Video Object Detection for Tractability with Deep Learning Method

Bing Tian ,Liang Li，Yansheng Qu，Li Yan
Information and Telecommunication Company of State Grid Shandong Electric Power Corporation
Jinan 250021, China
tianbingjn@126.com

***Abstract*—In this paper, a video based objection detection method is proposed for traceability system with deep learning method. The surveillance video is collected first, from which an annotated image database of target object such as people or vehicle was constructed to train convolutional neural network model off-line. With the trained model, a real-time target detection and recognition system is designed and implemented. The proposed method mainly includes three aspects: video processing, target detection and object recognition. It provides a variety of video interfaces to support the downloaded video and real-time video stream. The experimental results indicate that the proposed deep learning based detection method is efficient for the traceability application.**

Keywords—object detection; deep learning; traceability;

I. INTRODUCTION

Along with the wide deployment of video surveillance system, it is becoming increasingly important to extract the useful information from the video system automatically. However, the existing surveillance system do not supply the target identification function especially for the given scenario. In this paper, a deep learning based framework is proposed to detect the given objects from video.

In the video surveillance system, the targets change easily in different scenes with different lighting or shadow. The traditional object detection method in the case of complex scenes, cannot correctly identified the targets especially multiply types in real time. Along with the development of data mining method, deep learning has shown the advantages in model expression ability. It has good results on target detection and behavior identification. The convolution neural network in deep learning uses local perception, weights sharing and down sampling, which has good robustness on Displacement, scaling and distortion. Currently, the convolution neural network (CNN) has developed into an efficient image recognition method to be applied to many ways, such as license plate recognition and face detection.

In this paper, we apply the deep learning based object detection for video surveillance of production, warehousing z traceability system. The rest of the paper is structured as follows: related work is given in Section 2; Section 3 explains the proposed object detection algorithm; Experimental results are shown in Section 4 and Section 5 concludes the whole paper.

II. RELATED WORK

A. Target Detection

Fi Target detection is a basic research in computer vision field, mainly focus on two different detection tasks: target instance detection and target categories detection. Target detection task can be divided into two key sub-tasks: target classification and target location. Target Classification determine whether category of objects interested from the input image appear, and output a series of label with score indicating the possibility of objects we are interested in. The target location determine the Location and border of the object we are interested in from the input image, and output the information of Location and border of the object.

For instance target detection, it can be further subdivided into non-textured target instance detection and texture target instance detection. When the texture of the target is rich, the target instance can actually extract rich stable feature points and the corresponding feature descriptor so that they can be accurately identified and detected based on these feature points and feature descriptor.

Most non-textured target instances detection is based on template matching method. All binary image using by these methods is gotten through the edge extraction algorithm. So they are extremely sensitive to changes of light and noise. Hinterstoisser[1-2] have proposed two kinds of non-textured object detection algorithm based on image gradient direction as the feature by using a template matching technique. Rios-Cabrera[3-4] use machine learning to improve the discrimination of template feature to improve the detection accuracy. In order to strengthen the edge connectivity constraints, Hsiao[5] proposed a new shape matching algorithm. However, all algorithms failed to solve problems for detection accuracy attenuation caused by occlusion. Occlusion in various topics in the field of computer vision is a difficult problem.

The research for target categories detection has been a hot topic in computer vision. Viola[6] proposed framework based on AdaBoost algorithm. It is the first target category detection algorithm which timely process and give a good detection rate and mainly used in face detection. Felzenszwalb[7]proposed one

978-1-5386-1072-5/17 $31.00 © 2017 IEEE
DOI 10.1109/CBD.2017.75

of the most influential method of target category detection called deformation part multi-scale model (DPM), inherits the advantages of using HOG features and SVM classifier.

By 2012, the best algorithms are improved algorithms based on a variety of DPM framework. In 2012, Krizhevsky[8] proposed an image classification algorithm the depth of convolution neural network (DCNN) which is based on depth learning theory. Szegedy[9] saw the target detection problem as the regression of target mask and using DCNN as the regression prediction of the target mask in the input image . Erhan[10] used DCNN to do regression forecasting for the bounding box of the target, and gives confidence level for each bounding box containing class independent object. Sermanet[11] proposed a framework of DCNN OverFeat, integrated identification, positioning and inspection tasks. Different from OverFeat, R-CNN select sliding window search strategy to improve the detection efficiency. Currently, the depth of the convolutional neural network are the best results achieved on multiple target category detection data set.

B. Deep Learning

The concept of deep learning stems from the artificial neural networks, and its motivation is to establish mechanisms to simulate the human brain to learn data characteristics. In 2006 Hinton[12] proposed a basis to make it possible for depth structure of artificial neural networks to improve the learning ability of neural network. On this basis, Ranzato[13], Lee[14] use the sparse coding scheme optimized DBN. Tang[15] applied the method of noise reduction to RBM, making their learning performance improved further. In 2007, on the basis of DBN studies, Bengio[16] discussed the validity of the depth of learning when the CD training algorithm used in RBM replace as Stacked AutoEncoder. In the following years, an amount of work in connection with various applications to demonstrate the effectiveness of SAE algorithm. To optimize learning performance, a lot of improvement of the encoder is proposed. For example, Vincent[17] proposed Stacked Denoising AutoEncoder. Yu[18] proposed Sparse AutoEncoder and Rifai[19] proposed Contractive AutoEncoder. These improvements fundamentally add constraints so that the encoder can learn certain characteristics invariance.

In different fields of computer vision, the integration of deep learning and domain knowledge provides a new way of thinking for visual perception task. An important advantage for depth model to extract information from raw data on pixel level to abstract semantic concept layer by layer. This makes it have outstanding advantages in terms of global features and contextual information when extracting image.It bring the breakthrough for solving the task of tagging by pixel (such as image segmentation). Full convolution network receives image of any size as an input.

III. System IntroductionLing

The system mainly runs through two phases. In the model training phase, we need to collect the target data set from the video, construct and train neural network convolution. In target detection and recognition phase, trained CNN model with target location technology complete the task of target detection and identification.

A. Object Detection

Since there will be illumination changes in videos captured by surveillance cameras, high-frequencies background objects, camera oscillations and other disturbances. These disturbances can cause a lot of trouble when we split foreground objects from background. Inter-frame difference is very sensitive to camera shake. Once shake it is difficult to detect the moving foreground object. However, mixed Gaussian background model is sensitive to illumination changes and high-frequencies background objects. These changes will lead to a lot of noise in the foreground object. In addition, whether mixed Gaussian background model or non-parametric background model is used to split the background and foreground, but cannot determine whether there has been movement of foreground.

We use a combination of non-parametric background model and Inter-frame difference to get the number of pixels belonging to the change of foreground between adjacent two frames. According to whether the number of pixels changed is greater than a given threshold value, we determine whether there has movement so that we select frames we need.

This method detect whether movement occurs by subtracting adjacent frames after subtracting background model and largely overcome effect of camera shake on the inter-frame difference. It also eliminates the noise on the dichotomous image after background subtraction so that we can get more clear and accurate part of the movement foreground object.

After obtaining the key frames by background subtraction, we give different treatment depending on the phase.

1) Model training phase: We invite students classify these images and add tags through crowdsourcing platform and store images into database. These images will be used as the training dataset for CNN.

2) Target detection and recognition phase: When in this phase, we need to put key frames into trained CNN model for target detection and recognition. How to message between the target detection module and deep learning module has become our concern.We use redis to implement the data passing between the video processing module and deep learning module. First, we create a redis message queue. When there is new data to be pushed, we pop it from the queue and read the label of the image. Then we find the image according to the path we get from the label and do target detection and recognition. And the results will be saved to the database.

B. CNN model construction

First we introduce Hinton team's AlexNet Network in 2012.The input is 224 * 224 RBG picture with 3 channels. The convolution layer of 1st layer has 96 convolution kernel with the size of 11 * 11, and the sample layer use a 2 * 2's max-pooling; 2nd layer has 256 convolution kernel with the size of 5 * 5, and the sample layer use a 2 * 2's max-pooling, too; 3rd layer has 384 convolution kernel with the size of 3 × 3, which is same as the 4th layer. 5th layer has 256 convolution kernel with the size of 3 * 3, and the sample layer use a 2 * 2's max-pooling. The convolution layer of 1st layer use the output from 5th layer max-pooling changing into a one-dimensional vector as the input layer, with the total of 4096 dimension. Then go through 2nd

full connection layer with the same 4096 dimension. Finally it attach to softmax layer, output a 1000-dimension vector. Each dimension is the probability of image belonging to this category to give prediction of top-k type. CNN network structure of this paper is changed based on AlexNet, following is specific network structure:

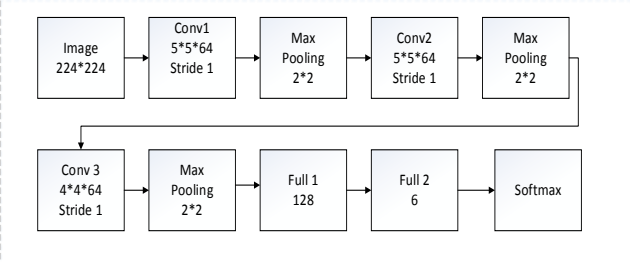


Fig. 1. Convolution neural network structure

The input is a three-channel RGB image, with a uniform size 224 * 224. 1st layer has 64 convolution kernel with the size of 5 * 5 and the sample layer use a 2 * 2's max-pooling; 2nd layer is same as the 1st layer. 1st layer has 64 convolution kernel with the size of 4 * 4 and the sample layer use a 2 * 2's max-pooling; The output of 1st full connection layer is a 128- dimension vector. The output of 2nd full connection layer is a 6- dimension vector. It finally attach to softmax layer too. The stride if all layers is 1.

IV. Experiment Results

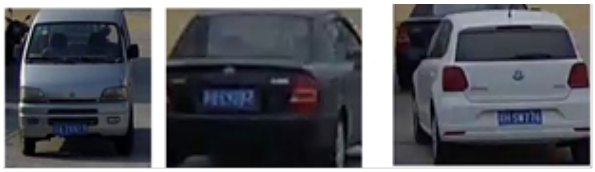
A.

We roughly divide effective target into six categories from traceable video {people, tractors, bicycles, cars, trucks, other} as shown in Fig 2. The training sample is mainly divided into these six categories, and each category select 6000 images. To make the trained model achieved good classification results in this scene, the sample should try to choose different angles, different colors and different lighting conditions, as much as possible to increase the diversity of the sample.

Training dataset generally consists of two parts, one is the image itself, the second is contained in the label of image data. We disrupt the image data sets by python script, and each category randomly select 5000 as the training set and 1000 as the validation set. At the same time we prepare a text file for each collection, which has a list of image information.

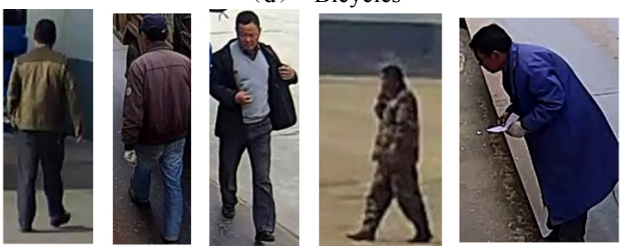
(a) Trucks

（b）Tractors



（c）Cars

（d） Bicycles



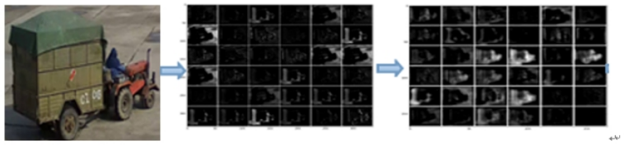
(e) People

(f) Others

Fig. 2 .The target object in the test scenario

We use Caffe platform for deep learning training. We use convert_imageset tool provided by Caffe totransfer data sets to lmdb formats and resize parameters of the tool can modify the dimensions of the image to any size, such as 224 * 224. Traditional machine learning to train the model needs to initialize the data, the depth of learning as well. Traditional machine learning for model training needs to initialize the data, as well as deep learning. In this paper, we use data equilibration and data normalized to preprocess data.

The weight of each convolution layers uses a Gaussian distribution to initialize, stochastic gradient descent for model learning and dropout to prevent over-fitting. The value of dropout is set to 0.5. Since the amount of data is not large, the learning rate cannot be set too high, so we set 0.001. Every 100 iteration carried out a test. Momentum is set to 0.9. The weight attenuation coefficient is set to 0.0005 weight and batch-size is set to 512.



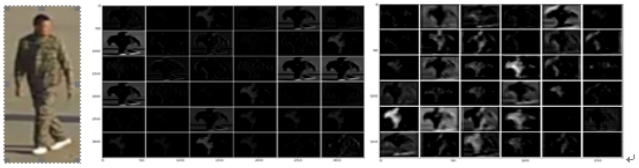


Fig. 3 .The first two convolution layer for tractors and people

B. *Analysis of CNN Training Results*

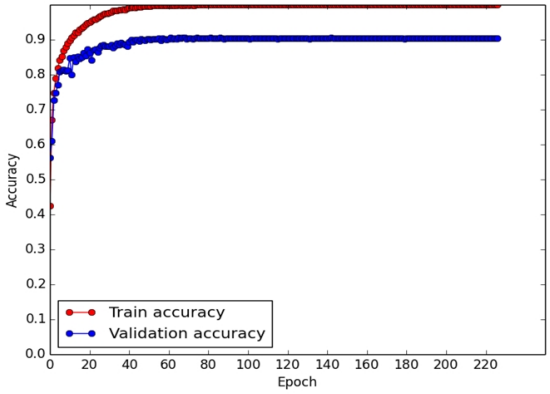


Fig. 4. Train Accuracy

Figure 4 is the training set and validation set accuracy figure in the training process corresponding to the epoch. Epoch refers to forward propagation and backward propagation of all the training data for one time. Red is the training set accuracy which comes to 99.9% after 50 epoch and the highest accuracy rate comes to 99.99%. The accuracy of verification set is probably stabilized after 50 epoch and the accuracy rate hovering at around 90.5%. Compared to the traditional methods, 90.5% have been greatly improved. This paper aims to design and implement an intelligent real-time target detection and identification system. 90.5% accuracy rate has been basically able to meet the demand for recognition and classification of this stage.

The Improvement for the accuracy has two main directions, one is the number increase of data, and another is to modify the CNN network model. The data we collected of the non-motor vehicles and trucks is relatively small, so in order to make the number of each category of image equal, we use only the 6000 picture for the training set and validation set of each category. The number is relatively small for the training deep learning, more data is required to be trained in order to improve the accuracy of the validation set. The CNN model of this paper is mainly based on modifications of AlexNet and without modifying the size of the input image. As a result, the motion area directly scaled to the size of 224 * 224. Such deformation process is likely to affect the expression of the network model on image features. After AlexNet there have been many great CNN network model so we can try to use other better network model.

Finally the real time traceability results is given in Fig.5 as follows.

Fig .5. Real time traceability

V. CONCLUSION

In this paper, we design and implement a target detection system based on motion detection and CNN. First, we give the overall solution of the system. Then we introduce two video collection services provided by the system. One is downloading video from the NVR servers by restful and the other is real-time live video stream. The part of Motion detection introduces a method of target deception which combines nonparametric modeling and frame difference and make a comparison to common methods. Data set part classifies and labels the object detected to prepare for further CNN training. In the part of CNN model, we detail the design of the CNN model, the process of CNN network model training and we analyze the results of CNN training.

REFERENCES

[1] Hinterstoisser S, Lepetit V, Ilic S, et al. Dominant orientation templates for real-time detection of texture-less objects[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 2257-2264.

[2] Hinterstoisser S, Cagniart C, Ilic S, et al. Gradient response maps for real-time detection of textureless objects[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2012, 34(5): 876-888.

[3] Rios-Cabrera R, Tuytelaars T. Discriminatively Trained Templates for 3D Object Detection: A Real Time Scalable Approach[C]//Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, 2013: 2048-2055.

[4] Rios-Cabrera R, Tuytelaars T. Boosting masked dominant orientation templates for efficient object detection[J]. Computer Vision and Image Understanding, 2014, 120: 103-116.

[5] Hsiao E, Hebert M. Gradient Networks: Explicit Shape Matching Without Extracting Edges[C]//AAAI. 2013.

[6] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. IEEE, 2001, 1: I-511-I-518 vol. 1.

[7] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, 32(9): 1627-1645.

[8] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[9] Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection[C]//Advances in Neural Information Processing Systems. 2013: 2553-2561.

[10] . Erhan D, Szegedy C, Toshev A, et al. Scalable Object Detection using Deep Neural Networks[J]. arXiv preprint arXiv:1312.2249, 2013.

[11] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv:1312.6229, 2013.

[12] . HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7) : 1527-1554 .

[13] . Ranzato, M.; Boureau, Y.; LeCun, Y. Sparse Feature Learning for Deep Belief Networks. NIPS 2007.

[14] . Ranzato, M.; Boureau, Y.; LeCun, Y. Sparse Feature Learning for Deep Belief Networks. NIPS 2007.

[15] . Tang, Y.; Salakhutdinov, R. and Hinton, G. Robust boltzmann machines for recognition and denoising. CVPR 2012: 2264–2271.

[16] Bengio, Y.; Lamblin, P.; Popovici, D. and Larochelle, H. Greedy Layer-Wise Training of Deep Networks. NIPS 2006: 153-160.

[17] Vincent, P.; Larochelle, H.; Bengio, Y. and Manzagol, P. Extracting and composing robust features with denoising autoencoders. ICML 2008:1096–1103.

[18] Yu, K.; Lin, Y. and Lafferty, J. Learning Image Representations from the Pixel Level via Hierarchical Sparse Coding. CVPR 2011:1713-1720.

[19] . Rifai, S.; Vincent, P.; Muller, X. et al. Contractive auto-encoders: Explicit invariance during feature extraction. ICML 2011.4. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)