Visual Loop Closing using Gist Descriptors in Manhattan World

Gautam Singh and Jana Košecká
Department of Computer Science
George Mason University

***Abstract*— We present an approach for detecting loop closures in a large sequence of omni-directional images of urban environments. In particular we investigate the efficacy of global gist descriptors computed for 360^o cylindrical panoramas and compare it with the baseline vocabulary tree approach. In the context of loop closure detection, we describe a novel matching strategy for panoramic views, exploiting the fact that the vehicle travels in urban environments where heading of the vehicle at previously visited locations and loop closure points are related by multiple of 90^o degrees. The performance of the presented approach is promising despite the simplicity of the descriptor.**

I. Introduction

The problem of generating metric and/or topological maps from streams of visual data has became in recent years a very active area of research. This increased interest has been to a large extent facilitated by improvements in large scale wide-baseline matching techniques and advances in localization by means place recognition. The problem of localization by means of place recognition, for purely appearance based strategies is typically formulated as an image based retrieval task. Namely given a database of views from certain geographical area, and set of new query views the goal was to determine the closest view from the reference database.

The problem of loop closure detection we investigate here is different in its nature in that it explicitly takes into account temporal ordering constraints among the views as opposed to considering the database as unorganized collection of views. The loop closure problem requires determining for two images whether they have been taken from the same place. In principle the problem of loop closure detection can be tackled using the same strategies as those used in location recognition. Namely given n views of the video sequences, loops are hypothesized by comparing all views to all other views.

The efficiency and scalability of the existing strategies depends on chosen image representation and the selected similarity measure. In this paper we investigate the suitability of the global gist descriptor as image representation and proposed a novel image panorama similarity measure between two views, which exploits the Manhattan world assumption stating that the vehicle heading at previously visited locations and current views are related by multiple of 90^o degrees. We will demonstrate that despite the simplicity and compactness of the global gist descriptor, its discriminability is quite high partly due to 360^o field of view.

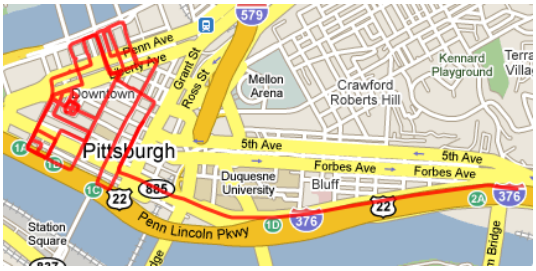


Fig. 1. Street view data set of panoramic views.

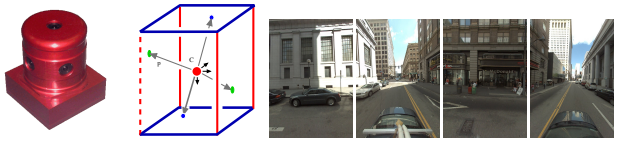


Fig. 2. Panorama acquisition device. (a) Point Grey *LadyBug* camera. (b) A panoramic piecewise perspective image as an outer surface of the prism.

II. Related work

There are several approaches for loop-closure detection in topological maps. The purely visual appearance based system for location recognition and topological localization is the FAB-MAP [1]. It uses bag of words image representation and explicitly models the dependencies between different visual words. In [2] authors use both odometry and appearance to maintain different hypothesis of topological maps. Another examples of approaches which uses also metric or odometry information are [3], [4].

Another class of methods uses only similarity matrix between all pairs of views and differ in how are the similarity scores computed, how are they used for loop detection and subsequent navigation. In [5] authors detect loop closure directly from the similarity matrix and formulate it as problem of detecting statistically significant sequences from the similarity matrix. In [6] authors formulate the loop closure detection in MRF framework and propose novel similarity measure for comparing two panoramas. The rotational invariance with respect to changes in heading is achieved by alignment of local features projected on horizontal plane using dynamic programming approach. In [7] authors used modified vocabulary tree approach for computation of image similarity scores. Localization using specifically omni-directional images has been shown very effective in the context of topological localization. The wide field of view

allows the representation of the appearance of a certain place (topological location) with a minimal amount of reference views. Examples of systems performing topological mapping and navigation are [8], [9], [10]. Additional approaches to both offline and online topological map building can be found in [11], [12], [13] or [14], [15].

III. Matching Panoramas

In our work we use Street View panoramas acquired by 360^o field of view LadyBug camera[1]. We create one panoramic image by warping the radially undistorted perspective images onto the sphere assuming one virtual optical center. One virtual optical center is reasonable assumption considering that the structure around the sensor is very far compared to the discrepancy between optical centers of all the cameras. The sphere is backprojected into a quadrangular prism to get a piecewise perspective panoramic image, see Fig. 2. Our panorama is composed of four perspective images covering in total 360^o horizontally and 127^o vertically. We do not use the top camera as there is not much information. The panorama is then represented by 4 views (front, left, back and right) each covering 90^o horizontal FOV. We discard the bottom part of all views, discarding the areas of the images that captured parts of the car acquiring the panoramas. In the following two sections we describe and subsequently compare two different panoramic image representations and two different panorama similarity measures for determining the loop closure.

A. Panoramas Similarity Measure

In the first stage we investigate the effectiveness of panorama matching using the gist descriptor. The gist descriptor [16] is a global descriptor of an image where statistics of different filter outputs are computed over 4 x 4 sub-windows. The feature vector corresponds to the mean response to steerable filters at 4 scales and 8, 8 and 4 orientations. The advantage of the descriptor is that it is very compact and fast to compute. Each image is represented by a 320 dimensional vector (per color band). The gist feature has been shown to be effective in holistic classifications of scenes into categories containing tall buildings, streets, open areas, highways and mountains etc [17] and has been used effectively for retrieving nearest neighbors from large scale image databases. In order to obtain the gist descriptor for the entire panorama, we compute the standard gist descriptor for each of the 4 views. In our previous work [18] we have explored the feasibility of the gist descriptor for place recognition with a different similarity measure and matching strategy.

The location recognition problems and loop closure detection are closely related, as they both require definition of image similarity metric and capability of handling and representing large amounts of data. One aspect where the problems differ is in how is the performance evaluated. In case of place recognition we typically assume that the database of images of the entire environment has been created and the query views are selected approximately along the same route as the vehicle travelled. In case of loop closure detection, each query view is compared to all previously visited views/locations, except the locations visited in previous 25 meters.

Given two panoramas, we define their similarity measure in the following way. Suppose the reference panorama is described by a 4-tuple of gist descriptors computed for left, front, right and back portion of the panorama denoted by $\mathbf{g}^r = [g_1^r, g_2^r, g_3^r, g_4^r] = \mathbf{g}_{1234}^r$ and similarly the query view with composite gist descriptor of $\mathbf{g}^q = [g_1^q, g_2^q, g_3^q, g_4^q] = \mathbf{g}_{1234}^q$, where the short hand index $_{1234}$ denotes the order of individual components of the descriptor. In order to compare the two panoramas, we want to take into account the possibility that they have been taken at different orientation headings. The type of loop closures we typically encounter in urban environments are:

1) portion of the street is traversed two times in the same direction,
2) portion of the street is traversed two times in the opposite direction,
3) vehicle passes an intersection with the street travelled previously.

In order to accommodate the level of viewpoint invariance required by the above mentioned changes in heading and assuming that the changes of heading are approximately multiples of 90^o, we propose to consider in matching the following permutations of the descriptors obtained by circular shifts $\mathbf{g}_{1234}$, $\mathbf{g}_{2341}$, $\mathbf{g}_{3412}$, $\mathbf{g}_{4123}$. The similarity measures between two panoramas is then defined in the following way

$$dist(\mathbf{g}^q, \mathbf{g}^r) = \min_k d_s(\mathbf{g}^q, \pi_k(\mathbf{g}_{1234}^r)) \quad (1)$$

where π_k is the k^{th} circular permutation of the gist component vectors (k = 1,2,3,4) and d_s is the sum of euclidean distances between individual components of the gist vector.

$$d_s(\mathbf{g}^r, \mathbf{g}^q) = \sum_{i=1}^{4} \|g_i^r - g_i^q\|.$$

B. Vocabulary Trees

In the following paragraph we describe the loop detection strategy using the vocabulary tree approach in Manhattan world. The vocabulary trees proposed by [19] use the concept of visual words. Local features in views are similar if they correspond to the same visual word in the tree.

For our experiments on vocabulary trees, we used SIFT features [20] as local features of the views. SIFT features correspond to highly distinguishable image locations which can be detected efficiently and have been shown to be stable across wide variations of viewpoint and scale. Views are added to a database as inverted files. The inverted files store the id-numbers of the views in which a particular node of the tree occurs, as well as for that view the number

[1]More details on the dataset are provided in the experimental section

of features that match this node. The closest match for a feature can be obtained by quantizing it using the vocabulary tree. This helps determine the closest visual word to which the feature corresponds. The use of inverted files for fast matching using the L_2 norm has been shown to be an efficient method by [19].

A node *i* in the tree is assigned the weight w_i and can be used to define both the query q_i and database vectors d_i as

$$q_i = n_i \times w_i$$
$$d_i = m_i \times w_i$$

where n_i and m_i are number of features of the query and database image, respectively, with a path through node *i*. The weight w_i is computed using entropy weighting

$$w_i = \ln \frac{N_t}{N_i}$$

where N_t is the total number of images currently in the database and N_i is the number of images in the database with at least one descriptor vector path through node *i*. This results in a TF-IDF (term frequency-inverse document frequency) scheme. The relevance score between a query and a database view is based on the normalized difference between the query q_i and database vector d_i in L_p norm. We use L_2 norm to calculate this difference.

In our method for loop closure detection, a database is constructed using the views corresponding to the initial locations of the vehicle's traversal. After that, whenever a new location is visited, its image views are queried against the database. To ensure that image views of immediately preceding locations are not returned in the results, we disregard the views from the previous 25 locations while matching. After computing the match results for the four views, these four views are subsequently added to the database as inverted files. One must note that the addition of the four views of the current location to the database will lead to an updating of the weights of all the visual words present in those images. The process of querying the views of a location and then adding them to the database is henceforth performed for all the remaining locations in the dataset.

The top N match results for a single view ranked $(1, 2, ..., N)$ are returned with their corresponding relevance scores $(s_1, s_2, ..., s_N)$ where $(s_1 < s_2 < ... < s_N)$. The results for the four views $(q_1^k, q_2^k, q_3^k, q_4^k)$ of a query location q_k are accumulated and used to determine the best match for that location. The database location which occurs most frequently in the results is considered the closest match for a location. In case of an equal number of occurrences for multiple locations, an aggregate score based on either the ranks or relevance scores is computed and the location with the lowest aggregate score is chosen as the closest match. Our two different aggregate score calculation strategies are

1) Rank aggregate for a database location l_i is $agg_i^{rank} = \sum_{j=1}^{4} rank(q_j^k, l_i)$
2) Relevance score aggregate for a database location l_i is $agg_i^{relevance} = \sum_{j=1}^{4} score(q_j^k, l_i)$

where $rank(q_j^k, l_i)$ is the ranking of the location l_i for query view q_j^k as returned from top N nearest neighbours and $score(q_j^k, l_i)$ is the SIFT similarity score of location l_i and query view q_j^k described at the beginning of this section. The results comparing these two scores are reported in Table I.

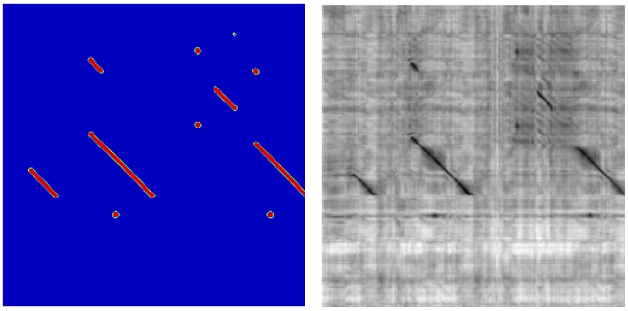


Fig. 3. Visualization of sub-matrix of the gist similarity score matrix: (left) the ground truth matrix, (right) the similarity matrix computed using the proposed measure. The sub-matrix is for a section of the dataset which was traversed more than once. The red diagonal lines correspond to streets that were visited in the same direction and red points correspond to an intersection being visited from a different direction

IV. Experiments

A. Ground truth and evaluation

We demonstrate the performance of our approach in a large dataset of 12,000 street view panoramas[2]. This data set is composed of a 13 mile long run in urban area and can be seen in Fig. 1. To extract the explained gist descriptor, we have used the code available on the web at http://people.csail.mit.edu/torralba/code/spatialenvelope/.
The locations in the dataset have been provided with GPS coordinates in degrees specifying the latitude and longitude of the vehicle. The distance d in meters between two locations with GPS coordinates (lat_1, lon_1) and (lat_2, lon_2) was calculated using the formula:

$$\begin{aligned} dx &= 69.1 \times (lat_2 - lat_1) \\ dy &= 69.1 \times (lon_2 - lon_1) \times \cos(lat_1 \times \pi/180) \\ d &= 1609.344 \times \sqrt{dx^2 + dy^2} \end{aligned} \quad (2)$$

For a given location all neighbouring locations within a distance threshold of 10 meters of the current location are considered to be the ground truth. In order to avoid considering nearby locations as loop closures, we use a window of 25 preceding frames to avoid considering views taken within short time of each other. This produced 3362 ground truth locations.

[2]Dataset provided for research purposes by Google™.

Algorithm	Loop Closures Identified	Accuracy
Rank aggregate	2302	68.47
Relevance score aggregate	2437	72.49

TABLE I

COMPARISON OF ACCURACY BETWEEN DIFFERENT AGGREGATE METHODS USED TO COMBINE INDIVIDUAL VIEW RESULTS ON QUERYING THE VOCABULARY TREE.

B. Vocabulary Tree comparison

Our experiments in using vocabulary trees were performed using a single hierarchical k-means tree. We thank Friedrich Fraundorfer for providing us this tree which was built offline on a set of arbitrary images from across the web, but including 2500 urbanscape images. One must note that the tree was built using images which are unrelated to the sequence for which loop closing is tested. The provided vocabulary tree had a branching factor k of 10 and the number of levels L is 6. This results in a tree with k^L = 1,000,000 leaf nodes. We sample 500 of the extracted SIFT features of each view for quantization purposes. We used N = 4 for our experiments. The number of locations used to build the intial database was set at 100. The accuracy statistics of using vocabulary trees for loop closure detection are tabulated in Table I.

C. Gist Similarity Measure

The gist similarity measure is very fast to compute and the storage requirements for the descriptors are minimal. Each panorama is comprised of 4 views and each image view is represented by a 320 dimensional gist descriptor vector. Storage of a single view's gist descriptor requires 1.3 KB of disk space and the database size of gist descriptors for the four views of the entire dataset of 12000 locations is 62MB. A subset of the gist similarity score matrix for a sequence of the dataset with loop closure in it is visualized in Figure 3.

We carried out several experiments evaluating the proposed similarity measure for loop closure detection. In our experiments, for a query location, the top 75 match results were returned in the ascending order of their gist similarity measure with this location. Accuracy statistics were collected for evaluation on the ground truth locations. In Figure 4a we show the effectiveness of the Manhattan world assumption and its associated similarity measure which compares each query view panorama with all rotational permutations of the reference views. In Figure 4b we demonstrate the effect of 360^o field of view for location discrimination. Notice that having the full FOV significantly improves the discrimination capability of the loop closure detection. Considering a smaller portion of the panorama, which resembles the localization with traditional cameras is shown to be detrimental to the overall performance.

Precision-Recall curves for the different gist similarity score methods are shown in Figure 4c . They were generated by thresholding the gist similarity score for the top-1 result returned in the search. Locations are labeled true positive if the gist similarity score is below the threshold and the distance between the query and the result is less than the threshold distance of 10 meters. A false positive occurs if the similarity score is below the threshold but the distance is more than 10 meters.

V. CONCLUSION

We introduced novel similarity measure between image panoramas and evaluated its efficiency for loop closure detection in urban environments. The approach is built upon gist descriptors which are a representations of parts of panoramas. We showed that this strategy can be used efficiently for detecting locations which are being revisited and its storage and computational efficiency of the matching stage are superior to the vocabulary trees. The capability of discriminate individual locations is largely related to the 360^o FOV. We are currently extending this baseline approach to incorporate temporal constraints and build topological model of this challenging environment, where nearly 25% of locations were visited more then once.

REFERENCES

[1] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *Int. J. of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[2] A. Ranganthan, E. Menegatti, and F. Deallert, "Bayesian inference in the space of topological maps," in *Proc. of IEEE CVPR*, 2006, pp. 2161–2168.

[3] N. Tomatis, I. Nourbaksh, and R. Siegwart, "Hybrid simultaneous localization and map building: a natural integration of topological and metric," *Robotics and Autonomous Systems*, vol. 44, no. 1, pp. 3–14, 2003.

[4] M. Bose, P. Newman, J. Leonard, M. Soika, W. Feiten, and S. Teller, "An atlas framework for scalable mapping," in *Proc. of IEEE CVPR*, 2006, pp. 2161–2168.

[5] K. L. Ho and P. Newman, "Detecting loop closure with scene sequences," *Int. J. of Computer Vision*, vol. 74, no. 3, pp. 261–286, September 2007.

[6] R. Anati and K. Daniilidis, "Constructing topological maps using markov random fields and loop closure detection," in *NIPS*, 2009, pp. 2161–2168.

[7] F. Fraundorfer, C. Engels, and D. Nistér, "Topological mapping, localization and navigation using image collections," in *In Proc. of IEEE/RSJ IROS*, 2007, pp. 3872–3877.

[8] J. Gaspar, N. Winters, and J. Santos-Victor, "Vision-based navigation and environmental representations with an omnidirectional camera," *IEEE Trans. on Robotics and Automation*, vol. 16, no. 6, pp. 890–898, 2000.

[9] E. Menegatti, T. Maeda, and H. Ishiguro, "Image-based memory for robot navigation using properties of the omnidirectional images," *Robotics and Autonomous Systems*, vol. 47, no. 4, pp. 251–267, 2004.

[10] A. Tapus and R. Siegwart, "Incremental robot mapping with fingerprints of places," *In Proc. of IEEE/RSJ IROS*, pp. 2429–2434, 2005.

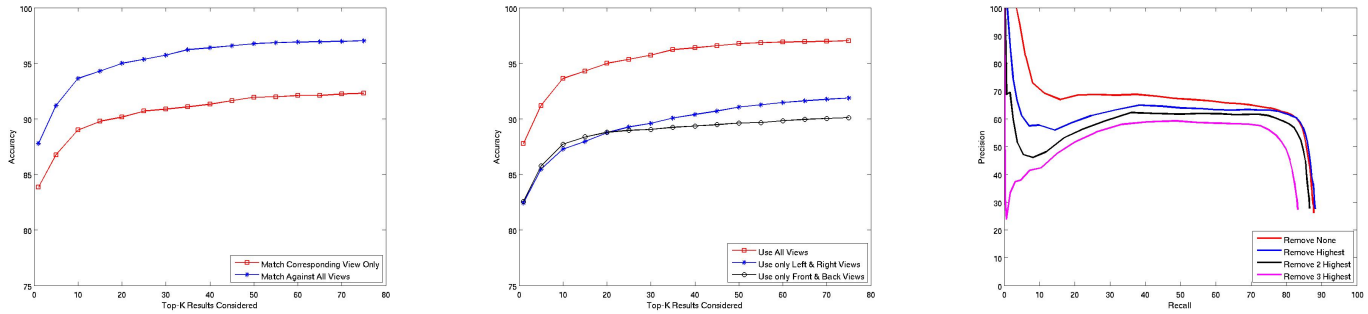
[11] T. Goedemé, M.Nuttin, T. Tuytelaars, and L. Van Gool, "Omnidirectional vision based topological navigation," *Int. J. of Computer Vision*, vol. 74, no. 3, pp. 219–236, 2007.

[12] C. Valgren, T. Duckett, and A. J. Lilienthal, "Incremental spectral clustering and its application to topological mapping," in *Proc. of IEEE ICRA*, 2007, pp. 4283–4288.

[13] Z. Zivkovic, O. Booij, and B. Krose, "From images to rooms," *Robotic and Autonomous Systems*, vol. 55, no. 5, pp. 411–418, 2007.

[14] C. Valgren and A. J. Lilienthal, "Sift, surf and seasons: Long-term outdoor localization using local features," in *European Conf. on Mobile Robots*, 2007, pp. 253–258.

[15] A. C. Murillo, C. Sagüés, J. J. Guerrero, T. Goedemé, T. Tuytelaars, and L. Van Gool, "From omnidirectional images to hierarchical localization," *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 372–382, 2007.



(a) (b) (c)

Fig. 4. Performance evaluation of the gist based similarity measure for loop detection. (a) Accuracy of the closest gps match amongst top-k results using gist aggregates of all views (b) Accuracy of the closest gps match amongst top-k results using gist aggregates of all four views and a subset of views (c) Precision Recall statistics for the first gist match result

top 1 top 5 top 75

Fig. 5. Visualization of correct loop closures and their associated accuracy. Green line are the correctly identified locations, red line are the locations of false negatives which our method missed. The above graph plots the results evaluating whether the closest reference view among top k nearest views to the query view.

[16] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," in *Visual Perception, Progress in Brain Research*. Elsevier, 2006, vol. 155.

[17] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Proc. of IEEE ICCV*, 2003, p. 273.

[18] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *IEEE Workshop on Omnidirectional Vision, Camera Netwoks and Non-Classical Cameras*, 2009, pp. 1–6.

[19] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. of IEEE CVPR*, 2006, pp. 2161–2168.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004, http://www.cs.ubc.ca/ lowe/keypoints/.