Multi-source image target detection technology based on salient feature fusion

Yuyong Cui, Shengzhe Wang, Xinyi Gao*, Zhongjian Wu, and Yang Yong
Southwest Institute of Technical Physics, Chengdu, Sichuan, 610041, China
*Corresponding author’s e-mail: 1721367938@qq.com

Abstract. The ground target can be easily affected by the complexity of ground objects and photoelectric interference, so the accuracy of target detection in complex background is not always satisfying. To this end, joint target detection of Infrared/Visible image, IR/VI, is adopted, and the feature representation is achieved by applying the method of deep image feature extraction. According to the different imaging characteristics of the target, the feature fusion is carried out to form a unified target feature representation. As a result, target detection based on deep learning can be achieved by the model established by such fused feature. Experiments show that the proposed method can effectively improve the accuracy of target detection.

1. Introduction
As a basic problem of computer vision, object detection has achieved remarkable development in intelligent monitoring, driverless vehicles, behaviour recognition, and other fields[1]. Traditional detection methods, especially image-based methods, for example, feature-based methods including Histogram of Oriented Gradient, HOG, Scale Invariant Feature Transform, SIFT[2], and window sliding feature method, can detect and distinguish from pedestrian, vehicles, and other objects. Such artificial feature selection often requires experience, but due to the different types of targets, there are some differences in feature construction. With the development of deep convolution network technology, typical target representation and target details can be extracted from higher-level features to obtain more accurate target detection results. The construction of standard libraries such as IMGENet and COCO[3] has promoted the rapid development of deep learning, which has significant advantages in the field of target detection. On the one hand, its deep features can simulate the human understanding process and complete the deep abstraction of object features. On the other hand, since the richness of the sample library has a great impact on the results of deep network architecture training, the deep network has a strong learning ability due to its numerous parameters. As a result, such vast number of training data helps to activate the neurons of the deep network and enables such neurons to have a deep understanding of the target from the shape, environmental attributes, and colour of the object. Therefore, deep network has unique advantages in target detection, and with the enrichment of samples, it can obtain more detection effect than ordinary methods.

For IR/VI under the same scene, two-stage mode is firstly applied to complete the depth feature extraction of the same target under different imaging conditions and to harvest typical features. Secondly, the inputs of deep learning network are established by the feature fusion mode. Thirdly, the concept of YOLO v3 is employed to meet the requirement of being applied to airborne platforms by a lightweight design of network to raise the efficiency of training and detecting. Finally, through the typical experimental verification, the productivity of the depth detection network based on multi-source image feature level fusion is improved.
2. Related Work
It is of great significance to use multi-source image features to complete typical target detection, especially for different types of target detection, and now target detection based on deep learning has become a research hotspot. Relying on the advantages of deep convolutional neural network feature description, it can effectively complete the detection of single objective or multiple targets. It is quite worth mentioning that such neural network has certain superiority when the target is disturbed[3]. OverFeat[4], a two-stage algorithm, uses Convolutional Neural Network, CNN, to complete the target analysis of typical video sequences and to finish the target training to form the separation of background and target. Reference[5] proposes a deep self-coding network and designs a loss function to effectively improve the background extraction effect. Jifeng with his team[6] introduces a video contrast mode into the high-level features of background to realize fast moving object detection at block level. In reference[7], a multi-layer self-organizing mapping network is constructed to complete the background model training, which effectively improves the effect of intelligent target detection. Based on R-CNN, Fast R-CNN carries out multiple feature iterations and adds convolution operation for many times to solve the problem of the overlapping of candidate regions, which results to the waste of resources. By using the idea of extracting features once and normalizing the expression, the detection efficiency has been greatly improved. Residual network strengthens the importance of network depth to network, but the deep network needs to avoid gradient explosion, which can be solved by constant connection[8,9]

3. Method
The way to reach our goal is to apply an intelligent target detection technology based on multi-source information fusion.

3.1. Target deep feature extraction
The deep learning model is used to construct the target feature network and a five-layer deep feature is constructed to meet the requirements, according to the experience of target detection in image. With the support of large-scale training level, the features from the first five-layer of AlexNet are used to extract the depth features of typical images. The initial parameters of the model are obtained by training ImageNet dataset, which is a kind of traditional datasets. The deep feature extraction of aim model can be achieved by training such dataset, the feature extraction mode of typical target can be improved by adopting representative pooling and convolution operations, and typical feature can be extracted by cancelling the fully-connected layer and the final convolutional layer of the typical network.

Through the feature extraction, visible and infrared images are transformed into a pyramid image which has three channels and seven layers. The output image as shown in Figure 1 is constructed according to the feature training of such typical image. Since the traditional network dimension is 224*224 but the dimension of the image used in this paper is 512*512, the network parameters of the first layer are extracted directly to obtain the accurate data.

The analysis results of typical feature are deduced by the mode of IR/VI fusion, which is detailed introduced in Chapter 3.2. Thanks to such fusion on the feature level, the typical single source image
extraction mode is formed based on the original image. Using the proposed extraction mode, the typical aircraft features are extracted.

According to the deep features of different image fusion, the typical target feature expression is produced. Since the essence of deep learning is to detect through the mixed components of the target, the typical target feature can be obtained by the component with the highest score, which is captured through training. As a result, such features can be equipped with the ability of distinguishing from target and background efficiently by representing features from different data source for various targets.

3.2. Deep fusion mode for features of IR/VI

For the requirement of accurate target attack and faithful information feature, fusion mode should be analysed by different target features, IR/VI fusion mode could be established, and finally, the feature-level fusion mode can be formed.

According to the advantages of infrared and photoelectric images in feature level detection, a typical feature detection fusion mode is adopted to form a multi-level feature detection deep fusion mode. In addition, through the salient region feature detection and fusion of the two kinds of image objects, the region of interest can be obtained effectively, the amount of data can be reduced, and the computational efficiency can be improved. However, due to the low signal-to-noise ratio of the infrared image in the fusion image and the obvious edge diffusion effect, the traditional segmentation algorithm based on edge extraction has some limitations in the accurate location and the actual size of the target. Deep features are used to extract typical targets and form salient target features. This process can not only accurately extract the feature information of the target, but effectively process the image with multiple significant targets, improve the processing efficiency, and further achieve the aim of removing the camouflage as well.

Salient target extraction uses different features to accomplish the registration before fusion, completes the fusion of typical targets, forms the pattern of typical target features, improves the target detail features, and finally helps to raise the target detection effect after fusion. Specific heterogeneous image fusion mode sets typical images and according to the different feature details of the same level, multiple feature association modes are used to evaluate, and the decision criteria is as shown in equation (1).

\[ f = \max(\alpha * F(a,b)) \]  
\[ \text{In which, } \alpha \text{ is the applied correlation parameter vector, which means the correlation degree of the two images is taken into consideration as well, and } F(a, b) \text{ stands for the fusion mode. In general, for different types of feature fusion, it is necessary to sort according to their features. Therefore, when the features of infrared and optoelectronic images are complementary, the feature vector can be considered sparse. To construct a typical feature fusion model, } F(a, b) \text{ can be expressed as equation (2).} \]

\[ F(a, b) = \sum_{i,j} p(a_{i,j}, b_{i,j}) \log \frac{p(a_{i,j}, b_{i,j})}{p(a)p(b)} \]  
\[ \text{In which, } p(a_{i,j}, b_{i,j}) \text{ can be regarded as the feature correlation metric to measure the fusion among such two images. Additionally, } p(a) \text{ is the feature detection accuracy of one input image and } p(b) \text{ represents the feature detection accuracy of another input image.} \]

3.3. Intelligent target detection based on fusion images

To achieve the intelligent target detection of typical fusion image, the mode based on multi-source information fusion is adopted, and the typical target detection is realized, which is based on the depth feature fusion of infrared and visible image. The overview of the whole technology is illustrated in Figure 2.
According to the different size and resolution of infrared and visible images, which are represented by $I_{op}$ and $I_{IR}$, regularization should be applied to the images, as shown in equation (3) and (4).

$$I'_{op} = f(I_{op})$$  \hspace{1cm} (3)

$$I'_{IR} = f(I_{IR})$$  \hspace{1cm} (4)

In which,

$$f(x) = e^{-\frac{x-x_{mean}}{\delta}}$$  \hspace{1cm} (5)

In image processing, the mean value, $x_{mean}$ is calculated by the mode of infrared and optoelectronic image disposing and $\delta$ is the error. After regularization, the size of different images is uniformly transformed to 512*512.

In the journey of feature fusion, the method given in chapter 3.2 is used to complete the fusion of typical features at different levels to form the feature representation mode of typical targets. The objective of candidate area location prediction is to point out which areas should be used as the centre to generate candidate areas, which is a binary classification problem. By feature fusion images, typical input images for YOLO v3 can be formed, which is the newest algorithms in the series of YOLO. Based on the original algorithm, YOLO v3 improves the prediction accuracy on the premise of maintaining the speed advantage, so as to strengthen the recognition ability of small objects. However, due to the direct use of the original framework for training, the effect is not ideal, the main reasons are as follows.

- In the same monitoring scene, there is a large scale difference between the target objects, which leads to the detection effect is not satisfying. The goal of this project is to identify different personnel targets and dangerous goods held by a person, such as a knife which is partially covered, so the size difference between the two is obvious. Therefore, the network must be improved so that objects of different sizes can be detected in one network at the same time.

- Due to the large scale difference of the target object in the application scene, the object positioning accuracy will also be reduced. This is because it is easy to produce polarization when region proposal is generated. On the one hand, for large-scale objects, the model tends to produce large-scale region proposal. On the other hand, small-scale object training samples will cause the model to be interested in small-scale region proposal, which makes the model vibrate and not easy to converge. The feature fusion used before the input image improves the accuracy of the features extracted at all levels, especially the fine features extracted at the low level and the abstract features extracted at the high level with semantic information. However, the features extracted in the low-level network are more abundant, so it is more suitable for extracting details, which has better effect on small objects, and more accurate positioning. Therefore, to extract small-scale objects, output should come from such network at a lower layer. Such an improvement not only preserves more detailed information of small-scale objects but obtains
more accurate positioning as well. At the same time, the early output of small-scale sample class makes the training converge more quickly, the feature extraction process of fusion image is added before the detector of the network, the multi-feature fusion input is used to extract the features of the same object in different scales and make good learning, and finally, the effectively of target detection is improved.

The mode of optimal confidence is applied to the output. Let \( I_oU \) to be the real box area and \( 0 \) to be the non-target area.

\[
y_b^* = \begin{cases} 
I_oU(\hat{b}, b^*) & \text{if } b \in B(M) \\
0 & \text{else}
\end{cases} \tag{6}
\]

In which, \( \hat{b} \) belongs to \( \hat{G} \) and \( b^* \) satisfies \((\hat{b}, b^*)\in M\), which means that the target candidate area is the same as the prediction area.

\[
\hat{B}(M) = \{\hat{b}|(\hat{b}, b^*)\in M\} \tag{7}
\]

\[
B^*(M) = \{b^*|{(\hat{b}, b^*)}\in M\} \tag{8}
\]

\( \hat{b} \) and \( b^* \) obeys equation (7) and (8), which stand for the registered detection set and the real label set.

4. Experiments and analysis

4.1. Environment

In this paper, we design the network model structure for the embedded application platform. However, due to the limited computing and storage capacity of the embedded hardware platform, it is difficult to complete the training and optimization task of the model effectively. To solve this problem, the model combining server and embedded system is applied to train and test such network. The above network model structure is firstly trained on the high-performance server platform, the actual performance test of the model is completed on hardware platform, and finally the model is imported into the embedded hardware platform. The details of the experimental environment are shown in Table 1.

| Hardware environment | Software environment |
|----------------------|----------------------|
| CPU                  | Memory               |
| Intel i5-7400        | 32G                  |
| System               | Software             |
| Windows 7 64bits     | Matlab 7.0           |

4.2. Results and analysis

Because it is difficult to obtain the IR/VI of the same scene under different conditions and backgrounds, and with the aim of testing the effectiveness of the algorithm, TNO Image fusion dataset is used fusion detection to improve the detection ability of the system in the case of interference and to raise the accuracy rate.

4.2.1. Comparison of model performance. The detection results of single-source image and fused image are calculated respectively, and the typical fusion mode is used to carry out multiple experiments based on complex backgrounds. For TNO video data, discrete time sampling is used to form a standard database, and through the mixed use of image rotation and the actual data collected by our team, the number of images reaches 20000, including 10000 infrared images and 10000 visible images. Then, this improved data set is used for simulation verification in typical scenarios. It is worth mentioning that there is less vehicle occlusion in TNO, so our team make a complementation to this condition. The
models are used to train the above two kinds of input data and to complete the performance test. Table 2 and 3 show the results.

Table 2. Comparison of model accuracy (%).

| Input data      | Pedestrian | Vehicle |
|-----------------|------------|---------|
| Infrared image  | 87.8       | 90.3    |
| Visible image   | 90.1       | 93.5    |
| Fusion feature  | 91.2       | 94.8    |

Table 3. Algorithm accuracy (%).

| Algorithm                  | Pedestrian | Vehicle |
|----------------------------|------------|---------|
| YOLO v3 (infrared image)  | 90.4       | 92.1    |
| YOLO v3 (visible image)   | 91.3       | 92.6    |
| Our method                | 92.6       | 96.5    |

Considering the application background of this algorithm and because of partial original data, the data are synthesized under the condition of being partially covered. At the same time, 8000 images are used for training and 2000 images are used for testing. It can be conducted that the detection accuracy is much higher in the feature fusion target detection model, because in the specific detection, although lots of key points are hidden after the pooling layer, the feature fusion detection mode does enhance the characterization ability of the pooling feature and does improve the performance of detection.

4.2.2. Comparison of model complexity. For real use, the evaluation of the model needs not only to be judged from the perspective of accuracy, but to take the correlation between resources and such accuracy into consideration as well. Therefore, the complexity, which is determined by the parameter calculating performance of the proposed models, is analyzed, and the specific results are shown in Table 4.

Table 4. Comparison of model complexity.

| Input data | Calculation time (ms) | Maximum memory (M) | Accuracy (%) |
|------------|-----------------------|--------------------|--------------|
| Infrared image | 12.5               | 1.5                | 91.2         |
| Visible image   | 10.6               | 1.6                | 90.1         |
| IR/VI         | 15.5               | 1.9                | 93.5         |

The calculation amount of our method is a little more than that of the single sensor, which is mainly due to the feature fusion method before detection, although some of the fusion features do not participate in the calculation. Therefore, the feature pruning method can be applied for further improvement although the existing time consuming is still controllable. Meanwhile, since the features need to be stored, the occupation of memory increases as well.

Compared with the single sensor, the accuracy of the model is increased by 3.3% compared with the infrared feature, and 2.42% compared with the visible image. The specific reason is that when the target is occluded and the feature is blurred, the characteristics of the visible image are more obvious in the daytime. In addition, due to the thermal radiation of the selected target, the infrared image is easy to be confused with the background when detected separately, resulting in low accuracy. Therefore, target detection by feature fusion image can effectively meet the requirements of intelligent transportation and typical military applications.
5. Conclusion

Aiming at the requirement of target detection in multi-source fusion conditions like typical military applications, intelligent transportation, and other common scenarios, IR/VI fusion algorithm is introduced to establish the image deep feature fusion mode. Through the deep feature fusion of infrared image and visible image, the integrity of target deep feature is improved, the effectiveness of feature detection at different levels is enhanced, and the purpose of target detection is achieved. In a word, the proposed IR/VI fusion model can effectively improve the detection probability of targets in battlefield.

For the aim of realizing the feature fusion in different layers and achieving the purpose of target detection, this paper adopts the mode of target deep feature fusion detection. However, it is often impossible to determine the fusion level through this mode, the fusion effect obtained at different levels may not be the best, and there might be some feature repetition among fusion process and the detection progress of YOLO v3. In the future, feature fusion in different levels and cooperation of different algorithms will be applied to reduce the reuse of features and to improve the accuracy of target detection.

References

[1] Wang, D., Bai, C., Wu, K. (2021) Survey of video object detection based on deep learning. J. Frontiers of Computer Sci. and Tech., 1-16.

[2] Dalal, N., Triggs, B. (2005) Histograms of Oriented Gradients for Human Detection. In: IEEE Conference on Computer Vision and Pattern Recognition. San Diego. pp. 886-893.

[3] Lowe, G.D. (2004) Distinctive Image Features from Scale-Invariant Keypoints. J. International Journal of Computer Vision. 60:91-110

[4] Ren, S., He, K., Girshick, R., Sun, J. (2015) Towards Real-Time Object Detection with Region Proposal Networks. In: 29th Conference and Workshop on Neural Information Processing Systems. Montreal. pp. 91-99.

[5] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. Columbus. pp. 580-587.

[6] Dai, J., Li, Y., He, K., Sun, J. (2016) R-FCN: Object Detection via Region-based Fully Convolutional Networks. In: Conference and Workshop on Neural Information Processing Systems. Barcelona. pp. 379-387.

[7] Russakovsky, O., et, al. (2015) ImageNet Large Scale Visual Recognition Challenge. J. International journal of computer vision, 115: 211-252.

[8] Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K. (2017) Aggregated Residual Transformations for Deep Neural Networks. In: IEEE Conference on Computer Vision and Pattern Recognition. Washington. pp. 5987-5995.

[9] Kong, T., Yao, A., Chen, Y., Sun, F. (2016) HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. In: IEEE Conference on Computer Vision and Pattern Recognition. Washington. pp. 845-853.