Combined Visually and Geometrically Informative Link Hypothesis for Pose-Graph Visual SLAM using Bag-of-Words

Ayoung Kim* and Ryan M. Eustice†
*Department of Mechanical Engineering
†Department of Naval Architecture & Marine Engineering
University of Michigan
Ann Arbor, Michigan 48109-2145
Email:{ayoungk, eustice}@umich.edu

***Abstract*— This paper reports on a method to combine expected information gain with visual saliency scores in order to choose geometrically and visually informative loop-closure candidates for pose-graph visual simultaneous localization and mapping (SLAM). Two different bag-of-words saliency metrics are introduced—global saliency and local saliency. Global saliency measures the rarity of an image throughout the entire data set, while local saliency describes the amount of texture richness in an image. The former is important in measuring an overall global saliency map for a given area, and is motivated from inverse document frequency (a measure of rarity) in information retrieval. Local saliency is defined by computing the entropy of the bag-of-words histogram, and is useful to avoid adding visually benign key frames to the map. The two different metrics are presented and experimentally evaluated with indoor and underwater imagery to verify their utility.**

I. Introduction

One of the most important and difficult problems in pose-graph visual simultaneous localization and mapping (SLAM) is determining a loop closure event. Loop closure in visual SLAM is obtained by recognizing previously viewed scenes. This recognition process involves identifying possible candidate image pairs and attempting to obtain a camera-derived relative-pose constraint. By correctly associating and registering corresponding image pairs, the uncertainty of both the map and the robot pose can be reduced and bounded.

Necessarily, this task involves choosing optimal loop-closure candidates because (*i*) the cost of estimating the camera-derived relative-pose constraint is computationally expensive (per loop closure candidate) and (*ii*) adding unnecessary/redundant measurements may result in over confidence [1]. In this paper, we focus on pose-graph visual SLAM methods where nodes in the graph correspond to image key frames and measurements appear as constraints (links) between nodes.

One way to intelligently hypothesize link candidates is to examine the utility of future expected measurements—an elegant method for measuring such uncertainty is to use information gain [1]. Typically, information gain refers to either Fisher information or mutual information. By definition, the Fisher information matrix is closely related to the Cramer Rao Lower Bound [2], while mutual information is derived from entropy [3] as defined by Shannon [4]. An example usage of information gain for control can be found in [3], [5], [6], where the proposed control scheme evaluates the information gain of possible measurements and leads the robot on trajectories that reduce the overall navigation and map uncertainty. Another application of information gain in SLAM can be found in the task of link hypothesis, where only informative links (measurements) are added to the pose-graph to significantly reduce the computational cost of inference [1].

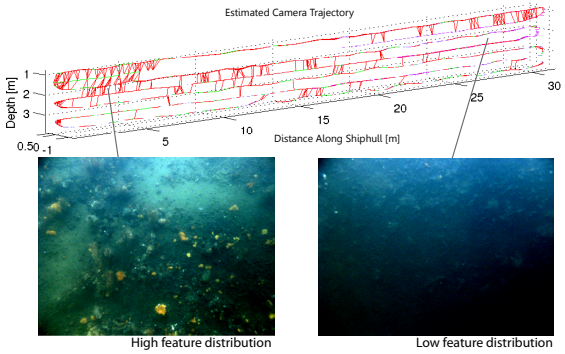


Fig. 1. Estimated trajectory of an underwater robot mapping the below-water surface of a ship hull. (top) Trajectory of the robot with successful cross-track camera registrations depicted as red lines. (bot) Representative images indicative of the image feature content within that area. Note that the density of pose-graph links is spatially correlated with feature content.

In the approaches described above, an equal likelihood of measurement availability is assumed. In other words, the information gain measure assesses the geometric value of adding the constraint *without regard to if, in fact, the constraint can be made.* In our scenario, camera-derived measurements may not be uniformly available within the environment due to the fact that the spatial distribution of visual features can vary greatly. For example, Fig. 1 depicts an underwater environment where there is a large variance in feature distribution (the surface of a ship hull). Here, successful camera-derived measurements (red links) occur when rich feature distributions are available, and in visually benign regions the camera produces few, if any, constraints.

Thus, as demonstrated by Fig. 1, the distribution of visual features dominates the availability of camera-derived mea-

surements, and hence, the overall precision of navigation. This indicates that successful camera measurements strongly depend upon the visual saliency of the scene, especially in feature-less areas or environments with repeated cues. Therefore, the ability to score meaningful key frames within a sequence of images is necessary.

A. Review of Bag-of-words

Our approach to scoring meaningful key frames begins with image representation. Foremost, the image needs to be represented in a compact way to reduce data storage and to facilitate later stages of feature-based image matching. One way to represent images is to use a feature detector and descriptor. Popular descriptors include the Scale Invariant Feature Transform (SIFT) [7] and Speeded Up Robust Features (SURF) [8]. These are known to be robust to scale, rotation and illumination changes. Another innovative way to describe features is to use a bag-of-words representation, which is the method we have adopted in this research. Originally developed for text-based applications, the bag-of-words approach was adapted and expanded to images by Sivic and Zisserman [9]. In this approach, each image is considered to be a document, and each feature descriptor found in the image corresponds to a word in the document. This representation reduces an image to set of vocabulary words, allowing for aggregate content assessment and enabling faster search. This approach has been successfully applied in areas such as image annotation [10], image classification [11], object recognition [12], [13] and also appearance-based SLAM [14]–[16].

B. Review of Visual Saliency

Once the conversion of images into a compact representation is decided, we wish to add another attribute for each image—namely the saliency of an image. The term "saliency" is a measure of how distinctive an image is, and has been defined by Kadir and Brady [17] using entropy. In their work, entropy was computed for a local patch and used in detecting the features of an image. If a patch has higher entropy, it likely has more randomness and thus is considered a feature. Similarly, Lee and Song [18] extended this entropy approach to color images, where the entropy of each Hue Saturation Value (HSV) intensity channel is computed. For grayscale images where no color channels are available, a Gabor-like function can be applied before the entropy is computed to obtain the saliency metric [19].

Entropy-based approaches can also be used when examining an entire image for global saliency. In [20], the author combined HSV channel entropy with Gabor filtered texture entropy to compute saliency for a color image in an underwater environment. This work successfully built saliency maps of the scene; however, its broad application was limited due to its reliance on source color imagery (i.e., excludes grayscale).

Recently, several bag-of-words saliency metrics have been explored [21]–[24]. Among the "words" (descriptors) appearing in an image, only the salient words are selectively captured and referred to as a bag-of-keypoints in [23]. In [24], a histogram of the distribution of words was used as a global signature of an image, and only salient regions were sampled to solve an object classification problem.

Although it is not specifically indicated as saliency, place recognition is another area where saliency is used in order to avoid perceptual aliasing. One remarkable development in the model-based place recognition approaches is Fast Appearance-Based mapping (FAB-MAP) [14]. In FAB-MAP, the model learns common words during an offline training phase, which it then uses to down-weight common (non-salient) words. Out of the non-model-based approaches, a popular statistics-based approach is called term frequency-inverse document frequency (tf-idf) [25]. This statistic is widely used in classification problems due to its simplicity and robustness. It emphasizes rare occurrences resulting in higher tf-idf scores for statistically salient words and typically can be learned online. In this paper, we use a statistics-based bag-of-words approach to define two saliency metrics—local and global. Global saliency is closely related to scene rarity, whereas local saliency refers to image texture richness.

II. Visually Augmented Navigation

The objective of this work is to combine measures of expected information gain with measures of visual-saliency in order to choose geometrically and visually informative loop closure candidates for pose-graph visual SLAM. In this section we succinctly review the pertinent aspects of our pose-graph monocular SLAM formulation, which we call visually augmented navigation (VAN) [26].

A. State Representation

We model the vehicle state using a six degree of freedom (DOF) representation for pose (position and Euler orientation), $\mathbf{x}_v = [x, y, z, \phi, \theta, \psi]^\top$, where pose is defined in a local-navigation frame. To approximate the time-evolution of vehicle state, we use a continuous-time constant-velocity kinematic model driven by white noise, which is then linearized and discretized to provide a linear time-varying discrete-time model of the vehicle dynamics.

We employ a pose-graph SLAM representation of the environment and therefore augment our state description to include a collection of historical vehicle poses sampled at regular spatial intervals throughout the environment. Each of these pose samples, $\mathbf{x}_{v_i}$, corresponds to a time instance t_i when an image key frame is stored by our visual perception process. Therefore, our augmented state representation for n key frames is denoted

$$\boldsymbol{\xi} = \left[\mathbf{x}_{v_1}^\top, \cdots, \mathbf{x}_{v_i}^\top, \cdots, \mathbf{x}_{v_n}^\top, \mathbf{x}_v^\top(t)\right]^\top.$$

The distribution of this augmented state representation is modeled as jointly Gaussian and is parameterized in the inverse covariance form as

$$\boldsymbol{\eta} = \Lambda\boldsymbol{\mu} \;\text{ and }\; \Lambda = \Sigma^{-1},$$

where $\boldsymbol{\eta}$ and Λ are the information vector and matrix, respectively, written in terms of the more familiar $\boldsymbol{\mu}$, Σ mean and covariance parameterization. We use an extended information filter (EIF) for inference, which is computationally efficient for sparse pose-graph representations. The reader is referred to [27], [28] for details.

B. Geometric Information Gain

We call the process of hypothesizing possible loop-closure candidates "link proposal", because a measurement will act as a "link" (i.e., constraint) between two nodes in our pose-graph framework. A simple way to propose candidate links would be to measure the expected Euclidean distance between two nodes in the graph and to statistically calculate whether or not their sensor field of views could have any overlap [29]. Candidate links are then those with large overlap ratios.

More intelligently, we could additionally calculate the predicted information gain that would be obtained from making such a measurement—retaining only those candidate links that have high expected information gain [1]. Adopting the mutual information definition of information gain as proposed in [1] and [6], which compares the current and predicted entropy after a measurement update between poses i and j, we can write the candidate link's predicted information gain as

$$\mathcal{I}_g = H(\boldsymbol{\xi}) - H(\boldsymbol{\xi}|\mathbf{z}_{ij}). \tag{1}$$

Here $H(\cdot)$ is entropy [4] and $\mathbf{z}_{ij}$ is the expected measurement between poses i and j.

For a Gaussian distribution, (1) simplifies to

$$\mathcal{I}_g = \frac{1}{2} \ln \frac{|\Lambda + \Delta\Lambda|}{|\Lambda|}, \tag{2}$$

where Λ is the information matrix as defined previously, $\Delta\Lambda = \mathrm{J}^\top \mathrm{R}^{-1} \mathrm{J}$ is the additive EIF measurement update, J is the Jacobian of the measurement model, and R is the (expected) measurement covariance. Ila et al. [1] showed that in the case of a pose-graph EIF, this information gain measure can be efficiently evaluated.

Note that this information gain measure is induced from geometry alone (through the Jacobian), and that the *act of perception* is not specifically accounted for. In our case, candidate links with high information gain may not be the best plausible camera links due to a lack of visual saliency. We argue that the registrability (i.e., act of perception) should play a key role in determining candidate image pairs. In what comes next, we describe a framework for *(i)* measuring visual saliency and *(ii)* combining it with geometric information gain.

III. Saliency

Visual saliency strongly influences the likelihood of making a successful pairwise camera estimate. When spatially overlapping image pairs fail to contain any locally distinctive textures or features—image registration fails. As a measure of the registration "usefulness" of an image, this paper adopts a bag-of-words scheme to represent images, and furthermore to score their saliency. We extract 128-dimension SURF key descriptors [8], and project them onto our vocabulary tree, which we build online. By doing so, we convert an image containing features to a document containing words, thus allowing us to leverage a rich set of statistics and research originally developed in the document search community.

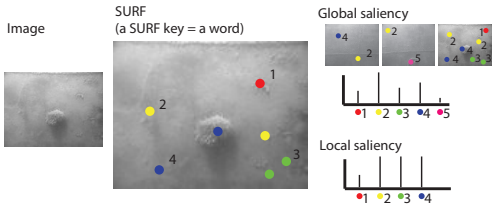


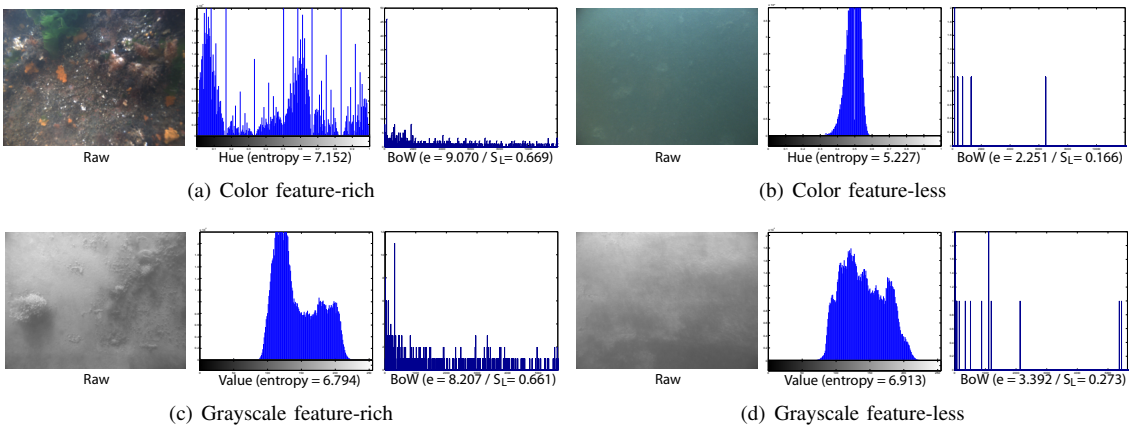
Fig. 2. Vocabulary construction and saliency score computation. SURF keys are extracted from an image and are used to update local and global histograms. Entropy from the local histogram detects feature richness, while idf from the global histogram captures rarity.

In this paper, we focus on two different measures of saliency: local (i.e., registrability) and global (i.e., rarity). Registrability refers to the intrinsic feature richness of an image (i.e., the amount of texture an image has). The lack of image texture, as in the case of mapping an underwater environment with a lot of benign regions (Fig. 1), prevents image registration from measuring the relative-pose between two locations. The amount of texture, however, is not the only term that defines saliency—an easy counterexample to this would be a checkerboard pattern or a repetitive brick wall. Images of these scenes would have high texture, but would likely fail to achieve a correct registration due to spatial aliasing of common features. As stated by Kadir and Brady [17], rarity should play a role in defining saliency.

A. Vocabulary Generation

Before defining our two measures of saliency, we first need to outline how we construct our vocabulary tree. Two concerns are relevant to our vocabulary building procedure: *(i)* we assume no prior knowledge of the environment; and *(ii)* the vocabulary must be representative.

Offline methods for vocabulary generation typically use a pre-processing stage with a clustering algorithm over a representative training data set [30]. Other studies have focused on online methods, which incrementally build the vocabulary tree during the data collection phase [15], [16]. One advantage of offline methods is that an even distribution of vocabulary words in descriptor space can be guaranteed; however, one disadvantage is that the learned vocabulary can fail to represent words collected from totally different data sets (e.g., using an indoor data set for underwater images). Online construction methods provide flexibility to adapt the vocabulary to incoming data, though, equidistant words are no longer guaranteed, which can lead to vocabulary fragmentation. Another concern about vocabulary construction is that the use of a single metric for an entire image, as we do in this paper, might be vulnerable to illumination and viewpoint changes [30].



(a) Color feature-rich (b) Color feature-less

(c) Grayscale feature-rich (d) Grayscale feature-less

Fig. 3. Local saliency test for color and gray-scale images. The leftmost column contains the raw image, the second column contains the histogram of the hue channel for color images and the intensity channel for gray images, and the third column contains the bag-of-words histogram. For color images ((a), (b)), both the hue channel and the bag-of-words histogram capture the feature richness properly. For gray-scale images ((c), (d)), the intensity histogram fails to detect feature richness, whereas the bag-of-words histogram still works well.

In this paper we have decided to pursue an online construction approach that initially starts from an empty vocabulary tree similar to [15], [16]. SURF features are extracted from the incoming image and are matched to existing words in the vocabulary using a Euclidean distance metric. Whenever the Euclidean distance is larger than a pre-specified threshold, we augment our vocabulary tree to contain the new word. To guarantee independent observations of words used in vocabulary generation, we only sample from images that are spatially distinct (i.e., without overlap). Fig. 2 depicts the overall process whereby two different histograms are calculated to quantify the image's saliency score—which is discussed next.

B. Local Saliency

We define local saliency as a measure of the ability of two images to be registered (i.e., texture richness). Underwater images contain few man-made features and often consist of empty or monotone sequences of images. To represent such images, many researchers have found that a bag-of-words histogram is useful. Under this scheme, we examine the entropy of the bag-of-words histogram to capture the diversity of the words (descriptors) appearing in an image:

$$e = -\sum_{i=1}^{|w|} p(w_i) \log_2 p(w_i) \tag{3}$$

where $p(w_i)$ is the empirical word distribution pdf computed from the image over words w_i and the size of the vocabulary is $|w|$.

Since we build the vocabulary online, the size of our vocabulary varies with time. Hence, to normalize our entropy measure, we look at the ratio of e to the maximum possible entropy[1] (which depends upon the size of the vocabulary) to yield a normalized entropy measure:

$$S_L = \frac{\sum_i p(w_i) \log_2 p(w_i)}{\log_2 |w|}. \tag{4}$$

[1]Note that the maximum entropy (i.e., $\log_2 |w|$) occurs when words are uniformly distributed for an image.

Fig. 3 depicts the local histogram and the normalized saliency score computed for color and grayscale images of an underwater scene. For the color images, the second column of Fig. 3 shows the hue channel histogram whereas the third column represents the bag-of-words histogram. We computed the hue channel histogram following [20], and verified that our normalized bag-of-words entropy score reveals a useful measure of image feature richness similar to the hue channel histogram. However, for grayscale images where no hue channel is available, the intensity channel histogram fails to distinguish the feature richness of a scene (Fig. 3(c), Fig. 3(d)), but as can be seen from the bottom row, our bag-of-words histogram measure works equally well for both grayscale and color imagery.

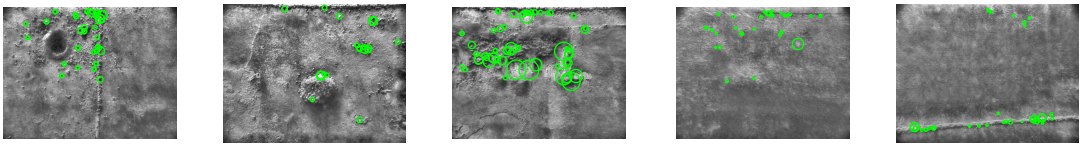
C. Global Saliency

In defining global saliency, we were motivated by a metric called term frequency-inverse document frequency (tf-idf), which is a classic and widely used metric in information retrieval [31]–[33]. This metric was first adopted in the computer vision community by [25] and subsequently shown to produce successful results in image-based classification [34], place recognition [35], and appearance-based SLAM [15], [16]. In a computer vision application, tf-idf for a word w is defined as:

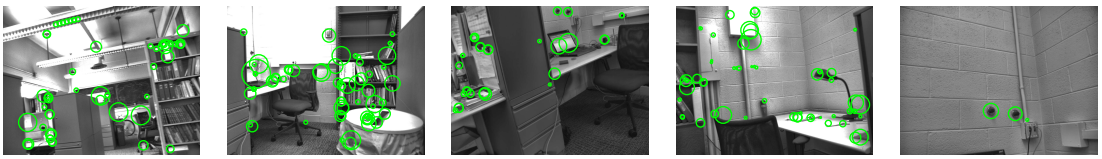
$$t_w = \frac{n_{wd}}{n_d} \log \frac{N}{n_w}, \tag{5}$$

where n_{wd} is the number of occurrences of word w in document d; n_d is the total number of words in document d; N is the total number of documents seen so far; and n_w is the number of documents with the occurrence of word w in the database so far. tf-idf captures the importance of a word (descriptor) appearing in a document (image) by penalizing common words.

Although tf-idf is prevalent in the text mining literature, its more fundamental form can be defined from inverse document frequency (idf) [32], [33], [36]; idf corresponds to the second term in (5), and has a higher value for words seen



(a) S_G=0.78 / S_L=0.81 (b) S_G=0.74 / S_L=0.77 (c) S_G=0.72 / S_L=0.78 (d) S_G=0.47 / S_L=0.74 (e) S_G=0.46 / S_L=0.65



(f) S_G=0.66 / S_L=0.73 (g) S_G=0.65 / S_L=0.76 (h) S_G=0.61 / S_L=0.73 (i) S_G=0.49 / S_L=0.61 (j) S_G=0.47 / S_L=0.68

Fig. 4. Global saliency test for underwater and indoor images. Salient features are marked with green circles. The global saliency score (S_G) and local saliency score (S_L) are provided below each image. Global and local saliency tend to agree on many cases; however, the global score can be low even with texture rich scenes (e.g., (d), (e), (i) and (j)).

less throughout a history. In other words, we expect high idf for features that are rare in the data set. In computer vision, Jegou et al. [37] uses a variation of idf to detect "burstiness" of a scene, noting idf's ability to capture frequency. In this paper, we use a sum of idf in an image to score global saliency for the image i:

$$s_i(t) = \sum_{w=1}^{n_d} \log_2 \frac{N(t)}{n_w(t)}. \tag{6}$$

The global saliency score is a function of time, where $N(t)$ is the total number of images and $n_w(t)$ is the number of images containing word w at time t. Since even a common word could be considered "rare" on its first occurrence, we use an inverted index scheme [31] to sparsely store and update the global saliency of all images that contain word w anytime word w is detected. This update happens with every image event to prevent the algorithm from scoring images taken early in the trajectory as being always globally salient.

Similar to our local saliency measure, we normalize the global saliency score to have a value between $[0, 1]$ using the maximum global saliency score encountered thus far:

$$S_{G,i}(t) = \frac{\sum_{w=1}^{n_d} \log_2 N(t)/n_w(t)}{\max_j s_j(t)}. \tag{7}$$

D. Comparison

Applying the two saliency metrics (i.e., local and global) to sample underwater and indoor office imagery, Fig. 4 shows the utility of these metrics in categorizing image saliency. As can be seen, our normalized local saliency score (S_L) is related to texture richness, whereas rarity of an image is related to our global saliency score (S_G). A feature-rich image tends to be locally salient; however, a high local saliency score does not always imply a high global entropy value. For example, the last two columns (Fig. 4(d), (e), (i), and (j)) show that global saliency can be low even with high texture. This is because several of the vocabulary words (e.g., bricks, marine growth) have been seen so many times throughout the image stream that they end up lowering their overall idf score.

E. Saliency Incorporated Information Gain

Based upon the two saliency metrics developed from the previous section, we now introduce a method for combining visual saliency with expected information gain to arrive at a combined visual / geometric measure. We found that the combined approach results in better link hypothesis in pose-graph visual SLAM—choosing candidate links that are both geometrically *and* visually informative.

1) IG with Local Saliency: The normalized global and local saliency measures are in the range of $[0, 1]$. Therefore, imposing local saliency on the geometric information gain can be written as

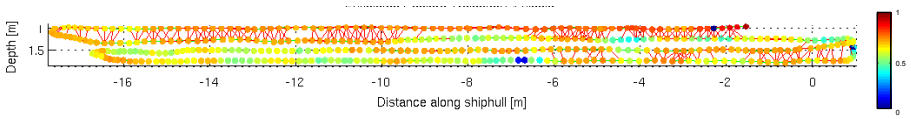
$$\mathcal{I}_L = \begin{cases} \mathcal{I}_g \cdot S_L & \text{if } S_L \geq S_{L,thres} \text{ and } \mathcal{I}_g \geq I_{g,thres} \\ 0 & \text{o.w.} \end{cases} \tag{8}$$

where $S_{L,thres}$ is a lower threshold for texture richness and $\mathcal{I}_g$ is the geometric information gain from (2). Strictly speaking, (8) is no longer a direct measure of information gain in the mutual information sense; however, it is a scaled version according to visual saliency. This allows us to quickly cull candidate image pairs whom have a low chance of perceptually matching due to low image feature content. In this paper, all results are shown for $I_{g,thres} = 0.2$ and $S_{L,thres} = 0.6$ (vertical line in Fig. 8).

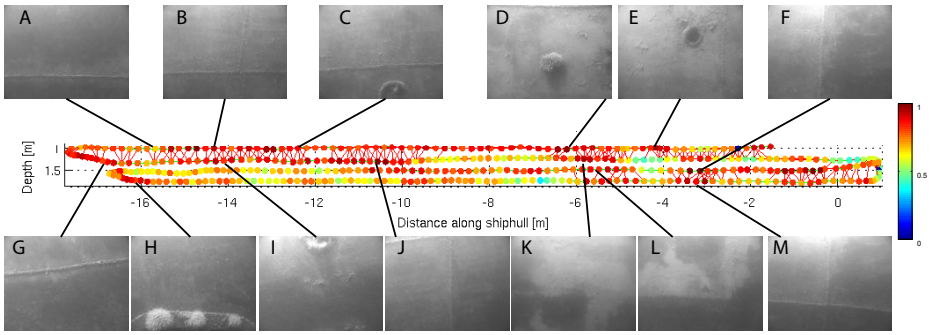
2) IG with Global Saliency: Global saliency incorporated information gain between the current and the ith pose is subsequently defined by multiplying the local information gain measure with the normalized global saliency score:

$$\mathcal{I}_{G,i} = \mathcal{I}_L \cdot S_{G,i}. \tag{9}$$

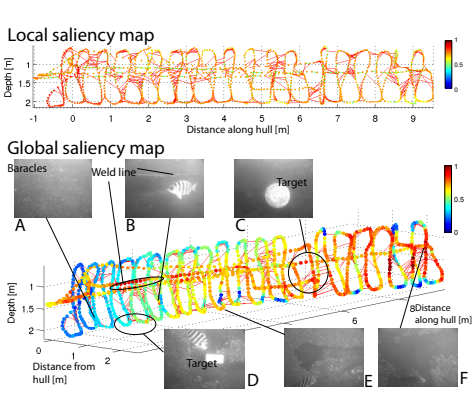
This results in a high information gain only for links that are both locally *and* globally salient, which are likely to be remarkable places in the environment that are also registrable. For example, when features are evenly distributed throughout the scene, this measure could be used to guide the vehicle to visually distinct areas of the environment for loop-closure.



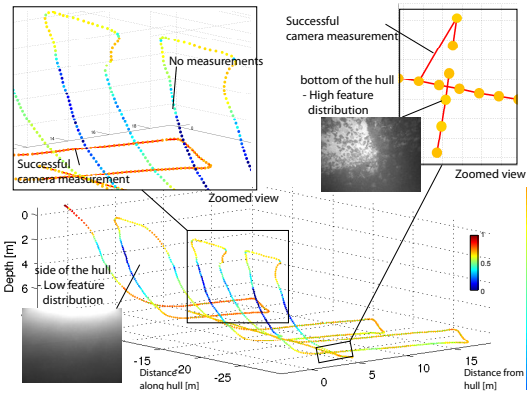
(a) Local saliency map on R/V Oceanus.



(b) Global saliency map on R/V Oceanus.



(c) Local / global saliency map on USCGC Venturous.



(d) Local saliency map on SS Curtiss.

Fig. 5. Three saliency maps overlaid on the pose-graph SLAM result. The colored circles represent the global / local saliency score as indicated by the color bar. Red lines represent successful pairwise camera measurements, which tend to coincide with high values of our local saliency measure. Global saliency, on the other hand, tends to capture the rarity of a feature.

IV. Experimental Results

In this section we illustrate the performance of the two different saliency metrics using real-world data collected from a series of underwater ship hull inspection surveys using the Hovering Autonomous Underwater Vehicle (HAUV) platform [38]. We surveyed three ship hulls: the Woods Hole Oceanographic Institution (WHOI) R/V Oceanus, the United States Coast Guard Cutter (USCGC) Venturous and the SS Curtiss (Fig. 6). The HAUV is equipped with a grayscale monocular camera that is actuated to maintain a nadir view to the hull. For each survey, we processed the data and built saliency maps for the underwater portion of the ship hull.

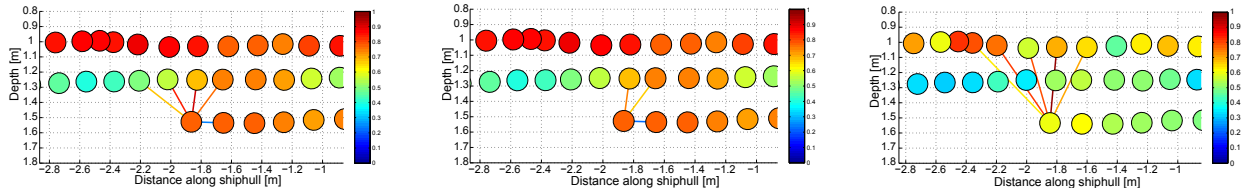
A. Local and Global Saliency Map

To verify the performance of the two saliency metrics (i.e., local and global), their respective normalized saliency maps have been overlaid atop our pose-graph visual SLAM results as depicted in Fig. 5. Red lines indicate successful pairwise image registrations, and the nodes in the graph have been color-coded by their saliency level.

(a) HAUV (b) Oceanus (c) Venturous (d) Curtiss

Fig. 6. Underwater hull inspection experiments conducted using the Bluefin Robotics HAUV shown in (a). Three different ship hulls were surveyed: R/V Oceanus (b), USCGC Venturous (c), and SS Curtiss (d).

Overlaying the local saliency map on the SLAM result shows the coincidence of successful camera registrations and areas with a high local saliency score. To have a successful pairwise camera measurement, both images need to be locally salient (i.e., texture rich). Note that successful measurements (red lines in Fig. 5) have been made only when both of the images have a high local saliency score. When either image lacks saliency, image registration fails (i.e., regions with missing edges in the graph). In terms of local saliency, images of the USCGC Venturous were abundant with features throughout the survey as evident from the top figure in Fig. 5(c), which shows evenly distributed



(a) IG. (b) IG with local saliency. (c) IG with global saliency.

Fig. 7. Saliency guided link proposal. By considering local saliency with geometric information gain, we can propose more effective candidate links. Global saliency incorporated link proposal provides unrealistic pairs, but shows regions where to find rare features in the environment.

camera registration. In contrast, the SS Curtiss shows a segregated local saliency map. The hull of the Curtiss was feature-less on the side, but covered by marine growth rich in texture at the bottom. Two zoomed views of Fig. 5(d) illustrate this location dependent saliency difference.

Thinking now about global saliency, we note that unlike local saliency the global saliency metric reacts to rare features. The global saliency map on the R/V Oceanus (Fig. 5(b)) does not necessarily correlate with successful camera measurements, but instead indicates the rarity of images. A similar result was found for the USCGC Venturous (Fig. 5(c)). Its hull was covered with barnacles in most of the regions (A in Fig. 5(c)), except for two locations where artificial targets (inert mines) were attached to the hull. High global saliency is reported at these two target positions (C and D in Fig. 5(c)) since they have a rare occurrence. Also, other visually uncommon scenes such as a weld line and a stray fish (B in Fig. 5(c)) scored high.

B. Link Proposal

Following §III-E, our saliency incorporated information gain metric was applied and tested using the R/V Oceanus data set. Fig. 7 shows three different cases of link proposal: information gain (IG) link proposal (Fig. 7(a)), IG with local saliency link proposal (Fig. 7(b)), and IG with local plus global saliency link proposal (Fig. 7(c)). The color of a depicted link indicates how informative the link is, while the color of a node represents how salient the scene is (i.e., local saliency score for Fig. 7(a) and 7(b), and global saliency score for Fig. 7(c)).

In the first case of IG-only link proposal, candidate links are computed using only geometry. Therefore, proposed links are geometrically symmetric in this case (Fig. 7(a)). However, when local saliency is considered, saliency incorporated information gain modifies the link proposal to select more feature-rich image pairs. Note that it prefers to process feature-rich image pairs that have less geometric information gain, rather than processing visually uninformative images with high geometric gain. In doing so, it proposes realistically plausible camera-derived candidate links.

The statistics in Table I reveal that a large number of non-plausible IG proposed links can be eliminated by considering local saliency. The results show that we can achieve almost the same number of successful cross-track camera-derived links while reducing the number of candidate proposals by ∼40–60%. Note that most of the removed candidates are

TABLE I

LINK PROPOSAL WITH LOCAL SALIENCY

thresh	N_s	N_c	ΔN_c	$\sigma_{\max}$	thresh	N_s	N_c	ΔN_c	$\sigma_{\max}$
0†	156	1048	0.0%	0.293	—	—	—	—	—
0.40	153	887	15.4%	0.293	0.60	127	590	43.7%	0.328
0.45	150	868	17.2%	0.294	0.65	115	379	63.8%	0.359
0.50	145	824	21.4%	0.295	0.70	86	224	78.6%	0.388
0.55	137	753	28.1%	0.302	0.75	54	90	91.4%	0.437

N_s / N_c = Number of successful / candidate cross-track camera links, respectively.
ΔN_c = Percentage of IG candidate links culled via local saliency measure (i.e., $S_{L,thres}$).
$\sigma_{\max}$ = Maximum node uncertainty in the resulting graph (i.e., $\max_i \sqrt{tr(\Sigma_i)}$) (expressed in meters).
†Note that a zero threshold corresponds to IG-only.

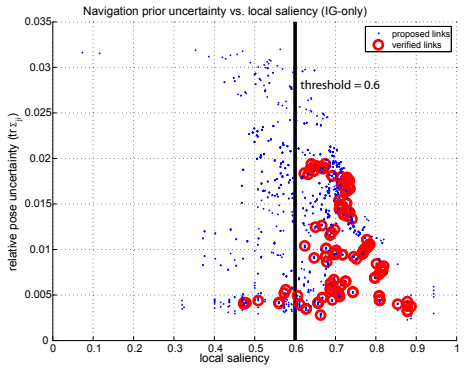


Fig. 8. Scatter plot of IG-only candidate image pairs versus relative-pose uncertainty and local saliency. Blue depicts all proposed links whereas red circles mark successfully registered pairs. Successfully registered image pairs tend to have higher local saliency scores.

from poses with low uncertainty, as depicted in Fig. 8. This is because we use our navigation prior as an epipolar search constraint to enhance the correspondence search. Under a strong motion prior, candidate pairs with low local saliency can thus be matched; however, when the motion prior is weak, the saliency score informs us as to which candidate pairs have the best chance of matching even under a weak motion prior.

Global saliency incorporated information gain, on the other hand, should be considered in a path planning sense rather than for link proposal. As depicted in Fig. 7(c), the resulting candidate links are unrealistic from an image overlap standpoint, but can be correctly interpreted as a recommended direction of travel to increase the chances of obtaining a visually unique camera measurement. We can use this metric to direct the robot to rare and distinguishable regions in the environment, which will be helpful for loop-closure data association. Coupling this saliency incorporated information gain into a path planning strategy is the focus

of our future work.

V. Conclusion

This paper reported on a method to improve pose-graph visual SLAM performance by proposing geometrically and visually informative loop-closure candidates. Two types of saliency metrics were proposed: local and global. The utility of these saliency metrics was tested with indoor and underwater images and shown to be general. Intelligent link hypothesis using combined information gain and visual saliency was presented and found to be effective in proposing registrable image pairs.

Acknowledgments

This work was supported through a grant from the Office of Naval Research (ONR) (Award #N00014-07-1-0791).

References

[1] V. Ila, J. M. Porta, and J. Andrade-Cetto, "Information-based compact pose SLAM," IEEE Trans. on Robotics, vol. 26, no. 1, pp. 78–93, Feb. 2010.

[2] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, Estimation with applications to tracking and navigation. New York: John Wiley & Sons, Inc., 2001.

[3] M. Bryson and S. Sukkarieh, "An information-theoretic approach to autonomous navigation and guidance of an uninhabited aerial vehicle in unknown environments," in Proceedings of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, Aug. 2005, pp. 3770–3775.

[4] C. E. Shannon, "A mathematical theory of communication," The Bell System Technical J., vol. 27, pp. 379–423, 623–656, July, October 1948.

[5] R. Sim and N. Roy, "Global a-optimal robot exploration in SLAM," in Proceedings of the IEEE Intl. Conf. on Robotics and Automation, Barcelona, Spain, 2005, pp. 661 – 666.

[6] T. Vidal-Calleja, A. Davison, J. Andrade-Cetto, and D. Murray, "Active control for single camera SLAM," in Proceedings of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, Orlando, FL, 2006, pp. 1930 –1936.

[7] D. Lowe, "Distinctive image features from scale-invariant keypoints," Intl. J. of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[8] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded-up robust features," in 9th European Conf. on Computer Vision, Graz, Austria, 2006, pp. 404 – 417.

[9] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in Proceedings of the IEEE Intl. Conf. on Computer Vision, vol. 2, Oct. 2003, pp. 1470–1477.

[10] L. Wu, S. Hoi, and N. Yu, "Semantics-preserving bag-of-words models and applications," IEEE Trans. on Image Processing, vol. 19, no. 7, pp. 1908 –1920, 7 2010.

[11] N. Lazic and P. Aarabi, "Importance of feature locations in bag-of-words image classification," in IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, vol. 1, 4 2007, pp. 641 – 644.

[12] M.-J. Hu, C.-H. Li, Y.-Y. Qu, and J.-X. Huang, "Foreground objects recognition in video based on bag-of-words model," in Chinese Conf. on Pattern Recognition, 11 2009, pp. 1 –5.

[13] D. Larlus, J. Verbeek, and F. Jurie, "Category level object segmentation by combining bag-of-words models with dirichlet processes and random fields," Intl. J. of Computer Vision, vol. 88, no. 2, pp. 238–253, 6 2010.

[14] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM - FAB-MAP 2.0," in Proceedings of Robotics: Science and Systems, Seattle, USA, June 2009.

[15] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," IEEE Trans. on Robotics, vol. 24, no. 5, pp. 1027–1037, oct. 2008.

[16] A. Kawewong, N. Tongprasit, S. Tangruamsub, and O. Hasegawa, "Online and Incremental Appearance-based SLAM in Highly Dynamic Environments," Intl. J. of Robotics Research, pp. 33–55, 2010.

[17] T. Kadir and M. Brady, "Saliency, scale and image description," Intl. J. of Computer Vision, vol. 45, no. 2, pp. 83–105, 2001.

[18] Y.-J. Lee and J.-B. Song, "Autonomous salient feature detection through salient cues in an hsv color space for visual indoor simultaneous localization and mapping," Advanced Robotics, vol. 24, no. 11, pp. 1595–1613, 2010.

[19] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," IEEE Trans. on Image Processing, vol. 11, pp. 467–476, 2002.

[20] M. Johnson-Roberson, "Large-scale multi-sensor 3d reconstructions and visualizations of unstructured underwater environments," Ph.D. dissertation, The University of Sydney, 2009.

[21] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in European Conf. on Computer Vision, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, vol. 3954, ch. 38, pp. 490–503.

[22] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in Intl. Conf. on Computer Vision (ICCV), 2009, pp. 2232–2239.

[23] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in Proceedings of the European Conf. on Computer Vision, 2004, pp. 1–22.

[24] R. Toldo, U. Castellani, and A. Fusiello, "A bag of words approach for 3d object categorization," in Proc. of the 4th Intl. Conf. on Computer Vision/Computer Graphics CollaborationTechniques. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 116–127.

[25] J. Sivic and A. Zisserman, "Video Google: Efficient visual search of videos," in Toward Category-Level Object Recognition, ser. LNCS, J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, Eds. Springer, 2006, vol. 4170, pp. 127–144.

[26] Ayoung Kim and R. M. Eustice, "Pose-graph visual SLAM with geometric model selection for autonomous underwater ship hull inspection," in Proceedings of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, St. Louis, MO, Oct. 2009, pp. 1559–1565.

[27] R. M. Eustice, H. Singh, and J. J. Leonard, "Exactly sparse delayed-state filters for view-based SLAM," IEEE Trans. on Robotics, vol. 22, no. 6, pp. 1100–1114, Dec. 2006.

[28] R. M. Eustice, H. Singh, J. J. Leonard, and M. R. Walter, "Visually mapping the RMS Titanic: conservative covariance estimates for SLAM information filters," Intl. J. of Robotics Research, vol. 25, no. 12, pp. 1223–1242, 2006.

[29] R. M. Eustice, O. Pizarro, and H. Singh, "Visually augmented navigation for autonomous underwater vehicles," IEEE J. of Oceanic Engineering, vol. 33, no. 2, pp. 103–122, Apr. 2008.

[30] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," Intl. J. of Robotics Research, vol. 27, no. 6, pp. 647–665, June 2008.

[31] C. D. Manning, P. Raghavan, and H. Schtze, Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press, 2008.

[32] G. Salton and C. S. Yang, "On the specification of term values in automatic indexing," J. of Documentation, vol. 29, pp. 351–372, 1973.

[33] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," J. of Documentation, vol. 28, pp. 11–21, 1972.

[34] C. Moulin, C. Barat, and C. Ducottet, "Fusion of tf.idf weighted bag of visual features for image classification," in Intl. Workshop on Content-Based Multimedia Indexing, 6 2010, pp. 1 –6.

[35] J. Knopp, J. Sivic, and T. Pajdla, "Avoding confusing features in place recognition," in Proc. of the European Conf. on Computer Vision, 2010, pp. 748–761.

[36] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for idf," J. of Documentation, vol. 60, pp. 503–520, 2004.

[37] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, 6 2009, pp. 1169 –1176.

[38] J. Vaganay, M. Elkins, D. Esposito, W. O'Halloran, F. Hover, and M. Kokko, "Ship hull inspection with the HAUV: US Navy and NATO demonstrations results," in OCEANS 2006, 9 2006, pp. 1 –6.