Selecting Promising Landmarks

Markus Knapek
Computer Science
Technical University of Munich

mknapek@brain.nefo.med.uni-muenchen.de

Ricardo Swain Oropeza David J. Kriegman
Computer Science
Beckman Institute
University of Illinois, Urbana-Champaign
{swain1, kriegman}@uiuc.edu

Abstract

Many approaches to visual servoing and mobile robot navigation are based on tracking feature points or landmarks in images. But are all features points equally effective as landmarks? Here we develop methods for selecting within an image those landmarks which are both perceptually salient and visually distinctive, and consequently are readily recognized in a second image acquired from a different viewpoint. Empirically, we characterize the performance of the recognition method and then demonstrate that the selection process does in fact choose the landmarks which are more likely to be recognized.

1 Introduction

Nearly all approaches to visual servoing are based on tracking feature points in an image sequence that correspond to the projection of viewpoint independent features of the 3-D scene or object [5, 6, 14]. Similarly, numerous approaches to vision-based mobile robot navigation recognize and possibly track landmarks [1, 7, 9, 12, 18]. Except within specific applications, most visual servoing implementations have either used a catalogue of model/application specific landmarks or relied on a person to initialize tracking. In most mobile robotics implementations, the robot's landmark recognition system is provided with a catalogue of domain-specific recognizable landmarks (e.g. lane boundaries, ceiling light, bar codes, door edges, etc.). Recently, it has become possible to track at frame rates a modest number of features (on the order of a dozen) on conventional personal computers using for example X-vision [2]. Yet in a scene or for an object of interest, there may be hundreds or thousands of feature points which could serve as landmarks. In fact, part of the process of most feature-based structure-from-motion algorithms is to establish correspondence amongst a large number of features. For real-time robot control using current methods, only a fraction of the possible features can be considered and tracked. Hence, one must select from the plethora of candidate image locations, those features which are very "trackable" (salient) and readily recognized (distinctive).

In this paper, we address the problem of selecting from a single monochrome image those landmarks which are both salient and distinctive. By salient, we mean that the landmark should readily "pop out" from the background by some detection mechanism. By distinctive, we mean that the landmark is unlikely to be confused during recognition. For example, a red fire extinguisher is likely to salient in almost any environment since it stands out from its background. However, it would be a dismal landmark in a fire extinguisher factory because even thought each extinguisher might stand out from the background, it would be difficult to distinguish one extinguisher from another.

Our methods are task and domain independent. This is both a strength and a weakness. By being domain-independent, the same method should apply to robots wandering indoors or outdoors. Yet there is more to "what makes a good landmark" than perception, and the set of chosen landmarks must be useful for the task. On the other hand, incorporating domain and task specific constraints can facilitate the landmark selection. On the whole, while domain and task specific visual control methods have yielded impressive performance in the laboratory, these methods are often brittle in practice, and so we choose to focus on domain-independent selection methods here.

There is too wide a literature on feature tracking, visual servoing, and landmark-based navigation to summarize here. There have also been a few papers on the process of selecting useful features points or landmarks in image data [12, 15, 18, 19, 20]. However, this approach is very strongly motivated by the image database indexing and recognition work of Schmid and Mohr [10, 11] which in turn builds on [16, 17].

2 Recognizing and Selecting Landmarks

Many methods for tracking have been developed including corner tracking, line tracking, region tracking, blob tracking, color tracking, 3-D model tracking, etc. In general, trackers continually estimate some parameter vector representing some attributes of the tracked object (e.g. image location, scale, lighting, 3-D pose, shape, etc.) which are presumed to be varying continuously. One class of trackers that is particularly useful for robot navigation provides over time the image location of the projection of a local (small) region or point of a 3-D scene. Tracking a modest number of such features can be used to localize the robot, navigate using visual servoing, or recognize a place. Typically, the local region is represented by a template. When tracking, a region of an image is searched for the location which minimizes the sum of squared differences (SSD) between the template and the image intensities about the location. Due to 3-D viewpoint changes however, the image pattern will differ from the template, and this is sometimes modeled as an affine image warp [2].

Here, our goal is to select distinctive templates (landmarks) from one image which can be readily recognized in a second image acquired from a different viewpoint. Let the irradiance (intensity) across the image plane be denoted by $I(x, y)$ where x and y are the image coordinates.

Since we are interested in tracking the projection of point-like features, we can characterize $I(x, y)$ locally about a point (x_0, y_0) by its differential structure. In particular, consider the vector of partial derivatives up k-th order which is known as the k-jet; for example, the 2-jet of $I(x, y)$ is given by:

$$\mathcal{F}(x, y) = \begin{bmatrix} I \\ I_x \\ I_y \\ I_{xx} \\ I_{xy} \\ I_{yy} \end{bmatrix}$$

The k-jet or some function of the k-jet can serve as a representation of a landmark. We now summarize the basic landmark selection and recognition method which directly follows elements of the recognition method of Schmid and Mohr [10]. To select the landmarks:

1. A detector is applied to the entire image to select potential landmarks which should be readily tracked (salient).
2. The potential landmarks are characterized by a feature vector derived from the k-jet.
3. The potential landmarks are ordered by distinctiveness, and the most distinctive ones are retained.

Similarly, the landmarks are recognized in a second image:

1. The same detector is applied to the image, but with lower thresholds, to identify candidate locations of landmarks.
2. Each candidate is again characterized by a feature vector computed from the K-jet.
3. Each selected landmark is recognized by nearest neighbor classification using a Mahalanobis distance.

2.1 Details

The motivation for separating the selection process into two steps with the two criteria of saliency and distinctiveness is computational cost. To sort l candidate landmarks based on their distinctiveness requires computing the similarity of all pairs of landmarks which is $O(l^2)$. Without a process for preselecting salient features, l could be the number of pixels n in an image, approximately 300,000. Instead, a set of sufficiently salient landmarks, which are expected to be readily tracked, are detected in the image using an $O(n)$ process. Typically a few hundred potential landmarks are detected, and the most distinctive ones are selected amongst the $l \approx 200$ candidates.

Following [10] candidate landmarks are detected using the Harris corner detector [3] which can be viewed as a successor to Moravec's interest operator [8]. The basic idea of the Harris detector is to compute a cornerness measure $c(x, y)$ from $I(x, y)$ which essentially determines the principal curvatures of the autocorrelation function; feature locations are taken as those locations $\mathbf{p} = (x, y)$ which are local maxima of $c(x, y)$ and exceed a threshold τ_h.

The neighborhood of each feature location $\mathbf{p}$ can then be characterized by its k-jet. However, the k-jet clearly depends on the location and orientation of the camera. To model the possible changes to the image pattern and to the k-jet, let us assume that $I(x, y)$ in the neighborhood of $\mathbf{p}$ is the projection of a planar Lambertian surface with non-constant albedo. It is well known that the change of coordinates between images of a plane acquired at different viewpoints is a projective transformation. However when the neighborhood is small, the change of coordinates

in the neighborhood can be approximated by an affine transformation $\mathbf{p}' = A\mathbf{p} + \mathbf{t}$. The effect of $A \in GL(2)$ can be characterized as a combination of rotation, independent scaling along the axis, and shearing, while the effect of $\mathbf{t}$ is clearly to translate the pattern. Assuming that the detector's response is insensitive to A (this has been confirmed for the Harris detector for moderate values of A in [11]), then we are really only interested in characterizing the changes to the k-jet under linear transforms A.

Two possibilities are to either model the variation in the k-jet as a function of A or compute a function of the k-jet which is invariant to A; we choose the later. If the camera motion is constrained or if elements of the camera motion can be directly measured, then we may only be interested in invariance to certain subgroups of $GL(2)$, e.g. image plane rotations, $SO(2)$. As discussed in [16, 17], one can compute functions of the K-jet which are invariant to different subgroups of A, so called differential invariants of $I(x, y)$. For example, the following complete set of 3rd-order differential invariants under $SO(2)$ was used by Schmid and Mohr [10] and will be used below:

$$\mathcal{F}^r(x,y) = \begin{bmatrix} I \\ I_x I_x + I_y I_y \\ I_{xx} I_x I_x + 2 I_{xy} I_x I_y + I_{yy} I_y I_y \\ I_{xx} + I_{yy} \\ I_{xx} I_{xx} + 2 I_{xy} I_{xy} + I_{yy} I_{yy} \\ I_{xxx} I_y I_y I_y - 3 I_{xxy} I_x I_y I_y + 3 I_{xyy} I_x I_x I_y \\ \quad - I_{yyy} I_x I_x I_x \\ I_{xxx} I_x I_y I_y - 2 I_{xxy} I_x I_x I_y + I_{xxy} I_y I_y I_y \\ \quad - 2 I_{xyy} I_x I_y I_y + I_{xyy} I_x I_x I_x + I_{yyy} I_x I_x I_y \\ I_{xxx} I_x I_x I_y + 2 I_{xxy} I_x I_y I_y - I_{xxy} I_x I_x I_x \\ \quad - 2 I_{xyy} I_x I_x I_y + I_{xyy} I_y I_y I_y - I_{yyy} I_x I_y I_y \\ I_{xxx} I_x I_x I_x + 3 I_{xxy} I_x I_x I_y + 3 I_{xyy} I_x I_y I_y \\ \quad + I_{yyy} I_y I_y I_y \end{bmatrix}$$

Of course, one can use similar differential invariants for other possibly relevant subgroups of $GL(2)$, e.g. rotation and scale changes, slant and scale, etc. In general, a K-jet at a point is represented as a vector with $1/2(k+1)(k+2)$ independent elements. For a subgroup with d degrees of freedom, the resulting differential invariant is composed of $1/2(k+1)(k+2) - d$ elements. If the camera/object motion is restricted, than it may be beneficial (lower error rates) to choose a subgroup with fewer degrees of freedom. This tradeoff will be seen in Section 3

Since digital images are discrete, one needs a means to compute the partial derivatives determining the k-jet. As is common practice, the discrete values are interpolated with a Gaussian, and derivatives are taken with respect to the interpolated signal. This is accomplished by filtering the image with a kernel given by partial derivatives of a Gaussian with some variance σ. The choice of the Gaussian kernel and the resulting scale space for different choices of σ is discussed in [16].

To compare two feature points $\mathbf{p}^1$ and $\mathbf{p}^2$ detected in two images, which are described by a feature vectors $\mathbf{f}^1$ and $\mathbf{f}^2$ (feature vector $\mathbf{f}^i$ could be computed from $\mathcal{F}(\mathbf{p}^i)$ or $\mathcal{F}^r(\mathbf{p}^i)$), we determine their Mahalanobis distance

$$d(\mathbf{p}^1, \mathbf{p}^2) = (\mathbf{f}^2 - \mathbf{f}^1)^t \Sigma^{-1} (\mathbf{f}^2 - \mathbf{f}^1). \qquad (1)$$

The covariance matrix Σ is taken as the pooled covariance of the feature vector computed from c corresponding points, ideally acquired over many image pairs having the range of conditions and features expected in the application domain. In our experiments, it was computed from 21 pairs of images with about 80 features per image. The covariance matrix is computed as:

$$\Sigma = \sum_{i=1}^{n} (\mathbf{f}_i^2 - \mathbf{f}_i^1)(\mathbf{f}_i^2 - \mathbf{f}_i^1)^t.$$

Recognition of landmark $\mathbf{p}^1$ is then simply determined by finding the feature in the second image whose Mahalanobis distance to $\mathbf{f}^1$ is smallest.

This also suggests a method for determining the most distinctive landmark. Assuming that the expected variation of a landmark's description as a feature vector is well characterized by the covariance matrix Σ, then two landmarks whose Mahalanobis distance is small are more likely to be confused (misclassified) in other images than two landmarks which are far apart. For a set of landmarks $\mathcal{P} = \{\mathbf{p}_j\}$, this suggests a distinctiveness measure for a landmark $\mathbf{p}_i \in \mathcal{P}$.

$$\delta(\mathbf{p}_i) = \min_{\mathbf{p}_j \in \mathcal{P}, \mathbf{p}_i \neq \mathbf{p}_j} d(\mathbf{p}_i, \mathbf{p}_j) \qquad (2)$$

The set of candidate landmarks can be sorted by $\delta(\mathbf{p}_i)$ where the most distinctive landmark has the largest value of $\delta(\mathbf{p}_i)$.

3 Experimental results

A series of experiments has been performed to characterize the performance of the recognition method and to assess the utility of selecting the most perceptually distinctive landmarks.

3.1 Experimental protocol

Three sets of images were gathered with a monochrome camera mounted on a Nomadics Superscout mobile robot. The algorithms were implemented in Matlab. In the first sequence, the robot moved

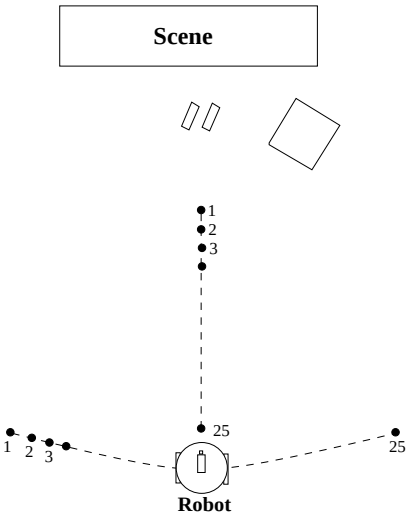

Figure 1: Experimental Setup: As the robot moved in 10cm steps along the two trajectories (a straight line away from the scene, and along a circular arc with the camera pointing toward the scene), 25 images were acquired.

along a 2.5m linear trajectory away from the scene, and images were acquired every 10cm; figure 2 shows four images from the sequence. The second sequence of twenty five images of the same scene was acquired as the robot moved along a circular arc as shown in Figure 1. The third sequence was acquired by rotating the camera about the optical axis in 15 degree steps.

3.2 Landmark Recognition

The goal of the first set of experiments is to characterize the performance of the landmark recognition process over the three sequences of images. From the linear sequence of 25 images, we considered pairs that were 20cm, 30cm,...,100cm apart; for each separation, selection and recognition was performed on 15 pairs, and the results were averaged.

Though most landmarks are correctly recognized, some are incorrectly matched. Because of the large number of landmark recognition tests being performed (e.g. each data point in Figure 4 represents the match of 750 landmarks), manual evaluation of the results is nearly impossible. And since ground truth is unavailable, we use the following automatic evaluation method. Given two images, the epipolar geometry is determined using a variant of Zhang's algorithm [21] which is based on the Hartley's 8-point algorithm [4] for estimating the fundamental matrix and RANSAC to be robust to outliers. We consider a selected landmark and the corresponding recognized landmark to

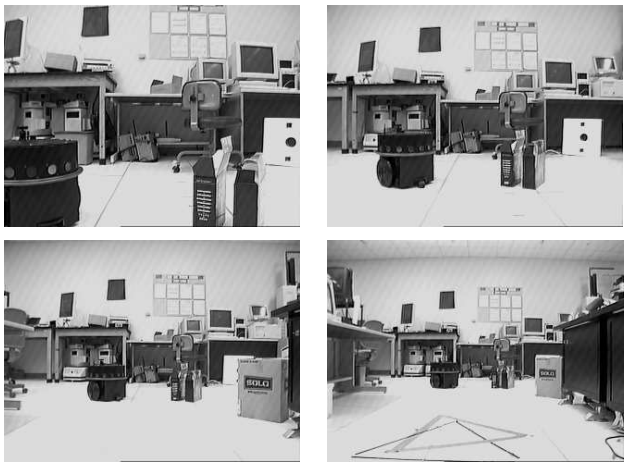

Figure 2: Four images of the sequence in which the robot moved along a linear trajectory in 10cm steps.

be false matches if they do not lie within two pixels of the corresponding epipolar lines. Clearly, this fails to detect false matches when they happen to lie on the same epipolar line, but considering that our images have 480 rows, only about 1% of the false matches will be mislabelled. In addition, some of the selected landmarks will not be recognized in the second image; in some cases, they are not detected or the Mahalanobis distance exceeds a threshold, but more often the selected landmark is not present in the second image due to occlusion or the change in the field of view. Unfortunately, there is no automatic way to ascertain the cause of the lack of a match, and so we report the error rate as the total number of incorrect matches divided by the number matched landmarks.

Two different feature sets were used for landmark selection and recognition, the third order k-jet and the rotation invariant. Figure 6 shows a pair of images with the selected and recognized landmarks while Figure 4 shows a plot of the error rate for landmark recognition over the linear image sequence. Not surprisingly, the error rate increases with the separation between the camera positions. Figure 5 shows the error rates for image pairs over the arc of camera positions. Figure 6 shows a pair of images with the selected and recognized landmarks, while Figure 7 shows the error rates for images where the camera rotated about the optical axis.

Interestingly, for the linear and arc sequences the error rate is higher when the invariants were used than when just the k-jets were used. For a mobile robot moving along the ground, there is little roll in the

camera motion, and so it seems that the loss of information by reducing the feature space from 10 independent elements to 9 elements has an impact on recognition performance. On the other hand, when there is significant rotation about the optical axis, the performance of the k-jet's was dismal whereas it was much better with the invariants. The clear lesson is to choose functions that are invariant to the expected transformation group, but not to a larger group.

3.3 Landmark Selection

To test the selection of perceptually distinctive landmarks, we performed the following test. The landmarks were sorted by distinctiveness according to (2). For each image pair, we took three subsets of the sorted landmarks, namely landmarks 1-10 (the most distinctive ones), landmarks 41-60, and landmarks 81-100. Over 26 pairs of images in the linear sequence with varying separation, the algorithm in Sec. 2 was used to recognize these landmarks. In some image pairs, a selected landmark from one image may not actually be present in the second image due to occlusion or because it falls outside of the field of view. For this experiment, we painfully checked all matches manually. Consequently, these results are limited to 1,300 landmark pairs. The error rate is reported in Figure 8. (Note that the k-jets were evaluated with $\sigma = 2$ pixels, which is smaller than that used for Fig. 4). We clearly see that the selected landmarks, that are declared to be more distinctive, are in fact more readily recognized.

4 Summary and Conclusions

We have presented a method for selecting and recognizing salient landmarks based on the method of Schmid and Mohr [10] and a criterion for selecting the most distinctive landmarks. The selection of the most recognizable landmarks can be important in real-time applications where it is not feasible to track all possible landmarks. Experimental results have characterized the performance of the method on indoor scenes where a mobile robot might typically operate.

It should be noted that there are numerous ways to improve the recognition performance. First, K-jets and invariants can be computed from color images and would offer greater discriminatory power, so long as the color of the lighting does not vary. If illumination color were to vary, than it would be interesting to merge the color invariants of Healey and Slater [13] with the local differential invariants. In our implementation, we only considered rotation invariants and for mobile robot navigation we in fact rarely have much rotation about the optical axis, and so we may want

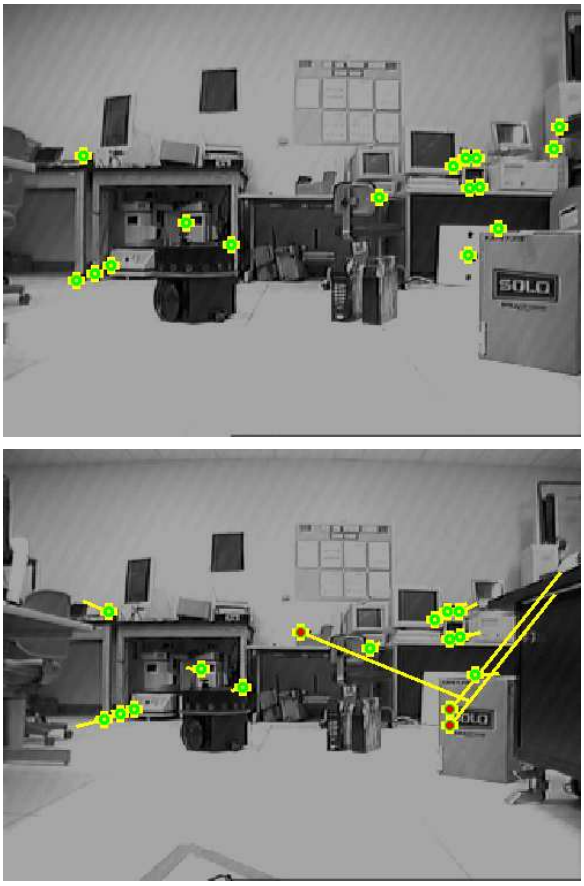

Figure 3: Two images in the linear sequence: Landmarks were selected in the upper image and recognized in the lower image. The lines denote correspondences. Note that in this pair, three mismatches occured.

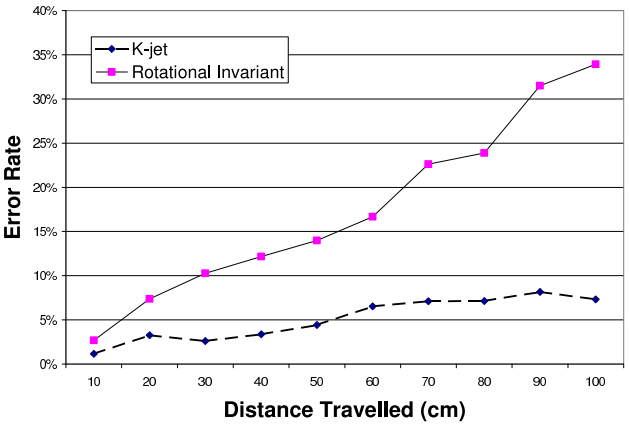

Figure 4: Error rate for recognizing the fifty most distinctive landmarks in the linear sequence using k-jets and local rotational invariants.

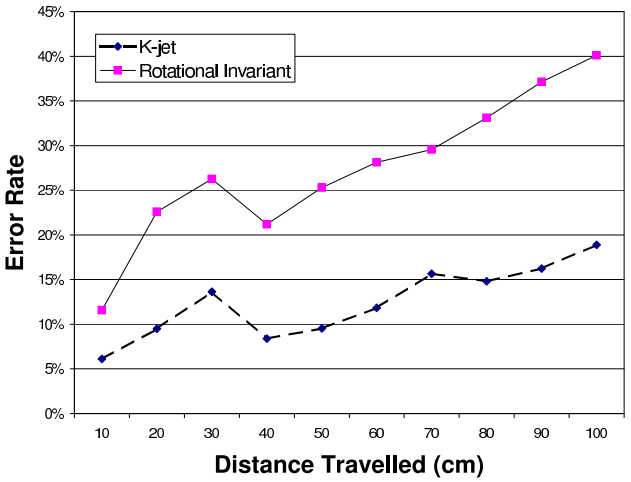

Figure 5: Error rate for recognizing the fifty most distinctive landmarks in the arc sequence using k-jets and local rotational invariants.

to explore differential invariants over other subgroups of $GL(2)$, particularly local invariants to scale. Alternatively, since derivatives are computed using a Gaussian filter, it is natural to explore recognition over scale spaces.

Features have been matched independently of each other. Clearly, between a set of correspondences in two images, the epipolar constraint must be satisfied, and this constraint can be used to both decrease the error rate and increase the speed. A related approach by Schmid and Mohr was to use a quasi-invariant of a collection of features when indexing, and this was shown to decrease the number of false matches [10]. For our projected use of natural landmarks in vision-based mobile robot navigation [1], the expected image location of the landmarks could be predicted which would reduce the search space and decrease the likelihood of false matches such that quasi-invariant indexing should be unnecessary.

Finally, perceptual distinctiveness should not be the sole criteria for selection. The task should also be taken into account; for example, it is probably desirable that the landmarks be well distributed in the image and not all correspond to coplanar scene features. Furthermore, during our experimentation, we found that the errors which occurred for distinctive features were very often due to selecting viewpoint dependent features such as T-junctions or patterns that cross an occlusion boundary. When the robot moves, the background in the neighborhood of the feature moves, and feature may no longer be detectable by Harris or its description by a K-jet/invariant may have changed. An

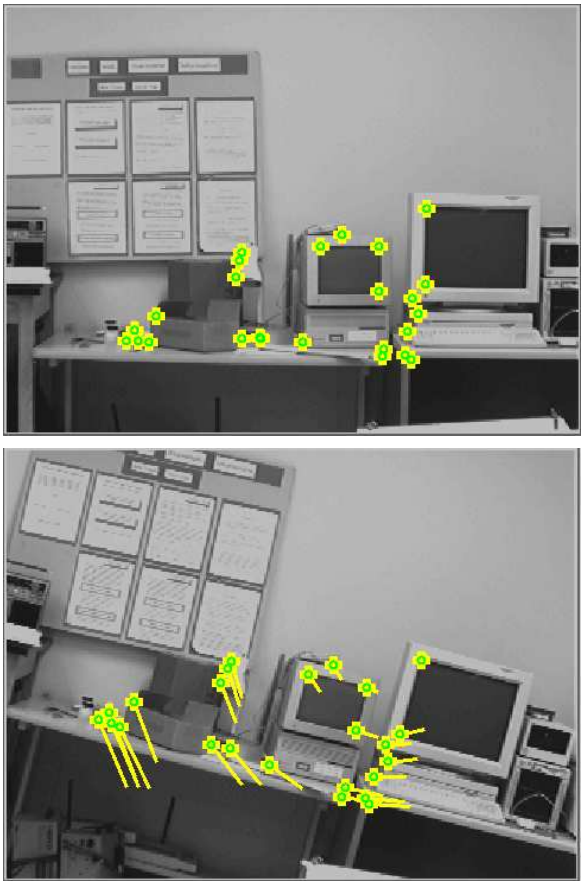

Figure 6: Two images in the rotation sequence: Landmarks were selected in the left image and recognized in the right image. The lines denote correspondences.

interesting avenue of investigation for handling both task constraints and pruning of viewpoint dependent features is to consider the way the image structure in the neighborhood of a feature changes as a robot moves.

Acknowledgments

This research was supported by the National Science Foundation under IRI-9711967 and by DARPA under DAAE07-98-C-L031.

References

[1] G. Hager, D. Kriegman, A. Georghiades, and O. Ben-Shahar. Toward domain-independent navigation: Dynamic vision and control. In *IEEE Conf. on Decision & Control*, 1998.

[2] G. Hager and K. Toyama. X vision: A portable substrate for real-time vision applications. *Computer Vision & Image Understanding*, 69(1):23–37, 1998.

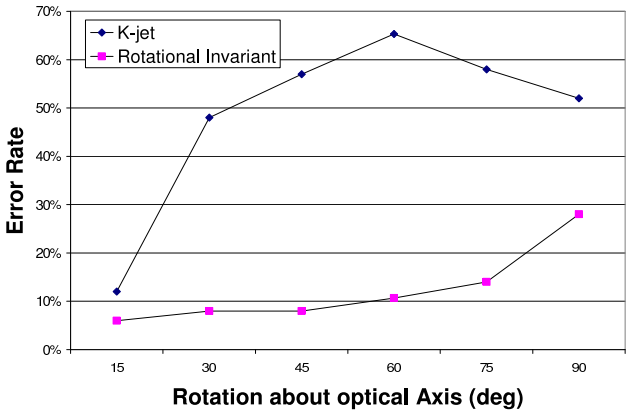

Figure 7: Error rate for recognizing the fifty most distinctive landmarks in the sequence of images with rotation about the optical axis using k-jets and local rotational invariants.

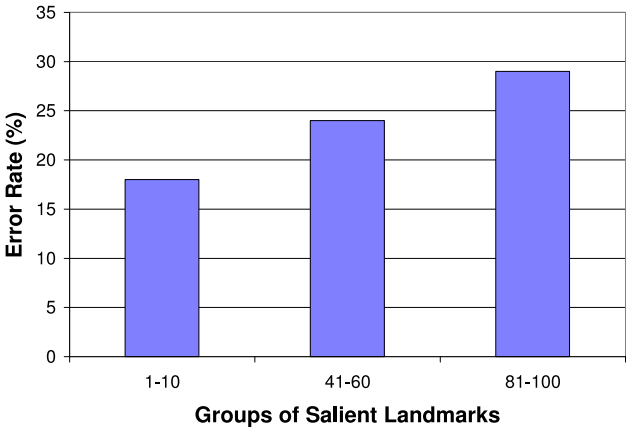

Figure 8: From 26 images, the error rate for three groups of landmarks ranked by distinctiveness.

[3] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[4] R. Hartley. Lines and points in three views and the trifocal tensor. *Int. J. Computer Vision*, 22(2):125–140, 1997.

[5] S. Hutchinson, G. Hager, and P. Corke. A tutorial on visual servo control. *IEEE Trans. on Robotics & Automation*, 12(5):651–670, 1996.

[6] D. Kriegman, G. Hager, and A. Morse. *The Confluence of Vision & Control*. Springer-Verlag, 1998.

[7] S. Li and S. Tsuji. Finding landmarks autonomously along a route. In *Int. Conf. on Pattern Recognition*, pages 316–319, 1992.

[8] H. P. Moravec. The Stanford cart and the CMU rover. *Proc. of the IEEE*, 71(7), July 1983.

[9] A. J. Munoz and J. Gonzalez. Two-dimensional landmark-based position estimation from a single image. In *Proc. IEEE Int. Conf. on Robotics & Automation (ICRA)*, 1998.

[10] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 19(5):1997, May 1997.

[11] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *Int. Conf. on Computer Vision (ICCV)*, 1998.

[12] S. Simhon and G. Dudek. Selecting targets for local reference frames. In *Proc. IEEE Int. Conf. on Robotics & Automation (ICRA)*, pages 2840–2845, 1998.

[13] D. Slater and G. Healey. The illumination-invariant recognition of 3d objects using local color invariants. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 18(2):206–210, Feb. 1996.

[14] R. Swain-Oropeza and M. Devy. Motion control using visual servoing and potential fields for a rover-mounted manipulator. In *Proc. IEEE Int. Conf. on Robotics & Automation (ICRA)*, Detroit, USA, 1999.

[15] Y. Takeuchi, P. Gros, M. Hebert, and K. Ikeuchi. Visual learning for landmark recognition. In *Proc. Image Understanding Workshop*, pages 1467–1473, 1997.

[16] B. M. ter Haar Romeny, L. M. Florack, J. J. Koenderink, and M. A. Viergever. Scale space: Its natural operators and differential invariants. In *Information Processing in Medical Imaging*, pages 239–255. Springer Verlag, 1991.

[17] B. M. ter Haar Romeny, L. M. Florack, A. H. Salden, and M. A. Viergever. Higher order differential structure of images. In *13th Int. Conf. on Information Precessing in Medical Imaging (IPMI93)*, 1993.

[18] S. Thrun. Finding landmarks for mobile robot navigation. In *IEEE Conf. on Robotics & Automation (ICRA)*, pages 958–963, 1998.

[19] C. Tomasi and J. Shi. Good features to track. In *Proc. IEEE Conf. on Comp. Vision & Patt. Recog. (CVPR)*, pages 593–600, 1994.

[20] E. Yeh and D. Kriegman. Toward selecting and recognizing natural landmarks. In *IEEE Int. Workshop on Intelligent Robots & Systems*, pages 47–53, 1995.

[21] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Technical report, INRIA, 1994.