Active Vision-Based Robot Localization and Navigation in a Visual Memory

Gian Luca Mariottini and Stergios I. Roumeliotis

***Abstract*— We present a new strategy for active vision-based localization and navigation of a mobile robot in a visual memory, i.e., within a previously-visited area represented as a large collection of images. Vision-based localization in such a large and dynamic visual map is intrinsically ambiguous, since more than one map-locations can exhibit the same visual appearance as the current image observed by the robot. Most existing approaches are passive, i.e., they do not devise any strategy to resolve this ambiguity. In this work, we present an *active* vision-based localization and navigation strategy that can disambiguate the true initial location among possible hypotheses by controlling the mobile observer across a sequence of highly distinctive images, while concurrently navigating towards the target image. The performance of our active localization and navigation algorithm is demonstrated experimentally on a robot moving within a large outdoor environment.**

I. Introduction

In order for a robot to autonomously navigate towards a target location, it must be able to solve a set of related sub-tasks; first, it must *globally localize* itself, i.e., it must estimate its location with respect to some environment representation (map) from little or no a priori pose information. Second, it must *plan a path* to the target and, finally, it has to reliably *navigate* along the path.

In order to achieve the above goals, robots often rely on GPS. However, GPS cannot be used for navigating indoors or in urban environments with tall buildings, due to the limited or absent line-of-sight to satellites. Time-of-flight laser scanners have also been used, but they are expensive, and their weight, volume, and power requirements limit their use to large-size robots. Finally, disambiguating between map locations using laser data is very challenging, especially when planar laser scanners are used. Instead, we are here interested in using vision sensors that are more informative, increasingly inexpensive, quite compact and can be used for large-scale map-based localization [1], [2]. In particular, in this work we are focusing on the problem of robot localization and navigation using a visual map (constructed from a pre-recorded sequence of images) of the area it navigates in. Robot localization in a large visual map exhibits several challenges. Among these is *perceptual aliasing*, which happens when the image database contains similarly appearing locations, and results in more than one location having the same visual appearance as the current robot image. As a consequence, location recognition from a *single* query image is often not sufficient to uniquely localize the robot within the visual map. Instead, it will only provide a set of *candidate* locations (or hypotheses).

Most existing approaches that deal with this issue are *passive*, i.e., they do not control the camera/robot towards additional non-ambiguous observations that will help to discriminate the true initial location among possible hypotheses.

In order to address the problem of perceptual aliasing, in this paper we present a new *active* localization strategy that can uniquely localize the camera/robot in a large *visual-memory* map (organized as a Vocabulary Tree - VT [2]), while visually navigating to a target image through highly-distinctive image paths. The first innovative contribution of this work is in the design of a sequential Bayesian approach that can discard false location candidates by collecting additional observations during the robot motion. The second contribution is the design of a path planning strategy based on entropy that guides the robot towards the target image across highly distinctive (i.e., low-entropy) images in a graph representation of the VT. The main advantages of this approach are the ability to discriminate the true robot location among multiple hypotheses, as well as increased robustness when re-localization is necessary. It is important to note that our active strategy is here applied to the case when no 3D scene or camera pose-prior is available, but can be easily extended to use such additional information. The performance of our vision-based active localization and navigation algorithm is demonstrated experimentally on a robot moving in a large outdoor environment.

The remainder of the paper is organized as follows. In Section II, we begin with an overview of the related literature. Section III presents a summary of the proposed active-localization and navigation algorithm. Section IV describes the proposed location-recognition algorithm using the VT. Our entropy-based planning and the navigation strategy are presented in Section V. Section VI describes the sequential Bayesian approach for location discrimination. Experiments are presented in Section VII. Finally, conclusions and future research directions are discussed in Section VIII.

II. Related Work

In what follows, we provide an overview of the representative literature on localization and navigation based on image collections, and compare our contribution with relevant approaches.

Royer *et al.* [3] presented a strategy for autonomous vision-based navigation along a previously learned path. The image features tracked in the training video are used to

G. L. Mariottini is with the Dept. of Computer Science and Engineering, University of Texas, Arlington, TX 76019, USA email:gianluca@uta.edu

S. I. Roumeliotis is with the Dept. of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA email:stergios@cs.umn.edu

compute off-line a *global* 3D map of the environment. When re-navigating the learned path, the robot computes its pose from 3D-to-2D correspondences between the map and the observed scene, respectively. Fontanelli *et al.* [4] recently presented a similar strategy to visually navigate the robot across a path connecting totally different initial and desired views. While their navigation approach can deal with the field-of-view constraints imposed by monocular cameras, their localization algorithm also requires an *accurate* 3D global map of this large environment.

Goedemé *et al.* [5] relaxed the above assumptions and built a wheelchair robotic system that can automatically navigate in a pre-computed visual map made of a sparse set of panoramic images. However, and differently from our approach, their navigation strategy uses the relative camera orientation and the (scaled) translation computed from epipolar geometry decomposition, which can be sensitive to image noise [6]. In addition, their algorithm still requires an on-line estimate of the 3D *local* map of the observed scene.

In general, 3D map- and pose-reconstruction is not necessary, since moving from one reference image to the next can also be done by relying *solely* on visual information [7]. Recently, an interesting quantitative comparison of the performance of some appearance-based controllers has been presented by Cherubini *et al.* [8]. However, all of these controllers assume that the camera/robot is moving with positive linear velocity. Additionally, an estimate of the distance to each observed feature is still necessary, thus affecting the convergence properties of the proposed strategies. In [9], the authors present an image-based robot navigation strategy that uses visual memory. Their closed-loop control law does not require global 3D reconstruction. However, and differently from our approach, their strategy does not make use of any efficient and scalable vocabulary tree (VT) scheme.

As an extension over the previous approaches, Fraundorfer *et al.* presented in [10] a vision-based localization algorithm that globally localizes the robot using a VT and allows the robot to navigate a large image map. This visual-memory map is represented as a graph, in which nodes correspond to training images, and links connect similar images. However, their navigation strategy does not guarantee asymptotic convergence to the next node. Moreover, their experimental results are limited to the case of a robot navigating along a limited indoor path.

At this point, it is important to remark a key difference between *all* previous approaches and our strategy: They are passive, i.e., they do not control the mobile observer to actively seek new images that can reduce the localization ambiguity caused by different locations exhibiting similar visual appearance. In contrast, addressing the problem of perceptual aliasing by actively controlling the robot/camera is our main contribution. Specifically, we introduce a planning strategy that allows the robot to visually explore those image paths that maximize discriminability, while leading to the target image. As new distinctive images are collected, a sequential Bayesian approach is used to infer the most likely robot location in a graph representation of the VT. In this regard, our work is also relevant to the literature in active localization and vision-based location recognition.

In [11], Jensfelt *et al.* presented an active global localization strategy that uses Kalman filtering (KF) to track multiple robot pose hypotheses. This is done jointly with a probabilistic approach for evaluating hypothesis correctness. Their approach provides improvements over traditional grid-based strategies [12] because it can be used even with incomplete maps and with computational complexity independent on the size of the environment. However, a key difference to our work is that their navigation strategy simply guides the robot to places with a high concentration of map features, without taking into account their distinguishability.

Arbel and Ferrie presented in [13] a gaze-planning strategy that moves the camera to another viewpoint around an object in order to recognize it. The new measurements, accumulated over time, are used in a one-step-ahead Bayesian approach that resolves the object recognition ambiguity, while navigating an entropy map. Differently from their work, our probabilistic approach seeks informative images over an *extended time-horizon*. More recently, LaPorte *et al.* [14] proposed a computationally efficient viewpoint-selection strategy that, jointly with sequential Bayesian recognition, can disambiguate among competing hypotheses on both object class and pose. However, and differently from the two strategies described above, our approach can visually navigate the robot to the target image without requiring any camera pose information.

Other researchers recently proposed to address the problem of image ambiguity in location recognition, by either querying twice the VT (in order to detect at the second time more distinctive features) [15], or by incorporating additional knowledge about the camera location among consecutive images [16]. Such additional information can also be used in our algorithm. However, to the best of our knowledge, our work is the first to introduce active localization in a large image database, that seeks to guide the robot towards the target through a path of highly-distinctive images.

III. Algorithm Description

Fig. 1 shows the block diagram of the proposed active-localization algorithm.

We hereafter assume that the robot has previously visited the environment and has collected a set of *training images*, $\{\overline{\mathcal{I}}_i\}$ $(i = 1, ..., N)$. All of these training images are used to create the *visual-memory* map: Specifically, a set of SIFT image descriptors [1] $\{\overline{\mathcal{Z}}_i\}$ is first extracted off-line from all the images and is used to build the VT (cf. Sec. IV). Then, and similarly to [5], [9], [10], a graph representation, $\mathcal{G}$, is obtained from this image collection by linking two nodes/images that share a minimum number of SIFT matches, thus indicating the possibility for the robot to visually navigate among similar images. In order to measure the distinctiveness of an image in the VT, an entropy measure is computed and assigned to each node.

[1] Our method uses SIFT keypoints [17], [18], but can be extended to use other types of descriptors.

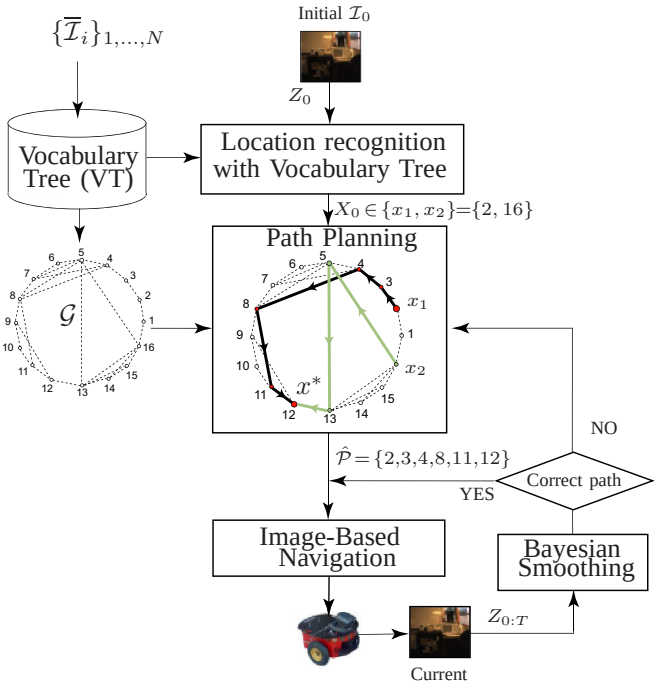


Fig. 1. Block diagram of our active vision-based localization in a visual map.

After this initial (off-line) phase, the robot is powered up somewhere in the environment and the active global localization phase starts. First, the on-board camera takes an *initial image* $\mathcal{I}_0$ and the SIFT descriptors, Z_0, are extracted from it and used to query the VT (cf. Sec. IV) to find the most similar candidate image in the visual map. Due to perceptual aliasing, this query may return a *set of* M_0 similar candidate images $\{\overline{\mathcal{I}}_j,\ j=1,...,M_0\} \subset \{\overline{\mathcal{I}}_1, ..., \overline{\mathcal{I}}_N\}$, $M_0 < N$, and the corresponding set of *candidate* node locations $\{x_j \in I\!N,\ j = 1, ..., M_0\} \subset \{1, ..., N\}$. We will represent a candidate location with the random variable X_0. For example, in Fig. 1, $M_0 = 2$ and $X_0 \in \{x_1, x_2\} = \{2, 16\}$.

After this global localization phase, a planning algorithm is used on the entropy-weighted graph $\mathcal{G}$, to find the least unambiguous image-path $\mathcal{P}_j$, from each candidate x_j to the goal node[2] x^* [cf. Sec. V]. Among all the M_0 possible paths, only the one with the lowest entropy, $\hat{\mathcal{P}}$, is selected and the *image-based navigation* algorithm starts to guide the robot across images in $\hat{\mathcal{P}}$.

As the robot moves from time step $t = 1$ to an intermediate time step $t = T$, a set of new SIFT-descriptor measurements $Z_{1:T} \triangleq \{Z_1, ..., Z_T\}$ is extracted from each new image and used to compute $p(X_0|Z_{0:T})$, i.e., the probability that the robot started from one of the initial candidate locations (or hypotheses). We do this by adopting *Bayesian smoothing* [19] over the set of the initial M_0 location hypotheses for X_0 (cf. Sec. VI). At time T, if the point of maximum of the posterior pdf still coincides with the index of the initially-chosen hypothesis for the starting location, then the visual navigation continues to the next node, and so on until the robot reaches x^*. Otherwise, the robot first visually navigates back to the initial image and then restarts the navigation process using a different hypothesis for the starting location.

Note that, instead of navigating back to $\mathcal{I}_0$, we could have chosen to re-plan the robot's path starting from the current location where it first realized the initial hypothesis was wrong. However, we chose not to do so for the following reasons: *(i)* our active localization strategy can rapidly detect a wrong hypothesis [cf. Sec. VII], so that the robot only needs to move back few steps; *(ii)* tracing back a traveled path is more robust compared to exploring a new (possible) path to the goal; *(iii)* fewer computations are required, since the paths from all the initial hypotheses have already been computed.

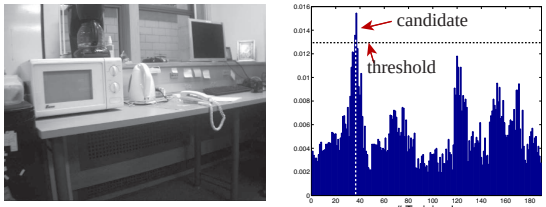
IV. Location Recognition Using a Vocabulary Tree

Given an initial image $\mathcal{I}_0$, our first step towards solving the global localization problem is to use Z_0 to query the visual map to find the most similar image. For this purpose, we employ a vocabulary-tree approach [2], due to its efficiency and speed, as well as its compact image representation and robustness to background clutter and occlusions. Our global localization algorithm uses a set of SIFT descriptors Z_0 to query the VT in order to determine X_0, i.e., the location (node) index of the training images most similar to $\mathcal{I}_0$.

We have used a tree with depth $L = 4$ and branching factor $k = 10$. The leaves of the tree correspond to quantized SIFT descriptors (visual words) obtained by clustering (with hierarchical $k-$means) all of the training data $\{\overline{Z}_i\}_{i=1,...,N}$. To each leaf node we assign a list of indices of the training images that had at least one descriptor assigned to it; this constitutes the *inverted file* [20]. Next, for each training image $\overline{\mathcal{I}}_i$, a sparse *document vector* $\mathbf{v}_{di} \in I\!N^W$, $i = 1, ..., N$, is created (with $W = k^L$). Each entry of $\mathbf{v}_{di}$ contains the number of descriptor vectors $\overline{Z}_i$ with a path through the corresponding leaf node, weighted using the inverse-document-frequency (IDF) [2]. The set Z_0 of descriptors extracted from robot's current image $\mathcal{I}_0$ is used to obtain the query vector $\mathbf{v}_0 \in I\!N^W$. After normalizing the vectors $\mathbf{v}_{di}$ and $\mathbf{v}_0$, we compute their L_2 distance $d_{0,i} = \|\mathbf{v}_0 - \mathbf{v}_{di}\|^2$, and use it as their similarity score. The output of each query process can be summarized by this score function that is defined over the domain of all the N training image indices.

In an ideal case, if queried with an image $\mathcal{I}_0 = \overline{\mathcal{I}}_i$, this score function will exhibit a unique peak localized at the index $X_0 = i$. However, due to presence of dynamic or commonly occurring objects (e.g., floor, brick walls, etc.) as well as changes in illumination and viewpoint, the similarity score may exhibit multiple peaks (see Fig. 2) or even a uniform distribution over a wide range of indices. These represent indeed the cases in which it is extremely difficult to uniquely localize the query image in the database. Fig. 3 shows an example in which the effects of perceptual aliasing are evident.

[2]We assume that the target image $\mathcal{I}^*$ is not ambiguous, so that the corresponding x^* is unique.



Query image ($\mathcal{I}_0$) $X_0 = 38^*$

(a) (b)

Candidate image ($\overline{\mathcal{I}}_{38}$)

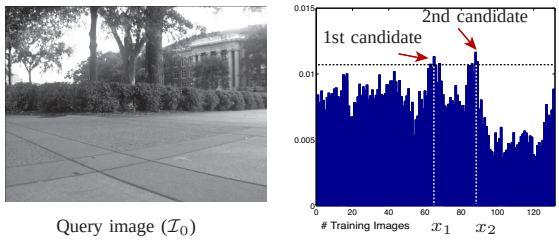
(c)

Fig. 2. Location recognition using a vocabulary tree with a *unique* match (indoor sequence). (a) The initial image $\mathcal{I}_0$ is used to query the VT ($N = 188$); (b) The resulting normalized score function shows only one candidate node ($X_0 = 38$) that exceeds a preset threshold (percentage of the maximum value). (c) The corresponding training image $\overline{\mathcal{I}}_{38}$, which is very similar to $\mathcal{I}_0$.

Note that the effect of perceptual aliasing may be reduced by using additional information (e.g., epipolar geometry constraints [15], [16]). However, there exists no method that can guarantee that the images will be matched unequivocally. To address this issue, in what follows we present our active-localization-based approach that controls the path of the robot so as to maximize the information acquired for disambiguating its location.

V. Appearance-Based Path-Planning and Navigation in the Entropy Space

In this section, we present a new strategy to plan a non-ambiguous visual path, from an initial image $\mathcal{I}_0$ to a target image $\mathcal{I}^*$.



Query image ($\mathcal{I}_0$)

(a) (b)

1st Candidate ($\overline{\mathcal{I}}_{64}$) 2nd Candidate ($\overline{\mathcal{I}}_{88}$)

(c) (d)

Fig. 3. Location recognition using a vocabulary tree with *multiple* matches (outdoor sequence). (a) Query image $\mathcal{I}_0$; (b) The score function shows two candidates $X_0 = \{x_1, x_2\} = \{64, 88\}$; (c)-(d) While $\overline{\mathcal{I}}_{64}$ appears very similar to $\mathcal{I}_0$, $\overline{\mathcal{I}}_{88}$ does not. However, both images have very similar feature representation in the VT.

Similarly to previous works [10], [5], we use a graph representation $\mathcal{G} = (V, E)$ of the image collection (see Fig. 1), in which each node $i \in V$ is associated to a training image $\overline{\mathcal{I}}_i$, ($i = 1, ..., N$). Moreover, a link between two nodes is established when the associated images share a minimum number of SIFT keypoint matches (20, in our implementation), thus indicating that the robot can effectively servo between the two associated images. Once the graph $\mathcal{G} = (V, E)$ is built, then it can be used to compute a traversable path from the starting to the target node/image.

A possible path-planning strategy would be to apply Dijkstra's algorithm [19] on the graph $\mathcal{G}$ using as distance between two nodes the inverse of the number of SIFT matches. However, such strategy cannot cope with perceptual aliasing. As an illustrative example, consider the case of Fig. 4 where a sequence of images was taken in front of two similarly-appearing buildings. Even if the image-pairs along the path $\mathcal{P}_1$ (dashed thick curve) share (among them) a large number of SIFT matches ($s\uparrow$), navigation through the set A can be confused with going through set B. This ambiguity can negatively affect the navigation performance, for example, when the robot needs to re-localize after unexpected disturbances (e.g., sudden illumination changes, image blur, etc.). If this re-localization process takes place within one of these ambiguous image sets, then the robot might start to follow erroneous paths that will drive it far from its goal.

In order to address this challenge, we propose to augment the graph with information, assigned to the edge weights in $\mathcal{G}$, about the entropy of the training images, and use it to plan a path through the most distinctive (i.e., low entropy) images. In the example of Fig. 4, $\mathcal{P}_2$ (continuous thick curve) will be chosen as the least ambiguous path, since its edge weights have low-entropy values w. In contrast, the edge weights for the images in the A- and B-segments of path $\mathcal{P}_1$ have high values of w (not shown in figure due to limited space).

The main steps of this algorithm are detailed hereafter. Specifically, to each node $i = 1, ..., N$ we assign a score that measures the *distinctiveness* of each node/image in the entire vocabulary tree. This score is given by the *entropy* $H_i \in \mathbb{R}$,

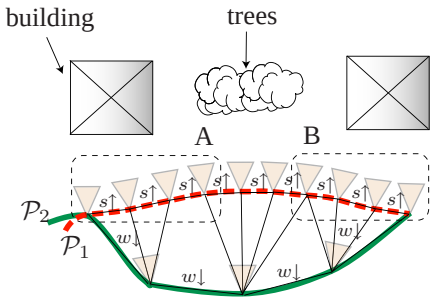


Fig. 4. The sets of views A and B along the path $\mathcal{P}_1$ (dashed thick curve) contain ambiguous images of two identical buildings. By using the entropy measure for setting the weights of the graph links (continuous black lines), the planner will correctly discard the ambiguous path $\mathcal{P}_1$ (even if its images contain a large number of SIFT descriptors $s\uparrow$). Instead, it will select path $\mathcal{P}_2$ (continuous thick curve) whose edges have low values of entropy ($w\downarrow$).

defined as

$$H_i \triangleq H(X|\overline{Z}_i) = -\sum_{j=1}^{N} p(X{=}j|\overline{Z}_i)\cdot \log p(X{=}j|\overline{Z}_i), \quad (1)$$

where $p(X|Z_i)$ is obtained as the result of querying the VT with each training image $\overline{\mathcal{I}}_i$ over all the graph nodes $X = j$, $(j = 1, ..., N)$. The result of the query is then normalized so as to represent a pdf.

A low value of entropy indicates a highly-distinctive image, while higher entropy values indicate a node associated to an image similar to many others (i.e., not an informative image). The entropy H_i is then used to compute an entropy-related weight w_{ij} for each edge in E between two nodes (i, j) that share a minimum number of SIFT descriptors. Each weight w_{ij} is computed as the average of the entropy at the nodes i and j, i.e.,[3]

$$w_{ij} = \frac{H_i + H_j}{2}. \quad (2)$$

Once the edge weights w_{ij} are computed, Dijkstra's algorithm is used to select the least uncertain path $\hat{\mathcal{P}}$.

So far we have considered the case of a single hypothesis for the initial location. When *multiple* initial-location candidates $X_0 = \{x_1, x_2, ..., x_{M_0}\}$ exist, we follow the same procedure described above to compute the weights w_{ij} for $\mathcal{G}$. Next, Dijkstra's algorithm is used to compute a set of M_0 possible paths, for all the initial hypotheses in X_0. Finally, only the path $\hat{\mathcal{P}}$ with the minimum entropy is chosen.

A. Image-based Visual Route Navigation

Once the candidate route $\hat{\mathcal{P}}$ is generated, a vision-based navigation strategy is used to navigate to the next node/image and towards the goal. To achieve this, we use our epipole-based visual servoing (IBVS) algorithm described in [21]. This algorithm has certain desirable advantages compared to alternative approaches. In particular, it is free from local minima and singularities (typically encountered in other IBVS schemes based on image Jacobians [7], [8]). Secondly, it does not require any additional geometric knowledge about the 3D scene. Finally, it guarantees *global* asymptotic convergence to the desired configuration, even in the case of unknown focal length.

VI. Location Discrimination

Among the set of candidate paths $\{\mathcal{P}_0, \mathcal{P}_1, ..., \mathcal{P}_{M_0}\}$ (each of them made of highly distinctive images), the strategy proposed in the previous section selects the path $\hat{\mathcal{P}}$ that is the least ambiguous path to travel. As the robot navigates along $\hat{\mathcal{P}}$, additional information is still needed in order to uniquely discriminate the correct *initial robot location* among all the M_0 initial hypotheses in X_0. In order to discriminate the assumed initial location in the visual map, our active strategy collects additional images from the moving on-board camera. This new data is used in a sequential Bayesian approach whose goal is to maximize the belief over X_0 for the initial location. In particular, as the robot moves, the camera collects *new* measurements, $Z_{0:T}$, that are used to evaluate the posterior $p(X_0|Z_{0:T})$, by formulating our problem as Bayesian smoothing [19]:

$$\begin{aligned} p(X_0|Z_{0:T}) &= \frac{p(X_0, Z_0, Z_{1:T})}{p(Z_0, Z_{1:T})} \\ &= \frac{p(Z_{1:T}|X_0)p(X_0|Z_0)}{p(Z_{1:T}|Z_0)} \\ &\propto p(X_0|Z_0)\, p(Z_{1:T}|X_0)\,, \end{aligned} \quad (3)$$

where $p(X_0|Z_0)$ is the prior pdf over the domain of the candidate nodes X_0, given the initial image. We assume $p(X_0|Z_0)$ is uniform, due to the selection of the initial location candidates by thresholding the normalized score function (cf. Sec. IV)[4]. This choice reflects the assumption that all the hypotheses X_0 have the same probability. While this assumption might change when other measurements are available (e.g., SIFT matches), we note that this will not cancel perceptual aliasing. The likelihood $p(Z_{1:T}|X_0)$ in (3) can be written as

$$p(Z_{1:T}|X_0) = \sum_{X_1} p(Z_{1:T}|X_1, X_0)\, p(X_1|X_0)\,, \quad (4)$$

where X_1 represents the nodes that can be reached by the robot starting from X_0 through visual servoing.[5] Employing the Markov assumption, (4) can be written as

$$p(Z_{1:T}|X_0) = \sum_{X_1} p(Z_1|X_1)\, p(Z_{2:T}|X_1)\, p(X_1|X_0). \quad (5)$$

Note that $p(X_1|X_0)$ represents the motion model in the planned path from nodes in X_0 to nodes in X_1. Since the motion is planned on a graph, and the visual servo algorithm is globally convergent [cf. Sec. V-A], we can assume that this is also uniform over all possible nodes in X_1 linked to nodes in X_0. $p(Z_{2:T}|X_1)$ represents the recursive call to (5) itself.

The pdf $p(Z_1|X_1)$ in (5) represents the measurement likelihood. In order to find an expression for it, consider the case in which the robot has moved to a specific node in X_1. In this case, a vocabulary-tree query using the current camera measurements Z_1 will return a normalized score function that matches with the one obtained by querying the VT using the training measurements $\overline{Z}_1$ (associated with a specific node in X_1). We use the Jeffrey divergence J [14] as a distance measure between the expected query $h' \triangleq p(X_1|\overline{Z}_1)$ and the current one $h \triangleq p(X_1|Z_1)$, i.e.,

$$J(h||h') = D(h||h') + D(h'||h) \quad (6)$$

where $D(h||h')$ is the Kullback-Leibler divergence [22] given by

$$D(h||h') = \sum_{X_1} h(X_1) \log \frac{h(X_1)}{h'(X_1)}\,, \quad (7)$$

[3]Note that the sum of the weights from an initial to a final node equals the sum of the entropies of all the intermediate nodes plus a constant term equal to the average entropy of the initial and final nodes.

[4]In our case the threshold is set as a percentage of the maximum value of the score function.

[5]In our implementation we considered nodes in the graph connected with X_0 up to distance 2.

and finally we can model the likelihood $p(Z_1|X_1)$ as

$$p(Z_1|X_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{J(h||h')^2}{2}}. \quad (8)$$

The peak of the posterior distribution $p(X_0|Z_{0:T})$ computed as in (3) from (4)-(8), will most likely be at the node hypothesis x_j which corresponds to the true initial location of the robot. In case that this value is different from the initial hypothesis selected for starting the path, the robot will use the images stored up to T to servo back to the initial image, and start to navigate along the new path which has x_j as its initial node.

VII. Experimental Results

In this section, we present experimental results to test the effectiveness of the proposed localization and image-based navigation algorithms. Our experimental platform is a nonholonomic Pioneer 3-DX. Our robot is only equipped with a IEEE 1394 camera that captures in real-time 1024×768 pixel images.

A. Vision-based localization performance

We first present the performance of the initial global localization algorithm using the VT. The robot was driven to capture two sets of images, from an indoor (office) and an outdoor sequence. In particular, the latter one is of approximately 200 m, with changing sunlight conditions due to trees and building shadows.

For each image sequence, a vocabulary tree is generated ($k = 10$, $L = 4$). Then, for a given number of realizations and for an increasing number of SIFT descriptors (randomly-chosen for each database image), the VT was queried and the best matching image obtained. The percentage of success in localizing the correct image for the office sequence is depicted in Fig. 5(a) and, as expected, shows that the best retrieval performance (97%) is obtained when *all* the descriptors in the image are used for the query. By decreasing the number of descriptors, the retrieval performances degraded to 70%. Similar performances is observed for the outdoor sequence (see Fig. 5(b)).

We also implemented a *stop-list* [2] which penalizes those images from the query that have a high number of descriptors commonly occurring in the database (e.g., carpet or pavement textures). This is done by blocking from scoring those inverted files that are above a certain length. In our implementation, the use of a stop-list led to an improvement in the retrieval quality, as shown in our results.

The obtained results suggest that trying to reduce the query time to the VT by decreasing the number of SIFT descriptors used, will increase the risk of perceptual aliasing, thus motivating the use of our *active* localization strategy. Other approaches (e.g., stop-list) can only alleviate, but not eliminate, the perceptual aliasing.

B. Active path planning and navigation

We hereafter present the results obtained for all three phases (location recognition, path planning and visual navigation) of our active localization algorithm (Sec. VI). In particular, we present two experiments that are representative of the single- and multiple-location candidate cases, respectively. The vocabulary tree used in this case is the same one generated for the results described in Sec. VII-A.

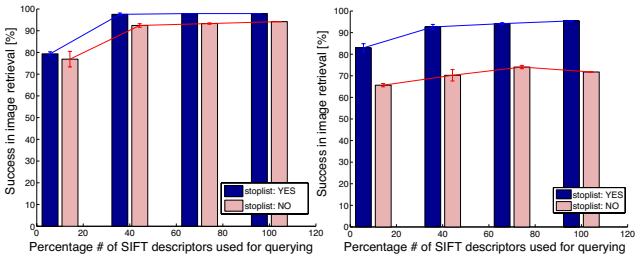
In the first experiment, the camera/robot observes an initial image $\mathcal{I}_0$ that is uniquely associated by the VT to node 54 in the graph (see Fig. 6(a)). Starting from this node, Dijkstra's algorithm computes a path $\mathcal{P}$ to the desired image $\mathcal{I}^*$ (node 88) comprising a sequence of indices of images to be visually navigated. The resulting robot/camera motion is shown in Fig. 6(a), superimposed to the satellite view of the experimental site.

In the second experiment, we consider the case in which the localization algorithm, queried with the initial view $\mathcal{I}_0$, provides *two* initial view hypotheses, $\overline{\mathcal{I}}_1$ and $\overline{\mathcal{I}}_{68}$, corresponding to the two nodes $x_1 = 1$ and $x_2 = 68$ in the graph (see Fig. 6(b)) (the correct result from the query should have been x_1). For each candidate, a path towards the desired image in the topological graph is generated and x_2 is erroneously selected as the hypothesis for the robot's initial location. Consequently, the path $\hat{\mathcal{P}} = \mathcal{P}_2$ is used for navigating towards the goal. This wrong initial belief on the robot's initial location makes the robot move forward (and away from the goal image $\mathcal{I}^*$), as it would have been necessary for going from $\overline{\mathcal{I}}_{68}$ to $\mathcal{I}^*$. This (initial) phase shows that perceptual aliasing can indeed defeat a simplistic visual navigation based *only* on queries to a VT. The forward motion for $T = 2$ steps is represented in the inset of Fig. 6(b). At this point, as described in Sec. VI, our Bayesian approach uses the collected data $Z_{0:2}$ and evaluates the posterior $p(X_0|Z_{0:2})$ for each of the two candidate nodes (cf. Sec. VI). This yields $p(X_0|Z_{0:2}) = \{0.938, 0.062\}$, which clearly indicates that the initial node location $x_1 = 1$ was the correct hypothesis (instead of $x_2 = 68$). The robot then visually navigates back to the initial image and selects the correct path $\mathcal{P}_1$, which finally leads the robot to $\mathcal{I}^*$, as shown in the bottom trajectory of Fig. 6(b).

Table I contains the posterior distribution $p(X_0|Z_{0:T})$ obtained for increasing time T and shows that the probability of the correct location x_1 increases rapidly as more images are considered for computing the posterior distribution.

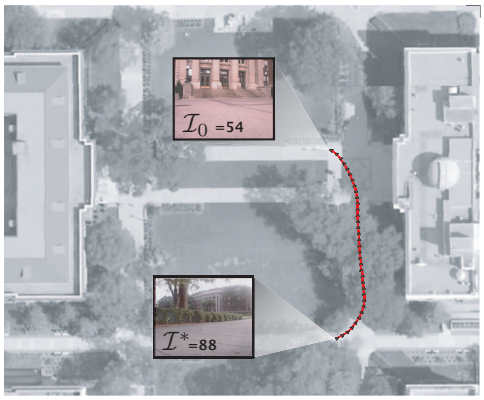
VIII. Conclusions

Location recognition in large and dynamic environments is intrinsically ambiguous. Existing approaches do not exploit

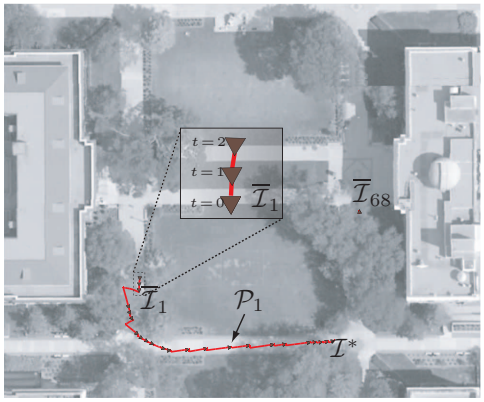


(a) (b)

Fig. 5. *Localization performance*. (a) Office sequence: percentage of success; (b) Outdoor sequence: percentage of success.



(a)



(b)

Fig. 6. *Experimental results*: (a) In the first experiment the initial image is correctly retrieved and a path generated. The robot then navigates (red line) along the selected way-points comprising the least ambiguous path. (b) In the second experiment the initial image $\mathcal{I}_0$ currently observed by the robot has two possible matches in the database, images $\{\overline{\mathcal{I}}_1, \overline{\mathcal{I}}_{68}\}$. Bayesian smoothing estimation is used to discriminate the correct position after $T = 2$ robot motion steps.

	$p(X_0 = x_1 \mid Z_{0:T})$	$p(X_0 = x_2 \mid Z_{0:T})$
T=0	0.5	0.5
T=1	0.820	0.180
T=2	0.938	0.062
T=3	0.962	0.038
T=4	0.976	0.024

TABLE I

POSTERIOR PDF FOR INCREASING TIME T, OVER THE DOMAIN OF ALL ($M_0 = 2$ IN THIS CASE) CANDIDATE NODES.

the possibility of controlling the camera/robot and do not leverage new camera measurements to disambiguate the true initial robot location. In this work, we presented a new method for active robot localization, planning, and navigation in a large-scale image map (represented as a vocabulary tree). In particular, we adopted a sequential Bayesian approach that allows us to eliminate the localization ambiguity by exploiting additional camera measurements over an extended time horizon, while navigating towards a target image, and along a non-ambiguous (i.e., low entropy) visual path. The performance of our localization, planning and navigation algorithms is demonstrated experimentally on a robot moving in a large outdoor scenario.

Our future work will address the extension of our experimental results to a city-scale scenario as well as to visit previously unexplored locations. We are also planning to design new methods for graph representation of the visual map that will include both geometric constraints between images, as well as discriminative features [16].

REFERENCES

[1] J. Wang, R. Cipolla, and H. Zha, "Vision-based global localization using a visual vocabulary," in *Proc. of the IEEE Int. Conf. on Robotics and Automation*, April 2005, pp. 4230 – 4235, barcellona, Spain.

[2] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. 2, Jun. 17-22 2006, pp. 2161–2168, New York, NY.

[3] E. Royer, M. Lhuillier, M. Dhome, and T. Chateau, "Towards an alternative GPS sensor in dense urban environment from visual memory," in *Proc. 14th British Machine Vision Conference*, Sept., 7-9th 2004, Kingston College, London, UK.

[4] D. Fontanelli, A. Danesi, F. A. W. Belo, P. Salaris, and B. A., "Visual servoing in the large," *International Journal of Robotics Research*, vol. 28, no. 6, pp. 802–814, June 2009.

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

[6] F. Chaumette, "Image moments: a general and useful set of features for visual servoing," *IEEE Trans. on Robotics*, vol. 20, no. 4, pp. 713–723, August 2004.

[7] F. Chaumette and S. Hutchinson, "Visual servo control, part i: Basic approaches," *IEEE Robotics and Automation Magazine*, vol. 13, no. 4, pp. 82–90, December 2006.

[8] A. Cherubini, M. Colafrancesco, G. Oriolo, L. Freda, and F. Chaumette, "Comparing appearance-based controllers for nonholonomic navigation from a visual memory," in *Proc. of the IEEE Int. Conf. on Robotics and Automation*, Kobe, Japan, May 12-17 2009.

[9] A. Remazeilles and F. Chaumette, "Image-based robot navigation from an image memory," *Robotics and Autonomous Systems*, vol. 55, no. 4, pp. 345–356, April 2007.

[10] F. Fraundorfer, C. Engels, and D. Nister, "Topological mapping, localization and navigation using image collections," in *Proc of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Nov. 2007, pp. 3872–3877, San Diego, CA.

[11] P. Jensfelt and S. Kristensen, "Active global localization for a mobile robot using multiple hypothesis tracking," *IEEE Trans. on Robotics and Automation*, vol. 17, no. 5, pp. 748–760, Oct. 2001.

[12] W. Burgard, D. Fox, and S. Thrun, "Active mobile robot localization by entropy minimization," in *Proc. of the Euromicro Workshop on Advanced Mobile Robots*, Los Alamitos, CA, 1997, pp. 155–162.

[13] T. Arbel and F. Ferrie, "Entropy-based gaze planning," *Image and Vision Computing*, vol. 19, no. 11, pp. 779 – 786, 2001.

[14] C. Laporte and T. Arbel, "Efficient discriminant viewpoint selection for active bayesian recognition," *Int. J. Comput. Vision*, vol. 68, no. 3, pp. 267–287, 2006.

[15] H. Kang, A. Efros, M. Hebert, and T. Kanade, "Image matching in large scale indoor environment," in *Proc. of the IEEE Comp. Vis. and Pattern Recognition*, June 2009, pp. 33–40, Miami, FL.

[16] F. Li and J. Kosecka, "Probabilistic location recognition using reduced feature set," in *Proc. of the IEEE Int. Conf. on Robotics and Automation*, May 2006, pp. 3405–3410, Orlando, FL.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Comput. Vision*, vol. 60, pp. 91–110, 2004.

[18] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[19] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ, 2003.

[20] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. of the Int. Conference on Computer Vision*, vol. 2, Oct. 2003, pp. 1470–1477, Beijing, China.

[21] G. L. Mariottini, G. Oriolo, and D. Prattichizzo, "Image-based visual servoing for nonholonomic mobile robots using epipolar geometry," *IEEE Trans. on Robotics*, vol. 23, no. 1, pp. 87–100, Feb. 2007.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.