

# Vision Based Topological Markov Localization

Jana Košecká

Department of Computer Science  
George Mason University  
Fairfax, VA 22030  
Email: kosecka@cs.gatech.edu

Fayin Li

Department of Computer Science  
George Mason University  
Fairfax, VA 22030  
Email: fli@cs.gmu.edu

**Abstract**—In this paper we study the problem of acquiring a topological model of indoors environment by means of visual sensing and subsequent localization given the model. The resulting model is inferred using simple global image descriptor and consists of a set of locations and neighbourhood relationships between them. Each location in the model is represented by a collection of representative views selected from a temporally subsampled video stream captured by a mobile robot during exploration. The spatial relationships between individual locations are modelled by a Hidden Markov Model (HMM). The quality of the acquired model is tested in the localization stage by means of location recognition: given a new view or a sequence of views, the most likely location where that view came from is determined.

## I. INTRODUCTION AND RELATED WORK

Visual sensing greatly enhances the functionality of mobile robots, facilitates variety of tasks and enables better human-robot interaction. The acquisition of unknown environment models, navigation and pose maintenance are one of the essential capabilities of a mobile robots. The two most commonly used representations which the vision-based approaches strive to attain are metric and topological models. Within these broad categories the existing techniques further differ depending on type of information extracted from images, assumptions about environment, means of model acquisition and localization. In the absence of prior geometric model of the environment the techniques for simultaneous localization and acquisition of metric models typically relied on the detection and tracking of the point features tied with the structure and motion recovery techniques in a recursive setting. Examples of these methods using single camera [1] or trinocular stereo [2] enabled continuous pose maintenance of the mobile robot. Alternative model based techniques focused on the problem of continuous matching of the observed data with the model and pose maintenance [3].

Among topological models the most commonly used topological models were those induced by visibility regions associated with the environmental landmarks. In the earlier approaches the model of landmarks was typically given and their location was either known a-priori or obtained during exploration [4]. In most of these instances artificial landmarks have been considered to simplify the issues of landmark recognition and simultaneously enable reliable estimation of the relative pose of the robot with respect to a landmark [5]. The techniques which tried to bypass the choice of artificial landmarks have been mainly motivated by approaches used

for object recognition. One of main concerns of these methods is the choice of image representation, which could guarantee some amount of invariance with respect to variations in pose, illumination and scale and be robust to partial occlusion and clutter. The resulting representations either comprise of descriptors computed locally at salient image locations or globally over the entire image. In the localization context, the local methods were typically used for landmark selection and recognition. The local representations proposed in the past comprised of a set of salient regions and their associated rotationally or affinely invariant feature descriptors [6], [7], [8], which would facilitate effective matching. Examples of global descriptors derived from local responses of filters at different orientations and scales [9] or multi-dimensional histograms [10] computed over the entire image. These led to more holistic types of representations of environments in terms of locations corresponding to regions with high appearance similarity. In both cases the final representations were commonly obtained by additional principal component analysis (PCA), clustering or alternative dimensionality reduction methods. In case of omni-directional views (local) PCA based techniques were applied successfully for topological model acquisition, thanks to small variations of the image appearance within a location [11], [12]. The obtained topological models represented by a graphs also varied in the semantics associated with individual nodes and edges. In some instances the nodes corresponded to segments of trajectories where the set of interest points can be successfully tracked [13] or visibility regions associated with the landmarks [4].

The problem of building a metric model and simultaneous localization (SLAM) using solely visual sensing has been demonstrated successfully in case of smaller environments (single room). The applicability of the existing methods to the problems of the scale comparable to those achieved by laser range sensors is very difficult due to the nature of visual measurements. It has been noted previously [14] that the representations of environments at different spatial scales are often advantageous, both from the perspective of model building and localization as well as navigation given the model. These types of hybrid models have been already explored previously using ultrasound sensing [15].

Motivated by need of different representations at different spatial scales, we strive to develop components which such hybrid models would be composed of. In the final model

we envision the coarse structure of the environment will be represented in terms of individual locations, each characterized by a set of representative views. Within the location we will endow the model with a local geometry relative to the set of representative views. In this paper we discuss a method for acquiring the coarse structure of the environment in terms of its topology with the localization being solved by means of location recognition.

In this paper we extend our previous work published in [16] where we reported recognition performance results based on a single view recognition. We extend the work by exploring slightly different image representations, carrying out larger experiment and demonstrating how to exploit the temporal context to improve the classification results. The use of temporal context is motivated and closely related to recently published work by [17] on using contextual information for place and object recognition. Their approach considered slightly different image representation and used hand labelled data set for learning the observation likelihood of individual locations. This strategy was appropriate in the context of wearable computing application authors considered.

## II. APPROACH

We propose to represent the large scale structure of the environment in terms of its topology captured by a *location graph*. In the presented work we focus on the localization scheme enabled by recognition of locations, which loosely correspond to the regions in the robot's work space which are similar in their appearance. The neighbouring locations are typically separated by regions where significant robot navigation decision have to be made; such as hallway intersections, corners and doorways. We first identify a simple image based representation and distance metric that enables us to compare two views. Towards this end we adopted gradient orientation histograms. The histograms are sufficiently discriminant between individual locations and also capture the similarity between the views which are perceptually close. In order to obtain higher discrimination capability and retain some of the spatial relationships present in the image, separate histograms are computed for 4 different image sub-regions (quadrants) and concatenated together.

The frames of the temporally subsampled video sequence obtained in the exploration stage are then partitioned and labelled as belonging to different locations. After obtaining a labelled set of views we carry out a vector quantization and obtain more compact representation for each location in terms of representative feature vectors. In the classification stage we determine given a previously unseen view, what is the location it most likely comes from. Low location likelihoods which in the presence of thresholds would yield classification errors are resolved in the second stage by exploiting the temporal context and spatial relationships between neighbouring locations modelled in terms of Hidden Markov Model (HMM).

### A. Appearance Based Distance Measure

The task of imposing a discrete structure on a quasi-continuous space of visual observations requires a definition of a distance measure in the image appearance space. Given two views  $I_1, I_2$  we denote the distance between them  $d(I_1, I_2)$ . In our case we exploit the constraints of man-made environments and seek an image descriptor and distance measure, which would suitably capture the variations we encounter. Exploiting the observation that the majority of gradient directions is aligned with the axes of the world coordinate frame [18], we use the information provided by image gradient orientation as distinguishing characteristic of an individual location.

In order to obtain image representation which captures the essential appearance of the location and is at the same time robust to changes in image brightness we explore two different strategies. In one case the gradient orientation histograms are computed only for the pixels with the magnitude in the top 4%. This is achieved by first computing the image derivatives  $[I_x, I_y]^T = [\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}]^T$  followed by non-maximum suppression and connected component analysis. Connected component analysis generates the final edge map and enables us to eliminate small connected components which do not correspond to essential features of the environment. In the second case we eliminate the gradient magnitude threshold by weighting the contribution of each pixel  $I(x, y)$  to orientation histogram by its gradient magnitude  $M(x, y) = \sqrt{I_x^2 + I_y^2}$ , which has been initially normalized to  $[0, 1]$ . Both choices yielded comparable results. Examples of representative views and their gradient orientation histograms are in Figure 1. The most notable characteristic of this simple feature is that

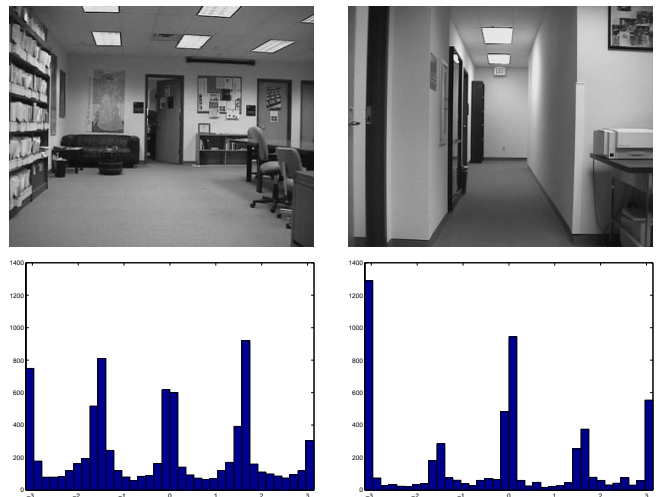


Fig. 1. Locations  $l5$  (left) and  $l3$  (right) of the 4<sup>th</sup> floor and their gradient orientation histograms.

it properly reflects the changes in image appearance due to portions of the environment leaving the field of view; characteristic which intuitively corresponds to the change of location. Alternative global descriptors have been used previously in the context of scene recognition [19], where

histograms of dominant texture elements have been used for categorization of images to different categories (e.g. bathroom, cities, mountains). In [9] authors used Gabor jets followed by dimensionality reduction step for recognition of indoor scenes. For indoors environments, the orientation histograms give us sufficient level of invariance with respect to variation in pose and illumination conditions. They also properly reflect commonly encountered presence of corners, doors, and bulletin boards and hence discriminate the locations with different appearance. Furthermore due to the dominant rectilinear structure and topology of man-made indoors environment, individual locations are typically approached with some canonical orientations and hence a complete rotational invariance is not necessary.

Once the gradient orientation histogram has been selected as a feature vector, we need to determine the best way to compare different features. Towards this end we use  $\chi^2$  empirical distance measure between two distributions, which is defined in the following way

$$\chi^2(h_i, h_j) = \sum_k \frac{(h_i(k) - h_j(k))^2}{h_i(k) + h_j(k)} \quad (1)$$

where  $k$  is the number of histogram bins. We have carried out several experiments in order to test the discrimination capability of the chosen feature vector and associated  $\chi^2$  distance measure. One set of experiments consisted of collecting set of images from the 4<sup>th</sup> floor of our building along the path depicted in Figure 2. The images were taken by still

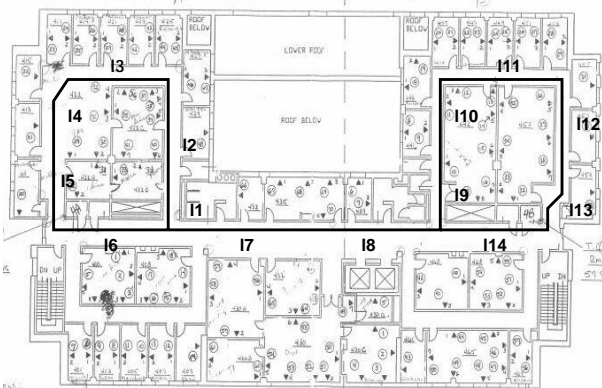


Fig. 2. Floor plan of the 4<sup>th</sup> floor; exploration route and labels associated with individual locations labelled by hand.

digital camera about 2 meters apart, with the orientation in the direction of mobile robot heading. In this dataset the heading direction was in most cases aligned with the principal directions of the world coordinate frame or perpendicular to it; assumption which we plan to relax in the future<sup>1</sup>.

<sup>1</sup>Additional processing is required in case the proposed method is being applied to a video sequence [16]. In case of video sequence the histograms were smoothed both spatially and temporally and the local temporal histogram mean has been subtracted from the data. This enabled us to eliminate some of the aliasing effects of the digital camcorder and transient effects corresponding to non-informative regions.

Our previous experiments [16] demonstrated that due to the global nature of the histograms certain locations are not well discriminable and in the presence of dynamic changes in the environment cause misclassification. Motivated by this observation we incorporated slightly modified similarity measure which retains some of the spatial information present in the image: instead of computing the histogram globally over the whole image, we divide the image into 4 quadrants and compute the local orientation histograms for each sub-image and stack them together. The discrimination capabilities of these two representations are depicted in Figure 3. The affinity matrices depict all pairwise comparisons between the views using  $\chi^2(h_i, h_j)$  and temporal distance profiles measure distances between two consecutive views of the sub-sampled video sequence  $\chi^2(h_{t-1}, h_t)$ . Note that the affinity matrices

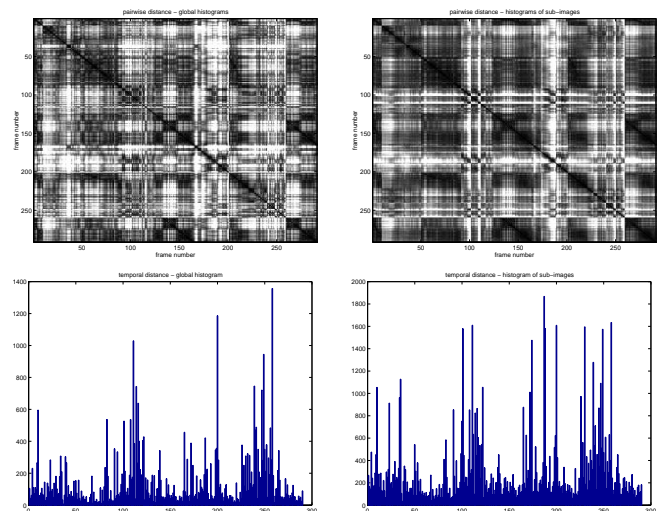


Fig. 3. The pairwise and temporal comparison of orientation histograms of images taken by still digital camera: global histograms image representation (left) and sub-image histograms (right).

in Figure 1 have clear distinguishing clusters corresponding to the images collected along particular path at locations  $l_1, l_2, \dots, l_N$ . The clear boundaries between the locations are represented by peaks in the temporal histogram comparison plot. These were typically caused by sudden change of robots' heading or more gradual change in the location appearance. Note that the affinity matrix reveals that certain sub-sequences are similar to each other in spite of the fact that they belong to different locations. This is not surprising given the structure of the office like indoors environments, where certain locations (e.g. corridors) appear very similar. Note that the clusters in the affinity matrix right are more pronounced. This indicates that the stacked sub-image histograms are more discriminative. Both the temporal distances and the affinity matrices reveal the fact that image orientation histogram capture reasonably well the appearance similarity between locations. In order to obtain sparser representation for each location in terms of fewer number of views and consequently solve the localization problem we next apply a vector quantization step.

## B. Learning phase

The assignment of individual views to clusters in our case can be induced from the temporal relationships acquired during exploration. We have examined two different methods for initial label assignment; automatic and by hand and obtained comparable recognition results which we report in the following section. The automatic location label assignment was obtained by searching for the peaks in the temporal histogram distance profile. First coarse peaks were detected and further refined using an adaptive threshold and the minimum separation distance criterion, yielding a set of dominant peaks. Note that in Figures 1 the dominant peaks are quite distinguishable, clearly separating images associated with the individual locations. In the experiments reported in this paper the location labels are assigned by hand due to the fact that the exploration path contains several cycles, which we do not currently detect automatically. In the near future we plan automate this step by exploiting the use of relative entropy measures for determining whether the location was visited previously. After generating prototype histograms of 291 frames captured along the path in 2, we partitioned this sequence into individual locations and obtained 30 classes. Due to the rectilinear structure of indoors environments and presence of large number of corridors, the orientation was coarsely quantized into 4 different directions (N, W, S, E). Although the orientation information is not being used in clustering stage, being at the same position with two dramatically different orientations corresponds to being at two different locations due to the often large change in image appearance. The goal of the learning stage in our case is to obtain representation for each class in terms of smaller number of prototype views.

For this purpose we have tested two different methods. First we used Learning Vector Quantization technique (LVQ). LVQ examines the data represented as vectors  $\mathbf{x}_i \in \mathbb{R}^n$  and in an iterative fashion builds a set of prototype vectors, called *codebook* vectors, that represent different regions in the  $n$ -dimensional feature space. Initially, the algorithm chooses randomly a set of prototype vectors to represent the data and refines them iteratively using the exemplars from the labelled dataset. Given an input sample  $\mathbf{x}_i \in \mathbb{R}^n$  the closest codebook vector  $\mathbf{m}_c$  is adjusted according to the following update rule:

$$\mathbf{m}_c(t+1) = \mathbf{m}_c(t) \pm \alpha(t)(\mathbf{x}_i - \mathbf{m}_c(t)) \quad (2)$$

where  $\mathbf{x}_i$  is the sample input and  $\mathbf{m}_c(t)$  represents the closest codebook at step  $t$  of the iteration. The sign  $\pm$  of the update depends of whether  $\mathbf{x}_i$  belongs to the same class as  $\mathbf{m}_c$  or not. As a result LVQ represents the final set of clusters with a small number of codebook vectors which are properly placed in each zone of the feature space in such a way that the decision borders between the zones are approximated by the nearest neighbour rule. The resulting codebook vectors do not try to approximate the density function of the class samples, but directly define class borders according to the nearest neighbour rule. We used the existing implementation of LVQ\_PAK package [20] with  $\chi^2$  statistics in place of the

distance function. In spite of the fact that  $\chi^2$  statistics is not a metric (triangle inequality does not hold), we chose to use it as our distance measure due to its good discrimination characteristics [21]. In the second methods we tested, all the views belonging to a particular location were first sampled uniformly, followed by K-means clustering stage. The number of samples varied depending on the location and number of clusters per location varied between 1 to 5.

## C. Recognition phase

In the learning and recognition phase we have randomly chosen 70%, 80% or 90% of total frames as the training data and the whole sequence is treated as testing data. For classes with small number of frames (3 or 4) one or more frames were left as the testing data. The recognition experiment was repeated 50 times for K-means and 10 times for representation obtained using LVQ. The recognition rate was recorded each time and averaged over all trials. In both representations we have used nearest neighbour classifier to determine the location which the view came from. The recognition rates for two vector quantization techniques are in Figures 4 and 5 are recorded as a function of total number of prototypes for all 30 locations. The number of prototypes per class however depends on distribution of data at each locations.

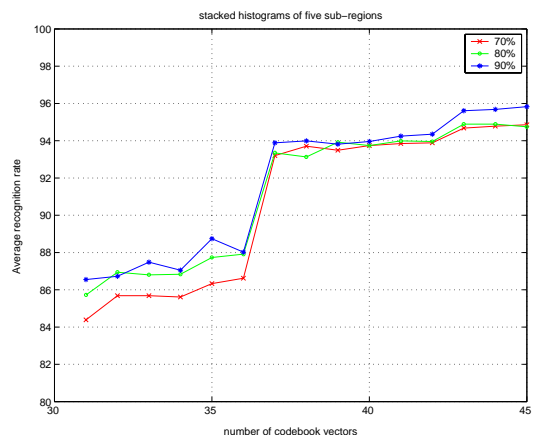


Fig. 4. Recognition rates using nearest neighbour classifier given the representation learned using LVQ method described above.

Some misclassification examples are shown in Figure 6. Note that in all instances although the misclassification occurred, the test location and assigned location are quite similar in their appearance. One possibility how to disambiguate these errors is to use more discriminative image representation or use the robots' odometry. We instead will demonstrate in the following paragraph how to use the temporal context in order to improve the location recognition accuracy, while still retaining the simple global image representation.

## D. Markov Localization

The result of the learning phase described above is a set of representative vectors for each class. The use of temporal context is motivated by the work of [17] which addresses

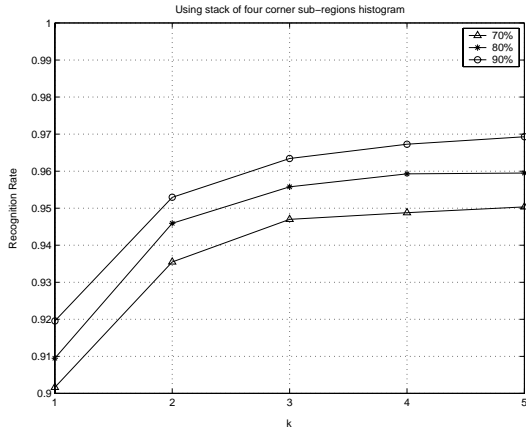


Fig. 5. Recognition rates using nearest neighbour classifier given the representation learned using kmeans method described above.



Fig. 6. Examples of test images which were misclassified in the recognition stage: the first row are the test images and the second row are the images which are closest to the nearest neighbour class center.

the place recognition problem in the context of wearable computing application. The temporal context is determined by spatial relationships between individual locations and is modelled 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. Since the states (locations) cannot be observed directly each location is characterized by its associated observation likelihood  $P(L_t = l_i | o_{1:t})$  denoting the conditional probability of being at time  $t$  and location  $l_i$  given the available observations up to time  $t$ . The problem of localization can then be formulated as a problem of estimating most likely location given all available measurements up to time  $t$ . The location likelihood can be estimated recursively using the following formula

$$P(L_t = l_i | o_{1:t}) \propto p(o_t | L_t = l_i) P(L_t = l_i | o_{1:t-1}) \quad (3)$$

where  $p(o_t | L_t = l_i)$  is the observation likelihood, characterizing how likely is the observation  $o_t$  at time  $t$  to come from location  $l_i$ . We have tested several methods for approximating the observation likelihood based on the recognition experiments carried out in the previous section. At the moment the probability that the observation comes from certain location  $p(o_t | L_t = l_i)$  is obtained by first finding the closest cluster

center among all classes based on Bayes rule. The chosen nearest cluster is then approximated with a spherical Gaussian distribution with the cluster center as the mean. The probability of the test image belonging to this cluster center then becomes the probability of the test image belonging to the location. Alternative representation of individual location models in terms of Gaussian mixtures has been proposed in [17]. We have found this soft assignment to be less effective given the choice of our image representation. The second term of equation 3 can be further decomposed

$$P(L_t = l_i | o_{1:t-1}) = \sum_j^N A(l_i, l_j) P(L_{t-1} = l_j | o_{1:t-1}) \quad (4)$$

where  $N$  is the total number of locations and  $A(l_i, l_j) = P(L_t = l_i | L_{t-1} = l_j)$  is the probability of two locations being adjacent. All the transition probabilities between individual locations were assigned non-zero values despite the fact that the transitions between certain locations did not exist. In the presence of a transition between two locations the corresponding entry was assigned value  $p_1$  and in the absence of the transition it was assigned value  $p_0$ . In the final stage all the rows of the matrix were normalized. The performance reported in the following experiments used the ratio of  $p_1/p_2 = 1.5$ . The ratio of values  $p_1$  and  $p_0$  affected the final recognition rate.

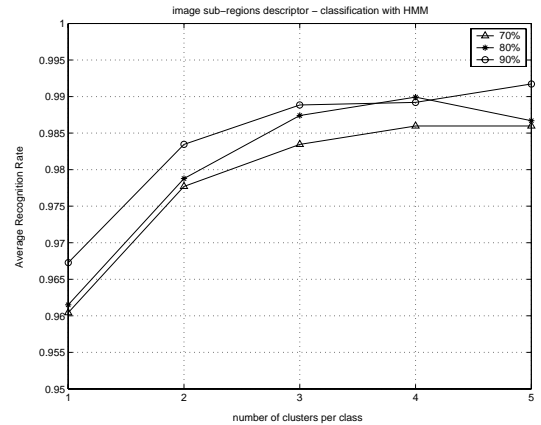


Fig. 7. Recognition rates for the entire sequence using HMM.

The recognition performance using temporal context enabled us to eliminate most of the previous classification errors and achieve classification rate around 99%. Towards this end we used a separate test sequence which visited all the locations in slightly different order. Figure 7 shows the recognition rates for the entire experiment averaged over 50 trials as a function of number of representative clusters. Figure 8 shows the correct labels of visited locations for the test sequence.

### III. CONCLUSIONS

Both the clustering experiments and the location recognition demonstrate promising performance despite the simple global appearance measure. While in the static case we have observed

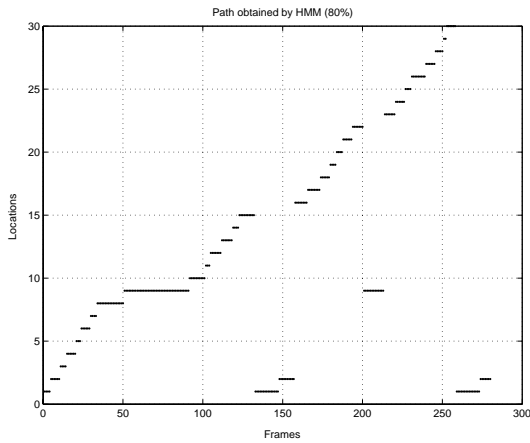


Fig. 8. The resulting test sequence of visited locations.

several classification errors, those were successfully eliminated using the temporal context modelled by HMM. We plan to further test the performance of the proposed method in less controlled environments, with various dynamic changes and fully automate the model acquisition stage. These steps are essential in for enabling simultaneous model acquisition and localization. The presented work only dealt with capturing the coarse spatial structure of the indoor environment. In parallel we are developing methods to enabling precise relative positioning within individual locations. We also plan to examine alternative choices of image representations and their applicability to ranges of indoors environment.

#### ACKNOWLEDGEMENTS

This work is supported by NSF grant IIS-0118732.

#### REFERENCES

- [1] A. Davidson and D. Murray, "Simultaneous localization and map building using active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 865–880, 2002.
- [2] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale invariant visual landmarks," *International Journal of Robotics Research*, 2002.
- [3] G. DeSouza and A. Kak, "Vision for mobile robot navigation: A survey," *IEEE Transactions of Pattern Recognition and Machine Intelligence*, vol. 24, no. 2, 2002.
- [4] C. J. Taylor and D. Kriegman, "Vision-based motion planning and exploration algorithms for mobile robots," *IEEE Transaction on Robotics and Automation*, vol. 14, no. 3, pp. 417–427, 1998.
- [5] A. Briggs, D. Scharstein, and S. Abbott, "Reliable mobile robot navigation from unreliable visual cues." in *In Fourth International Workshop on Algorithmic Foundations of Robotics, New Hampshire*, 2000.
- [6] R. Sims and G. Dudek, "Learning envirmental features for pose estimation," *Image and Vision Computing*, vol. 19, no. 11, pp. 733–739, 2001.
- [7] A. Pope and D.Lowe, "Probabilistic models of appearance for object recognition," *Internatinal Journal of Computer Vision*, vol. 40, no. 2, pp. 149–167, 2000.
- [8] J. Wolf, W. Burgard, and H. Burkhardt, "Robust vision based localization for mobile robots using image-based retrieval system based on invariant features," in *IEEE International Conference on Robotics and Automation*, 2003.
- [9] A. Torralba and P. Sinha, "Indoors scene recognition," in *AI Memo 2001-015*, 2001.

- [10] B. Schiele and J. L. Crowley, "Object recognition using multidimensional receptive field histograms." *International Journal of Computer Vision*, 2000.
- [11] M. Artac, M. Jogan, and A. Leonardis, "Mobile robot localization using an incremental eigenspace model," in *IEEE Conference of Robotics and Automation*, 2002, pp. 1025 – 1030.
- [12] J. Gaspar, N. Winters, and J. Santos-Victor, "Vision-based navigation and environmental representations with an omnidirectional camera," *IEEE Transactions on Robotics and Automation*, pp. 777–789, December 2000.
- [13] G. Bianco, A. Zelinsky, and M. Lehrer, "Visual landmark learning," in *IROS, Japan*, October 2000.
- [14] B. Kuipers and Y. T. Byun, "A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations," *Journal of Robotics and Autonomous Systems*, no. 8, pp. 47–63, 1991.
- [15] N. Tomatis, I. Nourbakhsh, and R. Siegwart, "Hybrid simultaneous localization and map building: a natural integration of topological and metric," *Robotics and Autonomous Systems*, 2002.
- [16] J. Košecká, L. Zhou, P. Barber, and Z. Duric, "Qualitative image based localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.
- [17] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *International Conference on Computer Vision*, 2003.
- [18] J. Košecká and W. Zhang, "Video compass," in *Proceedings of European Conference on Computer Vision*, 2002, pp. 657 – 673.
- [19] L. Walker and J. Malik, "When is the scene recognition just texture recognition?" *Vision Research*, no. (to appear), 2003.
- [20] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, and K. Torkkola, "LVQ-PAK - the learning vector quantization program package," Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finland, Tech. Rep. TR A30, 1996.
- [21] P. Barber, "Image-based localization for mobile robot navigation," Master's thesis, George Mason University, Department of Computer Science, 2002.