Research Article

Implementation of Deep Learning-based Object Recognition and Tracking for Intelligent Video Surveillance

Han-Jin Cho[1,]

[1]Department of Energy IT, Far East University, Eumseong-gun, Chungcheongbuk-do, 27601, Republic of Korea
*Corresponding author; Email: hanjincho@kdu.ac.kr

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: As research on artificial intelligence technology is actively conducted in recent years, research on deep learning technology that recognizes and classifies images in real time on behalf of humans is required. Object recognition has difficulties in finding an object of interest from a video clip or image, and classifying several detected objects for each object. To solve this problem, research is needed to detect and track objects using a CNN-based deep learning method. Among the CNN-based multi objects detection techniques, the most well-known methods are R-CNN and Faster R-CNN. These methods are based on ROI-based detection techniques to perform verification work to reduce candidate groups through pre-work in the ROI of the object. However, since the process of classifying objects for each region of interest is performed, the detector speed decreases. Real-time processing is not possible due to this speed problem. In this paper, to overcome these issues, we have proposed multi object detection, classification, and tracking method using YOLO, a single step technique that performs a single CNN to determine the location and type of objects in an image. Experimental results depict that it can detect and classify objects robustly in various environments, and that real-time tracking is possible because the calculation speed is faster than the conventional method.

Keywords: Deep Learning, Object Recognition, Object Tracking, Surveillance, Object Classification

1. Introduction

Recently, as artificial intelligence technology has become a hot topic and development, it has been applied in various fields. Like Google's AlphaGo, it is used for home appliances that play themselves against people and operate according to the situation at home[1]. In addition, famous companies such as Facebook, Google, and Samsung have created AI research teams and are conducting research professionally. In addition, fields using autonomous vehicles and drones using technologies such as computer vision are becoming more active, and the use of applied technologies is increasing[1,2]. CCTV, which provides image information for various purposes, is becoming intelligent, and the range of automation applications is increasing. In addition, many researches are being conducted in computer vision such as intelligent cars and video security for robust object detection and tracking. [3,4]. In research for detecting objects in images, identifying and tracking objects, a highly reliable detection and tracking method must be performed to accurately recognize objects, and a lot of methods are being researched for these[4,5].

Object detection is a field of research that allows a computer to identify and analyze objects on behalf of visual information that humans receive from images or images[5,6]. Object recognition has difficulty in finding an object of interest from a video clip or image and classifying several detected objects for each object. To overcome these issues, research is being conducted to detect and track objects using a CNN-based deep learning method[7,8]. Among the CNN-based multi objects detection techniques, the most well-known methods are R-CNN and Faster R-CNN. These methods are based on ROI-based detection techniques to perform verification work to reduce candidate groups through line work in the ROI of the object. [9-11]. However, it cannot use in real time because the detector speed is low because it requires the process of classifying objects by area of interest. In this paper, to overcome these issues, we have proposed multi object detection, classification and tracking method based on YOLO, a single-stage technique that determines the location and type of an object in an image by carrying out a single CNN.

Chapter 2 explains the introduction and structure of deep learning as a related study, and chapter 3 represents how to detect and track objects using YOLO, a single-stage technique. Chapter 4 explains the implementation results of our proposed process, and chapter 5 represents the conclusions and future connections.

2. Experimental Details

Methods of detection, classification and tracking objects have been researched continuously. A representative method for detecting an object is to detect an object by extracting a feature from the object and generating a feature point descriptor, and there are techniques such as SIFT(Scale Invariant Feature

Transform) and SURF(Speed Up Robust Features)[12,13]. However, the method based on the feature point has a disadvantage of slowing down the processing speed due to the operation on a high dimension, so real-time processing is not possible. To track an object, an object is tracked based on the density distribution such as feature points, corners, and colors of the data, such as MeanShift and CAMShift(Continuously Adaptive Mean Shift)[14,15]. However, it is not robust against various illumination environments and noise, and it cannot use in real-time due to a complex amount of calculation[9-11]. As a method to run out these issues, techniques of detecting and tracking objects using deep learning such as CNN(Convolutional Neural Networks), RNN(Region Based Convolutional Neural Networks), and YOLO(You Only Look Once) based on convolution operations are being actively studied.

2.1 Deep Learning

Deep learning is an algorithm that attempts to learn at multiple levels using artificial neural networks in machine learning. Artificial neural network represent the entire model with problem-solving ability by changing the strength of synaptic bonds through learning of artificial neurons that form a network through synaptic bonds. [9, 19,20]. Artificial neural networks have several problems. As the layer structure deepens, it can stay at the local minimum, model dissipation, overfitting, slow learning time, etc. depending on the initial setting. Deep learning algorithms have been developed to solve the weakness of these artificial neural networks.[20,21,22,23]. With the recent advancement in hardware, the time required for complex matrix calculations used in deep learning using a powerful GPU has been greatly reduced, and deep learning is drawing attention as it can be used for learning by synthesizing and analyzing a large amount of data. Recently, CNN-based deep learning algorithms have shown good results in research areas of computer vision and natural language recognition.

2.2 CNN(Convolutional Neural Network)

CNN can be utillized to create filters to extract certain features from images or video clips. That is, if 3 by 3 or more windows or masks are performed repeatedly throughout the image, depending on the count value of the mask, you can get an appropriate results[20,21]. Perform a conversion of the window area on the entire image and perform an operation on the mask area corresponding to the kernel of the conversion to obtain the result. The result can be acquired by carrying out the operation repeatedly while moving the window.

Neural network is a mathematical model constructed to handle large numbers organized into layers. CNN can be represented of two stages: feature extraction and classification. Feature extension is a step to find unique characteristics of input data, and when learning with only one original image, it is advantageous to learn with a variety of feature maps, so it is useful to find unique features. Classification is the process of choosing classes with the characteristics found. The CNN processing process does not consist simply of classifiers, but involves the extraction of features, enabling direct operation of raw images, and unlike conventional algorithms, no separate pretreatment steps are required. Filter and sub-sampling can be performed repeatedly to create feature extraction and topology invariance, which allows global access from the local feature.

3. Implementation

When a person looks at an image through vision, he or she immediately identifies the object information inside the image. R-CNN and others with complex processing processes have some parts that lack to emulate human visual systems due to their processing speed and difficulties in optimizing them. Object detection techniques based on CNN, such as R-CNN and Faster R-CNN, perform verification to decrease candidates for multi object's area of interest through pre-work with a detection technique based on the area of interest. However, the detector speed is reduced because the process of classifying objects in each area of interest is required. To overcome these issues, this paper embodied the method of detecting and tracking objects using YOLO, a single-stage technique that determines the location and type of objects in the images with one CNN performance.

The YOLO considers the probability of the Binding Box within the image and the type of object to be classified as Single Rejection Problem, estimating the class and bunding box of the object as 'one-time viewing' of the image. This method calculates the class probability for several bounding boxes through the single convolution network. The basic operation principle is expressed in Table 1.

Table 1. Operation of our thechnique

1. Convert input image to 448×448 and divide into S×S grid cells.
2. Get a tensor in the form of S×S+ (5)B+C) by passing through the single solution network.
3. Cut the class probability from the model into threads and print the result value.

YOLO uses S×S grid cells. Using each of these grid cells, it creates B bounding boxes and confidence scores for each bounding box, and generates C conditional class probability. If there is no object in the cell, the confidence score is 0. Expressed as an equation, it is the same as equation (1).

$$Pr(Obj) * IOU_{pred}^{truth} \tag{1}$$

Each grid predicts conditional probabilities based on five predicted values. This is an indicator of classification performance, indicating the probability that an object is a class that you want to classify when it exists in a grid. Thus, the value of 5 multiplied by B represents five counts of information, which means the relative coordinates (x, y) of the object for the entire image and the relative horizontal, vertical length, and probability values of that object. The class probability of the bonding box can be expressed by multiplying Equation 1 representing the confidence score. Through this probability, it can be used as an index to evaluate how well the bounding box predicted the position and size of an object, and this can be expressed as an equation (2).

$$Pr(Class_i|Obj) * Pr(Obj) * IOU_{pred}^{truth} = Pr(Class_i) * IOU_{pred}^{truth} \tag{2}$$

The metrics obtained from this are used to display the area of an object within the image and to classify which class the object belongs to. When the probability map of the class is output for S×S grid cells, calculate and draw the Binding Box. Object detection and tracking models use loss functions to find optimal weighted parameters through learning. The loss function can be represented by localization loss, constancy loss, and classification loss. The overall loss function can be calculated from each loss function. This function can be represented as shown in expression (3) and can find the presence of objects in each grid, classification, size, and location of objects for object detection and tracking.

$$\begin{gathered}
\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[(x_i - \hat{x_\iota})^2 + (y_i - \hat{y_\iota})^2 \right] \\
+ \lambda_{coord} \textstyle\sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \displaystyle\left[\left(\sqrt{\omega_i} - \sqrt{\widehat{\omega_\iota}}\right)^2 + \left(\sqrt{h_i} - \sqrt{\widehat{h_\iota}}\right)^2 \right] \\
+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left(C_i - \widehat{C_\iota}\right)^2 + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{noobj} \left(C_i - \widehat{C_\iota}\right)^2 \\
+ \sum_{i=0}^{S^2} I_{ij}^{obj} \sum_{c \in classes} \left(p_i(c) - \widehat{p_\iota(c)}\right)^2
\end{gathered} \tag{3}$$

4. Experimental Results

In this paper, multi object detecting, classifying, and tracking using deep learning technique were implemented with Python in 64bit Windows 10 OS environment equipped with GPU. The dataset used in the experiment was studied using VOC Pascal Data. In addition, the evaluation of the methods implemented in the paper used the Compute matrix as depicted in Table 2.

Table 2. Confusion matrix to evaluate the experiment

		Ground Truth	
		True	False
Predicated Result	True	True Positive	False Positive
	False	False Negative	True Negative

Performance metrics utilize AP(Average Precision) with precision and recall measurements widely used in object recognition problems. Expression (4) is the expression that calculates the average precision. Precision refers to the correct perceived proportion of all recognition results, and false positives (FP) in object recognition problems are when the boundary area of the object is too different from the actual boundary area or the predicted class is different. Reproducibility refers to the proportion of objects to be recognized properly, and False Negative (FN) refers to when the actual objects are not recognized or when the Confidence Score is very low.

$$Average\ Precision = \sum_n (R_n - R_{n-1}) \times P_n \qquad (4)$$

Figure 1 shows the dataset used in experiments to detect and classify objects in various environments and evaluate traces.

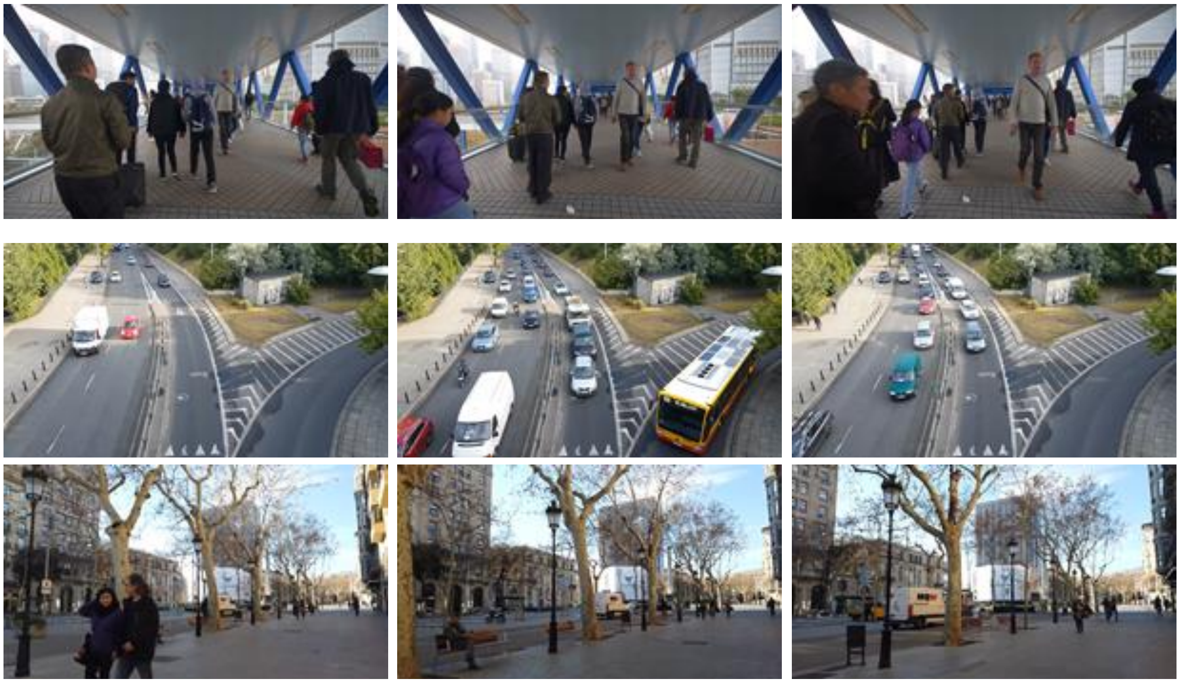

Figure 1. Datasets using our experiments for object detection, classification, and tracking

The experiment results in this paper represent that the average precision of object detection, classification and tracking model shows 95.5%, showing good results compared to other methods used in existing research. Table 3 shows the average process results compared to previous researches. Figure 2 also shows the results of our technique for detecting, classifying, and tracking objects in datasets used in experiments.

Table 3. Results of our technique in datasets

Method	Average Precision
R-CNN	90.7
Faster R-CNN	92.1
Our Method	95.4

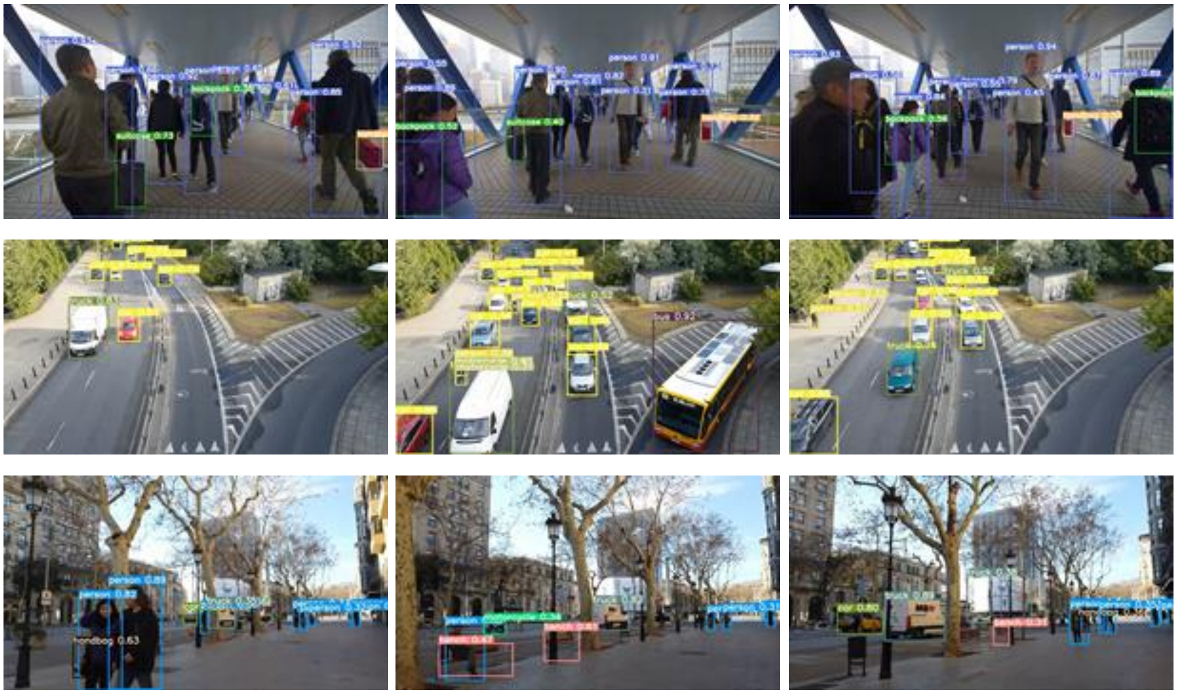

Figure 2. Results of our experiments for detecting, classifying, and tracking objects

5. Conclusion

Artificial intelligence technology has recently become a hot topic and has been applied in various fields. CCTV that provide video information for various purposes are becoming intelligent, and the scope of automated applications is increasing. Many studies are being conducted in computer vision such as intelligent cars and video security for robust object detecting, classifying, and tracking. In the research of detecting and identifying and tracking objects in images, reliable detection and tracking methods should be performed to accurately recognize them, and real-time processing should also be possible for high utilization. To address these issues , researches are being conducted using CNN-based deep learning methods to detect and track objects, however R-CNN or Faster R-CNN is required to classify multi objects in different areas of interest, making it arduous to use them in real time. To overcome these issues, in the paper, we have proposed a technique of detecting, classifying, and tracking objects using YOLO, a single-stage technique that determines the location and type of objects in the images with one CNN performance. The experimental results represent that objects can be detected and classified in a robust manner in various environments, and real-time tracking is possible. Future tasks will require research that can recognize and track cars on roads in real time and use them for intelligent traffic information systems that can provide traffic conditions by measuring speeds.

Acknowledgments

This work was supported by the 2020 Far East University Research Grant(FEU2020S03).

References

1. Valera, M., & Velastin, S. A. (2005). Intelligent distributed surveillance systems: a review. IEE Proceedings-Vision, Image and Signal Processing, 152(2), pp.192-204.
2. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.
3. Collins, R. T., Lipton, A. J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y. & Wixson, L. (2000). A system for video surveillance and monitoring. VSAM final report, 2000, pp.1-68.
4. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097-1105.
5. Joshi, K. A., & Thakore, D. G. (2012). A survey on moving object detection and tracking in video surveillance system. International Journal of Soft Computing and Engineering, 2(3), pp.44-48.
6. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems, pp. 568-576.
7. Liu, T., Fang, S., Zhao, Y., Wang, P., & Zhang, J. (2015). Implementation of training convolutional neural networks. arXiv preprint arXiv:1506.01195.
8. Chen, Y., Yang, X., Zhong, B., Pan, S., Chen, D., & Zhang, H. (2016). CNNTracker: Online discriminative object tracking via deep convolutional neural network. Applied Soft Computing, 38, pp.1088-1098.
9. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence, 38(1), pp.142-158.
10. Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448.
11. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pp. 91-99.
12. Bay, H., Tuytelaars, T., & Van Gool, L. (2006, May). Surf: Speeded up robust features. In European conference on computer vision, Springer, Berlin, Heidelberg, pp. 404-417.
13. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), pp.91-110.
14. Ahn, H., & Lee, Y. H. (2016). Performance analysis of object recognition and tracking for the use of surveillance system. Journal of Ambient Intelligence and Humanized Computing, 7(5), pp. 673-679.
15. Allen, J. G., Xu, R. Y., & Jin, J. S. (2004, June). Object tracking using camshift algorithm and multiple quantized feature spaces. In ACM International Conference Proceeding Series (Vol. 100), pp. 3-7.
16. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision. Springer, Cham, pp. 21-37.
17. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788.
18. Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271.
19. Ahn, H., & Cho, H. J. (2019). Research of multi-object detection and tracking using machine learning based on knowledge for video surveillance system. Personal and Ubiquitous Computing, pp.1-10.
20. Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: A benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2411-2418.
21. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9), pp.1904-1916.
22. Mallick, P.K., Satapathy, B.S., Mohanty, M.N. and Kumar, S.S., 2015, February. Intelligent technique for CT brain image segmentation. In 2015 2nd International Conference on Electronics and Communication Systems (ICECS) (pp. 1269-1277). IEEE.
23. Reddy, A.V.N., Krishna, C.P. & Mallick, P.K. An image classification framework exploring the

capabilities of extreme learning machines and artificial bee colony. Neural Comput & Applic 32, 3079–3099 (2020). https://doi.org/10.1007/s00521-019-04385-5