptor. Due to the compactness of the image descriptor and simplicity of particle filtering this method is highly scalable. However since the Gist descriptor is computed for an entire image, its robustness with respect to image transformations such as camera motion and illumination variation may hamper its effectiveness in image matching applications.

[7] Here CNN based image descriptors for LCD of SLAM was discussed. The above discussed image descriptors were based on handcrafted features. They lack robustness with respect to illumination changes and high computational cost. Here a CNN based image descriptors which learn features from the raw pixels of the input image without prior knowledge or human interaction is discussed. It uses a pretrained CNN model to extract CNN whole image descriptor easily one from each layer by travelling along the depth of the network. The feature vector from each layer is then normalized. The further and deeper into the CNN network, the more abstract representation of the input image. Due to this high level of abstraction it is able to extract semantic information and exhibits a high degree of invariance properties. Then for similarity measurement, the Euclidian distance of each descriptor to its nearest neighbor in the robot map is computed.

VI. SIMILARITY MEASUREMENT

Scene matching is achieved by computing a measure of similarity between the observations. Using some extraction processes an observation, *Si*, is represented by a set of descriptors, $\vec{S_i}$. A similarity function *Sim*($\vec{S_i}, \vec{S_j}$) $\in$[0,1] measures the similarity between observations *Si* and *S$_j$* and assigns a pair wise similarity score. A potential match between observations is allowed if the similarity value is above a certain threshold. Different similarity metrics may be appropriate for different descriptors. A brief overview of comparison techniques that are used in this work are explained below.

A. *Voting Algorithm*

The basis of a voting algorithm is to sum the number of matches between descriptors of two observations. Consider the case in which a query observation is compared with a database consisting of a set of observations S$_1$,…..,S$_k$. Each observation is described by a vector of descriptors $\vec{S_i}$. During the comparison process, each descriptor from one vector is matched against the descriptors from another vector. For example, similarities between two SIFT descriptors can be quantified using the Euclidean distance as follows:

$$d_{euc} = \left\| d_i - d_j \right\| = \sqrt{\sum_{k=1}^{128} (d_i(k) - d_j(k))^2}$$

Where d_{euc} is the Euclidean distance, d_i and d_j are SIFT descriptors from observation S_i and S_j respectively and *k* is the index number. If the distance, d_{euc}, is below a certain threshold, a match *id*(d_i, d_j) is considered found and a vote is added to the observation vector.

B. *Cosine Distance Function*

The similarity between two observations, S_i and S_j can be measured by calculating the cosine of angle (θ_c) between the two representative vectors of descriptors, $\vec{S_i}$ and $\vec{S_j}$ .A vector can be normalized by dividing each of it component by its length. This ensures that observations comprising of more descriptors do not score better just by virtue of the number of descriptors.

$$\mathrm{sim}(\bar{S}_i, \bar{S}_j) \blacksquare \mathit{cosine}(\theta_{ij}) \blacksquare \frac{\bar{S}_i \bar{S}_j}{|S_i||S_j|}$$

Where $\bar{S}_i$ $\mathit{and}\bar{S}_j$ are vectors of descriptors representing observations S_i and S_j respectively.

C . Combination of Similarity Measures

Different similarity metrics based on different characteristics of an observation can be used conjunctively to give an overall similarity score. The rationale behind using a combination of similarity metrics is that even though one comparison technique may accidentally allow an invalid match, it is unlikely several techniques will, given that these techniques are independent.

TABLE I. Comparison of various image descriptors

Paper	**Image descriptor**	**Characteristics**	**Advantages**	**Disadvantages**
A visual bag of words method for interactive qualitative localization and mapping	BoVW	Hand crafted local key point descriptor	➢ Invariance property	➢ Creation of visual dictionary is time consuming
Image classification with the fisher vector	FV	Hand crafted local key point descriptor	➢ Encode richer information	➢ Large storage requirements ➢ Impractical for large scale applications
VLAD- based loop closure detection for monocular SLAM	VLAD	Hand crafted local key point descriptor	➢ Less execution time ➢ Less storage requirements	➢ Encodes less information than FV
Visual loop closure detection with a compact image descriptor	GIST	Handcrafted global image descriptor	➢ Compact ➢ Highly scalable	➢ Limited robustness with respect to image transformations
Convolutional Neural Network-Based image representation for visual loop closure detection	CNN	CNN based image descriptor	➢ More abstract representation ➢ Can extract semantic information ➢ High degree of invariance properties	➢ Different representation for different CNN networks

VII. CONCLUSON

In this review paper the basic SLAM problem and its various solutions where discussed first. Then we focused on one of the important issue of SLAM which is the loop closure detection. For loop closure detection image description forms the major part. Here various image descriptors used for LCD of visual SLAM is studied and compared. Various descriptors here used for comparative study are BoVW, FV, VLAD, GIST and CNN image descriptors. Except CNN all others are hand crafted features. BoVW is an earliest feature descriptor which is based on a visual word dictionary of local key descriptors whose storage requirements is large. FV is adopted for obtaining richer information than BoVW but it suffers from the disadvantage of large storage space. VLAD is preferred when a tradeoff between performance and memory requirements is required. GIST is a global descriptor which is compact and highly scalable. Recent advancement uses CNN based image descriptors which can extract semantic information of an image and which offers high degree of invariance property.

REFERENCES

[1] Durrant-Whyte, Hugh, and Tim Bailey. "Simultaneous localization and mapping: part I." IEEE robotics & automation magazine 13.2 (2006): 99-110.

[2] Aulinas, Josep, et al. "The SLAM problem: a survey." CCIA 184.1 (2008): 363-371.

[3] Filliat, David. "A visual bag of words method for interactive qualitative localization and mapping." Robotics and Automation, 2007 IEEE International Conference on. IEEE, 2007.

[4] Sánchez, Jorge, et al. "Image classification with the fisher vector: Theory and practice." International journal of computer vision 105.3 (2013): 222-245.

[5] Huang, Yao, Fuchun Sun, and Yao Guo. "VLAD-based loop closure detection for monocular SLAM." Information and Automation (ICIA), 2016 IEEE International Conference on. IEEE, 2016.

[6] Liu, Yang, and Hong Zhang. "Visual loop closure detection with a compact image descriptor." Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on. IEEE, 2012.

[7] Convolutional Neural Network-Based image representation for visual loop closure detection.

[8] Ho, Kin Leong. *Loop Closing Detection in SLAM Using Scene Appearance*. Diss. University of Oxford, 2007.

[9] Bai, Dongdong, et al. "Matching-range-constrained real-time loop closure detection with CNNs features." Robotics and biomimetics 3.1 (2016): 15.