Int. J. Nonlinear Anal. Appl. 13 (2022) 2, 1495–1504
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.6525

A survey of deep learning-based object detection: Application and open issues

Shaymaa Tarkan Abdullah*, Bashar Talib AL-Nuaimi, Hazim Noman Abed

Department of computer science, College of science, University of Diyala, Baqubah, Iraq

(Communicated by Javad Vahidi)

Abstract

Object tracking and detection are among the most significant jobs in computer vision, having many applications in areas, which includes autonomous vehicle tracking, robotics, as well as traffic monitoring. Several studies have been conducted in past years. However, since detecting various problems, for instance, fast motion, illumination variations, as well as occlusion, study in this field persists. Furthermore, deep convolutional neural networks (DCNNs) have grown increasingly significant for object detection as deep learning (DL) techniques have advanced. As a result, numerous approaches for object detection are studied in this research, as well as a comprehensive. This project encompasses backbone networks, loss functions and training strategies, classical object detection architectures, complex problems, datasets and evaluation metrics, applications, future development directions, as well as a review and analysis of DL-based object detection techniques conducted in previous years. Experts in the field of object detection will benefit from this review article.

Keywords: Convolutional neural networks, Deep learning, Machine learning, Object detection
2020 MSC: 68T07

1 Introduction

Object tracking denotes a computer vision problem which involves analyzing videos to detect and track items adhering to one or more categories, for instance, inanimate objects, animals, cars, and pedestrians, despite any previous information of the target's appearance or quantity. Unlike object detection algorithms, which produce a collection of rectangular bounding boxes with coordinates, width and height, object tracking algorithms equate target identities (ID) with each box (recognized as a detection) in distinguishing intra-class objects. It resembles an excerpt of the object tracking algorithm's output.

The object tracking task is crucial in computer vision: from action recognition to crowd behavior analysis, from video surveillance to autonomous cars, many of these issues would gain from a robust tracking algorithm [24]. Nowadays, a growing number of these algorithms have begun taking advantage of deep learning's (DL) representational power. The capacity of deep neural networks (DNN) to discover rich representations as well as extract abstract and complex information from their input is their expertise. On the other hand, convolutional neural networks (CNNs) are state-of-the-art in spatial pattern extraction, which are then utilized in tasks involving image classification [5, 16, 38].

*Corresponding author
Email addresses: scicompms2114@uodiyala.edu.iq (Shaymaa Tarkan Abdullah), bashartalib6@gmail.com (Bashar Talib AL-Nuaimi), hazim_numam@uodiyala.edu.iq (Hazim Noman Abed)

Received: March 2022 *Accepted:* May 2022

This paper overviews the algorithms that employ the abilities of DL models to execute object tracking, with an emphasis on the diverse methodologies utilized for distinct object tracking applications and suggestions for improving the domain of object tracking and detection. The following is how this document is structured: Section 2 delves into the object tracking procedures and who is responsible for processing data to track an object. A quick overview of object detection methods relying on the DL approach is presented in Section 3. Sections 4 and 5 discuss the dataset and evaluation metrics widely employed in object detection. A basic description of important object tracking applications for resolving object detection difficulties in computer vision is provided in Section 6. Section 7 succinctly reviews several of the suggestions that were implemented to enhance and optimize DL models, as well as address a few of the issues that arise throughout testing and training.

2 Basic Stages of Object Tracking Framework

The first section of this study explains the various procedures required in tracking an object or a group of objects in a video sequence. Object detection and object classification are two phases that preface the tracking process and serve a critical role in enhancing tracking accuracy. Although this distinction between the three is minor and might be overlooked upon the first glimpse, it is critical to recognize them since each is distinct and requires its own research. Figure 1 illustrates a simplified flow diagram. The initial step determines which objects are visible in the video frame. Next, contingent on what we want to track, we need to categorize these objects. Afterwards, the actual tracking commences. The three phases are explained as follows, along with the various strategies utilized for each category.

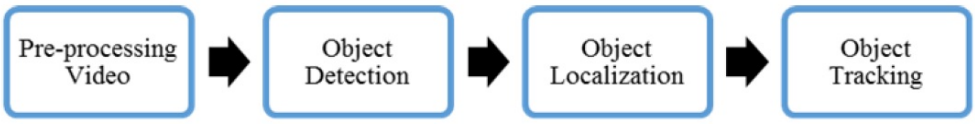

Figure 1: Flow Diagram of Basic steps in Object Tracking

Preprocessing Video Video is a collection of frames. Each frame represents the unique state of object status. Object detection from the frame and keep tracking the particular object throughout the video sequence [25].

Object Detection Object detection denotes a type of computer technology that detects occurrences of semantic objects with respect to a specific class (for example, cars, buildings, humans, etc.) in videos and digital images [29]. Several fundamental strategies for object detection include background subtraction, optical flow, and frame differencing.

Object Localization Object localization is the procedure of classifying the detected objects in the frame as to what object of interest they are. It is simply just figuring out what type of object it is. Object identification could be accomplished through a variety of factors, for instance, texture, color, motion, and shape. We can conduct texture-based classification, color-based classification, motion-based classification, or shape-based classification depending on the parameter employed.

Object Tracking Object tracking is a technique for determining how an object moves in relation to other things by following it over successive image frames. The greatest frequent method is to measure the displacement of the object's centroid in (x, y) in successive frames. Kernel-based tracking, point-based tracking, as well as silhouette-based tracking are the three types of object tracking [9, 30].

3 Object detection methods classification

Object detectors based on deep learning (DL) may be classified into two classifications: one-stage detectors and two-stage detectors. When it comes to two-stage detectors, the first stage generates a small number of fundamental suggestions. The second stage extracts feature vectors from these suggestions that are then encoded utilizing DNNs before generating predetermined class predictions. Moreover, a one-stage detector, on the contrary, considers all of the positions on the input image as potential target objects and attempts to classify each region of interest (RoI) as either a target object or merely background. Furthermore, one-stage detectors are faster and more practical for real-time object detection applications. However, they underperform when contrasted to two-stage detectors, which frequently show impressive findings on a variety of publicly accessible datasets [37]. Over time, a variety of algorithms have been created, and the distinct strategies employed in all these algorithms are outlined as follows:

3.1 TWO-STAGE DETECTORS

3.1.1 R-CNN

The Region-based Convolutional Neural Network (R-CNN) conducts a region search first, as the name implies, followed by classification. Here, the method of locating an object in an image is known as region search. The selective search technique is among the strategies for region search. However, subsequently in 2012, [42] devised an additional approach termed the exhaustive search method. Consequently, it starts by calculating smaller regions or parts of an input image and then groups them together in a hierarchical structure. As a result, the final group or hierarchical structure contains the complete image. When the discovered regions are sorted, two criteria are taken into account: similarity metrics and color space. As a consequence, the final image is established by combining smaller regions into a series of region suggestions. Using a selective search strategy, region proposals are discovered in R-CNN, and then deep learning (DL) is used to detect objects within those detected area proposals. Then, to match the Convolutional Neural Network (CNN)'s input size, CNN is employed. Moreover, the feature vectors of 4096-dimensions were extracted with the help of resizing specific region proposals. Various classifiers that accept these feature vectors as input are used to predict the likelihood of each class. The chance of recognizing items utilizing these feature vectors for specified classes utilizing a pre-trained support vector machine (SVM) is then forecasted. Linear regression may be utilized to reduce item localization inaccuracy in region recommendation, which changes the sizes and shapes of the bounding boxes.

3.1.2 FAST R-CNN

Here, the Fast Region-based Convolutional Network (Fast R-CNN) was created by [32] in 2015. In certain aspects, it resembles the R-CNN. However, its primary purpose is to cut down on time acquired to analyze every region proposal connected to a large number of models. Unlike R-CNN, which needs a CNN for every region proposal, fast R-CNN utilizes the whole picture as the CNN input, employing many convolutional layers. CNN's feature maps are subjected to a selective search technique to determine the RoI. The length and width of the zone of interest, as well as the length and breadth of the major criteria, were used to determine the exact RoI. The RoI pooling layer must be utilized to minimize the feature map size. Each RoI layer's output is forwarded into the fully-connected layers, producing a feature vector as a result. The objects are subsequently recognized utilizing these feature vectors and a softmax classifier, and their localization is modified using a linear regressor.

3.1.3 MASK R-CNN

[14] created the Mask Region-based Convolutional Network (Mask R-CNN) in 2017, which serves to identify bounding boxes while also assisting in predicting object masks. Apart from that, the mask is the process of breaking down items in an image into pixels. These strategies outperform other techniques in COCO challenges four times in a row, including keypoint identification, instance fragmentation as well as bounding box and object detection. The Mask R-CNN generates three outputs: a label for classes, an offset for bounding boxes, and an object mask, using a faster R-CNN software. It also generates bounding boxes using the region proposal network (RPN), and all three outputs are created simultaneously for each RoI. The major RoIPool layer is substituted by a RoIAlign layer in faster R-CNN. Following that, the partition of true RoI coordinates is eliminated, as well as the localization is assessed. Furthermore, the second branch is subsequently coupled to two convolutional layers, and the third branch is used to detect the mask of the object. The actions carried out by these three branches in relation to a set of loss functions are then integrated. Finally, because addressing the fragmentation task improves object localization, which improves the classification rate, this combined value is minimized to create improved performance.

3.2 ONE-STAGE DETECTORS

3.2.1 YOLOv1

[23] concentrated on increasing the object detector's speed. The object detection issue was regarded as a regression problem, as well as the region proposal stage in two-stage detectors was omitted. Rather than employing pre-defined anchors for object regions, it split input images into 7×7 cells, with each cell utilized to estimate the object's center falling into it. Bounding box locations, a score for each bounding box, as well as class probabilities were estimated by each cell. Convolutional layers were utilized to build the network, which was then supplemented by fully-connected layers. In addition, the sum of squared error loss was utilized to reduce classification and localization error. Apart from that, YOLO resembles a real-time object detector that could detect objects at 45 frames per second, which was extraordinarily quick when contrasted to other detectors. Class probabilities, on the other hand, were only estimated inside each cell. It does not operate well with items that are partially localized in one cell and cannot manage a large

number of ground truth objects, which face challenges in accurately predicting bounding box ratios and scales, all of which contribute to low localization accuracy.

3.2.2 YOLOv2

The fully connected layers were eliminated to improve recall, and the anchor boxes concept was utilized to estimate bounding boxes. Furthermore, unsupervised learning methods were employed to construct bounding box ratios and scales directly from training data. Rather than estimating solely one class probability per cell, it estimates both class and objectness for every bounding box, resulting in better detection of partly covered objects. Moreover, the bounding box regression estimated the location in relation to the cell's left top location, resulting in prediction bounds of 0 and 1. Multi-scale training, high-resolution classification, and batch normalization were among the other strategies offered. All of the strategies significantly increased detection accuracy while maintaining high speed [31].

3.2.3 YOLOv3

YOLOv3 was presented as a better alternative to YOLOv2. To address the scenario of several classes in one bounding box, it employed binary cross-entropy loss rather than softmax loss for class prediction. To estimate objects in three scales, it utilized a feature pyramid and multi-scale framework. For better speed and accuracy, a new backbone network with a ResNet module was suggested, particularly for small object detection [22].

3.2.4 YOLOv4

When compared to YOLOv3, YOLOv4 has significantly improved accuracy and speed. However, it just integrates YOLO with the methodologies given in previous years for other models [6]. CSP Darknet53, which integrates the cross-stage partial network (CSPNet) model having the residual block in Darknet53, serves as the YOLOv4 network's backbone. In addition, since its continuous, smooth, non-monotonic, self-regularized characteristics, the activation function in the convolution block is modified from Leaky ReLU to Mish. The path aggregation network (PAN) and spatial pyramid pooling (SPP) are employed in the neck part of the YOLOv4 network. SPP has the ability to dramatically expand the receptive area while also separating the crucial contextual elements. PAN can retrieve the image's multi-scale characteristics continuously. PAN features highly flexible RoI pooling than FPN and narrows the fusion path between feature maps in low and high-level layers [46]. The YOLOv3 head is utilized as the detection head in the new model, with the goal of estimating objects at diverse scales. Suppose there is a sum of three object classes, for instance. As a result, the number of filters is equal to $(\text{classes} + 5) \times 3 = 24$. SPP block, CSP Darknet-53, PANet, as well as the prediction head make up the skeleton of YOLOv4. To increase the variety of the learned features within various layers, CSP Darknet-53 gathers Darknet-53 and CSPNet, which contains the partial transition layer and the partial dense block. Moreover, the SPP block is utilized to expand the receptive field and segregate the crucial context elements without slowing down inference. The detection head has three scales, quite like YOLOv3. The detection head parameters in YOLOv4 are $64 \times 64 \times 24$, $32 \times 32 \times 24$, as well as $16 \times 16 \times 24$, accordingly, when the inputs are 512×512 [6].

3.2.5 Deconvolutional Single Shot Detector (DSSD)

By utilizing a broader network, DSSD was able to increase single shot multibox detector (SSD). ResNet-101, DSSD's deep and robust backbone network, surpassed the VGG network. To offer extra context information, a deconvolutional module was included. More crucially, the deconvolutional layers might be trained throughout the training process, allowing DSSD to be more flexible and achieve better results. To enhance the accuracy of the anchor ratios and scales, K-means clustering is employed to combine training boxes with squared root box areas as the distance measurement. Although DSSD increased SSD accuracy, particularly for small objects, balancing precision and real-time remains a concern [1].

3.2.6 Single Shot MultiBox Detector (SSD)

It resembles a single-shot detector having no stage for regional proposals. SSD conducted detection utilizing several layers to effectively capture multi-scale objects, unlike Faster R-CNN, which only utilized the final detection layer. SSD built numerous anchor scale ranges between layers, and given anchors were deployed to distinct feature maps. The scales of the lower levels were smaller, whereas the scales of the higher layers were bigger [21]. Since its design can manage a broad variety of objects, it has a greater recall rate. SSD employed completely convolutional layers to estimate localization offset as well as confidence score, unlike YOLO, which utilized fully connected layers for object

detection. SSD obtained state-of-the-art performance on numerous test datasets with few extra data augmentations and hard negative mining approaches. SSD, on the other hand, struggled with small objects caused by shallow layers that lacked deep semantic information.

4 Object Detection Datasets

Creating larger datasets with less bias is required for building better computer vision algorithms. Some prominent datasets and benchmarks in the object detection field have been produced in the last ten years, which include the MS-COCO Detection Challenge datasets, the ImageNet Large Scale Visual Recognition Challenge (for example, ILSVRC2014), as well as PASCAL VOC Challenges (for example, VOC2007, VOC2012).

Pascal VOC The PASCAL Visual Object Classes (VOC) Challenges were among the well-known competitions in the early computer vision area (from 2005 to 2012). PASCAL VOC comprises tasks, for instance, image classification [12], action detection, semantic segmentation, and object detection. The most frequently utilized Pascal-VOC versions for object detection are VOC07 and VOC12, with the former having 5k tr. images and 12k annotated objects while the latter having 11k tr. images and 27k annotated objects. These two datasets comprise 20 different types of everyday things (human: person; animal: sheep, horse, dog, cow, cat, bird; vehicle: train, motorcycle, car, bus, boat, bicycle, airplane; indoor: tv/monitor, sofa, potted plant, dining table, chair, bottle). As bigger datasets, for instance, ILSVRC and MS-COCO (to be launched), were developed in previous years, the VOC has progressively fallen out of favor, and it has now turned into a test-bed with respect to new detectors.

ILSVRC Generic object detection has reached a new level since the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Moreover, the ILSVRC was organized annually from 2010 to 2017. There is also an ImageNet image detection challenge there [35]. The ILSVRC detection dataset contains 200 visual object classes. It has a bigger number of image/object instances than VOC by two orders of magnitude. For instance, ILSVRC-14 includes 517k pictures and 534k annotated objects.

MS-COCO The MS-COCO dataset is presently the most difficult object detection dataset available. In 2015, there was a yearly competition centered on the MS-COCO dataset. In comparison to ILSVRC, there are lesser object categories. However, there are greater object instances. Apart from that, MS-COCO-17, for instance, has 164,700 pictures and 897,700 annotated objects organized into 80 categories. In addition, the most evident benefit of MS-COCO over ILSVRC and VOC is that each item is additionally tagged utilizing per-instance segmentation to help in exact localization in conjunction with the bounding box annotations [19]. Besides, in comparison to ILSVRC and VOC, MS-COCO includes more small items (those occupying less than 1% of the image) as well as objects which are densely arranged.

Open Images The Open Images Detection (OID) challenges, which followed in the footsteps of MS-COCO but on a considerably wider scale, were announced in 2018. The two objectives of Open Images: Object detection can be categorized into two categories: simple object detection as well as visual relationship detection, detecting paired objects in specified relations. Furthermore, the object detection dataset comprises 1910k photos having 15,440k bounding boxes over 600 object categories [17].

5 Evaluation Metric

The AP is the most popular metric utilized to quantify the accuracy of detections among various annotated datasets utilized by object detection competitions and the scientific community. In addition, before we explore the available AP versions, we now go through particular principles common among them. Hence, the following are the most important:

- **True positive (TP):** Detection of a ground-truth bounding box that is correct
- **False positive (FP):** An inaccurate detection of an object that does not present or a misplaced detection of an object that does present.
- **False negative (FN):** A ground-truth bounding box that is undetected. A true negative (TN) outcome is not appropriate in the object detection context since there exists an infinite number of bounding boxes that must not be identified within any particular image.

- **Precision** denotes the model's capability to recognize only relevant objects. Moreover, it is the proportion of positive estimates that are fulfilled.
- A model's **recall** is its capacity to discover all pertinent cases (all ground-truth bounding boxes).
- **mean AP (mAP)** is a statistic for assessing object detector accuracy across all classes in a database.
- The likelihood that an anchor box comprises an object is known as a **confidence score**. Therefore, a classifier is frequently employed to estimate it.
- The intersection area is split by the area of the union of a predicted bounding box (Bp) and a ground-truth box (Bgt) to obtain **Intersection over Union (IoU)**. Here, the parameters for determining if detection is a true positive or a false positive are both confidence score and IoU.

6 Application

Object detection, being among the three core tasks of computer vision, possess a broad variety of real applications. Object detection technology is implemented differently based on individual needs. This section examines significant applications of object detection, which include human tracking [39], medical object detection [20], and autonomous driving [33]. Surveillance, autonomous vehicles, and marketing in real-world applications such as both accuracy and speed need to be sufficiently high.

6.1 Human Tracking Detection

Scene understanding from a video is among the greatest obstacles in computer vision. Humans are often the center of attention in a scene, and tracking them in a video is a fundamental problem. Several methods have been recommended depending on the deep learning (DL) model using a different dataset. To accurately recognize human interactions in complex scenarios, lighting and illumination variations are supplied that are indifferent to color and texture alterations. Advanced computer vision sensors, for instance, depth sensors, are utilized for training the convolutional neural network (CNN) algorithm, which makes it simpler to track objects [43]. Head and shoulders detection has become a study hotspot that performs a substantial role in people counting [27] and crowd analysis, which could be utilized for many practical applications, which includes public transportation systems, logistics, and resource management coding, as well as surveillance [3]. The system utilizes the Hough Circular Gradient Transform to detect and track person-based heads and shoulders in a complex environment. Mainly scenarios covered by human tracking detection. Density detection and risk analysis, social distance detection and risk analysis, and face-mask detection and risk analysis are examples of situations. The YOLO, single shot multibox detector (SSD), and other methods are significate algorithms for these scenarios [34]. Multiple human tracking with their identities (IDs) over an image sequence is a difficult task for detecting the positions of multiple humans while maintaining position in real-world applications for instance, surveillance, autonomous vehicles, as well as marketing, both accuracy and speed need to be sufficiently high. The dispersion of interest points with optical flow is utilized to compute a tracking termination measure and a strong interest point selection inside human areas. The experts utilized video sequences taken from an overhead viewpoint to monitor the individuals in a real-time system, and compared them to other tracking algorithms with a 95% accuracy. By merging a mask branch with a fully convolutional twin neural network for target or person tracking, the SiamMask algorithm conducted segmentation of the target person [1].

6.2 Medical Object Detection

Medical image detection may facilitate doctors properly analyzing the lesion region, considerably enhancing medical diagnosis accuracy and lowering doctors' manual workload. DL for object detection is presently playing an essential role in the medical field. For the particle tracking challenge dataset, the DNN created for detection utilized microscope images predicated on object centroids. The network includes anchors, a layer for ensembling detection hypotheses throughout image scales, a Centroid Proposal Network, and a Feature Pyramid Network-based feature extractor and is trainable end-to-end [44]. In real-world applications, gathering and categorizing such a large-scale dataset may be unaffordable, time-consuming, as well as expensive. In the medical image analysis field, semi-supervised object detection has gotten a lot of interest. The authors [47] attempted to improve semi-supervised by reducing the time stands needed for building the DL object detection such as CNN algorithms and chiefly with photographic, medical datasets that are high resolution such as DeepLesion, and Nuclei Data. Other automatic anatomical structure detections and tracking in ultrasound scans employ a DL framework to a chosen anatomical target structure [2].

6.3 Autonomous driving

The challenge of object detection in autonomous cars is among the ongoing study topics that have maintained its appeal in the computer vision field. Note that 3D multi-object tracking has been recommended as a need for self-driving cars. Its goal is to predict the position, orientation, as well as size of all objects in the environment over time using a Kalman Filter technique [8]. The authors arranged a real-world dataset into 20-second sequences with data sampled at 10Hz from numerous sensors. Furthermore, each object in the collection has a unique identification that is constant throughout all sequences. They allow for the assessment of tracking data in 2D image view as well as 3D vehicle-centric coordinates. Current approaches are mostly focused on the detection-by-tracking pipeline, which ultimately necessitates a heuristic matching phase for cloud detection association. SimTrack was developed to ease the hand-crafted tracking paradigm by presenting an end-to-end trainable model for joint recognition and tracking from raw point clouds, according to the authors. Their goals were to estimate each object's first-appear location in a particular sample to determine the tracking identification and afterwards update the location using motion estimation from the Waymo Open Dataset [40]. The authors [4] started by examining self-driving architectures that include DL and neural networks, as well as the deep reinforcement learning approach. These approaches are the foundation for self-driving scene perception and motion planning. Tiny YOLOv3 proposed object detection-based DL for car detection, which combined with the Kalman filter. A small and real-time object detection system increases the model's accuracy without losing its speed [36]. The survey's objective was to investigate possible threats to the safety of the DL-based autonomous driving system's workflow, for instance, adversarial attacks, cyberattacks, and physical attacks. The physical attack was simple. However, it revealed several limitations that may be easily countered by defense tactics. The cyberattack proved challenging to carry out on a wide scale, but system defensive mechanisms were simple to build. The adversarial attack was successful, and new defensive measures were required to counter it, as standard defense tactics were ineffective in the setting of the self-driving context [11]. The CNN algorithms were developed and tested with the goal of modifying driving behavior and performing driving tasks instantaneously without the need for human involvement [45]. For the identification and categorization of on-road obstacles, for instance, vehicles, pedestrians, as well as animals, a DL system utilizing a region-based CNN trained with a PASCAL VOC image dataset were built. The system's execution on a Titan X GPU resulted in a processing frame rate of at least 10 fps for a VGA resolution image frame [28]. Relying on dynamic driving representation leveraging on audacity self-driving and gathered dataset as well as real-world environmental perceptions, a DNN was developed to automate direction prediction and steering angle [15]. The primary input is the driver-view images taken by the autonomous car's camera, and the deep Monte Carlo Tree Search algorithm can learn to manage the vehicle having no human input. By conducting virtual driving simulations, our program can anticipate driving maneuvers [7]. Current advances in DL and sensor technology have sped up the advancement of autonomous driving technology, which might enhance personal mobility, traffic efficiency, and road safety [10]. With the aid of modern artificial intelligence (AI) approaches, advances in information and signal processing technologies have had a substantial influence on autonomous driving (AD), boosting driving safety while decreasing human drivers' efforts. DL techniques have previously been utilized to handle a number of complex real-world occurrences [26]. The research discovered DL, which can address multiple issues at once via multitask learning. Upcoming possibilities encompass not just image "recognition" but also great hopes for the establishment of end-to-end learning and deep reinforcement learning technologies for autonomous vehicle "judgment" as well as "control" [13]. Scholars created a light enhancement net (LE-net) built on the CNN to solve the lack of clarity in images of road scenes in low-light circumstances. To begin, we presented a generation pipeline for converting daylight photographs to low-light images, which we then utilized to create image pairs for model enhancement [18]. Another aim was to improve truck or car detection utilizing support vector machines (SVMs) as well as CNNs [41]. Existing implementations of AD systems are limited to controlled and confined situations in limited quantities due to technical constraints and the expense of exteroceptive sensors. For the identification and categorization of on-road obstacles, for instance, animals, pedestrians, as well as vehicles, a DL system employing a region-based CNN trained to utilize the PASCAL VOC image dataset has been created. With respect to a VGA resolution image frame, the system's application on a Titan X GPU yields a processing frame rate of at least 10 fps. Therefore, this high frame rate combined with a strong GPU indicates the system's capability for autonomous car highway driving. The identification and classification findings on images from KITTI and inroads, as well as Indian roads, demonstrate the system's effectiveness in a variety of lighting and meteorological situations.

7 Recommendations in object detection

- The necessity to design a technology that may improve the recognition and tracking of groups of humans in the actual world due to overlaps between humans in video frames.

- Object detection is utilized in many security applications, especially in video surveillance. Scholars must be able to distinguish persons in prohibited or dangerous places, avert suicide or automate inspection tasks on remote sites utilizing computer vision.
- A confidence factor for each tracked individual must be identified to enhance the system's accuracy and robustness. In addition, we will create an algorithm to help us progress.
- For a more effective categorization system, the scholars want to investigate new and improved techniques for merging RGB and depth images. It also aims to train and test the suggested algorithm on more difficult datasets.
- Medical images are frequently plagued by data imbalance issues, making disease categorization extremely challenging. When a percentage of a certain type of disease in a dataset occurs in a tiny region of the overall dataset, this is known as an unbalanced dispersion of data in medical datasets. Therefore, it is vital to pay attention to cost-sensitive imbalance learning while building DL-based object detection.
- Although the model is only applicable to human diseases, few image detections for plant disease have lately been published; nonetheless, they may be used for a variety of fruit and crop detection tasks, general disease detection challenges, and many automated agricultural detection procedures.
- There is still a lot that can be done to make automated vehicles commercially viable. More diversified datasets, for instance, will enhance the performance of current object detection networks by simulating non-optimal real-life driving circumstances.

8 Conclusion

A comprehensive description of all object-tracking algorithms presented for the use of deep learning (DL) techniques is presented, focusing on four main steps to characterize a general object-tracking pipeline: data processing, object detection, object localization and tracking. DL has been investigated for application in object tracking. While the majority of the curriculum centered on various DL applications of learning as well as the application of current functions, there were a few exceptions. However, DL is only utilized in several approaches to lead the correlation algorithm. Object discovery is currently being explored, and the whole artificial intelligence (AI) area is built mostly on data. Although there is a growing demand for object-tracking applications to cover all real-life problems, the reality of application performance is not enough, and this is due to the lack of data collection capabilities to build and develop models. The information presented in this study can be considered to help researchers give an overview of the current advancements in the object tracking field for human, autonomous driving and medical detection.

References

[1] I. Ahmed and G. Jeon, *A real-time person tracking system based on SiamMask network for intelligent video surveillance*, J. Real-Time Image Process. **18** (2021), no. 5, 1803–1814.

[2] A.F. Al-Battal, Y. Gong, L. Xu, T. Morton, C. Du, Y. Bu, I.R. Lerman, R. Madhavan and T.Q. Nguyen, *A CNN segmentation-based approach to object detection and tracking in ultrasound scans with application to the vagus nerve detection*, Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, 2021, pp. 3322–3327.

[3] D. Anitta, *Human head pose estimation based on HF method*, Microprocess. Microsyst. **82** (2021), 103802.

[4] M.R. Bachute and J.M. Subhedar, *Autonomous driving architectures: insights of machine learning and deep learning algorithms*, Mach. Learn. Appl. **6** (2021), 100164.

[5] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi and P.H.S. Torr, *Fully-convolutional siamese networks for object tracking*, Eur. Conf. Comput. Vis. Springer, Cham., 2016, pp. 850–865.

[6] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, *Yolov4: optimal speed and accuracy of object detection*, arXiv Prepr. arXiv2004.10934, (2020).

[7] J. Chen, C. Zhang, J. Luo, J. Xie and Y. Wan, *Driving maneuvers prediction based autonomous driving control by deep monte carlo tree search*, IEEE Trans. Veh. Technol. **69** (2020), no. 7, 7146–7158.

[8] H.-K. Chiu, J. Li, R. Ambrus and J. Bohg, *Probabilistic 3d multi-modal, multi-object tracking for autonomous driving*, IEEE Int. Conf. Robotics and Automation (ICRA), 2021, pp. 14227–14233.

[9] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, R. Tagliaferri and F. Herrera, *Deep learning in video multi-object tracking: a survey*, Neurocomput. **381** (2020), 61–88.

[10] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li and D. Cao, *Deep learning for image and point cloud fusion in autonomous driving: a review*, IEEE Trans. Intell. Transp. Syst. **23** (2022), no. 2, 722–739.

[11] Y. Deng, T. Zhang, G. Lou, X. Zheng, J. Jin and Q.L. Han, *Deep learning-based autonomous driving systems: a survey of attacks and defenses*, IEEE Trans. Ind. Inf. **17** (2021), no. 12, 7897–7912.

[12] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn and A. Zisserman, *The pascal visual object classes (voc) challenge*, Int. J. Comput. Vis. **88** (2010), no. 2, 303–338.

[13] H. Fujiyoshi, T. Hirakawa and T. Yamashita, *Deep learning-based image recognition for autonomous driving*, IATSS Res. **43** (2019), no. 4, 244–252.

[14] K. He, G. Gkioxari, P. Doll and R. Girshick, *Mask r-cnn*, Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2961–2969.

[15] N. Ijaz and Y. Wang, *Automatic steering angle and direction prediction for autonomous driving using deep learning*, Proc. Int. Symp. Comput. Sci. Intell. Control. ISCSIC 2021, pp. 280–283.

[16] L. Kalake, W. Wan and L. Hou, *Analysis based on recent deep learning approaches applied in real-time multi-object tracking: a review*, IEEE Access **9** (2021), 32650–32671.

[17] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci and T. Duerig, *The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale*, arXiv 2018. arXiv preprint arXiv:1811.00982, (2018).

[18] G. Li, Y. Yang, X. Qu, D. Cao and K. Li, *A deep learning based image enhancement approach for autonomous driving at night*, Knowledge-Based Syst. **213** (2021), 106617.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C.L. Zitnick, *Microsoft coco: common objects in context*, Eur. Conf. Comput. Vision, 2014, pp. 740–755.

[20] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laa, B. Van Ginneken and C.I. Sánchez, *A survey on deep learning in medical image analysis*, Med. Image Anal. **42** (2017), 60–88.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu and A.C. Berg, *Ssd: single shot multibox detector*, Eur. Conf. Comput. Vision, 2016, pp. 21–37.

[22] Y. Liu, P. Sun, N. Wergeles and Y. Shang, *A survey and performance evaluation of deep learning methods for small object detection*, Expert Syst. Appl. **172** (2021), 114602.

[23] C. Liu, Y. Tao, J. Liang, K. Li and Y. Chen, *Object detection based on YOLO network*, Proc. 2018 IEEE 4th Inf. Technol. Mechatronics Eng. Conf. ITOEC 2018, pp. 799–803.

[24] X. Ma, W. Ouyang, A. Simonelli and E. Ricci, *3d object detection from images for autonomous driving: a survey*, arXiv preprint arXiv:2202.02980, (2022), 1–26.

[25] A. Makandar, D. Mulimani and M. Jevoor, *Preprocessing step–review of key frame extraction techniques for object detection in video*, Int. J. Curr. Eng. Technol. **5** (2015), no. 3, 2036–2039.

[26] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser and V.H.C. De Albuquerque, *Deep learning for safe autonomous driving: current challenges and future directions*, IEEE Trans. Intell. Transp. Syst. **22** (2021), no. 7, 4316–4336.

[27] M. Pervaiz, Y.Y. Ghadi, M. Gochoo, A. Jalal, S. Kamal and D.S. Kim, *A smart surveillance system for people counting and tracking using particle flow and modified som*, Sustain. 13 (2021), no. 10, 1–20.

[28] G. Prabhakar, B. Kailath, S. Natarajan and R. Kumar, *Obstacle detection and classification using deep learning for tracking in high-speed autonomous driving*, TENSYMP 2017 - IEEE Int. Symp. Technol. Smart Cities, 2017, pp. 3–8.

[29] A. Raghunandan, P. Raghav and H.R. Aradhya, *Object detection algorithms for video surveillance applications*, Proc. 2018 IEEE Int. Conf. Commun. Signal Process. ICCSP 2018, pp. 563–568.

[30] K. Ragland and P. Tharcis, *A survey on object detection, classification and tracking methods*, Int. J. Eng. Res.

Technol. **3** (2014), no. 11, 622–628.

[31] J. Redmon and A. Farhadi, *YOLO9000: better, faster, stronger*, Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog., (2017), pp. 7263–7271.

[32] S. Ren, K. He, R. Girshick and J. Sun, *Faster r-cnn: towards real-time object detection with region proposal networks*, Adv. Neural Inf. Process. Syst. 28 (2015).

[33] F. Rosique, P.J. Navarro, C. Fernández and A. Padilla, *A systematic review of perception system and simulators for autonomous vehicles research*, Sensors **19** (2019), no. 3.

[34] L. Rupasinghe and M.C. Liyanapathirana, *Human tracking and profiling for risk management*, Global J. Comput. Sci. Technol. **22** (2022), no. 1.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein and A.C. Berg, *Imagenet large scale visual recognition challenge*, Int. J. Comput. Vis. **115** (2015), no. 3, 211–252.

[36] A. Shafique, G. Cao, Z. Khan, M. Asad and M. Aslam, *Deep learning-based change detection in remote sensing images: a review*, Remote Sens. **14** (2022) , no. 4, 1–40.

[37] V. Sharma and R.N. Mir, *A comprehensive and systematic look up into deep learning based object detection techniques: a review*, Comput. Sci. Rev. **38** (2020), 100301.

[38] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. 2015, pp. 1–14.

[39] Z. Soleimanitaleb, M.A. Keyvanrad and A. Jafari, *Object tracking methods: a review*, 9th Int. Conf. Comput. Knowl. Eng. ICCKE 2019, pp. 282–288.

[40] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine and V. Vasudevan, *Scalability in perception for autonomous driving: waymo open dataset*, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2020, pp. 2443–2451.

[41] A. Uçar, Y. Demir and C. Güzeliş, *Object recognition and detection with deep learning for autonomous driving applications*, Simulation **93** (2017), no. 9, 759–769.

[42] K.E.A. Van De Sande, J.R.R. Uijlings, T. Gevers and A.W.M. Smeulders, *Segmentation as selective search for object recognition*, Proc. IEEE Int. Conf. Comput. Vis. 2011, no. 2, pp. 1879–1886.

[43] M. Waheed, M. Javeed and A. Jalal, *A novel deep learning model for understanding two-person interactions using depth sensors*, Int. Conf. Innov. Comput. (ICIC), IEEE, 2022, 1–8.

[44] T. Wollmann and K. Rohr, *Deep consensus network: aggregating predictions to improve object detection in microscopy images*, Med. Image Anal. **70** (2021), 102019.

[45] Y. Yin, *Design of deep learning based autonomous driving control algorithm*, 2nd Int. Conf. Consumer Electron. Comput. Engin. (ICCECE), IEEE. 2022, pp. 423–426.

[46] Z. Zhang, Y. Li, W. Wu, H. Chen, L. Cheng and S. Wang, *Tumor detection using deep learning method in automated breast ultrasound*, Biomed. Signal Process. Control **68** (2021), 102677.

[47] H.Y. Zhou, C. Wang, H. Li, G. Wang, S. Zhang, W. Li and Y. Yu, *SSMD: semi-supervised medical image detection with adaptive consistency and heterogeneous perturbation*, Med. Image Anal. **72** (2021), 102117.