BINOCULAR VISUAL ENVIRONMENT PERCEPTION TECHNOLOGY FOR UNMANNED SURFACE VEHICLE

Y. Wang [1], M. Peng [1,*], Z. Liu [1], W. Wan [1], K. Di [1], C. Hu [2,3], L. Liu [2], T. Lv [2], C. Yang [2]

[1] State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China – (wangyx716, pengman, liuzq, wanwh, dikc)@radi.ac.cn
[2] Beijing Institute of Aerospace Control Instruments, Beijing, China – 18910801138@189.cn, uuufly@126.com, lyu.teng@foxmail.com, yck2488585@163.com
[3] Pilot National Laboratory for Marine Science and Technology, Qingdao, China

Commission II, WG I/II, II/III

KEY WORDS: Stereo vision, Localization, Calibration, Deep learning, Target recognition, Unmanned surface vehicle

ABSTRACT:

Binocular vision system is an essential way for target localization in many fields, which has been widely used as payload of unmanned surface vehicles (USV). High resolution cameras, which can provide richer information, are utilized more often on a USV. This brings challenges of computing tremendous data for target detection and localization in real-time. In this paper, we propose an framework to automatically detect and localize target using high resolution binocular cameras for environment perception of USV. Instead of processing the whole image, the feature extraction and matching are executed within the target region of interest determined by a deep convolution network. Then the target can be localized using triangulation principle with calibrated binocular camera parameters. Experiments show that our proposed strategy can achieve both precise detection and high accurate localization results in real-time applications.

1. INTRODUCTION

Unmanned Surface Vehicles (USVs) are autonomous marine robots which have caused rising attention in recent years. As USVs can perform continuously without human interference to save labour and avoid casualty, they are and will be widely utilized in multiple marine tasks, such as environment monitoring, patrol, exploration, security, etc (Bertram, 2008; Breivik et al., 2008; Raboin et al., 2014).

Generally, in order to assure the safety of the platform and perform autonomously in complex environment, different types of payloads are mounted on a USV, such as radar, lidar, sonar, camera, GNSS and IMU, etc. (Heidarsson and Sukhatme, 2011; Ji et al., 2014; Liu et al., 2016b; Schuster, 2014; Shi et al., 2019). Correspondingly, target or obstacle detection methods by using data captured by at least one of these payloads have been studied in the past decade. Precise detection results can provide basic information for applications such as hazard avoidance and path planning of the USV (Chen et al., 2019; Liu et al., 2019).

Optical camera, which can provide rich information of the scene, is an essential payload of the USV for target and obstacle detection. Target detection using images has been studied for decades. In the early times, target detection algorithms were mostly about extracting edges and segmentation using thresholding. Then some algorithms under specific theories were also proposed such as clustering, active contour segmentation, level set, graph theory-based segmentation, super pixel, saliency detection, template matching, etc. Most of these methods only result in feature or region expression and extraction with no process of recognition or identification. With the development of the computing capabilities, machine learning methods have been used for image target recognition. In the early stage, shallow networks such as decision tree, random forests, support vector machines (SVM), Boosting and neural network, were mainly used to identify targets. At the same time, a number of feature description operators had emerged such as Haar, histogram of oriented gradient, local binary pattern, etc. As a branch of machine learning, deep learning raised around 2006 (Ian et al., 2016). In 2012, convolutional neural networks (CNN) gained great attention due to the significant advantages in target recognition and classification accuracy. Then many deep CNN networks such as AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015), etc., have been developed and applied in multiple fields. With these deep networks and subsequently proposed concepts such as regularized discarding, residual network, etc., numbers of target detection methods have been proposed such as Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2017), R-FCN (Dai et al., 2016), SSD (Liu et al., 2016a), Yolo v1~v3 (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018), etc. The number of the network layers have been significantly increased, meanwhile the detection accuracy has been improved.

Recently, target detection and obstacle avoidance using optical camera on USV have been reported (He et al., 2019; Ma et al., 2019). However, a single image can not restore depth information. In other vision fields, binocular vision structure is widely used for depth recovery. Normally, the stereo images are firstly rectified according to epipolar geometry, so that the corresponding points can be searched in the same horizontal lines of the images (Zhou et al., 2013). Then the 3D information

* Corresponding author

can be calculated by the matched points using triangulation principle. Alternatively, corresponding points can be extracted and matched directly on the unrectified stereo images to reduce the amount of calculation in epipolar rectification , so that 3D information of the sparse key points can be restored (Wan et al., 2017a). However, using the later way to process two large images is still time-consuming.

Nowadays, high resolution cameras are utilized for better visualization and detailed texture. Meanwhile, the data amount and computing cost of these large images may increase tens of times. This brings challenges of target detection and localization in real-time. In this paper, we propose a framework to automatically detect and localize target using binocular vision for USV in real-time applications.

The rest of this paper is structured as follows: Section 2 presents and specifies the proposed framework. Experimental results are presented in Section 3. Discussion and brief conclusions are given in Sections 4 and 5, respectively.

2. METHODS

The flowchart of the proposed approach is shown in Fig.1. Under the guidance of precision analysis results according to the stereo structure factors (Di and Li, 2007; Peng et al., 2014), binocular system with large baseline is established to capture stereo images. First, the captured stereo images are input into the deep network to extract features in order to obtain the targets types and their areas as regions of interest (ROIs). Then the ROIs are matched as candidate corresponding pairs on the left and the right images so that each pair contains the same target. For each ROI pair, multiple feature extraction algorithms are applied within the ROIs on both left and right image to gain feature points which are then matched as corresponding points. After exclusion of outliers, the rest corresponding points are utilized to compute three dimensional coordinates to localize the targets. The final outputs of the workflow are the targets types and positions.

Before the binocular system executing environmental perception, two tasks should be done. Firstly, the binocular system should be accurately calibrated to obtain the intrinsic parameters of the two cameras and structure parameters that reflect the transformation relations between them. Meanwhile, the deep convolution neural network should be trained using labelled samples.

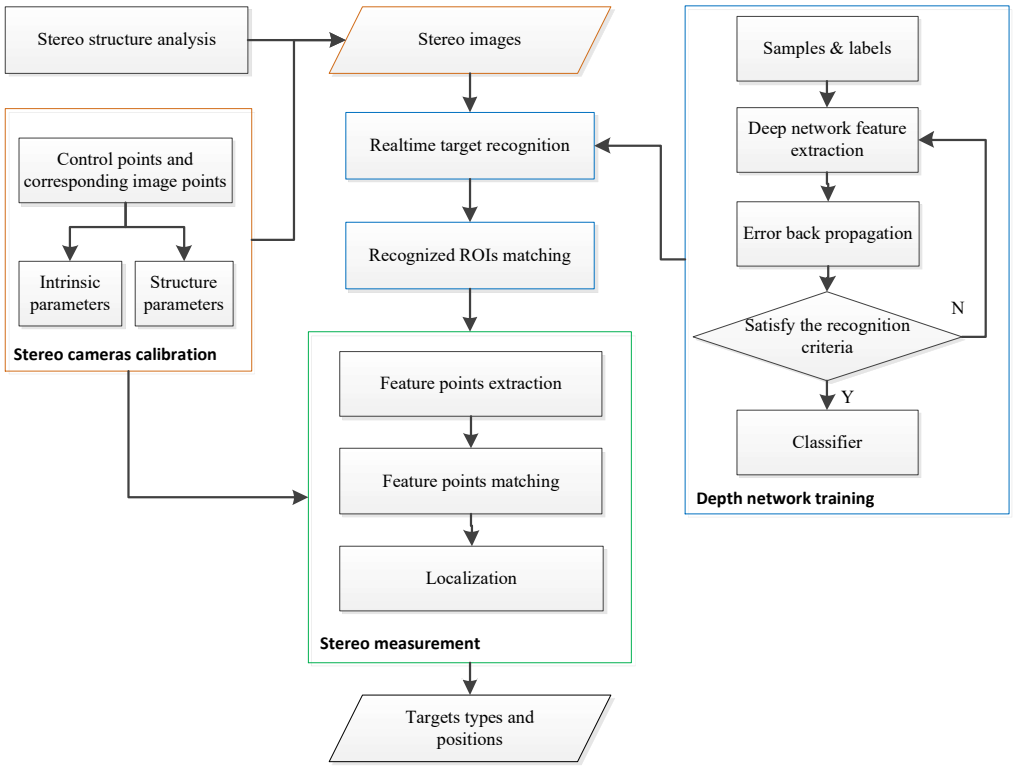


Figure 1. Flowchart of the proposed framework.

2.1 Real-time Target Recognition

By comparing multiple convolution neural network methods, such as Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2017), R-FCN (Dai et al., 2016), SSD (Liu et al., 2016a), Yolo v1~v3 (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018), etc., under consideration of recognition precision, computation speed and algorithm efficiency, we use Yolo v3 to detect and recognize targets in the images.

Yolo v3 adopts numerous residual layers instead of pooling layers to deal with negative effect brought by gradient, which can achieve better feature extraction performance with relatively less layers (He et al., 2016). In addition, this version of Yolo

predicts targets at three different scales with more accurate bounding boxes so that small objects are easier to be identified.

As the image size of both cameras is larger than the input layer, the images are resampled before input into the deep network. However, after resampling, some small or distant targets may not be recognizable. Therefore, a whole-to-part strategy is carried out during real-time recognition. First, the whole image adjusts the input layer of the deep network. If no targets are recognized, the whole image is cropped into four sub-images. The four sub-images are then input into the deep network in order to get recognition results. Each recognized target is described by type and location information of the rectangle area (image coordinates of the left corner and the width and height of the rectangle area).

The left and right images are first input into the deep network to extract features and detect targets. Information of the detected targets on both left and right images consists of the target types as well as their rectangle areas. On both images, there may be one or more identified targets with their rectangle areas. Before the target can be localized in 3D, the multiple rectangles should be matched so that they corresponding to the same target. The matching method concerns parameters of the stereo cameras, therefore will be detailed in the next subsection.

2.2 Three Dimensional Localization

In order to localize the detected target, triangulation principle is realized by parameters of the stereo cameras, including the intrinsic parameters of both cameras, the structure parameters which describe the rigid transformation relationship from the right camera to the left camera. Denote the coordinates system centered at optical center of the left camera as work coordinate system, control points with known three dimensional world coordinates are captured simultaneously by the stereo cameras for calibration purpose. The intrinsic and structure parameters of the stereo cameras are solved based on the image geometric model, the 3D world coordinates of control points and their 2D image coordinates on both left and right image.

With the stereo camera parameters, fundamental matrix $\boldsymbol{F}$ is calculated based on the epipolar geometry principle. Here the fundamental matrix can be utilized to match the rectangles detected from the previous sub-section. With the following constraint, the center point of each rectangle on the left image is calculated with center point of every rectangle with the same class on the right image.

$$\boldsymbol{x}_R^{\mathrm{T}} \boldsymbol{F} \boldsymbol{x}_L = 0 \tag{1}$$

where $\boldsymbol{x}_L$ and $\boldsymbol{x}_R$ represents the homogeneous image coordinates of the rectangle centers on left and right images, respectively. The pair that gives the least value and less than a setting threshold is considered a corresponding pair of ROIs.

After matching the rectangles, for each pair, the following feature extraction and matching processes focus only on the ROIs to reduce time consumption. Multiple feature point extraction algorithms, such as AKAZE, AKAZE_KAZA, ORB Brisk, and SIFT are applied in both ROIs to gain feature points. Then the points are matched using fundamental matrix constraint equation as well as the Euclidean distance of the feature descriptions. Then the matched results are refined using least square algorithm to obtain sub-pixel accuracy (Wan et al., 2017b). By the calibrated parameters of the stereo cameras, 3D coordinates in work coordinate system can be reconstructed using triangulation principle (Liu et al., 2015). Under security consideration, the point with the nearest distance with in the matched ROIs is chosen as the final localization result.

Note that the localization result is in work coordinate system whose origin is at the optical center of the left camera, which can provide the target position information with respect to the stereo system. By calibrating the transformation from the stereo system to other sensors, the localization result can be integrated into other coordinate system of the USV.

3. EXPERIMENT

To verify our proposed approach, a stereo system was built as a payload of USV. Two cameras, with resolution of 4096 pixel× 3072 pixel, are rigidly mounted on a horizontal carbon fiber mast with a baseline of 2m. To reduce water reflection effect, polarized lenses are attached in front of the camera lenses.

3.1 Target Recognition Result

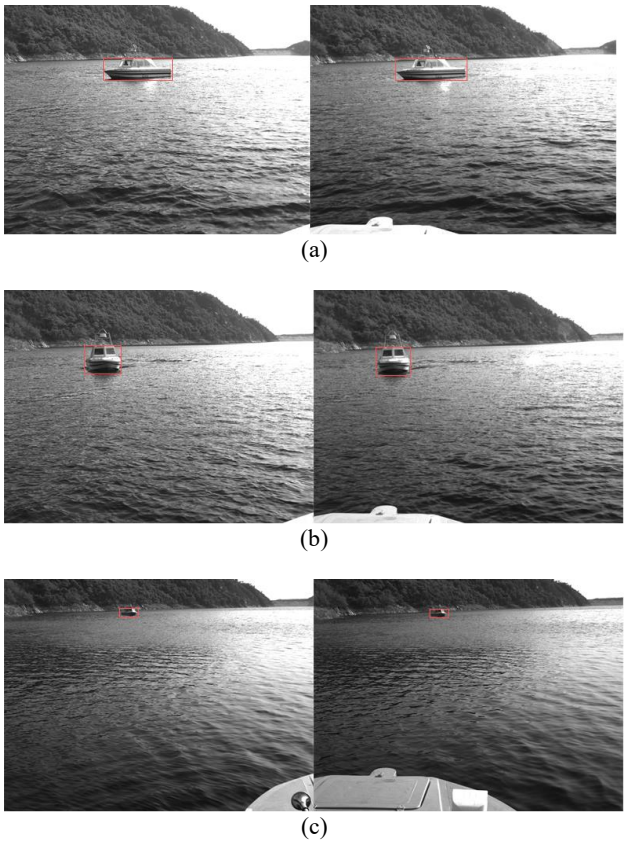
We select 3666 images from the images we captured at a reservoir and wharfs as training data. At the primary stage of our study, the main targets are boats. The samples contain boats in front, side and back views at different distances. To reduce the background interference, only the main body of the boat without the shelves and frames above the main body are labelled.

The samples integrate with the corresponding labels are utilized to train the deep network. The loss of the network declines rapidly in 20 thousand epochs, and becomes stable afterwards. According to the trend of the IoU values, the network doesn't change much after 200 million batches (64 samples a batch). Therefore, combine the above factors, the weights trained larger than 30 thousand epochs is used as the final weights for the network.

We have tested the trained weights using a 1000 images dataset; the test results are listed in Table 1. True positive means the targets on these images are correctly detected and identified, which reaches 98.5%. Only one false alarm is detected in this dataset, which is 0.1%. False detection occurs in several situations: wrong rectangle positions and ranges, repetitive detection of the same target, etc., which accounts for 0.5%. Targets in nine images (0.9% of the dataset) were not identified. The target detection speed is 5 FPS. Figure 2 shows the detection results (detection target areas are marked as red rectangles) of several images with target at different distances captured at the reservoir.

Numbers of images	True positive	False alarm	False detection	Dismissal
1000	985	1	5	9
Occupation (%)	98.5	0.1	0.5	0.9

Table 1. Test results of our trained deep network using a dataset of 1000 images.



(a) (b) (c)

Figure 2. Detection results of several images with target in different directions and distances captured at the reservoir. The detection results are marked as red rectangles.

3.2 Measurement Result

3.2.1 Calibration: The stereo cameras were calibrated by capturing images of the control points shown in Fig. 3 (black dots on the facade of the building). After extracting the control points on the image and establishing the map relationship between the 3D control points and their corresponding image points, the cameras intrinsic parameters and the structure parameters were solved. The calibration results are shown in Table 2.

Figure 3. Stereo images that capture the control points.

Paras.	x_0 (pixel)	y_0 (pixel)	f (pixel)	k_1
Left	2130.824	1481.890	7978.162	1.331e-9
Right	2248.419	1447.627	7986.934	1.220e-9
	k_2	p_1	p_2	α
Left	-2.7e-17	-6.787e-8	3.74e-8	-3.021e-4
Right	-9.0e-18	-5.597e-8	-4.052e-8	-2.277e-4
	β	Δx (m)	Δy (m)	Δz (m)
Left	4.741e-6	-999.608	-1002.580	101.261
Right	7.043e-5	-997.604	-1002.791	101.279
	$\Delta\omega$	$\Delta\varphi$	$\Delta\kappa$	
Left	-4.609	-1.165	-0.337	
Right	-5.480	-1.498	0.285	
	RMS: x (pixel)	RMS: y (pixel)		
Left	0.284	0.273		
Right	0.247	0.275		

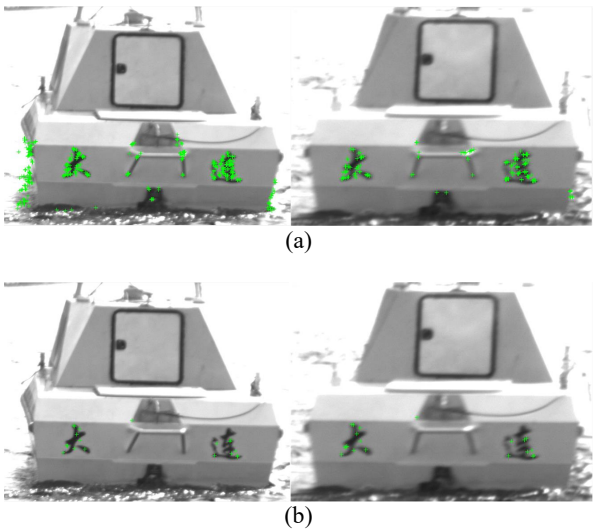
Table 2. Calibration results of the stereo cameras.

In order to test the accuracy of localization, we applied laser ranging results as reference values, and tested five localization results at different distances. The comparisons are listed in Table 3. As shown in Table 3, the measurement errors become larger as the distances from the cameras increase, which coincides with the theoretical error analysis of the stereo system.

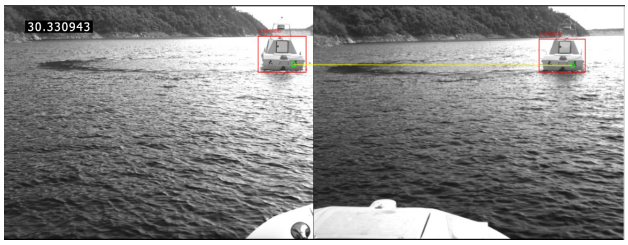
No.	Laser result	Stereo result	Error
1	39.7	39.671	0.036
2	114.8	112.143	2.657
3	210.3	214.865	-4.565
4	311.6	304.026	7.574

Table 3. Measurement results of the stereo cameras comparing with laser ranging.

3.2.2 Target Detection and Localization Result: As described in the previous sections, the detected targets areas are marked as ROIs, and feature point extraction and matching processes are carried out inside the ROIs. The feature point extraction results are shown in Fig. 4(a), and matched feature points in the ROIs are shown in Fig. 4(b). According to the calibration parameters and the localization method, three dimensional coordinates are calculated, and distance of the target from the USV calculated from the localization results is shown in Fig. 4(c). As feature extraction and matching process are time consuming, after adding this part, the speed of the whole process is 1.5 FPS.



(a) (b)



(c)

Figure 4. Results of feature extraction, matching and target localization.

4. DISCUSSION

According to the structure of Yolo v3 and the size of the stereo images, the minimum target that can be detected should be larger than approximately 100 pixel × 130 pixel. By the assistance of sub-image processing strategy, this range can be reduced to approximately 50 pixel×65 pixel. However, this is under ideal imaging conditions. Normally, the detection capability would decrease slightly in case of bad illumination, USV vibrations, and surface reflections from both water and the targets.

From the experiments, we found that especially when the target boat is not far, many extracted and matched points are on the water surface at the top half of the ROI, which would lead to wrong localization result. To avoid this situation, only the bottom half of the ROI is utilized to extract feature points. Besides, according to the most boat designs, the bottom part of the target boat is closer to the USV. In addition, with the distance of the target increases, the localization accuracy decreases fast because of the smaller ROI area, imaging quality reduction, vibration of the USV and reflection influence.

5. CONCLUSIONS

This paper proposed a framework of detecting and localizing target using stereo vision cameras mounted on USV. First, the targets are automatically recognized using deep neural network. Then the detected targets and their corresponding area are marked as ROIs and matched, where feature point extraction and matching are carried out. The localization results are obtained by triangulation principle using matched points and calibrated camera parameters. Experiments of both target detection and localization show that our developed system using stereo camera images achieved a high detection rate and accurate localization results. Future work may include detection and localization of more target types, further speed up the feature extraction and matching algorithms, etc.

ACKNOWLEDGEMENTS

This work was supported by National Key Research and Development Program of China (No. 2018YFB1305004) and National Natural Science Foundation of China (No. 41701489).

REFERENCES

Bertram, V., 2008. Unmanned surface vehicles –A survey, Proceedings of skib- steknisk selskab, Copenhagen, Denmark.

Breivik, M., Hovstein, V.E., Fossen, T.I., 2008. Straight-line target tracking for unmanned surface vehicles. Modeling, Identification and Control 29, 131-149.

Chen, Z., Zhang, Y., Zhang, Y., Nie, Y., Tang, J., Zhu, S., 2019. A Hybrid Path Planning Algorithm for Unmanned Surface Vehicles in Complex Environment With Dynamic Obstacles. IEEE Access 7, 126439-126449.

Dai, J., Li, Y., He, K., Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks. arXiv, 1605.06409.

Di, K., Li, R., 2007. Topographic Mapping Capability Analysis of Mars Exploration Rover 2003 Mission Imagery, 5th International Symposium on Mobile Mapping Technology (MMT 2007), Padua, Italy.

Girshick, R., 2015. Fast R-CNN, 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.

He, W., Xie, S., Liu, X., Lu, T., Luo, T., Sotelo, M.A., Li, Z., 2019. A novel image recognition algorithm of target identification for unmanned surface vehicles based on deep learning. Journal of Intelligent & Fuzzy Systems 37 4437–4447.

Heidarsson, H.K., Sukhatme, G., 2011. Obstacle detection and avoidance for an autonomous surface vehicle using a profiling sonar, IEEE international conference on robotics and automation, Shanghai, China, pp. 731–736.

Ian, G., Yoshua, B., Courville, A., 2016. Deep Learning. MIT Press.

Ji, X., Zhuang, J.Y., Su, Y.M., (). 1006 2014. Marine radar target detection for USV. Advanced Materials Research 1006, 863–869 .

Krizhevsky, A., I.Sutskever, Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks.

Liu, S., Wang, C., Zhang, A., 2019. A Method of Path Planning on Safe Depth for Unmanned Surface Vehicles Based on Hydrodynamic Analysis. Applied Sciences 9, 3228.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016a. Ssd: Single shot multibox detector, In European Conference on Computer Vision. Springer, Amsterdam, The Netherlands, pp. 21-37.

Liu, Z., Di, K., Peng, M., Wan, W., Liu, B., Li, L., Yu, T., Wang, B., Zhou, J., Chen, H., 2015. High precision landing site mapping and rover localization for Chang'e-3 mission. Science China-Physics Mechanics & Astronomy 58, 1-11.

Liu, Z., Zhang, Y., Yu, X., Yuan, C., 2016b. Unmanned surface vehicles: An overview of developments and challenges. Annual Reviews in Control 41, 71-93.

Ma, L., Xie, W., Huang, H., 2019. Convolutional neural network based obstacle detection for unmanned surface vehicle. Mathematical Biosciences and Engineering 17, 845-861.

Peng, M., Wan, W., Wu, K., Liu, Z., Li, L., Di, K., Li, L., Miao, Y., Zhan, L., 2014. Topographic mapping capbility analysis of Chang'e-3 Navcam stereo images and three-dimensional terrain reconstruction for mission operations. Journal of Remote Sensing 18, 995-1002.

Raboin, E., Švec, P., Nau, D.S., Gupta, S.K., 2014. Model-predictive asset guarding by team of autonomous surface vehicles in environment with civilian boats. Autonomous Robots 38, 1-22.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788.

Redmon, J., Farhadi, A., 2017. YOLO9000: Better, Faster, Stronger, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517-6525.

Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement. arXiv.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 1137-1149.

Schuster, M., 2014. Collision Avoidance for Vessels Using a Low-Cost Radar Sensor. IFAC Proceedings Volumes 47, 9673-9678.

Shi, B., Su, Y., Zhang, D., Wang, C., Abouomar, M.S., 2019. Research on Trajectory Reconstruction Method Using Automatic Identification System Data for Unmanned Surface Vessel. IEEE Access 7, 170374-170384.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint at Xiv: 1409.1556.

Szegedy, C., Liu, W., Jia, Y., P. Sermanet, Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, pp. 1-9.

Wan, W., Peng, M., Xing, Y., Wang, Y., Liu, Z., Di, K., Teng, B., Mao, X., Zhao, Q., Xin, X., Jia, M., 2017a. A Performance Comparison of Feature Detectors for Planetary Rover Mapping and Localization, International Symposium on Planetary Remote Sensing and Mapping, Hong Kong, China, pp. 149-154.

Wan, W., Peng, M., Xing, Y., Wang, Y., Liu, Z., Di, K., Teng, B., Mao, X., Zhao, Q., Xin, X., Jia, M., 2017b. A performance comparison of feature detectors for planetary rover mapping and localization. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLII-3/W1, 149-154.

Zhou, F., Wang, Y., Peng, B., Cui, Y., 2013. A novel way of understanding for calibrating stereo vision sensor constructed by a single camera and mirrors. Measurement 46, 1147-1160.