Real-Time Visual SLAM for Autonomous Underwater Hull Inspection using Visual Saliency

Ayoung Kim, *Student Member, IEEE,* and Ryan M. Eustice, *Senior Member, IEEE*

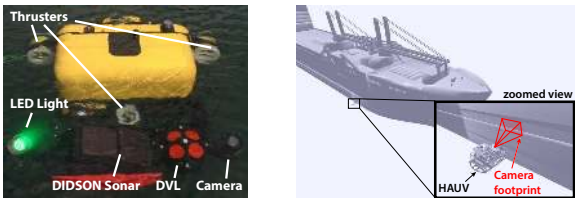
***Abstract*—This paper reports on a real-time monocular visual simultaneous localization and mapping (SLAM) algorithm and results for its application in the area of autonomous underwater ship hull inspection. The proposed algorithm overcomes some of the specific challenges associated with underwater visual SLAM, namely limited field of view imagery and feature-poor regions. It does so by exploiting our SLAM navigation prior within the image registration pipeline and by being selective about which imagery is considered informative in terms of our visual SLAM map. A novel online bag-of-words measure for intra- and inter-image saliency are introduced, and are shown to be useful for image key-frame selection, information-gain based link hypothesis, and novelty detection. Results from three real-world hull inspection experiments evaluate the overall approach—including one survey comprising a 3.4 hour / 2.7 km long trajectory.**

***Index Terms*—SLAM, computer vision, marine robotics, visual saliency, information gain.**

I. Introduction

MANY underwater structures such as dams, ship hulls, harbors, and pipelines need to be periodically inspected for assessment, maintenance, and security reasons. Among these, our interest is in autonomous underwater hull inspection, which seeks to map and inspect the below-water portion of a ship *in situ* while in port or at sea. Typical methods for port security and ship hull inspection require either deploying human divers [3], [4], using trained marine mammals [5], or piloting a remotely operated vehicle (ROV) [6]–[8]. Autonomous vehicles have the potential for better coverage efficiency, improved survey precision, and overall reduced need for human intervention, and as early as 1992 there was an identified need within the Naval community for developing such systems [9]. In recent times, effort in this area has resulted in the development of a number of automated hull inspection platforms [10]–[13].

Underwater navigation feedback in this context is typically performed using inertial measurement unit (IMU) or Doppler velocity log (DVL) derived odometry [12], [14], and/or acoustic beacon time-of-flight ranging [11], [15]. The main difficulties of these traditional navigation approaches are that they either suffer from unbounded drift (e.g., odometry), or they require external infrastructure that needs to be set up and calibrated (e.g., acoustic beacons). Both of these scenarios tend to vitiate the "turn-key" automation capability that is desirable in hull inspection.



(a) HAUV (b) Sensor FOV

Fig. 1. (a) The Bluefin Robotics Hovering Autonomous Underwater Vehicle (HAUV) used for hull inspection in this project. (b) Depiction of the HAUV's size in comparison to a typical large ship, and its camera's field of view (FOV) when projected onto the hull at a typical standoff distance of one meter.

For the past couple of decades now, a significant research effort within the mobile robotics community has been to develop a simultaneous localization and mapping (SLAM) capability. The goal of SLAM algorithms is to bound the navigational error to the size of the environment by using perceptually derived spatial information—a key prerequisite for truly autonomous navigation. For a historical survey of advancements in this field the reader is referred to [16], [17]. It is within this paradigm that nontraditional approaches to hull-relative navigation have generally sought to alleviate traditional navigation issues.

Negahdaripour and Firoozfam [8] developed underwater stereo-vision as a means of navigating an ROV near a hull; they used mosaic-based registration methods and showed preliminary results for controlled pool and dock trials. Ridao et al. [18] reported on the closely related task of automated dam inspection using an autonomous underwater vehicle; their solution uses ultra-short-baseline (USBL) and DVL-based navigation *in situ* during the mapping phase, followed by an offline image bundle adjustment phase to produce a globally-optimal photomosaic and vehicle trajectory. Walter, Hover and Leonard [19] reported the use of an imaging sonar for feature-based SLAM navigation on a barge and showed results for offline processing using manually-established feature correspondence. More recently, this work was significantly extended by Johannsson et al. [20] to work in real-time and to perform automatic registration of sonar hull imagery.

In parallel to these efforts we have, since 2007, collaborated with the authors of [20] and with Bluefin Robotics on an Office of Naval Research sponsored project for autonomous hull inspection (Fig. 1). Our part has been to develop a

Manuscript received February 16, 2012; revised October 31, 2012; accepted November 23, 2012. This work was supported by the Office of Naval Research under grants N00014-07-1-0791 and N00014-12-1-0092, monitored by Dr. T. Swean, M. Zalesak, V. Steward, and T. Kick. Portions of this work were presented in part at the 2009 and 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems [1], [2].

A. Kim and R. Eustice are with the Department of Naval Architecture & Marine Engineering, University of Michigan, Ann Arbor, MI, 48109, USA (e-mail: ayoungk@umich.edu; eustice@umich.edu).

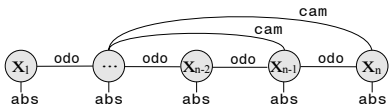


Fig. 2. Depiction of the pose-graph SLAM constraint graph. Odometry constraints (`odo`) are sequential whereas camera constraints (`cam`) can be either sequential or non-sequential. For each node, measurements of roll/pitch and depth are added as absolute constraints (`abs`).

real-time visual SLAM capability for hull-relative navigation in the open areas of the hull. Through collaboration with our project partners, we have developed an integrated real-time SLAM system for hull-relative navigation and control that has been recently demonstrated on the Bluefin Robotics HAUV (pronounced "H-A-U-V"). Specifications of the current generation vehicle design are documented in [21], and an overview of our integrated work in perception, planning and control is presented in [22].

In this paper, we report on the specific details of our real-time monocular visual SLAM solution for autonomous hull inspection. The contributions of this work are fourfold: *i)* the dissemination of a principled and field proven approach for exploiting available navigational and geometrical priors in the image registration pipeline to overcome the difficulties of underwater imaging, *ii)* the introduction of a novel and quantitative bag-of-words visual saliency metric that can be used for identifying visually informative key-frames to include in our SLAM map, *iii)* the development of a visually robust link hypothesis algorithm that takes into account geometric information gain as well as visual plausibility, and *iv)* the demonstration of a complete end-to-end real-time visual SLAM implementation on the HAUV with field results from three real-world deployments, which experimentally evaluates the overall approach.

II. System Overview

A. Hovering Autonomous Underwater Vehicle

For the autonomous hull inspection project, we use the Bluefin Robotics HAUV (Fig. 1) [21]. This vehicle was developed for explosive ordnance disposal (EOD) inspection, and is currently in production for the U.S. Navy [23]. For navigation the standard vehicle is equipped with a hull-looking 1200 kHz RDI Doppler velocity log (DVL), Honeywell HG1700 IMU, and Keller pressure sensor for depth, while for inspection the vehicle is equipped with a 1.8 MHz DIDSON imaging sonar [22]. Additionally, in collaboration with Bluefin, we have integrated a 520 nm (i.e., green) LED light source for optical imaging and a fixed-focus, monochrome, Prosilica GC1380 12-bit digital-still camera.

B. Pose-Graph Visual SLAM using iSAM

In our work, we estimate the vehicle's full six degree of freedom (DOF) pose, $\mathbf{x} = [x, y, z, \phi, \theta, \psi]^\top$, where the pose (position and Euler attitude) is defined in a local-level Cartesian frame referenced with respect to the hull of the ship. We use a pose-graph SLAM framework for state representation where the state vector, X, is comprised of a collection of historical poses. Each node in the graph, $\mathbf{x}_i$, corresponds to a camera event that we wish to include in our view-based map. Fig. 2 depicts the general topology of our resulting pose-graph, which consists of nodes linked by either odometry or camera constraints. For each node, measurements of gravity-based roll/pitch and pressure depth are added as absolute constraints, whereas absolute heading measurements are unavailable in our sensor configuration (note that magnetically-derived compass heading is useless near a ferrous hull). There exist many inference algorithms that solve the pose-graph SLAM problem [24]–[31], and in this paper we employ the open-source incremental smoothing and mapping (iSAM) algorithm due to its efficiency for real-time implementation and covariance recovery [31]–[33].

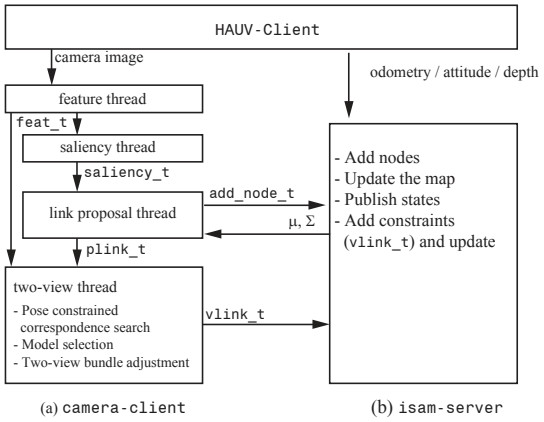


Fig. 3. Real-time SLAM publish/subscribe server/client software architecture using iSAM. The shared estimation server, `isam-server`, listens for add node message requests, `add_node_t`, from the `camera-client`. Extracted features, `feat_t`, are published by the feature thread. The saliency thread subscribes to these `feat_t` messages and computes a visual saliency score, which gets published as a `saliency_t` message. This score is used in the link proposal thread to determine node addition as well as link proposal events. Proposed link candidates are published as `plink_t` events, which the two-view thread then attempts to register. If successful, the camera thread then publishes the 5-DOF camera constraint as a verified link message, `vlink_t`, which then gets added to the pose-graph by `isam-server`.

We assume standard Gaussian process and observation models with independent control and measurement noise. The process model, $\mathbf{x}_i = f(\mathbf{x}_{i-1}, \mathbf{u}_i) + \mathbf{v}_i$, is a stochastic state transition model linking two sequential poses via control input $\mathbf{u}_i$ with noise $\mathbf{v}_i \sim \mathcal{N}(0, \Sigma_i)$. The observation model, $\mathbf{z}^k_{i,j} = h(\mathbf{x}_i, \mathbf{x}_j) + \mathbf{w}_k$, is a stochastic measurement model between two nodes i and j with measurement index k and noise $\mathbf{w}_k \sim \mathcal{N}(0, \Lambda_k)$.

C. Camera Constraints

In our SLAM framework, we model pairwise monocular image registration as providing a 5-DOF, relative-pose, modulo-scale constraint between nodes i and j. Here, the 5-DOF camera measurement is modeled as an observation of the baseline direction of motion azimuth, α_{ij}, and elevation angle, β_{ij}, and the relative Euler angles, $\phi_{ij}, \theta_{ij}, \psi_{ij}$, between the two poses [34],

$$h_{\mathsf{5dof}}(\mathbf{x}_i, \mathbf{x}_j) = \left[\alpha_{ij}, \beta_{ij}, \phi_{ij}, \theta_{ij}, \psi_{ij}\right]^\top. \qquad (1)$$

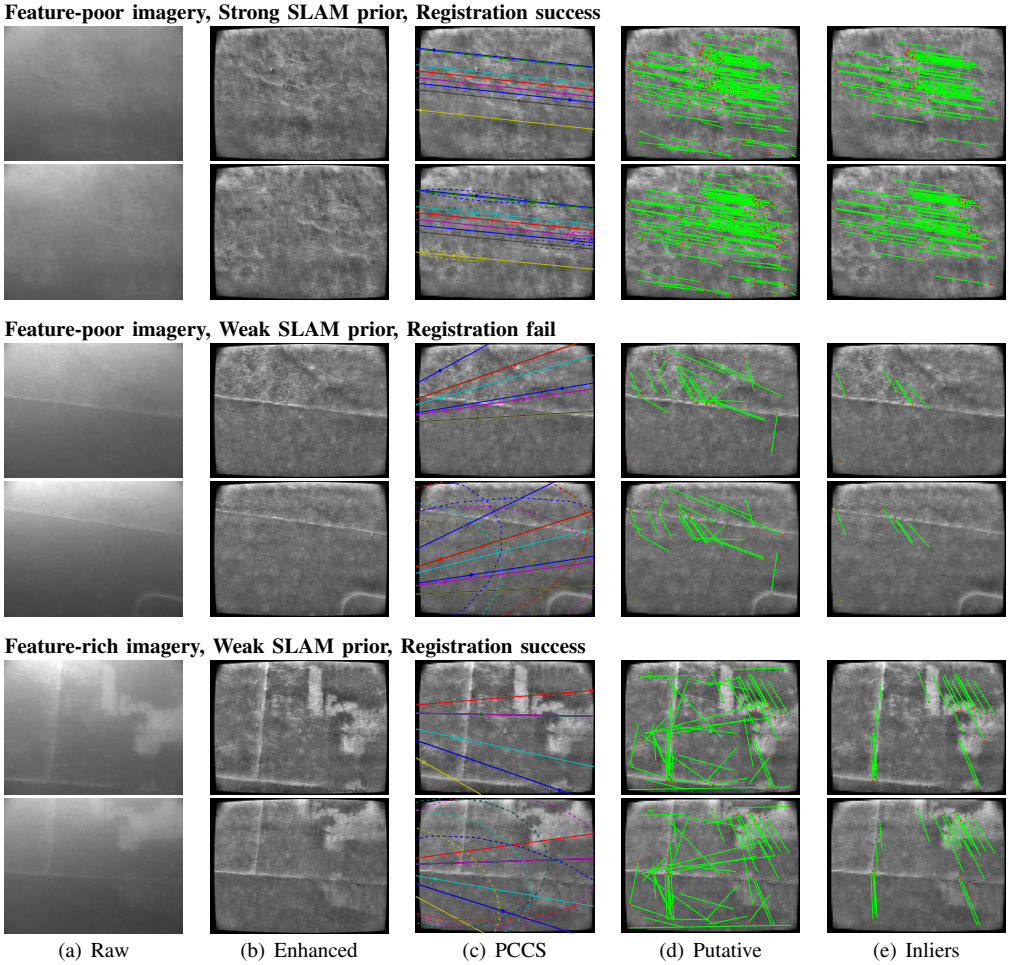


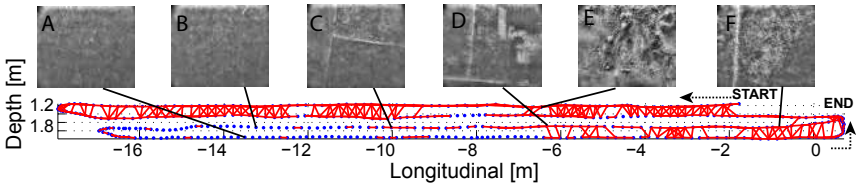
Fig. 4. Depiction of the `camera-client` underwater image registration process for typical hull imagery. (a) Raw images are (b) first radially undistorted and histogram equalized before extracting features. (c) A pose-constrained correspondence search (PCCS) using our SLAM pose prior is then applied to guide putative matching. Lines depict sample epipolar geometry induced from the SLAM pose prior with navigation uncertainty projected as 99.9% confidence ellipsoids in pixel space. (d) Putative correspondences are established within the PCCS search constraint using SIFT descriptors with a threshold on the ratio to the second best matching to obtain putative matches. (e) Inlier correspondences and motion model are then found from a RANSAC geometric model selection framework and optimized in a two-view bundle adjustment to determine the 5-DOF camera relative-pose constraint.
For the top row of imagery, because the PCCS search constraint is strong, correct correspondences are established despite the fact that the imagery is feature-poor. For the middle and bottom rows of imagery, we see two different cases—when the PCCS SLAM prior is weak and the imagery is feature-poor (middle row), image registration fails due to a dearth of correct putative correspondences. On the other hand, when the PCCS SLAM prior is weak but the imagery is feature-rich (bottom row), image registration succeeds because enough correct putative correspondences are established using visual similarity measures alone. Observation of this effect motivates the development of our novel image saliency metrics introduced in Section III.

These camera constraints are generated from a real-time visual SLAM perception engine, namely the `camera-client` process of Fig. 3. Fig. 4 depicts sample results from the `camera-client` processing pipeline, which consists of:

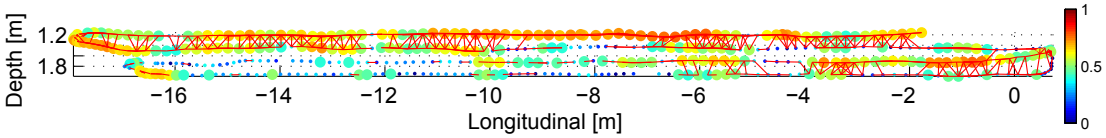
1) Images are first radially undistorted and enhanced using contrast-limited adaptive histogram specification (CLAHS) [35].
2) For feature extraction and description we use a combination of scale invariant feature transform (SIFT) [36] and speeded up robust features (SURF) [37]—real-time performance is enabled using a graphics processing unit (GPU) based implementation [38].
3) Correspondences are established using a pose-constrained correspondence search (PCCS) [34] and random sample consensus (RANSAC) geometric model selection framework [1].
4) Inliers are then fed into a two-view bundle adjustment to yield a 5-DOF bearing-only camera measurement (1), and a first-order estimate of its covariance [39].
5) This measurement is then added as a constraint to iSAM.

Three cases are interesting to note in Fig. 4. In cases where we have a strong prior on the relative vehicle motion (top row), for example due to sequential imagery with good odometry or when the SLAM prior is tight, then the PCCS search region provides a tight bound for putative matching and we can often match what would be otherwise feature-poor imagery. On the other hand, when we have a weak pose prior (middle row), for example due to poor odometry or when closing large loops, then the PCCS search constraint will be uninformative and registration will likely fail to find enough matches based upon visual similarity. However, if the hull imagery is sufficiently feature-rich (bottom row), then images may be matched even under a poor PCCS prior using purely appearance-based means. This indicates that image saliency

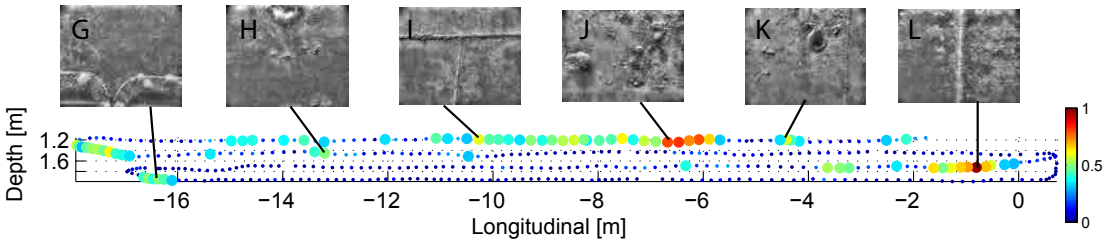
(a) *R/V Oceanus*



(b) SLAM pose-graph result for *R/V Oceanus*



(c) Local saliency map (S_L) on *R/V Oceanus*



(d) Global saliency map (S_G) on *R/V Oceanus*

Fig. 5. Motivation for the development of our local and global saliency metrics. Depicted are the hull inspection SLAM results for a survey of the port-side hull of the *R/V Oceanus*. (a) Picture of the *R/V Oceanus*' stern with the HAUV in view. (b) SLAM trajectory of the HAUV with successful cross-track camera registrations depicted as red edges. The histogram equalized images shown above are indicative of the type of imagery within that region of the hull. Qualitatively, note that the density of cross-track links is spatially correlated with what could be described as feature-rich imagery. (c) Our normalized local saliency measure, S_L, which spans from 0 to 1, is overlaid on top of the SLAM graph and correlates well with camera link density. Note that successful camera measurements typically correspond to nodes with a local saliency score of 0.4 or greater. (d) Our normalized global saliency measure, S_G, which also spans from 0 to 1, is overlaid on top of the SLAM trajectory and indicates image rarity. Global saliency can be used to identify visually rare (i.e., anomalous) scenes with respect to the rest of the hull. In both (c) and (d), for easier visualization, we have enlarged nodes with saliencies greater than 0.4.

plays a strong role in determining successful registration and could be exploited if quantified.

D. Software Architecture

Our real-time SLAM implementation is based on a publish/subscribe software architecture using the open-source Lightweight Communications and Marshalling (LCM) library [40] for inter-process communication. We run iSAM as a shared server process and each sensor client independently publishes measurement constraints to add to the graph; Fig. 3 depicts an architectural block-diagram. The server process subscribes to messages from the HAUV vehicle client to add DVL odometry constraints, absolute roll/pitch attitude measurements (from the IMU), and pressure depth observations.

Five DOF camera constraints are published to the server from the camera client process. The camera process is multi-threaded and organized into four main modules: a feature extraction thread, an image saliency thread, a link proposal thread, and a two-view image registration thread. The feature thread extracts robust features to be used for correspondence detection. The saliency thread then uses these extracted features to create a bag-of-words representation for the image and computes a visual saliency score. The link[1] proposal thread uses the visual saliency metric along with a calculation of geometric information gain to (*i*) add only salient nodes to the graph and (*ii*) to propose visually informative candidates for registration. The extracted features and proposed links are then fed to the two-view thread for attempted registration.

[1]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.

III. Visual Saliency

In our hull inspection scenario, camera-derived measurements are typically not uniformly available within the environment. Fig. 5 depicts a representative underwater visual SLAM result obtained on a clean hull (i.e., a hull with little or no bio-fouling). Here, successful camera registrations (i.e., red links) occur when feature-rich distributions are prevalent—in visually feature-poor regions, the camera produces few, if any, constraints. Thus, the distribution of visual features on the hull dominates the spatial availability of our camera-derived constraints, and hence, the overall precision of our SLAM navigation result. This indicates that visual saliency strongly influences the likelihood of making a successful pairwise camera measurement. When spatially overlapping image pairs fail to contain any locally distinctive textures or features—image registration fails. Hence, having a quantitative ability to evaluate the registration utility of image key-frames would greatly aid underwater visual SLAM. Fig. 5(c) and (d) depict sample results from our novel measures of image saliency, which are the subject of this section.

A. Overview of Our Approach

To tackle this problem, we focus on two different measures of saliency: local saliency (i.e., intra-image) and global saliency (i.e., inter-image). Both are computed using a bag-of-words (BoW) model for image representation. Registrability refers to the intrinsic feature richness of an image. The lack of image texture, as in the case of mapping an underwater environment with feature-poor regions (e.g., images A and B in Fig. 5(b)), prevents image registration from being able to measure the relative-pose constraint. However, texture is not the only factor that defines saliency—an easy counterexample is an image of a checkerboard pattern or a brick wall. Images of these type of scenes have high texture, but likely will fail registration due to spatial aliasing of common features. Thus, we develop local and global saliency as two different measures of image registrability in this section.

A brief illustration of the overall process is depicted in Fig. 6. We generate a coarse vocabulary online by projecting 128-dimension SURF descriptors to words using a BoW image model. Once mapped to a bag-of-words representation, we examine the intra-image histogram of word occurrence for the local saliency measure, and score the saliency level by evaluating its entropy. For global saliency, the inter-image frequency of word occurrence throughout all previously seen images is examined. This statistic is used to compute the global saliency score by measuring the so-called inverse document frequency.

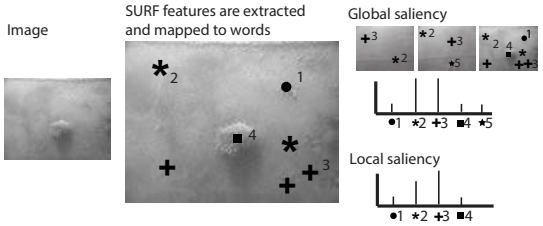


Fig. 6. Depiction of local and global saliency computation. Given an image stream, SURF descriptors are extracted and are used to compute local and global BoW statistics. Entropy from the local histogram (bottom right) detects intra-image feature richness, while inverse document frequency measures inter-image rarity (top right). Unlike local saliency, which is computed only from the current image, global saliency is computed by updating idf over a series of images.

B. Review on Saliency and Bag-of-Words

The term "saliency" refers to a measure of how distinctive an image is, and is related to seminal works by [41] and [42]. The authors of [43] extended [42]'s entropy approach to color images using the hue saturation value (HSV) color-space representation for detecting image features. Similarly, the author of [44] combined HSV channel entropy with a Gabor filter for texture entropy to compute a combined saliency score for color images. This approach was shown to produce usable saliency maps derived from down-looking underwater seafloor imagery; however, its application is limited to color imagery.

Alternatively to the above channel-based methods, several BoW saliency representations have recently been explored [45]–[48]. Originally developed for text-based applications, the general bag-of-words approach was first adapted and expanded to images by [49], [50], and [47], allowing for aggregate content assessment and enabling faster search. This approach has been successfully applied in diverse applications such as image annotation [51], image classification [52], object recognition [53], [54] and also appearance-based SLAM [55]–[59]. In connection to saliency, [47] explored the use of a BoW image model to selectively extract only "salient" words from an image and referred to them as a bag-of-keypoints. In [48], 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.

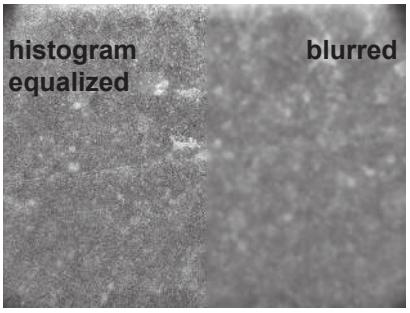
C. BoW Vocabulary Generation

Before defining our BoW saliency metric, we first need to outline how we construct our vocabulary. Offline methods for vocabulary generation typically use a clustering algorithm on a representative training dataset. An example method using this type of offline approach is the Fast Appearance-Based Mapping (FAB-MAP) algorithm, which has shown remarkable place recognition results using a pre-trained vocabulary [55], [56]. Other studies have focused on online methods, which incrementally build the vocabulary during the data collection phase [57]–[60]. Position Invariant Robust Feature (PIRF) based navigation [58] used this type of online approach, using only consistent SIFT descriptors to incrementally build the vocabulary, and showed comparable performance to other state-of-the-art appearance-based SLAM methods. In [59], in order to achieve fast and reliable online loop-closure detection, the authors used locality sensitive hashing to build the vocabulary *in situ*. Also, incremental online clustering schemes have been used by [60] to update the vocabulary clusters incrementally.

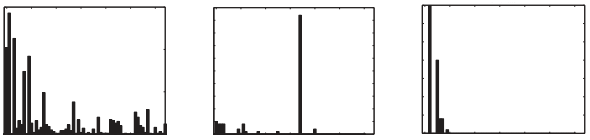
One advantage to offline methods is that an optimal distribution of vocabulary words (clusters) in descriptor space can be guaranteed; however, one disadvantage is that the learned vocabulary can fail to represent words collected from totally different datasets [58]. Online construction methods provide flexibility to adapt the vocabulary to incoming data, though equidistant words (clusters) are no longer guaranteed.

Two guidelines underpin our vocabulary building procedure: *(i)* we do not want to assume any prior appearance knowledge of the underwater inspection environment, and *(ii)* the vocabulary must be visually representative. With this in mind, we have decided to pursue an online construction approach that initially starts from an empty vocabulary set, similar to the algorithms in [57], [58]. SURF features are extracted from the incoming image and are matched to existing words in the vocabulary based on the Euclidean inner product (SURF descriptors are unit vectors). Whenever the direction cosine is larger than a threshold (0.4 in our experiments), we augment our vocabulary to contain the new word.

In terms of why we chose to use SURF features in our vocabulary construction, we evaluated the usage of both 128-dimension SIFT and 128-dimension SURF descriptors and found that SURF features tend to perform better for our saliency calculation. The SIFT descriptor is built by calculating the gradient orientation histogram, whereas the SURF descriptor is built from a set of Haar wavelet responses. Due to the noise sensitivity of the gradient orientation calculation,



(a) CLAHS / blurred



(b) CLAHS (c) Blurred (d) Scaled

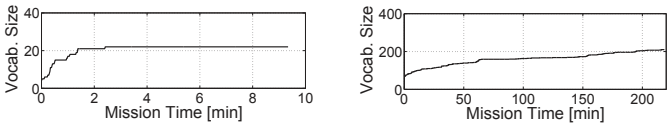
Fig. 7. Depiction of the effect of pre-blurring and scale-forced SURF detection for underwater image saliency. Image (a) shows the contrast-limited adaptive histogram specification (CLAHS) image on the left half and its blurred version on the right half. The BoW histogram showing intra-image word occurrence and its normalized entropy score (i.e., local saliency, S_L) are shown for the (b) CLAHS image (S_L=0.76), (c) the blurred image (S_L=0.35), and (d) the scale-forced SURF detection (S_L=0.48). Note that (c) and (d) have comparable entropy.

we found that SIFT's descriptor tends to assign two similar texture patches as two distinct words, whereas SURF's wavelet descriptor tends to assign them to the same type of word. (This is similar to what [44] noted when comparing a Gabor filter for texture detection versus gradient-based methods.)

An additional point worth noting is that we pre-blur imagery before running SURF. This is done to gently force it to return larger scale features. As shown in Fig. 7, we conducted a test to see the effect of this pre-blurring on underwater imagery. The depicted histogram-equalized sample image is "noisy" due to its accentuation of particulates in the water column and the effect of back-scattering. Processing the image at full scale makes the SURF descriptor sensitive to this high-frequency noise and, thus, its descriptors distinctive to each other. While this distinctiveness can be beneficial for putative correspondence matching, it is detrimental in vocabulary generation for the purpose of saliency detection. When the image contains particles and noise as in the sample image, these distinctive feature descriptors get mapped to different words, which artificially increases the entropy in our BoW histogram (Fig. 7(b)). However, this undesirable effect can be reduced by either pre-blurring the image (Fig. 7(c)), or (equivalently) by forcing SURF to return larger scale features (Fig. 7(d)). In practice, we found it easier to use the pre-blurring approach so that we could employ commonly available SURF libraries without modification.[2]

Typical BoW vocabulary sizes using our approach are relatively small—in our experience less than a couple of hundred words. This is in contrast to visual place recognition techniques, which typically have vocabulary sizes in the 4k to

[2]We use OpenCV's SURF implementation [61], which does not support direct scale-space thresholding.



(a) A small vessel (*R/V Oceanus*) (b) A large vessel (*SS Curtiss*)

Fig. 8. Online vocabulary size over the course of a hull inspection mission. The vocabulary size is plotted for two different vessels versus elapsed mission time in minutes. Because of the pre-blurring and coarse clustering, the resulting vocabulary size is small: 22 for the *R/V Oceanus* (a), and 210 for the *SS Curtiss* (b).

11k range or more [50], [55]–[58]. We note that the task of place recognition requires finer grain visual distinction than saliency detection does because vocabulary words are being used to uniquely index similar appearance imagery, whereas the goal of saliency detection is only to assess the visual variety of the scene. The pre-blurring and coarse clustering of our approach lead to small vocabulary sizes whose rate of growth plateaus in time as the vehicle collects enough visual variety to describe the inspection environment. Fig. 8 depicts the vocabulary sizes for two of the hull inspection missions reported in this paper.

D. Local Saliency

One of the original uses of BoW is for texture recognition [62], [63]. In these studies, an element of texture, a texton, can be expressed in terms of visual words using a BoW representation. These previous works mainly focused on recognition of texture using a texton representation, whereas the local saliency we develop here examines the diversity of the textures to assess image content richness. We define local saliency as an intra-image measure of feature diversity. We assess the diversity of words occurring within image I_i by examining the entropy of its BoW histogram:

$$H_i = -\sum_{k=1}^{W(t)} p(w_k) \log_2 p(w_k). \tag{2}$$

Here, $p(w)$ is the empirical BoW distribution within the image computed over the set of vocabulary words, $\mathcal{W}(t) = \{w_k\}_{k=1}^{W(t)}$, where $W(t)$ is the size of the vocabulary, which grows with time since we build the vocabulary online. We normalize the entropy measure with respect to the vocabulary size by taking the ratio of H_i to the maximum possible entropy to yield a normalized entropy measure, $S_{L_i} \in [0, 1]$, which we call local saliency:[3]

$$S_{L_i} = \frac{H_i}{\log_2 W(t)}. \tag{3}$$

This entropy-derived measure captures the diversity of words (descriptors) appearing within an image.

Fig. 9 shows sample results for color and grayscale underwater hull imagery. For comparison, following [44], we also compute the hue channel histogram as an alternative measure of saliency. The results show that our normalized BoW entropy

[3]The maximum entropy, $\log_2 W(t)$, corresponds to a uniform distribution over a vocabulary of $W(t)$ words.

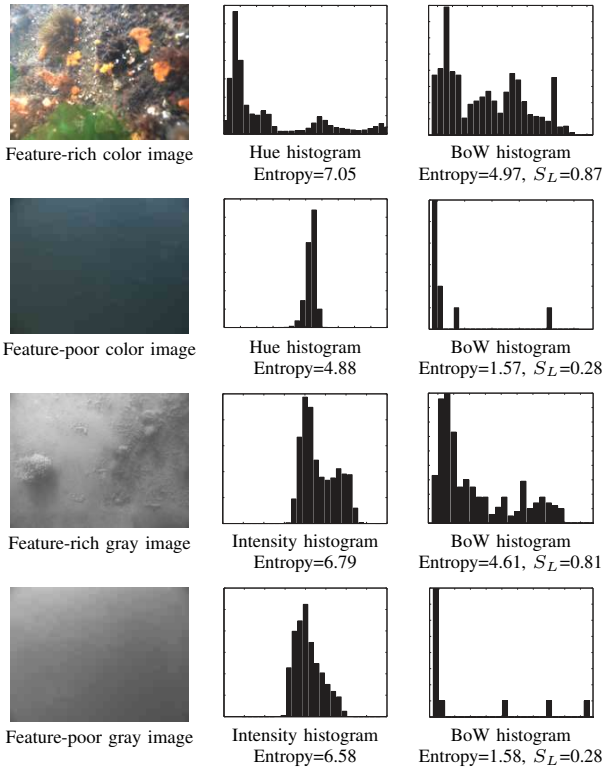


Fig. 9. Local saliency example for color and grayscale ship hull imagery of varying levels of feature content. In each result, the leftmost plot depicts the source image, the middle plot depicts the image intensity histogram (hue channel for color images and grayscale for monochrome images), and the rightmost plot depicts the bag-of-words histogram. For the color images, note that the hue channel histogram and the BoW histogram are both able to distinguish the feature richness of the scene. However, for the grayscale imagery, note that the image intensity histogram fails to detect feature richness, whereas the BoW histogram still works well.

score yields comparable results to [44] in terms of discriminating image saliency for color images, but moreover, our measure works equally well for grayscale imagery too (where no hue channel is available).

As a further example, Fig. 5(c) depicts the result of applying our local saliency score to the *R/V Oceanus* dataset. Note how our local saliency score shows good (predictive) agreement where the SLAM pairwise image registration engine was actually able to add cross-track camera constraints.

E. Global Saliency

We define global saliency as an inter-image measure of the uniqueness or rarity of features occurring within an image. The purpose of this measure is to identify unique regions of the hull that could be useful for guiding where the robot should revisit for attempting large scale loop-closure. In this scenario our SLAM prior will typically be weak and we will, therefore, have to rely upon visual appearance information only for successful pairwise image registration. Image D in Fig. 5 (same image as Fig. 4 bottom row) depicts such a case.

To tackle this problem, we were motivated by a metric called inverse document frequency (idf), which is a classic and widely used metric in information retrieval [64]–[66], and has a higher value for words seen less frequently throughout a history. In other words, we expect high idf for words (descriptors) that are rare in the dataset. In computer vision, Jegou et al. [67] used a variation of idf to detect "burstiness" of a scene, noting idf's ability to capture word frequency. Similar use is found in [68], where the authors used idf as a weighting factor in the definition of their min-Hash similarity metric.

In this paper, we use a sum of idf within an image, I_i, to score its inter-image rarity:

$$\mathcal{R}_i(t) = \sum_{k \in \mathcal{W}_i} \log_2 \frac{N(t)}{n_{w_k}(t)}. \tag{4}$$

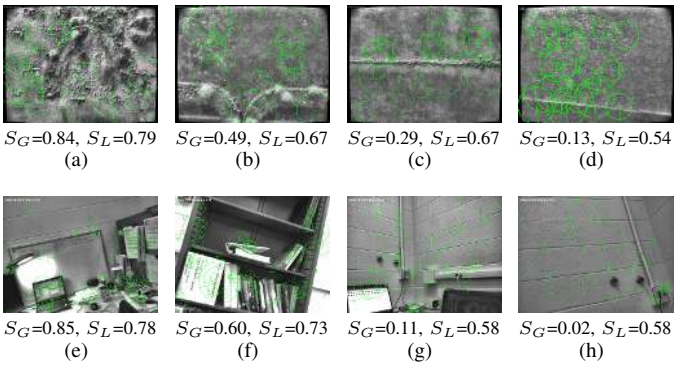
Here, $\mathcal{W}_i \subseteq \mathcal{W}(t)$ represents the subset of vocabulary words occurring within image I_i, $n_{w_k}(t)$ is the number of images in the vocabulary database containing word w_k, and $N(t)$ is the total number of images comprising the vocabulary database. The sum of idf in (4) makes the implicit independence assumption that words occur independently, similar to other BoW algorithms such as [50], [57], [58]. In cases where word occurrence is correlated (i.e., frequently occur together in the same images), this measure will overestimate the saliency of their combination, as denoted by [69]. In our application, we examined the co-occurrence of words in our vocabularies and found no significant correlation to exist between the appearance of words. To obtain independent sample statistics used in our idf database calculation, only spatially distinct images (i.e., non-overlapping) are used to update $n_{w_k}(t)$ and $N(t)$.

Since even a common word would be considered "rare" in (4) the first time it is observed (i.e., $n_{w_k} = 1$ on first occurrence in the database), $\mathcal{R}_i(t)$ needs to be updated through time. We use an inverted index update scheme combined with periodic batch updates to maintain $\mathcal{R}(t)$ for all images in the graph. The inverted index scheme [70] uses sparse bookkeeping for fast updates on the subset of $\mathcal{R}(t)$ who are impacted when changes in the statistics of $n_{w_k}(t)$ occur, and periodic batch updates that revise $\mathcal{R}(t)$ for all nodes in the graph when changes in the number of documents, $N(t)$, occur. At worst case this batch update is linear in complexity with the number of image nodes. Lastly, as was the case with our local saliency measure, we normalize the rarity measure for image I_i to have a normalized global saliency score $S_{G_i} \in [0, 1]$:

$$S_{G_i}(t) = \frac{\mathcal{R}_i(t)}{\mathcal{R}_{\mathsf{max}}}, \tag{5}$$

where the normalizer, $\mathcal{R}_{\mathsf{max}}$, is the maximum summed idf score encountered thus far.

Fig. 10 shows an example of applying global saliency to categorize sample underwater and indoor office imagery. As can be seen, the global saliency score, S_G, fires on the visual rarity of vocabulary words occurring within the image, whereas the local saliency score, S_L, fires on vocabulary diversity only. For example, the two rightmost figure columns (i.e., (c),(d) and (g),(h)) show that global saliency can be low even for locally salient imagery. This is because several of the vocabulary words (e.g., weld lines, bricks) occur frequently throughout the environment—lowering their overall idf score. As a further example, Fig. 5(d) depicts the result of applying our global saliency score to the *R/V Oceanus* dataset. Note



S_G=0.84, S_L=0.79 (a) S_G=0.49, S_L=0.67 (b) S_G=0.29, S_L=0.67 (c) S_G=0.13, S_L=0.54 (d)

S_G=0.85, S_L=0.78 (e) S_G=0.60, S_L=0.73 (f) S_G=0.11, S_L=0.58 (g) S_G=0.02, S_L=0.58 (h)

Fig. 10. Global saliency example for underwater (a)–(d) and indoor (e)–(h) images. Extracted features are marked with green circles. The global saliency score (S_G) and local saliency score (S_L) are provided below each image. In both datasets, images are arranged from left to right in order of decreasing global saliency. Note that the global saliency score can be low even for texture rich scenes (e.g., (c),(d) and (g),(h)), indicating that the vocabulary words appearing in those images are common in the environment and, therefore, not visually distinctive.

how the global saliency score identifies visually distinctive (i.e., rare) regions on the hull.

In separate work, we have reported the use of global saliency's rarity detection within an active SLAM paradigm for guiding the robot toward distinctive regions on the hull for attempting loop-closure [71], [72]—this represents one possible use of global saliency. Another possible application is anomaly detection on the hull, as supported later in the results of Fig. 19, which shows automatically identified foreign objects present on the hull. We present global saliency's formulation and evaluative results in conjunction with local saliency because it shares all of the same BoW vocabulary machinery and the two are fundamentally interrelated measures. Algorithm 1 provides a pseudo code description for the online vocabulary construction, and local and global saliency calculations.

Require: image I_i
Require: BoW vocabulary $\mathcal{W}(t)$ {$\emptyset$ on first use}
Require: idf statistics $N(t)$, $n_w(t)$
  Preblur and extract SURF features from I_i:
  $\mathcal{F}_i \leftarrow [f_1, f_2, \cdots, f_{n_f}]$

  {compute intra-image BoW statistics}
  initialize BoW histogram: $\mathcal{H}_i \leftarrow \emptyset$
  for each feature $f_j \in \mathcal{F}_i$ **do**
    find best vocabulary match $w_k \in \mathcal{W}(t)$
    if projection $f_j \cdot w_k >$ threshold **then** {augment vocab.}
      $\mathcal{W}(t) \leftarrow [\mathcal{W}(t), f_j]$, $w_k \leftarrow f_j$, $n_{w_k}(t) \leftarrow 1$
    end if
    increment histogram: $\mathcal{H}_i(w_k) \leftarrow \mathcal{H}_i(w_k) + 1$
  end for

  {update inter-image idf statistics}
  if I_i does not overlap with images already in $N(t)$ **then**
    increment the document database: $N(t) \leftarrow N(t) + 1$
    for each $w_k \in \mathcal{W}(t)$ **and** $\mathcal{H}_i(w_k) > 0$ **do**
      increment word occurrence: $n_{w_k}(t) \leftarrow n_{w_k}(t) + 1$
    end for
  end if

  {local saliency calculation}
  Compute image I_i BoW distribution: $p_i(w) \leftarrow \mathcal{H}_i / n_f$
  Compute image I_i BoW entropy: $H_i \leftarrow$ Eqn. (2)
  Compute image I_i local saliency: $S_{L_i} \leftarrow$ Eqn. (3)
  if $\mathcal{W}(t)$ was updated **then** {vocab. was augmented}
    Update S_L for all previous images
  end if

  {global saliency calculation}
  Compute image I_i rarity: $\mathcal{R}_i(t) \leftarrow$ Eqn. (4)
  Compute image I_i global saliency: $S_{G_i} \leftarrow$ Eqn. (5)
  if $N(t)$ or $n_w(t)$ were updated **then** {idf statistics changed}
    Update $\mathcal{R}(t)$ for all affected images
    Update maximum rarity $\mathcal{R}_{\mathsf{max}}$
    Update S_G for affected images
  end if

Algorithm 1: Online vocabulary and saliency calculation.

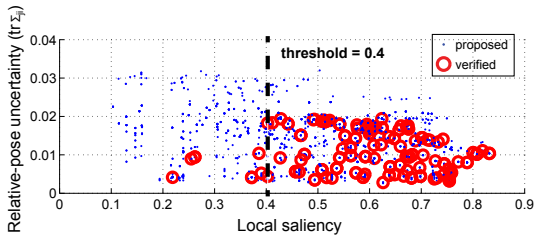
IV. Saliency-Informed Visual SLAM

One of the most important and difficult problems in SLAM is determining loop-closure events—in our visual SLAM framework this amounts to registering previously viewed scenes. Necessarily, this task involves intelligently choosing loop-closure candidates because (*i*) the computational cost of attempting the camera-derived relative-pose constraint (1) is not insignificant, and (*ii*) adding unnecessary/redundant edges to the SLAM pose-graph increases inference complexity and can also lead to overconfidence [73]. Using our previously defined local saliency measure, we can improve the performance of visual SLAM in two key ways:

1) We can sparsify the pose-graph by retaining only visually salient key-frames;
2) We can make link proposal within the graph more efficient and robust by combining visual saliency with geometric measures of information gain.

In the first step, we can decide whether or not a node should be added at all by evaluating its local saliency level—this allows us to decimate visually homogeneous key-frames, which results in a graph that is more sparse and visually informative. This improves the overall efficiency of graph inference and eliminates nodes that would otherwise have low utility in underwater visual perception.

In the second step, we can improve the efficiency of link proposal by making it "salient-aware". For efficient link proposal, the authors of [73] used expected information gain to prioritize which edges to add to the graph—thereby retaining only informative links. However, when considering the case of visual perception, not all camera-derived measurements are equally obtainable. Pairwise registration of low saliency images will fail unless there is a strong prior to guide the putative correspondence search (e.g., Fig. 4 top row), whereas pairwise registration of highly salient image pairs often succeeds even with a weak or uninformative prior (e.g., Fig. 4 bottom row). Hence, when evaluating the expected information gain of proposed links, we should take into account their visual saliency, as this is a good overall indicator of whether or not the expected information gain (i.e., image registration) is actually obtainable. By doing so, we can propose the addition of links that are not only geometrically informative, but also visually plausible.



(a) Scatter plot of relative-pose uncertainty vs. local saliency

$S_L^{\min}$	0.2	0.3	0.4	0.5	0.6	0.7	0.8
successful & retained [%]	100	98	95	82	55	22	5
failed & discarded [%]	5	15	32	51	70	88	99

(b) Effect of thresholding on local saliency

Fig. 11. Local saliency of image pairs that result in successful pairwise image registration for the *R/V Oceanus* dataset. (a) A scatter plot of relative-pose uncertainty versus local saliency for candidate image pairs satisfying a minimum overlap criteria. Blue dots represent all attempted pairs whereas red circles indicate those which were successfully registered. (b) Tabulated data showing what fraction of failed registrations are pruned and what fraction of successful registrations are retained when thresholding on different values for the minimum local saliency threshold, $S_L^{\min}$. For example, by using a threshold of $S_L^{\min} = 0.4$, we retain 95% of successful registrations, yet are able to prune 32% of failed match attempts.

A. Salient Key-Frame Selection

During SLAM exploration, image saliency can be used to pre-evaluate whether or not it would be beneficial to add a key-frame to the graph. Naively adding nodes to the graph can introduce a large number of meaningless variables, thereby making SLAM inference computationally expensive. When we have a measure of usefulness of the node, however, we can intelligently choose which set of nodes to include in the graph—only adding key-frames with high local saliency. For this purpose, we use a minimum threshold on local saliency, $S_L^{\min}$, as a criteria for adding key-frames to the graph.

To determine this threshold, we examined the local saliency score of underwater image pairs that resulted in successful pairwise image registration, while simultaneously examining the relative-pose certainty associated with their PCCS search prior. Fig. 11 displays a scatter plot from this analysis using data from the *R/V Oceanus* dataset (depicted earlier in Fig. 5). Plotted as dots are all attempted pairwise image registrations between nodes satisfying a minimum overlap criteria. Out of this set, those pairs which resulted in a successful pairwise image registration are circled. The results show a strong correlation between image registration success and local saliency. For those pairs which fall below a local saliency level of $S_L < 0.4$, we see that only a small fraction result in registration success, and for those that do, they have a strong PCCS search prior (i.e., low relative-pose uncertainty). Hence, by discarding images with low local saliency, we see that we can eliminate a large fraction of failed candidate pairs. In fact, the empirical evidence shows that we can eliminate 30–70% of the failed attempts by using a minimum saliency threshold somewhere between $S_L^{\min} = 0.4$–0.6.

B. Saliency Incorporated Link Hypothesis

One formal approach to hypothesizing link candidates is to examine the utility of future expected measurements—also known as information gain. For example, Ila et al. [73] use a measure of information gain to add only informative links (i.e., measurements) to the SLAM pose-graph. Other example uses can be found in control [74]–[76], where the control scheme evaluates the information gain of possible future measurements and leads the robot on trajectories that reduce the overall SLAM localization and map uncertainty.

Following [73], we express the information gain of a measurement update between nodes i and j as

$$\mathcal{I} = H(X) - H(X|\mathbf{z}_{ij}), \tag{6}$$

where $H(X)$ and $H(X|\mathbf{z}_{ij})$ are the entropy before and after measurement, $\mathbf{z}_{ij}$, respectively. For a Gaussian distribution, Ila et al. showed that this calculation simplifies to

$$\mathcal{I} = \frac{1}{2} \ln \frac{|\mathrm{S}|}{|\mathrm{R}|}, \tag{7}$$

where R and S are the measurement and innovation covariance, respectively. In the case of our 5-DOF camera observation model (1), the calculation of innovation covariance becomes

$$\mathrm{S} = \mathrm{R} + \begin{bmatrix} \mathrm{H}_i & \mathrm{H}_j \end{bmatrix} \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix} \begin{bmatrix} \mathrm{H}_i & \mathrm{H}_j \end{bmatrix}^\top, \tag{8}$$

where H_i and H_j are the non-zero blocks of (1)'s Jacobian and $\left[\begin{smallmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{smallmatrix}\right]$ is the marginal joint covariance between nodes i and j, which is efficiently recoverable within iSAM [33]. The utility of evaluating (7) is that it can be used to assess which edges are the most informative to add to the pose-graph—before actually attempting image registration.

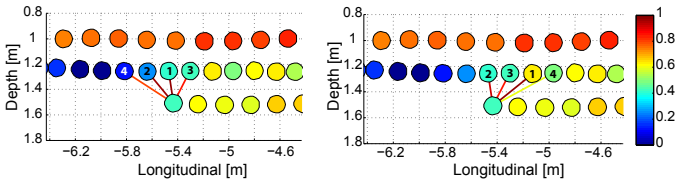
In the approach outlined above, an equal likelihood of measurement availability is assumed. In other words, (7) assesses the geometric value of adding the perceptual constraint *without regard to if, in fact, the constraint can be made.* As evident in our work, not all camera-derived constraints are equally obtainable, and are in fact largely influenced by the visual content within the scene. Candidate links with high information gain may not be the most plausible camera-derived links due to a lack of visual saliency. We argue that the act of perception should play an equal role in determining candidate image pairs.

Based upon the local saliency metric developed earlier, and noting that $S_L \in [0, 1]$, we combine visual saliency with expected information gain to arrive at a combined visual/geometric measure that accounts for perception:

$$\mathcal{I}_L = \begin{cases} \mathcal{I} \cdot S_L & \text{if } S_L \geq S_L^{\min} \text{ and } \mathcal{I} \geq \mathcal{I}^{\min} \\ 0 & \text{o.w.} \end{cases} \tag{9}$$

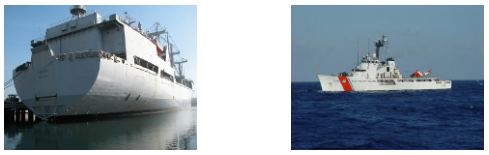
Strictly speaking, (9) is no longer a direct measure of information gain in the mutual information sense; however, it is a scaled version according to visual saliency. This allows us to prioritize candidate image pairs based upon their geometric informativeness as well as their visual registrability.

Presumably two images that have high saliency but low similarity have low probability of matching, so a similarity measure (which depends on the pair of images) seems like it would be better than just saliency, S_L, in (9), which depends only on one image. However, we found that implementing



(a) Link hypothesis using $\mathcal{I}$ (b) Link hypothesis using $\mathcal{I}_L$

Fig. 12. Sample result for link proposal using saliency incorporated information gain on the *R/V Oceanus*. Numbers in nodes indicate the relative ordering of how informative links are (i.e., 1 for the most informative link).



(a) *SS Curtiss* (b) *USCGC Venturous*

Fig. 13. Underwater hull inspection experiments conducted using the Bluefin Robotics HAUV on the hulls of the *SS Curtiss* and the *USCGC Venturous*.

similarity scores in (9), such as those reported by [50], [55] and [57], does not produce the desired result in our application for two main reasons:

1) Since our vocabularies are orders of magnitude smaller than place recognition methods ($\mathcal{O}(100)$ vs. $\mathcal{O}(10\text{k})$), we do not have enough visual variety in our quantization to accurately index imagery and support place recognition similarity measures.
2) Spatial overlap between neighboring imagery is small in our application—typically between 20% to 50%. We tested term frequency-inverse document frequency (tf-idf) similarity scoring as reported in [57], but found that our small overlap results in very low tf-idf scores due to common words occurring everywhere on the hull. Alternatively, when testing with the cosine distance between two BoW histograms, we found this yielded a large distance measure due to the histograms having inadequate intersection, also because of the small overlap.

In our hull inspection application, we found that the combined approach in (9) results in better link hypothesis than (7) alone—forcing the link proposal scheme to lean toward visually salient nodes among those that are equally informative. Fig. 12 depicts a sample result from the *R/V Oceanus* dataset. The color of a proposed link indicates how informative the link is (i.e., $\mathcal{I}$), while the color of a node represents how salient the imagery is (i.e., S_L). In the first case, only the geometry of the constraint is taken into account through the calculation of information gain. In the second case, the combined measure (9) guides the selection toward feature-rich image pairs, rather than processing visually uninformative images with high geometric gain. In doing so, it proposes realistically achievable camera-derived candidate links.

V. Results

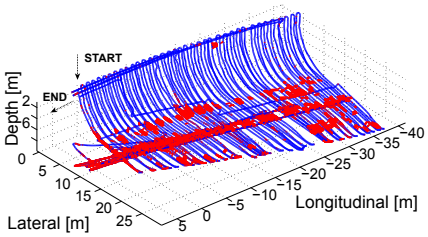
This section reports experimental results evaluating our real-time visual SLAM algorithm. The first dataset is from a February 2011 survey of the *SS Curtiss* (Fig. 13) using the HAUV. The *SS Curtiss* is a $183\ \mathrm{m}$ long single-screw roll-on/roll-off container ship currently stationed at the U.S. Naval Station in San Diego, California. The hull survey mission consisted of vertical tracklines, extending from the waterline to the keel, spaced approximately $0.5\ \mathrm{m}$ apart laterally. The survey started near the bow and continued toward the stern while maintaining a vehicle standoff distance of approximately $1\ \mathrm{m}$ from the hull using DVL measured range. This configuration resulted in approximately 30% cross-track image overlap for a $\sim 45^\circ$ horizontal camera field of view (in water). Occasionally the vehicle was commanded to swim back toward the bow, orthogonal to its nominal trackline trajectory, so as to obtain image data useful for time-elapsed loop-closure constraints. The total survey area comprised a swath of approximately $45\ \mathrm{m}$ along-hull by $25\ \mathrm{m}$ athwart hull for a total path length of $2.7\ \mathrm{km}$ and $3.4\ \mathrm{hr}$ mission duration. The camera was operated at a fixed sample rate of $2\ \mathrm{Hz}$, which resulted in a dataset of 24,773 source images. The dataset was logged using the LCM publish/subscribe software framework [40], which supports a real-time playback capability useful for post-mission software development and benchmark analysis. Results presented here are for post-process real-time playback using the visual SLAM algorithm implementation as described in this paper.

A. Saliency-Ignored SLAM Baseline Results

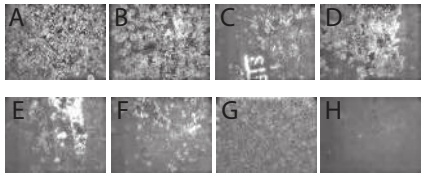
For these experiments we ran the visual SLAM algorithm in a "perceptually naive" mode to benchmark its performance in the absence of saliency-based key-frame selection and saliency-incorporated link hypothesis. For these tests we added image key-frames at a fixed spatial sample rate resulting in approximately 70% sequential image overlap, and used geometric information gain only (i.e., not saliency incorporated) for link hypothesis. We ran with three different levels of link hypothesis: $n_{\text{plink}} = 3$, $n_{\text{plink}} = 10$, and $n_{\text{plink}} = 30$, where n_{plink} represents the maximum number of proposed hypotheses per node. We refer to the $n_{\text{plink}} = 30$ case as the "exhaustive SLAM result", as all nominal nodes were added and all geometrically informative links where tried. This brute force result serves as a baseline for the number of successfully registered camera links that can be obtained in this dataset.

The resulting 3D trajectory for the exhaustive SLAM case is depicted in Fig. 14(a). It contains 17,207 camera nodes, 29,426 5-DOF camera constraints, and required a cumulative processing time of 10.70 hours (this includes image registration and iSAM inference). Fig. 14(c) shows a top-down view of the successful pairwise camera links (hypotheses), illustrating where they spatially occurred in the 3D pose-graph.

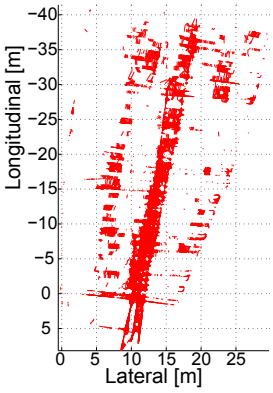
Using this exhaustive SLAM result as a baseline, we evaluate the performance of our saliency metrics by applying our local and global saliency algorithms to the exhaustive SLAM graph and then overlay their result. In particular, Fig. 14(d) shows that local saliency, S_L, correlates well where successful camera-edges occurred in the exhaustive SLAM graph. The bottom of the hull had a high concentration of marine growth (e.g., images A to F in Fig. 14(b)), making it visually feature-rich for pairwise image registration—it also independently



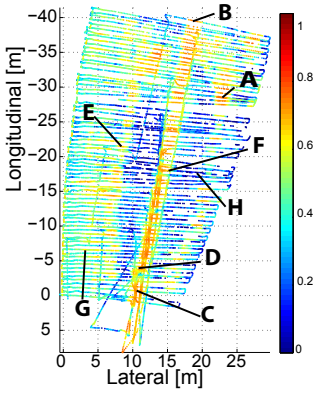
(a) 3D view of exhaustive SLAM result



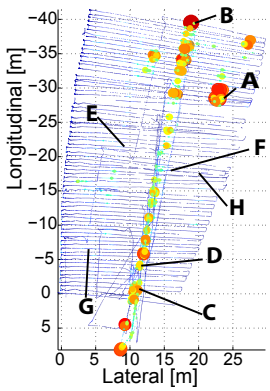
(b) Sample images



(c) Camera constraints



(d) Local saliency S_L



(e) Global saliency S_G

Fig. 14. Exhaustive, non real-time, baseline SLAM result for the *SS Curtiss* dataset for benchmark comparison. No saliency aiding is used in image key-frame selection nor in link hypothesis; camera nodes are uniformly added to the pose-graph based upon distance traveled. (a) The exhaustive SLAM graph consists of 17,207 nodes and 29,426 camera-derived edges (this includes along-track and cross-track edges); a link hypothesis factor of $n_{\text{plink}} = 30$ per node is used. (b) Sample imagery from along the hull—labels correspond to denoted locations in (d) and (e). (c) A top-down view of the pose-graph depicting where the successful pairwise camera-derived edges occur. (d) A top-down view of the pose-graph with our local saliency metric, S_L, overlaid. Note how S_L predicts well where successful camera registrations actually occur. (e) A top-down view of the pose-graph with our global saliency metric, S_G, overlaid. In addition to the colormap overlay, node size has been scaled by its saliency level for visual clarity. Note how S_G's character is distinctly different from the local saliency graph. Global saliency highlights only a handful of regions as being visually novel relative to the rest of the hull.

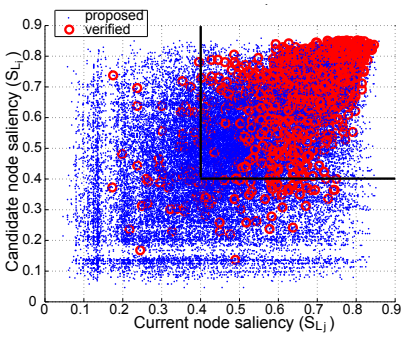


Fig. 15. Scatter plot depicting all attempted pairwise image hypotheses for the exhaustive SLAM result as viewed in saliency space. Each dot represents a single link hypothesis and indicates the (S_{L_i}, S_{L_j}) local saliency value for the image pair; successfully registered image pairs are circled. Note the strong positive correlation that exists between successfully registered pairs and their local saliency values. For reference, hypotheses that would be eliminated by a local saliency threshold of $S_L^{\text{min}} = 0.4$ lie outside the demarcated region.

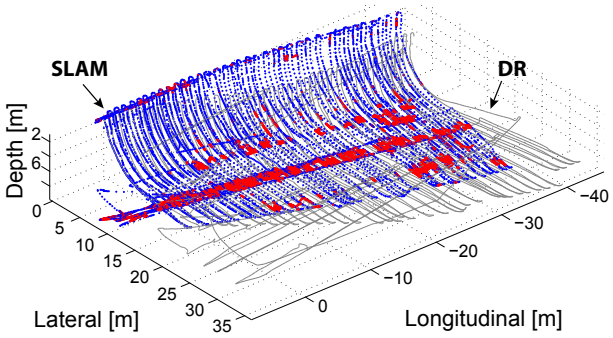
received a high local saliency score; this is where the majority of cross-track image registrations occurred. The vertical side of the hull was relatively clean and thus feature empty (e.g., images G and H in Fig. 14(b)), so relatively few pairwise registrations occurred in those regions—it also independently received a low local saliency score.

More quantitatively, Fig. 15 depicts a scatter plot, in local saliency space, of all proposed pairwise link hypotheses that were attempted by the exhaustive SLAM result. Each dot in the plot represents an attempted link registration between camera nodes $\mathbf{x}_i$ (candidate node) and $\mathbf{x}_j$ (current node), while each circle represents those pairs which resulted in image registration success. Each axis in the graph represents the individual local saliency levels (S_{L_i} and S_{L_j}) for the two images. The plot shows a positively correlated distribution in local saliency for registered links (i.e., circles). Successfully registered links are concentrated in the top-right corner of saliency space where both nodes have a high score. This distribution reveals that a large number of non-visually-plausible links could in fact be pruned from the SLAM process by incorporating local saliency into the key-frame selection and link hypothesis generation.

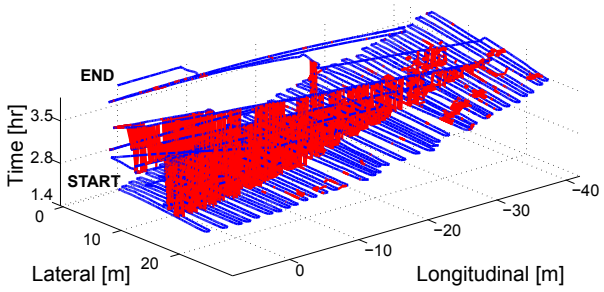
B. Saliency-Informed SLAM Result

For this experiment we ran the visual SLAM algorithm with saliency-based key-frame selection and saliency-informed information gain enabled. Based upon our earlier tests with the *R/V Oceanus* dataset (Fig. 11), we used a minimum saliency threshold of $S_L^{\text{min}} = 0.4$ for both image key-frame selection (demarcated region in Fig. 15) and link hypothesis. In non-salient regions, we used a minimum time threshold to add poses to the graph every 1 s for smoothed trajectory visualization. The resulting saliency-informed SLAM trajectory is depicted in Fig. 16. Using the saliency-based front-end, we reduced the total number of image key-frames from 17,207 (in the exhaustive set), to only 8,728—a 49.3% reduction by culling visually uninformative nodes from the graph. Moreover, the total processing time is only 1.31 hr, which is 2.6x faster than real-time. The tabulated values in Fig. 16(d) and Fig. 17(b) summarize the overall computational efficiency improvement.

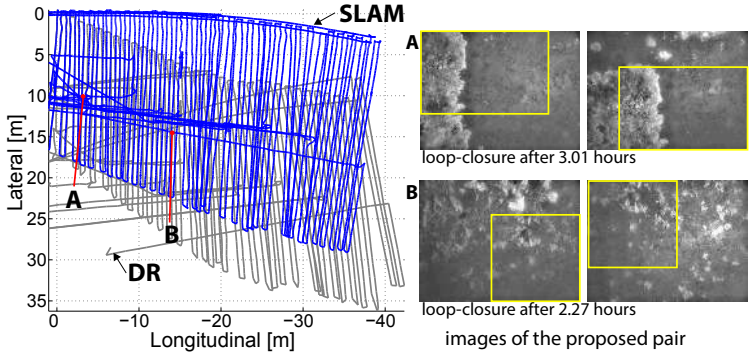
In terms of saliency's effect on SLAM performance, we note that even with far less nodes in the graph (just 8,728 versus saliency-ignored's 17,207), we were still able to achieve almost the same performance as the baseline exhaustive SLAM result in terms of estimated trajectory (Fig. 18), and better than saliency-ignored SLAM with a similar or comparable number of link proposals (i.e., $n_{\text{plink}} = 10$ and $n_{\text{plink}} = 3$). In



(a) 3D view of saliency-informed SLAM result



(b) Time elevation graph depicting loop-closure measurements



(c) 2D top-down view of SLAM versus DR

No. of	Saliency-ignored			Saliency-informed
Image key-frames	17,207	17,207	17,207	8,728
Hypoth. per node	30	10	3	3
iSAM CPU time (hr)	8.70	3.37	1.64	0.52
Image CPU time (hr)	2.00	1.05	1.31	0.79
Total CPU time (hr)	10.70	4.42	2.95	1.31
Speed up over real-time	0.3×	0.8×	1.1×	2.6×

(d) SLAM inference summary

Fig. 16. Real-time visual SLAM result for the *SS Curtiss* dataset using saliency driven image key-frame selection and saliency incorporated information gain for link hypothesis. The saliency-informed SLAM graph consists of 8,728 image nodes and used $n_{\text{plink}} = 3$ per node. The cumulative iSAM inference time in this case is 0.52 hours, and when accounting for image processing time, the entire SLAM result can be computed in less than 1.31 hours, which is 2.6× faster than the actual mission duration time of 3.4 hours. (a) The blue dotted trajectory represents the iSAM estimate with camera constraints depicted as red edges, while the gray trajectory represents dead-reckoned (DR). (b) The xy component of the SLAM trajectory estimate is plotted versus time, where the vertical axis represents mission time. This depiction makes it easier to visualize the elapsed duration between loop-closure camera measurements. (c) A top-down view of the SLAM estimate versus DR. The positions marked 'A' and 'B' are two examples of where large loop-closure events take place. The images on the right depict the key-frames and registered loop-closure event, verifying the overall consistency of the metric SLAM solution. For visual clarity, the yellow boxes indicate the common overlap between the two registered images. (d) A tabulated summary of the SLAM inference statistics. The actual mission duration was 3.40 hr and totaled 24,773 images at 2 Hz.

fact, Fig. 17(a) shows that saliency-informed SLAM's image registration success rate was nearly 60% out of links that it proposed whereas the saliency-ignored SLAM results were all less than 20%. Moreover, when comparing the amount of elapsed-time occurring between successful loop-closures (Fig. 17(b)) we see that in the case of image pairs with more than 1 hour of elapsed time between them that the saliency-informed SLAM result obtained 1275% more links than the comparable $n_{\text{plink}} = 3$ case of saliency-ignored SLAM.

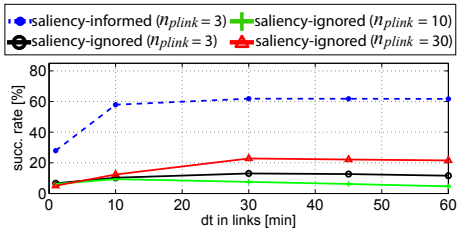
For easier loop-closure visualization, Fig. 16(b) depicts a time elevation graph of camera registration constraints—here the vertical axis indicates elapsed mission time. Camera measurements with large time differences indicate large loop-closure events—for example, the SLAM estimate was accurate enough to register image pairs with over three hours of elapsed time difference (events A and B in Fig. 16(c)). As Fig. 16(a) and Fig. 16(c) show, this is a significant improvement over the dead-reckoned odometry result. While saliency-ignored SLAM also shows reduced error over DR, saliency-informed SLAM substantially outperforms it by resulting in more verified links and less error relative to the baseline exhaustive SLAM result—despite cases where it used a lesser number of link proposals (e.g., $n_{\text{plinks}} = 3$ versus $n_{\text{plink}} = 10$). This is because the saliency-informed result actively takes into account the visual plausibility of imagery when considering its utility for SLAM.

C. Global Saliency Results

Unlike the local saliency metric, the global saliency metric reacts to rare or anomalous features. For evaluation, three different hull data sets were tested: the *R/V Oceanus* (Fig. 5(a)), the *SS Curtiss* (Fig. 13(a)), and the *USCGC Venturous* (Fig. 13(b)).

1) R/V Oceanus: Fig. 5(d) shows that the global saliency map on the hull of the *R/V Oceanus* can have low scores even for locally salient imagery (e.g., weld lines). This is because several of the vocabulary words (e.g., weld lines) occur frequently throughout the environment—lowering their overall idf score.

2) SS Curtiss: Fig. 14(e) shows that the global saliency map, S_G, has a macro scale character on the *SS Curtiss* distinctly different from local saliency, S_L. Global saliency's normalized idf score down-weights the inter-image occurrence of visually prevalent features and marks only a few regions as being globally rare relative to the rest of the hull (e.g., images A, B, C, and D in Fig. 14(b)). These images correspond to



(a) Image registration success rate versus elapsed time between key-frames

Δt		Saliency-ignored $n_{\text{plink}} = 30$ No. of	pct.	Saliency-ignored $n_{\text{plink}} = 10$ No. of	pct.	Saliency-ignored $n_{\text{plink}} = 3$ No. of	pct.	Saliency-informed $n_{\text{plink}} = 3$ No. of
1 min	plink	457,165	3%	124,653	12%	12,524	124%	15,553
	vlink	23,125	18%	7,772	56%	829	524%	4,348
	% success	3.6%	472%	1.9%	895%	6.6%	258%	17.0%
10 min	plink	133,282	3%	25,182	18%	2,848	160%	4,565
	vlink	16,476	16%	2,353	112%	293	902%	2,644
	% success	12.4%	469%	9.3%	620%	10.2%	568%	57.9%
1 hour	plink	38,701	6%	11,300	21%	1,001	239%	2,397
	vlink	8,348	18%	527	281%	116	1,275%	1,479
	% success	21.5%	286%	4.6%	1,324%	11.6%	532%	61.7%

(b) Tabulated success rate for a minimal elapsed time between key-frames of 1, 10, and 60 minutes

Fig. 17. Comparison of the link hypothesis success rate for the different SLAM results. Note that temporally sequential links are excluded from this analysis as we start the time difference at greater than 1 minute (i.e., at least 120 images apart at 2 Hz image sample rate). (a) Plot of the image registration success rate, defined as the number of verified links over the number of proposed links, for saliency-informed and saliency-ignored SLAM. The abscissa represents the amount of elapsed time occurring between the proposed image pairs (i.e., $dt = 30$ means 30 minutes of elapsed mission time between the two key-frames being attempted for registration). Links with a large time difference correspond to large loop-closure events. As can be seen, the saliency-informed link proposal yields a higher success rate as compared to the saliency-ignored results, which is because the saliency-informed SLAM link proposal takes into account the visual plausibility of attempted nodes. (b) A tabular comparison of proposed links (i.e., plink), verified links (i.e., vlink), and their resulting success rate for the different SLAM results. The column "pct." represents the percentage obtained by the saliency-informed result.

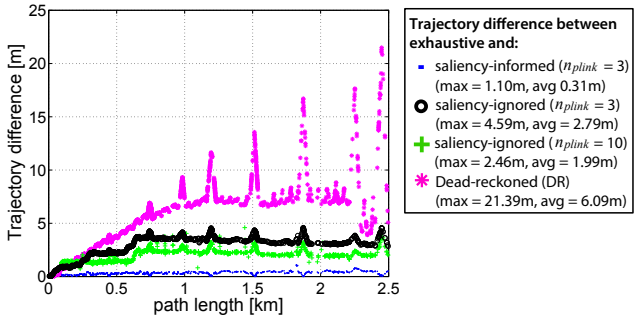
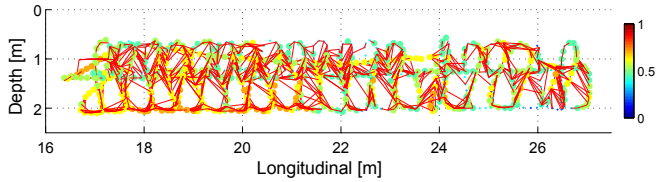
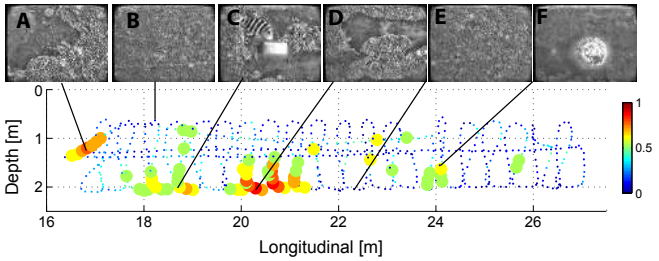


Fig. 18. A plot of the Euclidean distance between the different trajectory estimates relative to the baseline exhaustive SLAM result. The max difference between saliency-informed and exhaustive SLAM is 1.10 m, whereas the DR trajectory shows significantly larger discrepancy (21.39 m) due to navigation drift. The other two saliency-ignored SLAM results also show larger discrepancy relative to the exhaustive SLAM result throughout the mission.



(a) Local saliency map on *USCGC Venturous*



(b) Global saliency map on *USCGC Venturous*

Fig. 19. Local and global saliency maps for a survey on the hull of the *USCGC Venturous*. (a) Most of the hull is covered in texture-rich barnacles making the scene everywhere locally salient. In this case, camera measurements and locally salient nodes are evenly distributed everywhere on the hull. (b) Since the surface of the vessel is covered with marine growth (e.g., imagery B and E) the globally saliency score is low those regions. On the other hand, two artificial targets (images C and F), and distinguished scenes where there are no barnacles (images A and D), score high global saliency and are correctly denoted as rare areas on the hull.

regions of the hull where the scene content is distinct relative to the rest of the hull.

3) USCGC Venturous: Fig. 19 shows results for the *USCGC Venturous* survey, whose hull is covered with barnacles, yielding a high local saliency score everywhere on the hull (e.g., images B and E are representative of this barnacle growth). In two distinct locations there were artificial targets (inert mines) attached to the hull by divers for the inspection experiment. These regions scored a high global saliency score (i.e., images C and F) since they are rare relative to the rest of barnacle imagery seen on the hull. Moreover, other visually uncommon scenes, such as images A and D, also scored high due to their absence of full barnacle cover.

In all three different hull evaluations, *R/V Oceanus*, *SS Curtiss*, and *USCGC Venturous*, we see that global saliency identifies anomalous (i.e., rare) scenes with respect to the rest of the hull. For example, these visually distinctive regions can serve as useful locations for planning paths within an active SLAM framework for attempting loop-closure on the hull, as reported separately in [71], [72]. One observation worth noting is that global saliency does not necessarily imply texture-rich scenes, as demonstrated by images A and D of the *USCGC Venturous*. In those images, note that it is the *absence* of barnacle texture that designates those images as rare relative to the rest of the hull environment.

VI. Conclusion

This paper reported on a real-time 6-DOF monocular visual SLAM algorithm for autonomous underwater ship hull inspection. Two types of novel visual saliency measures were introduced: local saliency and global saliency. Local saliency was shown to provide a normalized measure of intra-image feature diversity, while global saliency was shown to provide a

normalized measure of inter-image rarity. Using three distinct hull inspection datasets we showed how local saliency can be used to guide key-frame selection, as well as how it can be combined with information gain to propose visually plausible links, and that global saliency can be used to identify visually rare regions on the hull.

ACKNOWLEDGMENT

We would like to thank J. Vaganay and K. Shurn from the Bluefin Robotics Corporation for their excellent support during the course of the experiments. We would also like to thank F. Hover, M. Kaess, B. Englot, H. Johansson, and J. Leonard from the Massachusetts Institute of Technology for their collaboration during the course of this project.

REFERENCES

[1] A. Kim and R. M. Eustice, "Pose-graph visual SLAM with geometric model selection for autonomous underwater ship hull inspection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, St. Louis, MO, Oct. 2009, pp. 1559–1565.

[2] ——, "Combined visually and geometrically informative link hypothesis for pose-graph visual SLAM using bag-of-words," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, San Francisco, CA, Sept. 2011, pp. 1647–1654.

[3] J. Mittleman and D. Wyman, "Underwater ship hull inspection," *Naval Engineers J.*, vol. 92, no. 2, pp. 122–128, Apr. 1980.

[4] J. Mittleman and L. Swan, "Underwater inspection for welding and overhaul," *Naval Engineers J.*, vol. 105, no. 5, pp. 37–42, Sept. 1993.

[5] R. B. Olds, "Marine mammals systems in support of force protection," in *SSC San Diego Biennial Review 2003*. San Diego, CA: Space and Naval Warfare Systems Center, San Diego, June 2003, ch. Chapter 3: Intelligence, Surveillance, and Reconnaissance, pp. 131–135.

[6] D. Lynn and G. Bohlander, "Performing ship hull inspections using a remotely operated vehicle," in *Proc. IEEE/MTS OCEANS Conf. Exhib.*, vol. 2, Aug. 1999, pp. 555–562.

[7] A. Carvalho, L. Sagrilo, I. Silva, J. Rebello, and R. Carneval, "On the reliability of an automated ultrasonic system for hull inspection in ship-based oil production units," *Applied Ocean Res.*, vol. 25, no. 5, pp. 235–241, Oct. 2003.

[8] S. Negahdaripour and P. Firoozfam, "An ROV stereovision system for ship hull inspection," *IEEE J. Ocean. Eng.*, vol. 31, no. 3, pp. 551–546, July 2006.

[9] G. S. Bohlander, G. Hageman, F. S. Halliwell, R. H. Juers, and D. C. Lynn, "Automated underwater hull maintenance vehicle," Naval Surface Warfare Center Carderock Division, Bethesda, MD, Tech. Rep. ADA261504, May 1992.

[10] S. Harris and E. Slate, "Lamp Ray: Ship hull assessment for value, safety and readiness," in *Proc. IEEE/MTS OCEANS Conf. Exhib.*, vol. 1, Seattle, WA, 1999, pp. 493–500.

[11] G. Trimble and E. Belcher, "Ship berthing and hull inspection using the CetusII AUV and MIRIS high-resolution sonar," in *Proc. IEEE/MTS OCEANS Conf. Exhib.*, vol. 2, 2002, pp. 1172–1175.

[12] J. Vaganay, M. Elkins, D. Esposito, W. O'Halloran, F. Hover, and M. Kokko, "Ship hull inspection with the HAUV: U.S. Navy and NATO demonstrations results," in *Proc. IEEE/MTS OCEANS Conf. Exhib.*, Boston, MA, 2006, pp. 1–6.

[13] L. Menegaldo, M. Santos, G. Ferreira, R. Siqueira, and L. Moscato, "SIRUS: A mobile robot for floating production storage and offloading (FPSO) ship hull inspection," in *IEEE Int. Workshop Advanced Motion Control*, Mar. 2008, pp. 27–32.

[14] L. Menegaldo, G. Ferreira, M. Santos, and R. Guerato, "Development and navigation of a mobile robot for floating production storage and offloading ship hull inspection," *IEEE Trans. Ind. Electron.*, vol. 56, no. 9, pp. 3717–3722, Sept. 2009.

[15] Desert Star Systems, *Ship Hull Inspections with AquaMap*, Desert Star Systems, Marina, CA, Aug. 2002.

[16] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–110, June 2006.

[17] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM): Part II," *IEEE Robot. Autom. Mag.*, vol. 13, no. 3, pp. 108–117, Sept. 2006.

[18] P. Ridao, M. Carreras, D. Ribas, and R. Garcia, "Visual inspection of hydroelectric dams using an autonomous underwater vehicle," *J. Field Robot.*, vol. 27, no. 6, pp. 759–778, Nov. 2010.

[19] M. Walter, F. Hover, and J. Leonard, "SLAM for ship hull inspection using exactly sparse extended information filters," in *Proc. IEEE Int. Conf. Robot. and Automation*, Pasadena, CA, May 2008, pp. 1463–1470.

[20] H. Johannsson, M. Kaess, B. Englot, F. Hover, and J. J. Leonard, "Imaging sonar-aided navigation for autonomous underwater harbor surveillance," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Taipei, Taiwan, Oct. 2010, pp. 4396–4403.

[21] J. Vaganay, L. Gurfinkel, M. Elkins, D. Jankins, and K. Shurn, "Hovering autonomous underwater vehicle — system design improvements and performance evaluation results," in *Proc. Int. Symp. Unmanned Untethered Subm. Tech.*, Durham, NH, 2009, pp. 1–14.

[22] F. S. Hover, R. M. Eustice, A. Kim, B. Englot, H. Johannsson, M. Kaess, and J. J. Leonard, "Advanced perception, navigation and planning for autonomous in-water ship hull inspection," *Int. J. Robot. Res.*, vol. 31, no. 12, pp. 1445–1464, Oct. 2012.

[23] L. G. Weiss, "Autonomous robots in the fog of war," *IEEE Spectrum*, vol. 48, no. 8, pp. 30–36, Aug. 2011.

[24] F. Lu and E. Milios, "Globally consistent range scan alignment for environment mapping," *Auton. Robot.*, vol. 4, pp. 333–349, Apr. 1997.

[25] K. Konolige, "Large-scale map-making," in *Proc. AAAI Nat. Conf. Artif. Intell.*, San Jose, CA, July 2004, pp. 457–463.

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

[27] E. Olson, J. Leonard, and S. Teller, "Spatially-adaptive learning rates for online incremental SLAM," in *Proc. Robot.: Sci. & Syst. Conf.*, Atlanta, GA, USA, June 2007.

[28] K. Konolige and M. Agrawal, "FrameSLAM: From bundle adjustment to real-time visual mapping," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1066–1077, Oct. 2008.

[29] G. Grisetti, D. Lodi Rizzini, C. Stachniss, E. Olson, and W. Burgard, "Online constraint network optimization for efficient maximum likelihood map learning," in *Proc. IEEE Int. Conf. Robot. and Automation*, Pasadena, CA, May 2008, pp. 1880–1885.

[30] M. Kaess, V. Ila, R. Roberts, and F. Dellaert, "The Bayes tree: An algorithmic foundation for probabilistic robot mapping," in *Int. Workshop Alg. Foundations Robot.*, Singapore, Dec. 2010, pp. 157–173.

[31] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," *IEEE Trans. Robot.*, vol. 24, no. 6, pp. 1365–1378, Dec. 2008.

[32] M. Kaess, H. Johannsson, and J. Leonard, "Open source implementation of iSAM," http://people.csail.mit.edu/kaess/isam, 2010.

[33] M. Kaess and F. Dellaert, "Covariance recovery from a square root information matrix for data association," *Robot. and Auton. Syst.*, vol. 57, no. 12, pp. 1198–1210, Dec. 2009.

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

[35] R. M. Eustice, O. Pizarro, H. Singh, and J. Howland, "UWIT: Underwater image toolbox for optical image processing and mosaicking in Matlab," in *Proc. Int. Symp. Underwater Tech.*, Tokyo, Japan, Apr. 2002, pp. 141–145.

[36] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[37] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 346–359, June 2008.

[38] C. Wu, "SiftGPU: A GPU implementation of scale invariant feature transform (SIFT)," http://cs.unc.edu/ ccwu/siftgpu, 2007.

[39] R. Haralick, "Propagating covariance in computer vision," in *Proc. Int. Conf. Pattern Recog.*, vol. 1, Jerusalem, Israel, Oct. 1994, pp. 493–498.

[40] A. S. Huang, E. Olson, and D. C. Moore, "LCM: Lightweight communications and marshalling," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Taipei, Taiwan, Oct. 2010, pp. 4057–4062.

[41] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.

[42] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Comput. Vis.*, vol. 45, no. 2, pp. 83–105, 2001.

[43] 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," *Adv. Robot.*, vol. 24, no. 11, pp. 1595–1613, Jan. 2010.
[44] M. Johnson-Roberson, "Large-scale multi-sensor 3D reconstructions and visualizations of unstructured underwater environments," Ph.D. dissertation, Univ. Sydney, 2010.
[45] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. European Conf. Comput. Vis.*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, vol. 3954, ch. 38, pp. 490–503.
[46] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 2232–2239.
[47] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. European Conf. Comput. Vis.*, 2004, pp. 1–22.
[48] R. Toldo, U. Castellani, and A. Fusiello, "A bag of words approach for 3d object categorization," in *Proc. IEEE Int. Conf. Comput. Vis.* Berlin, Heidelberg: Springer-Verlag, 2009, pp. 116–127.
[49] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 29–44, 2001.
[50] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, Oct. 2003, pp. 1470–1477.
[51] L. Wu, S. Hoi, and N. Yu, "Semantics-preserving bag-of-words models and applications," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1908 –1920, July 2010.
[52] N. Lazic and P. Aarabi, "Importance of feature locations in bag-of-words image classification," in *Proc. IEEE Conf. Acoustics, Speech, Signal Process.*, vol. 1, 4 2007, pp. 641–644.
[53] 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. Pattern Recog.*, 11 2009, pp. 1–5.
[54] D. Larlus, J. Verbeek, and F. Jurie, "Category level object segmentation by combining bag-of-words models with Dirichlet processes and random fields," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 238–253, June 2010.
[55] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, June 2008.
[56] ——, "Highly scalable appearance-only SLAM—FAB-MAP 2.0," in *Proc. Robot.: Sci. & Syst. Conf.*, Seattle, USA, June 2009.
[57] 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. Robot.*, vol. 24, no. 5, pp. 1027–1037, Oct. 2008.
[58] A. Kawewong, N. Tongprasit, S. Tangruamsub, and O. Hasegawa, "Online and incremental appearance-based SLAM in highly dynamic environments," *Int. J. Robot. Res.*, vol. 30, no. 1, pp. 33–55, Jan. 2011.
[59] H. Shahbazi and H. Zhang, "Application of locality sensitive hashing to realtime loop closure detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2011, pp. 1228–1233.
[60] T. Nicosevici and R. Garcia, "On-line visual vocabularies for robot navigation and mapping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Oct. 2009, pp. 205–212.
[61] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
[62] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, vol. 290, no. 5802, pp. 91–97, Mar. 1981.
[63] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2032–2047, Nov. 2009.
[64] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Documentation*, vol. 28, pp. 11–21, 1972.
[65] G. Salton and C. S. Yang, "On the specification of term values in automatic indexing," *J. Documentation*, vol. 29, no. 4, pp. 351–372, 1973.
[66] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for idf," *J. Documentation*, vol. 60, no. 5, pp. 503–520, 2004.
[67] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 6 2009, pp. 1169–1176.
[68] O. Chum, J. Philbin, and A. Zisserman, "Near duplicate image detection: min-hash and tf-idf weighting," in *Proc. British Mach. Vis. Conf.*, 2008, pp. 493–502.
[69] O. Chum and J. Matas, "Unsupervised discovery of co-occurrence in sparse high dimensional data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, June 2010, pp. 3416–3423.
[70] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.
[71] A. Kim and R. M. Eustice, "Next-best-view visual SLAM for bounded-error area coverage," in *IROS Workshop on Active Semantic Perception*, Vilamoura, Portugal, October 2012.
[72] A. Kim, "Active visual SLAM with exploration for autonomous underwater navigation," Ph.D. dissertation, University of Michigan, Ann Arbor, MI, August 2012.
[73] V. Ila, J. M. Porta, and J. Andrade-Cetto, "Information-based compact pose SLAM," *IEEE Trans. Robot.*, vol. 26, no. 1, pp. 78–93, Feb. 2010.
[74] R. Sim and N. Roy, "Global A-optimal robot exploration in SLAM," in *Proc. IEEE Int. Conf. Robot. and Automation*, Barcelona, Spain, Apr. 2005, pp. 661–666.
[75] T. Vidal-Calleja, A. Davison, J. Andrade-Cetto, and D. Murray, "Active control for single camera SLAM," in *Proc. IEEE Int. Conf. Robot. and Automation*, Orlando, FL, May 2006, pp. 1930–1936.
[76] M. Bryson and S. Sukkarieh, "An information-theoretic approach to autonomous navigation and guidance of an uninhabited aerial vehicle in unknown environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, Aug. 2005, pp. 3770–3775.

Ayoung Kim (S'08) received the B.S. and M.S. degrees in mechanical engineering from Seoul National University, Seoul, Korea, in 2005 and 2007, respectively, and a M.S. in electrical engineering and Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, MI in 2011 and 2012, respectively.

Currently, she is a Postdoctoral Scholar with the Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor. Her research interests include visual simultaneous localization and mapping, navigation, and computer vision.

Ryan M. Eustice (S'00–M'05–SM'10) received the B.S. degree in mechanical engineering from Michigan State University, East Lansing, MI in 1998, and the Ph.D. degree in ocean engineering from the Massachusetts Institute of Technology/Woods Hole Oceanographic Institution Joint Program, Woods Hole, MA, in 2005.

Currently, he is an Assistant Professor with the Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor, with joint appointments in the Department of Electrical Engineering and Computer Science, and in the Department of Mechanical Engineering. His research interests include autonomous navigation and mapping, computer vision and image processing, mobile robotics, and autonomous underwater vehicles.