USING SIFT ALGORITHM FOR VISUAL-BASED LOCALIZATION IN MULTI-AUVS FLEET

SILVIA BOTELHO*, PAULO DREWS JR*, MARIANE MEDEIROS*, TANIA MEZZADRI CENTENO†

*Fundação Universidade Federal do Rio Grande - FURG
Av. Itália Km 8 - 96201-000 - Rio Grande/RS - Brazil

†Universidade Tecnológica Federal do Paraná - UTFPR
Av. Sete de Setembro 3165 - 80230-901 - Curitiba/PR - Brazil

Emails: silviacb@ee.furg.br, paulo@ee.furg.br, mari@ee.furg.br,
mezzadri@cpgei.cefetpr.br

Abstract— The use of teams of Autonomous Underwater Vehicles for visual inspection tasks is a promising robotic field. The images captured by different robots can also be used to aid in the localization/navigation of the fleet. In previous works, we have proposed a distributed localization system based on the Augmented States Kalman Filter through the visual maps obtained by the fleet. In this context, this paper details a system for on-line construction of visual maps and its use to aid the localization and navigation of the robots. Different aspects related to the capture, treatment and construction of mosaics by fleets of robots are presented. The Scale Invariant Feature Transform (SIFT) algorithm is used as a method of extracting and describing keypoints between consecutive images which are robustly invariant to common image transforms. The developed system can be executed on-line on different robotic platforms. The paper is concluded with a series of tests and analysis, in different underwater conditions, for system validation.

Keywords— computer vision, robotics, underwater robots, mosaic, SIFT.

1 Introduction

Autonomous Underwater Vehicles ($AUVs$) can be applied to many tasks (Fleischer, 2000). In underwater visual inspection, the vehicles can be equipped with *down-looking cameras*, usually attached to the robot structure (Garcia et al., 2005). These cameras capture images from the deep of the ocean, supplying visual maps as the vehicle navigates. Each captured image is used to compose such map. Consecutive images are then aligned, generating a final map, also known as mosaic (Botelho et al., 2007). The generated mosaics can also be used as reference maps for the vehicle navigation system (Fleischer, 2000). This gives to the robot the ability to navigate in an autonomous way and in real-time. Given the robot altitude relative to the deep of the ocean (i.e. from the robot's altimeter) and the camera field of view, the covered area is known and real-time navigation is possible, using the described on-line construction.

In this context, we argue that simple robot fleets can be more efficient than a sophisticated AUV, in specific underwater exploration/inspection tasks. These simple vehicles could explore the same area in just a fraction of the time demanded by a single sophisticated AUV. More efficient mosaics can be generated and its visual information can be used to help the localization and navigation of the fleet.

In previous papers, an extension to the Augmented State Kalman Filter (ASKF) was presented, aiming the states estimation for a fleet of robots with *down-looking cameras* (Botelho et al., 2007). Besides, we have implemented a visual system based on (Garcia et al., 2005). In this approach, noises related to image acquisition were filtered out by using the Prewitt Filter (low-pass). Points of interest were obtained by a border detection algorithm (*corners*) (Shi and Tomasi, 1994). However this visual-based localization system needs that the z coordinate (deep) holds constant. Moreover, its does not deal very well with noises and undersea features.

Thus, in this paper we propose a new approach to extract and describe keypoints between consecutive images which are robustly invariant to common image transforms. This method is called Scale Invariant Feature Transform (SIFT) (Lowe, 2004). Our Distributed Visual System for on-line mosaic construction is presented. The system supply information for the ASKF estimator. The images can be captured by simple low-cost underwater robots, associated to a processing unit connected to the surface, the mosaic can be computed by the robot itself.

Initially the paper presents related works on underwater mosaic construction. Section 3 presents a detailed view of our approach with SIFT algorithm, followed by the implementation, test analysis and results with different undersea features. Finally, the conclusion of the study and future perspective is presented.

2 Utilizing AUVs for Visual Mapping

Vision system for mapping the deep of the ocean have been developed since the end of the 80's. In those systems the mosaic were used only as visual information for the users in the surface.

The Construction of Visual Maps In order to assemble the mosaic, the various captured images (also known as frames) must be successively overlayed, resulting in a single visual map. Normally, the overlaying of those images takes a few steps.

Initially, the images pass through a *Preprocessing* stage, when geometric deformations are corrected and inadequate information for processing are removed. After that, the displacement between consecutive frames need to be measured. In the literature we can find a set of works in the frequency (Rzhanov et al., 2000) or in the spacial domain (Olmos et al., 2000). The latter can be *feature-based* (search by points of interest) or *featureless* (no need for points of interest). The relative movement between consecutive captured images can be estimated by *Homography* (planar matrix transformation) (Szeliski, 1994).

Having the displacement information, *Mosaic update* takes place. In this stage the determination of *when* and *how* a new image must be incorporated to the mosaic (temporal/spacial interval) occurs. Finally the *Image overlaying and mosaic construction* phase is executed. The pixels are combined so they overlay in the time-line (for example, average in time, medium in time, most recent pixel or least recent pixel).

Utilizing Mosaics for AUV localization The use of visual maps to assist in the localization of the vehicle introduces a new phase on the process: path estimation from the captured images, taking into account the need for on-line processing.

The previous works concerning utilization of mosaics for AUV localization consisted on detecting and correlating points in successive images. The Kalman Filter was used to predict point positions on the next image. The vehicle path was estimated using Generalized Hough Transform (GHT). Another approach created visual maps in real-time by using special-purpose hardware for image manipulation (Marks et al., 1995).

(Gracias et al., 2002) uses the Harris and Stephes algorithm for the detection of border points as the points of interest, using first order derivatives of the images, estimating movement parameters (planar matrix transformation) for a sequence of images. A system for the generation of mosaics and real-time estimation of 3D displacement utilizing special-purpose hardware was developed by (Negahdaripour et al., 1998).(Garcia et al., 2005) applies texture operators and similarities measurement, as the Energy Filter, Co-occurrence Matrix, and others, making the correlation between consecutive images of the mosaic more precise. This approach was implemented directly in special-purpose hardware. The utilization of visual information to assist the dynamic stabilization of vehicles presented by (Perrier, 2005) should also be cited, as well as works in which the visual information is used in conjunction with informations acquired by other sensors (Kalyan et al., 2005).

(Se et al., 2005) propose a system for vision-based localization and mapping of mobile robots. This system uses SIFT to detect and correlate keypoints. The method computes a descriptor for the local image region that is invariant to image scale and rotation and highly distinctive, besides it is as invariant as possible to remaining variations, such as change in illumination or 3D viewpoint (Lowe, 2004).

In the multi-AUVs context, issues associated with architecture and supervision (Spenneberg et al., 2005), Mines Inspection using distributed sonar, new kinds of sensors, as *smart cables* (Yu and Ura, 2004) are presented in the literature. (Madhavan et al., 2002) proposes the utilization of Kalman Filters for state estimation of wheeled mobile robots and known structured environments.

In this paper, originally propose a fleet of AUVs, each one using visual information to improve on the localization task of the robots. The visual system is based on SIFT algorithm to treat undersea images captured by each robot. The system employs ASKF for fleet state estimation (Botelho et al., 2007). Next section details a distributed visual system for the construction of visual maps by sets of AUVs.

3 A distributed visual system for the construction of visual maps

This work starts with pre-processing and manipulation of correspondence points, then using homographic techniques and finally the on-line assembly of the distributed mosaic. The following subsections presents details of each one of these steps.

Pre-Processing The distortion caused by the camera lenses can be represented by a radial and tangential approximation. As the radial component causes a bigger distortion, most of the works developed so far corrects only this component (Gracias et al., 2002) . Thereby, the following equations 1 are used to correct radial distortion.

$$P_d = P(1 + k_1 * r^2 + k_2 * r^4) \qquad (1)$$

where $P_d = (x_d, y_d)$ are the corrected coordinates of the distorted point measure $P = (x, y)$, $r = (x^2 + y^2)$ and k_1 e k_2 are the coefficients of radial distortion, which are unique to the individual camera.

3.1 Detection of Keypoints and Correlating

The Scale Invariant Feature Transform (SIFT) is a robust filter to extract and describe interest points of images (Lowe, 2004). The algorithm has 4 major stages.

Scale-space extrema detection The first stage searches over scale space using a Difference of Gaussian function to identify potential interest points.

Keypoint localization The localization and scale of each candidate point is determined and keypoints are selected based on measures of stability.

Orientation assignment One or more orientations are assigned to each keypoint localization based on local image gradient directions. All future operations are performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, thereby providing invariance to these transformations.

Keypoint descriptor The local image gradients are measured at the selected scale in the neighborhood of each keypoint. These are transformed into a representation that allows for significant levels of local shape. The description vector is divided by the square root of the sum of squared components to obtain partially illumination invariance.

The SIFT feature algorithm is based upon finding localizations within the scale space of an image which can be reliably extracted. The first stage finds scale-space extrema located in $D(x, y, \theta)$, the Difference of Gaussians (DOG) function, which can be computed from the difference of two nearby scaled images separated by a multiplicative factor k, as shown in equation 2.

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

where $L(x, y, \sigma)$ is the scale space of an image, built by convolving the image I(x, y) with the Gaussian kernel $G(x, y, \sigma)$. Points in the DOG function which are local extrema in their own scale and one scale above and below are extracted as keypoints. Generation of extrema in this stage depends on the frequency of sampling in the scale space k and the initial smoothing σ_0. The keypoints are then filtered in search for more stable matches in more accurate scales and with greater subpixel precision using methods described in (Brown and Lowe, 2002).

Before a descriptor for the keypoint is constructed, the keypoint is assigned an orientation to make the descriptor invariant to rotation. This keypoint orientation is calculated from an orientation histogram of local gradients from the closest smoothed image $L(x, y, \sigma)$. For each image sample $L(x, y)$ at this scale, the gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ is computed using pixel differences, as shown in equation 3 and 4.

$$\begin{aligned} m(x,y) = (L(x+1,y) - L(x-1,y))^2 + \\ (L(x,y+1) - L(x,y-1))^2)^{1/2} \qquad (3) \\ \theta(x,y) = tan^{-1}((L(x,y+1) - L(x,y-1)) \\ /(L(x+1,y) - L(x-1,y))) \qquad (4) \end{aligned}$$

The orientation histogram has 36 bins covering the 360 degree range of orientations. Each sample added to the histogram is weighted by its gradient magnitude and by a Gaussian-weighted circular window with a σ that is 1.5 times that of the scale of the keypoint. Additional keypoints are generated for keypoint locations with multiple dominant peaks whose magnitude is within 80% of each other. The dominant peaks in the histogram are interpolated with their neighbors for a more accurate orientation assignment.

The local gradient data from the closest smoothed image $L(x, y, \sigma)$ is also used to create the keypoint descriptor. This gradient information is first rotated to align it with the assigned orientation of the keypoint and then weighted by a gaussian with σ variance. The weighted data is used to create a nominated number of histograms over a set window around the keypoint. Each histogram has 8 orientation bins each created over a support window of 4x4 pixels. The resulting feature vectors are 128 elements with a total support window of 16x16 scaled pixels.

The best candidate to correlate each keypoint is found by identifying its nearest neighbor in the database of keypoints from training images. The nearest neighbor is defined as the keypoint with minimum Euclidean distance for the invariant descriptor vector.

However, many keypoints from one image will have no good match to the second image. To eliminate false matches, the most effective method is to compare the smallest match distance to the second-best distance. A match should be selected only if this ratio is below a threshold.

3.2 Estimating the Homographic Matrix

The images correlation provide a set of relative displacement vectors between the points associated to the found correspondence pairs. The n pairs are used to determinate the homographic matrix H. This homographic matrix will provide the estimated displacement between such images, transforming the homogeneous coordinates into non-homogeneous. The terms are operated in order to obtain a linear system, as the equation 5:

$$\begin{bmatrix} x_1\prime & 0 & \cdots & x_n\prime & 0 \\ y_1\prime & 0 & \cdots & y_n\prime & 0 \\ 1 & 0 & \cdots & 1 & 0 \\ 0 & x_1\prime & \cdots & 0 & x_n\prime \\ 0 & y_1\prime & \cdots & 0 & y_n\prime \\ 0 & 1 & \cdots & 0 & 1 \\ -x_1.x_1\prime & -y_1.x_1\prime & \cdots & x_n.x_n\prime & y_n.x_n\prime \\ -x_1.y_1\prime & -y_1.y_1\prime & \cdots & -x_n.y_n\prime & -y_n.y_n\prime \end{bmatrix} .$$

$$\begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{21} & h_{22} & h_{23} & h_{31} & h_{32} \end{bmatrix} = \begin{bmatrix} x_1 \\ y_2 \\ \vdots \\ x_n \\ y_n \end{bmatrix} \quad (5)$$

3.3 Assembling the Mosaic

The mosaic construction concept of this paper allows a single vehicle to generate its mosaic and also multiple vehicles to construct it concurrently, using the same programming structure. The visual construction stages of the mosaic are detailed next.

Global Registry on the Mosaic Defined the projective transformation matrix between the previous image I and the current image $I\prime$, knowing the average value of the distances between interest points m and $m\prime$ from the images is different from zero, it is possible to register globally the current image $I\prime$, adding it and the H matrix to the mosaic structure, considering the H matrix is now referenced by ${}^{k}H_{k+1}$, where k references I and $(k+1)$ references I'. The global projective transformation [1] of the image $I\prime$ on the mosaic can be defined by the equation 6.

$${}^{1}H_{k+1} = \prod_{i=1,\dots,k} {}^{i}H_{i+1}, \quad (6)$$

The Distributed Construction Task To visually assemble the mosaic, the global projective transformation of each image belong to in the map must be known. As described by 6, ${}^{1}H_{k+1}$ represents this information and is used in $\tilde{m}^1 = {}^{1}H_{k+1} \cdot \tilde{m}^{(K+1)}$, which provides where each pixel ($m^{(K+1)}$ position) from the mosaic's image $(k+1)$ must be written on the final mosaic ($\tilde{m}^1$ position). The points $\tilde{m}^{(K+1)}$ vary within $x \in [1 \dots l]$ and $y \in [1 \dots c]$, assuming l and c the captured image dimensions.

When multiple vehicles are executing exploration jobs, the matrix ${}^{1}H_1$ from the first image of each vehicle is defined by the equation 7:

$${}^{1}H_1 = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \quad (7)$$

Notice that in this case each robot will be initially shifted by a position (t_x, t_y) from the inertial reference, so this position must be informed by the system operator on startup. Thus, considering that the matrix ${}^{1}H_1$ of each vehicle will be different, the homography ${}^{1}H_{k+1}$ of each one will provide the position on the final mosaic, therefore other transformations are not necessary.

Localizing the Fleet In (Botelho et al., 2007), the Augmented State Kalman Filter was extended to estimate the correct 3D position of the fleet and mosaic. The ASKF is fed with displacement informations obtained on the homography stage, correcting errors acquired throughout the process.

Capital cost of the system is limited to that of a camera, lighting, off-line storage and a capable processor (required for inspection tasks). No infrastructure such landmarks, beacons is necessary. However our system makes some assumptions: persistent appearance of the scene, majority flat 2D scene, restricted robot motion and initial position estimate of each robot.

The proposed Visual System was completely implemented. A set of tests are done, validating the proposal.

4 System Implementation, Tests and Results

We have design an underwater vehicle. These robots are equipped with a Tritech Typhoon - Colour Underwater Video Camera with Zoom, a Miniking sonar and a set of sensors (altimeters and accelerometers).

The developed platform consists of a software system for each vehicle and another for the central station. The implementation is based on a client-server architecture.The communication between the client and the server uses sockets, which utilizes TCP (Transmission Control Protocol). The client software is responsible for capturing the images obtained by the camera. Besides it process the SIFT correlation between those images. The central station is responsible for gathering the information from all the vehicles, assembling the distributed visual map. A set of tests and analysis were executed verify and validate the usage of the involved parameters and processes. Following there are some tests executed on desktop AMD Athlon X2 2800+ computers with 1Gb of RAM, both in the client and server side.

To validade the localization system, an experiment was conducted by coupling a camera to the robotic arm. The robotic arm consist of a *harmonic drive* actuator with a coupled encoder supplying angular readings each 0.000651 seconds. The data obtained by the encoder were compared with the information supplied by the vision system. The camera was pre-calibrated. Different

[1] relative to the first reference image added

Distortion	1	2	3	4
Light Source Distance	0.2	0.22	0.25	0.3
Attenuation Value	0.05	0.05	0.06	0.05
Gaussian Noise	2	2	2	4
Gray Level Minimum	20	30	20	20

Table 1: Undersea Features For Each Distortion and the tests.

undersea features were applied, like turbidity, sea snow, low illumination, and others.

Test 1 The first test is obtained by a rotation movement of the robotic arm around its own axis. The final mosaic is the result of rotation and translational movements of the camera coupled to the robotic arm. The test works with 200 captured and processed frames. While the mosaic is being constructed, real-time position information is acquired. Figure 1, at column one, shows in green the displacement returned by the SIFT vision system and in blue the reference obtained by the coupled encoder, respectively. Notice that when the arm inverts its movement (peaks on the curve), the small drift between the position returned by the vision and the reference is more evident. This is the result of the incapacity to detect the movement of the interest points in smooth movement situations.

The SIFT visual localization system follows the robotic system in a very good satisfactory manner. Notice that no information related to position or pre-established landmarks are necessary.

Test 2 Figure 1, at column two, shows in blue the speed of the arm, supplied by the encoder, and in green the speed obtained by the SIFT vision system. In this test the navigation time was considerable increased.

Test 3 The same test was conducted with four different undersea conditions, see table 1. The good results can be visualized in the figure 1, at column three.

Test 4 Tests were conducted with two robots connected to a central station. Figure 2 presents a visual map, generated in a distributed way with 195 images. Each robot was responsible for sending information to the central station at a three frames per second rate. The overlaying of images on the crossing region was satisfactory.

5 Conclusion

We intend to construct visual underwater maps using one or more AUVs. In this paper we have presented an original visual approach using SIFT and homography matrix to estimate the underwater robot's localization. Theses maps might assist on fleet localization throughout a exploration mission.

Several tests with different undersea features were conducted. The effectiveness of our propose was validated in a set of scenarios with different levels of turbidity, sea snow, low illumination, and others. The results shown advantage to using SIFT because it is invariant to scale (z coordinate) and rotation and highly distinctive. Moreover, the system was as invariant as possible to remaining variations, such as illumination variabilities.

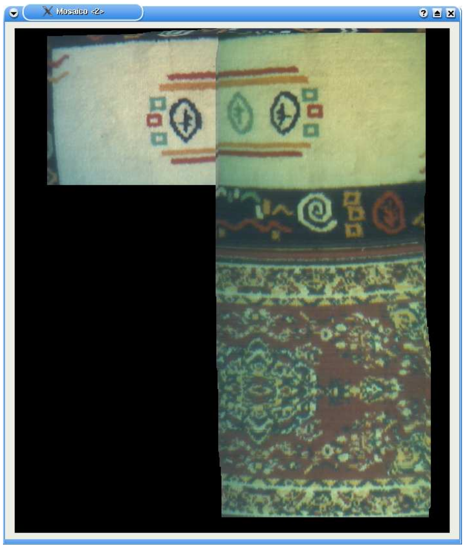

Figure 2: Final Mosaic generated by 2 robots.

As future works we intend to integrate the ASKF system (currently running on MatLab) with AUVs robotic platforms. Besides, we intend to use real-time GPU-SIFT (Sinha, 2006) and the improvements of the SIFT for a faster execution(Grabner et al., 2006)(Ledwich and Williams, 2004). Also the manipulation of stereoscopic images captured by stereo video heads is a medium-term goal.

References

Botelho, S., Drews, P. and Madden, C. (2007). A visual system for distributed mosaics using an auv fleet, *IEEE Oceans*, pp. 332–337.

Brown, M. and Lowe, D. (2002). Invariant features from interest point groups, *British Machine Vision Conference*.

Fleischer, S. (2000). *Bounded-error vision-based of autonomous underwater vehicles*, PhD thesis, Stanford University.

Garcia, R., Lla, V. and Charot, F. (2005). Vlsi architecture for an underwater robot vision system, *IEEE Oceans*, Vol. 1, pp. 674–679.

Grabner, M., Grabner, H. and Bischof, H. (2006). Fast approximated sift, *Asian Conference on Computer Vision*.

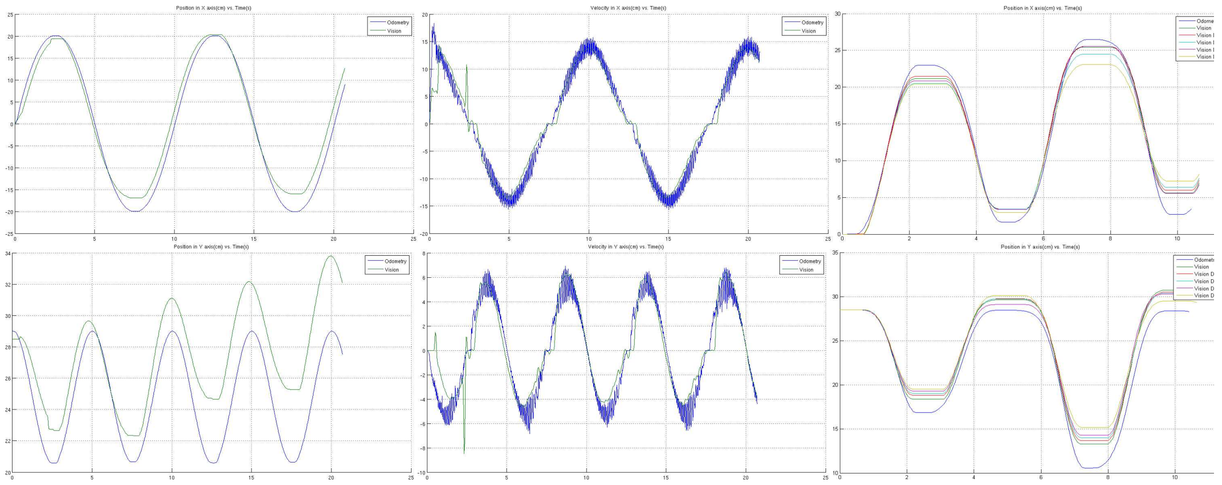

Figure 1: Column 1 shows the displacement in x and y with a trajectory generated by the rotation and translational movements - in blue the encoder output, in green the SIFT vision output. Column 2 present the velocities in x and y - in blue the encoder output, in green the vision output and column 3 is the displacement in x and y with a trajectory generated by four different undersea features and the reference in blue.

Gracias, N., Van der Zwaan, S., Bernardino, A. and Santos-Vitor, J. (2002). Results on underwater mosaic-based navigation, *IEEE Oceans Conference*, Vol. 3, pp. 1588–1594.

Kalyan, B., Balasuriya, A., Kondo, H., Maki, T. and Ura, T. (2005). Motion estimation and mapping by autonomous underwater vehicles in sea environments, *IEEE Oceans*.

Ledwich, L. and Williams, S. (2004). Reduced sift features for image retrieval and indoor localisation, *Australian Conf. on Robotics and Automation*.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* **60**(2): 91–110.

Madhavan, R., Fregene, K. and Parker, L. (2002). Distributed heterogeneous outdoor multi-robot localization, *ICRA*.

Marks, R., Rock, S. M. and Lee, M. J. (1995). Real-time video mosaicking of the ocean floor, *IEEE Journal of Oceanic Engineering* **20**(3): 229–241.

Negahdaripour, S., Xu, X. and Khamene, A. (1998). A vision-system for real-time positioning, navigation and video mosaicing of sea floor imagery in the application of rovs / auvs, *4th IEEE Workshop on Applications of Computer Vision*, pp. 248–249.

Olmos, A., Trucco, E., Lebart, K. and Lane, D. M. (2000). Detecting ripple patterns in mission videos, *IEEE Oceans*, pp. 331–335.

Perrier, M. (2005). The visual servoing system "cyclope" designed for dynamic stabilisation of auv and rov, *IEEE Oceans*.

Rzhanov, Y., Linnett, L. and Forbes, R. (2000). Underwater video mosaicing for seabed mapping, *IEEE Conference on Image Processing*.

Se, S., Lowe, D. and Little, J. (2005). Vision-based global localization and mapping for mobile robots, *IEEE TRA*, Vol. 21, pp. 364–375.

Shi, J. and Tomasi, C. (1994). Good features to track, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600.

Sinha, S. (2006). Gpu-based video feature tracking and matching, *Edge Computing Using New Commodity Architectures*.

Spenneberg, D., Waldmann, C. and Babb, R. (2005). Exploration of underwater structures with cooperative heterogeneous robots, *IEEE Oceans*, Vol. 2, pp. 782–786.

Szeliski, R. (1994). Image mosaicing for telereality applications, *IEEE Workshop on Applications of Computer Vision*, pp. 44–53.

Yu, S. and Ura, T. (2004). A system of multi-auv interlinked with a smart cable for autonomous inspection of underwater structures, *International Journal of Offshore and Polar Engineering* **14**(4).