Research on Detection Technology of Autonomous Landing Based on Airborne Vision

Jiahuan Li¹,a, Xinhua Wang¹,b, Hengrui Cui²,c, Ziyuan Ma¹,d

¹Nanjing University of Aeronautics and Astronautics, Automation College, Nanjing, China
²Nanjing University of Aeronautics and Astronautics, Electronic information engineering College, Nanjing, China

¹13057583913 @163.com, ²xhwang@nuaa.edu.cn, ³603113370 @163.com,
d310277566 @qq.com

Abstract. In this paper, aiming at the cooperative target detection problem in the process of unmanned helicopter sliding down, a detection method based on complementary filtering is proposed, which fuses improved SSD algorithm and related filtering KCF algorithm. The improved deep learning SSD model redesigns the feature extraction structure to improve the detection effect of small and medium targets for the small target size and large scale change in the landing scene. Then use the detection results of the SSD model to correct the KCF detection, adjust the weight parameters, and output the final fusion detection results. The test results show that the improved model detection accuracy is significantly improved, the detection accuracy in various environments reaches 93.3%, which is higher than 86.1% of the classic SSD model and 87.5% of the Faster-rcnn model. The final proposed fusion detection algorithm has a success rate of 91.1% and a processing speed of 91 hz, which basically satisfies the requirements of the ship.

1. Introduction

Shipborne Unmanned Aerial Vehicle (UAV) can perform many dangerous tasks at sea [1], such as battlefield rescue, reconnaissance, relay guidance, long-range strike and so on. It has become an ideal weapon in modern Maritime Warfare and has been vigorously promoted in many countries. Because of the influence of wave, ship, UAV performance and airflow in the landing area [2], unmanned helicopters are facing many dangers when they take off and land on naval vessels, especially on small and medium-sized ships.

Among many landing navigation modes, visual guidance has many advantages, such as anti-electromagnetic interference, small size, and light weight and so on. Vision-based UAV automatic landing technology has become an important research content in Shipborne UAV technology at home and abroad.[3] proposed installing infrared filters in front of airborne lenses and applied Negative Laplacian of Gaussian (NLOG) operator to detect and track centers of the infrared lamps in the images.[4] used feature points on deck to enhance traditional inertial measurements.[5] used Otsu method and Sobel operator to identify runway and align UAV according to runway position in critical landing stage. All the above schemes detected the feature points of cooperative targets at the end of
landing, while in the stage of landing, the detection effect of cooperative targets is poor due to the influence of sea conditions, distance, image jitter and so on. Therefore, in the process of landing, the whole cooperative target should be detected first, and the effective distance of visual guidance can be extended. For example, [6] proposed an online adaptive visual tracking algorithm, which used online incremental learning method and hierarchical tracking strategy to achieve pose resolution.

The traditional target detection method is affected by illumination and occlusion. The deep learning method adopts the convolutional neural network autonomous learning feature instead of the artificial design feature, which makes the algorithm more robust. The SSD (Single Shot MultiBox Detector) model [7] is a single-stage detection model based on deep learning that has emerged in recent years. Instead of explicitly generating the candidate region, it directly predicts the target category and location. Compared with the two-stage detection model faster-rcnn [8] with the process of generating the proposal, the detection speed is greatly improved, but the speed required for landing is still not reached. The SSD model extracts multi-layer feature maps with different resolutions to improve the detection ability of objects with large scale changes, but the detection accuracy of small targets is still poor. In the landing scenario, the feature extraction structure needs to be redesigned to improve the small and medium target detection capability of the SSD model.

In summary, this paper proposes a detection method based on complementary filter fusion to improve deep learning SSD model and KCF algorithm for the overall detection problem of the cooperation process of the ship landing process. The complementary filtering method uses the detection result of the SSD model to correct the detection result of KCF, reduce the KCF detection error, make up for the shortcoming of the slow detection speed of the SSD model, and improve the real-time detection. For the cooperation target in the landing scene, the feature extraction structure of the multi-scale information of the SSD model is redesigned with the characteristics of small field of view and large scale change, which improves the detection effect of small and medium targets.

2. Improvement of ssd detection model

The SSD model structure is shown in Fig.1. The basic network uses VGG-16-Atrous [9]. Large-scale feature maps contain more detailed information for detecting small objects, and small-scale feature maps extract high-level semantic information for detecting large objects. The SSD model extracts (4, 6, 6, 4, 4) priori boxes of different sizes on the feature maps of Conv4_3, Conv7, Conv8_2, Conv9_2, Conv10_2, and Conv11_2, for a total of 38*38*4+19*19*6+10*10*6+5*5*6+3*3*4+1*1*4=8732 bounding boxes for target detection.

![SSD basic model](image)

**Figure 1. SSD basic model**

2.1. Redesign Feature Extraction Structure

In order to improve the detection ability of small-scale objects, feature maps should take into account strong semantic features and location features. On the one hand, we should have stronger feature extraction ability and global semantic information, on the other hand, we should have enough resolution
to get the location information. In this paper, multi-scale features are fused to enhance the global semantic understanding ability of large resolution feature maps, so as to enhance the detection ability of small targets. Conv3_3 retains the shape details, but the convolution depth is shallow and the semantic features are few [10]. So this paper fuses Conv4_3, Conv7, Conv8_2 three-layer feature map information. Conv4_3, Conv7, Conv8_2 are reduced to 256 dimensions with a 1*1 convolution kernel, and 19*19 Conv7 and 10*10 Conv8_2 are up sampled to 38*38 feature maps of the same size as Conv4_3 by bilinear interpolation. The 3*3 convolution kernel multi-channel convolution is used for the concatenate feature map. Considering that the weights of high and low semantic feature maps are different in fusion, the weights of 3 *3 convolution kernel parameters can be learned by training, so the operation of adding feature maps is not used, and the aliasing effect of up-sampling can be alleviated [11]. Finally, the fusion features (38, 38, 256) are obtained. The feature fusion process is shown in Fig.2.

![Feature Fusion Diagram](image)

**Figure 2.** Feature Extraction Structure Fused with Multi-scale Information

2.2. Generate Priorbox

As the size of the feature map decreases, the scale of the a priori box increases linearly, covering small, medium, and large targets. The input image is 300×300, and the priori box scale of the six feature maps is 

\[
S_k = (30, 60, 111, 162, 213, 264)
\]

The aspect ratio scale generally has 

\[
a_r \in \left\{1, 2, 3, \frac{1}{2}, \frac{1}{3}\right\}
\]

and the width of the a priori frame is 

\[
w_k = S_k \sqrt{a_r}, h_k = S_k / \sqrt{a_r}
\]

At the same time, for each feature map, two square priori boxes with an aspect ratio of 1 but different sizes are set. A scale of 

\[
S_k' = \sqrt{S_k S_{k+1}}
\]

is added to each feature graph, and the width and height of the priori boxes are 

\[
w_k' = h_k = \sqrt{S_k S_{k+1}}.
\]

Analysis of the actual situation of UAV and deck in the landing scene: 1. When the proportion of UAV and deck is 1, 213, 264, UAV is close to the deck, then it can detect and calculate the position and
attitude of the feature points. 2. The target of landing cooperation is circular, and the aspect ratio of detection does not include 3, 1/3. Therefore, removing the tail end of the SSD model to detect large targets reduces the depth of the model, effectively alleviates over-fitting and reduces the difficulty of training. Four feature maps (30, 60, 111, 162) are tested. Each feature map extracts four sizes of priori frames as follows:

\[
\begin{align*}
    a_r &\in \left\{1, 2, \frac{1}{2}\right\} \\
    w'_k &= S_k \sqrt{a_r}, h'_k = S_k / \sqrt{a_r}, r \in \{1, 2, 3\} \\
    w'_k &= h'_k = \sqrt{S_k S_{k+1}}
\end{align*}
\]  

(1)

There are 16 kinds of priori boxes scales, and the scales on 300*300 images are shown in Fig.3.

---

**Figure 3.** Priorbox scale diagram

The improved model is shown in Fig.4. A total of \(38*38*4 + 19*19*4 + 10*10*4 + 5*5*4 = 7720\) priori boxes are found.

---

**Figure 4.** Improved SSD Model
3. Correlation filter tracking

The improved deep learning SSD algorithm improves the detection effect on small and medium targets, but the real-time performance of the algorithm is very important in the actual application of the ship. This paper designs a fusion detection algorithm based on complementary filtering. The correlation filter KCF tracking algorithm is modified by the result of the improved SSD detection module described above, and finally the fusion result is output.

3.1. Basic Principles of KCF

KCF is a discriminant tracking method [12]. The target position in the first frame is selected as a positive sample, and then the positive and negative samples are collected through the region around the circular displacement target. A target detector is trained by ridge regression of the collected sample set. Because the collected sample set is a circular matrix, the circular matrix can be diagonalized in Fourier space, so that the matrix operation can be transformed into the point multiplication of elements, which reduces the computational complexity and improves the computational speed. The target detector is used to detect whether the predicted position of the next frame is the target, and then the new detection result is used to update the training set and then update the target detector.

3.2. Fusion Detection Algorithm Based on Complementary Filtering

![Dual Thread Tracking Strategy](image)

**Figure 5.** Dual Thread Tracking Strategy
The fusion detection algorithm includes a target detection module, a target tracking module, and a complementary filter fusion output module. The target detection module is an improved deep learning SSD detection model to detect cooperative landing targets through contextual semantic information of a single picture; the target tracking module is a correlated filtering kcf tracking algorithm, which achieves continuous tracking of the landing cooperation target by utilizing the continuity of the video moving target.

Finally, the results of the above two modules are combined and output, which realizes the fusion of image time and spatial context information. The flow of designing the fusion detection algorithm is shown in Fig.5.

Complementary filtering is based on the different characteristics of SSD and KCF. The deep learning SSD detection module has high precision, low algorithm speed, and integrates multi-scale feature information to output the position of the ship cooperation target. In the short time, the KCF detection is accurate, and the algorithm rate is high. However, after a long time, the error of the tracking object is large, which causes the error to accumulate. Therefore, the KCF tracking algorithm is first modified by using the results of the SSD detection module: the deep learning SSD detection module runs at a low frame rate of 5 seconds/time. Each time it is 5s, the SSD_flag is set to 1, and the SSD detection module is started. The SSD detection module continuously detects that the detected targets are within the normal deck motion range, and sets the currently detected target position as a KCF tracker candidate area, and sets KCF_flag to 1, and initializes the KCF tracker. The target tracking module uses an associated filter tracking algorithm to calculate a filter template for searching for the location of the target in subsequent image frames based on the position information output by the SSD detection module. The filter template parameters are continuously updated during the tracking process. When the KCF tracking target is lost, set SSD_flag to 1, and start the SSD detection module to initialize the KCF tracker. The above target tracking process is shown in the right part of the Fig.5.

Finally, the SSD detection module and the KCF tracking module output target position results of different frequencies, and interpolate the SSD detection result to make it consistent with the KCF tracking module frequency. The detection and positioning results after the complementation of SSD and KCF are as follows:

\[ pos_{final} = pos_{std}k + pos_{kcf}(1-k) \]

\( pos_{final} \) is the final fusion output result, k is the detection positioning weight of the SSD algorithm, and \((1-k)\) is the detection positioning weight of the KCF algorithm. In this paper, k is taken as 0.5.

4. Experimental results and analysis

The experimental software environment is Ubuntu 16.04 operating system, and the deep learning software framework is pytorch. The experimental hardware environment CPU is Intel (R) Core (R) i7-8700K, and the processing frequency is 3.7GHz. GPU is NVIDIA TITAN Xp, memory is 32G, solid-state hard disk 500G. NVIDIA Jetson TX2 module is selected as the airborne embedded device. Its GPU adopts a new generation of Pascal architecture, and its memory is 6GB.

The design of the cooperative target is as follows: Fig. 6. The white concentric circle is the area of looking at the ship, the gray part is scaled grille, and the H pattern is the assistant sign of relative position and attitude in close range calculation.
The image of the landing cooperation target image actually collected by the camera was used as a
data set. The training set had a total of 900 samples, including different weather conditions such as
clear, fog, and strong light, and samples with different angles of the target rotation. Among them, 517
were small targets and 383 are medium targets. Some samples were shown in Fig.7. The data set was
made into the VOC2007 data set format. The training used a random gradient descent, the batch size
was 32, the learning rate was 0.001, and a total of 25,000 iterations were completed.

4.1. Analysis of Results of Improved SSD Model Detection
In order to verify the detection ability of the improved model's feature layer to the small target, the
detection effect was compared with the low-level feature layer for detecting the small target in the
trained original SSD model and the improved SSD model. The detection result is shown in Figure 8.
The original SSD model had an output detection probability of 0.71 for the small target in the figure,
and the improved SSD model output detection probability was 0.84, which shows that the model has
improved detection ability for small targets.

Under the same training set and test set, three main open source detection models, Faster-rcnn
model, Tiny Yolo model and original SSD model, are compared with the improved model in this paper,
and the accuracy of Ap [13] evaluation index is used to measure the detection effect of the model. And compare the processing speed of desktop devices equipped with TITAN graphics cards and onboard embedded devices. The experimental results are shown in Tab.1.

### Table 1. Comparison of algorithm accuracy and processing speed

| Table Head       | Ap   | Model size/MB | Speed/ms (titan) | Speed/ms (TX2) |
|------------------|------|---------------|------------------|----------------|
| Faster-rcnn      | 87.5%| 540           | 246              | 590            |
| Yolo-darknet     | 72.1%| 236           | 82               | 350            |
| Original SSD model | 86.1%| 95            | 112              | 450            |
| Improved SSD Model | 93.3%| 33            | 100              | 423            |

By comparison, the improved SSD model has an accuracy of 93.3% after the multi-scale features are combined. Compared to the accuracy of the Faster-rcnn model, the detection accuracy is increased by 5.8%, which is 21.2% higher than the yolo model, which is 7.2% higher than the original SSD model. Compared to the original SSD model processing speed, the improved model increased the speed by 10.7% on devices equipped with TITAN graphics cards and 6% on airborne embedded devices.

### 4.2. Complementary Filtering Fusion Algorithm

The three-segment continuous motion video is detected, for a total of 1512 frames. The confidence rate of the deep learning SSD detection algorithm is set to 0.7. When the ratio between the intersection and union of the predicted position frame and the real position frame is more than 80%, the detection is considered successful. The calculation method of detection success rate is: the number of successful frames / the total number of video frames. The tracking area is manually set up for the single KCF tracking algorithm to initialize the tracker. Compare the detection success rate and algorithm processing speed of the single KCF tracking algorithm, the single improved SSD algorithm and the dual thread detection algorithm.

### Table 2. Comparison of success rate and processing speed of each algorithm

| Table Head | SSD detection frames | KCF tracking frames | Success rate | speed / ms |
|------------|----------------------|---------------------|--------------|------------|
| ①          | 0                    | 563                 | 37.2%        | 9          |
| ②          | 1411                 | 0                   | 93.3%        | 42         |
| ③          | 24                   | 1353                | 91.1%        | 11         |

Among them, ① is single KCF tracking algorithm, ② is Improved SSD algorithm and ③ is dual thread detection algorithm. The comparison results are shown in Table.2. The single KCF algorithm is the fastest to process, but it is easy to lose the target after a short period of tracking. Although the single improved SSD algorithm has improved processing speed compared to the original SSD algorithm, it cannot meet the ship real-time requirements. The complementary filtering fusion detection algorithm proposed in this paper makes full use of time domain context and airspace context information. It combines the advantages of high precision of deep learning algorithm and fast real-time performance of KCF tracking algorithm. In the case of slightly sacrificing processing speed, the success rate is increased to 91.1%, and the average processing speed is 9ms. Compared with the single deeping learning detection model, the processing speed of 91Hz can fully meet the requirements of the ship.
5. Conclusion
For the cooperative target detection problem in the process of unmanned helicopter sliding down, there are several unique challenges: the cooperation target is small in field of view, large in scale change, image is easy to blur and jitter, and the real-time requirements for algorithm detection are high. This paper combines multi-scale information and redesigns the feature extraction structure to improve the detection ability of SSD model for small and medium-sized targets. Compared with the Faster-rcnn model, the Tiny yolo model and the original SSD model, which are known for their accuracy, the improved model detection accuracy is significantly improved. Furthermore, the complementary filtering fusion detection algorithm is proposed, combined with the improved deep learning model and the related filtering KCF technology, which improves the processing speed of the algorithm and basically meets the requirements of the ship.

Acknowledgments
Supported by “the Fundamental Research Funds for the Central Universities”, NO.NZ2019008.

References
[1] Miao Xiumei. Development Trend and Analysis of the Ship-based UAV Technology Abroad [J]. Ship Electronic Engineering, 2013, 33(12): 18-22.
[2] LI Qiang, ZHANG Shuli, MENG Wengong. Surveys of Carrier Landing Techniques for UAVs [J]. Unmanned Systems Technology, 2018, 1(02): 43-48.
[3] Gui Y, Guo P, Zhang H, et al. Airborne vision-based navigation method for uav accuracy landing using infrared lamps [J]. Journal of Intelligent & Robotic Systems, 2013, 72(2): 197-218.
[4] Zhou H, Lynch J, Zekkos D. Vision-based precision localization of UAVs for sensor payload placement and pickup for field monitoring applications[C]/Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2019. International Society for Optics and Photonics, 2019, 10970: 1097007.
[5] Nazir S, Aziz S, Khaliq Y, et al. Vision Based Autonomous Runway Identification and Position Estimation for UAV Landing[C]/2018 International Conference on Artificial Intelligence and Data Processing (IDAP). IEEE, 2018: 1-6.
[6] Changhong Fu, Carrio, A., Olivares-Mendez, M.A., Campoy, P.. Online learning-based robust visual tracking for autonomous landing of Unmanned Aerial Vehicles [P]. Unmanned Aircraft Systems (ICUAS), 2014 International Conference on, 2014.
[7] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]/European conference on computer vision. Springer, Cham, 2016: 21-37.
[8] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]/Advances in neural information processing systems. 2015: 91-99.
[9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:1409.1556, 2014.
[10] Li Z, Zhou F. FSSD: feature fusion single shot multibox detector [J]. arXiv preprint arXiv:1712.00960, 2017.
[11] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]/Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.
[12] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]/Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.
[13] Everingham M, Zisserman A, William C K I, et al. The Pascal Visual Object Classes Challenge[C]/International Conference on Machine Learning Challenges, Southamp-ton, UK, April 11-13, 2005. Heidelberg: Springer, 2005: 117-176.