2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Detection and Filtering of Landmark Occlusions using Terrain Spatiograms

Damian M. Lyons, *Member, IEEE*

Abstract— **A team of robots cooperating to quickly produce a map needs to share landmark information between members so that the local maps can be accurately merged. However, the appearance of landmarks as seen by members of the team can change dramatically due to the phenomenon of occlusion.**

We have previously presented an approach to landmark representation using *Terrain Spatiograms* – an extension to image spatiograms in which the spatial information relates to the scene rather than the image. Because this representation preserves depth structure, it is possible to identify and filter potential occlusions.

We present an approach to identifying and filtering occlusions using terrain spatiograms, and report experimental results on 20 landmark datasets for varying states of occlusion. We show that occlusion can be detected and filtered, resulting in improved landmark matching scores.

I. Introduction

In previous work [7][8], we have addressed the problem of communicating landmark information for collaborative area mapping by a team of mobile robots. A team of robots is deployed to cooperatively generate a map of the area, an area under reconnaissance or an urban disaster site, for example. The objective is to generate an accurate map showing hazards, obstacles, traversable routes, etc., very quickly and to communicate it back to a command center. The map will then be used by a combination of human and robot teams for effective operations in the mapped area.

The local maps generated by each member of the robot team need to be merged to generate the final map. Odometry can be used as a first estimate of the relative locations of the local maps, but this is rarely sufficient. Instead common features in the local maps also need to be identified and used to merge the local maps. Unfortunately, features seen by one robot in the team may appear in a different pose or in an occluded state when seen by another robot in the team.

As an important step towards this objective, in this paper we investigate the problems of filtering landmark occlusions using a combined image and terrain spatial representation, *Terrain Spatiograms*, first proposed in [7][8]. Occlusion can be defined naturally in terms of depth-related visibility constraints. Since terrain spatiograms retain depth structure, this definition allows us to propose an occlusion identification and filtering approach. Section II presents prior work. In Section III we review the terrain spatiogram notation used in the paper. In Section IV we present the proposed method for identifying and filtering occlusions. Section V reports our experimental results for landmark occlusion filtering for 20 single-landmark, single-pose occlusion datasets. Section VI discusses these results in the context of future work.

II. Existing Work

Prior approaches to robot map representation include the use of topological maps – maps based on places and their interconnection, and also metric maps – maps based on accurate spatial measurements [11]. We argue that both approaches are necessary to solve the problem set in the previous section. Features whose appearance is independent of scale and rotation, e.g., SIFT features [6], are commonly used in metric mapping [9]. A collection of these micro-landmarks are matched to localize the robot accurately. A *cognitive map* is a biologically inspired, primarily topological map composed of *natural landmarks* identifying places, the edges identifying routes between places and augmented with navigation and hazard information. Landmark recognition is important in topological maps, clearly, but also important in metric mapping for loop closure.

Zhang and Kosecka [13] represent images of buildings using localized color histograms collected along the vanishing directions (well-defined for man-made landmarks such as buildings) and use SIFT features to refine the matching of the histograms. Cummins & Newman [3] describe an appearance based-approach to localization and mapping using the bag-of-words approach popular in image retrieval [12] and employing SURF [1] features. Places are represented as a probabilistic distribution over the appearance words. An advantage to using SIFT or SURF features is a natural robustness to occlusion: If some of the features are mismatched due viewpoint change or partial occlusion, enough matches may remain for identification.

Ramos et al. [9] shows that a combination of depth and image information can be a powerful tool for landmark recognition. Another representation that combines depth and image information is Birchfield and Rangarajan [1]'s *spatial histogram* or *spatiogram.* The image spatiogram extends an image histogram with a Gaussian distribution per histogram bin that summarizes the image location for the image pixels

This work was supported in part by the U.S. Department of Energy under Grant DE-FG02-08CH11542.

D. M. Lyons is with the Robotics & Computer Vision Laboratory, Computer and Information Science Department, Fordham University, Bronx, NY 10458 USA (phone: 718-817-4480, email: dlyons@cis.fordham.edu).

that fall in that histogram bin. However, image spatial information is related in a rather complication fashion to the scene spatial information.

We proposed an extension to the spatiogram, called the Terrain Spatiogram [7][8], in which the image spatial information is replaced by terrain spatial information. In [7] we presented experimental results for mutual landmark recognition on two different model robots equipped with different stereo cameras, and with terrain spatiograms collected on one robot being used on the other. In [8] we extended the Gaussian spatial model to a mixture of Gaussians and presented results on combining multiple views of a landmark into a single terrain spatiogram.

These results don't indicate whether the usefulness of a terrain spatiogram would be severely reduced because of the phenomenon of occlusion. Clearly, a purely appearance-based histogram approach might suffer from this problem since there is no way to detect the presence of occlusion, and as mentioned a feature-based approach would not suffer from this issue as severely. However, since occlusion is a depth-based phenomenon, and since terrain spatiograms preserve terrain depth, it may be possible to detect and filter the occlusion.

III. Terrain Spatiograms

For the convenience of the reader, we briefly summarize the concept of terrain spatiograms in this section.

A. Spatiograms.

Let $I : P \rightarrow V$ be a function that returns the value $v \in V$ of a pixel at a location $p \in P$ in the image. The histogram of I captures the number of times each pixel value occurs in the range of the function I. Consider a set, B, of equivalence classes on V, a histogram of I, written h_I maps B to the set $\{0,\ldots,|P|\}$ such that $h_I(b)=n_b$ and

$$n_b = \eta \sum_{i=1}^{|P|} \delta_{ib}$$

where δ_{ib} is equal to 1 *iff* the i^{th} pixel is in the b^{th} equivalence class and 0 otherwise, and η is a normalizing constant. A *spatiogram* or *spatial histogram* adds information about where values occur in the image:

$$h_I(b) = \langle n_b, \mu_b, \Sigma_b \rangle$$

where μ_b , Σ_b are the spatial mean and covariance of the values in the class b defined in the standard way.

Birchfield & Ragajaran define a histogram as a first order spatiogram, a formulation that also allows for second and higher order spatiograms. They also introduce an approach to comparing two spatiograms as the spatially weighted sum of similarities

$$\rho(h,h') = \sum_{b=1}^{|B|} \psi_b \rho_n(n_b, n_b')$$

where ψ_b evaluates the spatial means of bins in h in the spatial distributions of h' and where ρ_n compares the bin values. O'Conaire and Smeaton [4] developed a normalized spatiogram comparison measure (one in which $\rho(h,h)=1$), making it much more intuitive to use ρ to match two spatiograms.

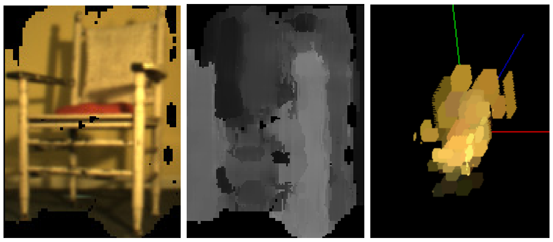

(a) (b) (c)

Figure 1: Image of chair landmark (a); monochrome stereo disparity for landmark (b); perspective view of color terrain spatiogram for landmark (c).

B. Terrain Spatiograms

The spatial dimensions used by Birchfield & Ragajaran and others are the spatial dimensions of the image and a primary use of spatiograms has been for color-based tracking in video images. There is nothing about the definition which constrains the spatial dimensions to be in the image. If, for example, the image information comes from a stereo camera, then the spatial information can be three-dimensional depth information.

The function $d(p)$ is introduced that maps a pixel at position p to its three-dimensional location in the viewed scene and the definition of the function δ_{ib} is modified so that $\delta_{ib} = 1$ *iff* the i^{th} pixel is in the b^{th} equivalence class and its stereo disparity is defined, 0 otherwise. The spatial moments for a *terrain spatiogram* then become:

$$\mu_b = \frac{1}{\sum_{j=1}^{|P|} \delta_{jb}} \sum_{i=1}^{|P|} d(p_i)\delta_{ib}$$

$$\Sigma_b = \frac{1}{\sum_{j=1}^{|P|} \delta_{jb}} \sum_{i=1}^{|P|} (d(p_i) - \mu_b)(d(p_i) - \mu_b)^T \delta_{ib}$$

For a robot to recognize a landmark, it computes a terrain spatiogram of the landmark and then compares that spatiogram with the terrain spatiograms of a list of stored landmarks. The spatial information must be landmark-centered rather than robot-centered [7] in order to effectively compare landmarks from different robots. We employ a variant on the normalized spatiogram measure introduced by O'Conaire and Smeaton [4] to compare two terrain spatiograms h and h':

$$\rho(h,h') = \sum_{b=1}^{|B|} \psi_b \sqrt{n_b n'_b}$$

where

$$\psi_b = 2(2\pi)^{0.5} |\Sigma_b \Sigma'_b|^{0.25} N(\mu_b ; \mu'_b, 2(\Sigma_b + \Sigma'_b))$$

is the normalized probabilistic spatial weighting term. In [8] we defined terrain spatiograms that employ a mixture of

Gaussians spatial distribution and the corresponding normalized comparison function.

C. Color Terrain Spatiograms

In [7][8], a color stereo image was represented as three channel terrain spatiograms, which is difficult to display accurately. In the current paper, we use a single color histogram where b_c bins are assigned to each color channel (b_c=25) and the histogram has $|B| = b_c^3$ bins in total.

If $s_b = \lceil 255/b_c \rceil$ then image pixel with color channel components $p=(r, g, b)$ is assigned to bin:

$$bin(p) = r + g\, s_b + b\, s_b^2$$

Fig. 1 shows an example color terrain spatiogram. Fig. 1(a) is the left image of a stereo pair taken using the SRI SmallVision [5] software and Videre digital stereohead[1]. Fig. 1(b) is a gray-level map of the stereo disparity, closer pixels being brighter. Fig. 1(c) shows a perspective view of the resulting color terrain spatiogram. The spatial and color content of the object in Fig. 1(a) is identifiable in the terrain spatiogram.

IV. Identifying and Filtering Occlusions

Landmark occlusion is a depth related phenomenon: a landmark is occluded when the occluding object hides a portion of the landmark image as a consequence of being between the image sensor and the landmark. Consider a landmark positioned at *p* relate to some Cartesian coordinate system. Let the XZ plane be the ground plane and Y the height. Let the image sensor be on the Z axis in the negative direction. If we look at the depth information, then we would expect to see a cluster of points representing the landmark itself, and additional clusters between the landmark and the image sensor representing occluding objects.

A. Identifying Occlusions

Fig. 2(a) is the left image of a stereo pair that shows a landmark (a table) occluded by a large box. Fig. 2(b) shows the image pixels mapped to depth and displayed in a perspective view. The Z axis is along the diagonal of the view. The occluding box is clearly separated out from the more distant table. A K-means clustering was applied to depth information in Fig. 2(b) projected to the XZ plane. Two clusters were identified and are shown in Fig. 2(c). A smaller occlusion case is shown in Fig. 2(d-f). The cluster weights were 0.45 and 0.53 for (a-c) and 0.45 and 0.47 for (d-f) indicating that between them the two clusters accounted for over 90% of the data. Since the terrain spatiogram preserves the spatial information, it's possible to determine what portion of the spatiogram corresponds to the landmark and what portion corresponds to the occlusion.

The landmark cluster can be separated from the occluding clusters by its weight (the landmark should be the principal object in view) and, if there are multiple clusters of the same weight, by its depth (the occluding cluster have to be in front of the landmark). We formalize this reasoning in the next section.

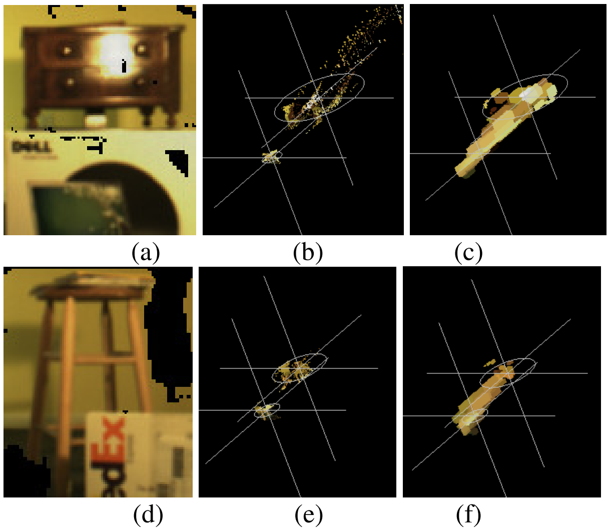


Figure 2: Occluded Landmark left image of stereo pair (a, d); perspective view of image pixels mapped to absolute depth (b, e); perspective view of terrain spatiogram with XZ cluster center and 1SD circle (c, f).

B. Filtering Occlusions

Even though the landmark cluster has been identified, and hence the portion of the terrain spatiogram can be determined, this is not sufficient to produce good matching (using ρ) in general for the following reasons:

1. The spatiogram is represented in landmark centered spatial terms, with origin the average depth of the landmark. However for an occluded landmark, the average depth will be moved away from the landmark origin and towards the center of the occluding object.
2. A portion of the landmark has been occluded. Therefore an occluded candidate landmark spatiogram may also be centered in X and Y planes differently from the landmark, since the visible portion of the landmark is different.
3. Finally, since a portion of the landmark has been occluded, the occluded candidate will be missing color and spatial structure unoccluded in the landmark spatiogram.

The unoccluded landmark spatiogram h_l is collected:

$$h_l = [\, N_l \;\; M_l \;\; S_l \,]$$

where *N, M* and *S* are the column vectors of n_b, μ_b and Σ_b respectively. Moments are calculated for the entire spatiogram as follows:

[1] Model STH-MDCS3

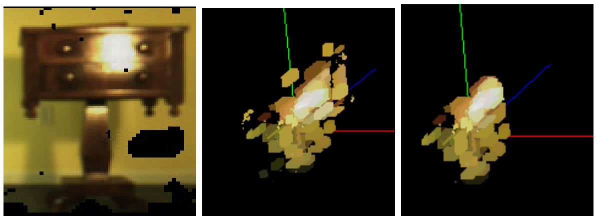


Figure 3: Terrain Spatiogram for unoccluded landmark (a); before (b), and after (c) trimming for outliers.

$$\overline{\mu_l} = \frac{1}{N_l 1^T} N_l M_l^T$$

$$\overline{\Sigma}_l = \frac{1}{N_l 1^T}(M_l - \overline{\mu_l})(M_l - \overline{\mu_l})^T$$

A visibility vector V_l is added to H_l to represent discarded portions of the spatiogram. For the landmark spatiogram, the visibility vector is used to trim outliers (Fig. 3) as follows:

$$v_b = \begin{cases} 1 & (\mu_b - \overline{\mu}_l)(\mu_b - \overline{\mu}_l)^T < n\overline{\Sigma}_l \\ 0 & \quad\quad else \end{cases}$$

The candidate, occluded landmark spatiogram h_o is collected:

$$h_o = [\ N_o\ M_o\ S_o\ ]$$

A K-means clustering is performed in the XZ plane and a set of cluster centers C_i and weights W_i produced:

$$(\ C_i\ ,\ W_i\ ),\ i \in \{\ 1, \ldots, K\ \}$$

The largest, rear-most cluster is identified as a candidate landmark cluster and other clusters are considered occlusions.

$$C_o = argmax\ (\ W_i\ )\ \ \&\ \ argmaxZ(\ \ C_i\ \ )$$

When h_o is matched against h_l the moments for the entire landmark spatiogram are used to calculate a visibility vector V_o for the candidate spatiogram and trim away the occlusions (Fig. 4).

$$v_b = \begin{cases} 1 & (\mu_b - C_o)(\mu_b - C_o)^T < n\overline{\Sigma}_l \\ 0 & \quad\quad else \end{cases}$$

Both the landmark spatiogram h_l and candidate landmark spatiogram h_o are translated to the origin for comparison, h'_l and h'_o respectively. However, since the X and Y components of the cluster center identified as the candidate landmark may reflect a distortion of the actual landmark center, only the depth (z) information is used to translate the candidate landmark:

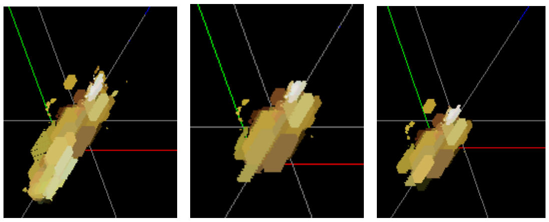


Figure 4: Candidate landmark spatiogram showing landmark cluster center (a); trimmed to landmark moments (b); and translated (c).

$$M'_l = (M_l - \overline{\mu}_l) \quad and \quad M'_o = (M_o - z(\overline{\mu}_l))$$

The translations won't effect the variances and hence

$$S_l' = S_l \quad and \quad S'_o = S_o.$$

Since both spatiograms now have hidden components, each needs to be renormalized about the intersection of their visibility vectors, calculated as:

$$V_I = \ V_l \circ V_o^T$$

$$N'_l = \frac{1}{N_l V_I^T} N_l \quad and \quad N'_o = \frac{1}{N_o V_I^T} N_o$$

Where ° is the Hadamard (component-wise) product. The filtered candidate and landmark can now be compared using the normalized comparison

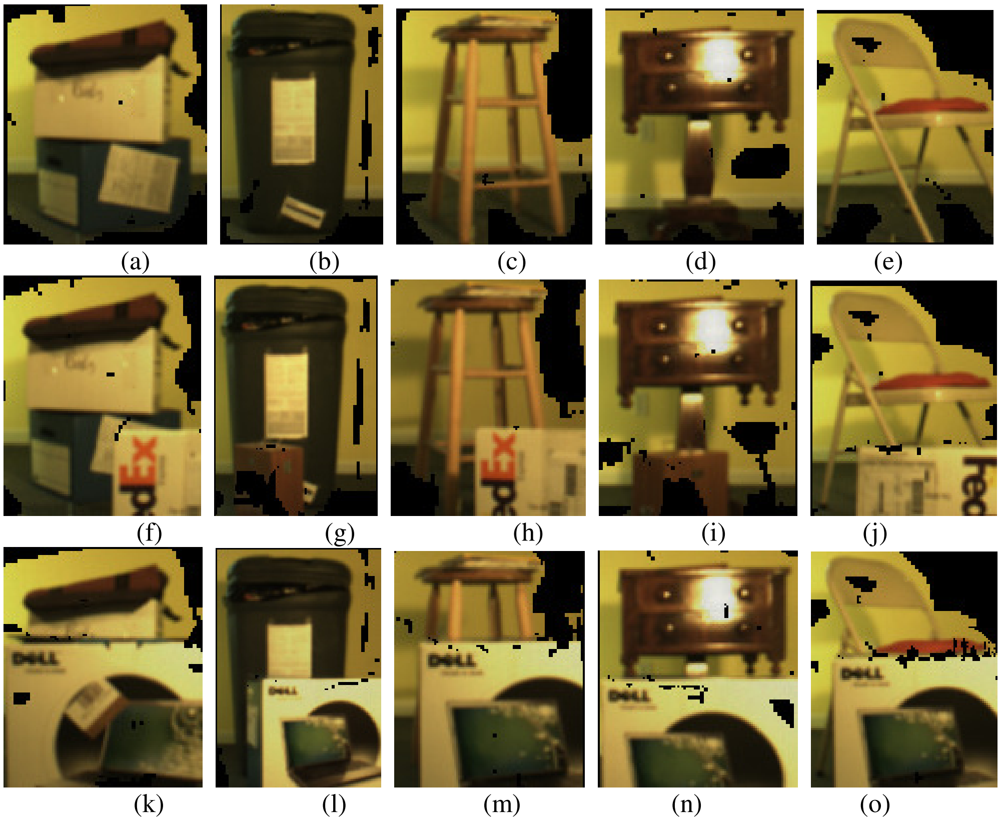
$$\rho(\ h'_l\ ,\ h'_o\ ) \in [\ 1, 0\ ]$$

V. Experimental Results

A. Method

The five landmarks shown in Fig. 5 were used to evaluate the approach. Image and disparity datasets were generated for the landmarks in an unoccluded state (Fig. 5(a-e)), a small occlusion state (Fig. 5(f-j)) and a larger occlusion state (Fig. 5(k-o)). A further set of 5 'invalid' occlusions datasets were also generated where a cloth was draped over a portion of the landmark, producing a non-depth detectible occlusion. Image and disparity information was gathered using the SRI SmallVision [5] software and Videre digital stereohead mounted on a Pioneer P3 robot positioned about a meter in front of the landmark in a well-lighted indoor location. The robot and (possibly occluded) landmark poses and lighting were substantially unchanged for all 20 measurements (See [7][8], for results relating to the comparison of unoccluded landmarks under differing pose and lighting conditions).

Terrain spatiograms where generated from the image and disparity information as follows. The SRI small vision


(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

Figure 5: Five landmarks used in occlusion experiments: top row (a-e), unoccluded objects; middle row (f-j), small occlusions; bottom row (k-o), larger occlusions.

software was used to generate depth estimates, and only image pixels with a defined disparity were collected (the regions of black in the images in Fig. 5 indicate regions of undefined disparity). The average depth of a landmark was estimated by sampling a window in the center of the image and filtering and normalizing the depth information around this point to generate landmark-centered depth information. A color spatiogram was generated for each case, using RGB information (since lighting was relatively invariant) and a color histogram size of b_c=*25* and $|B| = b_c^3$ bins.

B. Results

The five unoccluded terrain spatiograms were used to calculate a confusion matrix using the normalized comparison operation. The results are shown in Table 1. In general the mutual comparisons are quite low (with the exception of the stool and chair landmarks).

In Table 2 the occluded landmarks were directly compared to the unoccluded landmarks. The first three columns are the comparison of the landmark and two occlusion spatiograms with themselves. The fourth column is the comparison of the landmark for smaller occlusion and the fifth is the comparison for larger occlusion. The smaller occlusions show relatively good comparisons while the larger occlusions, not surprisingly, show much poorer comparisons.

Table 1: Confusion Matrix for Landmarks.

a	b	c	d	e	
1	0.434	0.463	0.385	0.416	**a**
0.483	1	0.417	0.459	0.335	**b**
0.486	0.351	1	0.545	0.61	**c**
0.41	0.4	0.533	1	0.449	**d**
0.485	0.258	0.61	0.486	1	**e**

Table 2: Direction Normalized Comparisons.

	ρ11	ρ22	ρ33	ρ12	ρ13
a	1	1	1	0.815	0.485
b	1	1	1	0.828	0.697
c	1	1	1	0.571	0.405
d	1	1	1	0.868	0.632
e	1	1	1	0.835	0.483

In Table 3, the results of the occlusion filtered comparisons are presented. The first column is the filtered comparison for the small occlusions and the second column the comparison for the larger occlusions. The third and fourth columns show the relative improvement over the respective Table 2 values.

Table 3: Occlusion Filtered Normalized Comparisons.

	ρ1'2'	ρ1'3'	ρ1'2' %change	ρ1'3' %change
a	0.905	0.694	11.113	42.86
b	0.893	0.885	7.871	26.92
c	0.632	0.549	10.721	35.628
d	0.917	0.812	5.687	28.574
e	0.914	0.611	9.536	26.455

It can be see that there are improvements in all cases, validating the approach. However, the improvements are most significant for the large occlusion cases. This is not surprising: the small occlusion comparison values were quite high anyway, so there is not always much additional evidence uncovered by removing the occlusion.

The most improvement was seen for the box landmark in Fig. 5(a, f, k). That may seem unusual since the landmark is heavily occluded in Fig. 5(k). However, the filtering process renormalizes the spatiograms around the intersection of the visibility vectors; that is, the comparison calculated is between the visible portion of the occluded landmark (no matter how small) and the corresponding portion of the unoccluded landmark.

Table 4: Normalized Comparisons with draped landmarks.

	ρ14	ρ1'4'	ρ1'4' %change
a	0.727	0.694	-4.53
b	0.83	0.864	4.095
c	0.867	0.92	6.034
d	0.748	0.799	6.738
e	0.623	0.581	-6.701

Finally, an additional set of five datasets were collected, in which the five landmarks were draped with an occluding red cloth. The red cloth seriously distorts the appearance of the landmark, but since its depth is not easily detectable for all the landmarks, it is more similar to 'repainting' a portion of the landmarks. This does not fit our definition of occlusion and the results are correspondingly poor.

VI. Conclusion

In this paper we have addressed the problem of detecting and filtering landmark occlusions using the terrain spatiogram representation. For the 15 landmark datasets showing valid levels of occlusion, the proposed approach does detect and improve landmark matching scores. For the 5 datasets with 'invalid' occlusions – occlusions that cannot be easily detected by depth analysis – the approach presents no overall improvement as expected.

There are several directions in which these results can be expanded. The detection and filtering approach presented here uses a single Gaussian model per bin. Extending it to handle a mixture of Gaussian model and corresponding normalized comparison [8] would make it more generally applicable but requires further work. From the experimental perspective, all the datasets reported here are from a single robot under very similar lighting and location conditions. In particular, there is no pose change between landmark and landmark candidates, and only a single landmark is visible in each scene. Our results need to be validated for a more comprehensive collection of datasets; the results reported in [7] for example were from different model robots, taken with differently configured stereo cameras, and taken in dramatically different lighting and scene conditions covering various outdoor and indoor locations at various times of day.

References

[1] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L.,"SURF: Speeded Up Robust Features", *Computer Vision and Image Understanding*, V110 N3, pp. 346-359, 2008.

[2] Birchfield, S., Rangarajan, S., Spatial Histograms for Region-Based Tracking, *ETRI Journal*, V29 N5, Oct. 2007.

[3] Cummins, M., and Newman, P., FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *Int. J. Rob. Research, V27 N6, 2009, pp.647-665.*

[4] O'Conaire, C., Smeaton, A.F., An Improved Spatiogram Similarity Measure for Robust Object Localization. *IEEE Int. Conf on Acoustics, Speech & Signal Proc.*, 15-20 Mar. 2007.

[5] Konolidge, K., Beymer, D., SRI SmallVision User's Manual v4.4d May 2007.

[6] Lowe, D., "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* 60, 2 pp. 91-110, 2004.

[7] Lyons, D. M., "Sharing and Fusing Landmark Information in a team of Autonomous Robots" *SPIE Defense and Security Symposium: Multisensor, Multisource Information Fusio*n, April 13-17, Orlando, FL 2009.

[8] Lyons, D. M., "Sharing Landmark Information using Mixture of Gaussian Terrain Spatiograms" IEEE/RSJ International Conference on Intelligent RObots and Systems *(IROS),* Oct 11-15, St. Louis MO, 2009.

[9] Ramos, F.T.; Nieto, J.; Durrant-Whyte, H.F, Recognising and Modelling Landmarks to Close Loops in Outdoor SLAM. *IEEE Int. Conf. on Robotics and Automation*, 2007.

[10] Se, S., Lowe, D., Little, J., "Local and Global Localization for Mobile Robots using Visual Landmarks," *Proceedings of the International Conference on Intelligent Robots and Systems* (IROS), pp.414—420, 2001.

[11] Thrun, S., Robot mapping: A Survey. Technical report CMU-CS-02-111, School of Computer Science, Carnegie-Mellon University, Pittsburgh PA Feb. 2002. In: *Exploring Artificial Intelligence in the New Millenium*, (Eds. Lakemeyer, G. and Nebel, B.) Morgan Kaufmann 2002.

[12] Yang, J., Ngo, C-W., Hauptmann, A., Jiang, Y-G., "Evaluating Bag-of-Visual-Words Representations in Scene Classification", ACM Multimedia Information Retrieval Workshop Augsburg, Germany, Sept. 28-29, 2007

[13] Zhang, W., and Kosecka J., Hierarchical Building Recognition, *Int. J. Comp. Vision, V25, 2007, pp 704-716.*