2015 IEEE International Conference on Robotics and Automation (ICRA)
Washington State Convention Center
Seattle, Washington, May 26-30, 2015

Location Graphs for Visual Place Recognition

Elena Stumm[1,2], Christopher Mei, Simon Lacroix[1,3], Margarita Chli[4]

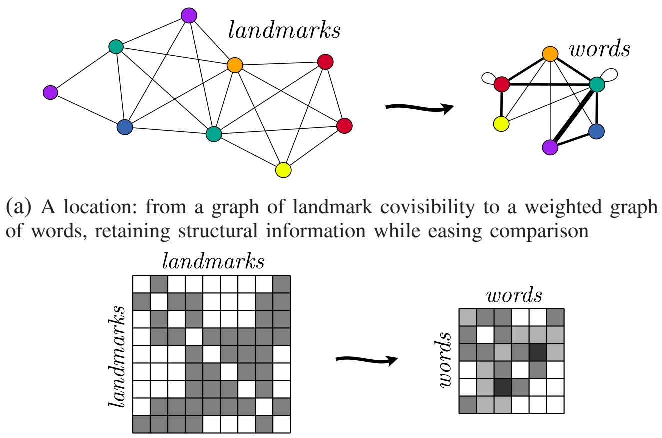
Abstract— **With the growing demand for deployment of robots in real scenarios, robustness in the perception capabilities for navigation lies at the forefront of research interest, as this forms the backbone of robotic autonomy. Existing place recognition approaches traditionally follow the feature-based bag-of-words paradigm in order to cut down on the richness of information in images. As structural information is typically ignored, such methods suffer from perceptual aliasing and reduced recall, due to the ambiguity of observations. In a bid to boost the robustness of appearance-based place recognition, we consider the world as a continuous constellation of visual words, while keeping track of their covisibility in a graph structure. Locations are queried based on their appearance, and modelled by their corresponding cluster of landmarks from the global covisibility graph, which retains important relational information about landmarks. Complexity is reduced by comparing locations by their graphs of visual words in a simplified manner. Test results show increased recall performance and robustness to noisy observations, compared to state-of-the-art methods.**

I. Introduction

As the robotics community is pushing towards the integration of robots in real scenarios, there is a growing need for robustness of operation. With vision-based approaches to Simultaneous Localization And Mapping (SLAM) gaining popularity due to the portability and applicability of image processing techniques, we have seen a range of impressive systems employing such approaches, from navigation of small aircraft [1] to the recently commercialized Dyson vacuum cleaner [2]. While locally, vision based navigation seems to have reached a certain maturity, appearance-based place recognition techniques still struggle for robustness.

With place recognition lying at the heart of the loop-closure and the kidnapped-robot problems, the computer vision and the robotics communities have been studying this challenge for a few years now. Current approaches vary between feature-based and global image representations. Full feature-based comparisons can be computationally heavy, and therefore most of the underlying structure and geometry between features is generally ignored, such as in the FAB-MAP framework [3]. As a result, such methods are prone to either perceptual aliasing or reduced recall, due to less discriminative observations. On the other hand, methods using global image representations, such as SeqSLAM [4], lack invariance and rely on using long sequences of images in order to escape perceptual aliasing.

?



(a) A location: from a graph of landmark covisibility to a weighted graph of words, retaining structural information while easing comparison

(b) Representing the location by its landmark adjacency matrix and word adjacency matrix

Fig. 1: Aiming to reduce the information encoded in images while preserving the most important visual and structural cues, each location, initially considered as a binary covisibility graph of visual landmarks, is converted to a weighted covisibility graph of visual words, where nodes correspond to unique visual words and edge weight to word covisibility count. In (1a), node colour represents the associated visual word, and the thickness of edges represents the relative weighting; (1b) shows the equivalent adjacency matrix representation, where cell shading represents the relative weighting. The word and landmark ordering used is arbitrary.

This paper examines structured comparison of locations based on appearance, in the context of mobile robotic place recognition. Considering a location initially as a constellation of visual landmarks, a graph is constructed, such that nodes are labeled with the associated visual words and edges correspond to binary covisibility of these words, i.e. connecting two nodes if the underlying landmarks have been co-observed in an image, as in [5], [6]. These location graphs reside as subgraphs in a larger *covisibility map* spanning the explored environment. Efficient comparison between two locations is achieved by converting their graphs of visual-landmark-covisibility, into weighted graphs of visual-word-connectivity (as shown in Figure 1) and comparing their corresponding sparse adjacency matrices. Working with graphs based on visual words rather than landmarks, we are able to circumvent the node alignment problem and exploit sparseness when evaluating matches, while maintaining more information than traditional bag-of-words techniques.

The complexity of the proposed method is $O(E)$, where E is the number of edges that two locations have in common, which is always less than or equal to $O(n^2)$, with n being the number of common words. The resulting approach requires no model parameters that need to be tuned, but relies on the

[1] CNRS, LAAS, 7 av du colonel Roche, F-31400 Toulouse, France; {stumm, cmei, simon} at laas.fr
[2] Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France
[3] Univ de Toulouse, LAAS, F-31400 Toulouse, France
[4] School of Informatics, University of Edinburgh, 10 Crichton Street Edinburgh EH8 9AB, U.K.; mchli at inf.ed.ac.uk

*This research was partly funded by the EC's Horizon 2020 Programme under grant agreement n. 644128 (AEROWORKS)

978-1-4799-6923-4/15/$31.00 ©2015 IEEE

good image representation to avoid aliasing.

input of sample images in order to avoid perceptual aliasing.

II. Related Work

A. Visual Feature Based Place Recognition

The introduction of visual bag-of-words techniques has allowed for efficient search and retrieval from vast amounts of images [7]. This technique relies on building a dictionary of visual words by clustering locally invariant feature descriptors, such as SURF [8], appearing in a set of model images and then representing each image as the set of visual words it contains. The use of this representation permits the analogous application of many theoretical developments such as tf-idf (term frequency×inverse document frequency) and probabilistic naive-bayes [9], from the fields of text retrieval and classification, on images [7], [3]. Such techniques apply well to place recognition for mobile robots, and are generally well established in the field, including extended generative models for location observations [10], [11], [6].

?? tf-idf 的运用?

One drawback of these approaches is that they discard most of the geometric information when comparing feature sets, therefore reducing the discriminative nature of the model and typically resulting in either perceptual aliasing or reduced recall. Following this realization, some previous works have investigated ways of incorporating some geometric information into the location models. For example, in [12] locations are represented by both visual landmarks and a distribution of the 3D distances between, given by range-finders or stereo cameras. Rather than using metric 3D data, [13] utilizes the relations between features in the 2D images to maintain a sense of geometry. This is done by extending the visual vocabulary to include a spatial dictionary as well the standard visual word dictionary.

3D implementation with distance in 3D

Alternatively, instead of using additional distance and position measurements, in [6] the implicit geometric relationships between features are given by covisibility information. Landmarks are tracked between successive images using a single camera, recording the covisibility between landmarks in a graph-based map of the world. The nodes in this graph represent landmarks, while edges indicate whether or not landmarks have been observed together in an image. The rationale is that when landmarks are consistently co-observed, it is an indication that these landmarks coexist in proximity in the structured physical world. As a consequence, by maintaining a sparse graph of landmark connectivity, this method retains a sense of geometrical structure between landmarks without requiring extraction and storage of exact position information. Covisibility is encoded in the same structure required for bundle adjustment and can thus be applied directly in Structure from Motion and SLAM frameworks.

Such a covisibility map has also been successfully used in [6] to dynamically extract and retrieve locations as clusters from the graph, which provide better matches to a query location than other methods relying on locations represented by fixed image frames. Building on this work, here we present a novel approach to efficiently incorporate more comprehensive structural comparisons between candidate locations, illustrating improvement in a variety of test cases as shown in Section IV.

B. Graph Matching

In the general case, the graph matching problem for undirected graphs is NP-hard. Finding node and edge correspondence is a combinatorial problem, which grows quickly with the number of nodes. In order to simplify this task and incorporate error tolerance, it is typical to use one of many inexact graph matching approaches.

One traditional method is graph edit-distance, which attempts to compute the minimum cost based on edit-operations (corresponding to deletion, insertion, substitution of nodes and edges) between two graphs [14]. It is important to note, however, that edit-distance relies on heuristic cost functions and finding the minimal edit-distance is still an NP-hard problem [15]. A more efficient method, is to work with the graph spectra rather than the graph itself, by decomposing a graph into the eigenvectors of the graph laplacian [16]. Spectral methods, however, have trouble coping with structural noise because the eigen-decomposition is sensitive to missing and spurious nodes [14].

Some review of graph kernel!

Graph kernels are a more recent and rather promising method of inexact graph comparison. Kernel methods can cope with non-linearities as well as allow graph structures to be used in traditional machine learning algorithms [14]. Essentially, graph kernels perform random walks across nodes to serialize the graph and then kernel methods can be used for classification [15]. Such random walk kernels can be computed at $O(n^3)$ [15], where n is the number of nodes in the graph, and have been implemented for applications related to computer vision [17] and scene characteristics [18]. However, $O(n^3)$ is still quite significant for larger graph sizes and limits tractability of online applications. As complexity is of great importance in place recognition and generally all processes concerning the navigation of a robot, this paper proposes a novel and more efficient method for inexact structural comparison between location graphs.

III. Location Graphs

Building on previous work in [6], we use the covisibility map of landmarks and their associated visual words to provide subgraphs, which serve as locations. Sections III-A and III-B provide a summary of this framework for completeness, followed by a description of novel location representations and comparison techniques in the remainder of this section. For more details regarding the framework of covisibility maps used, the reader is referred to [6].

A. The Covisibility Map

As the robot explores its environment, a global map $\mathcal{M}$ is maintained as a covisibility graph spanning all visual landmarks ever seen. Images are processed sequentially, detecting visual landmarks ℓ, which are represented by quantized visual words w from a dictionary $\mathcal{V}$ [7]. The dictionary is pre-trained using locally invariant descriptors such as SURF [8]. Features tracked across consecutive frames, are represented as the same landmark in the map. Whenever a landmark ℓ_i is observed together with another landmark ℓ_j, the two are connected by an edge E_{ij} in the graph. The map is implemented as a sparse matrix, and in addition, an inverted index of visual words and landmark observations is maintained for efficient look-up [7].

B. Location Retrieval

location retrieval (?use this instead?

Given a query location, the aim is to retrieve similar locations from the map and evaluate whether two locations actually correspond to the same place. Location retrieval is facilitated by the inverted index of visual words, which allows relevant cliques in the covisibility map to be efficiently identified. Once found, these cliques are expanded based on landmark connectivity to produce location subgraphs known as virtual locations [6]. Each virtual location is represented by its local covisibility graph of landmarks and their associated visual words (a simplified example of a virtual location is shown on the left side of Fig. 1). The candidate virtual locations retrieved in this process resemble the query at least partially, but still need further evaluation, which is done in the probabilistic framework as explained below.

C. Observation Graphs

The first step in evaluating candidate virtual locations is to compute the likelihood of the observations coming from the same place. This likelihood can be estimated by a normalized similarity score between the query and the candidate locations. As discussed in Section II, obtaining a score based on graph similarity and matching is generally a very complex optimization problem, primarily due to the graph alignment problem. In order to address this, we approximate location graphs of landmarks by their corresponding visual word graphs. This means that the graph consists of nodes representing visual words from the dictionary, rather than landmarks directly from the map. Working in the space of visual words allows the algorithm to bypass the alignment problem when comparing locations, as nodes can be easily matched one-to-one.

Figure 1 depicts the difference between graphs of landmarks and graphs of visual words. Nodes in the word graph correspond to words appearing in the location of interest, while edges are weighted according to the connectivity count between words (can be easily retrieved from the landmark graph). The corresponding word adjacency matrices are implemented as sparse matrices, and Algorithm 1 outlines the conversion process in pseudocode.

D. Graph Comparison

Following the representation of the query and the candidate virtual locations by their visual word graphs as described above, likelihood values can be formulated in a relatively straight-forward manner. Here, we assume that the sparse normalized cross-correlation between location adjacency matrices represents the observation likelihood $P(\mathcal{Z}|\mathcal{L})$ of an observation $\mathcal{Z}$ given a location $\mathcal{L}$, giving a parameter-free

Algorithm 1 Conversion between landmark and word adjacency matrices

```
norm = 0
n = 0
for i in range(num_landmarks):
    for j in range(i):
        # store row, column, and data values
        # in vectors, for efficient sparse
        # matrix creation
        # (only fill half sym. matrix)
        if landmark_adj[i, j] != 0:
            row[n] = min(landmark_words[i],
                         landmark_words[j])
            col[n] = max(landmark_words[i],
                         landmark_words[j])
            data[n] = landmark_adj[i, j]
            norm += landmark_adj[i, j]
            n += 1
# create sparse matrix
# (which sums duplicate entries)
word_adj = sparse(row, col, data)
# normalize:
word_adj /= norm
```

1 3 1 4 5 3 4 —
2 5 1 6 7 9 4 —
0 1 0 1 0 0 1 — -

将row与col具相乘且将相同的项相加.

approach, as shown in Equation (1).

$$P(\mathcal{Z}|\mathcal{L}) \approx \frac{\sum_{\{\mathcal{E}\}} E_{uv}^{\mathcal{Z}} \cdot E_{uv}^{\mathcal{L}}}{\sqrt{\sum_{\{\mathcal{E}\}} (E_{uv}^{\mathcal{Z}})^2 \sum_{\{\mathcal{E}\}} (E_{uv}^{\mathcal{L}})^2}} \tag{1}$$

where $E_{uv}^{\mathcal{Z}}$ and $E_{uv}^{\mathcal{L}}$ represent the edge weights between words w_u and w_v from the query observation and candidate location respectively, and $\{\mathcal{E}\}$ is the set of possible edges based on the dictionary $\mathcal{V}$. This calculation explicitly takes into account the presence of the edges in the graph, whereas methods like tf-idf only look at the presence of words while ignoring their connectivity. Using the likelihood calculation, the posterior probability of being in a location given the observation is subsequently given by Bayes' rule as follows,

$$P(\mathcal{L}|\mathcal{Z}) = \frac{P(\mathcal{Z}|\mathcal{L})P(\mathcal{L})}{P(\mathcal{Z}|\mathcal{L})P(\mathcal{L}) + P(\mathcal{Z}|\bar{\mathcal{L}})P(\bar{\mathcal{L}})} \tag{2}$$

where $P(\mathcal{Z}|\bar{\mathcal{L}})$ is calculated analogously to the observation likelihood, using a set of sample locations to represent the unknown world.

Due to the sparsity of visual words in each location, computing cross-correlation scores requires relatively few calculations, as only words common to both locations are involved in the numerator of Equation (1). This implies that the complexity of this computation is typically substantially less than $O(|\mathcal{Z}|^2)$, where $|\mathcal{Z}|$ is the number of words in the observation, as it depends on the number of common edges in the graph. However, even though only a subset of words from each location are involved in this sum, *all* words in a location have an impact on the final score, since edge weights are always normalized across the location (see Algorithm 1).

The likelihood of the query coming from another location $P(\mathcal{Z}|\bar{\mathcal{L}})$ can be calculated analogously using a set of sample locations as follows (see [6] for details),

$$P(\mathcal{Z}|\bar{\mathcal{L}}) \approx \sum_{s=1}^{N_s} \frac{P(\mathcal{Z}|\mathcal{L}_s)}{N_s} \tag{3}$$

with N_s corresponding the number of samples, and $\mathcal{L}_s$ to the s^{th} sample location. Note that the terms in the denominator of Equation (1) only need to be computed once for each location, which is of great importance in the case that the query and sample locations which are typically involved in multiple comparisons.

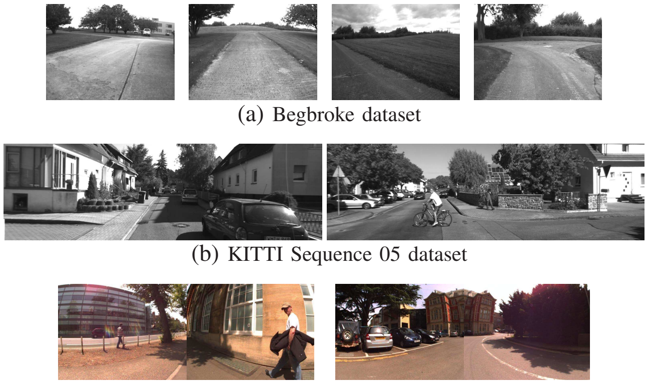
E. Matrix Weighting and Relation to tf-idf

As some words and edges occur more commonly than others, they provide different amounts of information about the location. This concept is well recognized in the text analysis field, and therefore terms are generally weighted according to a prior on their frequency, which is provided by known documents. As a result, more common terms tend to have less impact on the final results than rare terms. In the context of place recognition, an intuitive example is provided by comparing a brick wall to a statue. Since bricks are seen throughout cities, bricks would have a relatively low weighting, as they do not provide much context about the location, whereas the statue is unique and provides much more contextual information.

In the example of the commonly used tf-idf scoring method, each word is given by a value proportional to the number of times it was seen in that document (*term frequency*) and inversely proportional to the number of other documents which contained the same word (*inverse document frequency*). Each document is then represented by a vector of all its tf-idf word values, and similarity is given by the dot product between document vectors [7]. In this work, the word adjacency matrices can be viewed analogously to tf-idf vectors, comparing word connectivity (edges), rather than individual words. Note that working with this word connectivity, rather than a traditional bag of words, is the important factor for retaining structural information encoded in the observations. Each word adjacency matrix is then weighted according to the relative information content $-\ln P(E_{ij})$ provided by each edge, which is precalculated based on prior probabilities from a set of sample locations. Further study on the interpretation of tf-idf weighting can be found in [19].

IV. Experiments and Results

The proposed framework (hereby referred to as ‘Graph Covis’) was evaluated using real-world datasets from urban environments and compared to state-of-the-art methods. Each dataset consists of a stream of monocular images, with varying image characteristics and frame-rates between datasets. Examples of images from each dataset are shown in Figure 2, along with a brief overview in Table I, giving a sense of the environment and sequence parameters. The Begbroke dataset corresponds to the one used in the work of [5]; while the KITTI dataset is the fifth sequence from the odometry benchmark sequences, provided by [20]; and the City Centre dataset originates in the work of [3].



(a) Begbroke dataset

(b) KITTI Sequence 05 dataset

(c) City Centre dataset

Fig. 2: Example images from the datasets used for evaluation.

Dataset Name	Description	Sequence Length	Image Spacing	Image Specs
Begbroke	3 loops around a path surrounded by fields, trees, buildings, and cars.	approx. 1km, 1000 images	approx. 1m	greyscale, 512 × 384px
City Centre	University campus with many buildings, cars, roads, gardens, and people.	approx. 2km, 1200 images	approx. 1.6m	colour, 640 × 480px each
KITTI Seq. 05	Urban dataset containing mostly roads, houses, trees, and cars.	approx. 2.3km, 1400 images	approx. 1.6m	greyscale, 1226 × 370px

TABLE I: Overview of datasets used for testing.

During testing, each dataset is incrementally traversed, building a map over time and using the most recent location as a query on the current map, with the goal of retrieving any previous instances of the query location from the map. Precision-recall results for the three datasets are shown in Figure 3, comparing the method described in this paper (Graph Covis), to that of previous work which uses covisibility for location extraction but discards structure for a bag-of-words comparison (Naive-Bayes Covis) [6], and that of FAB-MAP which works with single-image locations and no location graphs [3]. For completeness, we also include results from the commonly used SeqSLAM framework [21], although the approach differs drastically from the one presented here. In this work, positive loop-closures are given by locations, which contain landmarks from within a given radius of the query location. The radius used for evaluation was set to $8m$, as errors in ground truth labels can reach several meters, and image spacing is frequently as far as $2m$. Each framework was provided with the same set of sample locations, which consist of images from streetview locations and other datasets (excluding the tested dataset). In addition, as is typically done during testing, no data associations were made based on loop closures [3], [6]. The same visual dictionary containing 10000 words (provided by [3]) was used for the implementation of all feature-based methods.

From Figure 3, one can see a general improvement from utilizing the structural information for both location extraction and comparison. Improvements are minor compared to the Naive-Bayes Covis framework on the Begbroke sequence, since the recall is near perfect already. Results are more significant in the other two datasets, where the Graph Covis framework provides a more significant boost in recall rates. Note that performance is lower on the City Centre dataset, as images tend to contain less overlap in features,

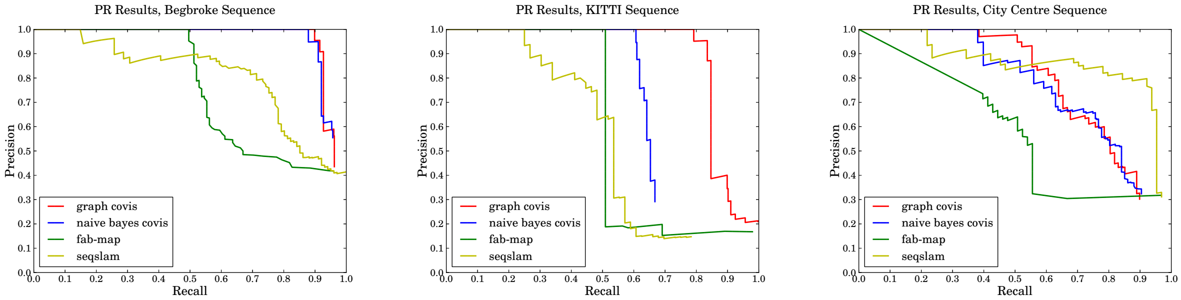
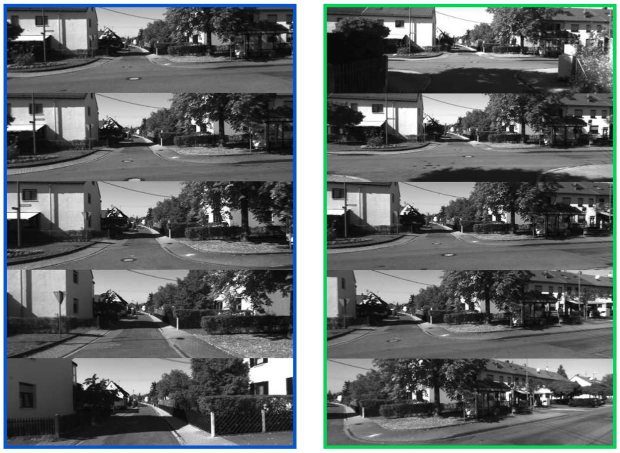
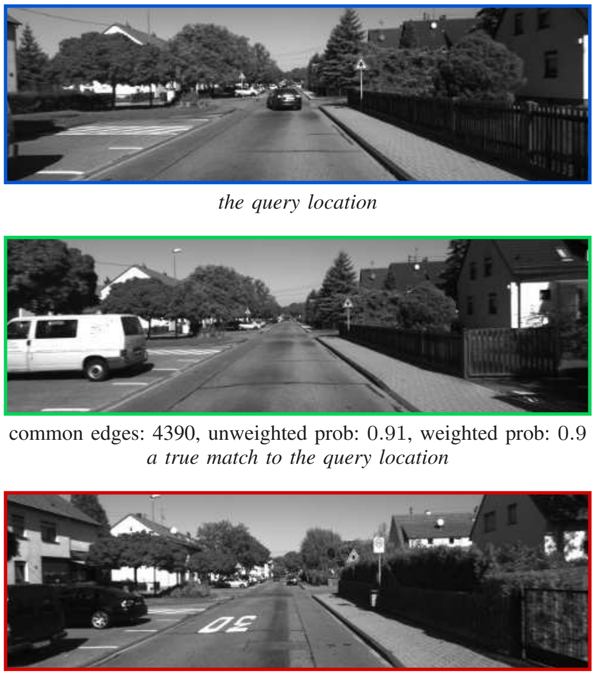


Fig. 3: Precision-recall results for the Graph Covis framework presented in this paper, the Naive-Bayes Covis method [6], and the FAB-MAP method [3], on three different datasets.



the query location *the candidate location*

Fig. 5: A query and a candidate location from the KITTI dataset shown by five representative images. The two locations are a true match, obtained from the same intersection; the query passing straight through and the candidate location turning right. Our graph-based covisibility framework assigns a match probability of 0.91, while the unstructured framework assigns a probability of 0.02.



the query location

common edges: 4390, unweighted prob: 0.91, weighted prob: 0.9
a true match to the query location

common edges: 3451, unweighted prob: 0.83, weighted prob: 0.66
a false match to the query location

Fig. 6: Example of a query and two retrieved locations from the KITTI dataset.In this case, both of the retrieved location graphs share many common edges with the query. As a result, the match probabilities when using unweighted term frequencies are high for both locations. However, when using term frequencies, which are weighted by the relative document frequencies from sample locations, the match probability of the false candidate drops significantly.

reducing the quality of the covisibility map built from them, and limiting the improvements from the proposed method. In addition, the City Centre sequence generally contains more variations and ground truth errors than the other two datasets. In the FAB-MAP framework, the probability normalization model differs compared to that in Equation (2), reducing recall significantly in the presence of multiple instances of the same location in the map (see [3] and [6] for more details). Furthermore, the use of single-image, bag-of-words location models limit the results in comparison to both the Naive-Bayes and Graph Covis frameworks.

In order to investigate the robustness of the location graph models, noise is incrementally added to locations and the behaviour is shown in Figure 4a. Taking independent locations from the KITTI dataset and adding varying amounts of noise, the noisy version is compared to the original location, plotting the resulting boxplots of the posterior match probabilities. Noisy locations are created by corrupting a certain percentage of the words associated to the location's landmarks, randomly swapping them with another word from the dictionary. This process implicitly alters the edge structure of the corresponding word adjacency matrices. Figure 4a also shows the highest posterior match probability achieved by a false loop-closure from the same dataset with grey shading, indicating the line above which perfect precision would be maintained. Figure 4b shows the results of the same experiment run using a traditional tf-idf comparison method, which uses words with no added edge structure. Note that the probabilities are significantly lower in this plot because locations are less discriminative under a structureless model, reducing the scores which are normalized using sample locations. Together, these plots illustrate the benefit of structured comparison. While the boxplot whiskers remain above the false-positive threshold up to 60% of added noise when using structured comparisons, this is only the case up to 30% of added noise when using unstructured comparisons.

A reprentative example of the improved recall of our method can also be seen in Figure 5, where a query and candidate location are compared using both the structured Graph

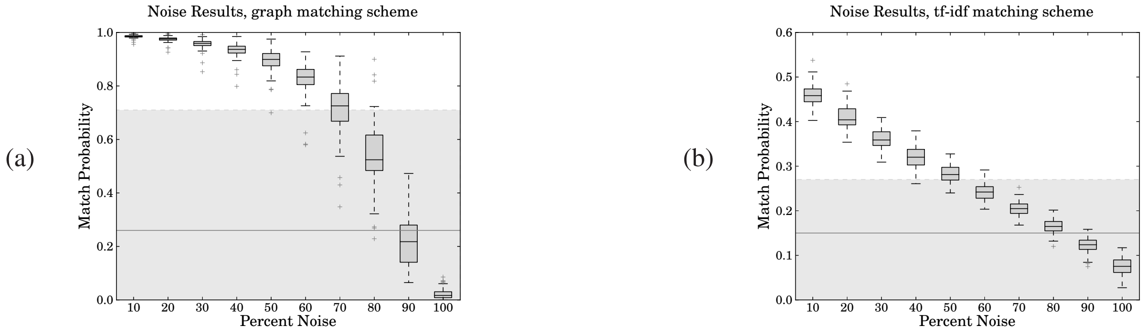


Fig. 4: Statistics for location similarity as the visual words of locations are corrupted by noise and compared to their original state. The grey shaded areas show the scores where the first false-positive from other retrieved locations occur, and the black lines show the median scores for these false-positives.

Covis and unstructured Naive-Bayes Covis frameworks. The two locations represent the same intersection, only traversed in different ways. This difference in traversal introduces enough differences to the word sets of each location for the unstructured method to assign a low match probability, while word connectivity remains consistent enough for the structured method to provide a relatively high probability.

The importance of edge weighting is clearly illustrated in Figure 6, which shows a query and two candidate locations (one matching and one false), along with the posterior match probabilities provided from using both unweighted and weighted word adjacency matrices as described in Section III-E. In this case, the unweighted probabilities remain fairly high for both locations, as they have a similar appearance and share a similar number of common edges with the query. When weighting the word adjacency matrices based on the edge frequencies in sample locations, however, the importance of frequently occurring edges is down-weighted, reducing the probability of the false location.

V. Conclusion

This article has presented a method for visual place recognition which exploits the covisibility of landmarks to account for geometric structure on top of appearance, in the search for matching locations. Employing the graph-based covisibility representation of locations introduced in [5], a novel approach for efficient graph matching is presented for comparing locations, and is demonstrated to outperform state of the art methods in place recognition. While only loose structural information is encoded in the initial covisibility graph, we show that this is enough to disambiguate across appearance-only matches, helping tackle perceptual aliasing, which is a common problem in existing methods. Our thorough evaluation on a variety of types of scenery and variable presence of noise reports increased recall at perfect precision. The added complexity of incorporating geometric cues, is minimized by employing efficient workarounds inspired by both graph-matching and place recognition literature.

Future work will focus on further investigation into the trade-off between computational complexity and richness of information encoded in various location models, and the effect they have on quality of location recognition.

References

[1] S. Weiss, M. W. Achtelik, S. Lynen, M. C. Achtelik, L. Kneip, M. Chli, and R. Siegwart, "Monocular Vision for Long-term MAV Navigation: A Compendium," *Journal of Field Robotics*, vol. 30, 2013.

[2] E. Ackerman. (2014) Dyson's robot vacuum has 360-degree camera, tank treads, cyclone suction. [Online]. Available: http://spectrum.ieee.org/automaton/robotics/home-robots/dyson-the-360-eye-robot-vacuum

[3] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *International Journal of Robotics Research*, 2008.

[4] M. J. Milford, "Vision-based place recognition: how low can you go?" *International Journal of Robotics Research*, 2013.

[5] C. Mei, G. Sibley, and P. Newman, "Closing loops without places," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[6] E. Stumm, C. Mei, and S. Lacroix, "Building location models for visual place recognition," December 2014.

[7] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," *International Conference on Computer Vision*, 2003.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," *Computer Vision and Image Understanding*, 2008.

[9] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, USA: Cambridge University Press, 2008.

[10] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *International Journal of Robotics Research*, 2011.

[11] A. Angeli, D. Filliat, S. Docieux, and J.-A. Meyer, "A fast and incremental method for loop-closure detection using bags of visual words," *IEEE Transactions on Robotics*, vol. 24, no. 5, 2008.

[12] R. Paul and P. Newman, "FAB-MAP 3D: Topological mapping with spatial and visual appearance," in *IEEE International Conference on Robotics and Automation*, 2010.

[13] E. Johns and G.-Z. Yang, "Feature-co-occurence maps: Appearance-based localisation throughout the day," in *IEEE International Conference on Robotics and Automation*, 2013.

[14] H. Bunke and K. Riesen, "Towards the unification of structural and statistical pattern recognition," *Pattern Recognition Letters*, 2012.

[15] K. M. Borgwardt, "Graph kernels," Ph.D. dissertation, Ludwig-Maximilians-Universität München, 2007.

[16] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, 2007.

[17] Z. Harchaoui and F. Bach, "Image classification with segmentation graph kernels," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[18] M. Fisher, M. Savva, and P. Hanrahan, "Characterizing structural relationships in scenes using graph kernels," in *ACM Transactions on Graphics*, vol. 30, no. 4, 2011.

[19] D. Hiemstra, "A probabilistic justification for using tf× idf term weighting in information retrieval," *International Journal on Digital Libraries*, vol. 3, no. 2, 2000.

[20] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research*, 2013.

[21] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE International Conference on Robotics and Automation*, 2012.