MOVING SHIP DETECTION AND MOVEMENT PREDICTION IN REMOTE SENSING VIDEOS

Yuhao Wang[1], Hangzhang Cheng[1], Xintong Zhou[1], Wei Luo[1,2,3], Haopeng Zhang[1,2,3]*

[1] Department of Aerospace Information Engineering, School of Astronautics,
Beihang University, 102206 Beijing, China - zhanghaopeng@buaa.edu.cn
[2] Beijing Key Laboratory of Digital Media, 102206 Beijing, China
[3] Key Laboratory of Spacecraft Design Optimization and Dynamic Simulation Technologies,
Ministry of Education, 102206 Beijing, China

Commission II, WG II/7, ICWG II/III

KEY WORDS: Ship Detection, Movement Prediction, Remote Sensing Videos, Deep Learning

ABSTRACT:

With the rapid development of remote sensing technology, it is possible to obtain continuous video data from outer space successfully. It is of great significance in military and civilian fields to detect moving objects from the remote sensing image sequence and predict their movements. In recent years, this issue has attracted more and more attention. However, researches on moving object detection and movement prediction in high-resolution remote sensing videos are still in its infancy, which is worthy of further study. In this paper, we propose a ship detection and movement prediction method based on You-Only-Look-Once (YOLO) v3 and Simple Online and Realtime Tracking (SORT). Original YOLO v3 is improved by multi-frame training to fully utilize the information of continuous frames in a fusion way. The simple and practical multiple object tracking algorithm SORT is used to recognize multiple targets detected by multi-frame YOLO v3 model and obtain their coordinates. These coordinates are fitted by the least square method to get the trajectories of multiple targets. We take the derivative of each trajectory to obtain the real-time movement direction and velocity of the detected ships. Experiments are performed on multi-spectral remote sensing images selected on Google Earth, as well as real multi-spectral remote sensing videos captured by Jilin-1 satellite. Experimental results validate the effectiveness of our method for moving ship detection and movement prediction. It shows a feasible way for efficient interpretation and information extraction of new remote sensing video data.

1. INTRODUCTION

Object detection in remote sensing images plays an increasingly important role in many application fields of remote sensing. With the rapid development of remote sensing technology in recent years, it is possible to obtain continuous video data from outer space successfully. For example, the Jilin-1 series of commercial satellites launched since October 7, 2015, are capable of providing remote sensing videos with high spatial resolution. Owing to the temporal characteristics of remote sensing videos, we can not only detect the target as in remote sensing images but also analyze the movement of the detected target. In recent years, more and more attention has been paid to this issue. A lot of studies have focused on object detection and movement prediction in natural images and videos, and promising results have been achieved.

The state-of-the-art object detection methods are deep-learning-based detectors, e.g. Faster RCNN (Faster Regions with Convolutional Neural Network) (Ren et al., 2015), YOLO (You Only Look Once) (Redmon et al., 2016), SSD (Single Shot MultiBox Detector) (Liu et al., 2016), etc. As an improvement of YOLO, YOLO v3 (Redmon, Farhadi, 2018) has better robustness to small targets due to the use of multi-scale prediction and the developed network structure. Besides, the detection speed of YOLO v3 is also fast, which fully meets the request for real-time detection. Since YOLO v3 can achieve fast and accurate object detection in one stage with end-to-end learning, it is very suitable for engineering applications. In terms of object detection in remote sensing images, Li et al. (Li et al., 2018) applied Ting-YOLO (Ma et al., 2017) to airport and aircraft recognition, and proposed a simplified Ting-YOLO algorithm to improve the detection speed. Kharchenko et al. (Kharchenko, Chyrka, 2018) applied YOLOv3 to the detection of airplanes on the ground. This method has high detection ability, positioning accuracy, and real-time processing speed. Chang et al. (Chang et al., 2019) used YOLO v2 (Redmon, Farhadi, 2017) for ship detection in SAR images. For video detection, the motion-guided propagation method mentioned in the T-CNN (Tubeless with Convolutional Neural Network) (Kang et al., 2017) can use the optical flow information to pass the detection result of the current frame forward and backward, effectively reduce missed detections, and sort the category scores successfully.

Multiple Object Tracking (MOT), which means to detect and identify multiple objects in the videos, has been widely applied in pedestrian tracking and vehicle detection. In recent studies, there are inter-frame difference method, optical flow method, background subtraction method, and many other algorithms. The GMM (Gaussian Mixture Model) proposed by Stauffer et al. (Stauffer, Grimson, 1999) compares the pixels of the input image with the background model and then uses the morphological method to extract the moving target. Besides, a new algorithm proposed in (Bewley et al., 2016) called SORT (Simple Online and Realtime Tracking) correlates and matches the objects of each frame to obtain the position coordinates in image sequences, and then uses the least square method to perform trajectory fitting to gain the motion trajectories of multiple

* Corresponding author

https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1303-2020 |

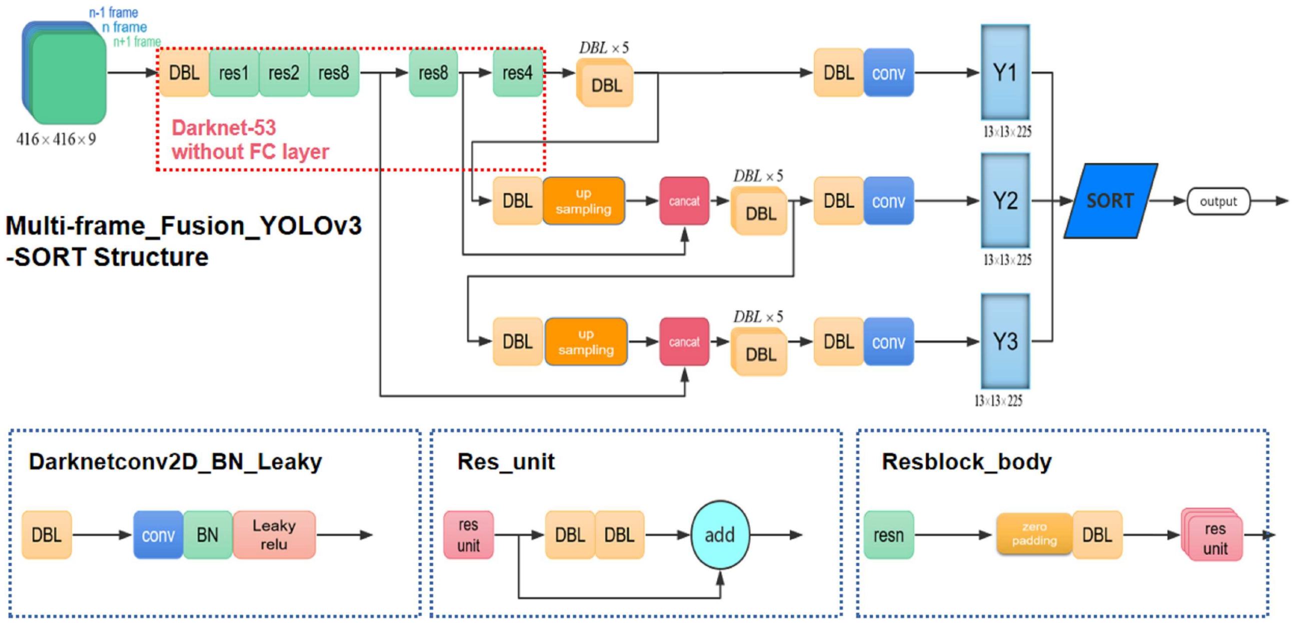

Figure 1. Overview of the proposed method based on YOLO v3 and SORT.

targets. This SORT method is simple and effective.

As mentioned above, there have been successful researches on object detection in high-resolution remote sensing images, especially the methods based on deep learning. However, researches on moving object detection and movement prediction in high-resolution remote sensing videos are still in its infancy, which is worthy of further study. In this paper, we improve the state-of-the-art one-stage deep learning-based object detection method YOLO v3, and use it to detect moving ships in remote sensing videos of Jilin-1 satellite and predict their movement information (i.e. trajectory, velocity, and direction) using SORT method. We have carried out experiments on multi-spectral remote sensing images selected on Google Earth, as well as real multi-spectral remote sensing videos captured by Jilin-1 satellite. The experimental results show that our improved multi-frame YOLO v3 model outperforms the singe-frame model by nearly 8% in terms of the average precision of object detection. Quantitatively, the average position detection error of our multi-frame YOLO v3 model is less than 3 pixels. In addition, our movement prediction method can achieve average prediction errors of trajectory, speed magnitude and speed direction less than 2.5 pixels, 0.2 pixels per second and 2.5 degrees, respectively.

The rest of this paper is organized as follows. Section 2 describes the details of our proposed moving ship detection and movement prediction method. Section 3 introduces the data for training and testing and reports the experimental results. Finally, Section 4 concludes the paper.

2. METHOD

2.1 Overview of the Proposed Method

Comprehensively considering detection accuracy and running speed, we improve the state-of-the-art one-stage deep-learning-based object detection method YOLO v3, and use it to detect moving ships in remote sensing videos of Jilin-1 satellite and predict their movement information, including trajectory, velocity and direction. We first train an optimized YOLO v3 model which can realize ship detection in a single frame of a remote sensing video to select better network structures and initial parameters for our task. Such a single-frame YOLO v3 model can also be regarded as the baseline for comparison. Then, we change the network structure to achieve multi-frame information fusion by expanding the number of the input channels from the original three to nine, to import three consecutive frames of the video together into the model for training. By such multi-frame training, we can utilize the information of continuous frames in a fusion way, and thus improve the detection performance. Moreover, to achieve movement prediction, we use the simple and practical multiple object tracking algorithm SORT to recognize multiple targets detected by multi-frame YOLO v3 model and obtain their coordinates. These coordinates are fitted by the least square method to get the trajectories of multiple targets. We take the derivative of each trajectory to obtain the real-time movement direction and velocity. The structure of our proposed method is shown in Figure 1.

2.2 Multi-Frame Fusion YOLO v3 for Ship Detection

In order to solve the problem of moving ship detection in multispectral remote sensing videos, we use YOLO v3 (Redmon, Farhadi, 2018) model and improve it by making full use of the rich contextual information of sequence images to achieve better detection results. To train a better YOLO v3 model for ship detection in remote sensing videos, we input 3 consecutive image frames into the model, which means the input of YOLO v3 will be changed to $416 \times 416 \times 9$ rather than the original $416 \times 416 \times 3$ for a single RGB input image. Such operation only changes the parameters of the first layer of the YOLO v3 network, while the number of convolution kernels and the size of the output feature map in the first convolutional layer will remain unchanged. In addition, the subsequent structure of the convolutional layer does not need to be changed as well. Therefore, it fully retains the effective performance of YOLO v3. By improving YOLO v3 to a multi-frame input model, the information of continuous frames can be better extracted by the network in a fusion way.

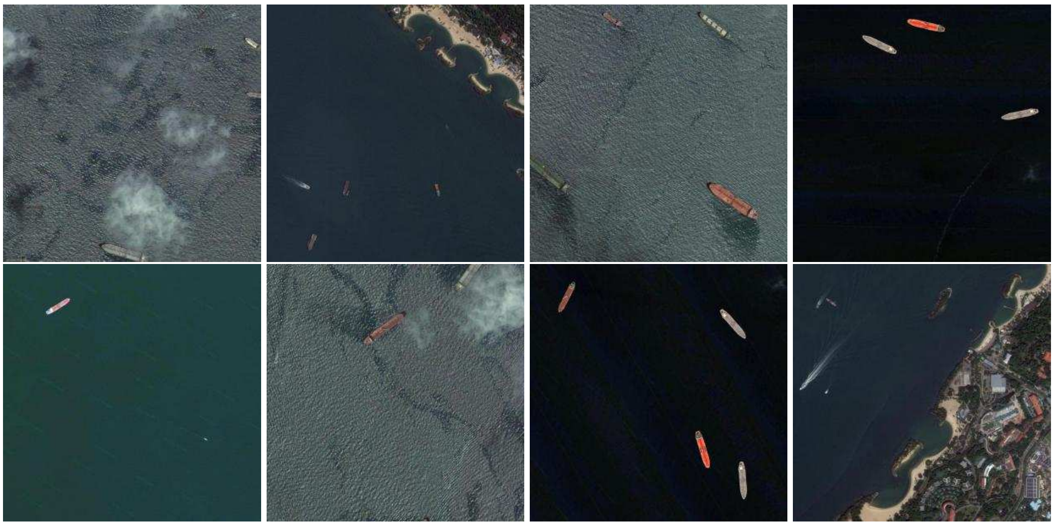

Figure 2. Sample images in the training set.

This method not only reduces much redundancy in the image sequence from the slow movement of the objects, but also enables the model to learn and link multiple frames of information during training.

At the same time, according to the characteristics of ships, we re-cluster the dataset by K-means and obtain 9 new clustering centers for bounding box priors. Considering that the larger size box will cause more errors for ship detection, we use the following formula to define the distance from the ground truth to the cluster center as

$$d(box, centriod) = 1 - IOU(box, centroid) \tag{1}$$

where $centroid$ is the border selected as the center during clustering, box is the border of a ground truth bounding box, and $IOU(\cdot, \cdot)$ is the operation to calculate the intersection over union (IOU) between box and $centroid$, which is a standard for measuring the accuracy of detecting corresponding objects. It is obvious that the distance decreases when the value of IOU increases.

It should be noticed that we adopt the Keras[1] framework for the implement of YOLOv3 model, since Keras makes it easy to adjust the network structure and achieve engineering applications.

2.3 SORT-Based Movement Prediction

To achieve the goal of matching objects between frames and getting their motion information, we choose a very simple and practical algorithm called SORT (Simple Online And Realtime Tracking) (Bewley et al., 2016) to correlate and match the tracking of objects of each frame. We connect SORT directly to the end of multi-frame YOLO v3 detection network mentioned in Section 2.2. If the bounding box matches the objects successfully in a frame, it will update the state of the objects. Then the velocity component will be optimally calculated by the Kalman filtering framework. If not, a linear velocity prediction model will be used to predict the objects. When allocating the bounding box to an existing target, the Kalman filter is used to predict the position of the potential bounding box that the object should appear in this frame. Then, the IOU of the predicted bounding box is calculated. If IOU is less than a threshold, the allocation will be rejected. We use the Pearson correlation coefficient as a standard to measure the tracks, then merge the related trajectories after judgment. The Pearson correlation coefficient is computed as

$$\rho_{XY} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2} \cdot \sqrt{N\sum Y^2 - (\sum Y)^2}} \tag{2}$$

where N is the number of the collected frames, X is the abscissa of the object in the image, and Y is the ordinate of the object in the image.

After obtaining the coordinates of the objects in a frame, we use the least square method to fit the motion trajectories. The instantaneous motion direction and speed are computed by derivation to achieve movement prediction.

3. EXPERIMENTS

3.1 Dataset

3.1.1 Training Set Our remote sensing ship detection requires a dataset that only contains ship objects. However, the number of ship images in publicly available remote sensing image dataset is not enough. Therefore, we select multi-spectral remote sensing images on Google Earth. There are 8022 images of 1m spatial resolution at the size of 1024×1024 from Google Earth. 2250 frames in 9 video clips at the size of 416×416 from a Jilin-1 satellite video (Video 1 in Table 2) are also included in our training set. Table 1 summarizes the information of the training set. Figure 2 shows samples in the training set. It can be seen that the training set contains ship targets under various scenarios including cloud interference, complex seas, moving and stationary ships, island or no island interference, etc.

Spatial Resolution	1m	2m
Size	1024 × 1024	416 × 416
Number	8022	2250
Source	Google Earth	Jilin-1

Table 1. The information of the training set.

[1] https://keras.io/

3.1.2 Testing Set To verify the robustness and performance of our method in different and complex practical environments, we use two real remote sensing videos to construct testing data, as shown in Table 2. These two videos are cropped from two real video data captured by Jilin-1 satellite in the sea area near Hong Kong, whose screen-shot is shown in Figure 3. The duration is about 30 seconds and the spatial resolution is 2 meters. The ships in the video are general civilian ships with a limited number of pixels, which makes the testing more difficult.

Source	Jilin-1 Satellite	
Name	Video 1	Video 2
Spatial resolution	2m	2m
Duration	30s	22s
Number of frames	750	550
Size	1200×1200	416×416
Number of ships	$20 \sim 60$	$2 \sim 5$
Movement	Unobvious	Obvious

Table 2. The information of Jilin-1 satellite videos.

Figure 3. Sample frame in Jilin-1 satellite videos.

In Video 1, each frame contains about 40 ships, and the ships are density with no obvious movement. However, in Video 2, the ships are scattered and less than Video 1, and some ships have obvious moving trajectories. Due to the limitation of memory, we divide Video 1 at the size of 1200×1200 to 9 video clips at the size of 416×416, and select 4,500 frames to test ship detection model while the rest 2250 frames are in the training set. We use Video 2 to test the motion prediction algorithm.

It is worth mentioning that 4500 frames and the single video clip (Video 2) in the testing set are all captured by Jilin-1 satellite, containing about 40 ship targets in various scenarios, such as clouds, complex or calm sea surface, island interference, etc. Therefore, the testing set is suitable to be used for better performance evaluation.

3.1.3 Data Augmentation The 1m resolution training images from Google Earth are resized to 416×416, in order to get similar spatial resolution as Jilin-1 data. In addition, Gaussian filtering has been applied to Google Earth images for data augmentation. Images after Gaussian filtering will be blur and more similar as the video frames of Jilin-1 data. Since performing Gaussian filtering does not change the position and size of the objects, the ground truth labels of ships will not change as well.

3.2 Evaluation Index

We use popular indexes of the Precision-Recall curve and average precision (AP) to evaluate ship detection performance. In terms of movement prediction, we define four indexes for performance evaluation. The first one is Detection Error (DE), i.e. the center coordinate error between the bounding boxes and the ground truth. The second one is Prediction Error (PE) defined as the average error between the predicted center coordinates after trajectory fitting and the ground truth. The third index called Velocity Magnitude Error (VME) is the error between the predicted velocity magnitude and the ground truth. The last index is Velocity Direction Error (VDE), calculated by the difference of predicted velocity direction angles and the ground truth. It should be noticed that all these four indexes are reported as the average values of the frames in the testing video clip.

3.3 Moving Ship Detection

3.3.1 Results of Single Frame YOLO v3 The single frame YOLO v3 model was pre-trained on the MS COCO dataset[2]. Then we fine-tuned it using our remote sensing image training set. We froze the first 249 layers of the network and activated the last 3 layers. The epoch of training at this stage was set to 25. Then the model entered the second stage where we activated all 252 layers, and the epoch at this stage was set to 35. The Precision-Recall curve of the single frame YOLO v3 is shown in Figure 4. The AP of single frame YOLO v3 on the testing set is 78.23%.

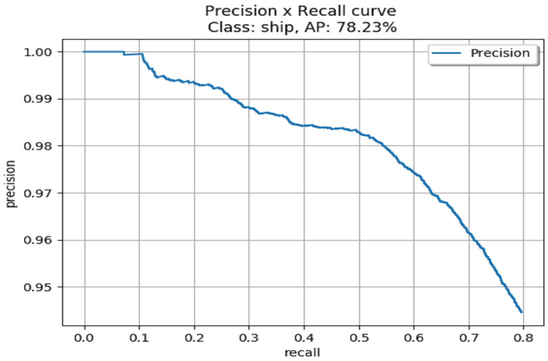

Figure 4. The Precision-Recall curve of the single frame YOLOv3 model.

3.3.2 Results of Multi-Frame YOLO v3 To train a multi-frame YOLO v3, we loaded the single frame YOLO v3 model and initialized the parameters of the first layer, while the parameters of other layers remained unchanged. Then we only activated the first and last three layers for training and froze the rest 248 layers of the network, with the epoch set to 25. After that, we activated all 252 layers and set the epoch to be 35. The Precision-Recall curve of multi-frame YOLO v3 is shown in Figure 6. The multi-frame YOLO v3 can achieve AP of 84.48%, nearly 8% higher than the single frame model. Such experimental results show that our improved multi-frame YOLO v3 model outperforms the original single frame one. The visual comparison between ship detection results of single and multi-frame YOLO v3 is shown in Figure 5.

[2] https://cocodataset.org/

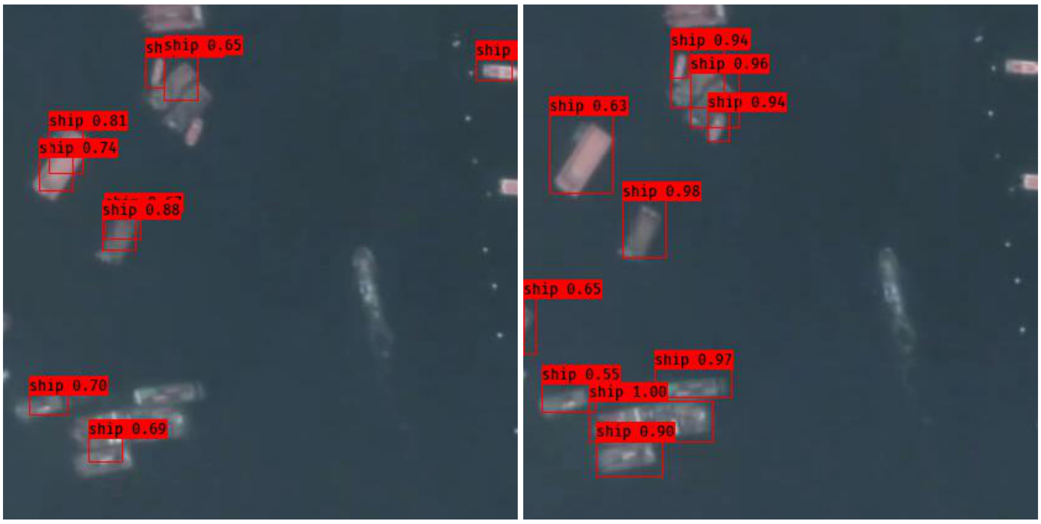

Figure 5. Visualization of ship detection results. The left image is obtained by the single frame YOLO v3 model, while the right one is from the multi-frame YOLO v3 model.

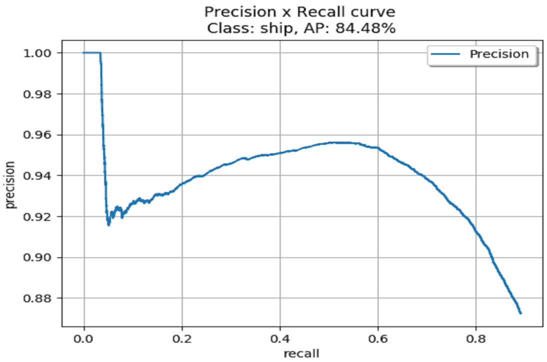

Figure 6. The Precision-Recall curve of the multi-frame YOLOv3 model.

3.4 Movement Prediction

Based on the detection results obtained from the multi-frame YOLOv3 model, we input the information of detection bounding boxes in each frame into the SORT algorithm. The multiple targets were correlated and matched firstly, and then the trajectory was fitted by least square method. We used a trajectory fitted by 400 continuous frames to predict the following 100 frames in the movement prediction process. The results are shown in Table 3 and Figure 7. The experiments show that our movement prediction method can achieve an average trajectory prediction error of less than 2.5 pixels, a speed magnitude prediction error of less than 0.2 pixels per second, and a speed direction prediction error of less than 2.5 degrees.

Object	Ship 1	Ship 2
DE (pixels)	2.5108	2.8539
PE (pixels)	0.8935	2.3113
VME (pixels/second)	0.124	0.068
VDE (degrees)	1.4	2.2

Table 3. the evaluation of the movement prediction

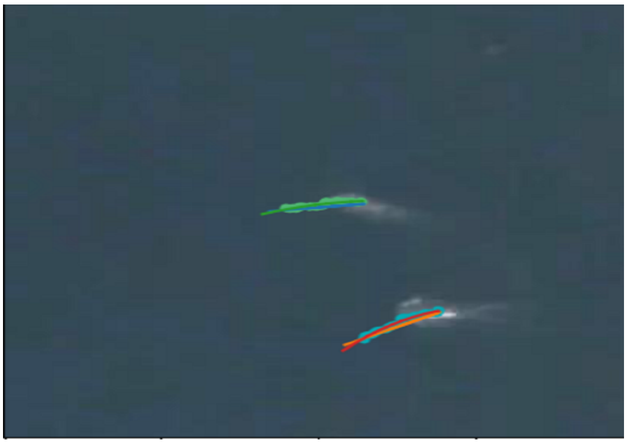

Figure 7. Visualization result of movement prediction. The blue and orange curves represent the real motion trajectories of the two targets. The green and red curves represent the prediction trajectories of the respective targets. The marked center points on the curves are the detection centers.

4. CONCLUSION

In summary, inspired by the methods of moving object detection and movement prediction in natural images, we have proposed a practical solution to achieve moving object detection and movement prediction in multiple spectral remote sensing videos. We performed experiments on multi-spectral remote sensing images and videos from Google Earth and Jilin-1 satellite, respectively. The experimental results validate the effectiveness of our method, which could contribute to providing a feasible way for efficient interpretation and information extraction of new remote sensing video data.

ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China (Grant Nos. 2016YFB0501300 and 2016YFB0501302), the National Natural Science Foundation of China (Grant No. 61501009), and the Fundamental Research Funds for the Central Universities.

REFERENCES

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking. *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 3464–3468.

Chang, Y.-L., Anagaw, A., Chang, L., Wang, Y. C., Hsiao, C.-Y., Lee, W.-H., 2019. Ship Detection Based on YOLOv2 for SAR Imagery. *Remote Sensing*, 11(7), 786.

Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X. et al., 2017. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10), 2896–2907.

Kharchenko, V., Chyrka, I., 2018. Detection of airplanes on the ground using yolo neural network. *2018 IEEE 17th International Conference on Mathematical Methods in Electromagnetic Theory (MMET)*, IEEE, 294–297.

Li, Y., Zhang, Y., Luo, Z., 2018. Airport and Airport Target Recognition Technology Using Deep Convolution Neural Network. *Journal of Chongqing University of Technology (Natural Science)*, 3.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C., 2016. Ssd: Single shot multibox detector. B. Leibe, J. Matas, N. Sebe, M. Welling (eds), *Computer Vision – ECCV 2016*, Springer International Publishing, Cham, 21–37.

Ma, J., Chen, L., Gao, Z., 2017. Hardware implementation and optimization of tiny-yolo network. *International Forum on Digital TV and Wireless Multimedia Communications*, Springer, 224–234.

Redmon, J., Divvala, S. K., Girshick, R. B., Farhadi, A., 2016. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Redmon, J., Farhadi, A., 2017. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517–6525.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (eds), *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., 91–99.

Stauffer, C., Grimson, W. E. L., 1999. Adaptive background mixture models for real-time tracking. *1999 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2, IEEE, 246–252.