CVPR 2003, June 16-22, Madison, Wisconsin (to appear)

Qualitative Image Based Localization in Indoors Environments

Jana Košecká, Liang Zhou, Philip Barber, Zoran Duric
Department of Computer Science, George Mason University
4400 University Drive, MS 4A5, Fairfax, VA 22030
kosecka@cs.gmu.edu

Abstract

Man made indoors environments posses regularities which can be efficiently exploited in automated model acquisition by means of visual sensing. In this context we propose an approach for inferring a topological model of an environment from images or the video stream captured by a mobile robot during exploration. The proposed model consists of a set of locations and neighbourhood relationships between them. Initially each location in the model is represented by a collection of similar, temporally adjacent views, with the similarity defined according to a simple appearance based distance measure. The sparser representation is obtained in a subsequent learning stage by means of Learning Vector Quantization (LVQ). The quality of the model is tested in the context of qualitative localization scheme by means of location recognition: given a new view, the most likely location where that view came from is determined.

1 Introduction and Related Work

Visual sensing greatly enhances the capabilities of mobile robots and their interaction with humans in a variety of applications. Depending on the requirements of robotic tasks different types of models are appropriate. A variety of 2-D maps and 3-D models proposed in the past were acquired by means of range and visual sensing. However even the full availability of 3-D information and capability of recovering exact pose of the mobile robot does not alleviate the difficulty of interpreting 3-D environments.

The existing models can be broadly partitioned depending on the amount of prior information about the environment and the type of model they attempt to capture. For landmark based approaches the topology of the environment has been induced by positions of the landmarks and the visibility regions associated with them. While the model of landmarks is typically given, their location was either known a-priori or obtained during exploration [14]. 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 [3]. A nice review of the existing techniques focusing mostly on the model based methods can be found in [5]. In the absence of a prior environment model visual sensing has been used successfully for the simultaneous map building and localization (SLAM) in the context of extended Kalman filter framework. The detection and tracking of the salient point features tied with the structure and motion recovery techniques in a recursive setting enabled continuous pose maintenance of the mobile robot [4]. Vision-based techniques which do not use any prior information about the environment vary depending on whether they try to recover full geometric structure of the scene or merely the salient environmental features, which can be subsequently recognized and used for localization [7]. The focus of these techniques is predominantly on exact localization and recovery of exact relative pose of the robot with respect to the environment. Alternative to the metric models are the models which represent environment topology. Topological representations are highly desirable since they impose a discrete structure on the otherwise continuous configuration space which often simplifies a variety of navigation tasks [9]. 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. The existing techniques have been motivated by appearance based approaches for object recognition. The main concern of the appearance based methods for recognition is the selection of image attributes (e.g. feature vectors characterizing the salient characteristics of the image), which could guarantee some amount of invariance with respect to variations in pose/viewpoint, illumination, scale and be robust with respect to partial occlusion and clutter. Commonly used representations are responses to a banks of filters, multi-dimensional histograms [12], local Fourier-transforms [13] and affine invariant feature descriptors [11]. These representations in the context of mobile robot navigation were most commonly obtained via principal component analysis (PCA) or alternative clustering techniques. In case of omni-directional views PCA based techniques were applied successfully for topological model acquisition, thanks to small variations of the image appearance within a location [1, 6].

2 Approach

In our approach the nodes of the topological model represent regions in space, so called locations, represented by a set of views. The qualitative localization scheme proposed here, will be enabled by recognition of locations, which loosely correspond to the regions in the robot's configuration space which are similar in their appearance. The neighboring 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 of the edge map, which captures the essential appearance information in each view. The histograms are sufficiently discriminant between individual locations and also capture the similarity between the views which are perceptually close. The frames of the temporally sub-sampled video sequence are partitioned and automatically labelled during the exploration phase by comparing temporal distance between consecutive views. After obtaining a labelled set of views we used a Learning Vector Quantization (LVQ) technique to obtain sparser representation for each location by selecting the representative feature vectors which best cover the class. In the classification stage we determine given a previously unseen view, what is the location it most likely comes from.

2.1 Measurement Stage

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 directions is aligned with the axes of the world coordinate frame [10], we use the information provided by image gradient orientation as distinguishing characteristic of an individual location. Gathering the gradient orientation in a histogram we obtain appearance based measure that is stable with respect to the changes in the lighting conditions and disambiguates different locations well. The most notable characteristic of this simple feature is that 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.

In order to obtain more robust measure, the gradient orientation histograms are computed only for the pixels with the magnitude in the top 4%. This empirical choice of threshold worked well for our image database yielding the dominant features of the environment. This is achieved by first computing the image derivatives $[I_x, I_y]^T$ 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. Examples of representative views, associated edge maps and gradient orientation histograms are in Figure 3.

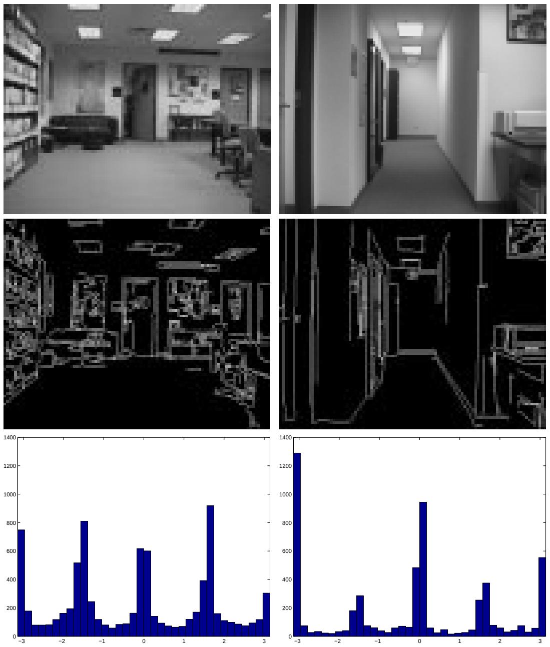

Figure 1. Locations H (left) and F (right) of the 4^{th} floor, their associated edge maps and gradient orientation histograms.

In the context of indoors environments, the orientation histograms of edge maps give us sufficient level of invariance with respect to variation in pose and illuminations conditions. The resulting pixels contributing to the histogram correctly reflect the visual cues that humans use to navigate the environment. Commonly encountered presence of corners, doors, and bulletin boards, is properly captured by orientation histogram and discriminates locations with different appearance. Furthermore due to the 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. Changes in the overall lighting intensity have virtually no effect on edge direction. Only at the far end of the saturation scale is its effect on edge direction detection significant. The level of intensity saturation has an effect on edge magnitude, but we mitigate that effect by considering only the pixels with the greatest edge strength in each individual image. This has a normalization effect since the pixels with the greatest edge strength tend to remain the same even if their absolute magnitudes change. Given the fact that the histogram is computed globally over the whole image, occlusions caused by walking people, misplaced objects have minor effect on the total histogram.

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

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

where k is the number of histogram bins. We have carried out several experiments in order to test the discriminatory capability of the chosen feature vector and associated χ^2 distance measure. One set of experiments consisted of collecting 140 images along 3 corridors of the 4^{th} floor of our building (see Figure 2). The images were taken by still digital camera about 2 meters apart, with the orientation in the direction of mobile robot heading. In this data set the heading direction was in most cases aligned, or perpendicular with the principal directions of the world coordinate frame; assumption which we plan to relax in the future. Second data set was acquired with a commercial digital camcorder mounted on the mobile robot during the environment exploration. The temporal and pairwise distances between the image histograms associated with individual views are depicted in Figures 3 and 4. Note that the affinity matrix of the first image collection in Figure 3 has clear distinguishing clusters corresponding to the images collected along particular trajectory segments at locations A, B, C, D, E, F, G, H. The clear boundaries between the clusters are due to the discontinuity in the heading of the mobile robot and sudden change in the location appearance. Note 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 some of the locations (e.g. corridors) appear very similar. Note that the temporal and pairwise χ^2 distances in case of video sequences have much smoother transitions between the clusters indicating some quasi-continuity of the space of visual observations. In case of video sequence the histograms were smoothed both spatially and temporally in order to eliminate some of the aliasing effects of the digital camcorder. As an additional processing step we have subtracted the local histogram mean, providing some normalization to the

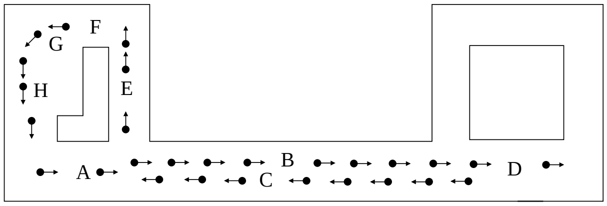

Figure 2. Floor plan of the 4^{th} floor; exploration route and label associated with individual locations labelled by hand.

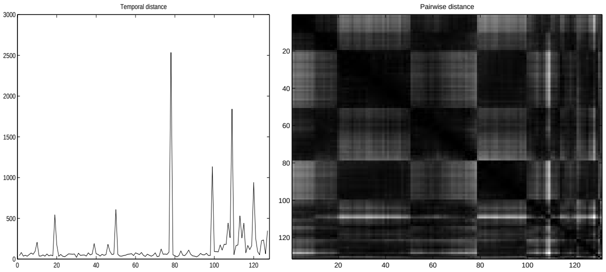

Figure 3. The temporal and pairwise comparison of orientation histograms of images taken by still digital camera.

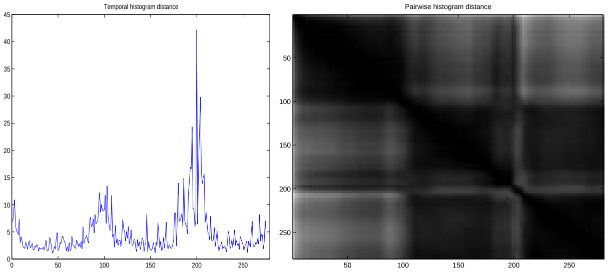

Figure 4. The temporal and pairwise comparison of orientation histograms of the video sequence.

histogram data. Figure 4 depicts the affinity matrix for 2800 frames long video sequence taken while the robot was travelling from locations $H \rightarrow G \rightarrow F \rightarrow E$. The sequence was temporally subsampled by factor of 10 and spatially subsampled by factor of 1 (240×320) . Both the temporal distances and the affinity matrices reveal the fact that image orientation histogram indeed captures well the appearance similarity of different locations. In order to obtain sparser representation for each location in terms of fewer number of views and consequently solve the localization problem we carry out an additional clustering stage.

2.2 Learning phase

Figures 3 and 4 demonstrate clear presence of individual clusters corresponding to different spatial locations. The assignment of individual views to clusters in our case can be naturally 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 label assignment was obtained by searching for the peaks in temporal histogram distance plot. 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 3 and 4 the dominant peaks are quite distinguishable, clearly separating images associated with the individual locations. 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 chosen Learning Vector Quantization technique (LVQ) which is well established in pattern recognition field. It is attractive due to its simplicity, effectiveness, and the fact that there is an existing implementation that meets our needs. The 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 a set of prototype vectors to represent the data. LVQ defines decisions surfaces between competing classes which are piece-wise linear hyperplanes. 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:

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

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 most notable difference between the existing applications of LVQ and our problem was the use of the distance measure. In the previously encountered cases the applications used the Euclidean distance between the feature vectors for the distance computation. In spite of the fact that χ^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 [2].

We used the existing implementation of LVQ_PAK package produced by the Laboratory of Computer and Information Science at the Helsinki University of Technology [8]. The number of codebook vectors for each class is determined in proportion to the prior probabilities of individual classes. After the initial prototype vectors are created, they are refined using one of the heuristic approaches proposed by Kohonen [8]. The refinement algorithm attempts to minimize distances between the original data vectors and the prototype vectors by removing vectors from over-represented groups and adding vectors to groups that are under-represented. The number of resulting prototypes varied depending on the size of the exemplar set and variability in the appearance within a location. The number of codebook vectors assigned to each class significantly affects the recognition performance. As the number of prototype vectors becomes small the ability of the system to produce correct results diminishes.

2.3 Recognition phase

After generating prototype histograms corresponding to individual locations the new images were classified using nearest neighbour test. Given a new image, the histogram h was generated and compared using χ^2 distance to each of the prototype vectors generated by LVQ . Computing distances to the nearest neighbour $h_{1^{st}}$ and the second closest vector $h_{2^{nd}}$ belonging to a different class, we obtain a confidence level defined as

$$C_\chi = \frac{\chi^2(h, h_{2^{nd}})}{\chi^2(h, h_{1^{st}})}.$$

The confidence level quantifies the reliability of the classification. If the confidence level is close to 1 (say 1-1.5), then the two distances are close enough indicating that the classification is not reliable. In our experience, confidence level greater than 1.6 was considered to be accurate. In case the classification was achieved with a low confidence the classification was further refined by comparing sub-images of the new image and the images in the database closest to the vectors in the database.

	global		sub-image	
magnitude%	6%	4%	6%	4%
images	91.12	92.35	92.37	95.40

Table 1. Recognition rates for the first data set of 185 images; global and sub-image histograms recognition strategies are compared.

3 Experimental results

The experimental results were carried out using several data sets acquired by both still digital camera and video camcorder. The initial data set of 185 images was taken at 7 different locations $A - F$ in Figure 2, the number of images associated with individual locations varied between 7 - 20 per location. After histogram computation images were assigned initial labels corresponding to different locations. Approximately 5% of histograms was randomly selected to act as testing set. The remaining images were used in the learning stage. This process was repeated 100 times, each time recording the percentage of correctly classified images. The number of resulting codebook vectors varied between 5 - 11 depending of the variation encountered within the class.

Sub-image comparison Once the confidence level was below the defined threshold the classification was refined by comparing sub-images of the new view with the sub-images of the nearest neighbours. We have considered five sub-images (one in the center and four quarters of the original image) to do the comparison. An example of a wrong classification reflected by low confidence level and re-classification using sub-image comparison is in Figure 5. Histograms of the test sub-images were generated and compared with the histograms of the nearest neighbours using χ^2 distance. The final classification was then based on the median of these distances. Using this additional test led to a slight improvement in the overall recognition rate as reported in Table 1. The accuracy of the recognition is reported as a function percentage of pixels which contributed to the unnormalized gradient orientation histogram.

Video Clustering Due to the larger size of the video data set, the training and testing was performed slightly differently in this case; 50 % of the examples were selected in the training phase and the remaining 50 % were used in testing. We varied the initial number of codebook vectors and recorded the overall accuracy of recognition as this number varied. The total recognition rates for the entire sequence are in Table 2. Examples of representative views from the entire data set covering larger number of locations are reported in Figure 6. The figure depicts the associated images of the first three prototype vectors from each class. Note that the resulting prototypes provide good coverage for the class.

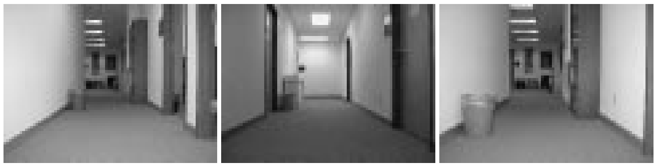

Figure 5. Example of an image from location F (left), misclassified as one from location E (middle) and then re-classified correctly as F (right) using sub-image comparison.

% of initial prototypes	10 %	20 %	30 %
video	84.17 %	92.81 %	99.28 %

Table 2. Recognition rates for the video sequences while varying the percentage of the initial number of prototype vectors.

4 Conclusions

In the current stage the experiments have been carried out using purely image appearance data, where the appearance was characterized by a simple gradient orientation histogram. Both the clustering experiments and the location recognition demonstrate promising performance. The most commonly encountered misclassification can be in many instances resolved by further sub-image comparison. We are currently investigating alternative image based representations and evaluating the performance of the method in the environment, which is both larger in scale, represented in terms of larger number of locations and exhibits bigger variations in image appearance. In addition to sub-image comparison methods we also integrating the temporal relationships between the views in the recognition phase. The resulting representation is appealing and correctly captures the notion of the location. We are currently using only visual data and assume in the model acquisition stage that each location has been visited exactly once. Detecting cycles requires an adoption of more elaborate inference procedure, which we are currently pursuing. If desired, given a set of views associated with each location, additional level of detail or geometric information can be computed from the representative views.

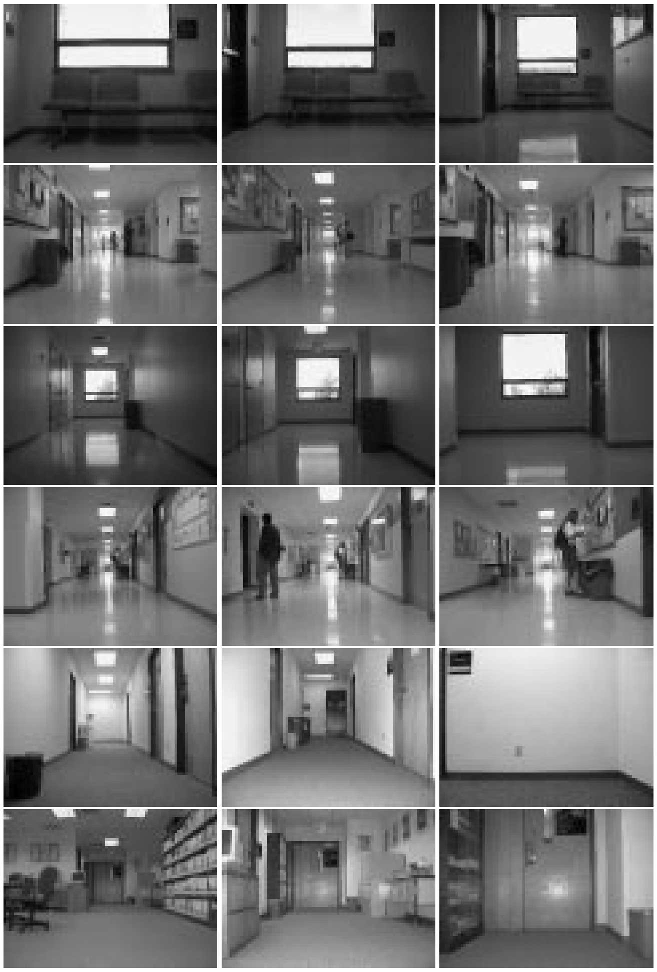

Figure 6. Examples of prototype vectors associated with individual locations A, B, C, D, E, F, G. Only first three prototype views are depicted.

Acknowledgments

This work is supported by NSF grant IIS-0118732.

References

[1] M. Artac, M. Jogan, and A. Leonardis. Mobile robot localization using an incremental eigenspace model. In *IEEE Conference of Robotics and Automation*, pages 1025 – 1030, 2002.

[2] P. Barber. Image-based localization for mobile robot navigation. Master's thesis, George Mason University, Department of Computer Science, 2002.

[3] 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.

[4] A. Davidson and D. Murrray. Simultaneous localization and map building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):865–880, 2002.

[5] G. DeSouza and A. Kak. Vision for mobile robot navigation: A survey. *IEEE Transactions of Pattern Recognition and Machince Intelligence*, 24(2), 2002.

[6] J. Gaspar, N. Winters, and J. Santos-Victor. Vision-based navigation and environmental representations with an omnidirectional camera. *IEEE Transactions on Robotics and Automation*, pages 777–789, December 2000.

[7] G. Hager and D. Kriegman. Image-based prediction of landmark features for mobile robot navigation. In *IEEE Conference on Robotics and Automation*, pages 1040–1046, 1997.

[8] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, and K. Torkkola. LVQ_PAK - the learning vector quantization program package. Technical Report TR A30, Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finland, 1996.

[9] J. Košecká. Visually guided navigation. *Robotics and Autonomous Systems*, 21(1):37–51, July 1997.

[10] J. Košecká and W. Zhang. Video compass. In *Proceedings of European Conference on Computer Vision*, pages 657 – 673, 2002.

[11] A. Pope and D.Lowe. Probabilistic models of appearance for object recognition. *Internatinal Journal of Computer Vision*, 40(2):149–167, 2000.

[12] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. *International Journal of Computer Vision*, 2000.

[13] R. Sims and G. Dudek. Learning envirmental features for pose estimation. *Image and Vision Computing*, 19(11):733–739, 2001.

[14] C. J. Taylor and D. Kriegman. Vision-based motion planning and exploration algorithms for mobile robots. *IEEE Transaction on Robotics and Automation*, 14(3):417–427, 1998.