Object Detection Based On Deep Learning

Mohanish Gopal Saim ,Vyankatesh Pujare
Sindhudurg Shikshan Prasarak Mandal's College Of Engineering,Kankavli, Maharashtra, India.

***Abstract*— The object detection based on deep learning is an important application in deep learning technology, which is characterized by its strong capability of feature learning and feature representation compared with the traditional object detection methods. The paper first makes an introduction of the classical methods in object detection, and expounds the relation and difference between the classical methods and the deep learning methods in object detection. Then it introduces the emergence of the object detection methods based on deep learning and elaborates the most typical methods nowadays in the object detection via deep learning. In the statement of the methods, the paper focuses on the framework design and the working principle of the models and analyzes the model performance in the real-time and the accuracy of detection. Eventually, it discusses the challenges in the object detection based on deep learning and offers some solutions for reference.**

Keywords-object detection; deep learning; framework design; model analysis; performance analysis;neural networks

I. INTRODUCTION

Object detection has already been the significant research direction and the focus in the computer vision , which can be applied in the driverless car, robotics, video surveillance and pedestrian detection . The emergence of deep learning technology has changed the traditional modes of object identification and object detection. In the Imagenet Large Scale Visual Recognition Challenge (ILSVRC) 2012, a global contest about computer vision, the AlexNet won the championship, which was the first successful deep convolution network in image recognition, and its Top5 accuracy surpassed the runner-up by 10%. Moreover, the methods of deep learning topped the list in the succeeding LSVRCs. In the year 2013, the LSVRC added the object detection, which facilitated the development of deep learning in object detection. The deep neural network has the strong feature representation capacity in image processing and is usually used as the feature extraction module in object detection. Deep models don't need special hand engineered features and can be designed as the classifier and regression device. Therefore, the deep learning technology is of great prospect in the object detection.

II. REVIEW OF OBJECT DETECTION

Object detection is an application to detect the object from the specified scenes by a certain measure or method. Before the emergence of deep learning technology, the methods of object detection are primarily accomplished by establishing the mathematical models based on some prior knowledge. At present, the common classical methods in object detection are as follows: Hough transform method, frame-difference method, background subtraction method , optical flow method, sliding window model method and deformable part model method. The Hough transform can transform image space into parameter space. Every pixel in the image space corresponds to a curve in the parameter space and the coordinates of the intersection of most curves by voting in the parameter space are the parameters of curve in image space. The common Hough transform is only applied to the object detection in which the object contour can be expressed with the analytic function, such as roundness, straight line, etc. The generalized Hough transform could detect the objects of any shape by combining the graphic edge information and the edge point direction information, which is of higher speed and more accuracy compared with the common Hough transform. Furthermore, the generalized Hough transform could not only carry out the shape detection, but also the category detection. For the frame-difference method, the principle is that the difference image results from subtracting the two adjacent frame images and is denoised by binaryzation processing and morphological filtering to get the object motion area. At present, the widely-adopted methods are two-frame difference method, three-frame-difference method and four-frame-difference method. The background subtraction method has three processes that are background modeling, object detection and background updating. The process of background subtraction method is similar to that of the frame-difference method and the difference is that the former needs to define a background frame and update in a timely manner. The background modeling technologies of defining the background frame are usually done by combining with the image features, which in general are luminance information, texture information and spatial information, such as Mixture of Gaussians Models (GMM) and Local Binary Patterns (LBP). The optical flow method is brought up by Horn and Schunck. The method assumes that the gray change is merely related with the object motion and delineates the motion of image pixel by establishing an optical flow equation, thereby delineating the motion of object. The sliding window model, by setting the sliding window of fixed sizes, slides on the image according to some strategies while extracting the features in the sliding window and then classifies them by some classifier. In the model, the features can choose color histogram, gradient histogram, SHIFT. And the classifier can choose SVM and the Adaboost classifier.

2017

30%. The Overfeat is proposed by LeCun's team, which extracts features with the improved deep convolutional model

AlexNet, enabling the offset and slide window to realize the goal of object classification by using images of various scales and locate objects by combining the regression network, thus

III. EMERGENCE OF OBJECT DETECTION BASED ON DEEP LEARNING

In the mode of region selection + feature extraction + classification adopted by object detection method based on deep learning, the region selection can be done according to the some strategy, the feature extraction can be achieved by the convolutional neural network and the classification can be realized by traditional SVM or the special neural network. The early typical modes of deep learning applied in object detection are DNN and Overfeat, which draw up the curtain for deep learning applying in object detection. The object detection by DNN has designed two subnetworks that include the classification subnetwork for recognition and the regression subnetwork for location. Originally, DNN is the deep neural network for classification. If the softmax layer in the rear is replaced with regression layer, DNN can work as the regression subnetwork and can accomplish the object detection task when combined with the classification subnetwork. The operation schematic diagram of DNN regression networks is shown in Fig. 1.

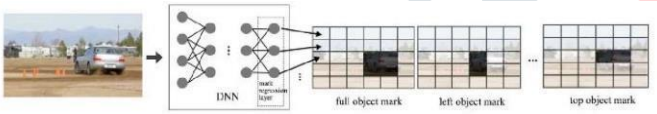


Figure 1. The operation schematic diagram of DNN regression networks

Just as shown in Fig. 1, DNN makes a regression by the binarizated grid combination covering the object region to locate the object. To distinguish the two adjacent objects, five ground truths marks are simultaneously adopted to make the regression. Those specific grids are designed to cover different parts of target by general target mark, left target mark, right target mark and bottom target mark. Even so, the accuracy of DNN is not satisfactory, and the mean Average Precision (mAP) in the VCO2007 test dataset is slightly over

can be briefly divided into: (1) the model based on region proposal; (2) the model based on regression.

A. Models based on region proposal

The deep learning object detection based on region proposal includes two main works: one is the extraction of region candidates; the other is the building of deep neural networks.

1) R-CNN

R-CNN is the convolution neural network based on the region proposal brought up in 2014 by Girshick who came up with the concept of region proposal for the first time. The accomplishing the object detection. The process diagram of Overfeat in object detection is shown in Fig. 2.

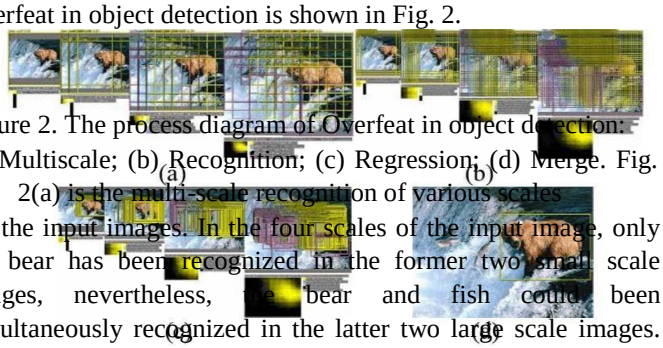


Figure 2. The process diagram of Overfeat in object detection: (a) Multiscale; (b) Recognition; (c) Regression; (d) Merge. Fig. 2(a) is the multi-scale recognition of various scales for the input images. In the four scales of the input image, only the bear has been recognized in the former two small scale images, nevertheless, the bear and fish could been simultaneously recognized in the latter two large scale images. Fig. 2(b) is the identification process, which gains more predictable results to increase the recognition precision by using offset operation and slide window operation; Fig. 2(c) is the regression process, which gains numerous object region proposals to enhance the location accuracy in the same way. Fig. 2(d) is the detection result after scale integration. However, the method of adopting offset operation and slide window operation is enormous computation in the Fig. 2(b) and Fig. 2(c). In terms of accuracy, the mAP in the test dataset of ILSVRC13 detected by Overfeat is 19.4%. Actually, most of the early object detection models based on deep learning employ the sliding window operation adopted by Overfeat to obtain the object candidates, and this blind and exhaustive method would result in the problem of data explosion. Later models try to solve the issue by improving the existing methods or proposing new ideas. Meanwhile, the early model design is not perfect and the model accuracy is unsatisfied. These disadvantages all chart a course for the succeeding model design of object detection based on deep learning.

IV. DEVELOPMENT OF OBJECT DETECTION BASED ON DEEP LEARNING

In recent years, with the development of deep learning in object detection, a massive deep detection models are proposed. Here, seven current mainstream deep learning models in object detection will be introduced and deeply analyzed according to the time order of emergency, and they

principle of R-CNN is that it utilizes the region segmentation method of selective search to extract the region proposals in the image, which include the possible object candidates, and loads them into convolution neural network to extract the feature vectors. Later, the classifier SVM will be used to classify the feature vectors to obtain the classification results in each region proposal. After merging by non-maximal suppression (NMS), the model outputs the precise object classifications and object bounding boxes to achieve object detection. The detailed process is shown in Fig. 3.

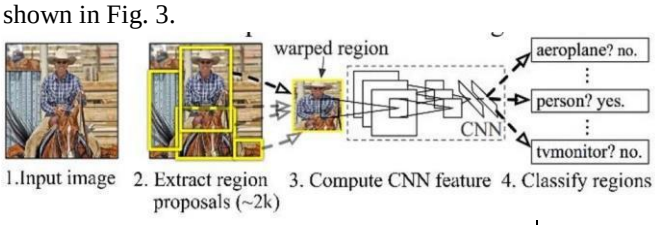


figure 3 ́The description of R-CNN framework

In the VOC2007 test dataset, the mAP of R-CNN object detection reaches 58.5% that is considerably lifted up compared with the former methods. Nevertheless, all the 2000 region proposals of R-CNN would pass through the convolution neural network in turn, resulting in that the realtime is poor. Even handling only one image on GPU needs tens seconds. Meanwhile, the data of the computation is great, for examples, the feature files of 5000 images producing from the convolution operation need to be stored on the hardware, and occupy hundreds of gigabyte storage space.

2) SPP-net

SPP-net is a deep neural network based on the spatial pyramid pooling proposed by MSRA He in 2014. The spatial pyramid pooling layer can get rid of the crop/warp operation on the input image in the former method. And it enables the input images of different sizes to connect with the full connection layer with the feature vector of the same dimension after passing the convolution layer. The crop/warp operation reshapes the sizes of input convolution neural network to the fixed size, which will lead to incompleteness of object image and object deformation, just as shown in Fig. 4.

Figure 4. Crop/warp operation of images: (a) Crop operation; (b) Warp operation.

Although SPP-net solves the problems of object image incompleteness and object deformation, it is still of colossal computation and poor real-time because its image processing is similar to that of R-CNN.

3) Fast R-CNN

Fast R-CNN is the upgrade of R-CNN proposed by Girshick and has the capability to solve the repetitive calculation problem of the 2000 region proposals passing through the convolution neural network in turn. The improvement of Fast R-CNN compared to R-CNN lies in that it maps the region proposal extracted by selective search algorithm in input image to the feature layer of convolution neural network and conducts the pooling on the mapped region proposal of feature layer by ROI pooling. The ROI pooling can help Fast R-CNN obtain the feature vector of fixed sizes, which is necessary to successfully connect with the full connection. The role of ROI pooling is just like the spatial pyramid pooling of SPP-net. The operation process of Fast R-CNN is as shown in Fig. 5.

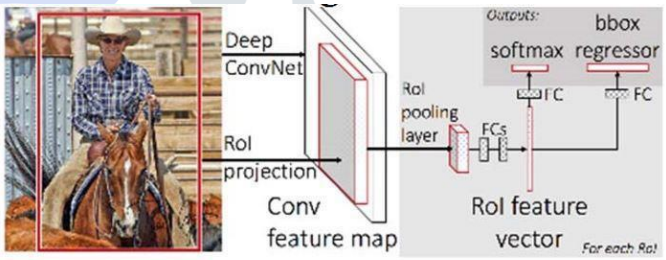


Figure 5. The operation process of Fast R-CNN

The method of mapping region proposal of input image to the feature layer in Fast R-CNN shares the convolution computation, which substantially reduces the calculation. In addition, in order to decrease the parameters of full connection, Fast R-CNN adopts truncated SVD to enable that the single fully connected layer corresponding to weight matrix is replaced by two small fully connected layers, which further lessens the network calculation. In the training stage, the speed of Fast R-CNN is 8.8 times that of R-CNN and 2.58 times that of SPP-net. In the test stage, the speed of Fast RCNN is 146 times that of R-CNN without the truncated SVD and 213 times that of R-CNN with the truncated SVD. When compared with SPP-net, the test speed of Fast R-CNN is 7 times that of the former without the truncated SVD and 10 times that of the former with truncated SVD.

4) Faster R-CNN

Faster R-CNN , proposed by Ren, He, Girshick, et al., is the upgrade version of the Fast-CNN. Faster R-CNN employs the region proposal network (RPN) to solve the issues of huge computation and poor real-time caused by the selective search method in R-CNN and Fast R-CNN. And Faster R-CNN is an end to end framework which can train the model easier. The function of RPN in Faster R-CNN is to replace the role of selective search in obtaining region proposals. RPN could divide the feature layer into n×n regions and obtain the feature regions of various scales and aspect ratios that are centered on the region, and the method is called anchors mechanism. The anchors in RPN are used to produce object proposals and then the proposals are sent to the rear classification and regression networks for the object recognition and location. The operation principle is shown in Fig. 6

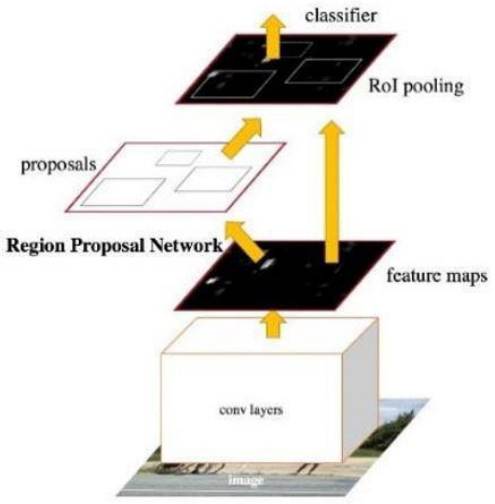


Figure 6. The operation principle of Faster R-CNN

After Faster RCNN adopts the RPN, the region proposals are reduced from 2000(by selective search) to 300, which significantly decreases the computation of the whole neural network. The experiment indicates that test speed of Faster RCNN reaches 5fp/s, 10 times that of Fast RCNN. Furthermore, the accuracy is also improved. The mAPs of Faster R-CNN in VOC2007 and VOC2012 dataset tests have been raised by 2% to 3% to reach 69.9% and 67.0% respectively compared with those of Fast R-CNN.

5) R-FCN

R-FCN, proposed by Dai, is a full convolution neural network based on regions, having solved the problem that RoI can't share the computation. The object detection framework of R-FCN also adopts RPN to generate candidate RoIs. With the position-sensitive score maps (khkh(C+1) dimensional convolution layer), R-FCN can record the response of every object in different locations. R-RCN defines the feature vector (C+1 dimension column vector) by voting according the RoIs and adopts softmax classification to classify the feature vectors in order to achieve the object recognition. The operation principle is shown in Fig. 7.

4

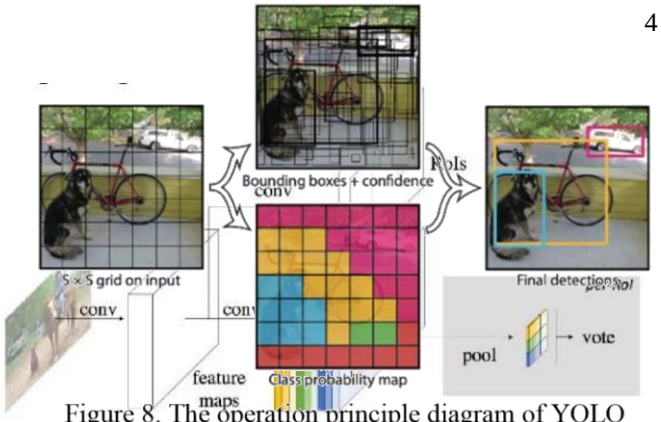


Figure 8. The operation principle diagram of YOLO

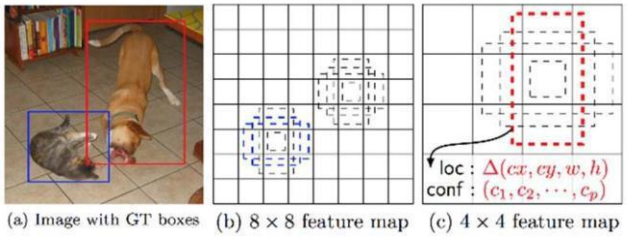
Figure 7. The operation principle of R-FCN

Moreover, the object location could be realized by appending a 4hkhk dimension convolutional layer to the above position-sensitive score maps and defining the feature vector (4 dimension column vector that represents the coordinates and width-height (tx, ty, tw, th) in the RoI region) by voting according to the RoIs. Therefore, R-FCN could do the recognition and location simultaneously to achieve object detection. As to the performance, the accuracy of R-FCN is similar to that of Faster R-CNN, but the test speed of R-FCN is 2.5 times that of the Faster R-CNN.

As to the object detection performance, YOLO's detection speed is 45 fps, and fast YOLO's can reach 155 fps to make the real-time detection possible, however, the YOLO's detection accuracy has decreased in a certain extent compared with other state-of-the-art object detection models of deep learning. In the same condition where the VGG is used as the convolution networks for feature extraction, the accuracy of YOLO is 66.4%, while the accuracy of Faster R-CNN is 73.2% [31]. The primary reason of the YOLO's accuracy decline is the cancel of region proposal.

2) SSD -

SSD is the single shot multi-box detector proposed by Liu Wei. The design of SSD has integrated YOLO's regression idea and Faster R-CNN's anchors mechanism. With the regression idea of YOLO, SSD simplifies the computation complexity of the neural network to guarantee the real-time. With the anchors mechanism, SSD can extract the features of different scales and aspect ratios to guarantee the detection accuracy. And the local feature extraction method of SSD is more reasonable and effective compared with the general feature extraction method of YOLO. What's more, because the feature representations in different scales are different, the method of multi-scale feature extraction has been applied in SSD, which contributes to promoting the detection robustness of different-scale objects. The operation principle is shown in Fig. 9



(a) Image with GT boxes (b) 8 × 8 feature map (c) 4 × 4 feature map

B. Models based on regression

At present, the object detection methods based on deep learning using region proposal gets satisfactory achievements, but the object detection is still of poor real-time that can't satisfy the application requirement.

1] YOLO

YOLO, came up with by Redmon, Divvala, Girshick, et al., is a convolution neural network for real-time object detection and can accomplish end to end training. Because of the cancel of RoI module, YOLO won't extract the object region proposal any more. The front end of YOLO connects a convolution neural network for feature extraction and the rear end connects two full connected layers for classification and regression in the grid regions. YOLO divides the input image scale into 7*7 grids, each of which will produce two bounding boxes. The bounding box will output a 4-dimnesional vector of coordinate information and the object confidence. Meanwhile, each grid also outputs 20 category probabilities, thus each grid produces a 30-dimentional vector including recognition information and location information. During the detection, YOLO filters the object proposals with low confidence by setting the threshold and wipes off the redundant object proposals to gain the detection results. The operation process is shown in Fig. 8.

In terms of object detection performance, the detection speed of SSD is 59 fps and its accuracy is 74.3%, which is the first object detection architecture having an accuracy rate more than 70% and satisfying the real-time requirement. However, it still has some disadvantages, one of which is its weak detection capacity to the small objects.

3] YOLOv3

Most classifiers assume output labels are mutually exclusive. It is true if the output are mutually exclusive object classes. Therefore, YOLO applies a softmax function to convert scores into probabilities that sum up to one. YOLOv3 uses multi-label classification. For example, the output labels may be "pedestrian" and "child" which are not non-exclusive. (the sum of output can be greater than 1 now.) YOLOv3 replaces the softmax function with independent logistic classifiers to calculate the likeliness of the input belongs to a specific label. Instead of using mean square error in calculating the classification loss, YOLOv3 uses binary cross-entropy loss for each label. This also reduces the computation complexity by avoiding the softmax function.

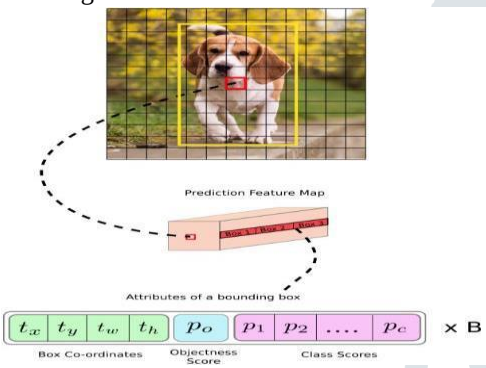


Fig10 : YOLOv3 operation principle

V. CHALLENGE AND SOLUTIONS

the poor realtime, there are two approaches to reduce the computation of network parameters as follows: (1) design the network architecture; (2) ameliorate the model algorithm. For example, Overfeat and R-FCN, they both take the measure of shared parameters to decrease the times of data passing the convolution neural network. Faster R-CNN applies the RPN structure to obtain region proposal by learning instead of the selective search algorithm. PVANET decreases the computation by improving the method of feature extraction. The paper in adopts a tiny network as the front end to preprocess the data and then send it to the object detection network to decrease the unnecessary computation.

B. The Robustness

The object detection based on deep learning is good at robustness compared with traditional detection. However, the deep models is hard to train. During the design of deep learning models, there are various tricks to improve the detection abilities and model robustness. To prevent the overfitting of deep model, Hinton, by putting forward the concept of Dropout, randomly controls the weight switch of neural network to enhance its generalization ability. Revising the activation function of neuron nodes, such as Relu and Maxout

With the further advancement of deep learning technology, the object detection model based on deep learning has been continuously ameliorated. Currently, the object recognition is of better and more active prospect than object detection in deep learning. Whereas, it is observed that object detection based on deep learning lays huge dependency on object recognition, thus the elevation of object recognition capability will promote the lift-up of object detection capability. Now various kinds of visual recognition and detection contests are held to further boost the advancement of deep learning in object recognition and object detection. Meanwhile, the ever-changing hardware update also drives the applications of deep learning technology. According to the overall analysis of current object detection methodologies based on deep learning, it can be discovered that there are two primary challenges which are the real-time and robustness of the computation. In the future, the object detection will certainly develop in better real-time and robustness.

A. The Real-time

Deep learning has been applied in object detection due to its powerful feature representation. Whereas, because of great amounts of network parameters and huge computation algorithm, most of the deep learning models are poor real-time in terms of the existing computation capability. Therefore, the real-time has become the bottleneck for the application of deep learning in object detection. To improve could increase the fitting ability of whole network. To adapt to the object detection of various scales, Hypernet has made use of the multi-level feature fusion to combine the feature of diverse resolution, which is adopted by SSD as well. In terms of training strategies, the hard negative mining is conducive to the precision by increasing the negative sample proportion, which reinforces the detection capability of the hard samples. The consideration of context link when the networks are designed can also help to increase the model accuracy. As to the details is little in high-level feature layer, the DSSD has added a deconvolution module on the basis of SSD, lifting up its recognition ability to the small objects. YOLO V2 the upgrading version of YOLO, applies the anchor mechanism in extracting the object candidates to improve the accuracy in YOLO.

VI. CONCLUSION

This paper firstly introduces the classical methodologies of object detection, discusses the relation and differences between the classic methodologies and the deep learning methodologies in object detection. Then it clarifies the ideas of model design and the limitations of deep learning method by overviewing the early object detection methods based on deep learning. Afterwards, it elaborates on the common object detection model based on deep learning, during whose process, it makes a

detailed interpretation of the framework design and operation principle of the model and points out its innovation and performance assessment. Finally, this paper makes a further analysis of the challenges in object detection based on deep learning, and offers some solutions for reference. With the innovation of deep learning theories and computer hardware upgrading, the performance of object detection based on deep learning will be ceaselessly enhanced and the applications of it will be widely ranged. Spatially, the development and application of current embedded systems in deep learning will pave a promising prospect for object detection based on deep learning.

REFERENCES

[1] D. Erhan, C. Szegedy, A. Toshev, et al, "Scalable object detection using deep neural networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2155-2162.

[2] A. Borji, M. M. Cheng, H. Jiang, et al, "Salient object detection: A benchmark," IEEE Transactions on Image Processing, vol. 24, Dec 2015, pp. 5706-5722.

[3] Y. Tian, P. Luo, X. Wang, et al, "Deep learning strong parts for pedestrian detection," 2015 IEEE International Conference on Computer Vision, 2015, pp. 1904-1912.

[4] P. Ahmadvand, R. Ebrahimpour and P. Ahmadvand, "How popular CNNs perform in real applications of face recognition,2016 24th Telecommunications Forum (TELFOR), 2016, pp.1-4.

[5] W. Ouyang, X. Wang, X. Zeng, et al, "Deepid-net: Deformable deep convolutional neural networks for object detection," 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2403-2412

[6] P. M. Merlin, D. J. Farber, "A parallel mechanism for detecting curves in pictures," IEEE Transactions on Computers, vol. C-24, Jan 1975, pp. 96-98.

[7] N. Singla, "Motion detection based on frame difference method," International Journal of Information & Computation Technology, vol. 4, no. 15, 2014, pp. 1559-156

[8] D. S. Lee, "Effective Gaussian mixture learning for video background subtraction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, May 2005, pp. 827-832

[9] B. K. P. Horn, B. G. Schunck, "Determining optical flow," Artificial intelligence, vol. 17, 1981, pp. 185-20

[10] J. L. Barron, D. J. Fleet, S. S. Beauchemin, et al, "Performance of optical flow techniques," 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992: pp. 236-242.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, pp. I-511-I-518.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, et al, "Object detection with discriminatively trained part-based models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, 2010, pp. 1627-1645.

[13] P. Felzenszwalb, D. McAllester, D. Ramanan, "A discriminatively trained, multiscale, deformable part model," 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8.

[14] M. Ulrich, C. Steger, A. Baumgartne, "Real-time object recognition using a modified generalized Hough transform," Pattern Recognition, vol. 36, Nov. 2003, pp. 2557-2570.

[15] J. Xu, X. Sun, D. Zhang, et al, "Automatic detection of inshore ships in high-resolution remote sensing images using robust invariant generalized Hough transform," IEEE Geoscience and Remote Sensing Letters, vol. 11, Dec. 2014, pp. 2070-2074.

[16] B. Leibe, A. Leonardis, B. Schiele, "Robust object detection with interleaved categorization and segmentation," International Journal of Computer Vision, vol. 77, 2008, pp. 259-289.

[17] A. Shrivastava, A. Gupta, R. Girshick, "Training region-based object detectors with online hard example mining," 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 761-769.

[18] K. H. Kim, S. Hong, B. Roh, et al, "PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection," arXiv preprint arXiv:1608.08021, 2016. [

[19] K. Dai, G. Li, D. Tu, et al, "Prospects and current studies on background subtraction techniques for moving objects detection from surveillance video," Journal of Image and Graphics, vol. 11, July 2006, pp. 919-927

[20] Q. Ji, S. Yu, "Object detection algorithm based on surendra background subtraction and four-frame difference," Computer Applications and Software, vol. 31, Dec. 2014, pp. 242-244.

[21] C. Stauffer, W. E. L. Grimson. "Adaptive background mixture models for real-time tracking," 1999 IEEE Computer Society Conference on
7

[22] R. Girshick, J. Donahue, T. Darrell, et al, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014, pp. 580-587.

[23] Y. Li, K. He, J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," Advances in Neural Information Processing Systems, 2016, pp. 379-387

Computer Vision and Pattern Recognition, 1999. pp. 246-252.