Visual Localization in Highly Crowded Urban Environments

A. H. Abdul Hafez[1], 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 useful features from multiple runs in highly crowded 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 of 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 reduce the size of representation, and make it more reliable in crowded scenes. We tested the proposed method on data sets collected from highly crowded in Indian urban outdoor settings. Experiments have shown that with the help of a small subset (10%) of the detected features, we can reliably localize. We achieve superior results in terms of an localization error 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 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 experience. Images are represented using local or global features. The utility of global visual features was explored for mobile robot exploration, navigation, and localization [5], [6]. Local features [7] are 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 the problem for large databases, 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 popular techniques such as the Inverted Index and Min-Hash [9] for fast matching and retrieval of images.

Localization in outdoor environment is difficult because, queries are often at different viewpoint, scale and illumination from the previous visual experience [3]. In crowded and cluttered out door settings, the problem is further challenging [10]. Dynamic objects could constitute most of the visible region and thus complicates the localization (see Fig. 1). The results are often erratic since the robot may not 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, collected multiple runs of data from crowded Indian roads. We aim at learning to localize 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 camera fixed to moving vehicle.

Fig. 2. Random frames from [10] dataset (first row) and our dataset (second row) depicting extent of occlusion and large difference in useful visible background.

Knopp *et al.* developed in [11] 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] reduce the database size by selecting a small portion of the total detected features. They call them useful features. The removed features are not useful and not representative. The matching performance is as accurate as using the full set.

Cummins and Newman [4], [13] describe a probabilistic, called FAB-MAP, for recognizing places based on their visual appearance. Their algorithm is suitable for online loop closure detection in mobile robotics. Milford and Wyeth [3] presented a new approach, is called SeqSLAM, for visual localization under different environment and illumination conditions. Instead of obtaining the single most likely location given a current view, the system calculates the best candidate matching sequence. However, both methods pre-

[1] International Institute of Information Technology, Gachibowli, Hyderabad-500032, India.

Fig. 3. The first row shows visual features of images after one training run of environment. We tested Achar's [10] formulation on these set of features. The second row shows the same frames with only useful features after fifth run.

sented above, i.e. FAB-MAP and SeqSLAM, are not tested in highly crowded urban environments, they also neither use the concept of useful features nor they identify dynamic object in the scene.

Achar *et al.* [10] investigated the problem of localization in urban environments, and conducted experiments to verify their methodology on roads in India. They propose a novel method to identify dynamic scene elements from the base run of the robot, and filter out features obtained from dynamic elements in order to facilitate robust localization of the robot. In contrast, a highly crowded environment containing dynamic elements, like standing 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 the robot.

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

The remaining of this paper is organized as follows. Section II shows an overview of the proposed localization method, while Section III focuses on learning useful features when a new data is available from new run of the robot through the environment. Section IV explains the results from conducting a set of experiments support our proposed localization method. Finally, we conclude with some remarks and future work proposals.

II. Qualitative visual localization

Qualitative visual localization uses past experiences of the robot to determine a set of images in close neighbourhood of its current pose. The motivation of our work arises through observing that a majority of extracted visual features belongs to the dynamic environment Fig. 3. Many recent works [10], [11], [12] have argued that only a small percentage of the extracted features are useful. These however haven't 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. Rather than using available object detectors, we propose that the ability to learn the useful subset of environment can be enhanced by traversing the same path multiple times.

The extraction of interest points, called features from the available images using suitable detectors is the first step in visual localization. Many works have suggested the use of the available features in image retrieval but the Bag of words method [8] has been observed to be highly efficient. The method represents the images as a set of unordered visual words which are used to build an inverted index which quantifies the occurence of each feature. A query image, captured at robot's current pose is then quantized to the exisiting vocabulary tree and mapped to a set of visual words. The weighted retrieval method, which identifies discriminative features, returns a set of images that represent the current pose.

A. Bag of words model

The robot explores the environment during one traning run and captures the frames referred as ground truth data. SIFT features [7] are extracted from the keyframes and a vocabulary tree of K branches and L depth containing K^L leaf nodes is constructed using hierarchical k-means clustering [2]. The inverted index stores for each visual word the occurrence of a feature and maintains its history through weights . The occurrence of a feature has been quantified by tf-idf [8] in some earlier works, but in this paper we use the weighted retrieval which assigns higher weight to highly discriminative features while lower weight to others.

B. Distinctive features

The distinctiveness of a given feature z with respect to robot 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 given by the $P(Z)$. This can be interpreted as looking for features which appear in all images about some specific robot pose, but rarely appear elsewhere. The concept of information gain used in [2] captures this concept of distinctive features.

Information gain $I(X|Z)$ 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 with respect to the entropy $H(X)$ and conditional entropy $H(X|Z)$ of distributions $P(X)$ and $P(X|Y)$. For two random variables, X and Z, the entropy of X is given by:

$$H(X) = -\sum_{x \in X} P(X = x) \log P(X = x), \tag{1}$$

while the conditional entropy is

$$H(X|Z) = -\sum_{z \in Z} P(Z = z) H(X|Z = z). \tag{2}$$

Hence, the information gain is defined as

$$I(X|Z) = H(X) - H(X|Z). \tag{3}$$

The information gain is always considered with respect to a specific robot pose x_i and a specific feature z_k. In other words, the distinctive weight w_k of the kth feature z_k in the vocabulary tree is computed as

$$w_k = I(x_i|z_k) = H(x_i) - H(x_i|z_k). \tag{4}$$

More detailed computation of the distinctive weight, i.e. information gain, is available in [2], [10], [14], [15], where this concept is successfully used.

C. Querying

When the robot explores the environment again, it captures a new set of frames, called query data. The extracted features from each query image are quantized to visual words using the available vocabulary tree and inverted index by using greedy N-best paths algorithm [2]. Since the vocabulary tree was built from ground truth data captured under different conditions, a highly efficient method is required for both vocabulary tree construction and image retrieval. By assisgning desciminative weights to the visual words corresponding to query feature, a score is calculated for each ground truth image by searching for visual words corresponding to the image. To reduce the complexity of search, the position of features relative to the principal point of the camera is used for filtering irrelavant images [10].

The normalized score for the few relevant images is computed as:

$$Score(img_n) = \frac{\sum_{z_k \in Z_n \cap Z_q} W_k}{|Z_n|}, \tag{5}$$

where Z_n is the set of SIFT descriptors in the n^{th} 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 feature discussed later in Eq. (12).

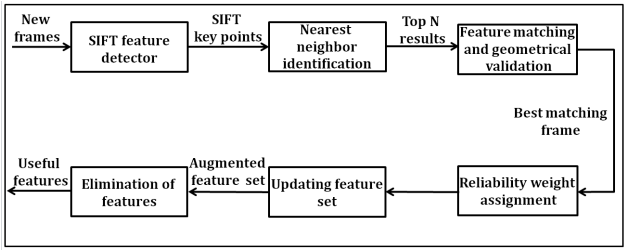

Fig. 4. Flowchart showing the gradual learning process of system. The input to the system are newly observed frames and output are useful features

The top N images based on the normalized score are returned as the best matches for direct image retrieval. In case of global localization, the feature correspondences between the top matches and the query image are geometrically verified using epipolar geometry and results are filtered based on number of inliers.

The weighted bag of words retrieval method, however suffers from some disadvantages in crowded urban environments. The inclusion of non-distinctive features like sky, lane markings and trees in the feature set results in problems like perceptual aliasing which may lead to large localization errors. As discussed above, the vocabulary tree also provides inefficient output in such conditions which necessiates stricter geometrical validation of matching features, thereby increasing retrieval run time. This paper addresses the problem by identifying useful features of environment through multiple experiences of robot which not only reduces the size of vocabulary tree but also improves the performance after each experience.

III. Learning useful features

Through multiple experiences, the objective is twofold: Firstly, to observe a set of new features which may be hidden previously and secondly to identify the useful features from previous experiences using current observation. Thus, the expectation is that through each new training run, the robot enhances its knowledge of the environment. Figure 4 depicts a block diagram of the learning process of our proposed method. From each available frame in the current run, SIFT key points are extracted. To develop the relationship between newly observed features and the existing feature set, the best matching frame from current run corresponding to each frame from the first training run (henceforth called base run) is determined. The available GPS data defines a small neighbourhood 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 some initial weight.

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 using features is increased. Every training run is classified as one of the following:

- Base run: The robot experiences the environment for the first time and unable to determine the useful features.
- Augmentation run: The knowledge of environment is enhanced through new experience and the useful features are learnt. Some newly observed features are also augmented to the feature set.
- Elimination run: The useful features are retained based on their reliability weights

A carefully designed methodology to use both augmentation and elimination run is the critical to the learning process.

A. Initial reliability weights in base run

The robot explores the environment for the first time and assignes equal weight to each feature. This weight, called reliability weight indicates the usefullness of a feature in the environment based on its ability to reoccur the neighbourhood of the same pose under different conditions over multiple runs. Assuming M_1 features are initially extracted and each kth feature is assigned reliability weight, $w_k^0 = 1$, the normalized weight is given as:

$$\hat{w}_k^0 = \frac{1}{\sum_{k=1}^{M_1} w_k^0} = \frac{1}{M_1} \tag{6}$$

where $\sum_{k=1}^{M_1} \hat{w}_k^0 = 1$. Alongwith the reliability weights, the robot also saves visual features and a GPS tag for each pose which is used in future experiences.

B. Augmentation of features

As the robot explores the enviornment again under different conditions, it observes a set of features which may or may not have been recorded previously. To develop a relationship between these newly observed features and the exisiting features, first the best visually matching frame from the current run corresponding to every frame of the base run is determined. We realize that matching all the newly observed features with exisiting may not be a realistic approach. So based on the available GPS data, a set of $\mathcal{K}$ images from current run are indentified which represent nearest neighbours of the image belonging to the base run. This is justified since the useful features observed in an image of the base run at a particular pose cannot be observed at an entirely different pose in the current run. Thus, the problem of $N_1.N_t$ computations is reduced to only $N_1.\mathcal{K}$ computations and the image with maximum number of geometrically validated matching features is labelled as the best matching frame. Here N_1 and N_t are the number of images from base run and the current run respectively.

Any feature D_1 from the selected best image is labelled as a match to a feature D_2 of the base image if and only if:

$$\frac{d(D_1, D_i)}{d(D_1, D_2)} \leq T, \tag{7}$$

where $d(D_i, D_j)$ is the Euclidean distance of descriptor D_i to descriptor D_j . This formula can be interpreted as that the Euclidean distance $d(D_1, D_2)$ multiplied by a threshold T is smaller than the distance of descriptor D_1 to all other descriptors D_i in the image from base run [16]. Selecting $T = 1.5$ show best satisfaction.

The matching visual features represent the previously observed features, newly observed features as well as features belonging to dynamic objects. These features are assigned the visual stability weight given as:

$$w_{vs} = \frac{\mathcal{F}}{\mathcal{N}}, \tag{8}$$

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

The feature correspondences contain many negative results, dynamic objects which are discarded by identifying spatially consistent features through epipolar geometry. Thse are simply the inliers estimated by fitting a fundamental matrix using RANSAC algorithm and assigned the spatial consistency weight as:

$$w_{sc} = \frac{\mathcal{G}}{\mathcal{N}}, \tag{9}$$

where $\mathcal{N}$ is the number of features of key frame of base run, $\mathcal{G}$ is the number of geometrically validated matches.

A large portion of the newly observed data consists of the features belonging to the dynamic objects while a smaller percentage, though useful belonging to to the static part hidden during the previous run. Despite the knowledge that inclusion of these hidden static features would imply inclusion of large number of useless features as well, we augment the exisitng data because learning these hidden features has been our objective throughout. The newly observed features are assigned a reliability weight, w_k equal to the weight possessed by the minimum weighted feature in the database during the particular run.

It may be argued that a certain percentage of these hidden useful features, though distinctive may never reappear even under a different conditions, thus making it impossible for robot to learn them. Such features which have a tendency to be hidden for prolonged period of time are labelled as non useful features and treated similar to the dynamic features. A similar argument could be made for some dynamic objects which may always be present in the environment and may mislead the system into being considered as useful. For such features, we use the idea of the distinctiveness to negate their relative importance.

C. Enhancing the knowledge through elimination

The features belonging to dynamic objects and augmented to the exisiting knowledge of robot pose the problem of decreasing the efficiency of localization process. However, it can be analyzed that only the useful features would reoccur at the same pose under different conditions, i.e. gain higher reliability weight. Thus, the features with reliability weights greater than the minimum weighted feature are retained while the remaining are eliminated. To ensure that a visual feature is not discarded unfairly, we continue to augment feature set till third run. Though expensive, the idea is to provide

Fig. 5. Gradual learning process through base, augmentation and elimination run. Note the growth in number of features over frame 2 and 3 and finally retention of only useful features in last frame.

new features sufficient support to prove theirusefullness. The features unable to prove their importance to the process are then eliminated but at the same time, the newly observed features in the third run are retained for future explorations. This ensures that the newly observed features during any training run are verified substantitally before elimination. Thus, after t runs, any kth useful feature will have a reliability weight:

$$w_k^t = \hat{w}_k^{t-1} + w_{vs} + w_{sc} \tag{10}$$

Then, it is normalized to produce the weight, $\hat{w}_k^t$ is given as:

$$\hat{w}_k^t = \frac{w_k^t}{\sum_{k=1}^{M_t} w_k} \in (0,1) \tag{11}$$

where M_t is number of features after tth run Thus, the total weight assigned to each feature of base run is given as:

$$W_k = w_1.w_2 \tag{12}$$

where $w_1 = I(X;Y)$ as in Eq. 4 and $w_2 = -\frac{1}{log(\hat{w}_k^t)}$ Here, the total weight of a feature is directly related to its reliability and distinctiveness.

D. Amount of Learning

The robot learns the environment gradually with each run through augmentation and elimination of feature set. Since the data in the proposed work is collected over a shorter duration, the number of useful visual features is limited in the environment. Hence, it is fair to assume that the system's incremental learning ability will reduce with each training run. Since 39.31% features have been classified as useful after seventh run (Table I), the system may not be expected to learn further due to lack of static visual features. It will be shown experimentally later that the mean error also saturates after fifth run (Fig. 7). Rather, in some cases it increases slightly which is due to unintentionally learning of reoccuring non useful features. This improvement in performance is also closely related to the accuracy of GPS data. The performance of the system will never exceed the minimum error of GPS receiver.

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 x 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 10fps, the base run then having 2596 frames. The motion blur and unpredictable motion of traffic added to the uniqueness of collected data. A GPS receiver of permissible error was used to record the current position of the vehicle and used later as ground truth data. Based on the time stamp associated with the GPS information and the captured video, GPS tags were assigned to the extracted key frames.

The total number of features increases linearly with each training run and if allowed to continue would affect the computation time with no improvement in retrieval performance. By applying the proposed elimination concept, the number of features was reduced by 90.99%, 84.63% and 80.94% for third, fifth and seventh run respectively which highlights the effectiveness of our approach in retaining only the small subset of useful features and eliminating the rest.

B. Testing

Training video from eighth run of the same environment was sampled at 10 frames per second to obtain 3734 key frames. Though there was a large variation in 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-C, the inverted index of the quantized visual words was searched for matches and the score was assigned to each database frame. The top 10 results based on the score were returned as best matches and labelled as direct retrieval results. The matched features in this set of 10 top results were geometrically verified to remove the mismatches and top results were again obtained based on the number of geometrically validated features. GPS ground truth data was used to check localization performance by measuring 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 implemented on three datasets. For all the tests, the query images were chosen randomly assuming independent and identically distributed samples.

1) Test 1: The first 500 key frames of all the eight runs covering (approx) one-fifth of the total path were chosen. The key frames from the first seven runs were used to build database of useful features. A random 200 frames extracted from the first 500 frames of eighth run were used as query.

TABLE I

VARIATION IN NUMBER OF VISUAL FEATURES WITH EACH TRAINING RUN. THE NUMBER OF USEFUL FEATURES RETAINED BY SYSTEM AFTER ELIMINATION ARE ALSO SHOWN. THE DRASTIC 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	3.38M	6.02M	925k	3.52M	6.01M	1.14M

Fig. 6. Sample query frames from eighth training run.The repeating dynamic elements, pedestrians and the clogged roads make the localization process difficult.

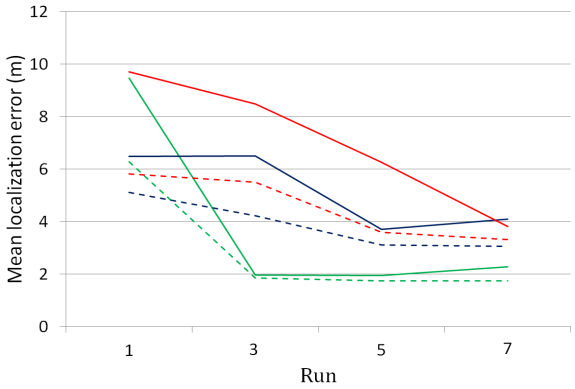

Fig. 7. Variation of mean localization error for both direct retrieval (solid lines) and global localization (dashed lines) corresponding to the three tests is shown. The performance can be observed getting better with increase in number of runs. However, a slight increase in mean error is observed for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

2) Test 2: To check for accuracy over larger number of frames, the database of useful features was built using all the keyframes from the first seven training runs. After building the vocabulary tree, a random 1000 key frames were chosen from the eighth run and queried for both direct retrieval and global localization.

3) Test 3: To check the scalability of our proposed work and the effectiveness of larger vocabulary tree, database of useful features was built as in test 2 from the first seven runs but the full eighth run, i.e. 3734 key frames covering the full path were used as query.

C. Discussion

Despite occlusion of useful features, the returned results have shown good visual performance. As can be seen from adjoining Fig. 8, good performance was achieved after fifth training run with only 30% visual features whereas tests after first run with all the features based on the work of [10] failed to give desired results. The failure was due to large occlusion of useful features as well as similarity in appearance of scenes. In our case, even the total 30% features contained a major portion belonging to sky and roads. So, the actual available useful features were far less, on an average 10%.

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 through the proposed method after fifth run using only 10% useful features

Over the full eighth run, our formulation continued to return results similar in appearance to the query and belonging to the same pose.

The quantitative analysis (Table II) in terms of distance measure has shown remarkable improvement. The direct retrieval and global localization results showed an improvement of 60.72% and 43.12% respectively after seventh run for Test 3 (Fig. 7). The reduction in variance is much larger with 95.53% and 98.46% respectively for both the cases. Considering the inherent GPS error and usage of only 10% useful features, the currently achieved results can be termed as excellent and support our approach to gradually learn the useful features through multiple experiences.

12 frames from a total of 3734 queries 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 belong to the non distinctive elements like trees, lane markings and sky which being repetitive, are incorrectly classified as useful because of the weights. This error can be attributed to the inability of Achar's method to effectively assign low weights to non-distinctive features. Our argument is aided by the fact that no matching feature in these frames belonged to the dynamic clutter, implying that our proposed method is successful.

V. CONCLUSION AND FUTURE WORK

The basic objective of the paper to achieve good localization performance using only the set of useful features that are not more than 10% of the total feature excluding the sky and road features, has been surpassed with extremely positive results and large improvement in both visual appearance as well as quantitative measurement. The learning principle has been effectively applied to the system and the challenging task of searching useful features amongst clutter has been

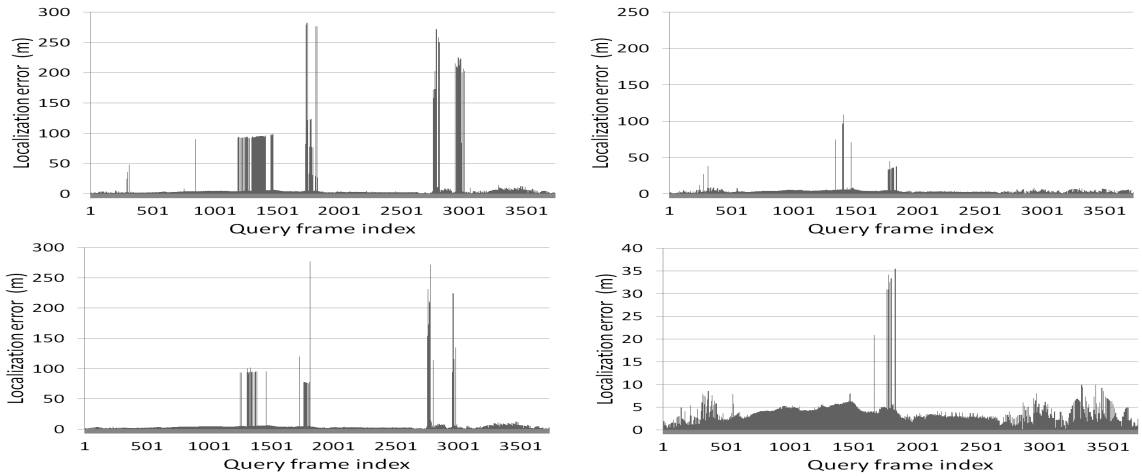

Fig. 10. The first row shows the mean localization error comparison for direct retrieval between Run 1 and Run 7 corresponding to Test 3. The second row compares the same for global localization. The few frames with large errors even after seven runs are shown in Fig. 9

Fig. 9. Shown are the matching features of the frames returning large errors. Most of these matching features correspond to non distinctive elements like trees and lane markings

reduced to a simple cyclic process of augmentation and elimination. At the same time, we have observed that the unintentional learning of non-useful features by the system needs to be avoided by carefully monitoring its change in performance after every run. Though the performance of system has been excellent over the majority of query frames, large error has been observed for a small percentage of frames (0.003%) due to failure of distinctiveness. This is a concern that needs to be addressed by designing a robust mathematical framework for recognition of distinctive features in such densely cluttered environments.

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 (V). 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 (%)	V	M (m)	P (%)	V
$T1_{R1}$	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, pp. 1–7.

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

[15] A. Dame and E. Marchand, "Using mutual information for appearance-based visual path following," in *Robotics and Autonomous Systems*, 2013.
[16] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008.