Probabilistic Location Recognition using Reduced Feature Set

Fayin Li and Jana Košecká
Department of Computer Science
George Mason University, Fairfax, VA 22030
Email: {fli,kosecka}@cs.gmu.edu

***Abstract*— The localization capability is central to basic navigation tasks and motivates development of various visual navigation systems. In this paper we describe a two stage approach for localization in indoor environments. In the first stage, the environment is partitioned into several locations, each characterized by a set of scale-invariant keypoints and their associated descriptors. In the second stage the keypoints of the query view are integrated probabilistically yielding an estimate of most likely location.**

The novelty of our approach is in the selection of discriminative features, best suited for characterizing individual locations. We demonstrate that high location recognition rate is maintained with only 10% of the originally detected features, yielding a substantial speedup in recognition and capability of handling larger environments. The ambiguities due to the self-similarity and dynamic changes in the environment are resolved by exploiting spatial relationships between locations captured by Hidden Markov Model.

I. Introduction

The problem of localization is of interest in several applications including augmentation of human navigation capabilities and mobile robot localization. Two main variations of the localization problem have long been established in the robotics community and are known as global localization (also known as robot kidnapping problem) and pose maintenance. In this paper we will focus on the global localization aspect and demonstrate how to solve it by means of location recognition.

A. Related Work

Due to the different nature of the location recognition task compared to object recognition , several representations of locations were proposed in the past. Previously proposed approaches frequently used local image features and their descriptors for location representation, which were invariant to affine transformations or rotationts [1], [2] or local Fourier transforms of salient image regions [3]. Due to the locality of these image features, the recognition can naturally handle large amounts of clutter and occlusions. The sparser set of descriptors can be obtained by principal component analysis or various clustering techniques. Several instances of pose maintenance and acquisition of metric environment models have been successfully solved in smaller scale environments [4], [5]. The applicability of these methods to large dynamically changing environment poses additional challenges and calls for alternative models.

B. Approach Overview

We propose to tackle the location recognition and localization problem by using a model of the environment represented by a set of locations and spatial relationships between them. Each location is represented by a set of views and their associated local scale invariant features. We present a novel technique for identifying most discriminative features for individual locations reducing the feature database to 10% of its original size, without forgoing the recognition accuracy. An associated likelihood model characterizing each location is then used in the Hidden Markov Model framework which enables us to resolve misclassification due to the self-similarity and dynamic changes in the environment. Once the most likely location is determined we can compute the relative pose of the camera, with respect to the reference view. We will report on the localization experiments in indoor environment with 18 locations.

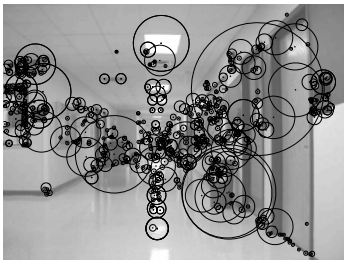
II. Location Representation

As a starting point of our method, we use the environment model obtained in the exploration stage. Given a temporally sub-sampled sequence acquired during the exploration, the sequence is partitioned into $N = 18$ different locations. The locations in our model correspond to hallways, sections of corridors and meeting rooms approached at different headings. The initial model was obtained by a mobile robot, which was guided through the environment. The number of views per location vary between 5 to 20 depending on the appearance variation within the location. The transitions between the locations occur either at places where navigation decisions have to be made or when the appearance of the environment changes suddenly. The images were taken approximately every 2-3 meters. The representative views of some locations in Figure 1 demonstrate the variability of our dataset. More details about the model acquisition stage can be found in [6].

Individual locations are represented by scale-invariant (SIFT) keypoints described in [1]. The SIFT features represent distinguishable image locations, which are stable across variations in scale. Each feature is endowed with a 128 dimensional descriptor, which captures the orientation information of local image region centered at each keypoint, is rotationally invariant and has been shown to be robust with respect to large variations in viewpoint and scale.

Fig. 1. Examples of representative views of 6 out of 18 locations.

Figure 2 shows the features detected in one of the representative views of Location 1. The number of features detected in each image varies from hundreds to thousands. Despite the large overlap between consecutive views, only a small number of features detected in consecutive views are matched. Some features have better capability to handle variations in scale and viewpoint and match stably in several different views of the same location. Selecting such features can keep the environment model more compact and save the computational cost for future localization. This observation brings to forefront



Location 1

Fig. 2. 700 SIFT features detected in the representative view of Location 1.

the issue of feature selection in the model acquisition stage.

A. Feature Selection

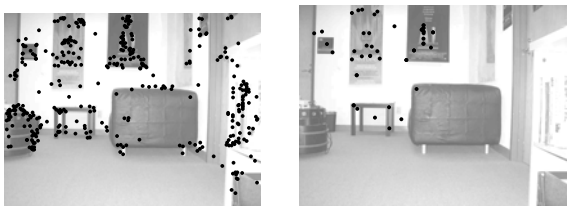
Previously proposed techniques for reducing the feature pool include k-means clustering, greedy techniques or boosting [7], [8], [9], [10]. Questions focusing on the model compactness [11] as well as trade-off between the complexity of features and the complexity of classifiers were explored in [12]. In [11] authors estimate the posterior of each feature with respect to each object and the Shannon entropy is used to select the discriminative regions. Our method for feature selection is similar to the method proposed by [11] but with a different selection criterion. Suppose location L_i has N_i training images with total K_i of detected features $G_i = \{g_k^i\}_{k=1\cdots K_i}$. To obtain the information content of each feature g_k^i with respect to location identification, we need to estimate the posterior probability $P(L_l|g_k^i), l = 1\cdots N$. The posterior probability at feature g_k^i is estimated using only features g_j inside a Parzen window of a local neighborhood $Z = \{g_j | \ \|g_k^i - g_j\| \leq \epsilon, j = 1\cdots z\}$, where ϵ determines the size of the window. We weight the contribution of specific feature g_j^l in Z (labeled by location L_l), which should increase the posterior estimate $P(L_l|g_k^i)$, by a Gaussian kernel function $N(\mu, \sigma)$ in order to favor the features with smaller distance to the feature g_k^i, with $\mu = g_k^i$ and $\sigma = \epsilon/2.5$. Then the posterior probability at feature g_k^i is estimated as

$$P(L_l|g_k^i) \propto \sum_{g_j^l \in Z} exp(-\frac{\|g_j^l - g_k^i\|^2}{2\sigma^2}). \quad (1)$$

After the posterior probabilities of all features are obtained, we could proceed by calculating the information entropy of each feature and use it in the selection process. In our data set, however, each location has different number of training images and the number of features detected in each image varies largely as it can be seen in Figure 2c. If one location l has few features detected and the Parzen window size is not accurate, even though a good feature g_k^l has a large posterior probability $P(L_l|g_k^l)$, it may also have a large posterior probability with respect to another class $P(L_m|g_k^l)$, because location m has large number of detected features. The posterior probability estimates will hence be biased and the entropy will not successfully capture the right information content.

Due to this reason we use directly the estimate of $P(L_l|g_k^l)$ in each image for ranking the features based on their discrimination capability. The number of features we keep is specified as a percentage η of detected features. For a training image from location L_i with M_i features detected, only top $max(M_i\eta, N_0)$ features are selected based on their posterior $P(L_l|g_k^i)$ ranking. N_0 is the minimal number of features selected to avoid discarding too many features from the images of locations with few detected features. Because we select the features based on the rank of their posterior, the feature with high rank of posterior with respect to location L_i may also have large posterior with respect to another location L_j. That feature can be shared by several locations and can distinguish them from others. At the same time it can introduce ambiguities in discriminating locations. Currently, instead of fully exploiting the shared features, we properly account for their contribution in the feature matching stage.

In our data set, there are 296 training images for 18 locations with 112,705 detected features. We chose $\eta = 10\%$ and $N_0 = 50$ to select features from each image. Each feature has the location (x, y), a scale s, an orientation and a 128 dimensional descriptor. Figure 3 shows the total number of detected SIFT



a) b)

Fig. 3. a) The total of 480 features detected in one of the representative views of the location and b) 50 informative features selected by our method.

Fig. 4. The top 16 features selected for three of the locations.

features and the informative features selected in the image by our method. The features belonging to posters have good discrimination capability. Figure 4 shows the top 16 selected features for different locations. Each feature is cropped from the training image centered at (x, y) with radius $r = 6 \times s$, where s is the scale of the SIFT feature. The patches are normalized to 64×64.

Note that the selected features often have large scale and capture the global information about individual locations. For example for Location 1, which has large depth variation some the selected discriminative features are centered in the middle of the corridor and entail the entire view of the corridor. This also demonstrates that the suitable choice of the representation for a location varies largely and for certain locations the global descriptors are indeed highly discriminative. Figure 5 shows some examples of discarded features. For example features belonging to the ceiling lights are discarded except when the feature contains multiple lights.

III. Reduced Feature Set Matching

In order to demonstrate the feature selection process is effective, we compare the performance of location recognition using the reduced feature set, with the standard voting approach which uses all features. In this experiment the i-th

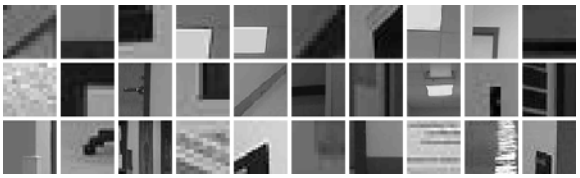


Fig. 5. The features discarded in the selection.

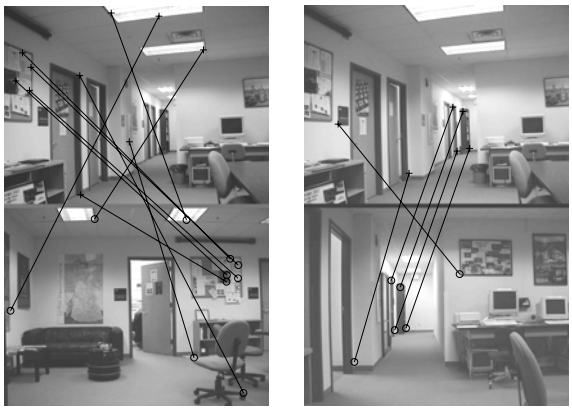
sequence (# of frames)	original set	reduced set
No.1 (296)	100.0%	98.0%
No.2 (134)	82.1%	79.9%
No.3 (130)	83.1%	73.9%

TABLE I

Recognition performance in term of percentage of correct localization using voting scheme.

location is represented by a number of representative views $\{I_n^i\}$ and their associated original and reduced SIFT feature sets $\{g_k(I_j^i)\}$ and $\{\tilde{g}_k(I_j^i)\}$. For a new query image Q and its associated features $\{g_k^Q\}$, a set of matches between Q and each model view I_j^i is determined by matching each feature in $\{g_k^Q\}$ against the model database features and choosing the nearest neighbor based on the Euclidean distance between two descriptors. As suggested in [1] only the point matches whose nearest neighbor is at least 0.6 times closer than the second nearest neighbor are considered. The model view I_j^i with the highest number of matched keypoints with Q is considered to be the correct result. To evaluate the proposed feature selection mechanism we compare the recognition performance using the model database of all detected features $\{g_k(I_j^i)\}$ and the reduced feature set $\{\tilde{g}_k(I_j^i)\}$.

Table I shows the recognition rates for the training sequence and two test sequences using the original and reduced feature set, respectively. The results are reported on the training sequence of 296 images from which the model views and features were selected and two test sequences of 134 and 130 images. The two test sequences, taken at different days and times of day, exhibit larger deviation from the path traversed during the training and several locations underwent dynamic changes which changed their appearance. In few instances the reduced features yields correct recognition, while the wrong decision was made using the original feature set (see Figure 6). Relatively poor performance on the test sequences was due to several changes in the environment between the training and testing stage as demonstrated in Figure 8. Most SIFT features belong to objects some of which are not inherent to particular locations. In the next section we describe how to replace the voting scheme by a simple probabilistic model and propose how to resolve the remaining issues by explicitly modeling spatial neighborhood relationships between individual locations.



a) b)

Fig. 6. Selected features yield better matches than original feature set. a) the misclassification using original set b) correct recognition with one false match using the reduced set.

IV. Probabilistic Location Recognition

The results of voting approach described in the previous section demonstrate that the feature selection is very effective. The misclassified locations are often due to the self-similarity of the environment (e.g. similarity of the appearance of corridors or hallways belonging to different locations), large changes in the pose between the query view and model views or dynamic environment changes. The classification performance can be improved by either exploiting more elaborate recognition scheme or additional information about the environment which would help to reduce the ambiguities due to the self-similarity of the environment. In the following section we demonstrate how to improve the classification by formulating the location recognition probabilistically and by exploiting the spatial relationships between the locations modeled by Hidden Markov Model.

The probabilistic formulation of the classification prediction entails computation of the posterior probability $P(L_l|\{g_k^Q\})$ of each location given the selected features from the query view. Such computation requires likelihood model for the matched features, which explicitly accounts for the quality of individual matches and hence is expected to be superior to the simple voting approach. The likelihood model can then be naturally used in the Hidden Markov Model (HMM) to achieve more reliable and robust system.

When local descriptors are used as observations, several models of class posteriors have been proposed in the context of probabilistic approaches to object recognition [13], [14]. The proposed likelihood models account for the feature density and spatial relationships between features and have been shown to improve overall recognition rate. Those approaches use very complex parametric model and need large number of training examples to learn its parameters. Furthermore the location recognition problem is notably simpler then the object recognition problem due to the background clutter not being so prominent [1]. We propose a non-parametric method to estimate the $P(L_l|\{g_k^Q\})$ from training data directly without modeling the decision function. The essential features of this probabilistic method, which we describe next, is the selection of relevant features in the matching stage and integration of the evidence they provide for individual locations through a strangeness measure. As the result of the feature selection stage the features from the representative views of location i are joined to form a model of that location denoted by $\tilde{G}_i = \{\tilde{g}_k^i\}$. Given the query image Q with detected features $\{g_k^Q\}$ in order to determine most probable location, we need to compute the posterior probabilities $P(L_i|Q) = P(L_l|\{g_k^Q\})$ for $l = 1 \cdots N$. Similarly as in the model building stage, many features in $\{g_k^Q\}$ are not informative and have no evidence for classification label. They may confuse the prediction if such features are considered during prediction, especially when cluttered background is present. We need to select *good* and relevant features from $\{g_k^Q\}$ for estimation of the posterior. The selection criterion not only gives the number of matched features, but also yields the confidence of the match. The procedure is based on the hypothesis test. Given a set of features in the query image $\{g_k^Q\}$, we first define the so called strangeness parameter α_k^i, which characterizes the discrimination capability of k-th feature, with respect to i-th location

$$\alpha_k^i = \frac{min_{g_j \in \tilde{G}_i}(\|g_k^Q - g_j\|)}{min_{g_j \notin \tilde{G}_i}(\|g_k^Q - g_j\|)}. \quad (2)$$

α_k^i is the ratio of minimal intra-distance within the class and minimal inter-distance to features from other classes. If α_k^i is greater than 1, the feature g_k^Q is not contributing to classification of Q as label L_i. The smaller the α_k^i is, the more discriminative is the feature for the purpose of classifying Q as i-th location. If $g_j^* = argmin_{g_j \in G_i}(\|g_k^Q - g_j\|)$ is a shared feature among a set of locations S, the strangeness is very close to 1. Since we want the shared features to be considered, in the selection stage the strangeness of the shared feature g_k^Q with putative label l is re-computed as follows

$$\alpha_k^l = \frac{min_{g_j \in \tilde{G}_l}(\|g_k^Q - g_j\|)}{min_{g_j \notin \tilde{G}_l \wedge g_j \notin \tilde{G}_{i \in S}}(\|g_k^Q - g_j\|)}. \quad (3)$$

Hence in the case of shared features we do not consider the inter-distance from the features in shared locations $t \in S$ in this special case. The shared features can help to distinguish the location subset S from other locations but are useless for discriminating the locations in subset S. The computation of α-values has the same computational complexity as the nearest neighbor ratio computation in the standard voting scheme. For the computation of the likelihood $P(\{g_k^Q\}|L_l)$ we select only top R features from the query image, ranked by their strangeness, under current hypothesis test. Note that don't consider the features who have strangeness measure $\alpha_k^l \geq 1$.

[1]The probabilistic models used in the object recognition, must also account for the fact that large number of detected features comes from background and not the object.

Sequence (# of frames)	Maximal Likelihood	HMM
No.1(296)	99.0%	100.0%
No.2(134)	85.8%	95.5%
No.3(130)	80.8%	95.4%

TABLE II

RECOGNITION PERFORMANCE IN TERMS OF PERCENTAGE OF CORRECT LOCALIZATIONS BASED ON α-VALUES.

The likelihood of feature g_k^Q has the putative label L_l is defined as

$$P(g_k^Q|L_l) = P(\alpha_k^l|L_l) \; = \; exp(-\frac{{\alpha_k^l}^2}{2\sigma^2}). \tag{4}$$

Since we do not know how many features belong to location L_l among top R features, we need to integrate the evidence over all possible hypotheses. A hypothesis in our case indicates that a subset $h_j \neq \emptyset$ of top R features is classified as location L_l. Assuming the selected features are independent, we can now compute the probability of a single hypothesis h_j conditioned on location L_l

$$P(h_j|L_l) = \prod_m P(\alpha_m^l|L_l) \prod_n (1 - P(\alpha_n^l|L_l)). \tag{5}$$

Index m ranges over features which belong to location l with certain probability, n ranges over features which do not belong to location L_l, where $m + n = R$ is the number of selected features, i.e. the length of a hypothesis. Then the probability $P(L_l|\{g_k^Q\})$ can be computed as

$$P(\{g_k^Q\}|L_l) = \sum_{h_j} \prod_m P(\alpha_m^l|L_l) \prod_n (1 - P(\alpha_n^l|L_l)). \tag{6}$$

It can be simplified to

$$P(\{g_k^Q\}|L_l) = 1 - \prod_{k=1}^{R} (1 - exp(-\frac{{\alpha_k^l}^2}{2\sigma^2})). \tag{7}$$

In our experiments we use $\sigma = 1/3$. Assuming the location prior is uniform and $R = 10$, we tested the training and two test sequences again using the above maximum likelihood criterion. The recognition rates are shown in Table II. The performance is very close to the one using the original training feature set while reducing the matching computational cost by about 90%. If we consider the informative factor of each training feature and can reliably estimate their likelihood, we can estimate the posterior by

$$P(L_l|\{g_k^Q\}) \propto P(L_l)(1 - \prod_{k=1}^{R} (1 - P(g_k^*|L_l) exp(-\frac{{\alpha_k^l}^2}{2\sigma^2}))), \tag{8}$$

where $g_k^* = argmin_{g_j \in G_l}(\|g_k^Q - g_j\|)$ is the feature from location l matched with g_k^Q. In our method, the minimal number of features selected N_0 in each representative view is an important parameter to make the feature selection reliable. We vary N_0 from 20 to 70 and test the recognition performance again. The recognition rates will be almost constant after N_0 is greater than 60, which means using more features helps little in increasing the performance.

A. Exploiting Neighborhood Relationships

We propose further to deal with the dynamic changes in the environment by incorporating additional knowledge about neighborhood relationships between individual locations. The rationale behind this choice is, that despite the presence of ambiguities in recognition of individual views the temporal context should be instrumental in resolving them. The temporal context is determined by spatial relationships between individual locations and is modeled by a Hidden Markov Model (HMM). In this model the states correspond to individual locations and the transition function determines the probability of transition from one state to another. We have already explored the use of this representation using original feature set and slightly different likelihood function [6]. Since the locations cannot be observed directly, each location is characterized by the location observation likelihood $P(o_t|L_t = L_l), l = 1 \cdots N$ at time t during the exploration. The most likely location is at each instance of time obtained by maximizing the conditional probability $P(L_t = L_l|o_{1:t})$ of being at time t and location L_l given the available observations up to time t. The location likelihood can be estimated recursively using the following formula

$$P(L_t = L_l|o_{1:t}) \propto P(o_t|L_t = L_l) P(L_t = L_l|o_{1:t-1}) \tag{9}$$

where $P(o_t|L_t = L_l)$ is the observation likelihood, characterizing how likely is the observation o_t at time t to come from location L_l. The conditional probability $P(o_t|L_t = L_l)$ that a query image Q_t at time t characterized by an observation $\{g_k^{Q_t}\}$ comes from the location l is simply the likelihood $P(\{g_k^{Q_t}\}|L_t = L_l)$ introduced in Equation IV

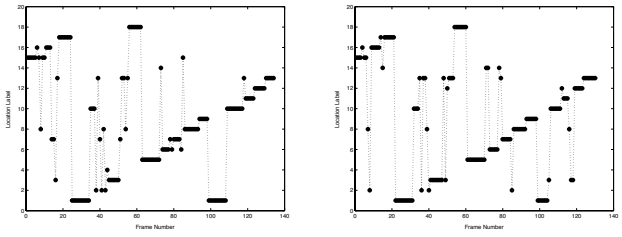
$$P(o_t|L_t = L_l) \propto 1 - \prod_{k=1}^{R} (1 - exp(-\frac{{\alpha_k^l}^2}{2\sigma^2})). \tag{10}$$

The second term of equation (9) can be further decomposed to explicitly incorporate the location neighborhood relationships

$$P(L_t = L_i|o_{1:t-1}) = \sum_{j=1}^{N} A(i,j) P(L_{t-1} = L_j|o_{1:t-1}), \tag{11}$$

where N is the total number of locations and A is a $N \times N$ matrix, where $A(i,j) = P(L_t = L_i|L_t = L_j)$ is the probability of two locations being adjacent. In the presence of a transition between two locations the corresponding entry of A was assigned a unit value and in the final stage all the rows of the matrix were normalized.

The results of location recognition employing this model are in Figure 7. For each frame of two test sequences, Figure 7 shows the location label which had the highest probability. The recognition rate with HMM for sequence 1 was 95.5% and for sequence 2 it was 95.4%. While in both cases some images were misclassified the overall recognition rates are a great improvement compared to the rates of single view location recognition. The dynamic changes, which make the single



a) Testing 1 without HMM b) Testing 2 without HMM

c) Testing 1 with HMM d) Testing 2 with HMM

Fig. 7. Classification results with for Test Sequence 1 and Sequence 2 with (bottom row) and without (top row) considering the spatial relationships modeled by HMM. The black circles correspond to the location labels assigned to individual frames of the video sequence.

a) L_{15} training b) L_{15} test

Fig. 8. Appearance changes of location 15 between training and testing. There is no chair and the door is closed in the training view. The posters on the board are different between training and testing.

view recognition fail are resolved successfully using HMM model. Figure 8 shows the example of dynamic changes. Despite some classification errors in test sequences, the order of visited locations was correctly determined. For test sequence 2, where we exhibited some intentional deviations between the path taken during training and testing, the classification of frames 69-70 as location 14 is incorrect (Figure 7d) . The effect of HMM model can be examined by making all the probabilities in the transition matrix A uniform and essentially neglecting the knowledge of location neighborhood relationships. For comparison this is depicted in Figure 7a and 7b. Once the most likely location has been determined, we can estimate the relative pose of the camera with respect to the most likely representative view. This can be done by exploiting geometric relationship between two views captured by epipolar geometry. The detailed description of this stage in the context of the proposed application can be found in our earlier work [6].

V. Summary and Conclusions

We have demonstrated an approach for location recognition in indoor office like environments. The model of the environment is partitioned to individual locations and neighborhood relationships between them in the exploration stage. The individual locations were represented by SIFT features and location recognition was approached by feature matching between query and model views. We have presented a novel feature selection strategy, which exploited local information content and discriminability of the individual features and their associated descriptors. We have shown that by reducing the feature pool to 10% of the original size, we can achieve comparable performance to methods which use the original feature set. Further improvements were demonstrated by interpreting the quality of the features matches probabilistically and by endowing the environment with the HMM structure which exploits spatial relationships between locations. We are currently evaluating the effectiveness of the feature selection strategy in the context of other object and category recognition data sets.

Acknowledgments

The authors would like to thank D. Lowe for making available the code for detection of SIFT features. This work is supported by NSF grants IIS-0118732 and IIS-0347774.

References

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[2] J. Wolf, W. Burgard, and H. Burkhardt, "Using and image retrieval system for vision-based mobile robot localization," in *Proc. of the International Conference on Image and Video Retrieval (CIVR)*, 2003.

[3] R. Sims and G. Dudek, "Learning enviromental features for pose estimation," *Image and Vision Computing*, vol. 19, no. 11, pp. 733–739, 2001.

[4] A. Davidson and D. Murrray, "Simultaneous localization and map building using active vision," *IEEE Transactions on PAMI*, vol. 24, no. 7, pp. 865–880, 2002.

[5] S. Se, D. Lowe, and J. Little, "Global localization using distinctive visual features," in *Proc. of International Conference on Robots and Systems*, 2002, pp. 153–158.

[6] J. Kosecka and X. Yang, "Location recognition and global localization based on scale-invariant keypoints," in *Workshop on statistical learning in vision, ECCV*, 2004.

[7] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[8] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 264–271.

[9] A. Torralba, K. Murphy, and W. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *International Conference on Computer Vision and Pattern Recognition*, 2004.

[10] S. Helmer and D. Lowe, "Object class recognition with many local features," in *Workshop on Generative Model based Vision, CVPR*, 2004.

[11] G. Fritz, L. Paletta, and H. Bischof, "Object recognition using local information content," in *ICPR*, 2004.

[12] M. V. Naquet and S. Ullman, "Object recognition with informative features and linear," in *ICCV, Nice, France*, 2003.

[13] A. Pope and D. Lowe, "Probabilistic models of appearance for 3-d object recognition," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 149–167, 2000.

[14] C. Schmid, "A structured probabilistic model for recognition," in *Proceedings of CVPR, Kauai, Hawai*, 1999, pp. 485–490.