مجلة

كلية التـــراث الجامعة

مجلة علمية محكمة

متعددة التخصصات نصف سنوية

العدد الثامن والثلاثون

15 شباط 2024

ISSN 2074-5621

رقم الايداع في دار الكتب والوثائق 719 لسنة 2011

مجلة كلية التراث الجامعة معترف بها من قبل وزارة التعليم العالي والبحث العلمي بكتابها المرقم (ب 3059/4) والمؤرخ في (4/7 /2014)

A Survey on Real Time Object Recognition Methods In Deep Learning

Mohamed Riyad Saleh
Mustansiriyah University

Abdul-Wahab Sam
Mustansiriyah University

ABSTRACT

Object recognition is tough in image processing. The success of many computer vision applications depends on object recognition. For years, object recognition techniques and systems have been developed to solve this challenge. This article reviews computer vision and object recognition methods. In this survey we discuss challenges faced by the computer vision, Objectives of Research in this field ,Types of Architecture in CNN, point detector, frame differencing to detect objects. The article discusses object detecting strategies from several researchers.

Keywords: deep Learning , cnn , computer vision

الملخص

يعد التعرف على الكائنات أمرًا صعبًا في معالجة الصور. يعتمد نجاح العديد من تطبيقات الرؤية الحاسوبية على التعرف على الأشياء. لسنوات عديدة، تم تطوير تقنيات وأنظمة التعرف على الأشياء لحل هذا التحدي. تستعرض هذه المقالة رؤية الكمبيوتر وطرق التعرف على الأشياء. تقوم هذه المقالة بمراجعة وتحليل خوارزميات الكشف عن الكائنات. نناقش في هذا الاستطلاع التحديات التي تواجه الرؤية الحاسوبية، وأهداف البحث في هذا المجال، وأنواع الهندسة المعمارية في شبكة سي ان ان وكاشف النقاط، واختلاف الإطارات لكشف الأشياء. يناقش المقال استراتيجيات الكشف عن الأشياء من قبل العديد من الباحثين.

1. INTRODUCTION

In recent times, there has been a notable proliferation of surveillance systems on a global scale. Consequently, there has been a corresponding rise in the demand for intelligent systems to enhance and refine their functionalities, both in civil and military contexts. As a result, numerous detection and tracking technologies have been devised and disseminated to effectively identify and differentiate various objects, regardless of their state of motion. The significance of these systems resides in their intrinsic utility, encompassing the identification and monitoring of mobile entities in both civilian and military domains, surveillance of pedestrians or diverse objects, as well as the tracing of crucial data and information, human-computer interaction, virtual reality development, motion analysis, and numerous other applications[1]. The discipline of computer engineering, among others, encompasses a wide range of areas. Therefore, the significance of doing research in this particular domain cannot be overstated, given the multitude of studies, research endeavors, and specialist publications that have emerged. These have not only captivated many academics but have also spurred the development of sophisticated and contemporary algorithms and methodologies [1, 2].

The current identification and tracking systems use two fundamental technologies, namely:

1- One often used method for object detection and tracking is radar-based technology, sometimes referred to as signal processing.

2- One approach that utilizes image processing, often referred to as Computer Vision (CV) in contemporary literature, is used to identify and monitor the item in question.Computer Vision is a field that focuses on the extraction and analysis of valuable information from pictures or sequences of images. The proliferation of computer vision technology has facilitated the integration of complex applications into everyday devices, enabling functionalities that were previously unattainable. Examples of these applications include facial identification on mobile phones and laptops, as well as pedestrian and vehicle detection in autonomous vehicles. These technologies have not only facilitated the completion of everyday chores but have also enhanced security in both physical and cyber domains[3].

2. Computer vision technologies hurdles

1. The presence of variability in data and environments: Real-world images and videos exhibit significant variations in lighting conditions, viewpoints, weather patterns, and other factors. Computer vision systems must possess sufficient robustness to effectively manage the inherent variability present in the data.

2. The availability of annotated data is sometimes limited : when training computer vision models, since it requires big datasets with precise annotations. Acquiring datasets of this kind may prove to be a laborious and costly endeavor, particularly when catering to specialized applications.

3. Generalization: Although models may exhibit strong performance on the data they were trained on, the task of assuring their ability to generalize to unfamiliar data and a wide range of situations is a persistent problem.

4. Variations in lightning conditions may occur throughout the day. Additionally, the atmospheric conditions have the potential to impact the illumination of a picture. Indoor and outdoor photographs of the same item may exhibit different lighting conditions. The presence of shadows inside an image has the potential to impact the distribution and intensity of light within such picture. Regardless of the lighting conditions, it is essential for the system to possess the capability to accurately identify and classify objects inside any given picture.

5. Rotation : allows for the picture to be shown in a rotated format. The system must possess the capability to effectively manage such challenges[4]. As seen in Figure 1, the character "A" has the potential to manifest in several forms. However, it is essential that the orientation of the letter or picture does not impact the identification of the character "A" or any other item shown.

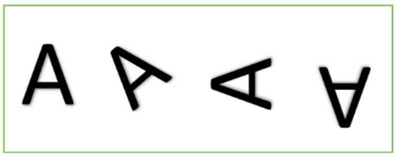


Figure 1 Rotation the character ‘’A’’

6. The object recognition system must be able to identify the mirrored image of any given item.

7. Scale invariance: The object recognition system should maintain its accuracy regardless of any variations in the size of the item. The aforementioned are a few of the challenges that might

potentially occur in the context of object recognition. The aforementioned challenges may be overcome in order to construct a highly effective and resilient object detecting system.

The task of object recognition is widely acknowledged as a challenging problem within the field of image processing. The identification of objects has significant importance within the field of computer vision, as it exhibits a strong correlation with the efficacy and accomplishment of several computer vision applications. Category recognition and detection are integral components of the broader field of object recognition. The primary goal of category recognition is to accurately assign an item to a certain category from a predetermined set of categories. The primary objective of detection is to differentiate and discern things from their surrounding context. There are a multitude of issues pertaining to the identification of objects. In general, the detection of objects requires the ability to discern them amongst backdrops that are crowded and noisy, as well as other objects, all of which may be subject to varying lighting and contrast conditions[4].

3. Literature Survey

Numerous literary pieces are associated with the suggested methodology, whereby a certain group has been chosen in the following manner:

- In the year 2018, J. Redmon et al. [5] introduced YOLO version 3 (YOLOv3) as a means of achieving improved detection accuracy, but at a lower level compared to Fast R-CNN. Nevertheless, it continues to exhibit rapid processing capabilities for all of these tasks. In the same year, an algorithm was proposed by T. Mahalingam and M. Subramoniam [6]. This algorithm consists of three stages: the Foreground segmentation phase, which utilizes the Mixture of Adaptive Gaussian model. the tracking step, which employs blob detection; and the evaluation stage, which involves classification based on feature extraction. In the same year, Andrew G. and colleagues [7] proposed a small and versatile real-time tracker that utilizes the front-backlight detector known as MobileNets. The technology exhibits compromises in terms of accuracy and demonstrates robust performance in scenarios involving significant scale and delay.

- In a study conducted in 2019, B. Shijila et al. [8] proposed the development of a novel algorithm that utilizes learning templates to improve performance in low-level, dispersed, and subtle regions. The results of their research indicate that this algorithm outperforms traditional approaches. The technique has been further enhanced to address both noise removal and object recognition. In their work, Ray and Chakraborty [9] (2019) introduced a method for representing moving objects as clusters of spatial and temporal points. This approach utilizes the Gabor 3D filter to perform spatial and temporal analysis on consecutive videos. The resulting clusters are then connected while using the Minimum Spanning Tree technique. In the year mentioned, B. Blanco-Filgueira et al. [10] put forth the use of an autonomous learning algorithm in drones for the purpose of object recognition and tracking in real time. This was achieved by using the integrated camera or a low-power computing system. The researchers integrate a scaled-down version of the Faster-RCNN algorithm with the Kernelized Correlation Filters (KCF) tracker in order to perform object tracking on an unmanned aerial vehicle (UAV). This article presents a detection technique that exhibits accuracy but is characterized by sluggish computational speed and high mathematical complexity. Although tracking algorithms exhibit high speed, they lack precision, particularly when dealing with fast-moving objects or abrupt changes in motion. In their study, Hossain et al. [11] introduced an application that utilizes a deep learning approach. This application was designed to be deployed on a computer system that is coupled with a drone, enabling real-time tracking of objects. The effectiveness of multi-rotor aircraft was proved to be satisfactory via experimental investigations. The object in question was assessed based on its comparable characteristics. However, it deviated from accurately tracking a previously identified feature and instead treated it as a distinct target. The efficacy of these algorithms may be compromised when confronted with certain obstacles encountered in real-time applications.

- In the year 2020, Zhao, H. and colleagues introduced a novel approach called Mixed YOLOv3-LITE, which is a computationally efficient real-time object identification framework suitable for deployment on both non-graphics processing units (GPUs) and mobile-devices [12]. The findings demonstrate that Mixed YOLO-v3-LITE has superior efficiency and performance on mobile terminals and other devices. In their study, R. Chen et al. [13] introduced a novel detection technique that utilizes Fast PCP and Motion Saliency as its foundation. It demonstrates superior accuracy outcomes while taking into consideration the computational expenses.
- In their 2021 study, Modwel et al[14] . introduced a hybrid approach that integrates three key algorithms in order to minimize the scanning workload for each frame. The method for recursive Density Estimation (RDE) determines the frame that requires scanning. The You Look at Once (YOLO) algorithm performs detection and identification inside the chosen frame. The items that have been detected are tracked using the Speed Up Robust Feature (SURF) method in order to track these things in consecutive frames.

4. CNN's Types of Architecture

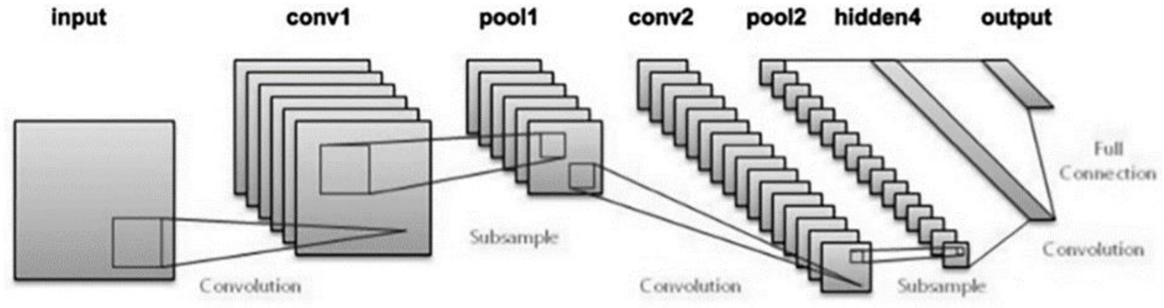
4.1 LeNet-5

The LeNet-5 it was created by LeCun et al. in 1998. It was the first technique used to categorize handwriting numerals and digits. the LeNet-5 design gets its name from the fact that it has five levels. Three of the five layers are convolutional layers, with pooling layers between each one. The two layers are the only ones that are entirely linked. Finally, the softmax classifier is utilised to divide the pictures into classes. This design is the most common since it is a very simple method [15].

Figure 2 the architecture of LeNet-5[15]

4.2 AlexNet

Alex Krizhevsky et al. created the AlexNet architecture in 2012. AlexNet's design is similar to LeNet's, however the depth of the network in AlexNet is greater. The AlexNet architecture is divided into eight levels. Five of them are convolutional layers with a maximum pooling



layer, while the remaining three are fully connected layers. Except for the output layer, the ReLU activation functions are introduced to each layer. Overfitting in the network may be prevented by using dropout layers[15].

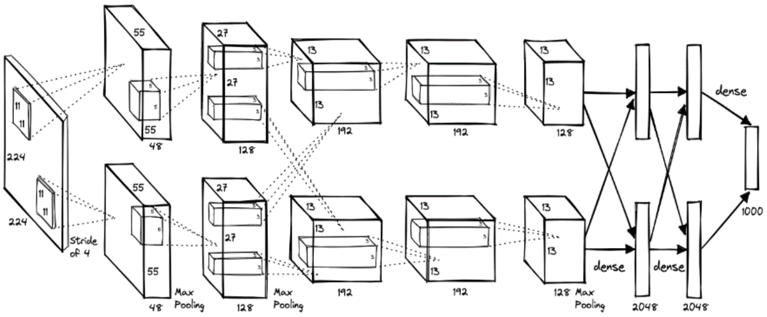


Figure 3 The AlexNet architecture diagram[15]

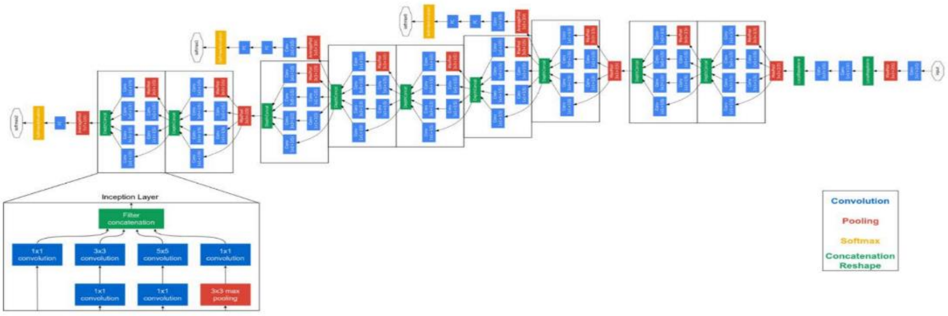
4.3 GoogLeNet/Inception

GoogLeNet's design varies from the previous architectures in that it employs 1*1 convolution and global average pooling to construct deeper networks. The number of parameters utilised in convolution is reduced, resulting in enhanced network depth. The use of the global average pooling improves classification accuracy. For regularisation, the fully connected layer with the ReLU activation function is utilised, as well as the dropout layer. The softmax classifier is useto categorise pictures or data. Figure 5 depicts the GoogLeNet block diagram[15]

Figure 4 the block diagram of GoogLeNet[15]

4.4 VGGNet

Simonyan and Zisserman created VGGNet in 2014. The VGGNet design contains 16 convolutional layers in overall. As with AlexNet, the number of filters is enhanced. The 3*3



filters are used to boost the network's depths. Next the pooling levels, the three completely linked levels are added[15].

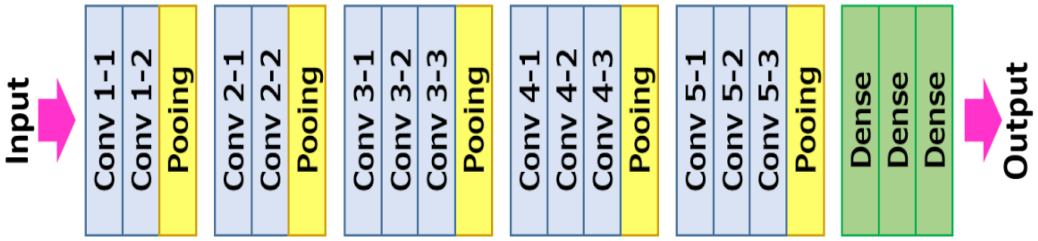


Figure 5 VGGNet architecture[15]

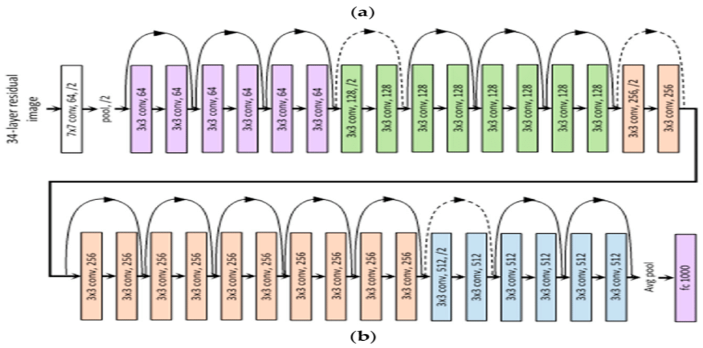
4. ResNet

Kaiming He et al. created the ResNet in 2015. ResNet is used to eliminate the disappearing gradient. The ResNet system employs the skipped connection approach. The skip connection works by skipping a few levels of training and connecting all the other layers to the result layer [15].

Figure 6 architecture of ResNet[15]

4.6 You Only Look Once (YOLO)

You-Only-Look-Once (YOLO) is a real-time object identification system that use neural networks. This algorithm is well-known for its speed and precision. It has been used to identify traffic lights, pedestrians, parking metres, and animals in a variety of applications. To identify objects in real-time, the YOLO method leverages CNN. To identify objects, the approach needs just one forward propagation through a neural network, as the name implies .This indicates that the complete picture is predicted in a one algorithm run. The CNN is used to predict several class probabilities and bounding boxes at the same time [16].the YOLO algorithm has many variations. Tiny YOLO and YOLOv3 are two popular ones.



YOLO architecture is similar to GoogleNet. As illustrated Figure 4, it has overall 24 convolutional layers, four max-pooling layers, and two fully connected layers.

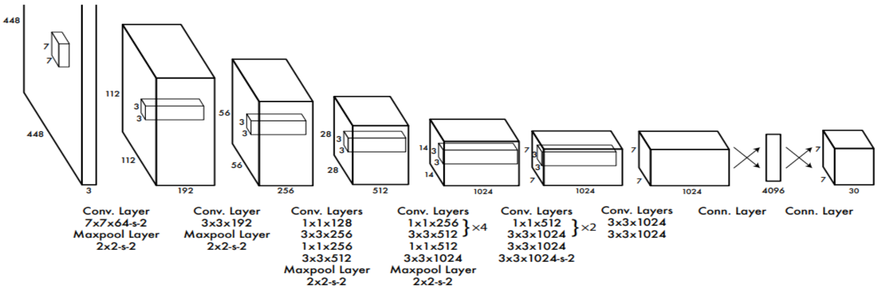


Figure 7 YOLO architecture[16]

5. Data Sets Used For Evaluation

Microsoft COCO stands for Microsoft Common Objects in COntext. It has 91 common item categories and 328,000 photos with a total of 2,500,000 occurrences. A precise pixel level segmentation provides the spatial position of each item. Furthermore, a key feature of this collection is the amount of labelled occurrences per picture. This might help in learning contextual knowledge[17].

CIFAR-10 and CIFAR-100 are two types of CIFARs. These subsets are produced from the Tiny Image Dataset, with the pictures more precisely tagged. The CIFAR-10 collection has 6000 samples of each of the ten classes, whereas the CIFAR-100 set contains 600 examples of each of the hundred classes. The resolution of each picture is 32*32.

OpenImages is a big dataset that includes 9 million training pictures, 41,620 validation samples, and 125,456 test samples. It has 9,600 trainable classes and is largely annotated. The Open Images Dataset V4: Scalable unified picture categorization, object recognition, and visual connection detection.

6. Metrics

Object detectors utilise various metrics to assess its performance, including frames per second (FPS), accuracy, and recall. The most popular assessment statistic, however, is mean Average Precision (mAP). Precision is calculated as the ratio of the area of overlap and the area of union between the ground truth and the projected bounding box (IoU). To assess whether the detection is accurate, a threshold is specified. If the IoU is more than the threshold, it is considered True Positive; otherwise, it is considered False Positive. False Negative[18] occurs when the model fails to identify an item in the ground truth. Precision is the proportion of right forecasts, while recall is the percentage of correct predictions in relation to the ground truth[19].

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

According to the above equation, average precision is computed separately for each class. To compare performance between the detectors, the mean of average precision of all classes, called mean average precision is used, which acts as a single metric for final evaluation [20].

7.CONCLUSION

8. Finally, the review emphasises major advances in object identification systems within the deep learning paradigm. The combination of novel architectures, large-scale datasets, and transfer learning has moved the area ahead, with real-world applications affecting a wide range of sectors. all the methods, structures, and previous studies presented are good at detecting objects, but we still have not reached the speed required for real-time detection , While obstacles remain, the future of deep learning object identification is hopeful, thanks to continued research and cooperation across fields.

REFERENCES

[1] B. Cyganek, "Object-detection and recognition in digital images : theory and practice", John Wiley & Sons, 2013.

[2] S. Karakaş, "Detecting and tracking moving objects with an active camera in real-time", Middle East Technical University, The Master thesis, 2011.

[3] Andrew Yan-Tak Ng, “Computer Vision - deeplearning.ai | Coursera.” [Online]. Available: https://www.coursera.org/learn/convolutional-neuralnetworks/lecture/Ob1nR/computer-vision. [Accessed: 04-Jun-2018].

[4] C. M. Sukanya, R. Gokul, and V. Paul, “A Survey on Object Recognition Methods,” Ijcset, vol. 6, no. 1, pp. 48–52, 2016.

[5] J. Redmon, and A. Farhadi, "Yolov3: An incremental improvement", 1804.02767, 2018.

[6] T. Mahalingam, and M. Subramoniam, "A robust single and multiple moving-object-detection, tracking and classification", Applied Computing and Informatics, 2018.

[7] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko W. Wang, T. Weyand, and H. Adam,"Mobilenets: Efficient convolutional neural networks for mobile vision applications" :1704, 2018.

[8] B. Shijila, J. T. Anju, and N. G. Sudhish , "Simultaneous denoising and moving object-detection using low rank approximation", Future Generation Computer Systems, Vol. 90, pp 198-210, 2019.

[9] K. S. Ray, and S. Chakraborty, "Object-detection by spatial temporal analysis and tracking of the detected-objects in a video with a variable background", Journal of Visual Communication and Image Representation, Vol. 58, pp. 662-674, 2019.

[10] B. Blanco-Filgueira, D. García-Lesta, M. Fernández-Sanjurjo, V. Brea, and P. López, "Deep learning-based multiple object visual tracking on embedded system for IoT and mobile edge computing applications", IEEE Internet of Things Journal, Vol. 6, No.3, pp. 5423-5431, 2019.

[11]S. Hossain, and D. J. Lee, "Deep learning-based real-time multiple-object detection and tracking from aerial imagery via a flying robot with GPUbased embedded devices", Sensors, Vol. 19, No.15, pp. 3371-3395, 2019

[12] H. Zhao, Y. Zhou, L. Zhang, Y. Peng, X., Hu, H. Peng, and X. Cai, "Mixed YOLO-v3-LITE: A lightweight real-time object-detection method", Sensors, Vol. 20, No. 7, p.186, 2020.

[13] R. Chen R, Y. Tong, J. ang, and M. Wu, "Video foreground detection algorithm Based on fast principal component pursuit and motion saliency", Computational Intelligence and Neuroscience, 2020.

[14] G. Modwel, A. Mehra, N. Rakesh, and K. K. Mishra. "A robust real time object detection and recognition algorithm for multiple objects", Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), Vol. 14, No. 1, pp.331-338, 2021.

[15]A. Géron, Hands-On machine learning with scikit learn,keras,and tensorFlow. 2019.

[16]J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-December, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.

[17]T.-Y. Lin et al., “LNCS 8693 - Microsoft COCO: Common Objects in Context,” Comput. Vision–ECCV 2014 13th Eur. Conf. Zurich, Switz., pp. 740–755, 2014.

[18] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 658–666.

[19]M. A. Yousif, “Tongue Print Recognition Based on Ex treme Learning Machine and Convolutional Neural Network,” 2019.

[20]N. Syafri, Edi; Endrizal, “Deep Learning for Computer Vision with Python,” J. Chem. Inf. Model., vol. 53, no. 9, pp. 1689–1699, 2013.