Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.020995

ARTICLE

A Multi Moving Target Recognition Algorithm Based on Remote Sensing Video

Huanhuan Zheng[1,*], Yuxiu Bai[1] and Yurun Tian[2]

[1]School of Information Engineering, Yulin University, Yulin, China

[2]ZTE Communication Co., Ltd., Xi'an, China

*Corresponding Author: Huanhuan Zheng. Email: zhenghuan@yulinu.edu.cn

Received: 11 December 2021 Accepted: 11 February 2022

ABSTRACT

The Earth observation remote sensing images can display ground activities and status intuitively, which plays an important role in civil and military fields. However, the information obtained from the research only from the perspective of images is limited, so in this paper we conduct research from the perspective of video. At present, the main problems faced when using a computer to identify remote sensing images are: They are difficult to build a fixed regular model of the target due to their weak moving regularity. Additionally, the number of pixels occupied by the target is not enough for accurate detection. However, the number of moving targets is large at the same time. In this case, the main targets cannot be recognized completely. This paper studies from the perspective of Gestalt vision, transforms the problem of moving target detection into the problem of salient region probability, and forms a Saliency map algorithm to extract moving targets. On this basis, a convolutional neural network with global information is constructed to identify and label the target. And the experimental results show that the algorithm can extract moving targets and realize moving target recognition under many complex conditions such as target's long-term stay and small-amplitude movement.

KEYWORDS

Deep learning; remote sensing images; moving target; recognition; salient

1 Introduction

Remote sensing images can intuitively display wide view scene information, and are widely applied in the fields. It can be applied to multi-target recognition and tracking scenes such as battlefield reconnaissance, border patrol, post-disaster rescue, public transportation and more [1]. However, it is difficult to obtain spaceborne remote sensing images, which limits the in-depth research [2]. In recent years, with the development and popularization of UAV technology, UAV technology presents the characteristics of high efficiency, flexibility and low cost, and the imaging equipment carried tends to be mature. It has been widely used in battlefield reconnaissance, border patrol, post disaster rescue, public transportation, etc. [3,4], which makes the aerial photography data present a blowout situation. In the face of so many aerial remote sensing data, how we can obtain useful information from aerial video is an important direction in the field of computer vision [5].

Target detection is the basis of tracking, recognition and image interpretation [6]. There are multiple moving targets in the video, just because the video field angle captured by UAV monitoring system is large. Thus, it is difficult to detect moving targets quickly and accurately. Xiao et al. [7] proposed a low frame rate aerial video vehicle detection and tracking method based on joint probability relation graph according to probability relation. Andriluka et al. [8] applied UAV technology to rescue and search. Cheng et al. [9] established a dynamic Bayesian network from the perspective of pixels to realize vehicle detection. Gaszczak et al. [10] established a model from the perspective of thermal imaging to detect pedestrians and vehicles in aerial images. Lin et al. [11] improved the traditional Hough transform to realize road detection. Rodríguez-Canosa et al. [12] established a model to solve the problem of aerial image jitter to a certain extent. Zheng et al. [13] established GIS vector map to realize vehicle detection. Liang et al. [14] guided target detection according to background information. Prokaj et al. [15] built a dual tracker to realize the construction of background and prospect. Teutsch et al. [16] established a model under the condition of low contrast to realize vehicle detection in aerial monitoring images. Chen et al. [17] optimized the tracking window to realize multi-target detection. Jiang et al. [18] introduced the prediction module to realize target tracking in order to enhance the stability of tracking. Poostchi et al. [19] established semantic depth map fusion for moving vehicle detection. Tang et al. [20] established single revolutionary neural networks to predict vehicle direction. Aguilar et al. [21] used cascade classifiers with mean shift to realize pedestrian detection. Xu et al. [22] established a prediction and feature point selection model to find moving targets in multi-scale on infrared aerial images. Hamsa et al. [23] established a cascaded support vector machine and Gaussian mixture model to realize vehicle detection (SVM + GMM). Ma et al. [24] established the rotation invariant cascaded forest (RICF) to meet the target detection in complex background. Mandal et al. [25] established simple short and shallow network to realize rapid target detection. Song et al. [26] built a model according to the time and space relationship of the target, and then built the regulated AdaBoost recognition model to realize target recognition. Qiu et al. [27] built a deep learning network to track moving targets. Feng et al. [28] used mean shift algorithm to track high-speed targets. Wan et al. [29] used Keystone Transform and Modified Second-Order Keystone Transform to achieve moving target tracking. Lin et al. [30] used multiple drones to achieve multi-target tracking.

The above algorithms have achieved certain results in target detection, but there are still deficiencies: The model considers the limited interference of noise, tree disturbance and other factors, which leads to the inaccurate extraction of moving targets. Insufficient mining target attributes lead to inaccurate recognition. The innovation of our algorithm shown as following aspects: firstly, from the perspective of Gestalt vision, we propose a target motion extraction algorithm based on the saliency graph theory; secondly, in order to achieve fast and accurate target recognition, we constructed the convolutional neural network structure with global information, especially take global information into consideration.

2 Algorithm

Gestalt visual school believes that the reason why things are perceived is the result of the public action of eyes and brain. Firstly, the image is obtained through the eyes, and then the objects are combined according to some rules to form an easy to understand unity. If it cannot be combined, it will appear in a disordered state, resulting in incorrect cognition. Based on this principle, a multi-target recognition process in accordance with the Gestalt vision principle is established, as shown in Fig. 1. Firstly, the salient graph mechanism is established to extract moving targets, and then the convolution neural network structure based on global information is used to realize target recognition.

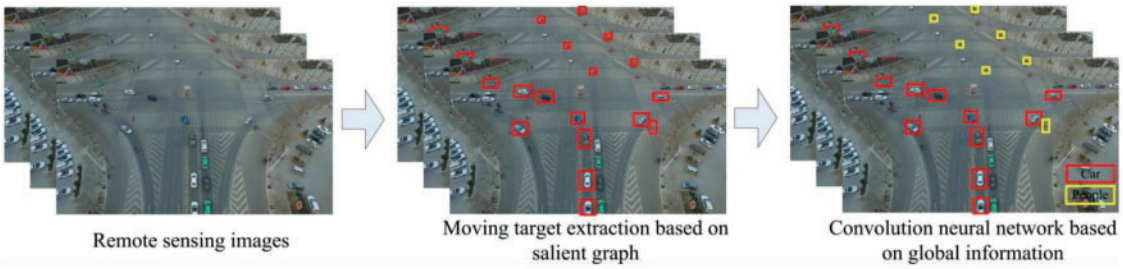

Figure 1: Algorithm pipeline

2.1 Moving Target Extraction Based on Salient Gragh

In the video scene, a large amount of short-term motion information is included between consecutive image frames, and ignores the information that does not move temporarily. A video with long time contains a lot of long-time motion information, so the conditional motion salient graph includes the motion saliency of the target and the motion saliency of the background. In addition, the interference of noise makes it difficult to distinguish the significance.

A time series group containing short-term motion information and long-term motion information is constructed on the time scale. By calculating the motion significance probability, the significance of the moving target is highlighted and the background significance is suppressed, as shown in Fig. 2.

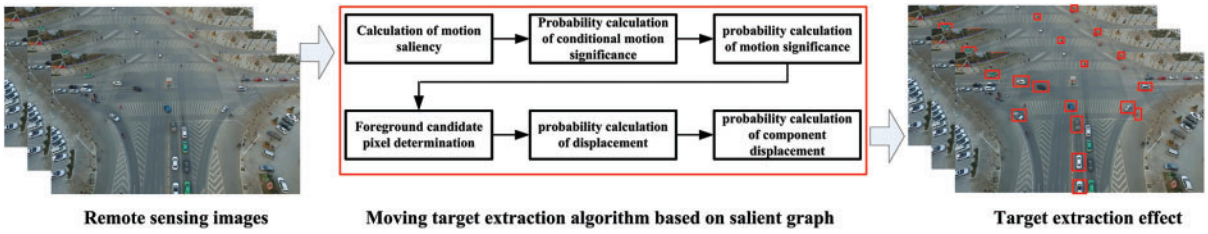

Figure 2: Pipeline of moving target extraction algorithm based on salient graph

Temporal Fourier transform (TFT) is a motion saliency detection algorithm based on time information which uses pixels at the same position in consecutive frames to integrate in time to form a time series. The waveform is reconstructed by Fourier transform and inverse Fourier transform, and the maximum point is marked on the salient graph. The significant values indicate the probability that the target belongs to the foreground. Let the time series be

$$D^i_{x,y}(t) = \left\{I_{x,y}(t), I_{x,y}(t+1), \ldots I_{x,y}(t+i)\right\} \tag{1}$$

Its corresponding Fourier transform is obtained as

$$f^i_{x,y}(t) = F\left(D^i_{x,y}(t)\right) \tag{2}$$

$$p^i_{x,y}(t) = angle\left(f^i_{x,y}(t)\right) \tag{3}$$

where F represents Fourier transform and $p^i_{x,y}(t)$ represents the phase spectrum of $f^i_{x,y}(t)$.

$$I^i_{x,y}(t) = g(t) * F^{-1}\left(p^i_{x,y}(t)\right) \tag{4}$$

where F^{-1} represents inverse Fourier transform, and $g(t)$ is Gaussian filter. The larger the amplitude of $I^i_{x,y}$ change, the larger the time scale change. Thus, the motion significance of the time series is

constructed as

$$\varphi_{x,y}^{i} = \max\left(\left\|I_{x,y}^{i}(t)\right\|\right) \tag{5}$$

where $\varphi^{i}{}_{x,y}$ represents the significant value of the pixel (x, y) in the i-th viewing angle sequence, $\|\ \|$ is F2 norm.

Conditional motion saliency probability refers to the probability that pixels belong to the foreground in the time series. To unify the scale, $\varphi_{x,y}^{i}$ is normalized as

$$P_F\left(I_{x,y}(t)\,|D_{x,y}^{i}(t)\right) = \frac{\varphi_{x,y}^{i} - \min(\varphi^{i})}{\max(\varphi^{i}) - \min(\varphi^{i})} \tag{6}$$

where $P_F\left(I_{x,y}(t)\,|D_{x,y}^{i}(t)\right)$ represents the normalized value of $\varphi_{x,y}^{i}$, and φ^{i} is the time series significant image composed of the current frame at time t and the i-th time slice. The motion saliency value is positively correlated with the probability that the corresponding point belongs to the foreground.

The motion saliency probability represents the probability that the pixel belongs to the foreground by the motion saliency of the pixel. Under the guidance of the full probability formula, it can be calculated as follows:

$$P_F\left(I_{x,y}(t)\right) = \sum_{i=1}^{l} P_F\left(I_{x,y}(t)\,|D_{x,y}^{i}(t)\right) P\left(D_{x,y}^{i}(t)\right) \tag{7}$$

The motion saliency probability graph uses the long-term and short-term motion information to enhance the saliency of the moving target in the current frame and suppress the motion saliency of the background and historical frame. Due to the background interference, there is false detection in the detection results, which makes the local significance high and difficult to remove by traditional methods. We use the correlation between adjacent pixels to construct a histogram algorithm to segment the saliency graph, then model the spatial information, and use the spatial information modeling to calculate the displacement probability, which has achieved the purpose of eliminating interference.

The histogram based threshold method is used to segment the motion saliency probability graph to obtain the candidate pixels,

$$S = \begin{cases} 1 & P_F\left(I_{x,y}(t)\right) \geq T \\ 0 & other \end{cases} \tag{8}$$

where T is obtained by the traditional Ostu algorithm. When $S = 1$, it indicates foreground candidate pixels. When $S = 0$, it indicates background candidate pixels. When the background, i.e., trees, moves, there is a risk of detection as the target. However, its inter frame motion amplitude is limited. We construct the function:

$$\begin{aligned} P_N\left(I_{x,y}(t)\right) &= \max_{(x,y)\in N(x,y)}\left(P_B\left(I_{x,y}(t)\right)\right) \\ P_B\left(I_{x,y}(t)\right) &= 1 - P_F\left(I_{x,y}(t)\right) \end{aligned} \tag{9}$$

where $P_B\left(I_{x,y}(t)\right)$ is the probability that the pixel (x, y) belongs to the background. The above algorithm can suppress the background, but it will also eliminate the real moving target pixels. We limit the detected connected region and measure it by considering the neighborhood position of the whole foreground target. Define the component displacement probability as the probability of a detected

connected component:

$$P_c\left(I_{x,y}(t)\right) = \prod_{(x,y)\in C} P_N\left(I_{x,y}(t)\right) \tag{10}$$

For the connected component of the real target, the probability of the component displacement from the background is very small, and the threshold t_h is set to distinguish:

$$S_c(x,y) = \begin{cases} 1 & P_c(x,y) \le t_h \\ 0 & P_c(x,y) > t_h \end{cases} \tag{11}$$

where $S_c(x, y) = 1$ represents the foreground and $S_c(x, y) = 0$ represents the background.

2.2 Convolution Neural Network Based on Global Information

Convolutional neural network (CNN) is a kind of feedforward neural network, whose neurons carry out corresponding control on the units within the coverage. It has excellent performance in the field of large-scale image processing. Therefore, we process remote sensing images based on traditional CNN.

The basic structure of CNN includes feature extraction layer and feature mapping layer. Feature extraction layer: The input of each neuron is connected to the local acceptance domain of the previous layer, and the local features are extracted. When the local feature is extracted, the position relationship between it and other features is also determined. Feature mapping layer: Each computing layer of the network is composed of multiple feature maps. Each feature mapping can be regarded as a plane, and the weights of all neurons on the plane are equal. Because the feature detection layer of CNN learns from the training data, it avoids explicit feature extraction and implicitly learns from the training data. Because the weights of neurons on the same feature mapping surface are the same, the network can learn in parallel. Subsequent scholars have carried out a lot of research on the basis of CNN: Bayar et al. [31] constrained the convolution layer to meet the image target detection. Li et al. [32] used a double-layer CNN structure to detect targets. The above algorithms have improved CNN from different aspects and achieved certain results.

According to the particularity of remote sensing video, traditional CNN cannot be directly applied to multi-target tracking based on remote sensing video, which is not enough to capture global information. Therefore, a new convolution neural network framework based on global information is proposed by effectively combining global average pool and Atrous convolution, as shown in Fig. 3.

The image sequence is convoluted and pooled to reduce the size of the image and increase the receiving domain. The merged image is restored by up sampling to the original size prediction of the image. The information in the original image is lost while zooming out and adjusting. To solve this problem, Atrous Convolution is applied, as shown in Fig. 4. Atrus revolution is a convolution idea proposed to solve the problem of image semantic segmentation in which down sampling will reduce image resolution and lose information. The advantages are: on the condition of loss information without pooling and the same calculation conditions, the receptive field is increased so that each convolution output contains a large range.

Global information plays a key role in image classification or target detection. In order to obtain more global context information, the global average pool is combined with the atrus revolution. Fig. 5 shows the main structure of the network.

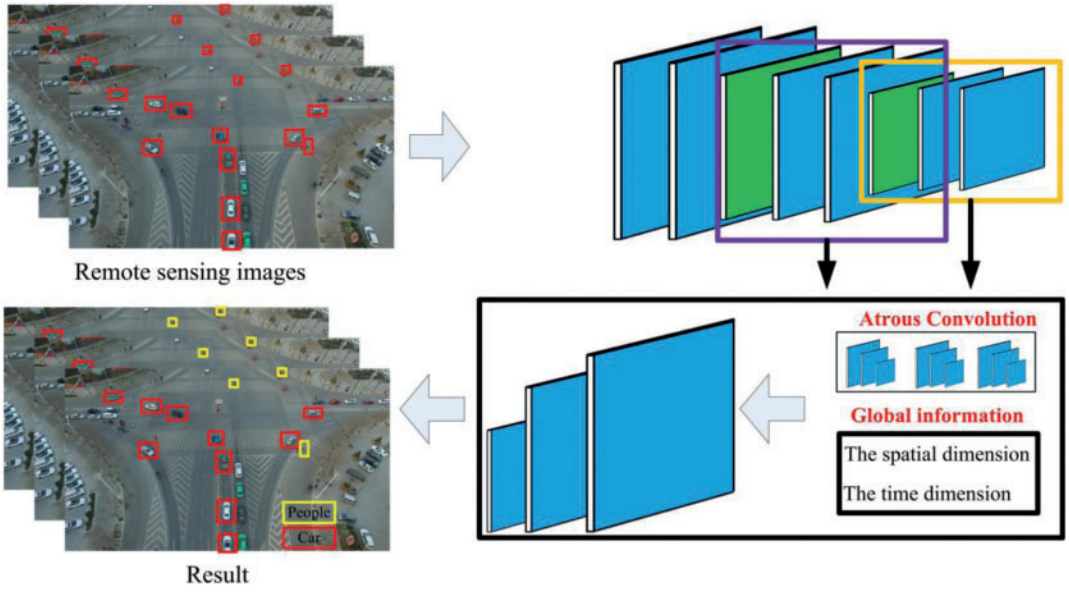

Figure 3: Convolutional neural network graph based on global information

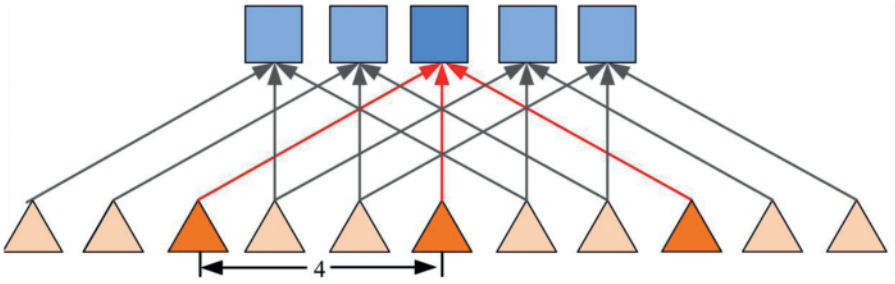

Figure 4: Atrous convolution

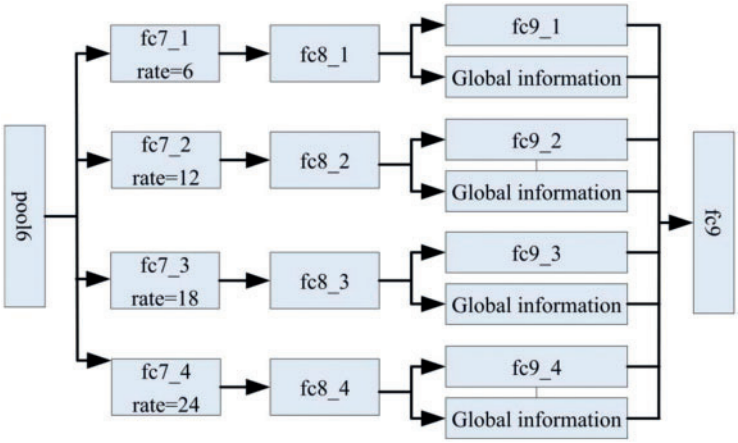

Figure 5: The network structure

When extracting global features for feature fusion, due to different data scales at different levels, it is necessary to regularize the feature data.

$$\mathbf{y} = \frac{\mathbf{X}}{\|\mathbf{X}\|_2} \tag{12}$$

We use L2-normalize to normalize the input feature x, where y represents the normalized output vector. ||• | is L2 norm.

Because the value distribution of the eigenvector is uneven, the scale parameter is introduced as

$$\mathbf{z} = \alpha \mathbf{y} \tag{13}$$

In the training process, L2 norm propagation is used to calculate the scale parameters through the chain method:

$$\begin{cases} \dfrac{\partial l}{\partial y_i} = \dfrac{\partial l}{z_i}\alpha \\ \dfrac{\partial l}{\partial \alpha} = \sum\limits_j \dfrac{\partial l}{\partial z_j} y_j \\ \dfrac{\partial y_i}{x_i} = \dfrac{\|\mathbf{x}\|_2^2 - x_i^2}{\|\mathbf{x}\|_2^3} \end{cases} \tag{14}$$

where l is the objective function. L2 norm normalizes the extracted features and the added global features.

The median frequency balance strategy is used, and the cross entropy loss function is

$$l(p, p^*) = -\frac{1}{n}\sum_i p^* \log p_i \,,\; p_i = \frac{A_i}{\sum\limits_c A_i^c} \tag{15}$$

as the objective function in the network training process. It is to determine the distance between the actual output P and the desired output p*, where c is the class of the tag.

Because the number of pixels in each category in the training data varies greatly, different weights are required according to the actual category. In order to obtain better results, the median frequency balance is proposed, and Eq. (4) is rewritten as follows:

$$\begin{cases} l(p, p^*) = -\dfrac{1}{n}\sum\limits_i p^* \log(p_i) w_c \\ w_c = \dfrac{M(f(c) \,|\, c \in C)}{f(c)} \end{cases} \tag{16}$$

where w_c is the adjustment weight, $f(c)$ is the proportion of pixel value c to the total number of pixels. Through the different loss weights of real classes in the data to balance the categories, we can achieve better classification results.

In the process of deep neural network training, dropout is used in the full connection layer of convolutional network to prevent network over fitting [33]. However, in FCN, dropout layer cannot improve the network generalization ability. To solve this problem, dropblock is introduced

$$\gamma = \frac{1 - kp}{bs^2} \frac{fs^2}{(fs - bs + 1)^2} \tag{17}$$

where γ is used to control the number of channels removed from each convolution result. bs is used to control the block size to 0. kp is the dropout parameter. fs is the dimension of the characteristic

diagram. dropblock makes the training network learn more robust features and greatly improves the generalization ability of the network.

3 Experiment and Result Analysis

The algorithm proposed in this paper is programmed and applied in the window system, VS2010 platform. Window 7, Intel® Core i5-6500 CPU, 3.20 GHZ, 16.0 GB and uses the deep network to extract features. We normalized the image to 512×512. At present, the average processing time of a single frame image is 1.2 s, which temporarily cannot meet the needs of real-time computing, and further research will be conducted in the future.

3.1 Database

3 datasets: The UA-DETRAC [34] dataset is a 10 h video shot by Canon EOS 550D camera at 24 different locations with a frame frequency of 25 fps and a resolution of 960×540 pixels. The UAV-DATA [35] dataset contains scenes such as trees and highways, including the characteristics of large-scale and multiple moving targets. The total video is 4.89 GB, and the minimum and maximum image resolution are 1920×1080 and 3840×2160, respectively. The minimum and maximum frame rate is 4 fps and 25 fps, respectively. The campus environment database is the shooting data over the playground, which contains a large number of small targets, as shown in Fig. 6.

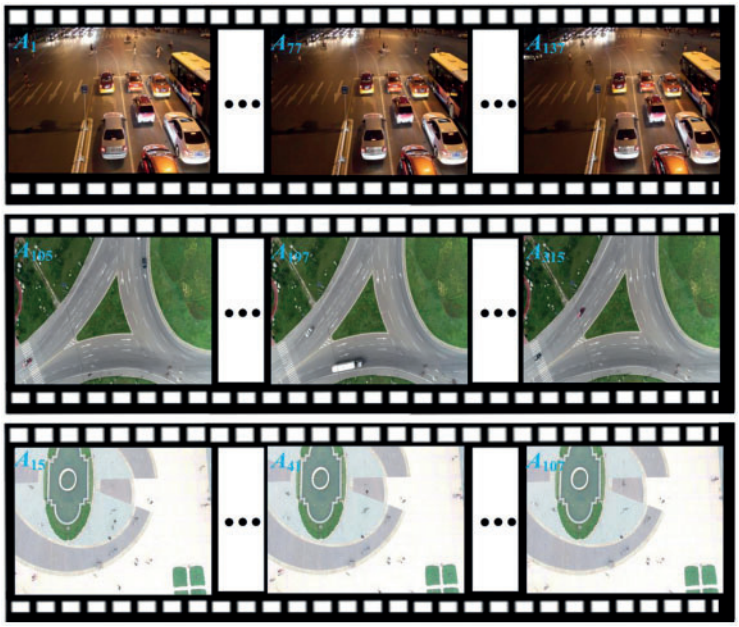

Figure 6: Data display

3.2 Detection Accuracy

To measure the effectiveness of the algorithm, AOM and ROC curves are introduced

$$\mathrm{AOM}(\gamma, \nu) = \frac{\mathrm{S}(\gamma \cap \nu)}{\mathrm{S}(\gamma \cup \nu)} \times 100\% \tag{18}$$

where γ represents the result of manual annotation and ν represents the detection result of the algorithm.

Based on the dataset described above, we have divided the data into **Data 1:** The first frame is a pure background image, and the target is moving all the time. **Data 2:** Small moving targets. **Data 3:** It remains stationary for a long time after the target moves.

As shown in Table 1 and Fig. 7, with the complexity of the environment, the performance of the algorithm shows a downward trend. SVM + GMM algorithm uses GMM model to extract motion region, and SVM is used to judge whether the region is foreground. RICF detects the target according to the rotation invariance of the target. Time space [26] extracts the moving region according to the relationship between time and space, and establishes AdaBoost model to realize multi-scale target recognition. Under the guidance of gestalt vision, the proposed algorithm establishes a saliency graph mechanism to extract moving targets, and then realizes target recognition based on the convolution neural network structure of global information. Although the operation speed is slightly lower than SVM + GMM, AOM is the highest.

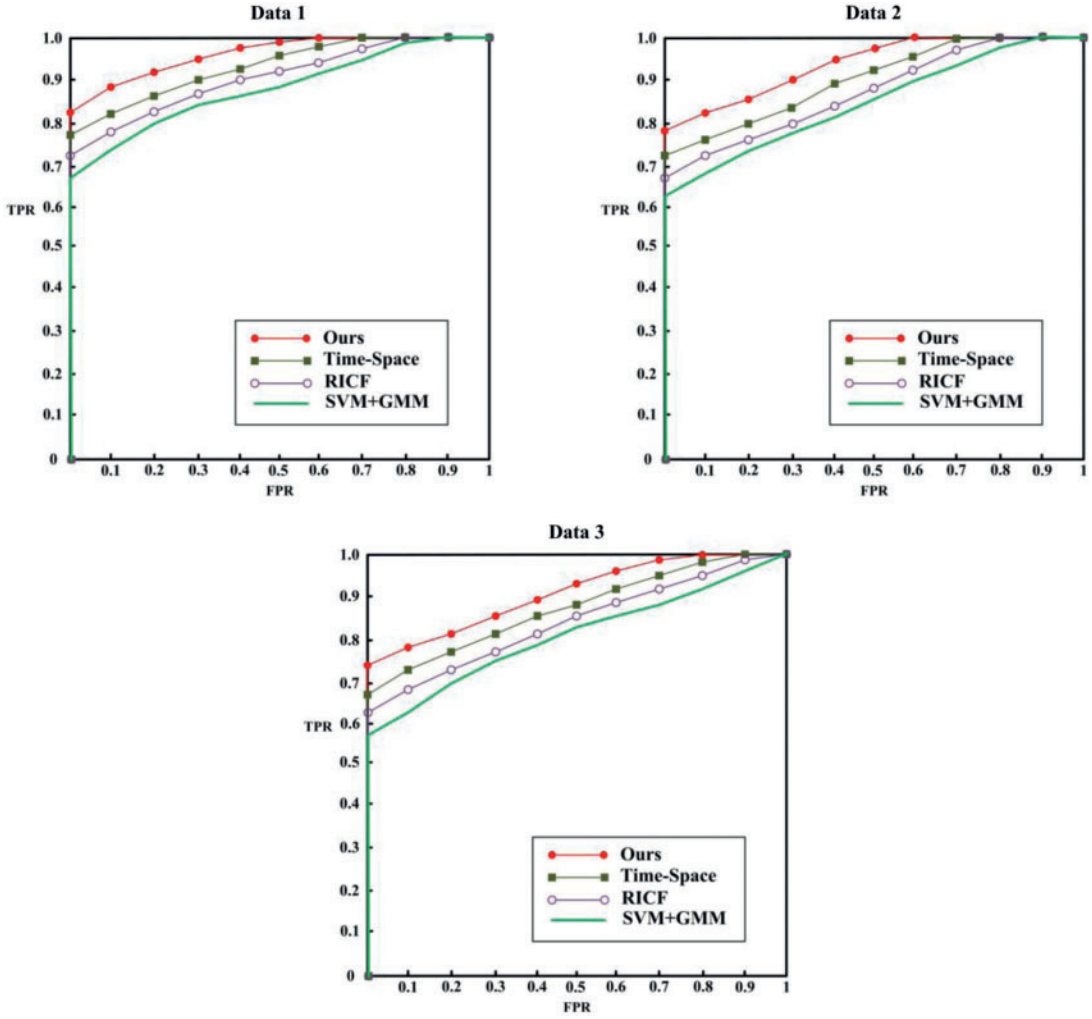

Figure 7: Region of interest curves

Table 1: AOM and operation time

Algorithm	Data 1	Data 2	Data 3	The average time (S/frame)
SVM+GMM	80.1	76.2	72.5	0.03
RICF	81.8	78.5	74.8	0.08
Time-Space	82.5	79.2	75.6	0.09
OURs	83.1	80.3	78.1	0.05

3.3 Effect of Target Extraction

In order to intuitively show the effect of our algorithm, some detection results are selected, as shown in Fig. 8. The proposed algorithm can effectively detect the target, and has good detection effect in the face of complex background, unstable target motion and small targets.

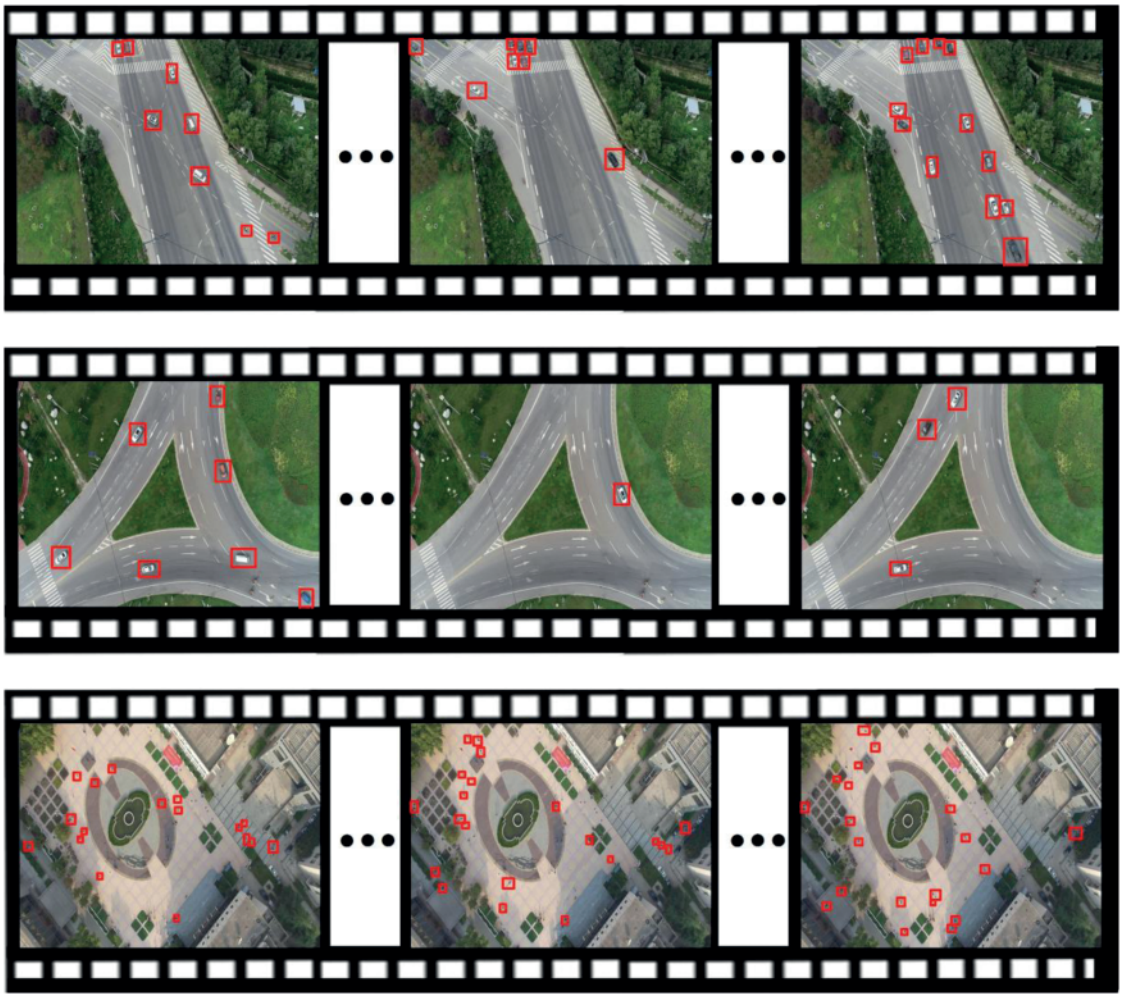

Figure 8: Detection results images

4 Conclusion

Remote sensing images can obtain ground and target's information intuitively so as to provide accurate basis for decision-making. To solve the problem that it is difficult to accurately extract moving objects under complex conditions such as long-term stay and small-amplitude motion of moving objects in remote sensing videos, so we proposed a multi-moving object recognition algorithm based on remote sensing videos. Firstly, the problem of moving target detection is transformed into the problem of salient region probability, and the saliency map is constructed to extract moving targets. Secondly, by analyzing the global and local information of multiple targets, a convolutional neural network with global information is constructed to identify the target. Experiments show that the research results have a better effect on multi-target extraction in complex environments, and provide a new method for multi-target tracking in remote sensing images. So, on this basis, follow-up research on ground feature analysis can be carried out accordingly.

Funding Statement: This work is supported by Yulin Science and Technology Association Youth Talent Promotion Program (Grant No. 20200212).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Shao, Z., Wang, L., Wang, Z., Deng, J. (2019). Remote sensing image super-resolution using sparse representation and coupled sparse autoencoder. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(8),* 2663–2674. DOI 10.1109/JSTARS.2019.2925456.
2. Ma, J., Jiang, J., Zhou, H., Zhao, J., Guo, X. (2018). Guided locality preserving feature matching for remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing, 56(8),* 4435–4447. DOI 10.1109/TGRS.2018.2820040.
3. Cao, X., Jiang, X., Li, X., Yan, P. (2016). Correlation-based tracking of multiple targets with hierarchical layered structure. *IEEE Transactions on Cybernetics, 48(1),* 90–102. DOI 10.1109/TCYB.2016.2625320.
4. Farmani, N., Sun, L., Pack, D. J. (2017). A scalable multitarget tracking system for cooperative unmanned aerial vehicles. *IEEE Transactions on Aerospace and Electronic Systems, 53(4),* 1947–1961. DOI 10.1109/TAES.2017.2677746.
5. Chao, H., Cao, Y., Chen, Y. (2010). Autopilots for small unmanned aerial vehicles: A survey. *International Journal of Control, Automation and Systems, 8(1),* 36–44. DOI 10.1007/s12555-010-0105-z.
6. Gleason, J., Nefian, A. V., Bouyssounousse, X., Fong, T., Bebis, G. (2011). Vehicle detection from aerial imagery. *2011 IEEE International Conference on Robotics and Automation*, pp. 2065–2070. Shanghai, China, IEEE. DOI 10.1109/ICRA.2011.5979853.
7. Xiao, J., Cheng, H., Sawhney, H., Han, F. (2010). Vehicle detection and tracking in wide field-of-view aerial video. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 679–684. San Francisco, USA, IEEE. DOI 10.1109/CVPR.2010.5540151.
8. Andriluka, M., Schnitzspan, P., Meyer, J., Kohlbrecher, S., Petersen, K. et al. (2010). Vision based victim detection from unmanned aerial vehicles. *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1740–1747. Taipei, Taiwan, IEEE. DOI 10.1109/IROS.2010.5649223.
9. Cheng, H. Y., Weng, C. C., Chen, Y. Y. (2011). Vehicle detection in aerial surveillance using dynamic Bayesian networks. *IEEE Transactions on Image Processing, 21(4),* 2152–2159. DOI 10.1109/TIP.2011.2172798.

10. Gaszczak, A., Breckon, T. P., Han, J. (2011). Real-time people and vehicle detection from UAV imagery. *Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques,* vol. 7878. San Francisco Airport, California, USA. DOI 10.1117/12.876663.
11. Lin, Y., Saripalli, S. (2012). Road detection from aerial imagery. *2012 IEEE International Conference on Robotics and Automation*, pp. 3588–3593. Saint Paul, MN, USA, IEEE. DOI 10.1109/ICRA.2012.6225112.
12. Rodríguez-Canosa, G. R., Thomas, S., Del Cerro, J., Barrientos, A., MacDonald, B. (2012). A Real-time method to detect and track moving objects (DATMO) from unmanned aerial vehicles (UAVs) using a single camera. *Remote Sensing, 4(4),* 1090–1111. DOI 10.3390/rs4041090.
13. Zheng, Z., Zhou, G., Wang, Y., Liu, Y., Li, X. et al. (2013). A novel vehicle detection method with high resolution highway aerial image. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(6),* 2338–2343. DOI 10.1109/JSTARS.2013.2266131.
14. Liang, P., Ling, H., Blasch, E., Seetharaman, G., Shen, D. et al. (2013). Vehicle detection in wide area aerial surveillance using temporal context. *Proceedings of the 16th International Conference on Information Fusion*, pp. 181–188. Istanbul, Turkey, IEEE.
15. Prokaj, J., Medioni, G. (2014). Persistent tracking for wide area aerial surveillance. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1186–1193. Columbus, USA.
16. Teutsch, M., Krüger, W., Beyerer, J. (2014). Evaluation of object segmentation to improve moving vehicle detection in aerial videos. *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 265–270. Seoul, Korea, IEEE. DOI 10.1109/AVSS.2014.6918679.
17. Chen, B. J., Medioni, G. (2015). Motion propagation detection association for multi-target tracking in wide area aerial surveillance. *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6. Karlsruhe, Germany, IEEE. DOI 10.1109/AVSS.2015.7301766.
18. Jiang, X., Cao, X. (2016). Surveillance from above: A detection-and-prediction based multiple target tracking method on aerial videos. *2016 Integrated Communications Navigation and Surveillance (ICNS)*, pp. 4D2–1. Herndon, USA, IEEE. DOI 10.1109/ICNSURV.2016.7486348.
19. Poostchi, M., Aliakbarpour, H., Viguier, R., Bunyak, F., Palaniappan, K. et al. (2016). Semantic depth map fusion for moving vehicle detection in aerial video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 32–40. Las Vegas, USA.
20. Tang, T., Zhou, S., Deng, Z., Lei, L., Zou, H. (2017). Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks. *Remote Sensing, 9(11),* 1170. DOI 10.3390/rs9111170.
21. Aguilar, W. G., Luna, M. A., Moya, J. F., Abad, V., Parra, H. et al. (2017). Pedestrian detection for UAVs using cascade classifiers with meanshift. *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pp. 509–514. San Diego, USA, IEEE. DOI 10.1109/ICSC.2017.83.
22. Xu, W., Zhong, S., Yan, L., Wu, F., Zhang, W. (2018). Moving object detection in aerial infrared images with registration accuracy prediction and feature points selection. *Infrared Physics & Technology, 92,* 318–326. DOI 10.1016/j.infrared.2018.06.023.
23. Hamsa, S., Panthakkan, A., Al Mansoori, S., Alahamed, H. (2018). Automatic vehicle detection from aerial images using cascaded support vector machine and Gaussian mixture model. *2018 International Conference on Signal Processing and Information Security ICSPIS*, pp. 1–4. Dubai, United Arab Emirates, IEEE. DOI 10.1109/CSPIS.2018.8642716.
24. Ma, B., Liu, Z., Jiang, F., Yan, Y., Yuan, J. et al. (2019). Vehicle detection in aerial images using rotation-invariant cascaded forest. *IEEE Access, 7,* 59613–59623. DOI 10.1109/ACCESS.2019.2915368.
25. Mandal, M., Shah, M., Meena, P., Vipparthi, S. K. (2019). Sssdet: Simple short and shallow network for resource efficient vehicle detection in aerial scenes. *2019 IEEE International Conference on Image Processing*, pp. 3098–3102. Taipei, Taiwan, IEEE. DOI 10.1109/ICIP.2019.8803262.
26. Song, P., Si, H., Zhou, H., Yuan, R., Chen, E. et al. (2020). Feature extraction and target recognition of moving image sequences. *IEEE Access, 8,* 147148–147161. DOI 10.1109/ACCESS.2020.3015261.

27. Qiu, S., Cheng, K., Cui, L., Zhou, D., Guo, Q. (2020). A moving vehicle tracking algorithm based on deep learning. *Journal of Ambient Intelligence and Humanized Computing,* 1–7. DOI 10.1007/s12652-020-02352-w.
28. Feng, Z. (2021). High speed moving target tracking algorithm based on mean shift for video human motion. *Journal of Physics: Conference Series, 1744(4),* 42180. IOP Publishing.
29. Wan, J., Tan, X., Chen, Z., Li, D., Liu, Q. et al. (2021). Refocusing of ground moving targets with Doppler ambiguity using keystone transform and modified second-order keystone transform for synthetic aperture radar. *Remote Sensing, 13(2),* 177. DOI 10.3390/rs13020177.
30. Lin, C., Shi, J., Huang, D., Zhang, W., Li, W. (2022). Tracking strategy of multiple moving targets using multiple UAVs. *Advances in Guidance, Navigation and Control,* vol. 644, pp. 1417–1425. Singapore: Springer. DOI 10.1007/978-981-15-8155-7_118.
31. Bayar, B., Stamm, M. C. (2018). Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security, 13(11),* 2691–2706. DOI 10.1109/TIFS.2018.2825953.
32. Li, W., Dong, R., Fu, H., Yu, L. (2019). Large-scale oil palm tree detection from high-resolution satellite images using two-stage convolutional neural networks. *Remote Sensing, 11(1),* 11. DOI 10.3390/rs11010011.
33. Achille, A., Soatto, S. (2018). Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12),* 2897–2905. DOI 10.1109/TPAMI.2017.2784440.
34. UA-DETRAC. http://detrac-db.rit.albany.edu/.
35. Yang, T., Li, D., Bai, Y., Zhang, F., Li, S. et al. (2019). Multiple-object-tracking algorithm based on dense trajectory voting in aerial videos. *Remote Sensing, 11(19),* 2278. DOI 10.3390/rs11192278.