Hindawi
Journal of Sensors
Volume 2022, Article ID 1732532, 13 pages
https://doi.org/10.1155/2022/1732532

Research Article

Research on Multitarget Recognition and Detection Based on Computer Vision

XiaoYing Zhang and Yiran Chen

College of Big Data and Artificial Intelligence, Chongqing Institute of Engineering, Chongqing, 400056, China

Correspondence should be addressed to XiaoYing Zhang; 00250@cqie.edu.cn

Received 13 February 2022; Revised 10 March 2022; Accepted 15 March 2022; Published 8 April 2022

Academic Editor: Yuan Li

Information is obtained from human eyes for thinking divergence, and further associated with computer equipment, so human beings endow computers with the ability of "vision" to convey and feel information. This field has developed for many years, and many aspects can be in line with other research directions, such as artificial intelligence, which has become popular in recent years, and pattern recognition, which has been applied a lot. In order to sort out the structure and content of multitarget recognition smoothly, this paper starts from the perspective of shallow vision, uses theory and practical experiments, and chooses the core technology with the largest weight from massive computer technologies, so that the recognition algorithm can compare with the recognition algorithm. The research shows that (1) CNN shows its unique feature ability and incomparable detection accuracy from many models, and the error rate can be reduced from 28.07% to 18.40%. (2) The method of candidate region is complex, and the larger the region, the more difficult it is to calculate. The method based on regression is far beyond it in both precision and speed and is more suitable for the research of this subject. (3) When the mAP increases, the speed is forced to slow down. If the image resolution is high with the same model, the mAP will be high (SSD and YOLO models are often used). Experiments show that the recognition effect is obvious. At the end of the article, the advantages and disadvantages of this study are summarized. In the field of computer vision, people need to do more in-depth research. Follow-up can optimize multitarget recognition and detection and strive to improve the accuracy.

1. Introduction

When it comes to computer vision, we have to mention the background of deep learning. If we want new development, we need to study some old classic problems repeatedly (such as the target detection problem in this paper) and find blind spots that others cannot find from the details, resulting in qualitative changes, and "blind spots" become "breakthrough points." When you think of the word "vision," you can think of the meaning that it makes the machine detached and get the perception ability. However, this goal is difficult to achieve. Researchers have wasted a lot of effort and made a lot of useless work, and it was not until the introduction of deep learning that they got a positive meaning that cannot be omitted. Target recognition is the core, most valuable, and most basic task step. If this part of work is not done well, all subsequent steps will be greatly affected. In this paper, a variety of computer technologies are used, such as deep learning and image processing, combing and summarizing the relevant information of the better target recognition algorithm based on computer vision so far and selecting the multitarget visual recognition method (target detection algorithm based on convolution neural network) which meets the actual needs and is most suitable for this study. According to the latest cutting-edge science, we have inquired a lot of information and literature related to computer vision and cited them reasonably. The latest progress of computer vision algorithm and the improvement of computer performance in reference [1] have new functions. Reference [2] designs real-time computer vision and machine learning systems for modeling and recognizing human behavior in visual surveillance. Literature [3] investigates head posture estimation in computer vision. Reference [4] analyzes and calculates the development of horizontal

setting method and fast marching method of interface motion in computer vision. Reference [5] provides 19 enlightening basic achievements in computer vision research. In literature [6], computer vision introduces flexible decision forest models to deal with huge and diverse image and video analysis tasks. Literature [7] extracts moving targets from complex background scenes and designs the background code manual model of target motion detection algorithm. Reference [8] uses wavelet transform to detect human moving objects based on computer vision. Literature [9] investigates the ability of six different algorithms to track subpixel targets in moving background and noise. Literature [10] discusses the key algorithms of video target detection and recognition in intelligent transportation system based on computer vision technology. In reference [11], an improved algorithm combining background subtraction and mixed frame difference method is introduced on the basis of moving target detection algorithm. Literature [12] solves the problem of detecting moving objects by using tree search algorithm for object detection in image sequences. Literature [13] introduces the method of moving multitarget detection and tracking in stable scene. Literature [14] realizes target tracking and real-time obstacle detection of autonomous vehicles based on computer vision. Literature [15] develops a new m-sequence target and circular correlation processing technology based on computer vision for real-time displacement measurement.

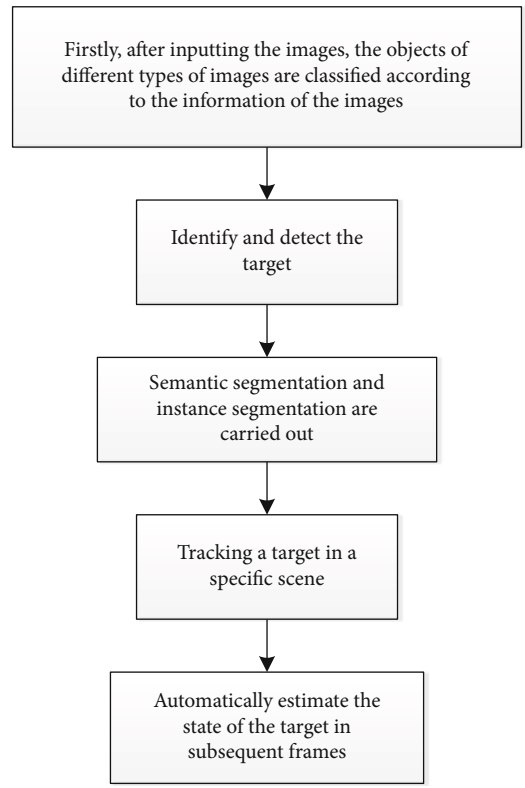

Figure 1: Typical processing flow of computer vision.

2. Theoretical Basis

2.1. Computer Vision. Computer Vision [16] is a science that studies how to make machines "see". At the beginning of last century, the function that computer vision can realize is only to analyze and recognize two-dimensional digital images. With the continuous expansion of this field, people are no longer easily satisfied with processing simple images, and people prefer that computers can recognize and even understand what they see through vision. Machine analysis extracts image features and then analyzes them. Image understanding [17] is an extension and expansion of computer vision research. The process flow is shown in Figure 1.

Main research directions of computer vision is shown in Table 1.

2.2. Deep Learning. Turing test [19], a method to test whether a machine has human intelligence, was put forward by mathematician Turing in 1950s, and it can well judge whether a machine has the same perception ability as human beings. The core idea of the Turing test is to ask computers to disguise themselves as human beings as much as possible when they take questions from people. This coincides with the desire of people in our target detection algorithm that computers disguised as human "eyes" can obtain, process, and convey information. Whether it is Turing test, deep learning, or artificial intelligence, the contents of their research topics are similar and integrated, and a lot of knowledge is interlinked. Many algorithms cannot achieve this goal, and the research and success of deep learning algorithm bring hope. Deep learning is a big science, which aroused the research wave of deep learning as early as the 1940s, but it did not achieve great success until 2012. Deep learning [20] is to learn the internal law and representation level of sample data. The feature extraction method of directly calculating the original data replaces the old method. The feature expression of unsupervised learning can be divided into three types according to different components. As shown in Table 2, there are various methods of feature expression.

2.2.1. Limit Boltzmann. DBN [21] consists of "Constrained Boltzmann Machines," as shown in Figure 2.

Let n be the visible layer node, m be the hidden layer node, v be the visible unit, and h be the hidden unit. Energy of the system:

$$E(v, h \mid \theta) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{n}\sum_{j=1}^{m} v_i W_{ij} h_j. \quad (1)$$

When the model state is constant, the joint probability distribution is

$$P(v, h \mid \theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)}, Z(\theta) = \sum_{v,h} e^{-E(v,h|\theta)}. \quad (2)$$

Table 1: Main research directions of computer vision.

Image classification	Challenges such as viewpoint change, scale change, intraclass change, image deformation, image occlusion, lighting conditions, and background clutter; nowadays, the popular image classification architecture is convolution neural network.
Object recognition and detection [18]	Subdivision detection algorithms such as face detection, vehicle detection, and character recognition are derived. Commonly used models are R-CNN and fast R-CNN.
Semantic segmentation	Every pixel of the input image is classified, and its inner meaning can be clearly described with a picture. Commonly used models are full convolution network (FCN), SegNet, and so on.
Motion and tracking	Generally speaking, large-scale convolution neural networks can be trained as classifiers and trackers. The representative tracking algorithms are full convolution network tracker (FCNT) and multidomain convolution neural network (MD net).
Visual question and answer	The purpose of this study is that users ask questions according to the input images, and the algorithm automatically answers questions according to the content of questions.
Motion recognition	In practical applications, accurate motion recognition is helpful for public opinion monitoring, advertising, and many other tasks related to video understanding.
Three-dimensional reconstruction	In the field of 3D vision, geometry-based methods are still the main methods, such as 3D reconstruction and visual SLAM.

Table 2: Feature expression method based on learning.

Category	Method name
Feature expression based on deep learning	CDBN, SBM, DeCAF, R-CNN, fast R-CNN, NIN, SPPNet, segDeepM, MatchNet, OverFe-at, SuperCNN……

The activation probability of hidden unit is

$$p\left(h_j = 1,|v,\theta\right) = \sigma\left(b_j + \sum_j v_i W_{ij}\right). \tag{3}$$

Visible cell activation probability is

$$p(v_i = 1,|h,\theta) = \sigma\left(a_i + \sum_i h_j W_{ij}\right). \tag{4}$$

The sigmoid activation function is

$$\sigma(x) = \frac{1}{1 + \exp\ (-x)}. \tag{5}$$

The output of the activation function hidden layer node is

$$v^{k+1} = \omega^k h^k + b^k, h^k = \sigma\left(v^k\right). \tag{6}$$

Softmax function:

$$p_s = \frac{\exp\ \left(v_s{}^{N+1}\right)}{\sum_j \exp\left(v_s{}^{N+1}\right)}. \tag{7}$$

When $d_s = 0$, the cross-entropy function:

$$L = -\sum_s d_s \log p_s. \tag{8}$$

2.2.2. Self-Coding Machine. As shown in Figure 3, the main structure of the self-encoding machine is as follows.

Input an N-dimensional signal x, through the input layer to the middle layer, the signal changes y, the principle of self-coding machine.

$$y = s(Wx + b). \tag{9}$$

The signal y is decoded by the decoding layer to the output layer of n neurons, and the signal becomes z.

$$z = s\left(W'y + b'\right). \tag{10}$$

Matrix transposition:

$$W' = W^T. \tag{11}$$

Typical square error:

$$L(xz) = \|x - z\|^2. \tag{12}$$

Cross entropy method:

$$L_H(x,z) = -\sum_{k=1}^{n}[x_k \log z_k + (1 - x_k) \log\ (1 - z_k)]. \tag{13}$$

2.2.3. Convolution Neural Network. Convolutional neural networks (CNN) [22] s a kind of feedforward neural networks with depth structure including convolution calculation. It is a very important deep learning algorithm. We give an example of a single-layer convolution neural network, which includes two processes: convolution and subsampling. Introducing

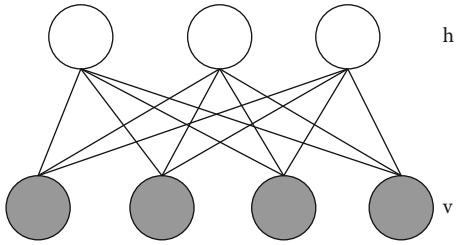

Figure 2: Constrained Boltzmann machine.

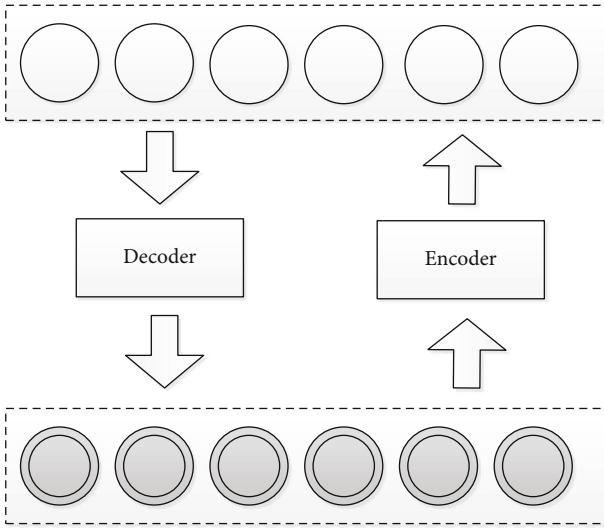

Figure 3: Self-coding machine.

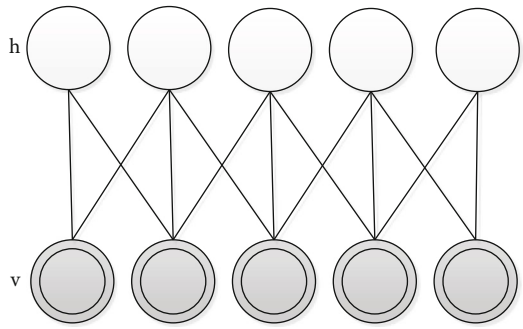

Figure 4: Single-layer convolution neural network.

different convolution kernels to extract different features and observe specific patterns of input signals; subsampling can reduce the dimension of feature map (using average pooling or maximum pooling operation), which reduces the resolution of feature map, but can keep the feature description. In the graph, every two nodes have various connections, which represent the process of convolution from the input node and then subsampling into the output node, as shown in Figure 4.

$$Z^{l+1}(i,j) = \left[Z^l \otimes \omega^{l+1}\right](i,j) + b, \tag{14}$$

$$Z^{l+1}(i,j) = \sum_{k=1}^{K_l}\sum_{x=1}^{f}\sum_{y=1}^{f}\left[Z_k^l(s_0 i + x, s_0 j + y)\omega_k^{l+1}(x,y)\right] + b, \tag{15}$$

$$L_{l+1} = \frac{L_l + 2p - f}{s_0} + 1\ (i,j) \in \{0, 1, \cdots L_{l+1}\}, \tag{16}$$

$$L_{l+1} = \frac{L_l + 2p - f}{s_0} + 1(i,j) \in \{0, 1, \cdots L_{l+1}\}, \tag{17}$$

$$L^{l+1} = L. \tag{18}$$

2.3. Target Detection Algorithm. Application fields of target detection and tracking are shown in Table 3.

If we want to detect the target, we need to use the corresponding algorithm. Because the traditional algorithm has great limitations, it must be based on image processing and hand-designed feature extraction, and it is greatly influenced by external factors such as noise and resolution, so its performance is not enough to meet people's needs. CNN's accuracy is nearly 30% higher than before [23]. The flow of traditional target detection algorithm is shown in Figure 5.

2.4. Image Processing. As shown in Figure 6, it is a flowchart about image processing.

(1) Pixel units are converted to centimeters:

$$L_{(\text{Actual})} = L_{(\text{Pixel})} * \left(\frac{100}{L_{(\text{Ruler})}}\right) cm. \tag{19}$$

(2) Image grayscale depends on RGB color scale operation [24].

Floating point algorithm:

$$\text{Gray} = R * 0.3 + G * 0.59 + B * 0.11. \tag{20}$$

Integer method:

$$\text{Gray} = \frac{(R * 30 + G * 0.59 + B * 11)}{100}. \tag{21}$$

Shift algorithm:

$$\text{Gray} = (R * 76 + G * 151 + B * 28) \gg 8. \tag{22}$$

Average method:

$$\text{Gray} = \frac{(R + G + B)}{3}. \tag{23}$$

Table 3: Application fields of target detection and tracking.

Application field	Specific application
Virtual reality	Interactive virtual world, game control, virtual studio, character animation, teleconferencing, and so on
Autonomous navigation	Vehicle navigation, robot navigation, space probe navigation, etc.
Robot vision	Industrial robots, home service robots, restaurant service robots, space probes, etc.
Advanced human-computer interaction	Sign language translation, gesture-based control, information transmission in noisy environment, etc.
Intelligent monitoring	Public safety monitoring, parking lots, supermarkets, department stores, vending machines, ATMs, access control of outsiders, traffic scenes, old and young care, etc.
Motion analysis	Content-based sports video retrieval, personalized training of golf and tennis, clinical research of patients, etc.

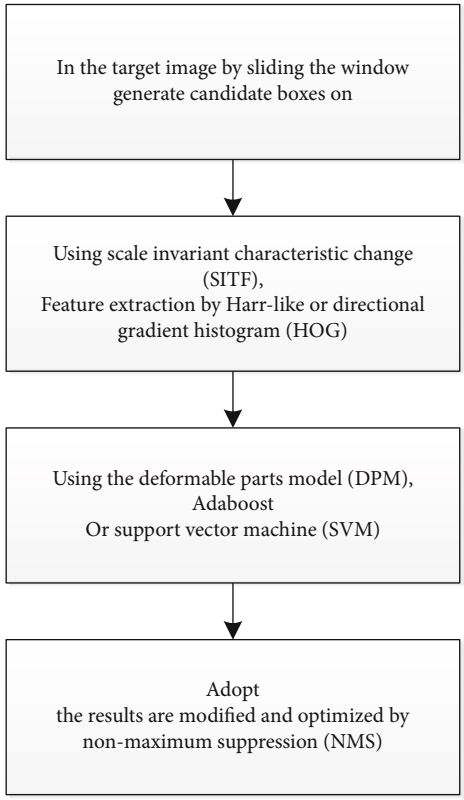

Figure 5: Traditional target detection algorithm.

Green method only:

$$\text{Gray} = G. \tag{24}$$

(3) The binarization of the image makes the image either black or white [25]. It can reduce the difficulty of subsequent image processing and further reduce the occupation of storage space

$$g(x,y) = \begin{cases} 0(\text{Gray value is less than threshold}(T)), \\ 255(\text{Gray value is greater than threshold}(T)). \end{cases} \tag{25}$$

(4) Enhancement and sharpening

$$d(x,y) = f(x,y) - g(x,y). \tag{26}$$

(5) Edge detection

$$\nabla^2 f(x,y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2}, \tag{27}$$

$$\nabla^2 f(x,y) = f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) - 4f(x+1,y). \tag{28}$$

3. Target Detection Algorithm Based on Convolution Neural Network

3.1. Overview of Convolution Neural Networks. All the research in this paper is mainly based on the field of computer vision, in which feature extraction and classification are indispensable parts. In the traditional image processing, the feature extraction method is usually designed by hand, which is modified on the image by human beings according to relevant theory and experience knowledge, far from being "intelligent;" moreover, this method cannot completely extract the target information, and its limitations are too great. In convolution neural network, convolution can represent feature extractor, and neural network is classifier, which meets the requirements of target recognition algorithm. Moreover, convolution neural network adopts average pooling or maximum pooling operation, which can imitate the visual and perceptual mechanism structure of living bodies and carry out supervised learning and unsupervised learning. Add three concepts to convolution neural network to reduce limitations, as shown in Table 4.

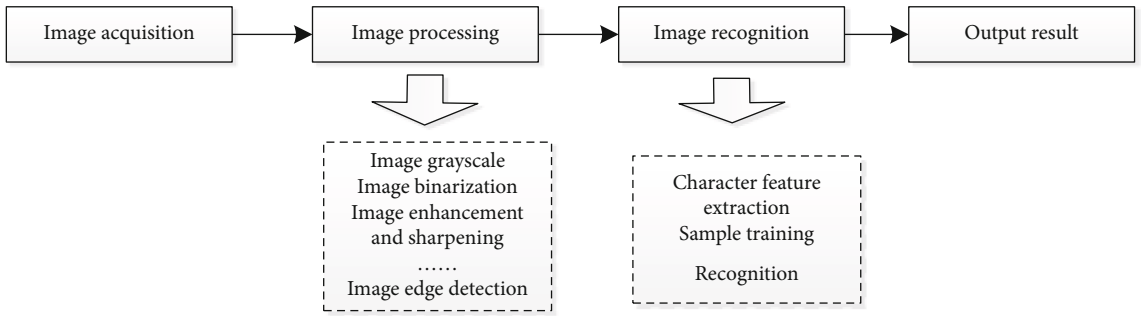

Figure 6: Image processing flow.

Table 4: Supplement of convolution neural network.

Purpose	New concept
It makes the convolution neural network more suitable for special processing of image data, and greatly reduces the limitations of traditional methods.	Local sensory angle Sparse weight Parameter sharing

Table 5: Flow comparison of two algorithms.

Object detection algorithm based on regression (one-stage algorithm)	Target detection algorithm for candidate areas (two-stage algorithm)
This kind of algorithm omits the candidate region generation stage	Select candidate regions on the input image
Feature extraction, target classification, and target regression are implemented directly in the same convolution neural network, which truly realizes end-to-end	Feature extraction and classification of candidate regions by convolution neural network

3.2. Selection of Target Detection Algorithm. As shown in Table 5 is the flow comparison of the two algorithms.

Then, as shown in Table 6, several main models under the two-stage algorithm are briefly summarized.

The one-stage algorithm can omit the candidate region generation stage. Therefore, in this study, according to the limitations of various literatures and experimental conditions, we decided to mainly select SSD model for multitarget recognition and detection based on computer vision; at the same time, we will also select YOLO model similar to SSD model to compare their performance and specific recognition work, which can better explain the situation and function of multi-target recognition and detection.

3.3. SSD Target Detection Algorithm

(1) SSD300 network architecture

SSD inherits the idea of regression from YOLO and has the running speed of YOLO and the detection accuracy of faster RCNN. SSD can complete target location and classification at one time. Among them, the shallow feature map plays a great role in target positioning, which contains position information. The deep feature map is of great significance to the classification of image objects, and it contains a lot of semantic information.

The SSD target detection algorithm adopts the feature pyramid structure detection way and utilizes each kind of feature function diagram of different size, such as Conv1, Conv2, Conv3, Conv4, Conv5, and Conv6. It can predict targets on characteristic maps of different receptive fields.

The reason why SSD model is chosen for target detection is mainly because SSD is improved on the basis of YOLO, which has more details optimized by YOLO, and SSD is widely used in many fields, so there are many data for reference, as shown in Table 7.

Feature pyramid prediction and SSD network architecture are shown in Figures 7 and 8:

Shallow feature maps are used to predict small-sized targets, while deep feature maps are used to predict large-sized targets, which improves the detection effect of small-sized targets. However, due to the lack of semantic information in the shallow feature map, SSD is still very poor in small target detection. At the same time, due to the nonuniformity of positive and negative samples, it is difficult to train.

(2) SSD training strategy

Training SSD is based on the fact that the output of ground truth information depends on the output of a fixed set of specific detectors, which can determine the end-to-end loss function and back propagation. Mapping

Table 6: Some main models of the two algorithms.

One-stage algorithm	Performance summary	Two-stage algorithm	Performance summary
YOLO series	YOLOv1 is very fast and can be monitored in real-time. The recognition effect of small targets is not good, and pictures with fixed size.	R-CNN	Ross Girshick proposed in 2014. Selective search algorithm is used instead of sliding window, which solves the problem of window redundancy and reduces the time complexity of the algorithm. Convolution neural network replaces the traditional hand-made feature extraction part, which can extract the image features more effectively and improve the external anti-interference ability.
SSD series	YOLOv2 solves the problem of difficult convergence and uses high-resolution pictures to fine-tune the network; anchor frame and convolution for prediction.	SPPNet	In 2015, Kaiming He and others proposed. The feature map is obtained by running convolution layer only once from the whole image, which greatly reduces the time consumed by feature extraction. Reduce the loss of image information and avoid repeated calculation of convolution features. The lifting speed is about 24 times to 64 times.
M2Det	YOLOv3 uses Darknet-53 as the network backbone and adopts FPN architecture.	Mask, R-CNN	In 2017, He et al. proposed Mask R-CNN, which combines faster R-CNN and FCN. The multiscale feature extraction ability of the model is strengthened, and the recognition of small target objects is more accurate. The detection speed is about 5 pieces per second.
CentripetalNet	YOLOv4 uses CSPDarknet 53 and many pervasive algorithms to achieve the best experimental results.	D2Det	Cao et al. proposed in 2020. At the same time, it solves the problems of accurate positioning and accurate classification. Dense local regression and DRP are introduced to extract accurate target feature regions from the first stage and the second stage, respectively, thus improving performance.

Table 7: SSD network classification.

Classification	Action
Standard network	This part of the network can classify images and remove the layers related to classification.
Pyramid network	This part of the network can be used to detect multi-scale feature mapping layer, so as to complete the detection of different sizes of targets.

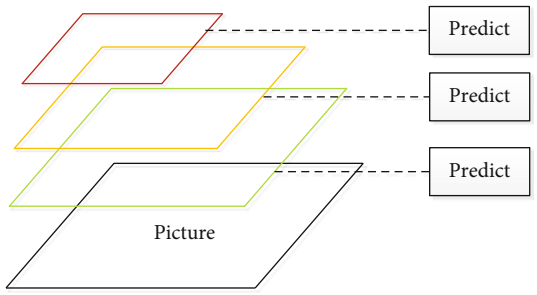

Figure 7: Feature pyramid prediction.

annotation information to default boxes is critical for some network models with default boxes. The total loss function includes position loss (loc) and confidence loss (conf).

$$L(x, c, l, g) = \frac{1}{N}\left(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)\right). \tag{29}$$

Losses include classification and regression. The cross-entropy loss function of classified loss is shown in formula (30). SmoothL1 of regression loss is shown in formula (31).

$$L_{conf}(x, c) = -\sum_{i \in Pos}^{N} x_{ij}^{p} \log\left(\widehat{c}_i^{p}\right) - \sum_{i \in Neg} \log\left(\widehat{c}_i^{0}\right) \quad \text{where} \quad \widehat{c}_i^{p} = \frac{\exp\left(c_i^{p}\right)}{\sum_{p} \exp\left(c_i^{p}\right)}, \tag{30}$$

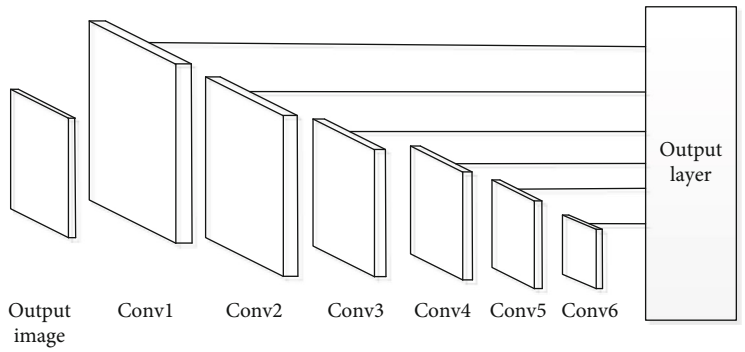

Figure 8: SSD network architecture.

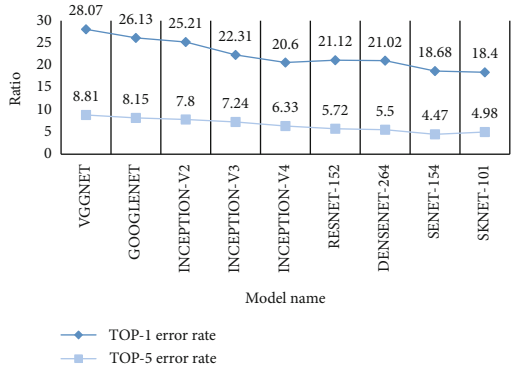

Figure 9: Performance comparison of some "CNN" models.

Table 8: Main evaluation indexes of COCO data set.

Indicators	Meaning
AP	The value of AP when IOU = 0.5: 0.05: 0.95
AP^{50}	The value of AP when IOU = 0.5
AP^{75}	The value of AP when IOU = 0.75
AP_S	AP value of small objects
AP_M	AP value of medium object
AP_L	AP value of large objects

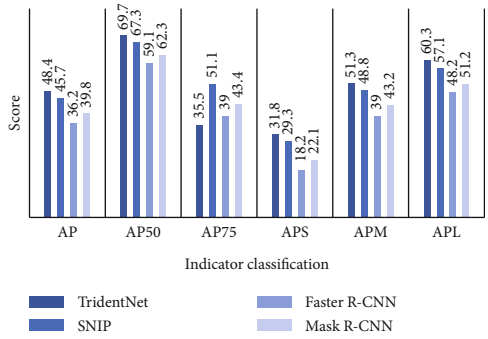

Figure 10: Target detection algorithm based on candidate regions.

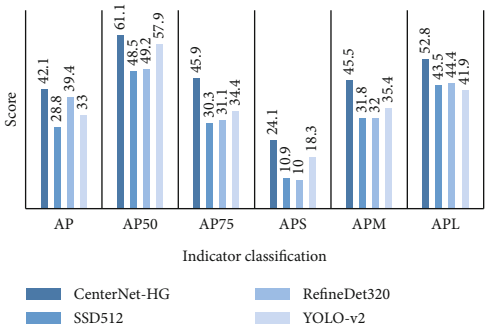

Figure 11: Object detection algorithm based on regression.

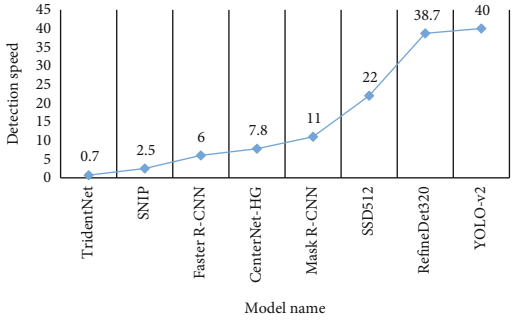

Figure 12: Comparison of detection speed of models.

$$L_{loc}(x,l,g) = \sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}} x_{ij}^{k}\text{smooth}_{L1}\left(l_i^m - \widehat{g}_j^m\right). \tag{31}$$

The receptive field in the shallow layer of SSD is smaller; for the deep layer, its receptive field is larger. The output of SSD is actually a convolution of multiple single outputs. Its training can be equivalent to splitting a complete image, and then dividing the image into countless small subimages to make a single output classification and positioning. SSD predicts the object category and position in the window; predict the background if there are no objects.

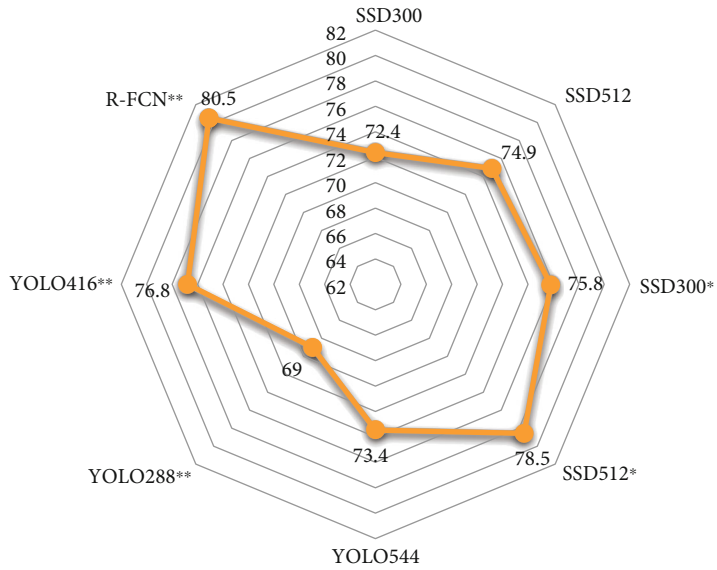

Figure 13: mAP performance metrics comparison.

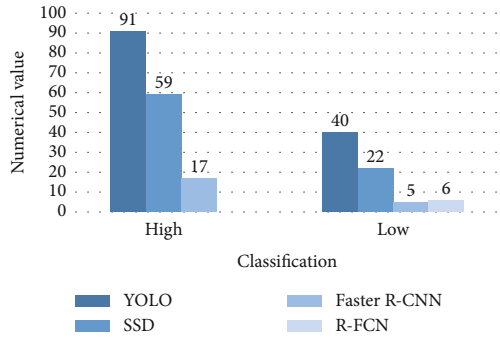

Figure 14: Maximum and minimum FPS under the same mAP.

4. Experimental Analysis

4.1. "CNN" Model Performance Comparison. We can use partial convolution neural network model to make a performance comparison, so as to highlight the strength of convolution neural network and the superiority of convolution method.

From the data shown in Figure 9, it is obvious that the error rate based on TOP-1 has dropped sharply from the original 28.07 to 18.40%; the error rate based on TOP-5 also decreased from 8.81% to 4.98%. Therefore, whether from the target demand or technical superiority, choosing a well-developed "CNN" model can help reduce the pressure and burden of target detection.

4.2. Performance Comparison on COCO Datasets. Evaluate the performance of the two algorithms, as shown in Table 8.

The details are shown in Figures 10 and 11. We can find from Figure 10 that if the algorithm chooses the candidate region method, with the refinement and improvement of the model, the accuracy of the detected results is obviously higher and higher, but the real-time situation cannot be detected all the time, which leads to the great drawback that cannot be overcome by this method. The model for the selection of the region is too complex, which virtually adds a lot of burden to the algorithm. The larger the model, the deeper the computation of the algorithm is. We can find from Figure 11 that the accuracy of regression algorithm is superior.

In order to evaluate the superiority of speed, we select some models of the two algorithms (TridentNet, SNIP, Faster R-CNN, CenterNet-HG, Mask R-CNN, SSD512, RefineDet320, and YOLO-v2) to compare the detection speed. As shown in Figure 12, the YOLO-v2 model has the fastest detection speed, and the TridentNet model has the slowest detection speed. Comparing the data of the two graphs, we can find that the speed of most regression algorithms is faster than that of candidate regions.

4.3. Comparison of SSD and YOLO. SSD and YOLO series models are the "darlings" in regression algorithms. In order to study the performance of models in multitarget recognition and detection, we choose a series of models of these two methods to compare and explain several performances.

(1) First, we select PASCAL VOC 2007 and PASCAL VOC 2012 data, and then use them to train SSD model and YOLO model. "∗" means that the model applies small target data enhancement; "∗∗" indicates that the model data results are tested by the VOC 2007 test set. One problem that needs attention is that because the test results of VOC 2007 are generally better than those of 2012, we specially selected the results of R-FCN in VOC 2007 in order to add cross-reference content. Finally, we can see from the diagram shown in Figure 13 that high-resolution images of the same model will have better mAP, but its processing speed will become slower

Figure 15: Model processing flow.

Table 9: Results analysis.

Results	Analysis
The training results based on VOC07 + 12plus (240000) are more accurate	This is because the training amount of 240,000 can make SSD have higher detection accuracy
The training results based on VOC07 + 12 (120000) have some error recognition	The training amount of 120000 is insufficient, so SSD generally uses multiple scale predictions of multiple scale feature maps and aspect ratio in images to determine the range of recognition targets, so as to separate and predict them.

Figure 16: Training result 1.

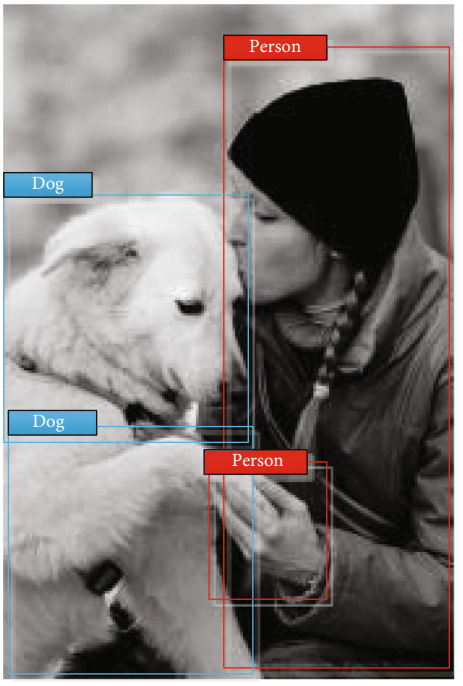

Figure 17: Training result 2.

(2) As shown in Figure 14, the highest FPS and the lowest FPS of YOLO, SSD, faster R-CNN, and R-FCN were compared under the same mAP. The input image resolution and feature extractor will affect the speed. In the figure, whether it is the highest FPS or the lowest FPS, YOLO's FPS is the highest, followed by SSD; however, the FPS data of the other two models, faster R-CNN and R-FCN, are very similar. By comparison, their FPS is very low, far less than the data of YOLO and SSD

4.4. Multitarget Recognition and Detection Based on SSD. In this section, choose clone-recursive directly on github, configure it according to the instructions and run it; GitHub provides a VOC-based model. Therefore, after careful consideration, we decided to use SSD300 model which has been trained on VOC 2007 and realized the detection of several pictures after simple prediction of trial operation results.

(3) As shown in Figure 15

The analysis of processing results is shown in Table 9: The processed picture is shown in Figures 16 and 17.

(4) We choose the model based on VOC07+12plus (240000). After the picture is processed, the situation of multi-target recognition and detection is shown in Figures 18–21

According to the test results, when there are multiple targets in an image, we can easily find that some targets in

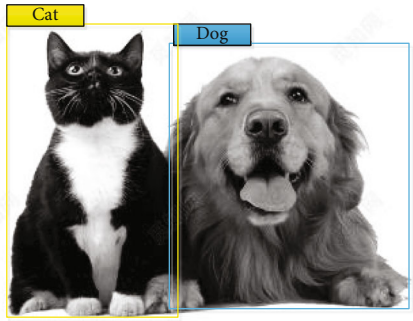

Figure 18: Test 1.

Figure 19: Test 2.

Figure 20: Test 3.

the image cannot be detected because of various problems: it may be a clear problem, and the edge of the target is blurred and difficult to distinguish. It may be that the target is too small to detect its existence. It is difficult to capture effective features due to incomplete recognition due to occlusion by other targets. It is difficult to fully capture features during movement, and there are omissions in real-time tracking… These problems are waiting for us to actively solve.

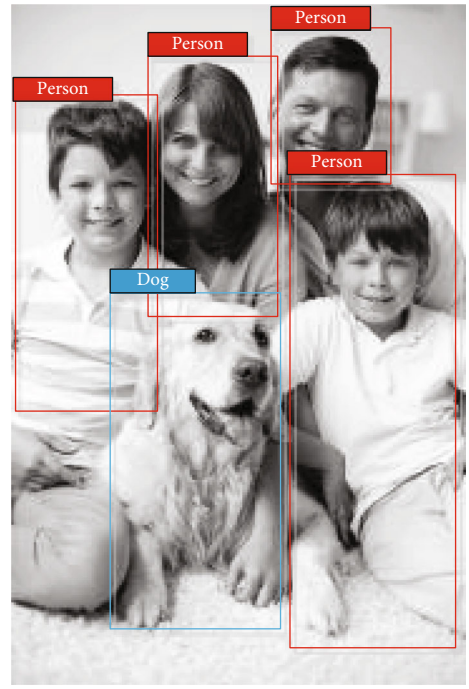

Figure 21: Test 4.

It is difficult to detect small objects: the main reason is that the resolution of the image is very low, the feature expression ability is weak, and it is difficult to extract. The detection target is too small, resulting in low IOU, which makes the target have no corresponding anchor frame. In the training data set, large and medium-sized targets are more inclined to be detected, and small objects are difficult to be included.

The occlusion problem leads to the incomplete expression of features, which increases the difficulty of classification and location.

When detecting densely distributed targets, the position of prediction box of similar targets close to each other is difficult to determine. After NMS processing, it is easy to be regarded as a single object.

5. Conclusion

This paper combs the effects of several typical target detection algorithms, gives their performance comparison on number sets, and highlights the superiority of target detection algorithms based on convolution neural network. The research results of this paper show that

(1) CNN shows its unique feature ability and incomparable detection accuracy from many models, and the error rate can be reduced from 28.07% to 18.40%

(2) The selection of candidate region method in target algorithm is complex and difficult to control, and the larger the region, the more difficult it is to

calculate. The method based on regression is far beyond it in both precision and speed and is more suitable for the research of this subject

(3) When the mAP increases, the speed is forced to slow down. If the image resolution is high with the same model, the mAP will be high (SSD and YOLO models are often used)

(4) In image training, using VOC07 + 12plus model, the VOC detection with 240,000 training amount is most in line with the regulations, and the final effect is good

The effect of target detection is good, but based on the technology of computer vision, the existing target detection algorithm has more research space, which can be optimized and improved to improve the detection efficiency and accuracy of the algorithm, so that people can choose the most suitable algorithm when studying experiments. At present, target detection is mainly aimed at the recognition and detection in specific scenes. In complex natural life scenes, the recognition and detection of multiple targets are easily interfered by various self or environmental factors, such as small targets, blocked by other objects, too dense objects, which make it difficult to find targets, and difficult to identify and track them in real-time in moving scenes… These are a series of problems that researchers are waiting to solve.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declared that they have no conflicts of interest regarding this work.

Acknowledgments

This research was supported by the Scientific and Technological Research Program of Chongqing Municipal Education Commission (Grant nos. KJZD-K202001901 and KJZD-K201901902), Chongqing Municipal Natural Science Foundation (Grant no. cstc2020jcyj-msxmX0666), and Chongqing Municipal Education Commission Humanities and Social Sciences Research Project (Grant no. 20SKGH277).

References

[1] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*, O'Reilly Media, Inc, 2008.

[2] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, 2000.

[3] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: a survey," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[4] A. M. Andrew, "Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science," *Kybernetes*, vol. 29, no. 2, pp. 239–248, 2000.

[5] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face recognition by humans: nineteen results all computer vision researchers should know about," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948–1962, 2006.

[6] A. Criminisi and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis*, Springer, London, 2013.

[7] K. Chen, X. Han, and T. Huang, "Target detection algorithm based on the movement of codebook model," *Science*, vol. 5, no. 2, 2012.

[8] F. Li, Fang Shuai, and X. Xinhe, "Human motion target detection based on computer vision," *Journal of Ordnance Engineering*, vol. 26, no. 6, pp. 766–770, 2005.

[9] D. Casasent, B. Kumar, and Y. L. Lin, "Subpixel target detection and tracking," *Journal of the American Veterinary Medical Association*, vol. 245, no. 9, pp. 992–992, 1987.

[10] Q. Pan and H. Zhang, "Key algorithms of video target detection and recognition in intelligent transportation systems," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 34, 2020.

[11] C. F. Zhang, Z. Y. Fang, and H. Q. Qu, "Information technology in an improved moving target detection algorithm," *Advanced Materials Research*, vol. 886, pp. 560–563, 2014.

[12] S. D. Blostein and T. S. Huang, "A tree search algorithm for target detection in image sequences," in *Proceedings CVPR'88: The Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 690–695, Ann Arbor, MI, USA, 1988.

[13] G. Tao, Z. Yao, W. Ping, C. Wang, and J. Yang, "Automatic stable scene based moving multi-target detection and tracking," *Journal of Computers*, vol. 6, no. 12, pp. 2647–2655, 2011.

[14] R. Fang and C. Cai, "Computer vision based obstacle detection and target tracking for autonomous vehicles," *MATEC Web of Conferences*, vol. 336, article 07004, 2021.

[15] Y. D. Hu, Q. Xia, and R. Hou, "Computer vision-based displacement measurement with m-sequence target," *Smart Structures and Systems*, vol. 27, no. 3, pp. 537–546, 2021.

[16] C. Szegedy, W. Liu, Y. Q. Jia et al., "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, 2015.

[17] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: an overview," *Computer Science Review*, vol. 11-12, pp. 31–66, 2014.

[18] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: an experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1442–1468, 2014.

[19] W. L. Ouyang, P. Luo, X. Y. Zeng et al., "DeepID-Net: multi-stage and deformable deep convolutional neural networks for object detection," https://arxiv.org/abs/1409.3505.

[20] D. Maturana and S. Scherer, "VoxNet: a 3D convolutional neural network for real-time object recognition," in *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).*, pp. 922–928, Hamburg, Germany, 2015.

[21] S. F. He, R. W. H. Lau, W. X. Liu, Z. Huang, and Q. X. Yang, "SuperCNN: a superpixelwise convolutional neural network for salient object detection," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 330–344, 2015.

[22] P. Dhankhar and N. Sahu, "A review and research of edge detection techniques for image segmentation," *International Journal of Computer Science and Mobile Computing*, vol. 2, no. 7, pp. 86–92, 2013.

[23] T. H. Rassem and B. E. Khoo, "Object class recognition using combination of color SIFT descriptors," in *Proceedings of the 2011 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 290–295, Penang, Malaysia, 2011.

[24] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1030–1037, San Francisco, CA, USA, 2010.

[25] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.