Review for Region Localization in Large-Scale Optical Remote Sensing Images

Shoulin Yin, School of Information and Communication Engineering, Harbin Engineering University, China*
https://orcid.org/0000-0002-5367-1372

Lin Teng, Shenyang Normal University, China

ABSTRACT

For the massive large-scale visible image data obtained by satellite, unmanned aerial vehicles, and other reconnaissance platforms, if only relying on manual visual interpretation, there will be problems such as heavy workload, low efficiency, high repeatability, strong subjectivity, and high cost, which cannot meet the demand of modern society for efficient information. Therefore, in order to improve work efficiency, it is necessary to study the rapid automatic region localization in large-scale remote sensing images. That will play an important role in change detection, temperature retrieval, and other files. The development and present situation of the region localization algorithms are analyzed. This paper summarizes the development, improvement, and deficiency of the traditional algorithm, as well as the difficulties and challenges. And the authors make a comparison to the deep learning-based methods. Finally, a possible development direction is prospected.

KEYWORDS

Deep Learning, Large-Scale Remote Sensing Image, Region Localization

1. INTRODUCTION

Region localization is one of the important research directions in the field of computer vision. The goal of region location is to determine the position of an object region in the remote sensing image. At present, the common localization method is to use the supervised learning algorithm (Bahrami et al., 2021; Liu et al., 2016; Teng et al., 2017) to complete the object region localization in the test set according to the region category and location information training. The task of object region localization is to find out all interested regions in the image and determine their position and size, which is one of the core problems in the field of machine vision. Region localization is used in many scenarios, such as airport, harbor detection. Due to various regions have different appearance, shape, posture, imaging light and shielding factors, region localization has been the most challenging problem in machine vision. In many practical applications, such as small object detection (Elakkiya et al., 2021; Gong et al., 2021), traffic object detection (Fan et al., 2021; Ye et al., 2020), multi-modal object detection (Yin et al., 2018), medical object detection (Xi et al., 2020) and other tasks, data shortage and numerous missing marks cannot meet the requirements of neural network detection

DOI: 10.4018/IJISTA.306654 *Corresponding Author

tasks. And in these applications, the data set differences between categories are bigger, the part of the unlabeled objects seriously polluted the background feature space. The classifier is difficult to distinguish the differences between the known categories and the current object. It makes an error classification, which confuses the judgment ability of the supervision model and leads to the lower accuracy of the model. As shown in figure 1, traditional region detection methods are divided into three steps. 1) Region selection (Tang et al., 2000; Yin & Li, 2020). That is, it uses sliding window to cover a part of the image to be tested as a candidate region; 2) Feature extraction (Guan et al., 2021; Karim et al., 2020; Li et al., 2022). The features are related to the candidate region, such as HOG, SIFT; 3) Classification. It uses the classifier completed by training to classify classes, such as the commonly used Support Vector Machine (SVM) model, Adaboost, DPMC6, RF(Rodriguez-Galiano et al., 2012) (random forest).

However, these algorithms all need to manually obtain the relevant region feature information from the original image with many limitations (Daura-Oller et al., 2009; Yin & Li, 2021; Zhou et al., 2002):

1. Poor portability. For specific detection tasks, different methods need to be designed manually. For different regions or different forms of the same region, designers have high requirements for experience.
2. The classification of feature extraction and training is a common problem of traditional detection models. If the extraction of artificial features occurs omission phenomenon in the design process, the missing useful information will not be recovered from classification training, thus affecting the detection results.
3. The traditional methods mostly adopt sliding window to conduct traversal search, and divide the image into small blocks of various sizes, and then identify the image blocks. It retains the part with high probability, and merges or deletes the part with low probability. The complexity of this method is high. There are a large number of redundant small pieces, which seriously affect the operation speed, and it is difficult to achieve in reality. Therefore, since the rise of deep learning in the field of target detection in 2013, it has quickly replaced the status of traditional algorithm.

In the traditional object region detection, Fe et al. proposed the Deformable Part Model (DPM) in 2008. DPM carried out performance extension on the basis of HOG and SVM, making full use of the advantages of HOG and SVM and achieving important breakthroughs in image processing, face recognition and other tasks. However, the traditional object region detection algorithm has two main defects (Jiang et al., 2013):1) The region selection strategy based on sliding window has no pertinence with high time complexity and window redundancy. 2) features of manual design are not very robust to changes in diversity.

DPM model has high complexity and low object detection speed and accuracy. Although the rise of deep learning has improved the accuracy of object detection, the effect has been difficult to break through. In 2013, Grshick et al. proposed R-CNN (Girshick et al., 2014), whose mAP of VOC 2007 test set was increased to 48%. In 2014, the mAP was increased to 66% by modifying the network structure, and the mAP of ILSVRC 2013 test set was also increased to 31.4%. Since 2014, after Girshick et al. References (Jiang & Yin, 2021; Wang et al., 2017; Yin et al., 2021) put forward the R-CNN that has achieved breakthrough results in the field of target detection, SPP-Net (He et

Figure 1. Flowchart of traditional region localization

al., 2015), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2017), R-FCN (Dai et al., 2016), YOLO (Cai et al., 2020), SSD1 and other algorithms have appeared successively. Deep learning-based target detection algorithm has become one of the hotspots in machine learning field.

2. TRADITIONAL OBJECT REGION DETECTION METHODS

The traditional algorithm can be roughly divided into two categories: target instance detection and traditional target category detection:

1. Object region detection usually uses the template and stable feature points of image to obtain the corresponding relationship between the template and the object in the scene, finally detect the object region instance. Object region detection only focuses on the region itself. The rest of the objects in the image are irrelevant.
2. Traditional region classification uses AdaBoost algorithm framework, HOG feature and support vector machine, etc. and other methods to detect a limited number of categories according to selected features and classifiers.

2.1 SIFT-Based Methods

The SIFT algorithm proposed by Lowe. It is a widely used key point detection and description algorithm to find the characteristic points which are not easily affected by illumination, noise and affine transformation. In this algorithm, the scale space is realized by using Gaussian fuzzy, the extreme value is detected by the Difference of Gaussian function, and then the unstable points of edge response are screened through the determination of edge principal curvature (Piccinini et al., 2012; Wang et al., 2021). The key points with matching stability and strong anti-noise ability are obtained. Finally, the direction histogram is used to calculate the neighborhood gradient and direction of key points, and the descriptor is obtained.

Through a series of methods, the SIFT algorithm ensures that the extracted features are invariant in translation, scaling and rotation. It is also robust to light, noise and a small number of Angle changes, and has a good recognition rate for partial occlusion. However, SIFT algorithm has such problems as high complexity, slow detection speed, and difficulty in extracting effective feature points from fuzzy images and smooth edges.

2.2 SURF-Based Methods

SURF algorithm is an improved algorithm based on SIFT, and Hessian matrix is the core of the algorithm (Li & Zhang, 2013). The Gaussian filter is used to guarantee the scale independence, and the box filter is used to replace the Gaussian filter to simplify the calculation. By constructing Hessian matrix, the key point location is obtained. In addition, in the scale space, instead of building images of different scales with SIFT, SURF kept the size of images unchanged and only changed the size of the filter, reducing the amount of computation. In short, SURF algorithm uses approximate Hessian matrix to reduce the sampling process, and quickly constructs scale pyramid to realize target detection speed.

2.3 AdaBoost-Based Methods

AdaBoost is a machine learning algorithm based on Boosting. Initially, n samples in the training set are assumed to have the same weight. After each training, it adjusts the data weight in the training set, increasing the weight of the error samples, so that the next classifier can focus on training the error samples. After N training, N weak classifiers are integrated, and corresponding weights are allocated according to the performance of each classifier to form a classifier with high accuracy and low error (Jun-Feng & Luo, 2009).

2.4 Viola-Jones Methods

Viola-jones algorithm is the first algorithm that can process face detection in real time with better effect. The proposed algorithm marks the entry of face detection into the practical application stage. Viola-jones detection algorithm (VJ) uses Haar features to describe the window (Viola et al., 2005), reflects the change of light and dark in local areas, and adopts the idea of integral graph to solve the disadvantages of heavy calculation and repetition in Haar feature extraction. At the same time, the idea of cascade is introduced. As shown in Figure2, VJ arranges classifiers according to their complexity and computational cost. The higher the classification cost, the fewer images they need to classify, thus reducing the classification workload.

2.5 PCA-SIFT-Based Methods

For the existing problems of SIFT, Ke et al. (Ke & Sukthankar, 2004) proposed the PCA-SIFT algorithm. This algorithm improves the last step of SIFT. Principal component analysis (PCA) method was introduced to replace histogram with PCA to reduce dimensionality of description subvectors to improve matching efficiency. Compared with SIFT, PCA-SIFT has fewer dimensions, it is more flexible and variable, and its detection speed is about 3 times than SIFT. However, some information is lost due to dimensionality reduction, which leads to better effect only for representative images, but has some limitations.

2.6 Summary

This paper makes a comparative summary of traditional algorithms. In general, the purpose of these algorithms is to quickly calculate and predict features on the premise of ensuring the extraction of rich and accurate features. However, the features extracted by traditional algorithms are basically low-level and manually selected features, which are relatively more intuitive, easy to understand, and more targeted to specific objects. But it cannot well express a large number of multi-class targets.

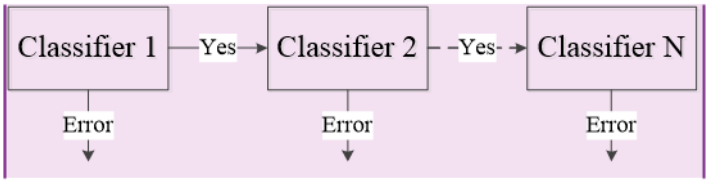
3. DEEP LEARNING METHODS FOR OBJECT REGION DETECTION

Since AlexNet has greatly improved the accuracy of image classification by using convolutional neural network in competitions, some scholars have attempted to apply deep learning to the detection of object region. Convolutional neural network can not only extract higher-level features with better expression ability, but also complete the extraction, selection and classification of features in the same model. In this aspect, there are two main algorithms: one is the classification-based RCNN series object region detection framework combining region proposal and CNN network. The other is single stage, which converts object region detection into regression problem (Shi et al., 2021).

3.1 RCNN Algorithm

Region proposal is based on Selective Search algorithm. According to the information of texture, edge and color in the image, less regions are detected while higher recall rate is guaranteed. Ross Girshick et al. proposed the RCNN model. RCNN uses Selective Search to obtain candidate regions

Figure 2. Flowchart of Cascade



(about 2000). Then the candidate region size is normalized and used as the standard input of CNN network. Reusing AlexNet to gain the features of the candidate area, the multiple SVM classification and the linear regression are adopted to fine-tune positioning frame (Bounding box).

RCNN significantly increased the detection rate from 24.3% of OverFeat to 31.4% (ILSVRC 2013 dataset), and achieved an accuracy rate of 58.5% on VOC2007 dataset. However, RCNN extracted features from nearly 2000 candidate regions respectively, and there were many repeated regions between the candidate regions, leading to a large number of repeated and slow operation. The average processing time of each image was 34s. At the same time, data storage space is lost much. In addition, the normalization of candidate regions will affect the final results.

3.2 SPP-Net Algorithm

In view of the shortcomings of RCNN in extracting features from all candidate regions, SPP-Net performs convolution operation on the whole picture to extract features at one time. As a result, feature extraction has changed from nearly 2000 times of RCNN to one time for the whole picture feature extraction, greatly reducing the workload.

In addition, SPP-Net added Space pyramid pooling layer (SPP layer) after the last convolutional layer and before the full connection layer to extract feature vectors of fixed size to avoid the complicated operation of normalization of candidate region size. The above two improvements make the detection speed of SPP-Net 38-102 times faster than that of RCNN, and solve the problem of candidate region normalization. SPP-Net has replaced the convolutional network, but its accuracy is almost the same. At the same time, SPP-Net still does not solve the problem of RCNN storage space consumption, and the steps of determining candidate regions, feature extraction, object classification and positioning correction are still separated.

4. GRADIENT PYRAMID MODEL

In order to adapt to the complex environment and data conditions that may exist in the target location, for example, in the visual tasks such as the small amount of data information and unmarked data. Based on the features of the convolutional network structure, it has an internal multi-scale pyramid shape and computes the feature hierarchy layer by layer. The method not only pays attention to the deep semantic information, but also gives consideration to the target's superficial texture and edge information, and enriches the feature space. We make full use of convolution hierarchy pyramid structure network characteristics, created in all scales with strong semantic features, the gradient in each hierarchy feature map. Through the path of the top-down by combining and transverse connection, gradient pyramid class mapping model was constructed, the model increased the intensity of sexual characteristics under different scales.

In this paper, our structure uses the fused gradient information to understand the characteristics of different dimensions. First, the output of each level after the current image feed-forward calculation is $\{C_1, C_2, \cdots, C_l\}$, where l corresponds to different convolution layers, and the output of each level is directly taken as the returned characteristic map $\{F_1, F_2, \cdots, F_l\}$. Because the first layer is too close to the input image and its network discrimination information is insufficient, the first layer is not used. Then, the network calculates the predicted output category c, and calculates the gradient score of each category c relative to all feature layers, that is, the partial derivative $\frac{\partial y^c}{\partial F_l^k}$ of output y^c to the feature map F_l^k of convolution layer l. The partial derivative information is processed by global average pooling operation to obtain $(\omega_{F_1^c}, \omega_{F_2^c}, \cdots, \omega_{F_l^c})$. Where, the pooling range corresponding to the sub-block k of each feature map is (i, j), so:

$$\omega_{F_{l\_i}^c} = \frac{1}{Z}\sum_m\sum_n \frac{\partial y^c}{\partial F_{i\_mn}^k} \tag{1}$$

For the different feature maps under each level, the corresponding length, width and channel $\{m,n,k\}$ of a single map. Through the activation function ReLU layer:

$$S_{F^l}^c = \mathrm{Re}\,LU(\sum_k \omega_{F_l^c} F_l^k) \tag{2}$$

We can get the feature score for each layer of the current gradient pyramid:

$$Score = \{S_{F^1}^c, S_{F^2}^c, \cdots, S_{F^l}^c\} \tag{3}$$

In each gradient feature map $S_{F^l}^c$, we perform a two-step operation. First, it is sampled twice to make it the same shape as the gradient map of the next layer. Then, it is crosswise connected with the gradient intensity map $S_{F^{l-1}}^c$ of the next layer to enhance the shallow feature intensity and the deep feature intensity for fusion. The operation between each layer is:

$$S_{F^l}^c = \varphi_{ups}(S_{F^{l-1}}^c) + S_{F^{l-1}}^c \tag{4}$$

For the gradient feature map at the bottom, we can get:

$$S_{F^l}^c = S_{F^1}^c + 2^1 S_{F^2}^c + \cdots + 2^{i-1} S_{F^i}^c \tag{5}$$

The high level feature map has more bigger weight than the low level feature map. Because the semantic information of high-level feature map is more concentrated, and more visual structures can be captured. The horizontal connection between layers makes the gradient stronger and stronger. It can be seen that the feature information based on gradient pyramid is more abundant, which provides more judgment basis for image processing tasks.

3.3 Fast RCNN Algorithm

Fast RCNN algorithm is based on the SPP-Net, the SPP is simplified to ROI Pooling layer, and the output of connect layers is performed with SVD decomposition to get two output vector: the classification score of the softmax and the window regression of Bounding box. This improvement combines the classification problem with the border regression problem using softmax to replace SVM. All the features are stored in the video memory, reducing the occupation of disk space. SVD decomposition can greatly accelerate the detection speed without affecting the accuracy.

Fast RCNN uses VGG16 to replace AlexNet, with an average accuracy of 70.0%. The training speed is 9 times higher than that of RCNN, and the detection speed reaches 0.3s per image (excluding the region proposal stage). Fast RCNN still uses the Selective Search method to select candidate regions, which involves a lot of calculation. When running on a CPU, it takes an average of 2s to get candidate regions for each image. It can be seen that the improvement of Selective Search is the key to improve the speed of Fast RCNN.

3.4 Faster RCNN Algorithm

Although SPP-Net and Fast RCNN can reduce the workload from the perspective of feature extraction, the slow speed problem of candidate regions selection by Selective Search is still not solved. For Faster RCNN, Region Proposal Networks are used to replace the Selective Search algorithm for the purpose of achieving true end-to-end computation for target recognition. The RPN network maps the pre-scale anchor frame to the original image by window drawing on the feature map to obtain candidate regions. RPN network inputs feature map and full connection layer feature map sharing the calculation. The use of RPN enables Faster RCNN to complete candidate region, feature extraction, classification, location correction and other operations within a network framework.

RPN makes Faster RCNN need only 10ms in the region proposal stage, the detection speed reaches to 5f/s(including all steps), and the detection accuracy is also improved to 73.2%. However, Faster RCNN still uses ROI pooling, resulting in the loss of translation invariance of subsequent network features and affecting the final positioning accuracy. After ROI pooling, each area passes through multiple full connection layers. There are many repeated calculations. Faster RCNN uses anchor point frame to correspond to the original map on the feature map. The anchor point frame, after multiple down-sampling operations, corresponds to a large area in the original map, resulting in a poor effect of Faster RCNN in measuring small objects.

3.5 R-FCN Algorithm

Target detection includes two problems: classification and location problem. The former has translation invariance and the latter has translation sensitivity. R-FCN uses full-convolution network ResNet27 instead of VGG to improve the effect of feature extraction and classification. To overcome the inadaptability of the full convolutional network to translation sensitivity, the algorithm adopts a specific convolutional layer to generate a Position Sensitive Score Map containing the target spatial Position information. The full connection layer will not be connected after the ROI Pooling layer to avoid double calculation.

The accuracy rate of R-FCN reached to 83.6%, and the average cost of testing each image was 170ms, which was 2.5 to 20 times faster than faster RCNN. However, R-FCN needs to generate a channel number that increases linearly with the number of categories before obtaining Score map. Although this process improves the target detection accuracy, it slows down the detection speed, making it difficult to meet the real-time requirements.

3.6 Mask RCNN Algorithm

Mask RCNN is an improved algorithm based on Faster RCNN, which increases the attention to instance segmentation. In addition to classification and location regression, the algorithm adds parallel branches about instance segmentation. The losses of the three are trained together.

The instance segmentation requires the accuracy of instance positioning to reach the pixel level. However, error is introduced into the constant scaling process of ROI Pooling layer for Faster RCNN, resulting in rough space quantization and inaccurate positioning. Mask RCNN proposed the bilinear difference value RoIAlign to obtain more accurate pixel information, which improved the accuracy of Mask by 10% to 50%. Mask RCNN also uses ResNeXt basic network, and the detection speed on COCO data set is 5 f/s, and the detection accuracy is improved from 19.7% of Fast r-cnn to 39.8%.

Mask RCNN achieves the highest level in detection accuracy and instance segmentation. Some of the subsequent algorithms have been improved in performance, but remain at the same level. However, the detection speed of this algorithm is still difficult to meet the real-time requirements, and instance segmentation is still facing the problem of too expensive labeling.

3.7 YOLO Algorithm

From RCNN to Faster RCNN, target detection always follows the idea of “region proposal and classification”. The two training models will inevitably lead to the increase of parameters and training

amount, which will affect the speed of training and detection. Therefore, YOLO proposed a "single-stage" idea. As shown in figure 8, YOLO segments the image into $S \times S$ grid (cell), the grid only is responsible for test center on the goal of the grid, each grid needs to predict two dimensions of the bounding box and category information. One-time forecasts all areas including the bounding box of of target, the target confidence as well as the category complete detection probability.

YOLO adopts cell-based multi-scale region to replace region proposal. It abandons some accuracy in exchange for a large increase in the detection speed. The detection speed can reach 45 f/s, which is enough to meet the real-time requirements.The detection of RCNN is 73.2%, with a large gap.

YOLO greatly improves the detection speed, but there also exists the following problems: (1) because each grid predicts two bounding box with same category, so for center at the same time, multiple objects in the environment have more missed detection; (2) since the determination of YOLO on the positioning box is a little rough, the accuracy of its object region localization is not as good as Fast-RCNN; (3) for objects with unconventional appearance, the detection effect is not good.

3.8 SSD

Faster-RCNN detection has a high detection accuracy with a slow speed, while YOLO detection has a low accuracy but a fast detection speed. SSD combines the advantages of both, and uses the idea of RPN for reference on the basis of YOLO to ensure a high precision detection while taking into account the detection speed. The feature map of specific layer only needs to train object detection of corresponding scale. Therefore, SSD combines the feature map of the high level and the low level, using multi-scale Regression of regional features.

The mAP of SSD can reach to 73.2%, which is basically the same as Faster RCNN (VGG16), and the detection speed reaches to 59 f/s, which is 6.6 times faster than Faster RCNN. However, SSD has the following problems:(1) the small target corresponds to a very small area in the feature map and cannot be fully trained, so the detection effect of SSD on small target is still not ideal; (2) when there is no candidate region, it is difficult to make regional regression, and it is difficult to converge. (3) the feature maps of different layers of SSD serve as independent input of the classification network, leading to simultaneous detection and repeated operation of the same object by different size boxes.

By adding batch normalization, multi-scale training and k-mean dimensional clustering after each convolutional layer, YOLOv2 further improves the detection speed and accuracy. The algorithm can achieve the detection speed of 67 f/s with 76.8% accuracy and 40 f/s with 78.6% accuracy. The performance of this algorithm basically represents the most advanced level in the industry. In the same paper, YOLO9000 was also proposed. The algorithm adopts wordTree hierarchical classification, mixed detection data and recognition data set, and simultaneously trains on the classification and detection data set to realize the detection of 9418 classes. Both YOLO series and SSD algorithms follow the RCNN series algorithm, which first conducts classification pre-training on large data sets and then uses fine-tune method on small data sets. However, the fine-tune pre-training model has the following problems:(1) the pre-training model is often unable to be transferred to specific data such as medical images; (2) the structure of the pre-training model is basically fixed and difficult to modify; (3) there is a difference between the pre-training sample and the final detection target, and the model obtained may not be the best model for the detection target.

4. METHODS FOR IMPROVING OBJECT REGION LOCALIZATION PERFORMANCE

RCNN-based and YOLO object detection framework are two basic frameworks for target detection. Based on these frameworks, the researchers propose a series of methods to improve the performance of target detection from other aspects:

1. Hard Negative Mining. RCNN used the idea of Hard sample Mining in training SVM classifier, but Fast RCNN and Faster RCNN did not use Hard sample Mining due to the end-to-end training strategy (only the proportion of positive and Negative samples was set and randomly selected). Experimental results show that OHEM (Online Hard Example Min ing) mechanism can improve the mAP of Fast RCNN algorithm on VOC2007 and VOC2012 by about 4%.
2. Multi-layer feature fusion. Both Fast RCNN and Faster RCNN make full use of the features of the final convolutional layer for target detection. However, as the features of the convolutional layer in the upper layer have a lot of details (pooling operation), the positioning is not very accurate.
3. When extracting the Region Proposal features with the context information, the Region Proposal context information (Yin et al., 2020) is combined to achieve better detection effect.

5. PROSPECT AND OUTLOOK

Deep learning based target detection, which has greatly improved the detection accuracy and speed compared with traditional methods, but still faces some problems.

For small data volume, the current framework may not get good results. At present, most of the algorithms use transfer learning, that is, existing big data sets are used for training, and the trained "semi-finished products" are used for fine-tune operation. If the target data is not in the data set such as ImageNet, the training effect depends on the correlation between the target and the big data set. Although DSOD algorithm designed a zero-based training network and achieved good results, its detection speed is still to be improved.

Deep learning has poor explanatory power, especially at a deeper level. In many cases, it can only rely on tests and experience to guess the reasons for its effectiveness or ineffectiveness. For the middle process, it looks like a black box.

High calculation strength. The use of GPU has improved the computing power of computers, but many operations are still too large. How to simplify and reuse calculation while ensuring accuracy as much as possible may be a point of innovation.

Insufficient use of original information in video, such as scene information and semantic information, results in loss of some effective information.

Whether it is RCNN series or SSD algorithm or not, it is always unable to achieve satisfactory results in the detection of small targets. As far as the current algorithm is concerned, in order to ensure the detection speed, the image of feature pyramid usually reduces the amount of calculation, but this will inevitably lead to insufficient training of small targets on the feature map. For example, R-SSD increases the number of feature maps and loses the detection speed. This problem has something in common with problem (3).

In the future, we will do the following works.

More and more comprehensive data sets. At present, there are two solutions: one is manual annotation. For small data volumes, the operation is simple and can ensure a high accuracy rate. But for large data volumes and accurate annotation of object segmentation, it fails. Another approach is to use parallel vision. it uses artificial scenes to simulate the actual scenes, design and evaluate the model through computational experiments, and perform parallel online optimization of the visual system. If parallel vision is realized, it will solve the problem of insufficient annotation data set and promote the development of target detection.

More computing sharing. Both RCNN series and regression-based detection algorithms are designed to allow the calculation between different ROIs.

How to use contextual information, scene information and semantic information will be an important research direction of target detection in the future. If the idea of parallel vision is feasible, the problems of difficult data set annotation and insufficient data volume will be better solved. In addition, how to better solve the problem of small data set detection which is not closely related to the training set is also a relatively important research direction.

ACKNOWLEDGMENT

The author would like to thank the two reviewers for their anonymous review.

REFERENCES

Bahrami, S., Dornaika, F., & Bosaghzadeh, A. (2021). Joint auto-weighted graph fusion and scalable semi-supervised learning. *Information Fusion*, *66*(1), 213–228. doi:10.1016/j.inffus.2020.09.007

Cai, Y., Li, H., & Yuan, G. (2020). *YOLObile: Real-Time Object Detection on Mobile Devices via Compression-Compilation Co-Design*. arXiv:2009.05697. 10.17760/D20398279

Dai, J., Li, Y., & He, K. (2016). *R-FCN: Object Detection via Region-based Fully Convolutional Networks*. Curran Associates Inc.

Daura-Oller, E., Cabre, M., Montero, M. A., Paternain, J. L., & Romeu, A. (2009). Specific gene hypomethylation and cancer: New insights into coding region feature trends. *Bioinformation*, *3*(8), 340–343. doi:10.6026/97320630003340 PMID:19707296

Elakkiya, R., Teja, K., & Deborah, L. J. (2021). Imaging based cervical cancer diagnostics using small object detection - generative adversarial networks. *Multimedia Tools and Applications*, 1–17.

Fan, S., Zhu, F., & Chen, S. (2021). FII-CenterNet: An Anchor-Free Detector With Foreground Attention for Traffic Object Detection. *IEEE Transactions on Vehicular Technology*.

Girshick, R. (2015). *Fast R-CNN*. arXiv e-prints, arXiv:1504.08083. 10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 580-587. doi:10.1109/CVPR.2014.81

Gong, H., Zheng, S., & Wu, Z. (2021). Spatial-Prior-Guided Attention for Small Object Detection in Overhead Catenary System. *Journal of Circuits, Systems, and Computers*.

Guan, S., Wang, X., Hua, L., & Li, L. (2021). Quantitative ultrasonic testing for near-surface defects of large ring forgings using feature extraction and GA-SVM. *Applied Acoustics*, *173*, 107714. doi:10.1016/j.apacoust.2020.107714

He, K., Zhang, X., Ren, S., & Sun, J. (2015, September 1). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916. doi:10.1109/TPAMI.2015.2389824 PMID:26353135

Jiang, , & Yin, . (2021). Facial expression recognition based on convolutional block attention module and multi-feature fusion. *International Journal of Computational Vision and Robotics*.

Jiang, P., Ling, H., Yu, J., & Peng, J. (2013). Salient Region Detection by UFO: Uniqueness, Focusness and Objectness. *2013 IEEE International Conference on Computer Vision*, 1976-1983. doi:10.1109/ICCV.2013.248

Jun-Feng, G. E., & Luo, Y. P. (2009). A Comprehensive Study for Asymmetric AdaBoost and Its Application in Object Detection. *Acta Automatica Sinica*, *35*(11), 1403–1409.

Karim, S., Zhang, Y., Yin, S., Bibi, I., & Brohi, A. A. (2020). A Brief Review and Challenges of Object Detection in Optical Remote Sensing Imagery. *Multiagent and Grid Systems.*, *16*(3), 227–243. doi:10.3233/MGS-200330

Ke, Y., & Sukthankar, R. (2004). PCA-SIFT: a more distinctive representation for local image descriptors. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2004.1315206

Li, J., & Zhang, Y. (2013). Learning SURF Cascade for Fast and Accurate Object Detection. In Computer Vision & Pattern Recognition. IEEE Computer Society.

Li, Y., Wang, S., Yang, Y., & Deng, Z. (2022). Multiscale symbolic fuzzy entropy: An entropy denoising method for weak feature extraction of rotating machinery. *Mechanical Systems and Signal Processing*, *162*(7), 108052. doi:10.1016/j.ymssp.2021.108052

Liu, Z., Li, X., & Luo, P. (2016). Semantic Image Segmentation via Deep Parsing Network. In *IEEE International Conference on Computer Vision*. IEEE.

Piccinini, P., Prati, A., & Cucchiara, R. (2012). Real-time object detection and localization with SIFT-based clustering. *Image and Vision Computing*, *30*(8), 573–587. doi:10.1016/j.imavis.2012.06.004

Ren, S., He, K., Girshick, R., & Sun, J. (2017, June 1). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. doi:10.1109/TPAMI.2016.2577031 PMID:27295650

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*(Jan), 93–104. doi:10.1016/j.isprsjprs.2011.11.002

Shi, Q., Yin, S., Wang, K., Teng, L., & Li, H. (2021). Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation. *Evolving Systems*. Advance online publication. doi:10.1007/s12530-021-09392-3

Tang, H., Wu, E., Ma, Q., Gallagher, D., Perera, G. M., & Zhuang, T. (2000). MRI brain image segmentation by multi-resolution edge detection and region selection. *Computerized Medical Imaging and Graphics*, *24*(6), 349–357. doi:10.1016/S0895-6111(00)00037-9 PMID:11008183

Teng, L., Li, H., & Yin, S. (2017). Modified pyramid dual tree direction filter-based image denoising via curvature scale and nonlocal mean multigrade remnant filter. *International Journal of Communication Systems*, (3), e3486.

Viola, P., Jones, M. J., & Snow, D. (2005). Detecting Pedestrians Using Patterns of Motion and Appearance. *International Journal of Computer Vision*, *63*(2), 153–161. doi:10.1007/s11263-005-6644-8

Wang, D., Wang, X., & Yin, S. (2021). A New Recursive Neural Network and Center Loss for Expression Recognition. *International Journal of Electronics and Information Engineering.*, *13*(3), 97–104.

Wang, X., Shrivastava, A., & Gupta, A. (2017). A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3039-3048. doi:10.1109/CVPR.2017.324

Xi, P., Guan, H., Shu, C., Borgeat, L., & Goubran, R. (2020). An integrated approach for medical abnormality detection using deep patch convolutional neural networks. *The Visual Computer*, *36*(9), 1869–1882. doi:10.1007/s00371-019-01775-7

Ye, T., Zhang, Z., & Zhang, X. (2020). Autonomous Railway Traffic Object Detection Using Feature-Enhanced Single-Shot Detector. *IEEE Access*.

Yin, Li, & Laghari. (2021). A Bagging Strategy-Based Kernel Extreme Learning Machine for Complex Network Intrusion Detection. *EAI Endorsed Transactions on Scalable Information Systems, 21*(33). 10.4108/eai.6-10-2021.171247

Yin, S., & Li, H. (2020). Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 5862–5871. doi:10.1109/JSTARS.2020.3025582

Yin, S., & Li, H. (2021). GSAPSO-MQC: Medical image encryption based on genetic simulated annealing particle swarm optimization and modified quantum chaos system. *Evolutionary Intelligence*, *14*(4), 1817–1829. doi:10.1007/s12065-020-00440-6

Yin, S., Li, H., & Teng, L. (2020). Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images. *Sensing and Imaging*, *21*(1), 49. Advance online publication. doi:10.1007/s11220-020-00314-2

Yin, S., Zhang, Y., & Karim, S. (2018). Shahid Karim. Large Scale Remote Sensing Image Segmentation Based on Fuzzy Region Competition and Gaussian Mixture Model. *IEEE Access: Practical Innovations, Open Solutions*, *6*, 26069–26080. doi:10.1109/ACCESS.2018.2834960

Zhou, J., Hinz, M., & Tönnies, K. D. (2002). *Focal Region-Guided Feature-Based Volume Rendering*. 1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT 2002), Padova, Italy.