Feature Tracking of Objects in Underwater Video Sequences

Prabhakar C J & Praveen Kumar P U
Department of P.G. Studies and Research in Computer Science
Kuvempu University, Shankaraghatta - 577451
Karnataka, India
psajjan@yahoo.com, praveen577302@gmail.com

***Abstract*—Feature tracking is a key, underlying component in many approaches to 3D reconstruction, detection, localization and recognition of underwater objects. In this paper, we proposed to adapt SIFT technique for feature tracking in underwater video sequences. Over the past few years the underwater vision is attracting researchers to investigate suitable feature tracking techniques for underwater applications. The researchers have developed many feature tracking techniques such as KLT, SIFT, SURF etc., to track the features in video sequence for general applications. The literature survey reveals that there is no standard feature tracker suitable for underwater environment. We proposed to adapt SIFT technique for tracking features of objects in underwater video sequence. The SIFT extracts features, which are invariant to scale, rotation and affine transformations. We have compared and evaluated SIFT with popular techniques such as KLT and SURF on captured video sequence of underwater objects. The experimental results shows that adapted SIFT works well for underwater video sequence.**

***Index Terms*—Features tracking, SIFT, 3D reconstruction, Underwater Video**

I. Introduction

Feature tracking is an important step in 3D surface reconstruction of underwater objects using video sequences. The researchers have developed many techniques for feature tracking, as many algorithms rely on accurate computation of correspondences through a sequence of images. These techniques works well for clean, out-of-water images; however, when imaging underwater, even an image of the same object can be drastically different due to varying water conditions. Moreover many parameters can modify the optical properties of the water and underwater image sequences show large spatial and temporal variations. As a result, descriptors of the same point on an object may be completely different between different underwater video images taken under varying imaging conditions. This makes feature tracking between such images is a very challenging problem.

The 3D reconstruction of underwater objects such as underwater mines, ship wreckage, coral reefs, pipes and telecommunication cables is an essential requirement for recognition, localization of these objects using video sequences. Feature tracking is applied to recover the 3D structure and motion of objects in video sequence. To recover the 3D structure and motion, it is required to find out the same features in different frames and match them. The efficiency of the feature tracking techniques depends upon the invariant nature of the extracted features. Further the extracted distinctive invariant features can be used to perform reliable matching between different views of an object or scene. The features extracted should be invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. The features are highly distinctive, in the sense that a single feature can be correctly matched with high probability against a large database of features from many images. In underwater environment, extracting the invariant features from the sequence of images is a challenging due to optical properties of the water, which varies the same feature in sequence of images.

3D reconstruction for underwater applications is a relatively recent research area with a higher complexity than the 3D reconstruction for general applications. This is a very active and growing field of research that brings together challenging problems from the underwater environment and powerful 3D reconstruction techniques from computer vision. Many articles have been published during the last years involving feature tracking from video sequences for 3D reconstruction of underwater objects demonstrates the increasing interest in this topic and its wide range of applications. E. Trucco et al. [1] have developed a robust tracker based on efficient outlier rejection rule suitable for subsea video sequences. They extended Shi-Tomasi-Kanade tracker in several ways: first by introducing an automatic scheme for rejecting bad features, using a simple, efficient, model-free outlier rejection rule, called X84. Secondly, they developed a system for running indefinite length of time and the tracker runs in real time on non-dedicated hardware. The Shi-Tomasi-Kanade [10][9] feature tracker is based on Sum-of-Squared-Difference (SSD) matching and assuming affine frame-to-frame warping. The system tracks small image regions and classifies them as good (reliable) or bad (unreliable) according to the residual of the match between the associated image regions in the first and current frames. K. Plakas [2] have adapted Shi-Tomasi-Kanade as a feature tracker for 3D shape reconstruction from uncalibrated underwater video sequences.

Anne Sedlazeck et al. [3] have adapted KLT (Kanade-Lucas-Tomasi) feature tracker to reconstruct 3D surface of ship wreck using underwater monocular video sequence.

DOI: 02.CEM.2011.01.529

Andrew Hogue and Michael Jenkin [4] also adapted KLT as a feature tracker for 3D shape reconstruction of coral reefs using stereo image sequences. The KLT tracker is a good solution for tracking problem in the case where the input image sequence consists of dense video i.e. the displacement between two consecutive images is very small. The main drawback of KLT is that it cannot be applied for stereo image sequences with large baseline. The 3D reconstruction of coral reefs is developed by V. Brandou et al. [5] using Scale-Invariant Feature Transform (SIFT) technique [6] for extracting and matching features in stereo video sequences. They have captured coral reefs using two video cameras, which are aligned to capture stereo video sequences. P Jasiobedzki et al., [8] have adapted SIFT feature tracker for extraction of feature from images and recognizing the object in subsequent images.

In this paper, we proposed to adapt SIFT method for extraction of features from underwater monocular video sequences. SIFT is an algorithm in computer vision to detect and describe local features in images. The SIFT method is very suitable in the case where the interest points are invariant to image scaling and rotation, and partially invariant to changes in illumination and 3D camera viewpoints. For every keypoint a descriptor is defined, which is a vector with 128 dimensions based on local image gradient directions in the keypoint neighborhood. SIFT is more robust than other feature tracking techniques for images with large translation and rotation it still can track the features.

The remaining sections of the paper are organized as follows: section II describes adapted SIFT feature tracker in detail. The experimental results are presented in the section III. Finally, the section IV concludes the paper.

II. SIFT

This algorithm is first proposed by David Lowe in 1999, and then further developed and improved [7]. SIFT features have many advantages. SIFT features are all natural features of images. They are favorably invariant to image translation, scaling, rotation, illumination, viewpoint, noise etc. Good speciality, rich in information, suitable for fast and exact matching in a mass of feature database. Relatively fast speed. The extraction SIFT features from video sequences can be done by applying sequentially steps such as Scale-space extrema detection, Keypoint localization, Orientation assignment and finally Generation of keypoint descriptors.

A. Scale-space extrema detection

The first stage of computation searches over all scales and image locations. It is implemented efficiently by means of a Difference-of-Gaussian function to identify potential interest points that are invariant to orientation and scale. Interest points for SIFT features correspond to local extrema of Difference-of-Gaussian filters at different scales. Given a Gaussian-blurred image described as the formula

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \qquad (1)$$

Where $*$ is the convolution operation in x and y, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}.$$

To efficiently detect stable keypoint locations in scale space extrema in the difference-of-Gaussian function convolved with the image, $D(x, y, \sigma)$, which can be computed from the difference of two nearby scales separated by a constant multiplicative factor k:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma). \qquad (2) \end{aligned}$$

which is just be different from the Gaussian-blurred images at scales σ and $k\sigma$.

The first step toward the detection of interest points is the convolution of the image with Gaussian filters at different scales, and the generation of difference-of-Gaussian images from the difference of adjacent blurred images. The rotated images are grouped by octave (an octave corresponds to doubling the value of σ), and the value of k is selected so that we can obtain a fixed number of blurred images per octave. This also ensures that we obtain the same figure of Difference-of-Gaussian images per octave (Fig. 1).

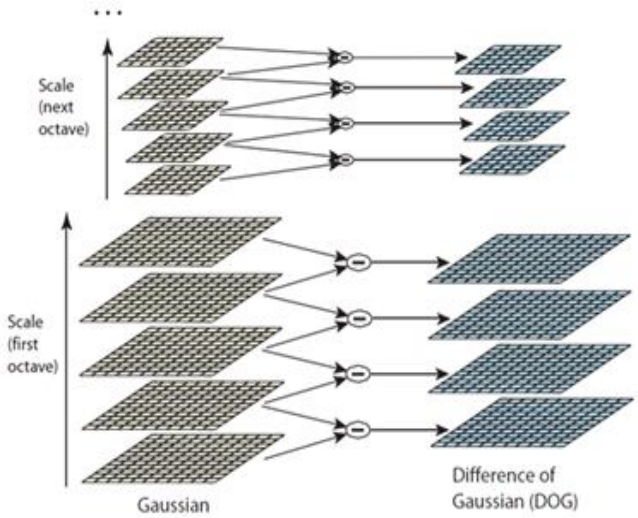

Figure 1. The blurred images at different scales, and the computation of the difference-of-Gaussian images

Interest points (called keypoints in the SIFT framework) are identified as local maxima or minima of the DoG images across scales. Each pixel in the DoG images is compared to its 8 neighbors at the same scale, plus the 9 corresponding neighbors at neighboring scales. If the pixel is a local maximum or minimum, it is selected as a candidate keypoint (Fig. 2).
For each candidate keypoint:
1) Interpolation of nearby data is used to accurately determine its position;
2) Keypoints with low contrast are removed;
3) Responses along edges are eliminated;
4) The keypoint is assigned an orientation.

DOI: 02.CEM.2011.01.529

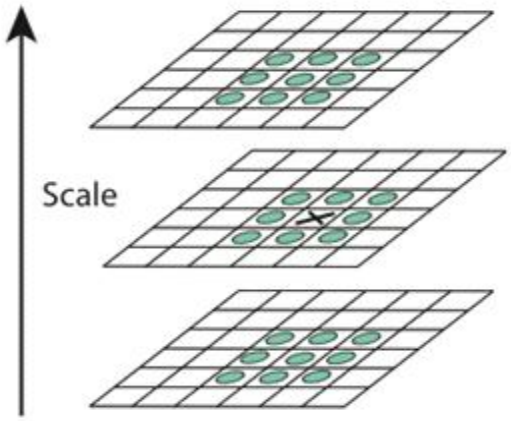

Figure 2. Local extrema detection, the pixel marked X is compared against its 26 neighbors in a $3\times3\times3$ neighborhood that spans adjacent DoG images

To determine the keypoint orientation, a gradient orientation histogram is computed in the neighborhood of the keypoint (using the Gaussian image at the closest scale to the keypoint's scale). The contribution of each neighboring pixel is weighted by the gradient magnitude and a Gaussian window with a σ that is 1.5 times the scale of the keypoint. Peaks in the histogram correspond to dominant orientations. A separate keypoint is rerated for the direction corresponding to the histogram maximum, and any other direction within 80% of the maximum value. All the properties of the keypoint are measured relative to the keypoint orientation, this provides invariance to rotation.

B. Locating Keypoints

At each candidate location, a detailed model is fit to determine scale and location. Keypoints are selected on basis of measures of their stability. The Fig. 3 shows a whole process on how to find and describe the SIFT feature points. In Fig. 3, we can find that, if we want to find and describe the SIFT feature points, we should follow these steps:

1) Input an image ranges from [0, 1].

2) Use a variable-scale Gaussian kernel $G(x, y, \sigma)$ to create scale space $L(x, y, \sigma)$.

3) Calculate Difference-of-Gaussian function as an approximate to the normalized Laplacian is invariant to the scale change.

4) Find the maxima or minima of Difference-of-Gaussian function value by comparing one of the pixels to its above, current and below scales in 3×3 regions.

5) Accurate the keypoints locations by discarding points below a predetermined value.

$$D(\hat{X}) = D + \frac{1}{2}\frac{\partial D^T}{\partial x}\hat{X}. \qquad (3)$$

Where $\hat{X}$ is calculated by setting the derivative $D(x, y, \sigma)$ to zero.

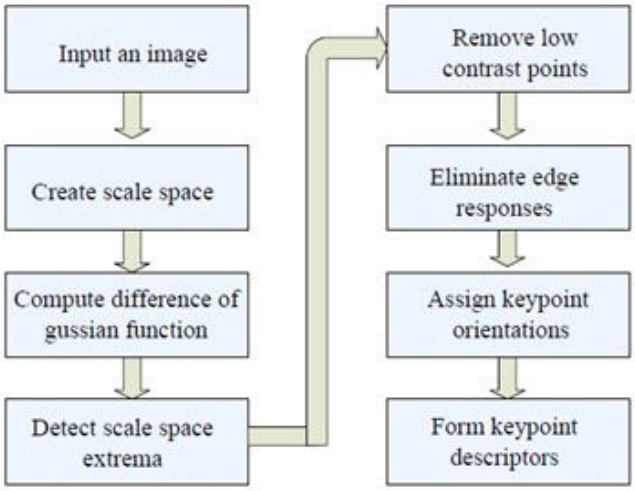

Figure 3. The diagram of keypoints location process

6) The extremas of Difference-of-Gaussian have large principal curvatures along edges, it can be reduced by checking

$$\frac{Tr(H)^2}{Det(H)} \langle \frac{(r+1)^2}{r}. \qquad (4)$$

Evidently, H in (4) is a 2×2 Hessian matrix, r is the ratio between the largest magnitude and the smallest one.

7) To achieve invariance to rotation, the gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ are pre-computed as the following equations.

$$m(x, y) = [(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2]^{1/2} \qquad (5)$$

$$\theta(x, y) = \tan^{-1}\left(\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}\right) \qquad (6)$$

8) Take a feature point and its 16×16 neighbors round it. Then divide them into 4×4 subregions, histogram every subregion with 8 bins.

C. SIFT feature representation

Once a keypoint orientation has been selected, the feature descriptor is computed as a set of orientation histograms on 4×4 pixel neighborhoods. The orientation histograms are relative to the keypoint orientation, the orientation data comes from the Gaussian image closest in scale to the keypoint's scale. The contribution of each pixel is weighted by the gradient magnitude, and by a Gaussian with σ 1.5 times the scale of the keypoint. Histograms contain 8 bins each, and each descriptor contains an array of 4 histograms around the keypoint (Fig. 4). This leads to a SIFT feature vector with $4\times4\times4=128$ elements. This vector is normalized to enhance invariance to changes in illumination.

DOI: 02.CEM.2011.01.529

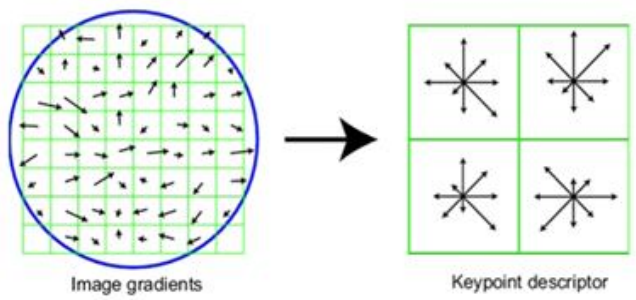

Figure 4. The keypoint descriptor is generated and weighted by a Gaussian window (blue circle)

D. Orientation assignment

Direction parameters to the keypoints are determined to quantize the description. Lowe formulated the determination with the norm and the angle in Euclidean space, with the direction of key points used as normalized the gradient direction of the key point operator in the following step. After an image revolvement, the identical directions demanded can be worked out.

E. Keypoint matching

The next step is to apply these SIFT methods to video frame sequences for object tracking. SIFT features are extracted through the input video frame sequences and stored by their keypoints descriptors. Each key point assigns 4 parameters, which are 2D location (x coordinate and y coordinate), orientation and scale. Each object is tracked in a new video frame sequences by separately comparing each feature point found from the new video frame sequences to those on the target object. The Euclidean distance is introduced as a similarity measurement of feature characters. The candidates can be preserved when the two features Euclidean distance is larger than the threshold specified previous. So the best matches can be picked out by the parameters value, in the other way, consistency of their location, orientation and scale. Each cluster of three or more features that agree on an object and its pose is then subject to further detailed model verification and subsequently outliers are throwed away. Finally the probability that a particular set of features indicates the presence of an object is computed, considering the accuracy of fit and number of probable false matches. Object matches that pass all of the above tests can be recognized as correct with high confidence.

III. Experimental results

The adapted SIFT approach is compared with other popular feature tracking techniques such as KLT [9][10], SURF [12] for tracking features of objects in underwater video sequence. The comparison is based on evaluation of number of feature points matched in pair of successive frames. The video sequence of underwater objects is captured in a small water body in natural light. The scene includes several objects at distance [1m, 2m], near the corner of the water body. The monocular video sequence of underwater objects is captured using Canon D10 water proof camera at a depth of 2m from the surface level of water. The captured monocular underwater video sequence is diminished due to optical properties of the light in underwater environment. The captured video images are suffered from non-uniform illumination, low contrast, blurring and diminished colors. The Fig. 5 shows one pair of video frames suffered from these effects. These effects are preprocessed sequentially by applying homomorphic filtering, wavelet denoising and anisotropic filtering technique [11]. The Fig. 6 shows result of preprocessing on a pair of video frames. The proposed SIFT technique is employed on preprocessed pair of video frames to track the same features. The Fig. 7 shows the extracted features using SIFT method. The features extracted from each frame is stored in a feature descriptor, the euclidian distance is employed as a similarity measurement for each feature descriptor using suitable threshold found empirically. Finally, the features are matched based on Best-Bin-First nearest neighbor approach. The Fig. 8 shows matched features between corresponding consecutive image pair. Table I gives the comparison of matched feature points obtained using SIFT, KLT and SURF. From Table I we can see that SIFT method is best compared to KLT and SURF, because the number of matched feature points is 419 using SIFT compared to matched feature points 397 and 403 obtained using KLT, SURF respectively. This due to the fact that SIFT extracts most invariant features from the video sequence where the disparity between the frames is very high. The underwater video sequence has such property that the lighting variations between the frames are very high due to optical properties of the water. Therefore the SIFT is most suitable feature tracker for underwater video sequence.

Figure 5. Successive video frames of captured underwater video Sequence

Figure 6. The result of pre-processing on diminished video images

DOI: 02.CEM.2011.01.529

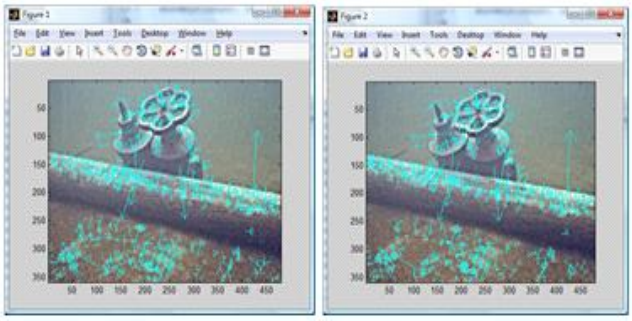

Figure 7. Extracted features using SIFT on a pair of successive video frames

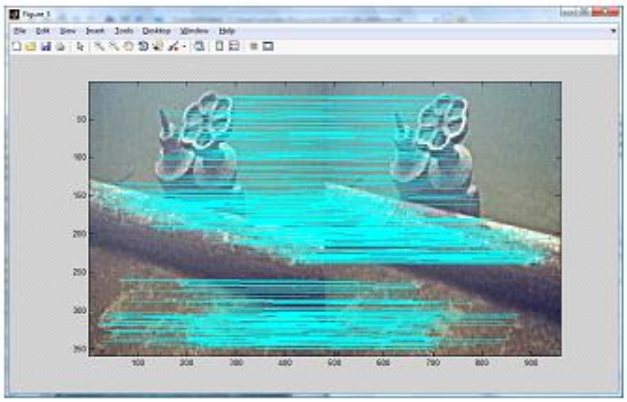

Figure 8. The result of SIFT for feature matching on a pair of a video frames

TABLE I. COMPARISON OF FEATURE TRACKING TECHNIQUES BASED ON NUMBER OF FEATURES MATCHED

Method	Features in frame1	Features in frame2	Matches
SIFT [6]	868	896	419
KLT [9] [10]	873	893	397
SURF[12]	756	824	403

III. CONCLUSIONS

In this paper, we proposed to adapt SIFT technique as a feature tracker in the video sequence of underwater objects for the purpose of 3D surface reconstruction. Since there is no benchmark video data available for comparing proposed technique with other popular techniques. We have conducted experiment using captured video sequence obtained from monocular video camera. To evaluate the efficacy of the proposed method, we have compared with popular tracker such as KLT and SURF techniques based on number of feature points matched. The experimental results shows that SIFT method gives comparably better performance than KLT and SURF tracker. SIFT features are invariant to rotation, scale changes and affine transformations. KLT is fast and good for illumination changes, but not suitable for affine transformation. SURF is fast and has good performance as the same as SIFT, but it is not stable to rotation and illumination changes, which is common in underwater environment.

ACKNOWLEDGMENT

This work has been supported by the Grants NRB/SC/ 158/2008-2009, Naval Research Board, DRDO, New Delhi, India.

REFERENCES

[1] E. Trucco, Y. R. Petillot, I. Tena Ruiz, K. Plakas, and D. M. Lane, “Feature Tracking in Video and Sonar Subsea Sequences with Applications”, *Computer Vision and Image Understanding*, vol. 79, pp. 92-122, February 2000.

[2] K. Plakas, E. Trucco, and A. Fusiello, “Uncalibrated vision for 3-D underwater applications”, In Proceedings of OCEANS ’98, vol. 1, pp. 272 - 276, October 1998.

[3] Anne Sedlazeck, Kevin Koser, and Reinhard Koch, “3D Reconstruction Based on Underwater Video from ROV Kiel 6000 Considering Underwater Imaging Conditions”, OCEANS 2009 - EUROPE, pp. 1-10, 2009.

[4] Andrew Hogue and Michael Jenkin, “Development of an Underwater Vision Sensor for 3D Reef Mapping”, In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5351- 5356, October 2006.

[5] V. Brandou, A. G. Allais, M. Perrier, E. Malis, P. Rives, J. Sarrazin, and P. M. Sarradin, “3D Reconstruction of Natural Underwater Scenes Using the Stereovision System IRIS”, OCEANS 2007 - Europe , pp. 1-6, 2007.

[6] David G Lowe, “Distinctive image features from scale-invariant keypoints”, *International Journal of Computer Vision*, vol. 60(2), pp. 91-110, 2004.

[7] David G Lowe, “Object recognition from local scale-invariant features”, Proceedings of the International Conference on Computer Vision, vol. 2, pp. 1150-1157, 1997.

[8] Piotr Jasiobedzki, Stephen Se, Michel Bondy, and Roy Jakola, “Underwater 3D mapping and pose estimation for ROV operations”, OCEANS 2008, pp. 1-6, September 2008.

[9] C. Tomasi and T. Kanade, “Detection and tracking of point features”, Technical Report CMU-CS-91-132, Carnegie Mellon University, Pittsburg, PA, April 1991.

[10] J. Shi and C. Tomasi, “Good features to track”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 593-600, June 1994.

[11] Prabhakar C. J. and Praveen Kumar P.U., “Underwater image denoising using adaptive wavelet subband thresholding”, In Proceedings of IEEE International Conference on Signal and Image Processing (ICSIP), 2010, pp. 322-327, December 2010.

[12] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, “SURF: Speeded Up Robust Features”, *Computer Vision and Image Understanding*, vol. 110(3), pp. 346-359, 2008.

DOI: 02.CEM.2011.01.529