Selection and Recognition of Landmarks using Terrain Spatiograms

Damian M. Lyons, *Member, IEEE*

***Abstract*— A team of robots working to explore and map an area may need to share information about landmarks so as to register their local maps and to plan effective exploration strategies. In previous papers we have introduced a combined image and spatial representation for landmarks: terrain spatiograms. We have shown that for manually selected views, terrain spatiograms provide an effective, shared representation that allows for occlusion filtering and a combination of multiple views.**

In this paper, we present a landmark saliency architecture (LSA) for automatically selecting candidate landmarks. Using a dataset of 21 outdoor stereo images generated by LSA, we show that the terrain spatiogram representation reliably recognizes automatically selected landmarks. The terrain spatiogram results are shown to improve on two purely appearance based approaches: template matching and image histogram matching.

I. Introduction

The application domain considered in this paper consists of team of robots deployed to cooperatively generate a map of a specific 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. This map will then be used by a combination of human and robot teams for effective operations in the mapped area.

In previous work [4][5][6], we have proposed a combined image and terrain spatial representation for landmarks, the *terrain spatiogram*. However, in that work, the input images were manually windowed. In this paper, we introduce a saliency-based architecture, LSA, for automatically generating candidate landmarks. LSA follows a model of landmark saliency initially proposed by Rauball & Winter [11] for human way-finding.

Using 21 stereo datasets collected with LSA, we show that the terrain spatiogram approach can effectively recognize landmarks in a range of poses and scales. An image template matching approach and a color histogram matching approach is applied to the same dataset with inferior results.

This paper is laid out as follows. Previous work is reviewed in Section II. In Section III, the Landmark Saliency Architecture (LSA) is introduced. We recap the terrain spatiogram notation in Section IV. Experimental procedure and results are reported in Section V followed by discussion and conclusions in Section VI.

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).

II. Prior Work

Appearance-based approaches to landmark recognition include Zhang and Kosecka [14] representing images of buildings using localized color histograms collected along the vanishing directions, and Cummins & Newman [1] employing a SURF-based, bag-of-words approach for mapping and localization. Ramos et al. [10] show that a combination of depth and appearance information can be a powerful tool for landmark recognition to implement loop-closure for outdoor SLAM.

In [4][5], we introduced an approach combining image and spatial information based on Birchfield & Rangarajan [9]'s *spatial histogram* or *spatiogram.* With range sensing equipment, it is possible to relate the image positions of the spatiogram to Cartesian coordinates relative to the robot. A spatiogram using terrain rather than image spatial information is called a *terrain* spatiogram (or TSG). We have shown that the TSG is an effective approach to sharing information between robot platforms [4], combining multiple views of a landmark [5], and detecting and filtering landmark occlusions [6] when the input images are windowed to manually selected landmarks.

Some automatic landmark selection approaches are specific to the landmark representation being used, e.g., quadrangular patches in [3], and 2D patterns in [7]. On the other hand, saliency approaches [9] use information about visual attention [11][12] to determine general candidate image areas. A TSG represents a spatially compact portion of the environment and its appearance (color) information. This constraint is easily captured with saliency concepts. Furthermore, our objectives include sharing landmarks with humans – another reason for pursuing a saliency approach.

Rauball & Winter [11] present a formal model of landmark saliency for human travellers consisting of visual attraction, structural attraction and semantic attraction components. Their visual attraction component is what is usually seen in robot saliency architectures [9]. However, their structural attraction component allows the definition of the spatial compactness criteria for TSGs. Their semantic attraction component supports a well-defined communication channel for more general and task-related landmark selection, allowing different landmark selection criteria to apply when exploring, constructing a quick topological map of a new area, or constructed a metric map for a local region.

III. Landmark Saliency Architecture (LSA)

The purpose of the Landmark Saliency Architecture is to extract TSG landmark candidates from image and depth

views. The saliency criteria need to include visual and spatial regions that can be represented well by a TSG. However, we also want our landmarks to be useful for humans, so we include some criteria that relate to human visual attention.

A. *Model of Landmark Saliency*

Following Rauball & Winter [11]'s formal model of landmark saliency for human travelers, we consider the saliency of a landmark to consist of three components:

- Visual attraction,
- Structural attraction, and
- Semantic attraction.

We consider *visual attraction* to refer to iconic image properties, while *structural attraction* will refer to region properties. *Semantic attraction* captures the relevance of a landmark to an ongoing task, modifying the relative importance of the various image and feature properties in selecting a landmark.

B. *Visual Attraction*

The sensory input to the saliency architecture consists of an $n \times m$ visual image I_c and an $n \times m$ spatial image I_d registered as follows:

$$I_c = \{ c_{ij} = (v_1, v_2, v_3) \mid i \in 1..n, j \in 1..m \}$$
$$I_d = \{ d_{ij} = (x_1, x_2, x_3) \mid i \in 1..n, j \in 1..m \}$$

where d_{ij} is the spatial location in the terrain associated with the visual pixel c_{ij}.

Many aspects of human color preferences can be accounted for by considering a color space, based on retinal cone responses, that roughly corresponds to Red-Green and Blue-Yellow axes [9]. In their recent study, Schoss & Palmer [9] found that irrespective of gender, Green and Blue were in general the preferred end of these two axes, but that prior positive reward experience played an important role in personal color preferences. Based on this, we have selected the CIELab color opposition space for I_c where the a component corresponds to a Red-Green axis and the b component to a Blue-Yellow axis. In general, low a and high b values will be considered salient. However, semantic attractiveness (the 'prior experience' reported by [9]) needs to be able to modify this.

The visual attraction module of LSA is shown in Figure 1 with example stages in its processing shown in Figure 2. This module carries out iconic operations on the input image (it is applied in parallel to I_c and I_d) to segment smooth regions of high saliency. The first stage in the module applies a filter M_α, $\alpha_v \in \{-1,0,1\}$ to each plane of I, where -1 inverts the values on that plane (e.g., change from high-saliency red to high saliency green in I_c), 0 masks that plane (e.g., mask width and high and process only depth in I_d), and 1 passes that plane unchanged.

The module subsamples the filtered image at a scale s and computes the average Av_s and variance Var_s of the sub-sampled regions. Figure 2(a, b) and Fig. 2(a, d) show these images at $s=2$ for I_d and I_c respectively for the scene shown in Fig. 2(i). The variance image is thresholded with τ_v to establish the salient level of smoothness and to produce a binary image V_s used to remove unsmooth areas as follows:

$$R_s(I) = Av_s(I) \,.\, V_s(I)$$

Figure 2(c) and 2(f) show the salient smooth regions $R_s(I)$ for I_d and I_c respectively.

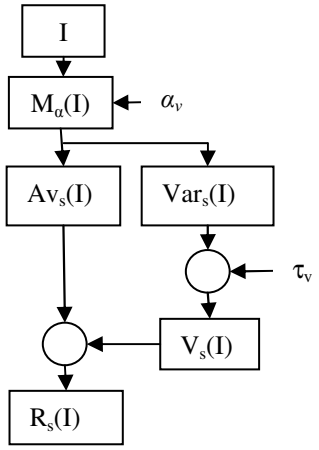


Figure 1: Visual Attraction Module

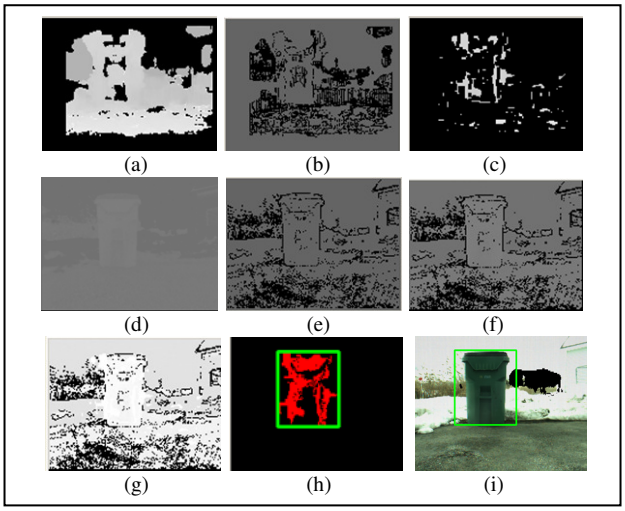


Figure 2: Landmark Saliency Example
(a-c): $Av_s(I_d)$, $Var_s(I_d)$, and $R_s(I_d)$;
(d-f): $Av_s(I_c)$, $Var_s(I_c)$, and $R_s(I_c)$;
(g-i): Fused Conspicuity map, Top saliency region, original image showing top region. Brighter is more salient in a - g.

The visual attractiveness module results are next processed for structural attractiveness. The settings of α_v and τ_v are part of the semantic attractiveness (subsection D).

C. *Structural Attraction*

The images $R_s(I_c)$ and $R_s(I_d)$ are the input to the structural attractiveness module, which focuses on salient region properties. The structural attraction module of LSA is shown in Figure 3.

The two images are linearly combined to form a fused conspicuity map [9] as follows:

$$FM(I) = w_c R_s(I_c) + w_d R_s(I_d)$$

where $w_c + w_d = 1$ and $0 \leq (w_c, w_d) \leq 1$. Fig. 2(g) shows the fused map. A connected components algorithm is used to generate a list of regions r, $r \in 1..k$, from the fused image.

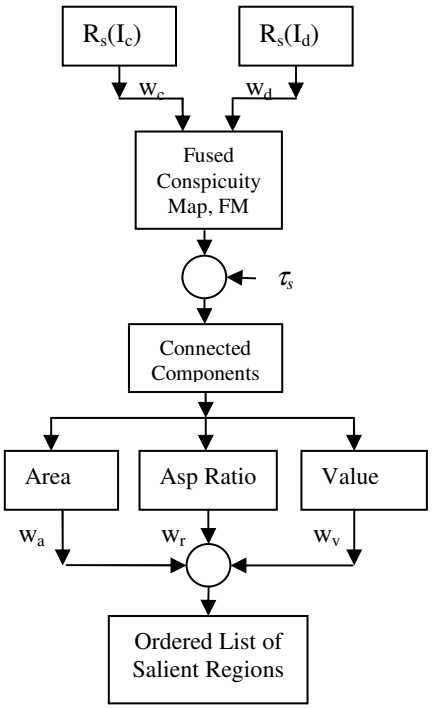


Figure 3: Structural Attractiveness Module

Three structural attractiveness properties are measured:

1) Region Area, a_r
2) Region aspect ratio, ar_r (the height of the region bounding box divided by its width)
3) Average fused attractiveness, v_r over the region

These properties are linearly combined for each region r to produce an overall saliency score for the region:

$$ss_r = w_a\, a_r + w_r\, ar_r + v_r\, w_v$$

where $w_a + w_r + w_v = 1$ and $0 \leq (w_a, w_r, w_v) \leq 1$. Figure 2(h) shows the top-ranked salient region for the example in Figure 2.

D. Semantic Attractiveness

The Semantic Attractiveness is the settings for the masks, thresholds and weight parameters for the visual and structural attractiveness modules. Rather than having these be fixed values, or 'tuning' parameters hidden in the architecture, we have chosen to make these explicitly visible so that LSA's selection of landmarks can be modified by the needs of the task at hand.

The following are the seven parameters of semantic attractiveness and a discussion of their settings:

$$(\alpha_v, \tau_v, w_c, w_d, \tau_s, w_a, w_r, w_v)$$

1. α_v : This parameter allows the salience of the input components to be reversed or masked, For example, human preference has Green preferred over Red. However, there are tasks, for example driving, where the Red of traffic signs should be more salient. With spatial information, a closer landmark might be preferred for constructing metric maps whereas a more distant (and hence more widely visible) might be preferred for topological mapping.
2. τ_v : This controls how smooth surfaces need to be to show up as salient. Man-made objects (walls, garbage bins) tend to be smoother than natural objects (bushes and trees).
3. w_c, w_d : These two mutually dependent parameters indicate how important spatial information is relative to color information. For the TSG landmark representation, a smooth spatial region is more useful than a smooth colored region.
4. τ_s : This controls how salient a fused region needs to be to appear in the list of regions. A task that is looking for many candidate landmarks (e.g. local, metric mapping) should set this low, whereas a task looking for few, but highly salient landmarks (e.g., global, topological mapping) should set this high.
5. w_a, w_r, w_v : These three mutually dependent parameters control the relative attractiveness of large regions versus small regions, vertical regions (tall) versus horizontal (squat) regions and high versus low fused visual attractiveness. Large, tall, close landmarks were preferred as candidates for TSG landmarks: Large, so that sufficient samples could be taken for the spatiograms; Tall, so that the samples were very compact with respect to the ground plane, and Close so that good depth precision was possible.

IV. Terrain Spatiograms

In this section, we briefly review the material on Terrain Spatiograms (TSG) from [4][5][6] for the benefit of the reader.

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

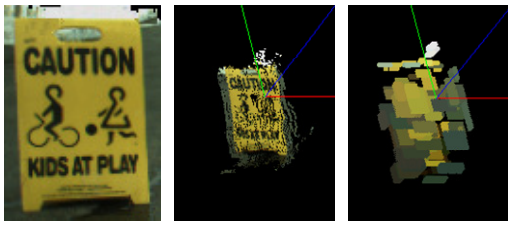
$$n_b = \eta \sum_{i=1}^{|P|} \delta_{ib}$$

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. Birchfield & Ragajaran define a

histogram as a first order spatiogram, a formulation that also allows for second and higher order spatiograms.



(a) (b) (c)

Figure 4: Terrain Spatiogram (TSG) Example
(a) Original image; (b) pixels mapped to depth; (c) TSG

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. Note that 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.

In [4] 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* (TSG) 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 TSG of the landmark and then compares that TSG with the TSGs of a list of stored landmarks. The spatial information must be landmark-centered rather than robot-centered [4] in order for it to be shared. We employ a variant on the normalized spatiogram measure introduced by [7] to compare two TSGs h and h':

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

where

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

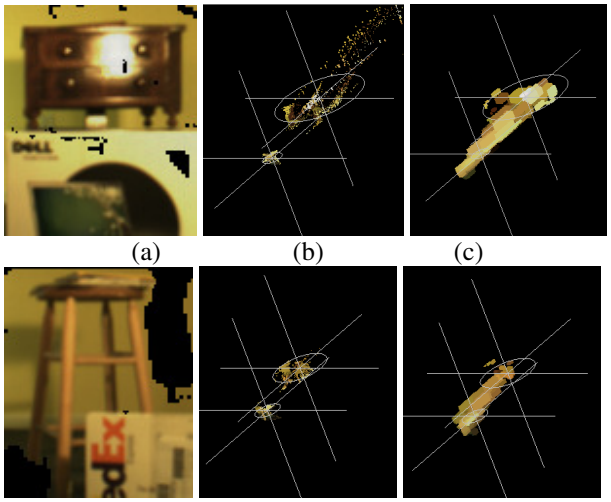
is the normalized probabilistic spatial weighting term.[1] In [5] we defined TSGs that employ a mixture of Gaussians spatial distribution and the corresponding normalized comparison function, and demonstrated how this could be used to combine multiple views of a landmark into a single TSG as well as to share landmarks between robots.

C. *Color Terrain Spatiograms*

In [4][5] a color stereo image was represented as three channel terrain spatiograms. This is quite difficult to display accurately. In the current paper as in [6] 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. Figure 4 shows an example color terrain spatiogram for one of the landmarks in this paper, a yellow road sign. Fig. 4(a) is the left image of a stereo pair taken using the Videre digital Stereohead[2]. Fig. 4(b) shows the image pixels mapped to their spatial location. Fig. 4(c) shows a perspective view of the resulting color terrain spatiogram. The spatial and color content of the object in Fig. 4(a) is identifiable in the terrain spatiogram.

D. *Identifying and Filtering Occluded Landmarks*

An advantage to using SIFT or SURF features for landmark representations, e.g., [1][14], is a natural robustness to occlusion: If some of the features are mismatched due to viewpoint change or partial occlusion, enough matches may remain for identification.



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

Figure 5: 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).

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

[1] It can be easily verified that $\rho(h,h)=1$.

[2] Model STH-MDCS3

itself, and additional clusters between the landmark and the image sensor representing occluding objects.

Figure 5 (a) is the left image of a stereo pair that shows a landmark (a table) occluded by a large box. Fig. 5(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. In [6], K-means clustering was applied to depth information in Fig. 5(b) projected to the XZ plane. Two clusters were identified, shown in Fig. 5(c). A smaller occlusion case is shown in Fig. 5(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 becomes possible to determine what portion of the spatiogram corresponds to the landmark and what portion corresponds to the occlusion.

V. Experiments

A. Experimental Procedure

The experiments were conducted on a Pioneer AT3 robot equipped with a Videre Stereocamera (6mm lenses) on a Biclops PT base. The robot was instructed to follow a loop around an outdoor traverse area in which there were a variety of objects. The robot stopped at regular distances along its traverse and collected sets of image and depth information from the Stereocamera, with pan angle set to 80°,90 °,100 ° (i.e., three side views). This resulted in a variety of views of the objects in the traverse area.

The traverse area was on a 7m×10m outdoor parking lot. The surface was blacktop and the pan angles used resulted in the robot always looking away from the parking lot over some grass and snow covered areas around the periphery of the lot.

The objects around the lot were mostly natural occupants of the area augmented with some additional candidate objects. A key issue for place detection in topological mapping and in loop-closure for SLAM is *perceptual aliasing* [2] – for this reason a number of similar appearing landmarks were chosen: the garbage bins in Figure 6 (a), 6(c) and 6(h). Additional candidate landmark objects included a large compressor (Fig. 6(e)) and a yellow sign (Fig. 6(f)). In total, LSA extracted eight landmarks at a variety of poses and scales, some of which are shown in Figure 6. Between four and ten poses for each landmark were generated by LSA.

B. TSG Landmark Recognition Results

Single Gaussian TSGs where generated for each LSA landmark candidate extracted (46 TSGs in total). These were filtered to the group of three best matches per landmark provided the match was above 0.6 (to eliminate poor landmarks). This resulted in one landmark candidate being discarded at all poses, leaving seven reliable landmarks, each with three poses. All the images in Figure 6 are best poses.

The 21 remaining TSGs were used to generate a confusion matrix, shown in Figure 7(a) as a 21×21gray-level image. The axes are the consecutive landmark and pose indices. The darker colors represent poorer matches. Figure 7(b) is a side view of a surface plot of the matrix, looking along the diagonal.

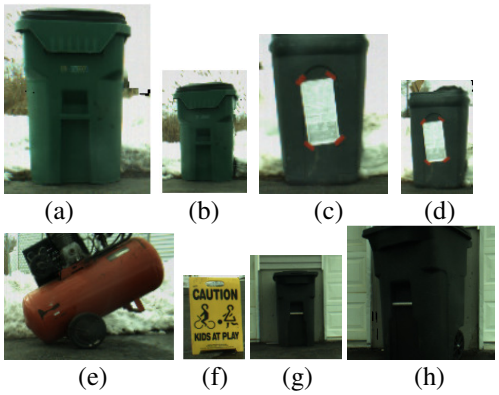


Figure 6: Some salient landmarks extracted by the LSA

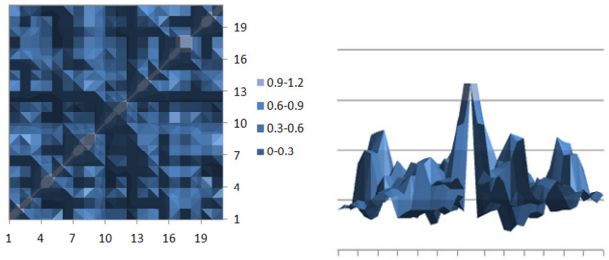


Figure 7: Confusion Matrix for TSG Comparisons

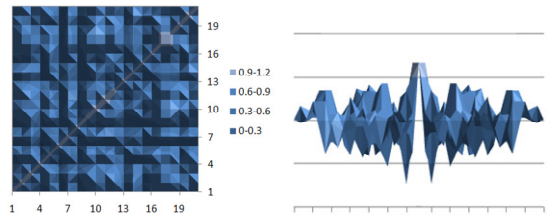
The strong diagonal band (of 7 3×3 submatrices) shows that different poses of a landmark are well recognized and well distinguished from other landmarks – despite the somewhat similar shape and color of the three garbage bins (Figure 6 (a), 6(c) and 6(h)) for example. This result from automatic landmark selection by the LSA reinforces our previous results for manually selected landmarks, documented in [4][5].

To illustrate the difficulty of the landmark recognition problem with this data set, two other approaches to landmark recognition from LSA results were used: a template-based approach and an image histogram based approach.

C. Image-based Landmark Recognition Results

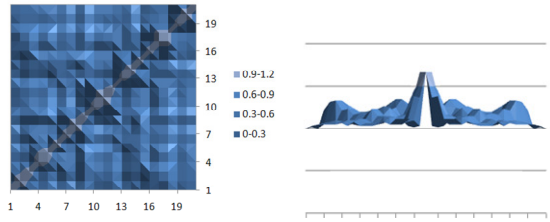
The template-based recognition approach normalized the rectangular image region produced by the LSA to a 90×60 template for each of the 21 landmark poses. A normalized confusion matrix was calculated for a Squared Sum of Differences comparison of templates with 1.0 being the best match and 0.0 being the worst. This is shown in Figure 8 with the same scale as Figure 7.

The image histogram approach extracted a normalized color histogram for each rectangular image region produced by the LSA. A confusion matrix was produced by using a Bhattacharayya histogram comparison operation modified to produce a 1.0 for the best match and 0.0 for the worst. This is shown in Figure 9 with the same scale as Figure 7.



(a) (b)

Figure 8: Confusion Matrix for SQDIFF Comparisons



(a) (b)

Figure 9: Confusion Matrix for Histogram Comparisons

The results of the three methods can be quantified further by removing the (1×21) diagonal and looking at the mean of the 3×2 of the remaining diagonal submatrix terms and the mean of the off-diagonal terms. This is shown in Table 1.

Table 1: Comparison of Confusion Matrix Means

Method	Diagonal	Off-Diagonal
TSG	0.79	0.37
SQDIFF	0.59	0.59
HISTO	0.84	0.68

Table 2: Variance Ratios for each method

Method	Var Ratio
TSG	0.77
SQDIFF	0.41
HISTO	0.68

The SQDIFF approach provides little help in distinguishing between landmarks, showing no statistical difference between poses of the same landmark and other landmarks. The HISTO approach identifies poses of the same landmark well, but the range between means is small, only 17% as opposed to 51% for the TSG approach. Finally, the ratio of variance of the matrix divided by sum of the variances for the diagonal and off-diagonal terms yields a measure of the discriminative power of each method for these landmarks. This is shown in Table 2.

VI. DISCUSSION

We have introduced a landmark saliency architecture, LSA, based on Raubal & Winter's model of landmark saliency. In addition to the visual attraction component modeled by most saliency architectures, this includes a structural attractiveness component, capturing the spatial conciseness criteria for candidate TSG landmarks, and a semantic attractiveness component, a channel by which the task at hand can influence landmark saliency.

We show that landmarks selected automatically by LSA can be recognized reliably when represented as TSG landmarks. However, when template matching or image histogram approaches are used, the recognition is less reliable.

This result supports our previous results [4][5][6] for terrain spatiograms. However, this paper's results were based on unshared, single view, and non-occluded landmarks. Future work will need to evaluate LSA used to build multiple view landmarks and to share landmarks. This latter is not trivial since the landmarks will need to appear salient on both robot platforms. The interaction of LSA with occluded landmarks may also be an issue, since both occluder and landmark may need to appear salient.

REFERENCES

[1] 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.*

[2] Dudek, G., and Jenkin, M., Computational Principles of Mobile Robotics. Cambridge Press 2000.

[3] Hayet, J., Lerasle, F., Devy, M., A visual landmark framework for mobile robot navigation, *Image and Vision Computing V25 N8*, Aug. 2007.

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

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

[6] Lyons, D., Detection and Filtering of Landmark Occlusions using Terrain Spatiograms. *IEEE Int. Conference on Robotics and Automation,* Anchorage, Alaska, May 2010.

[7] Mata, M., Armingol, J., de la Escalera J., and Salichs, M., Mobile Robot Navigation Based on Visual Landmarks Recognition, *IFAC Symposium on Intelligent Autonomous Vehicles*, Sapporo-shi, Japan, 2001.

[8] O'Conaire, and 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.

[9] Ouerhani, N., von Wartburg, R., Hugli, H., and Muri, R., Empirical Validation of the Saliency-based Model of Visual Attention , *Electronic Letters on Computer Vision and Image Analysis* 3(1):13-24, 2004.

[10] 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.

[11] Raubal, M., and Winter S., Enriching Wayfinding Instructions with Local Landmarks, *2nd Int. Conf. on Geographic Information Science, (GIScience)* Boulder, CO, USA, September 25-28, 2002.

[12] Schloss, K., and Palmer, S., Valence Theory of Human Color Preferences. *Journal of Vision*, *9*(8):358, Aug. 2009.

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

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