Research and Application on Smart Courtyard Monitoring System

Zhen-hua LI*, He-he LI, Bo-fang LIU

School of Control Science and Engineering, Shandong University, Jinan 250100, China
zhenhua.li@sdu.edu.cn, hehe.li@mail.sdu.edu.cn, bofang.liu@mail.sdu.edu.cn

Abstract: Residential safety has always been a basic need for people’s lives. In the vast suburbs and villages of China, quantities of courtyard houses are distributed, and the level of security monitoring is related to the people’s security and happiness. This study takes the characteristics of the human body with a large aspect ratio into account. By improving the structure of the Faster-RCNN network, a method for enhancing human body features is designed to enhance the network's ability to recognize the human body, and covers various human postures such as standing and lying down, and eventually improves the accuracy of the original human body recognition by 7%. In addition, this study designs and deploys a complete set of real-time monitoring system, which continuously saves images when detecting human intrusion, and sends alarm messages to the user's mobile phone WeChat in real time, which meets the real-time and adds convenience. The actual deployment shows that the system can find the target within 120 milliseconds when the intruder enters the sight and send out alarm messages to meet the real-time design requirements.

1. Introduction
As a basic life guarantee, residential safety has always been a key consideration for residents. There are quantities of residential areas with courtyards in the suburbs of cities and large rural areas in China. Due to the high price of monitoring software and equipment and early people's disregard for housing monitoring, the number of monitoring facilities has grown slowly. Lowering the price of monitoring software while improving the identification efficiency and ensuring a certain degree of accuracy will effectively increase the penetration rate of community monitoring facilities.

Traditional monitoring mostly uses the method of sensor and camera fusion. Lin[1] turns camera to the direction of abnormal point and collects images according to the abnormal pressure value of the pressure sensor. Chen Xi [2] used the abnormal temperature values obtained by infrared images to switch the camera in real time. Although traditional surveillance method can play a certain role, it does not use the large amount of information in the video itself.

As a core component of a new generation of intelligent monitoring systems, target detection has always been a hot topic of research. Since the deep learning [3,4] algorithm shined in the 2012 ImageNet competition, with the subsequent improvement of hardware computing power represented by GPU (Graphics Processing Unit), target detection has turned to a method based on statistical learning [5,6]. As a pioneering work in the application of deep learning in the field of target detection, Girshick [7] et al. proposed R-CNN (Regions with CNN features), using the Selective Search algorithm to extract a large number of candidate regions, and then classify each region. Later, Girshick [8] and others proposed a pyramid pooling layer with only one layer, using a pooling layer called ROI Pooling to divide candidate regions of different scales into M×N small blocks, and use max pooling in each small block.
Later, there were improved one-step methods based on the above two-step methods, such as the YOLO series [10-13] and SSD [14,15] algorithms, which also performed well in target detection.

2. Improved Faster-RCNN method
Ren [9] et al. proposed Faster-RCNN, adding a Region Proposal Network (RPN). Among them, the area generation network is divided into two forward paths, and the first path is classified using Softmax method. The second path is used to calculate the regression offset of the upper border of the anchor point. After that, another layer was added to integrate the above classification results and offset, and then combined with image information to complete the target positioning.

2.1 Regional Proposal Network
The most important thing in RPN is the setting of anchor points, as shown in Figure 1.

![Figure 1. Original anchor shape.](image)

The so-called anchor points are 9 rectangles divided into 3 shapes, and the aspect ratios are 1:2, 1:1, and 2:1. When the image sample is input, the image will be resized to a fixed size, and the largest side length in the anchor point is the side length of the image. In this way, the 9 anchor points basically cover the various sizes and shapes in the image. Next, traverse the feature map obtained through the convolutional layer, calculate the frame of the anchor point for each point, and then correct the frame position through frame regression.

2.2 Improve the Faster-RCNN network structure
Since the original Faster-RNN was designed to cover the shape features of each object to the maximum, it did not specifically optimize the slender shape features of the human body.

This research expands the anchor points in the original region generation network to 0.2, 0.2, 0.4, 0.4, 1, 2, 2.5, 4, 4. The size of anchor expands to 4, 6, 8, 10, 12, 16, 20, 24, 32, 36. The number of detections for areas is increased with higher and lower aspect ratios, so that the frame of the anchor point is more suitable for the relatively slender shape of the human body. As shown in Figure 2, the anchor points of the same size have been deformed to facilitate analysis.

![Figure 2. Improved anchor point shape](image)
2.3 Experimental process design

![Experimental flowchart]

As shown in Figure 3, this experiment is divided into two parallel timelines: one is the deep convolutional network part, and the other is the alarm software design part.

2.4 Training

Based on the open source deep learning framework Pytorch [16], the improved Faster-RCNN network structure is used as a model, the momentum is set to 0.9, the weight attenuation coefficient is 0.0005, the initial learning rate is 0.001, and the learning rate adopts a multi-level decay strategy. The Gamma coefficient is 5.

2.4.1 Pre-trained convolutional neural network

First select the pre-trained ResNet101 [17] convolutional neural network to extract feature maps. Resnet uses a residual network structure. The residual architecture in the network can integrate the output of the previous module with the output of a certain hidden layer, thereby improving the ability of the deeper network to suppress gradient degradation.

2.4.2 Training set

Since deep convolutional neural networks need to learn features of human limbs from a large amount of data, it is difficult to obtain sufficient sample features if the number of samples is too small.

The training set of this study uses the Pascal VOC2007 data set, which has a total of 9,963 images divided into 20 categories, covering humans, animals, vehicles, and indoor furniture. Among them, humans have 8566 target bodies, distributed on 4087 pictures, and have a variety of postures, including walking, running, jumping, cycling and other sports forms, as well as standing, lying on the side, full body, portrait and other postures and parts. Clothing styles vary greatly, covering different genders, different races, large scale changes and varying degrees of occlusion. The data set is divided into two parts: training set and test set.

2.5 Deployment

2.5.1 Hardware and system

The CPU version is Intel i5 9300H, the memory is 8G, the GPU version is Nvidia GeForce GTX1660Ti, and the camera model is HP Wide Vision HD. The operating system is the Linux series Ubuntu 18.04.

2.5.2 Monitoring system design

The system is implemented using Pycharm programming, written in python language, and the acquisition and storage of video content is completed using OpenCV [18]. The WeChat port adopts Deepin-wechat and runs based on Wine, which can not only meet the requirements of real-time transmission of alarm information, but also provide a user experience that facilitates receiving alarm information.
Figure 4. Interactive interface of courtyard monitoring system

As shown in Figure 4, the system user interface is designed using QT. The interface is divided into three parts, namely the title area, the display area and the button area. Among them, there are 3 buttons in the keypad, namely "Start", "View Screenshot" and "Exit". At the beginning of use, click the "Start" button to start the program, then the target detection subsystem will be called, and a real-time video monitoring window will pop up to display real-time monitoring images.

Figure 5. Software design flow chart of real-time monitoring subsystem

The flowchart of the subsystem software design is shown in Figure 5. The real-time monitoring subsystem adopts multi-threaded programming to complete. One thread contains three steps: completing Faster-RCNN network parameter loading, using OpenCV to obtain video pictures, and detecting whether there is human intrusion. Another thread is used to send alarm information to the user's mobile phone WeChat.

3. Test Results and Discussions
The test results show that by improving the Fater-RCNN network structure, the original network’s recognition rate of the human body in the Pascal-VOC data set has been increased by 7% to 78%. Improve the detection accuracy for the human body.
As shown in Figure 6, pictures a1-a6 are the top view angles of the scene (a), which are the detection results of the back, front, and side of the person without objects. pictures b1-b2, c1-c2, d1-d2, e1 -e2 are the recognition results in different scenarios.

As shown in Figure 7, once the system detects an intrusion, it will call the WeChat client to send an alert to the designated user, meeting the user's requirements for the real-time and convenience of receiving alarm information.

In actual use, within 120 milliseconds of the appearance of the human target, the system will automatically detect it and send an alarm message to the user. Every 0.5 seconds, the system will record the mark map of the intruder and save it in a specific folder. Click the "View Screenshot" button to view the captured image of the intruder.
4. Conclusions
This paper designs a real-time monitoring and alarm system based on the improved Faster-RCNN network. Redesign the anchor points of the regional proposal network, and enhance the characteristics of the human body, which increases the recognition accuracy by 7% compared with the previous one, meeting the accuracy requirements. A complete set of courtyard monitoring automatic alarm system contains user interaction interface. After detecting the intruder, the image of the person is saved and the alarm information is sent to the user's WeChat client to meet the real-time alarm requirements and convenience user experience of the monitoring system.

References
[1] Lin, Y., Wang, H., Tong, X. (2015) Research and implementation of courtyard intelligent monitoring system. J. Electronic Testing., 11: 9-12.
[2] Chen, X., (2020) The application of fire warning intelligent video monitoring system in general cargo terminal. J. Science and Technology Innovation., 18: 99-100.
[3] Lecun, Y., Bengio, Y., Hinton, G. (2015) Deep learning. J. nature., 7553:436-521.
[4] Schmidhuber, J. (2015) Deep learning in neural networks: An overview. J. Neural Networks., 61:85-117.
[5] Zhang, S., Lu, Y., Luo D. (2018) Overview of target detection algorithms based on deep learning. In: 22nd Network New Technology and Application Annual Conference Proceedings. Beijing. pp. 129-132+141.
[6] Cao, Y., Li, H., Wang, T. (2020) A review of research on target detection algorithms based on deep learning. J. Computer and Modernization., 05: 63-69.
[7] Girshick, R., Donahue, J., Darrell, T., et al. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Columbus. pp. 580-587.
[8] Girshick, R. (2015). Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. Boston. pp. 1440-1448.
[9] Ren Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. Montreal. pp. 91-99.
[10] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas. pp. 779-788.
[11] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.
[12] Gao, Z., Li, S. (2018) Pedestrian detection method based on YOLO network J. Computer Engineering, 487:221-225.
[13] Wang, D. (2020) YOLOv3 pedestrian detection algorithm based on depth separable convolution J. Computer Applications and Software, 37: 218-223.
[14] Liu, W. (2016) SSD: Single shot multibox detector. In: Computer Vision. 16:21-37.
[15] Li, H. (2020) Pedestrian head detection method based on SSD.J. Computer Engineering and Design, 41: 827-832.
[16] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Desmaison, A. (2019). Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. Vancouver. pp. 8026-8037.
[17] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence. San Francisco.
[18] Kaehler, A., & Bradski, G. (2016). Learning OpenCV 3: computer vision in C++ with the OpenCV library. O'Reilly Media, Inc. Sevastopol.