ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 156-163
doi: 10.18178/wcse.2019.06.023

A Novel Object Detection Algorithm in Video

Shengyu Lu, Junhao Liu, Beizhan Wang+, Wenxi Liu

Software School, Xiamen University Xiamen City, Fujian, China
18059204016, 13774661806, 13959238599, 18150085352
lu.s.y@foxmail.com, 935771982@qq.com, wangbz@xmu.edu.cn, 1729670798@qq.com

Abstract. Deep learning technology performs effect in feature extraction of images. Nowadays, with the development of video monitoring, the application of deep learning technology to surveillance video has profound implications. The effects of traditional video recognition are not satisfactory, but deep learning methods perform effect in many scenes of image classification. This paper proposed a novel object detection algorithm in video. It combined the traditional methods of extracting feature and deep learning algorithm to realize vehicle identification based on surveillance video. The method used the frame difference method and background subtraction to preprocess the image, and then trained a network model based on YOLO to perform object detection and obtain the categories and location information of the monitored vehicle. Compared with the existing object detection algorithms faster RNN, our method can achieve higher accuracy and can significantly shorten the time for detection, which can recognize the object of vehicle video quickly and efficiently. The method can meet the requirements of real-time detection.

Keywords: Object detection; video; YOLO network; real-time;

1. Introduction

Nowadays, the pressure on traffic has been increasing. In order to ease the pressure, the smart transportation system emerged. Well, object detection is the basis of smart transportation. Through identify the object information in the video and classify the object statistics, we can better control the road conditions, and reduce traffic congestion. It will provide powerful information support for informational command and traffic monitoring.

The current research on object detection mainly based on virtual point technology. The main challenge of traditional object detection technology is instability. When the bad weather emerges or the camera shakes, it is difficult to recognize objects accurately. With the development of deep learning, more and more laboratories are applying deep learning technology to the work of object inspection. Deep learning methods perform effect in ImageNet's competition based on the powerful ability of feature extraction. However, the deep learning methods are meeting the problem of long training time, and it is difficult to detect in real-time for surveillance video.

This paper uses the deep learning to complete the task of object detection and ensure the accuracy and stability of the detection. In the meantime, it applied traditional detection algorithms of frame difference algorithms to improve the detection efficiency and prepare for real-time detection.

The key of the algorithm is to preprocess the video and build an efficient deep learning model. The basis of deep learning algorithm is neural network, and convolutional neural network (CNN) performs well in the recognition of image. The difficulty lies in training an available convolutional neural network, which

+ Corresponding author. Tel.: 13959238599
E-mail address: wangbz@xmu.edu.cn

requires selecting the appropriate parameters and convolutional layers. In the meantime, we need to adjust the network structure based on our training data. The figure1 shows the framework of the method.

The method combines the traditional method with deep learning method to detect objects in video. Traditional video recognition methods include background subtraction, background difference, etc. The disadvantages of these methods are that the accuracy and stability are not good. However, they can have a fast speed in the progress of images. Therefore, it also combines both of them to improve the performance of detection.

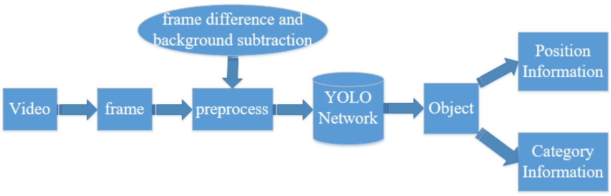

Fig. 1: The framework of the method.

2. Background and Related work

Object detection aims to find the position of target objects in the image and judge the categories of them. The traditional methods extract the features of images [1], such as SIFT[2], HOG[3], etc., However, these methods have the drawbacks of high time consuming and poor generalization ability.

Deep learning methods have the powerful ability of feature extraction. Object detection methods based on deep learning can divide into methods based on regional candidates and end-to-end methods.

The region candidate methods replace the traditional sliding window by searching a possible region of interest as the candidate region. In 2014, Ross Girshick team proposed the RCNN model[4]. The basic approach is to selectively obtain multiple candidate regions on the input image, and use CNN method to extract the features of regions, finally obtain the best region by non-maximal suppression. The advantage of RCNN is to make full use of the auxiliary information and select some windows to keep the recall rate at a high level. By using candidate windows from different images, it can improve the adaptability of features and solve some challenges such as window redundancy. Shaoming He and others proposed the SPP-net method for the complex steps and high time consuming of RCNN [5], it uses the strategy of spatial pyramid sampling and reduces the time consuming by the operation of convolving full maps, finally maps the location information of candidate regions to the feature map. In 2015, the Ross Girshick team improved the RCNN model and proposed the Fast RCNN, which mainly implemented the candidate window duplication and integrated all the models. However, it still adopted the strategy of selective search for candidate regions, and did not solve the long time consuming. In order to solve this problem, Girshick team designed the Faster RCNN algorithm[6], and they innovatively proposed using Region Proposal Network (RPN) to generate candidate regions. They made the regional candidate, regression and other methods share the convolutional features [7] so that the efficiency of object detection has improved significantly.

The end-to-end method is based on non-regional candidate, the two representative algorithms are the YOLO [8] and SSD [9]. They obtain candidate regions of the image by means of uniform segmentation operation. The method of comparing region candidates is faster and can implement the high accuracy of detection.

3. Method

3.1 Image preprocessing

The task of image preprocessing is to capture each frame of the video, and remove the random noise of the image to separate the background and foreground objects. At present, the background extraction methods for still cameras mainly include background subtraction [10] and frame difference method. The disadvantage of the background subtraction is that the background image needs to be preprocessed in advance, and it is extremely sensitive to environmental changes such as light. The frame difference method is to perform the difference between two adjacent frames. Since the change caused by the environment between the two frames is extremely weak, it can extract the shape features of objects more accurately. However, overlap occurs when the objects move too slowly. In this paper, it uses the method that combines frame difference and background subtraction to perform simple object detection and extraction.

Since the vehicle video is about 25 frames per second. The background and the moving object's pixel points change more significantly, so it can consider that the brightness of a particular pixel satisfies the Gaussian distribution [10]. It lets (x,y) represent the coordinates of a pixel, the brightness of the pixel satisfy B(x,y) ~ N(μ, σ2), where μ denote the mean difference of the Gaussian distribution at a certain moment, and σ2 denote the variance of the Gaussian distribution.

In the background model [11], each pixel has two parameters. The two parameters can define as:

$$\mu(\mathrm{x},\mathrm{y}) = \frac{1}{N}\sum_{i=0}^{N-1} F_i(x,y) \tag{1}$$

$$\sigma(\mathrm{x},\mathrm{y}) = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}[F_i(\mathrm{x},\mathrm{y}) - \mu(\mathrm{x},\mathrm{y})]^2} \tag{2}$$

By modeling each pixel in Gaussian method, it can calculate the mean and variance of each pixel in a period of time. The pixel points in the new frame image are used to differentiate with the corresponding pixel points in the Gaussian background. It is considered to exist the target object if it exceeds a threshold, and the background is considered if the value is lower than the threshold value. This completes the modeling of the background.

$$D_k(x,y) = |\ f_k(x,y) - f_{k-1}(x,y)\ | \tag{3}$$

By this way, it removed the similarity between the two frames. Each pixel in the obtained difference image is fitted with the corresponding Gaussian model. It adjusted the background model, and detected the specific target with the background subtraction [12].

The current object detection methods applied to practical projects incorporates a variety of existing algorithms. The advantages of multiple models can improve the performance of object detection. The advantage of applying the hybrid method of frame difference and background subtraction in this project is that it can significantly improve the accuracy of detection for the target object and avoid environmental interference. In the current study, there are still many problems that need to be further resolved.

3.2 YOLO network

YOLO network is an improved CNN [13]. The whole network mainly consists of convolution layers, pooling layers and full connection layer. YOLO network is an end-to-end object detection algorithm, which integrates the three processes of generating candidate regions, extracting regional features and classifying in object detection [14]. Compared with traditional methods, YOLO can implement the object detection of images quickly.

The YOLO network is shown below, including 24 layers. Where, the $L^{(3)}$ layer to the next layer and is similar to the $L^{(1)}$ layer and the $L^{(2)}$ layer. There are two convolution layers and max-pooling layers appearing alternately in each layer in YOLO network. The convolution kernel of all convolution layers are size=3, stride=1, pad=1. The activation function is ReLU(). All the max-pooling layers are size=2, stride=2. The output layer is a tensor with 7*7 (5*5+5) dimensions. The framework of YOLO network shows as Figure2.

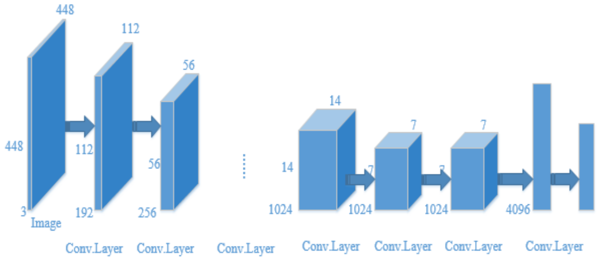

Fig. 2: The framework of YOLO network.

3.3 Loss function

It uses the sum of the squared errors (SSE) as the loss function [15]. The YOLO algorithm divides images into S*S grids, and the output of each grid is (B*5+C) dimensions, which include the location information, the confidence of the border box[16], and the number of categories. However, these factors have different effects on the accuracy of the object recognition of vehicles. In addition, many of the segmented grids do not contain objects. It is that the confidence of the boundary boxes built by these grids is 0. In general, the gradient in the training process of such grids is much higher than the grids containing objects, which will lead to unstable training and easy divergence [17]. For these objects, it sets different weights for different grids. The α_coord represents the weight of the grids which contain the center of an object, which is equal to 5. The α_noobji represents the weight of the grids which have no object.

As for the error of size and position of the boundary frame, the effect of size is more sensitive. Therefore, it replaces w, h with the $\sqrt{w}$,$\sqrt{h}$. It specifies that each border box only select one object. In detail, it calculates the value of IOU of the current bounding box and all reference bounding boxes, and selects the object with the maximum IOU value as the object detected by the current bounding box. The loss function is as following:

$$\mathrm{L} = \alpha_{coord} \sum_{i=0}^{S_2} \sum_{j=0}^{B} I_{ij}^{obj} \left[(x_i - \hat{x})^2 + (y_i - \hat{y})^2\right] +$$

$$\alpha_{coord} \sum_{i=0}^{S_2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(\sqrt{w_i} - \sqrt{\widehat{w_\iota}}\right)^2 + \left(\sqrt{h_i} - \sqrt{\widehat{h_\iota}}\right)^2\right] +$$

$$\sum_{i=0}^{S_2} \sum_{j=0}^{B} I_{ij}^{obj} \left(C_i - \widehat{C_\iota}\right)^2 + \alpha_{noobj} \sum_{i=0}^{S_2} \sum_{j=0}^{B} I_{ij}^{nobj} \left(C_i - \widehat{C_\iota}\right)^2 +$$

$$\sum_{i=0}^{S_2} I_i^{obj} \sum_{c \in classes} (p_i(c) - \widehat{p_\iota}(c))^2 \tag{4}$$

where, I_i represents whether contain an object in the center of grid i. if contain an object that the I_i is 1, otherwise it is 0. I_{ij} represents whether the object exists in the boundary box j of the grid i. If so it is 1, otherwise it is 0.

In the training process, it used the gradient descent method with small batches and impulse to converge quickly. Based on the derivative of the loss function, it used the inverse propagation method to update parameters continuously until the value of the loss function converges. The updated formula as follows:

$$M_{w(l)} = \mu M_{w(l)}(t-1) + \beta\left(\frac{L(t)}{w^{(l)}} + \gamma w^{(l)}\right) \tag{5}$$

$$w^{(l)} = w^{(l)}(t) - M_{w(l)} \tag{6}$$

where, $w^{(l)}(t)$ represents the impulse of $w^{(l)}$ in the t time; μ represents the rate of the impulse; β represents the learning rate and γ represents the weight attenuation.

It uses the impulse of the previous iteration to calculate the current impulse, which can avoid falling into the local minimum and converge faster.

4. Experiment

In this section, it uses several data sets and comparative methods to experiment [18]. It uses different evaluation metrics to describe the experimental results in detail, and then demonstrates the effectiveness and adaptability of our method. The main equipment of the experiment include: E5 processor, four-channel GPU (GTX-1080) and 64 GB memory. The algorithm implements by Python programming language and deep learning framework of tensorflow.

4.1 Data sets

The data sets were from surveillance video of urban smart transportation. It preprocessed surveillance video by a combined method of the frame difference method and background subtraction. It collected 6,000 images with a resolution of 1280*720. There are 4,000 images in the training set and 2,000 in the test set. In the training set, we marked 6428 positive samples for 20 categories of vehicles including cars, buses, trucks, etc. In the meantime, it also marked 8245 negative samples for 5 categories of other objects including motorcycles, pedestrians, etc.

It uses the samples with labels to train model, then inputs the surveillance video into the model. It gets the vehicle category and location information and implement the object detection in real time [19]. As shown in the figure3 below:

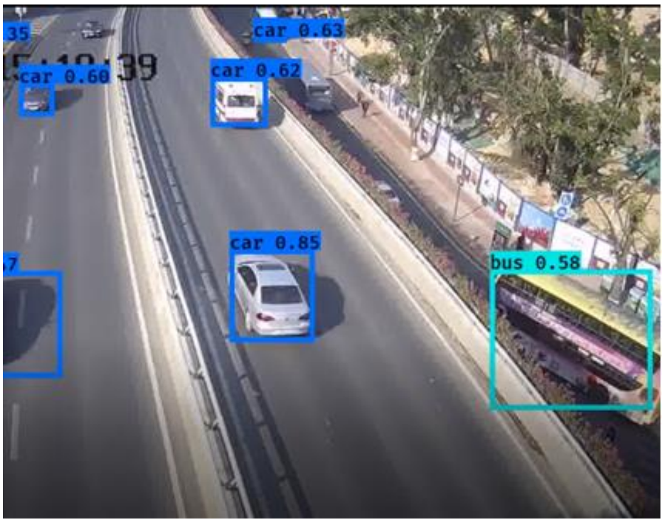

Fig. 3: Vehicle detection image.

4.2 Evaluation metrics

It used the accuracy rate and recall rate to evaluate the experimental results. Besides, calculates the IOU values of the detection boundary boxes and the reference boundary boxes to judge whether the result is true or false.

$$\text{IOU} = \begin{cases} \geq 0.5 \ \ ture \\ < 0.5 \ false \end{cases} \tag{7}$$

The accuracy rate and recall rate defines as:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{9}$$

where, TP, FP, and FN respectively represent the number of real cases, false positive cases and false negative cases.

4.3 Baseline methods

The approaches compared to our model as follows:.

1 Slide Window: Divide the image into several grids and use the sliding window to extract the features of each grid and predict the categories of objects respectively [10].

2 CNN: Image features are extracted by convolution operation, according to color, edge, texture and so on [16].

3 RCNN: Divide the image into several local areas, and input these regions to the CNN and get the regional features, then take them to the classifier to concludes that the area is an object or background [17].

4 Faster R-CNN: On the basis of RCNN, the extracted feature areas are mapped to the feature map of the last convolutional layer of CNN so that extract the feature of images only by once [6].

5 SSD: Different scales of feature mapping is extracted from the output of different layers. Divide the meshes into different scales and conclude the object categories in the grids [9].

4.4 Results and analysis

In the experiment, it used five data sets to test the performance of the algorithm and respectively obtained the precision rate, recall rate and the frames of recognition per second (FPS) of the YOLO algorithm and baseline methods from different data sets. It respectively calculated the average of each evaluation metric in the different data sets, as shown in the Table1.

Table 1: experimental results of evaluation metrics

Algorithm	**Precision (%)**	**Recall (%)**	**FPS(f/s)**
Slide window	70.59	72.91	1/15
CNN	80.82	78.38	3/10
RCNN	84.19	83.69	2
Faster R-CNN	83.96	82.65	6
YOLO	**88.34**	**86.22**	**30**
SSD	86.82	84.10	25

As can seem in the Table1, the YOLO algorithm can recognize 30 frames per second, far more than other baseline methods. The traffic surveillance video plays at a speed of 25 frames per second, so the YOLO method can realize real-time object detection if the device allows. Compared with other algorithms, the precision rate and the recall rate of YOLO algorithm can achieve better results, especially in terms of the recall rate. In the experiment, in order to ensure the real-time performance of YOLO algorithm, it sets the grid parameter S equal to 8. Therefore, it reduces the recall rate in some degree, because it may miss any object without in the center of the boxes.

Table 2: The precision rate of some categories of positive samples

Algorithm	car	trunk	bus
Slide window	65.35	70.26	68.65
CNN	81.88	75.38	79.98
RCNN	75.38	80.21	85.36
Faster RCNN	76.24	84.32	85.32
YOLO	**87.64**	84.58	**89.99**
SSD	80.25	84.29	83.58

Table 3: The precision rate of some categories of negative samples

Algorithm	motor	person
Slide window	68.34	72.16
CNN	80.21	72.31
RCNN	82.52	85.33
Faster RCNN	74.81	80.25
YOLO	81.21	88.42
SSD	76.82	80.16

It counted the precision rate of some categories including positive samples and negative samples in Table2 and Table3. From Table2 and Table3, it can see the precision rate of some kinds of categories by using different algorithms. Compared with other algorithms, the precision rate of YOLO algorithm for different objects is higher obviously. From the perspective of the recognition for different categories, the precision of the car and bus types are significantly higher than the trunk. It has a relationship with our data sets. The probability of the emergency of the cars and bused in our surveillance video is bigger than the trunks. With more number of samples with tags, it can train a more accurate model. The two categories of motors and person are similar. There are more negative samples in the data sets, and the precision rate of negative samples are higher than the positive samples.

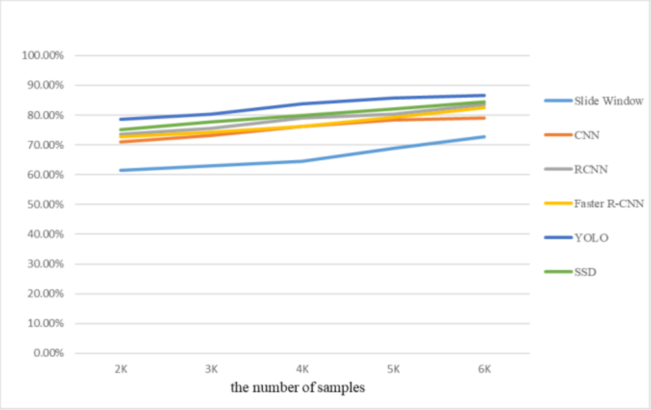

Fig. 4: The precision rate of data set

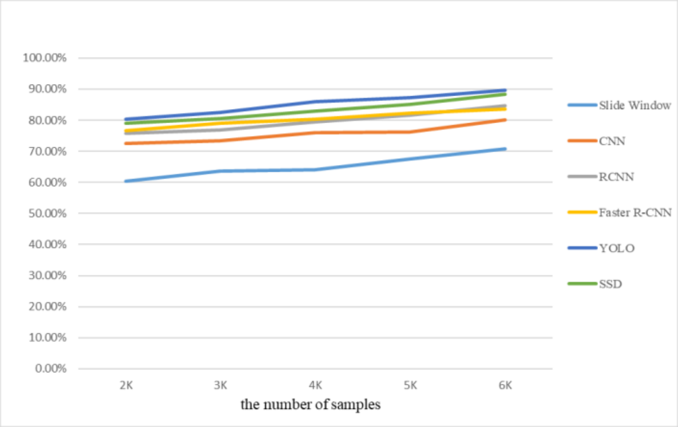

Fig. 5: The recall rate of data set

As can be seen from Figure4 and Figure5, YOLO algorithm and other baseline methods show an upward and stable trend in both precision rate and recall rate as the number of samples change from 2000 to 6000. When the number of training samples are small, the model is underfitting. With the increasing of the number of samples, it can make the model more fit the distribution of samples. When the number of samples reaches 6000, the model begins to converge essentially. The YOLO algorithm is always the best results in different cases.

5. Conclusion

This paper proposed a novel object algorithm for object detection in video based on YOLO network. The method combined the frame difference method with background subtraction to filter out the background of the images to improve the efficiency of detection. Then trained the deep learning network model based on YOLO to object detection and obtain the object categories and location information. The method made full of the advantages of traditional methods and deep learning techniques, and ensured the accuracy and speed of object detection. Compared with existing algorithms, the algorithm can implement high accuracy based on deep learning network, and reduce the time consuming greatly for image detection. The algorithm can implement the object detection of video in real-time, and obtain object information quickly and accurately. In the future, it will further optimize the YOLO network to improve the recognition accuracy and shorten the time of detection.

6. References

[1] D. M. Ramík, C. Sabourin, R. Moreno, and K. Madani, ‘A machine learning based intelligent vision system for autonomous object detection and recognition’, Applied Intelligence, vol. 40, no. 2, pp. 358–375, Mar. 2014.

[2] W.-L. Zhao and C.-W. Ngo, ‘Flip-Invariant SIFT for Copy and Object Detection’, IEEE Transactions on Image Processing, vol. 22, no. 3, pp. 980–991, Mar. 2013.

[3] K. Mizuno, Y. Terachi, K. Takagi, S. Izumi, H. Kawaguchi, and M. Yoshimoto, ‘Architectural Study of HOG Feature Extraction Processor for Real-Time Object Detection’, 2012, pp. 197–202.

[4] A. R. Pathak, M. Pandey, and S. Rautaray, “Application of Deep Learning for Object Detection,” Procedia Computer Science, vol. 132, pp. 1706–1717, 2018.

[5] G. Li, J. Liu, C. Jiang, L. Zhang, M. Lin, and K. Tang, ‘Relief R-CNN: Utilizing Convolutional Features for Fast Object Detection’, in Advances in Neural Networks - ISNN 2017, vol. 10261, F. Cong, A. Leung, and Q. Wei, Eds. Cham: Springer International Publishing, 2017, pp. 386–394.

[6] S. Ren, K. He, R. Girshick, and J. Sun, ‘Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks’, p. 14..

[7] S. Lu, “An Image Retrieval Learning Platform with Authentication System,” in 2018 13th International Conference on Computer Science & Education (ICCSE), Colombo, 2018, pp. 1–5.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ‘You Only Look Once: Unified, Real-Time Object Detection’, 2016, pp. 779–788.

[9] D. Biswas, H. Su, C. Wang, A. Stevanovic, and W. Wang, “An automatic traffic density estimation using Single Shot Detection (SSD) and MobileNet-SSD,” Physics and Chemistry of the Earth, Parts A/B/C, Dec. 2018.

[10] P. Sudowe and B. Leibe, ‘Efficient Use of Geometric Constraints for Sliding-Window Object Detection in Video’, in Computer Vision Systems, vol. 6962, J. L. Crowley, B. A. Draper, and M. Thonnat, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 11–20.

[11] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, ‘Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning’, arXiv:1602.07261 [cs], Feb. 2016.

[12] W. Shuigen, C. Zhen, and D. Hua, ‘Motion Detection Based on Temporal Difference Method and Optical Flow field’, 2009, pp. 85–88.

[13] K. F. Wallis, ‘The Two-Piece Normal, Binormal, or Double Gaussian Distribution: Its Origin and Rediscoveries’, Statistical Science, vol. 29, no. 1, pp. 106–112, Feb. 2014.

[14] S. S. Seferbekov, V. I. Iglovikov, A. V. Buslaev, and A. A. Shvets, ‘Feature Pyramid Network for Multi-Class Land Segmentation’, arXiv:1806.03510 [cs], Jun. 2018.

[15] S. Lu, B. Wang, H. Wang, and Q. Hong, “A Hybrid Collaborative Filtering Algorithm Based on KNN and Gradient Boosting,” in 2018 13th International Conference on Computer Science & Education (ICCSE), Colombo, 2018, pp. 1–5.

[16] S. Shinde, A. Kothari, and V. Gupta, “YOLO based Human Action Recognition and Localization,” Procedia Computer Science, vol. 133, pp. 831–838, 2018.

[17] J. Li, Z. Su, J. Geng, and Y. Yin, “Real-time Detection of Steel Strip Surface Defects Based on Improved YOLO Detection Network,” IFAC-PapersOnLine, vol. 51, no. 21, pp. 76–81, 2018.

[18] S. Lu, H. Chen, X. Zhou, B. Wang, H. Wang, and Q. Hong, “Graph-Based Collaborative Filtering with MLP,” Mathematical Problems in Engineering, vol. 2018, pp. 1–10, Dec. 2018.

[19] B. Wang, S. Tang, J.-B. Xiao, Q.-F. Yan, and Y.-D. Zhang, “Detection and tracking based tubelet generation for video object detection,” Journal of Visual Communication and Image Representation, vol. 58, pp. 102–111, Jan. 2019.