Visual Localization in Highly Crowded Urban Environments

A. H. Abdul Hafez[1,2], Manpreet Singh[1], K Madhava Krishna[1], and C.V. Jawahar[1]

***Abstract*— Visual localization in crowded dynamic environments requires information about static and dynamic objects. This paper presents a robust method that learns the useful features from multiple runs in highly crowded urban environments. Useful features are identified as distinctive ones that are also reliable to extract in diverse imaging conditions. Relative importance of features is used to derive the weight for each feature. The popular Bag-of-words model is used for image retrieval and localization, where query image is the current view of the environment and database contains the visual experience from previous runs. Based on the reliability, features are augmented and eliminated over runs. This reduces the size of representation, and makes it more reliable in crowded scenes. We tested the proposed method on data sets collected from highly crowded Indian urban outdoor settings. Experiments have shown that with the help of a small subset (10%) of the detected features, we can reliably localize the camera. We achieve superior results in terms of localization accuracy even when more than 90% of the pixels are occluded or dynamic.**

I. Introduction

The localization problem tries to provide an answer to the question "Where am I?". In other words, it is the process of computing the current pose of robot in the environment [1], [2]. Appearance-based localization [2], [3], [4] is a variant of the popular content-based image retrieval. The query image, in the case of robotic localization, is the current view as seen by the robot, and database is the past visual experience. Images are represented using local or global features. Utility of the global visual features was explored for mobile robot exploration, navigation, and localization [5], [6]. Local features [7] are possibly more effective and can be computed at interest points on the image. However, the number of descriptors as well as the computations required to match explode with large databases. To handle this problem, Sivic *et al.* [8] quantized the features into a set of visual words, popularly known as the Bag-of-Words technique. It is used along with techniques such as the Inverted Index and Min-Hash [9] for fast matching and retrieval of images.

Localization in outdoor environments is challenging because, queries are often from different viewpoint, scale and illumination than the previous visual experience [3]. In crowded and cluttered outdoor settings, the problem is further challenging [10]. Dynamic objects could constitute most of the visible region and thus complicate the localization (see Fig. 1). The results are often erratic since the robot may not be able to distinguish the features from static and dynamic parts of the image. We are interested in designing a robust localization solution in highly crowded urban environments. For this purpose, we collected multiple runs of data from crowded Indian roads. We aim at learning to localize the camera better by finding out reliable and useful features.

Fig. 1. Change in dynamic objects with time. The four images were captured at the same pose at different instances from a forward facing camera fixed to moving vehicle.

Fig. 2. This figure depicts the extent of occlusion and how much our dataset (first row) is crowded in comparison with dataset used by [10] (second row).

Knopp *et al.* [11] developed an automatic method to detect and suppress confusing (useless) features. The method significantly improved the performance and reduced the database size. Turcot and Lowe [12] reduced the database size by selecting a small portion of the total detected features. They call this subset as 'useful' features.

Cummins and Newman [4], [13] describe a probabilistic framework, called FAB-MAP, for recognizing places based on their visual appearance. Milford and Wyeth [3] presented a new approach, called SeqSLAM, for visual localization under different environments and illumination conditions. However, both the methods presented above, i.e. FAB-MAP and SeqSLAM, are not tested in highly crowded urban environments. Also, they neither use the concept of useful features nor identify dynamic objects in the scene.

Achar *et al.* [10] investigated the problem of localization in urban environments. They proposed a novel method to

[1] International Institute of Information Technology, Gachibowli, Hyderabad-500 032, India.

[2] Dept. of Computer Engineering, Hasan Kalyoncu University, 27100 Sahinbey, Gaziantep, Turkey.

identify dynamic scene elements from the base run, and filter out features obtained from dynamic elements to facilitate robust localization of the robot. In contrast, a highly crowded environment containing dynamic elements, like pedestrians, movable objects and parked vehicles is considered in our work as shown in Fig. 2. Such elements may not always be present at the same pose in the environment, and hence, can misguide the localization process of robot.

We propose a method to learn the elements that are "truly" static for a given scene and hence, improve the localization performance. These useful features are learnt incrementally over a period of time. We assume that every time the robot runs through a certain spatial locality, it learns new features with the possibility to discard a part of features which were learnt during the previous runs. Over time, the number of features converges to very small portion of the total features. This removal of non-useful features reduces the memory and computational requirements.

The remainder of this paper is organized as follows. Section II presents an overview of the proposed localization method, while Section III focuses on learning useful features when a new training run is available. Section IV explains the experiments and results to support our proposed localization method. Finally, we conclude with some remarks and proposals for future work.

II. Qualitative visual localization

Qualitative visual localization uses past experiences of the robot to determine a set of images in the close neighborhood of its current pose. Our work is motivated by the fact that a majority of the extracted visual features belongs to the dynamic environment. Many recent works [10], [11], [12] have argued that only a small percentage of the extracted features are useful. These however have not been tested in highly crowded environments, which is necessary for future autonomous systems. Our proposed framework is a localization methodology for these highly crowded urban environments through multiple experiences. Our work proposes that the system's ability to learn the useful subset of features can be enhanced by traversing the same path multiple times.

The interest points, called features are extracted from the available images. Bag-of-words method [8] has been observed to be highly efficient in image retrieval. The method represents the images as a set of unordered visual words. These are used to build an inverted index which quantifies the occurrence of each feature. The features of a query image, captured at the robot's current pose are quantized using the existing vocabulary tree and mapped to a set of visual words.

The robot explores the environment during a training run and captures the frames referred to as the ground truth data. SIFT features [7] are extracted from the key frames and a vocabulary tree of K branches and L depth containing K^L leaf nodes is constructed using hierarchical k-means clustering [14]. For each visual word, the inverted index stores the occurrence of a feature and maintains its history through weights. The utility of a feature has been quantified by tf-idf [8] in some earlier works, but in this paper we use the weighted retrieval method which assigns higher weight to highly discriminative, *i.e.* useful and distinctive features while lower weight to others.

A. Distinctive features

The distinctiveness of a given feature z with respect to the robot's pose x is a measure of the information that is added to our knowledge about the pose. Assume that our knowledge about the robot pose is represented by the distribution $P(X)$, while the amount of information given by the measurement is $P(Z)$. This can be interpreted as looking for features which appear in all the images about some specific robot pose, but rarely appear elsewhere. The information gain captures this concept of distinctive features.

Information gain $I(X|Z) = H(X) - H(X|Z)$ is a measure of how much uncertainty is removed from a pose distribution $P(X)$ given some specific additional knowledge about features $P(Z)$. It is defined using the entropy $H(X)$ and conditional entropy $H(X|Z)$ of distributions $P(X)$ and $P(X|Z)$ respectively. The information gain is always considered with respect to a robot's specific pose, x_i and a specific feature, z_k. In other words, the distinctive weight, w_k of the kth feature in the vocabulary tree is computed as:

$$w_k = I(x_i|z_k) = H(x_i) - H(x_i|z_k). \tag{1}$$

More details on computation of the distinctive weight, i.e. information gain, is available in [2], [10], [15], [16].

B. Querying

When the robot explores a previously observed environment, it comes across unknown views which are called query images. The extracted features from each query image are quantized to visual words using the vocabulary tree and inverted index. The greedy N-best paths search [2] is used to improve the quantization. A score is calculated for each indexed image by searching for visual words corresponding to the query image. The normalized score for the few relevant images is computed as:

$$Score(img_n) = \frac{\sum_{z_k \in Z_n \cap Z_q} W_k}{|Z_n|}, \tag{2}$$

where $|Z_n|$ is the number of elements in the SIFT descriptors set Z_n in the nth key frame of base run, Z_q is the set of SIFT descriptors in the query image, and W_k is the total weight of the kth feature, z_k computed later in Eq. (6). The top N images based on the normalized scores are returned as the best matches for direct image retrieval.

In the case of global localization, the feature correspondences between top matches and the query image are geometrically verified using epipolar geometry and results are filtered based on the number of inliers. The weighted bag of words retrieval method, however suffers from some disadvantages in crowded urban environments. The inclusion of features corresponding to dynamic elements in the environment leads to lower localization accuracy. Our work addresses the problem by identifying the useful features of

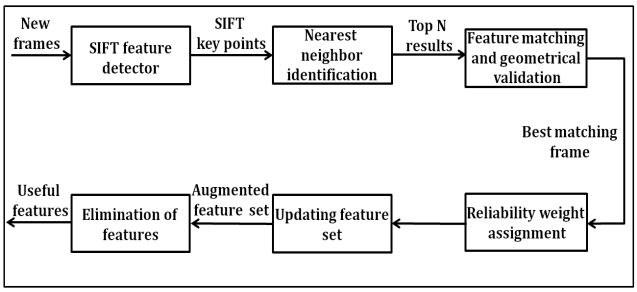

Fig. 3. Pictorial representation of the learning process. The input to the system are newly observed frames and output are the useful features.

environment through multiple experiences. This not only reduces the size of vocabulary tree but also improves the performance after each experience.

III. Learning Useful Features

Factors such as illumination, clutter and weather conditions play a key role in the performance of visual localization methods. By traversing the same path multiple times, the effect of these factors is minimized and the probability of observing the useful features is increased. Thus, the expectation is that with each new run, the robot enhances its knowledge of the environment. Figure 3 depicts a block diagram of the learning process. From each available frame in the current run, SIFT key points are extracted. A neighborhood is defined in which the best frame is searched through feature matching followed by geometrical validation through epipolar geometry. The repetitive features (useful features) are assigned reliability weights whereas the new features are augmented to the feature set with equal initial weight.

Every training run is classified as either base run in which the robot experiences the environment for the first time, or augmentation run in which the knowledge about the environment is enhanced through learning new and useful features, or elimination run in which the useful features are retained based on their reliability weights while the non-useful features are eliminated.

A. Initial reliability weights in base run

The robot explores the environment for the first time and assigns equal reliability weight to each feature. The initially extracted M_1 features are assigned reliability weight, $w_k^1 = 1$, where w_k^t refers to the kth feature after tth run. The normalized weight of features in the base run is given as $\hat{w}_k^1 = \frac{1}{M_1}$.

B. Augmentation of features

As the robot explores the environment again under different conditions, it observes a set of features which it may or may not have seen previously. The best visually matching frame from the current run corresponding to every frame of the base run is determined. Matching all the newly observed features with existing set may not be a realistic approach. Hence, using the available GPS data, a set of $\mathcal{K}$ images

Fig. 4. The left image shows the visually stable features (matching static and dynamic elements) while the right image shows the spatially consistent features (static elements of environment).

from the current run is determined, which in turn represents nearest neighbors of the image belonging to the base run.

A number of features from the selected best matching image are considered as matches to the features from the base image [17]. The matching visual features (see Fig. 4) contain the previously observed static features as well as features belonging to the dynamic objects. These features are assigned the visual stability weight given as:

$$w_{vs} = \frac{\mathcal{F}}{\mathcal{N}}, \tag{3}$$

where $\mathcal{N}$ is the number of visual features in an image of the base run, and $\mathcal{F}$ is the number of feature correspondences between an image of the base run and its best matching image from the current run.

The feature correspondences contain many negative results, dynamic objects which are discarded by identifying spatially consistent features (see Fig. 4). These are simply the inliers estimated by fitting a fundamental matrix using RANSAC algorithm and assigned the spatial consistency weight as:

$$w_{sc} = \frac{\mathcal{G}}{\mathcal{N}}, \tag{4}$$

where $\mathcal{N}$ is the number of features in an image of the base run, and $\mathcal{G}$ is the number of geometrically validated matches.

Both the equations 3 and 4 are indicative of similarity between the elements of environment observed at the same pose during different training runs. Thus, higher number of feature correspondences ($\mathcal{F}$) or geometrically validated matches ($\mathcal{G}$) imply reoccurrence of a larger number of static elements in the two views and vice-versa. Therefore, every matching feature of an image from the base run is assigned an equal weight which is a function of the similarity between images. Apart from the matching features, a number of features are newly observed in the current training run. These features are assigned a reliability weight, w_k equal to the minimum weighted feature during the particular run and augmented to the useful feature set.

C. Enhancing the knowledge through elimination

Since only the useful features would reoccur at the same pose under different conditions and gain higher reliability weight, augmenting the features that belong to dynamic objects decreases the efficiency of localization process. Thus, the features with reliability weights greater than the minimum weighted feature are retained while the remaining are eliminated (see Fig. 5). To ensure that a visual feature is

Fig. 5. Gradual learning of the useful features through base, augmentation and elimination run. Observe the growth in the number of features at a particular pose (from left to right) over run 1, 2, 3 and finally retention of only useful features in the last frame. Note that a small number of features in the last frame correspond to non-distinctive elements like sky. The effect of these features is negated using distinctive weights.

not discarded incorrectly, we continue to augment the newer features till the third run. Though expensive, the idea is to provide new features sufficient support to prove their usefulness. The features unable to prove their importance to the process are then eliminated. Similarly, the newly observed features in the third and subsequent runs are retained for future explorations till their usefulness to localization process is verified. A certain percentage of the useful features may remain hidden for most of the training runs. These features, though distinctive are not useful for localization since their visibility in a query is highly uncertain. Such features are labeled as non-useful and eliminated due to low reliability weight.

After t runs, a kth useful feature will have the reliability weight as:

$$w_k^t = \hat{w}_k^{t-1} + w_{vs} + w_{sc}. \tag{5}$$

Then, it is normalized to produce the weight, $\hat{w}_k^t = w_k^t/(\sum_{k=1}^{M_t} w_k)$, where M_t is number of features after the tth run.

The total weight assigned to the kth feature of the base run after t training runs is given as:

$$W_k = w_1.w_2, \tag{6}$$

where $w_1 = I(X;Y)$ as in Eq. 1 and $w_2 = -\frac{1}{log(\hat{w}_k^t)}$ as in Eq. 5.

Here, the total weight of the feature is directly related to its reliability and distinctiveness. For the features that are incorrectly labeled as useful, such as sky in Fig. 5, the reliability weight is negated with low distinctiveness weight.

D. Limits of Learning

The robot learns the environment gradually with each run through augmentation and elimination of features. Since significant structural changes do not occur in an urban environment within a short span of time, the number of useful visual features is limited. Hence, it is fair to assume that the system's incremental learning ability will reduce with each training run. Later, it will be experimentally shown that the mean error saturates after fifth run (Fig. 7). Rather, in some experiments, it increases slightly due to incorrect labeling of non-distinctive reoccurring features such as sky as useful features. The improvement in the localization performance with multiple experiences for a robot is closely related to the accuracy of the GPS receiver which is used as ground truth data for verification.

IV. Experiments and Results

A. Dataset creation by applying learning principle

The visual environment forming the dataset was captured in the highly cluttered Koti area of Hyderabad, India using a forward facing digital camera attached to a moving vehicle. The 640×480 resolution data containing 68700 frames was recorded at 30 fps under varying conditions of traffic and illumination. The key frames were obtained by sampling the training data at 10 fps; the base run containing 2596 frames. The motion blur and unpredictable motion of the traffic added to the uniqueness of collected data. A GPS receiver of permissible error was used to record the current position of the vehicle for later use as ground truth data.

The number of features increases linearly with each training run and if allowed to continue would affect the computation time without improvement in the localization accuracy. By applying the proposed elimination method, the number of features was reduced by 90.99%, 87.54% and 82.37% after third, fifth and seventh run respectively. This highlights the effectiveness of our approach in retaining only the small subset of useful features and eliminating the rest.

B. Testing

Training video from the eighth run of the same environment was sampled at 10 fps to obtain 3734 key frames. Though there was a large variation in the number of useful features with each training run (Table I), the size of the vocabulary tree was kept constant at 537K without any observed degradation in performance. As already explained in II-B, the inverted index of the quantized visual words was used for searching matches and the score was assigned to each database frame. The top 10 results based on the score were returned as best matches and labeled as direct retrieval results. For global localization, the matched features from the top results were geometrically verified using epipolar geometry to remove the mismatches. The best matching frame was obtained based on the number of geometrically validated features.

GPS ground truth data was used to check the localization performance by measuring the mean error for both direct retrieval and global localization. The returned results were deemed to be correct if the localization error was less than 7.5m. The proposed formulation was tested on three datasets. For all the tests, the query images were randomly chosen assuming independent and identically distributed samples.

TABLE I

VARIATION IN THE NUMBER OF VISUAL FEATURES WITH EACH TRAINING RUN. THE NUMBER OF USEFUL FEATURES RETAINED BY SYSTEM AFTER ELIMINATION ARE ALSO SHOWN. THE LARGE REDUCTION IN NUMBER OF FEATURES IS INDICATIVE OF THE AMOUNT OF CLUTTER PRESENT.

Training Run	1	2	3	3-E	4	5	5-E	6	7	7-E
Number of features	2.90M	5.53M	8.15M	734k	6.03M	8.67M	1.08M	6.29 M	8.74M	1.54M

Fig. 6. Sample queries from the eighth training run. The repeating dynamic elements, pedestrians and the clogged roads influence the localization performance.

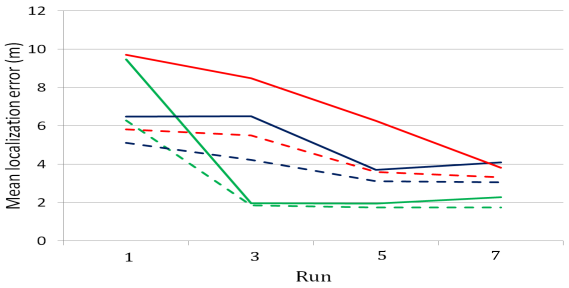

Fig. 7. Variation of mean localization error for both direct retrieval (solid lines) and global localization (dashed lines) corresponding to the three tests is shown. The performance improves with each training run. In two cases, a slight increase in mean error is observed for seventh run which can be attributed to unintentional learning of dynamic features by the system. (Green=Test 1, Blue=Test 2, Red=Test 3)

1) Test 1: The first 500 key frames from all the eight runs were chosen. The frames from the first seven runs were used to build the database and a random 200 frames from the eighth run were used as query.

2) Test 2: To check for accuracy over larger number of frames, the database of useful features was built using all the key frames from the first seven training runs. A random 1000 key frames were chosen from the eighth run as query.

3) Test 3: The database of useful features was built similar to test 2 but the full eighth run, i.e. 3734 key frames covering the complete path were used as query.

C. Discussion

Despite occlusion of useful features (see Fig. 6), the results have shown good visual localization performance. As can be seen from Fig. 8, better performance was achieved after fifth training run with only 30% visual features than after first run with all the features [10]. In our case, even the used 30% features contained a major portion belonging to sky and roads. Therefore, the actual available useful features were far less, on an average 10%. Over the full eighth run, our formulation returned results similar in appearance to the query and in close neighborhood of the query pose.

The quantitative analysis (Table II) in terms of the distance measure has shown considerable improvement. The direct retrieval and global localization results improved by 60.72% and 43.12% respectively after seventh run for Test 3 (Fig. 7). The reduction in variance is much larger, i.e. 95.53% and 98.46% respectively for both the cases. Considering the inherent GPS error and usage of only 10% features, the current results are extremely positive and support our approach to learn the useful features through multiple experiences.

Fig. 8. The first row shows the query frames. The second row shows the retrieved results using the formulation of [10] whereas the third row shows the retrieved results using our proposed method after fifth run using only useful features.

Fig. 9. Shown are the matching features from two sample frames returning large errors. The matching features correspond to non-distinctive elements like trees, sky and lane markings.

Twelve out of 3734 query frames returned extremely large errors (greater than 20m) that can be observed both visually (Fig. 9) and quantitatively (Fig. 10). The matching features in these queries correspond to the non-distinctive elements like trees, lane markings and sky. Here, the poor localization performance can be attributed to the distinctive weights which were unable to negate the corresponding reliability weights of such features in Eq. 6. However, our proposed method to learn and use the subset of useful features for visual localization is a success since no matching feature in these frames corresponds to the dynamic elements (Fig. 9).

V. CONCLUSION AND FUTURE WORK

The objective of this work to achieve good localization performance using only the set of useful features that are not more than 10% of the total features excluding the sky and road features, has been surpassed with extremely positive

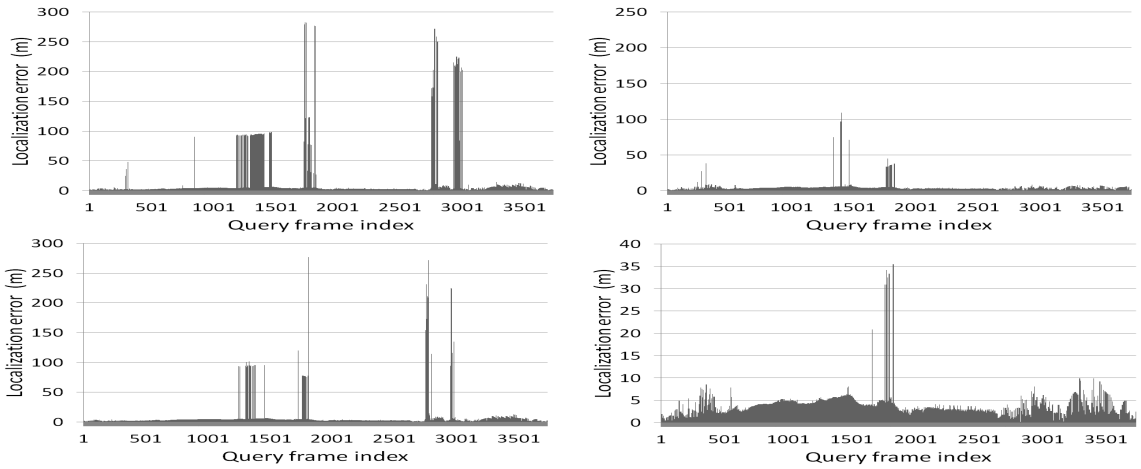

Fig. 10. The first row compares the mean localization error for direct retrieval between Run 1 and Run 7 corresponding to Test 3. The second row compares the same for global localization. Two samples from the twelve frames with localization error greater than 20 m after seven runs are shown in Fig 9.

results. A large improvement in both visual appearance and quantitative measurement of localization accuracy has been proved. The learning principle has been effectively applied to the system and the challenging task of identifying the useful features in highly crowded urban environments has been reduced to a simple cyclic process of augmentation and elimination of features. Our method successfully eliminates the non-useful features of environment and reduces the memory and computational requirements.

Though the performance of system has been excellent over the majority of query frames, large error (greater than 20 m) has been observed for a small percentage (0.321%). This problem needs to be addressed by developing a better method for the recognition of distinctive features in such densely cluttered environments.

VI. Acknowledgement

The authors would like to thank Aayush Bansal and Supreeth Achar for their valuable inputs and insights. This work was supported through the grants made available by Dept. of Science and Technology, India, SR/S3/EECE/0114/2010.

TABLE II

THE QUANTITATIVE PERFORMANCE IS SHOWN FOR BOTH DIRECT IMAGE RETRIEVAL AND GLOBAL LOCALIZATION. THE PARAMETERS INCLUDE MEAN ERROR (M), PERCENTAGE ERROR (P) (ERROR LARGER THAN THE THRESHOLD OF 7.5M) AND VARIANCE. HERE T_i INDICATES THE QUERY DATASET USED WHILE R_j IS THE RUN NUMBER WHERE $i = 1, 2, 3$ AND $j = 1, 3, 5, 7$.

	Direct Retrieval			Global Localization		
	M (m)	P (%)	Variance	M (m)	P (%)	Variance
$T1_{R1}$	9.47	19.50	305.28	6.28	8.50	229.44
$T1_{R3}$	1.97	0.50	1.88	1.86	0.00	0.39
$T1_{R5}$	1.95	1.00	11.73	1.75	0.00	0.42
$T1_{R7}$	2.27	1.50	20.79	1.74	0.00	0.68
$T2_{R1}$	6.48	4.60	640.90	5.12	3.90	355.09
$T2_{R3}$	6.50	6.40	419.60	4.22	3.40	142.92
$T2_{R5}$	3.70	2.20	104.53	3.12	1.80	7.08
$T2_{R7}$	4.09	3.10	81.36	3.05	1.40	8.70
$T3_{R1}$	9.70	8.16	879.77	5.82	4.65	333.14
$T3_{R3}$	8.48	8.67	545.31	5.50	4.49	386.49
$T3_{R5}$	6.26	5.32	286.71	3.59	2.03	25.56
$T3_{R7}$	3.81	2.00	39.24	3.31	0.88	5.13

References

[1] J. Leonard and H. Durrant-Whyte, "Simultaneous map building and localization for an autonomous mobile robot," in *IROS Workshop*, 1991, pp. 1442–1447.

[2] G. Schindler, M. Brown, and R. S. Szeliski, "City-scale location recognition," in *CVPR*, 2007.

[3] M. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *ICRA*, 2012, pp. 1643–1649.

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

[5] D. Santosh, S. Achar, and C. V. Jawahar, "Autonomous image-based exploration for mobile robot navigation," in *ICRA*, 2008, pp. 2717–2722.

[6] C. Zhou, Y. Wei, and T. Tan, "Mobile robot self-localization based on global visual appearance features," in *ICRA*, 2003, pp. 1271–1276.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.

[8] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, 2003, pp. 1470–1477 vol.2.

[9] O. Chum, J. Philbin, and A. Zisserman, "Near duplicate image detection: min-hash and tf-idf weighting," in *BMVC*, 2008.

[10] S. Achar, C. V. Jawahar, and K. M. Krishna., "Large scale visual localization in urban environments," in *ICRA*, 2011, pp. 5642–5648.

[11] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *ECCV*, 2010.

[12] P. Turcot and D. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *ICCV Workshop*, October 2009, pp. 2109–2116.

[13] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *IJRR*, vol. 30, no. 9, pp. 1100–1123, August 2011.

[14] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *CVPR*, 2006.

[15] T. Vidal-calleja and J. Andrade-cetto, "Active control for single camera slam," in *ICRA*, 2006, pp. 1930–1936.

[16] A. Dame and E. Marchand, "Using mutual information for appearance-based visual path following," in *Robotics and Autonomous Systems*, 2013.

[17] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008.