A Pedestrian detection method Based on YOLOv3 model and Image enhanced by Retinex

Hongquan Qu, Tongyang Yuan*, Zhiyong Sheng, Yuan Zhang
North China University of Technology
Beijing, China

Abstract—**Pedestrian detection is a basic technology in the field of intelligent traffic video surveillance. It is also help for the optimization design of rail transport. It is known that the deep learning technology can achieve considerable performance on pedestrian detection. However, this kind of methods demand a large number of high-quality samples. In addition, the quality of data sample in the subway station is usually sensitive to the background environment, such as variant illumination or pedestrian density, which can significantly affect performance of the deep neural network. To solve this problem, this paper adopts an image enhancement policy based on the Retinex theory to preprocess training samples to reduce the influence of light changes. Firstly, we use the image enhancement method to enhance the contrast of the image and highlight the color of the object itself. Next, we put the initial sample into darknet frame with YOLOv3 to train the detection model 1 and put the enhanced sample into the YOLOv3 to train the detection model 2. Finally, we tested these two models with 200 pedestrian pictures of four different scenarios. The experimental results show that the model trained by Retinex image enhancement has a more accurate detection rate of 94% compared with the model without the enhancement sample trained.**

Keywords-Retinex; Image enhancement; YOLOv3; pedestrain detection

I. INTRODUCTION

Pedestrian detection is applied to many scenes, such as intelligent driving assistance, passenger flow monitoring, pedestrian behavior analysis and so on. In recent years, the object detection based on deep learning has developed dramatically [1]. The common target detection methods divided into two species. They are detection methods based on region proposal and Single-pass detector. YOLOv3 (you only look once) belong to Single-pass detector. It is a fast and well-detected object detection technology. Compared with Faster R-CNN and SSD, YOLOv3 has a small lower detection accuracy than Faster R-CNN on very small targets, but the detection speed is much faster and can be better use for engineering [2]. At the same time, the detection accuracy of YOLOv3 is like Faster R-CNN when targets are not very small [7]. YOLOv3 is also better than SSD in aspects of detection speed and accuracy. Therefore, we select YOLOv3 as the target detection method in this paper.

However, the method of obtaining detection model by training a large number of samples is greatly influenced by the quality and quantity of samples. In situations where the light is weak, the object is not clearly compared with the background. Then we can't get a good detection model. Therefore, in this paper, we use Retinex enhanced to samples which have poor quality. These images may be due to high exposure then lead to overwrite or camera dust accumulation then lead to image fog sensation [5]. After Retinex enhanced the image contrast is improved and the sample quality improved with defogged.

In this paper, we collected samples at Beijing subway station, then we use image enhancement by Retinex to deal with the image. Then we got samples with clearer object and object more contrast with the background. After that, we trained pedestrian detection model by using the method of YOLOv3. We got two models. The model1 is trained with the original samples and the model2 is trained with enhanced samples. Then we tested the detection effect of the two models and analyzed the experimental results.

II. THEORY

A. Image Enhancement Based on Retinex Theory

First, Color images contain rich details and color information, but due to the influence of light, environment and other factors, the color images sometimes have the problems of low contrast and blurred image details. In order to obtain a clear color image, image enhancement techniques are often used to improve the image quality. Commonly used image enhancement methods are histogram equalization, low-pass filter high-pass filtering. The basic idea of histogram equalization is to improve the visual effect of the image by expanding the gray scale range of the image distribution so as to achieve the purpose of enhancing the contrast. However, this method synthesizes the multiple gray levels of the original image into one gray level, resulting in the loss of image detail. The high-pass filter or low-pass filter can only achieve the image smoothing or sharpening [6]. In image enhancement, it is usually necessary not only to adjust the dynamic range of the image, but also to highlight the details of the image. Color constant image enhancement technology is an image-based visual effects enhancement method, through the digital image computing and transformation to reflect the authenticity of the image color and enhance the contrast. Here to introduce the principle of Retinex.

Retinex is a synthetic word made up of Retina and Cortex. This theory is about how the human visual system adjusts the perceived color and brightness of an object. In the Retinex model, the image S (x, y) is composed of two parts, one is the

This work was supported by the National Key R & D Program Of China [grant number 2016YFB1200402] and National Natural Science Foundation of China (Grant No. 61806008)

978-1-5386-7604-2/18/$31.00 ©2018 IEEE

bright brightness component of the object and the other is the reflection image of the object [5]. They are represented by R (x, y) and L (x, y). L (x, y) denotes incident light and R (x, y) denotes the reflection property of the object. The imaging process can be expressed by

$$S(x,y) = R(x,y) \bullet L(x,y) \tag{1}$$

If we can separate the brightness image and the reflection image in the original image, we can achieve the purpose of image enhancement by changing the ratio between the brightness image and the reflection image. The Retinex schematic diagram is shown in Fig.1

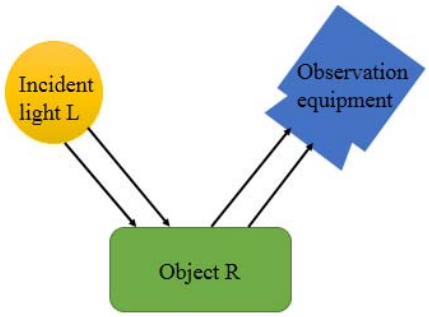

Figure1. Retinex schematic

It is impractical to obtain the reflected part of the object directly but S is known. It is possible to estimate the incident light by estimating the luminance image of the incident part and then obtaining the reflected image R by the (1). Jobson proved that the Gaussian function can estimate the luminance image from the known image S very well. The Gaussian function formula can be expressed as

$$F(x,y) = k \bullet e^{\frac{x^2+y^2}{c^2}} \tag{2}$$

K is the normalization factor and 'c' is the scale parameter of the Gaussian function. The logarithmic transformation of (1) can transform the complex product into a simple addition or subtraction form and the logarithmic form is closer to the perceived ability of the human eye to brightness. The logarithmic transform of (1) is as follows

$$r(x,y) = s(x,y) - l(x,y) \tag{3}$$

$$r(x,y) = \log(R(x,y)), s(x,y) = \log(S(x,y)). I(x,y) = \log(L(x,y)) \tag{4}$$

Then we can get enhanced model as below

$$r(x,y) = \log[R(x,y)] = \log\frac{S(x,y)}{L(x,y)} = \log[S(x,y)] - \log[S(x,y) \bullet F(x,y)] \tag{5}$$

After further linear transformation can be enhanced image. Retinex's mathematical formula is further summarized as

$$R_i(x,y) = \log S(x,y) - \log[F(x,y) * S_i(x,y)] \tag{6}$$

$S_i(x,y)$ is the i color band of the original image, * is the convolution operation, $F(x,y)$ is the surround function. We generally select the Gaussian surround function. The constraint of Gaussian function is by

$$\iint F(x,y)dxdy = 1 \tag{7}$$

Therefore, the convolution in the single-scale Retinex algorithm can be regarded as the calculation of the illumination image in space. Its physical meaning can be expressed as: estimating the illumination variation in the image by calculating the weighted average of the pixel points in the image and the surrounding area, this illumination component is removed, and finally only the reflection properties of the image are retained, so as to achieve the enhancement purpose.

B. The introduction of YOLOv3

The YOLOv3 predicts bounding boxes using dimension clusters as anchor boxes and predicts 4 coordinates for each bounding box [3]. It predicts an object score for each bounding box using logistic regression and use sum of squared error loss [4]. This should be 1 if the bounding box prior overlaps a ground truth object by more than any other bounding box prior. If the bounding box priories not the best but does overlap a ground truth object by more than some threshold it will ignore the prediction [2]. It uses darknet-53 network to extract features. This new network is a hybrid approach between the network used in YOLOv2, Darknet-19, and that newfangled residual network stuff.

The network structure of YOLOV3 is divided into the following modules. The first layer is convolutional layer, we put a picture into the layer. The channel number of the picture is 3 and the pixel of the picture is 416*416.We use 32 convolutional layers to the picture to extract features. Each convolution kernel size is 3*3 and the step size is 1. After convolution operation, we get 32-channel feature map of 416*416.

Second layer is res layer, the res layer is derived from resent. In order to solve the phenomenon of gradient dispersion or gradient explosion of the network, it is proposed to change the layer-by-layer training of the deep neural network into phase-by-stage training, and divide the deep neural network into several sub-segments, each of which contains relatively shallow The number of network layers, and then use the shortcut connection method to make each small segment train the residuals, each part learns a part of the total difference (total loss), and finally reaches the overall small loss, at the same time, a good control gradient Spread to avoid situations where gradients disappear or explosions are not conducive to training.

The third part is darknet-53: from the 0th layer to the 74th layer, there are 53 convolutional layers, and the rest are res layers. As the main network structure for yolov3 feature extraction. The structure uses a series of 3*3 and 1*1 convolutional convolution layer. These convolutional layers are obtained by integrating convolutional layers with better performance from various mainstream network structures. The structure of darknet53 is as below. It is much better than darknet-19. At the same time, it is 1.5 times more efficient than resnet-101 in the case of better performance. It almost doubles the efficiency of resnet-152 with the same effect as resnet-152.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure2. Darknet-53

The last layer is the characteristic interaction layer of the yolo network. It is divided into three scales. In each scale, local feature interaction is realized by convolution kernel, which acts like a fully connected layer but through a convolution kernel. It is the way to achieve local feature interaction between feature maps, and finally complete classification and regression on this foundation [3].

YOLOv3 predicts boxes at 3 different scales. Then extracts features from those scales using a similar concept to feature pyramid networks. At last it predicts some tensor encoding bounding box and object. In our experiments we have two classes and predict 3 boxes at each scale so the tensor is N × N × [3 ∗ (4 + 1 + 2)] for the 4 bounding box offsets, 1 object prediction and 2 class predictions. After train, we will get the target box and category probability, by setting the threshold value we can predict the target.

III. EXPERIMENTS

In this paper, we do experiments on the subway passenger flow data and the picture size is 640*480. The experimental data comes from a sample set we made by ourselves. We get pictures of pedestrians by the camera installed in the subway. The total data number is 10000. We selected 9000 images among them randomly as the training set and the rest of the images constitute the test set. The experimental environment was p100 GPU server, ubuntu16.04 operating system, darknet network framework. And we use detection rate and false alarm rate as experimental evaluation indexes.

A. Experiments on image enhancement

First using python's OpenCV module to write code about Retinex enhancement. Retienx algorithm implementation process is as follows:

- We read into the original image, our sample are color image so each color image processing separately, the pixel value of each component is converted from an integer value to a floating-point number, and converted to the logarithm domain for easy follow-up calculation. Original picture and RGB components are as Fig.3.

Figure3. (a) The original image
(b) The blue component of the image
(c) The green component of the image
(d) The red component of the image

- enter the scale C; under discrete conditions, the integral is converted to sum, while further determining the value of the scale parameter K.
- According to the (6), calculated $r_i(x, y)$ y
- convert $r_i(x, y)$ from the logarithmic domain to the real number field to obtain the output image $R(x, y)$.
- linear stretch of $R(x, y)$ to get the corresponding output format display.

The comparison of unused image enhancement and image enhancement is as Fig.4, We can observe the (b) picture, the contrast is enhanced and picture become more clearly.

Figure4 Image enhancement results

(a) picture without enhancement

(b) picture with enhancement

B. Training results by YOLOv3

- Prepare the data and put the data into the same folder, including the label file and the image file.

- Create an index file, list the addresses of each image, put them into box_train.txt and box_test.txt, and normalize the data.
- Download pre-trained weights on ImageNet
- Modify file cfg/voc.data. Set classes=1, the category name is “head shoulder” and “ignore”. Modify the path of the training sample set, verify the path of the sample set and the path of the image name file.
- Create a new folder named backup under the darknet folder. We use it to store the model obtained during the training and the final model.
- Modify data/voc.name to the tag name of the sample set.
- Modify cfg/yolov3-voc.cfg. Set the working mode to training mode, the network input width = 640, height = 416, channels = 3, momentum = 0.9, decay = 0.0005, saturation = 1.5, exposure = 1.5, hue=.1, max batches = 50200, policy = steps, steps = 40000,45000, scales = 0.1,0.1 number of convolution kernels filters=32, convolution kernel size size=3, convolution kernel step long stride=1, activation function activation=leaky, ignore thresh = .5, no target threshold is 0.5, truth thresh = 0.9, target threshold 0.9. codify cfg/yolov3-voc.cfg. Set the working mode to training mode, the network input

The experimental training obtained two models, the model1 is a model obtained by using the original image training, and the model2 is a model trained by the Retinex-enhanced image sample. The test results of the two models are tested using a variety of pedestrian pictures, the experimental results are as follows

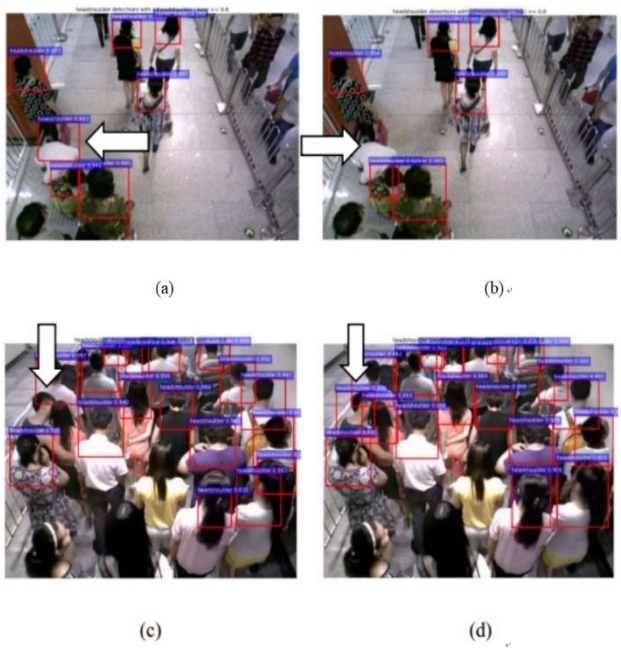

Figure5. test results

In Fig.5 (a) and (c) show the results of testing using the model obtained from the training of non-enhanced pictures and (b), (d) show the results of testing using the model obtained from the enhanced picture training. In the lower left corner of figure (a), the bag is detected as a human being. Figure (b) does not exist this wrong so that the enhanced model reduces the detection of false detection. In the figure (c), Two lovers to the left were not detected but were detected in figure (f). It can be learned that the enhanced model enhances the detection effect.

After we use a lot of image to test model one and model two. We can observe that the model obtained with the enhanced image training shows advantages in two aspects. Firstly, it reduces the false alarms when the same number of pedestrians can be detected. That is, some backpack trash can be misjudged as pedestrians. Secondly, some pedestrians whose color cannot be distinguished from the background color, the target can’t be detected by the original model and the detection effect has also been improved.

TABLE I. DETECTION RATE AND FALSE ALARM RATE OF MODEL 1 AND MODEL 2

test results	**The evaluation index**			
	detection rate of model1	***detection rate of model2***	***false alarm rate of Model 1***	***false alarm rate of Model 2***
value	90%	94%	5%	2%

C. Conclusion

Based on deep learning and Retinex image enhancement, this paper uses the YOLOv3 to train the pedestrian detection model and improve the detection accuracy. The test results show that the average recognition rate of the model is 94%. The detection effect is good and the false detection rate is much lower than the model without image enhancement. Therefore, the use of Retinex image enhancement method to pretreatment subway passenger flow pictures can make YOLOv3 learn better model. Moreover, the trained model has a precise detection rate for the case where the number of single-image pedestrians is less than 25. So, our method is feasible and this model can be applied to many scenes.

[1] S Walk, N Majer , K Schindler ,B Schiele, “New features and insights for pedestrian detection” IEEE/Computer Vision & Pattern Recognition , 2010 , 119 (5) :1030-1037

[2] Joseph Redmon, Ali Farhadi “YOLOv3: An Incremental Improvement”, Computer Vision and Pattern Recognition,2018, arXiv:1804.02767

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “ You Only Look Once: Unified, Real-Time Object Detection” Computer Vision and Pattern Recognition,2016, arXiv:1506.02640

[4] Joseph Redmon, Ali Farhadi, “YOLO9000:Better, Faster, Stronger”, Computer Vision and Pattern Recognition,2017, arXiv:1612.08242

[5] Z Rahman, DJ Jobson, GA Woodell, “Multi-scale Retinex for color image enhancement”, International Conference on Image Processing , 2002 , 3 :1003-1006 v

[6] KEAVD Sande , T Gevers , CGM Snoek, “Evaluating Color Descriptors for Object and Scene Recognition”, IEEE /Transactions on Pattern Analysis & Machine Intelligence, 2010 , 32 (9) :1582-96

[7] S. Ren, K. He, R. Girshick, and J. Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks”. 2015 arXiv:1506.01497

[8] Yihong L I, Zhou X. L “Multi-resolution and multi-scale retinex for color image enhancement”. Computer Engineering & Applications, 2017.

[9] Hanumantharaju M C, Ravishankar M, Rameshbabu D R, et al. Color Image Enhancement Using Multiscale Retinex with Modified Color Restoration Technique[M]. IEEE, 2011..