REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
23-02-2012

2. REPORT TYPE
Conference Proceedings

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
Image Feature Detection and Matching in Underwater Conditions

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
0602435N

6. AUTHOR(S)
K. Oliver, Weilin Hou, S. Wang

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
73-6369-00-5

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Research Laboratory
Oceanography Division
Stennis Space Center, MS 39529-5004

8. PERFORMING ORGANIZATION REPORT NUMBER
NRL/PP/7330-10-0246

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
One Liberty Center
875 North Randolph Street, Suite 1425
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution is unlimited.

13. SUPPLEMENTARY NOTES

20120307032

14. ABSTRACT

The main challenge in underwater imaging and image analysis is to overcome the effects of blurring due to the strong scattering of light by the water and its constituents. This blurring adds complexity to already challenging problems like object detection and localization. The current state-of-the-art approaches for object detection and localization normally involve two components: (a) a feature detector that extracts a set of feature points from an image, and (b) a feature matching algorithm that tries to match the feature points detected from a target image to a set of template features corresponding to the object of interest. A successful feature matching indicates that the target image also contains the object of interest. For underwater images, the target image is taken in underwater conditions while the template features are usually extracted from one or more training images that are taken out-of-water or in different underwater conditions. In addition, the objects in the target image and the training images may show different poses, including rotation, scaling, translation transformations, and perspective changes. In this paper we investigate the effects of various underwater point spread functions on the detection of image features using many different feature detectors, and how these functions affect the capability of these features when they are used for matching and object detection. This research provides insight to further develop robust feature detectors and matching algorithms that are suitable for detecting and localizing objects from underwater images.

15. SUBJECT TERMS
Underwater Imaging, Object Detection, Object Recognition, Feature Detection, Feature Description, Point Spread Function

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Weilin Hou
Unclassified	Unclassified	Unclassified	UU	12	19b. TELEPHONE NUMBER *(Include area code)* 228-688-5257

Image feature detection and matching in underwater conditions

Kenton Oliver[a], Weilin Hou[b], and Song Wang[a]

[a]University of South Carolina, 201 Main Street, Columbia, South Carolina, USA;
[b]Naval Research Lab, Code 7333, 1009 Balch Blvd., Stennis Space Center, Mississippi, USA

ABSTRACT

The main challenge in underwater imaging and image analysis is to overcome the effects of blurring due to the strong scattering of light by the water and its constituents. This blurring adds complexity to already challenging problems like object detection and localization. The current state-of-the-art approaches for object detection and localization normally involve two components: (a) a feature detector that extracts a set of feature points from an image, and (b) a feature matching algorithm that tries to match the feature points detected from a target image to a set of template features corresponding to the object of interest. A successful feature matching indicates that the target image also contains the object of interest. For underwater images, the target image is taken in underwater conditions while the template features are usually extracted from one or more training images that are taken out-of-water or in different underwater conditions. In addition, the objects in the target image and the training images may show different poses, including rotation, scaling, translation transformations, and perspective changes. In this paper we investigate the effects of various underwater point spread functions on the detection of image features using many different feature detectors, and how these functions affect the capability of these features when they are used for matching and object detection. This research provides insight to further develop robust feature detectors and matching algorithms that are suitable for detecting and localizing objects from underwater images.

Keywords: Underwater Imaging, Object Detection, Object Recognition, Feature Detection, Feature Description, Point Spread Function

1. INTRODUCTION

Detection, description, and matching of discriminative image features are fundamental problems in computer vision and have been studied for many years. Algorithms to solve these problems play key roles in many vision applications, such as image stitching,[1,2] image registration,[3,4] object detection,[5] object localization,[6] and object recognition.[7] In practice, feature descriptors are made to be invariant to certain spatial transformations, such as scaling and rotation.

Geodesic Invariant Histograms (GIH)[8] model a grayscale image as a 2D surface embedded in 3D space, where the height of the surface is defined by the image intensity at the corresponding pixel. Under this surface model a feature descriptor, based on geodesic distances on the surface, is defined which is invariant to some general image deformations. A local-to-global framework was adopted in[9] where multiple support regions are used for describing the features. This removes the burden of finding the optimal scale as both local and global information is embedded in its descriptor. The Scale-Invariant Feature Transform (SIFT)[10] is a well-known choice for detecting and describing features. Comparison studies[11] have shown that SIFT and its derivatives[11–13] perform better than other feature detectors in various domains. SIFT is invariant to rotation and scaling, and has been shown to be invariant to small changes in illumination and perspective (up to 50 degrees).

Further author information: (Send correspondence to K.O.)

K.O. :	E-mail: oliverwk@cec.sc.edu,	Telephone: (803) 777-8944
W.H.:	E-mail: weilin.hou@nrlssc.navy.mil,	Telephone: (228) 688-5257
S.W. :	E-mail: songwang@cec.sc.edu,	Telephone: (803) 777-2487

Ocean Sensing and Monitoring II, edited by Weilin (Will) Hou,
Robert A. Arnone, Proc. of SPIE Vol. 7678, 76780N · © 2010 SPIE
CCC code: 0277-786X/10/$18 · doi: 10.1117/12.852339

All of these feature detectors and descriptors only address invariance in the spatial domain. They are not invariant when the considered image undergoes a destructive intensity transformation, where the image intensity values change substantially, inconsistently and irreversibly. Such transformations often significantly increase the complexity in discerning any underlying features and structures in the image, as shown in.[11] A typical example is intensity transformation introduced by underwater imaging. Light behaves differently underwater.[14] The added complexities of impure water introduce issues such as turbulence, air bubbles, particles (such as sediments), and organic matter that can absorb and scatter light, which can result in a very blurry and noisy image. Since available feature descriptors are not invariant under such intensity transformations, matching the features detected from an underwater image and a clean out-of-water image, or the features detected from two underwater images taken in different underwater conditions, is a very challenging problem.

In this paper we investigate the performance of current high level detectors and descriptors in underwater images. We look at detectors based on corner and blob detection, Harris and Hessian respectively, and two well-known feature descriptors SIFT and Gradient Location and Orientation Histograms[11] (GLOH). We quantitatively look at both detector and descriptor performance independently and jointly using a measure of detection repeatability and matching precision and recall. The rest of the paper is organized as follows: Section 2 gives a brief overview of problems associated with underwater imaging and vision and the model used to simulate these effects. Section 3 briefly introduces the region detectors used in this study and Section 4 explains the region descriptors. Section 5 explains our approach to evaluating detector and descriptor performance. Section 6 presents our results.

2. UNDERWATER IMAGING

Underwater Imaging is an area with many applications including mine countermeasures, security, search and rescue, and conducting scientific experiments in harsh, unreachable environments. On a clear day, a person can see miles to the horizon out-of-water, but in many underwater conditions one cannot see more than a few meters, and what can be seen is blurred and difficult to discern. This reduction in visibility is due to the absorption and scattering of light by the water and particles in the water. There are numerous particles such as sediment, plankton, and organic cells in the water which cause light scattering and absorption. Even optical turbulence and bubbles effect how light is transmitted. Light that is spread out by this scattering is the source of the blurriness and fuzziness common in underwater images.

This absorption and scattering of light in water can be modeled mathematically, and much work has been done to develop robust models to this effect by Jaffe,[15] Dolin,[16] and Hou.[17,18] These models are typically some form of a *point spread function* (PSF) which models a system's response to an impulse signal (point source). For this work, we use a simplified version of Dolin's PSF model,[16,17] to simulate underwater conditions. Given an out-of-water image, convolution with the PSF creates a synthetic underwater image. Dolin's model takes the form

$$G(\theta_b, \tau_b) = \frac{\delta(\theta_q)}{\pi\theta_q}e^{-\tau_b} + 0.525\frac{\tau_b}{\theta_q}e^{(-2.6\theta_q^{0.7}-\tau_b)} + \frac{\beta_2^2}{2\pi}[2-(1+\tau_b)e^{-\tau_b}]e^{(-\beta_1(\beta_2\theta_q)^{\frac{1}{3}}-(\beta_2\theta_q)^2+\beta_3)} \tag{1}$$

$$\tag{2}$$

$$\beta_1 = \frac{6.857 - 1.5737\tau_b + 0.143\tau_b^2 - 6.027\cdot 10^{-3}\cdot\tau_b^3 + 1.069\cdot 10^{-4}\cdot\tau_b^4}{1 - 0.1869\tau_b + 1.97\cdot 10^{-2}\cdot\tau_b^2 - 1.317\cdot 10^{-3}\cdot\tau_b^3 + 4.31\cdot 10^{-5}\cdot\tau_b^4} \tag{3}$$

$$\beta_2 = \frac{0.469 - 7.41\cdot 10^{-2}\tau_b + 2.78\cdot 10^{-3}\tau_b^2 + 9.6\cdot 10^{-5}\cdot\tau_b^3}{1 - 9.16\cdot 10^{-2}\tau_b - 6.07\cdot 10^{-3}\cdot\tau_b^2 + 8.33\cdot 10^{-4}\cdot\tau_b^3} \tag{4}$$

$$\beta_3 = \frac{6.27 - 0.723\tau_b + 5.82\cdot 10^{-2}\cdot\tau_b^2}{1 - 0.072\tau_b + 6.3\cdot 10^{-3}\cdot\tau_b^2 + 9.4\cdot 10^{-4}\cdot\tau_b^3} \tag{5}$$

where, θ_q is the scattering angle—the angle at which light is refracted away from its original direction—and $\tau_b = \tau\omega$, where τ is the optical depth and ω is the single scattering albedo, the ratio of light scattered to the total light attenuation. More details can be found in.[16,17] In this paper we will use the notation $\mathrm{PSF}(\cdot, \tau, \omega)$ to refer to the operation of convolution with a PSF with parameters τ and ω.

3. REGION DETECTION

For our experiments we examined region detection schemes based on the Harris[19] detector along with its scale and affine invariant extensions, and the Hessian matrix along with its scale and affine invariant extensions. For more detailed information and an extensive study of current region detectors on various spatial transforms please refer to Tuytelaars.[20]

3.1 Interest Point Detectors

The Harris detector is based on the second moment matrix which describes local gradient information around a point. It is defined as

$$M = \sigma_D^2 \, \mathrm{g}(\sigma_I) * \begin{bmatrix} I_x^2(\mathbf{x}, \sigma_D) & I_x(\mathbf{x}, \sigma_D) I_y(\mathbf{x}, \sigma_D) \\ I_x(\mathbf{x}, \sigma_D) I_y(\mathbf{x}, \sigma_D) & I_x^2(\mathbf{x}, \sigma_D) \end{bmatrix}. \tag{6}$$

The image's local derivatives are estimated with a Gaussian kernel with scale σ_D, and the derivatives are smoothed over the neighborhood with a Gaussian of scale σ_I.

$$I_x(\mathbf{x}, \sigma) = \frac{\partial}{\partial x} \mathrm{g}(\sigma) * I(\mathbf{x}) \tag{7}$$

$$\mathrm{g}(\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{8}$$

Cornerness is then measured as the difference of the determinant and the trace:

$$\det(M) - \lambda \, \mathrm{trace}(M) \tag{9}$$

As an interest point detector, non-maximum suppression is used to extract local corner maxima.

The Hessian is the second matrix issued from the Taylor expansion of the image intensity function:

$$H = \begin{bmatrix} I_{xx}(\mathbf{x}, \sigma_D) & I_{xy}(\mathbf{x}, \sigma_D) \\ I_{xy}(\mathbf{x}, \sigma_D) & I_{yy}(\mathbf{x}, \sigma_D) \end{bmatrix} \tag{10}$$

with

$$I_{xy}(\mathbf{x}, \sigma) = \frac{\partial}{\partial y} \frac{\partial}{\partial x} \mathrm{g}(\sigma) * I(\mathbf{x}). \tag{11}$$

Local maxima of the trace and determinant give strong responses to blob- and ridge-like structures, with the trace being the Laplacian. One approach to obtain more stable respones is to find points which achieve a local maximum of the determinant and trace simultaneously.

3.2 Scale and Affine Invariant Detectors

For this paper we look at two extensions of the above detectors, the Harris-Laplace and Hessian-Laplace, which are scale-invariant. Harris-Affine and Hessian-Affine are the affine-invariant extensions.[20,21] The Harris-Laplace/Hessian-Laplace detector uses a multiscale Harris or Hessian detector to locate local features. Scale selection is based on the idea proposed by Lindeberg,[22] where the characteristic scale of a local structure is determined by searching for extrema of a function in scale-space, which is the convolution of the function with Gaussian kernels of various sizes.

The Affine extension[20,21] applies an iterative process to points detected by the Harris/Hessian-Laplace detectors to estimate elliptical affine regions proposed by Lindeberg:[23]

1. Initial region detection with Harris/Hessian-Laplace
2. Use second moment matrix to estimate the shape
3. Normalize affine region to a circular region
4. Re-detect next location and scale on normalized image
5. Repeat from 2 if eigenvalues of the second moment matrix are not equal

4. REGION DESCRIPTION

SIFT is a very well-known and popular choice of image descriptor. SIFT and its extensions have bccn shown to perform better than other descriptors in comparison studies.[11] For these reasons we chose to focus on the performance of SIFT and one of its extensions, GLOH.

4.1 Scale-Invariant Feature Transform

SIFT[7,10] features are rotation, scale, and translation invariant, and have been shown to bc robust against some lighting and viewpoint transformations. There are four stages associated with SIFT: (1) scale-spacc extrema detection, (2) region keypoint localization, (3) orientation assignment, and (4) region descriptor gcncration. For this work, we are only interested in the latter two stages (namely, SIFT's ability as a descriptor) so we replace thc first two stages with the Harris- and Hessian-based detectors. SIFT assigns orientation by building a histogram of gradient-magnitude-weighted gradient orientations. The gradients are computed ovcr the region at the selected region scale. Peaks in the histogram are detccted by selecting the highest bin and any other bin with 80% of the highest. In this manner, a single region can yield multiple detectors. The region descriptor is then represented relative to the location, characteristic scale, and dominant orientation(s) of the region. To get the descriptor thc region gradient magnitude and orientations are again calculated, relative to the dominant orientation, at thc sclected scale. The descriptor is a $4x4$ array of 8 bins each, organized by gradient orientation, cach bcing a gradient-magnitude-weighted histograms yielding a 128 dimcnsional descriptor vector.

4.2 Gradient Localization and Orientation Histogram

The GLOH[11] descriptor is an extension of SIFT using radial histogram binning and PCA to reducc the descriptor dimension. The SIFT descriptor is computed for a log-polar location grid with 3 radial and 8 angular bins resulting in 17 location bins. Gradient orientations are then quantized into 16 bins resulting in a 272 bin histogram. This is reduced from 272 to 128 with PCA using the 128 largest eigenvectors.

5. BENCHMARK

To test the performance of the various detection and description schemes, we use the same benchmark as Mikolajczyk[11] and Tuytelaars.[20] The authors in these works examined the performance of detectors and descriptors for a range of geometric and photometric transforms. They measured performance of the detectors and the region selection consistency. The descriptors were then matched using distance thresholding, nearest neighbor, and ratio of nearest neighbor and second nearest neighbor schemes. These schemes were then rated based on the number of correct descriptor matches.

Region detection performance is based on the repeatability metric—that the same regions, under transform, are found in the original and transformed images. For example, if T is a geometric transform and I, J are images such that $J = T(I)$ and $R \subset I$ and $S \subset J$ are regions such that $S = T(r)$ then S repeats R, and R, S both cover the same scene area in their respective images.

To measure the repeatability, the benchmark requires that the homography transform between two images be known. A region $R \subset I$ and a region $S \subset J$ correspond if the projection of S onto I, $S^{'} = T(I)$ and R have small overlap error, i.e. $1 - \frac{R \cap S^{'}}{r \cup S^{'}} < \delta$. Since, for our purposes, we are only interested in photometric transforms, T is the identity, and R and S are compared directly. Repeatability is then measured as the ratio of the number of corresponding regions to the number of regions in I. For regions that have multiple correspondences, thc onc with the least overlap error is chosen.

To measure descriptor performance, we look at the precision and recall for a matching between two image's rcgion dcscriptors. Precision gives an indication of how well a set of matches is with respect to itself, while recall is a global measure of the matches. These measures are given by:

$$precision = \frac{\text{\#of correct matches found}}{\text{\# of matches found}}, \text{ and } recall = \frac{\text{\# of correct matches found}}{\#of correctmatches}. \qquad (12)$$

These measures don't provide much information on their own. For example, it is possible to have a set of matches which are all correct (high precision) but fail to find very many of the possible matches (low recall). In this case, looking only at the precision the would give false confidence in the quality of the matching, which is why it is common to look at precision given a certain recall. The F-score, which combines precision and recall, is a good overall measure and is given by the harmonic mean:

$$F\text{-}score = \frac{2 * precision * recall}{precision + recall}. \tag{13}$$

To calculate the precision, recall, and F-scores, a ground truth matching is needed. This is obtained from the correspondences determined from the region overlap error, which was used in the repeatability performance. Two descriptors should be matched if their regions have small overlap error. For matching descriptors we looked at three different matching techniques. The first is a simple threshold matchingi. Given two regions R, S and their respective descriptors $\mathbf{r}, \mathbf{s}$, R and S are matched according to the following criteria

$$\text{match}_t(R, S) = \begin{cases} 1 & \|\mathbf{r} - \mathbf{s}\|_2 \leq t \\ 0 & \|\mathbf{r} - \mathbf{s}\|_2 > t. \end{cases} \tag{14}$$

This approach to matching has the added difficulty that a good threshold must be chosen to obtain accurate matchings. Also, under this approach, a region can have multiple matchings.

Another approach is to match based on the nearest neighbor (NN). Let $R^1 \ldots R^N$ be the set of regions detected from an image and S a region detected from another image. These regions have corresponding descriptors denoted $\mathbf{r}^1 \ldots \mathbf{r}^N$, and $\mathbf{s}$. The nearest neighbor matching is then defined as

$$\text{match}_{\text{NN}}(R^k, S) = \begin{cases} 1 & \text{if } k = \underset{i=1,\ldots,N}{\text{argmin}} \|\mathbf{r}^i - \mathbf{s}\|_2 \\ 0 & \text{if } k \neq \underset{i=1,\ldots,N}{\text{argmin}} \|\mathbf{r}^i - \mathbf{s}\|_2 \end{cases} \tag{15}$$

The third approach attempts to address problems with the nearest neighbor method. Given $R^1 \ldots R^N$, S and their descriptors $\mathbf{r}^1 \ldots \mathbf{r}^N, \mathbf{s}$, by definition $\mathbf{s}$ will always have a nearest neighbor in $\mathbf{r}^1 \ldots \mathbf{r}^N$, but the descriptors may still be distant from each other resulting in a noisy match. The ratio of nearest neighbors (RNN) approach works under the assumption that if a NN match is noisy then the distance between the nearest neighbor and second nearest neighbor should both be relatively large. Whereas a good nearest neighbor should be sufficiently closer than the second nearest neighbor. This matching criteria is formulated as

$$\text{Define } l(k) = \underset{i=1,\ldots,N; i\neq k}{\text{argmin}} \|\mathbf{r}^i - \mathbf{s}\|_2 \tag{16}$$

$$\text{match}_{\text{RNN}}(R^k, S) = \begin{cases} 1 & \text{if } k = \underset{i=1,\ldots,N}{\text{argmin}} \|\mathbf{r}^i - \mathbf{s}\|_2, \text{ and } \frac{\|\mathbf{r}^k - \mathbf{s}\|_2}{\|\mathbf{r}^{l(k)} - \mathbf{s}\|_2} < t \\ 0 & \text{if } k \neq \underset{i=1,\ldots,N}{\text{argmin}} \|\mathbf{r}^i - \mathbf{s}\|_2 \text{ or } k = \underset{i=1,\ldots,N}{\text{argmin}} \|\mathbf{r}^i - \mathbf{s}\|_2 \text{ and } \frac{\|\mathbf{r}^k - \mathbf{s}\|_2}{\|\mathbf{r}^{l(k)} - \mathbf{s}\|_2} \geq t \end{cases} \tag{17}$$

6. RESULTS

To test the performance of the detectors and descriptors we captured an original out-of-water image and apply different PSFs to simulate different underwater conditions. Using each region detector, regions are detected for the original and each PSF-convoluted image. The regions are then evaluated by the benchmark. In the following figures $\tau_b = \tau\omega$ where ω is the optical depth and τ is the scatter-absorption ratio. These parameters are used to generate PSFs from Dolin's model as described in Section 2. For all of our tests we use two images; the first is a stuffed teddy bear chosen because of its textured fur, the second is a Secchi disk which is a well-known tool for measuring visibility.

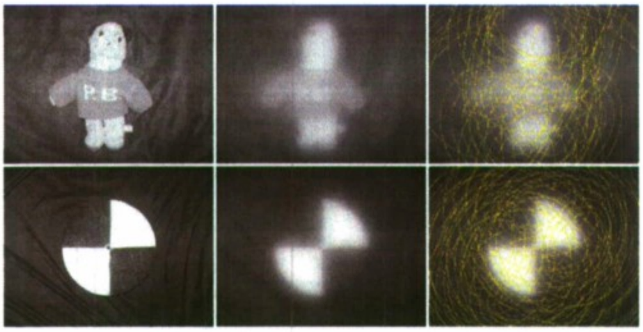
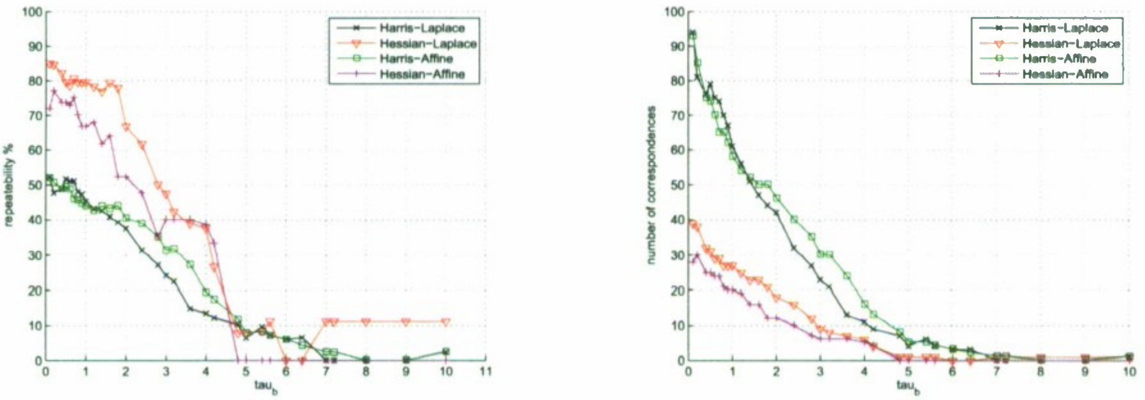


Figure 1. The two test images, a bear and a Secchi disk, images convoluted with $PSF(\cdot, \tau = 1, \omega = 1.0)$, the convoluted images with detected regions.

6.1 Region Detection Performance

The repeatability performance for the different region detectors is shown in Figure 2(a) and Figure 3(a) and the raw number of correspondences in Figure 2(b) and Figure 3(b) for the bear and Secchi disk images respectively. These figures show that the rate of detecting repeatable regions drops off very quickly as water conditions become worse, for all detectors.

In terms of pure repeatability the Hessian-based detectors clearly outperform the Harris-based detectors on both test images. This outcome seems to agree with the intuition that a blob detector would be more robust against blurring, whereas corners, which are finer details, would be obscured in more turbid water. In terms of the raw number of correspondences on the bear image, the Harris-based dectors perform better, which is to be expected since the bear has textures from the fur and sweater to respond to the corner detection. When looking at a more structured scene such as the Secchi disk, this advantage disappears. Overall, the choice seems to depend on the application or water conditions encountered, however all methods fail as water clarity decreases.



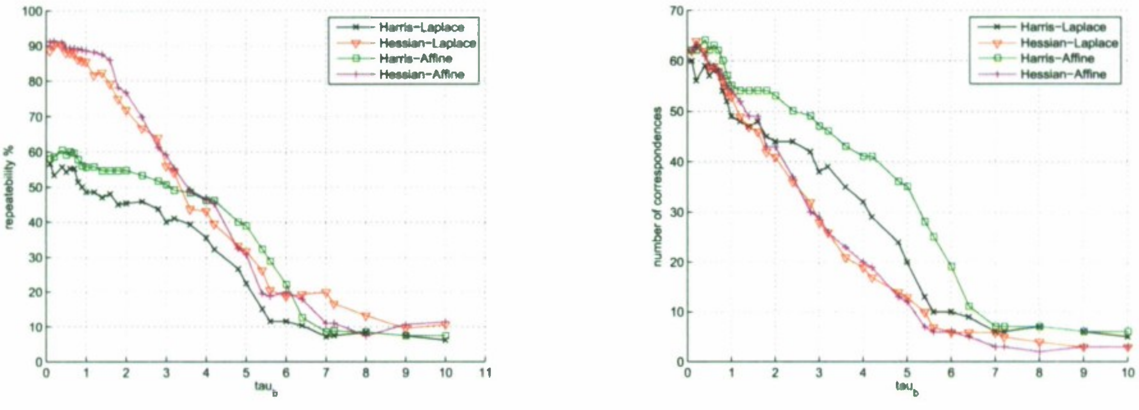
(a) Repeatability rates for the region detectors across a range of underwater conditions.

(b) Number of region correspondences based on region overlap error, across a range of underwater conditions.

Figure 2. Repeatability and raw number of correspondences for the bear image.

6.2 Region Description Performance

To test the descriptor performance we conduction two experiments. First descriptors are built from the detected regions in each image; however, as shown in Section 6.1, repeatability of regions falls off very quickly as image conditions worsen. While this gives a more accurate picture of overall performance, we would like to also isolate the descriptors to have an idea of their performance alone. To accomplish this we assume that the region detectors have 100% repeatability. Regions are detected on the original image then, for the different PSF convoluted images,



(a) Repeatability rates for the region detectors across a range of underwater conditions.

(b) Number of region correspondences based on region overlap error, across a range of underwater conditions.

Figure 3. Repeatability and raw number of correspondences for the Secchi disk image.

descriptors are built for the regions detected in the original. Since our only transformation between the images is photometric (PSF convolution) and no spatial transformations are introduced this simulates our assumption of a 100% repeatable detector. In general, if spatial transforms were introduced, this assumption could still be made but regions from the original would need to be transformed using the homography and then descriptors built on them.

Figures 4, 5, 6, 7 show the matching performance for Nearest Neighbor, Threshold, and Ratio of Nearest Neighbors matching. The Nearest Neighbor and Threshold curves were genereated by thresholding matches based on descriptor distances of the matching while the Ratio of Nearest Neighbors was thresholded by ratios 1 to 1.5 by 0.05, 1.8 to 3.4 by 0.2. The maxmimum F-score achieved is shown in the legend. The point on the curve where the maximum F-score is achieved is denoted by the large marker.

Figures 4, 6 show the descriptor performance with a 100% repeatable detector for SIFT (left side) and GLOH(right side) of the stuffed bear and Secchi disk images respectively. It is evident that for all instances, the descriptors built on the Hessian-based regions perform much better than the descriptors on Harris regions. However, they still perform fairly poorly overall, with a trend for hitting a wall in recall score with NN matching, with the best performance hitting this wall around 0.5. This best performance is achieved by Hessian-Affine on an image with $\tau_b = 0.1$, which is the clearest test parameter, yet we still only get a performance yielding half of the total correct matches.

For the experiements where the real detected regions are used, the overall conclusion is the same, though here the Hessian-based regions have more separation from the Harris-based regions for the stuffed bear. They are noticibly better on the Secchi disk as well, though the separation from Harris-base regions is not as distinct. They still appear to reach a limit aroun $0.5 - 0.6$ recall for the stuffed bear and $0.3 - 0.4$ for the Secchi disk.

7. CONCLUSION

This work asses what problems need to be addressed in the area of underwater feature detection, description and matching in order to use computer vision techniques for object detection and recognition in underwater environments. Our results show that all three components have major limitations when dealing with photometric transformations introduced when imaging underwater. The Hessian based detectors performed best, though their performance is not great and trails off quickly as the water gets murkier. The descriptors do not perform any better, with recalls consistently below 0.5 on regions which are already few in number. While more robust and photometric invariant descriptors are needed, the problem might also be approached by developing novel matching techniques which take advantage of, or see through, the murkiness of the features.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Office of Naval Research and the U.S. Naval Research Laboratory under Base Program PE 62782N.

REFERENCES

[1] Brown, M., Hartley, R., and Nister, D., "Minimal solutions for panoramic stitching," in [*IEEE Confcrence on Computer Vision and Pattern Recognition*], 1–8 (2007).

[2] Jin, H., "A three-point minimal solution for panoramic stitching with lens distortion," in [*IEEE Conferencc on Computer Vision and Pattern Recognition*], 1–8 (2008).

[3] Hess, R. and Fern, A., "Improved video registration using non-distinctive local image features," in [*IEEE Conference on Computer Vision and Pattern Recognition*], 1–8 (2007).

[4] Medioni, G., "Retinal image registration from 2d to 3d," in [*IEEE Conference on Computer Vision and Pattern Recognition*], 1–8 (2008).

[5] Torralba, A., Murphy, K., and Freeman, W., "Sharing visual features for multiclass and multiview object detection," *IEEE Transactions on Pattern Analysis and Machine Intclligence* **29**, 854–869 (May 2007).

[6] Lampert, C., Blaschko, M., and Hofmann, T., "Efficient subwindow search: A branch and bound framework for object localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**, 2129–2142 (December 2009).

[7] Lowe, D., "Object recognition from local scale-invariant features," in [*International Conference on Computer Vision*], 1150–1157 (1999).

[8] Ling, H. and Jacobs, D., "Deformation invariant feature matching," in [*International Conference on Computer Vision*], 1466–1473 (2005).

[9] Cheng, H., Liu, Z., Zheng, N., and Yang, J., "A deformable local image descriptor," in [*IEEE Conferencc on Computer Vision and Pattern Recognition*], 1–8 (2008).

[10] Lowe, D., "Distinctive image features from scale-invariant keypoints," *Intcrnational Journal on Computer Vision* **60**(2), 91–110 (2004).

[11] Mikolajczyk, K. and Schmid, C., "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 1615–1630 (October 2005).

[12] Ke, Y. and Sukthankar, R., "Pca-sift: A more distinctive representation for local image descriptors," in [*IEEE Conference on Computer Vision and Pattern Recognition*], 506–513 (2004).

[13] Mortensen, E. N., Hongli, D., and Shapiro, L., "A sift descriptor with global context," in [*IEEE Confcrencc on Computer Vision and Pattern Recognition*], 184–190 (2005).

[14] Mobley, C., [*Light and Water: Radiative Transfer in Natural Waters*], Academic Press (1994).

[15] Jaffe, J., "Monte carlo modeling of underwater-image formation: Validity of the linear and small-angle approximations," *Applied Optics* **34**(24), 5413–5421 (1995).

[16] Dolin, L., Gilbert, G., Levin, I., and Luchinin, A., [*Theory of imaging Through Wavy Sca Surface*], Russian Academy of Sciences, Institute of Applied Physics, Nizhniy Novgorod (2006).

[17] Hou, W., Gray, D., Weidemann, A., and Arnone, R., "Comparison and validation of point spread models for imaging in natural waters," *Optics Express* **16**(13) (2008).

[18] Hou, W., "A simple underwater imaging model," *Optics Express* **34**(17) (2009).

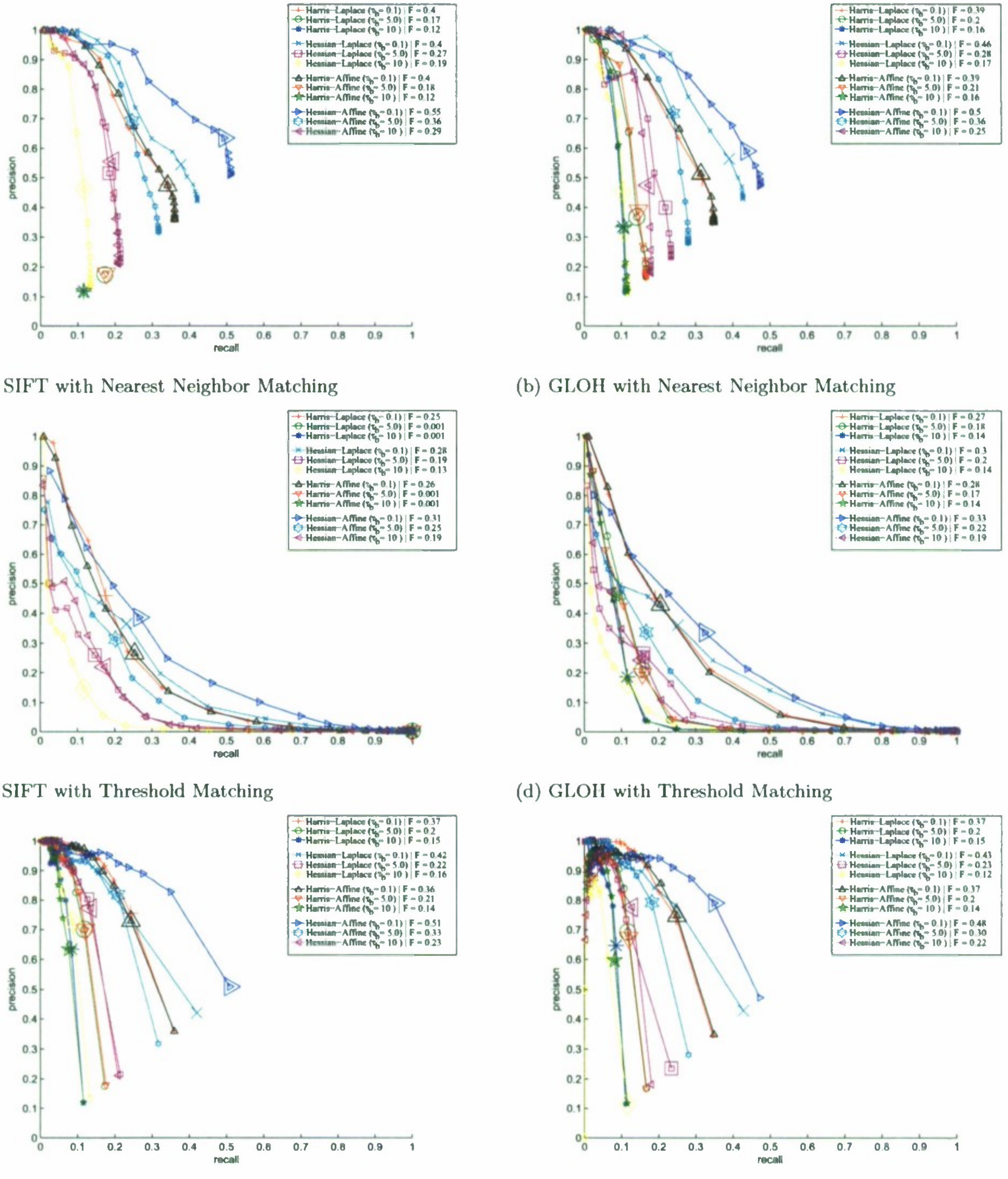
[19] Harris, C. and Stephens, M., "A combined corner and edge detector," in [*Alvey Vision Conference*], 147–151 (1988).

[20] Tuytelaars, T. and Mikolajczyk, K., "Local invariant feature detectors: A survey," *Computer Graphics and Vision* **3**(3), 177–280 (2008).

[21] Mikolajczyk, K. and Schmid, C., "Scale and affine invariant interest point detectors," *International Journal on Computer Vision* **60**(1), 63–86 (2004).

[22] Lindeberg, T., "Feature detection with automatic scale selection," *International Journal on Computer Vision* **30**(2), 79–116 (1998).

[23] Lindeberg, T., "Direct estimation of affine image deformations using visual front-end operations with automatic scale selection," in [*International Conference on Computer Vision*], 134–141 (1995).



(a) SIFT with Nearest Neighbor Matching

(b) GLOH with Nearest Neighbor Matching

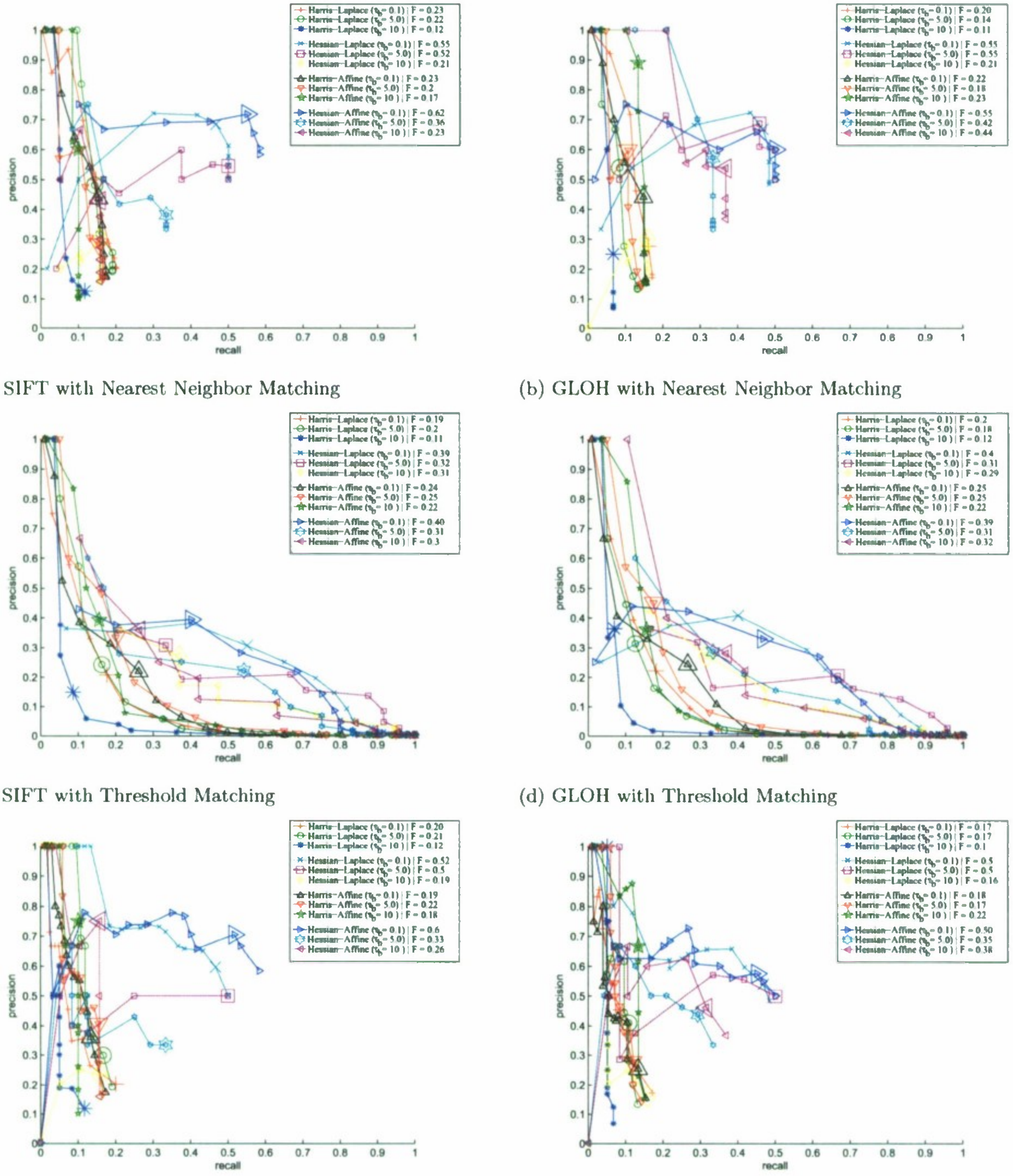
(c) SIFT with Threshold Matching

(d) GLOH with Threshold Matching

(e) SIFT with Ratio of Nearest Neighbors Matching

(f) GLOH with Ratio of Nearest Neighbors Matching

Figure 4. Descriptor performance for the bear image with simulated 100% repeatability.



(a) SIFT with Nearest Neighbor Matching

(b) GLOH with Nearest Neighbor Matching

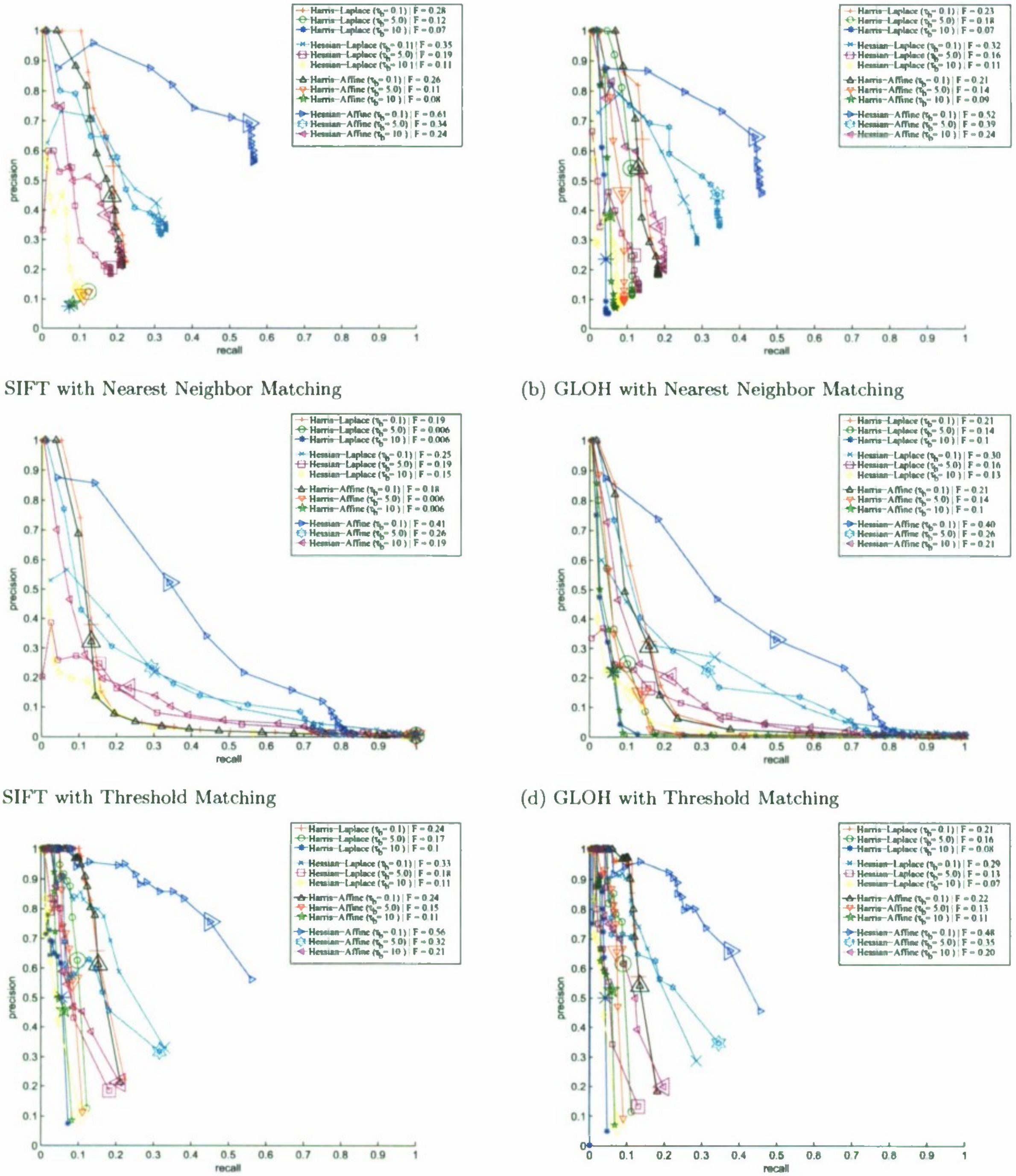
(c) SIFT with Threshold Matching

(d) GLOH with Threshold Matching

(e) SIFT with Ratio of Nearest Neighbors Matching

(f) GLOH with Ratio of Nearest Neighbors Matching

Figure 5. Descriptor performance for the bear image with actual repeatability.



(a) SIFT with Nearest Neighbor Matching

(b) GLOH with Nearest Neighbor Matching

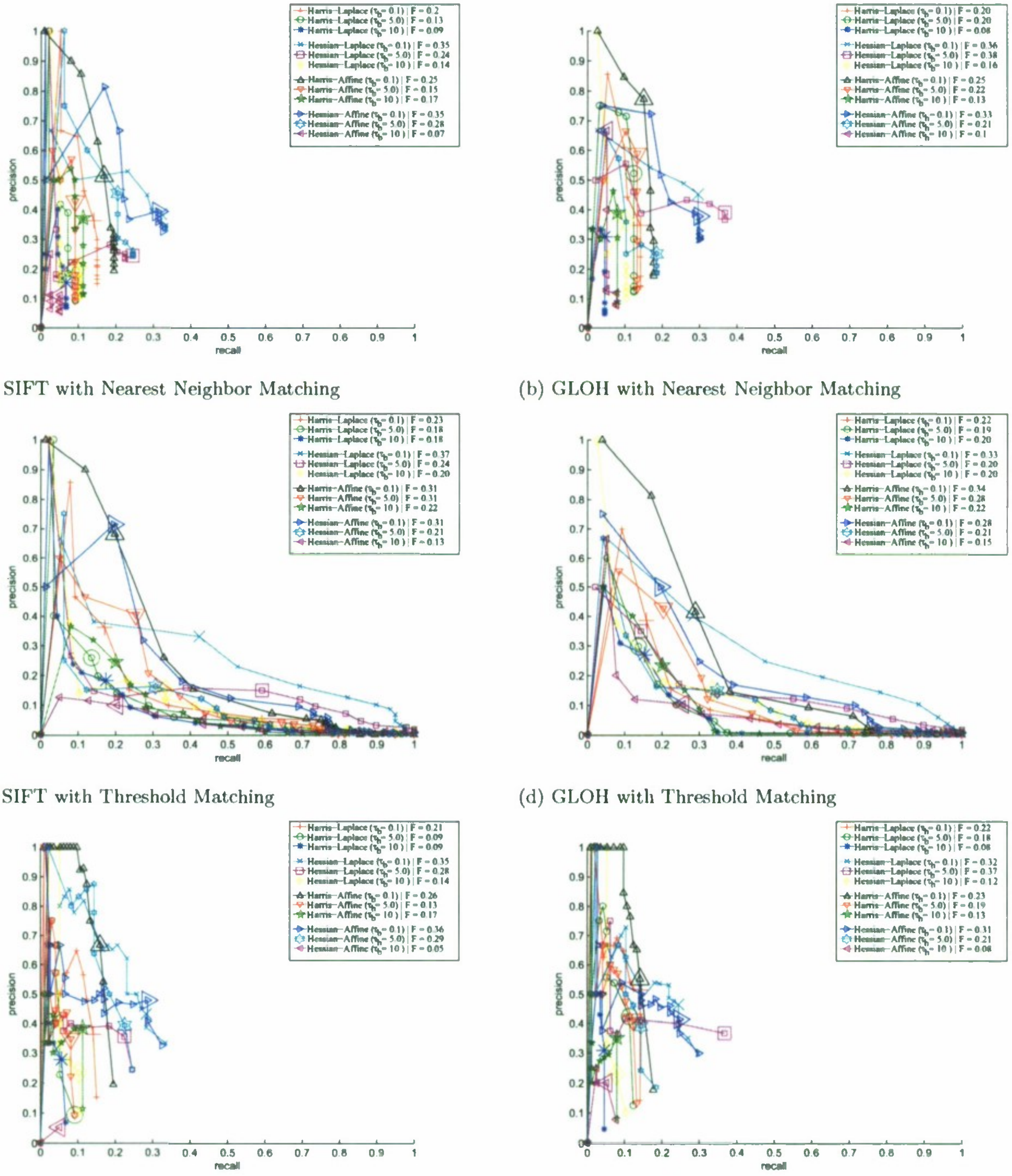
(c) SIFT with Threshold Matching

(d) GLOH with Threshold Matching

(e) SIFT with Ratio of Nearest Neighbors Matching

(f) GLOH with Ratio of Nearest Neighbors Matching

Figure 6. Descriptor performance for the Secchi disk image with simulated 100% repeatability.



(a) SIFT with Nearest Neighbor Matching

(b) GLOH with Nearest Neighbor Matching

(c) SIFT with Threshold Matching

(d) GLOH with Threshold Matching

(e) SIFT with Ratio of Nearest Neighbors Matching

(f) GLOH with Ratio of Nearest Neighbors Matching

Figure 7. Descriptor performance for the Secchi disk image with actual repeatability.