One Framework to Detect Them All: A Cross-Industry Object Detection System Using Detectron

Yash Sharma
Department of Computer Engineering
Thakur College of Engineering and Technology

Eshan Save
Department of Computer Engineering
Thakur College of Engineering and Technology

Omkar Shetty
Department of Computer Engineering
Thakur College of Engineering and Technology

Vaishali Nirgude
Department of Computer Engineering
Thakur College of Engineering and Technology

Abstract

Object detection and tracking is one of the most important and demanding fields in computer vision and have been widely distributed applied in various fields such as health care monitoring, autonomous control, anomaly detection and so on. With rapid development deep learning (DL) networks and GPU computing power, the performance of detectors and object trackers was very high improved. Understand the main development status of the object detection and monitoring of pipelines thoroughly, in this survey, we have critically analyzed the existing DL object network methods detection and tracking and described various reference datasets. This includes recent developments in granular DL models. We have primarily provided a comprehensive overview of the various species both general object detection models and specific object detection models [1]. We have obtained various benchmarks to get the best ones detector, tracker and their combinations. In addition, we have stated traditional and new object detection and tracking applications showing its development trends. Finally, challenging problems, including the relevance of granular computing, in the domain mentioned are developed as a future scope of research along with some concerns [1].

I. INTRODUCTION

In recent years, object detection and tracking has gained increasing attention due to a wide range of applications and recent breakthrough research. In both real-world and academic applications, object detection and tracking are of equal importance [1]. Imaging technology has made tremendous progress in recent years. At the same time, computing power IS increasing dramatically as proposed in. In recent years, computing platforms have focused on parallelization, such as multi-core processing and graphics processing units (GPUs). Such a hardware version enables CV for object detection and tracking for real-time implementation. Rapid development in deep convolutional neural network (CNN) and improved GPU computing power arethe main reasons for the rapid development of CV-based object detection and tracking [1].

In a detection network, a deep CNN is used as a backbone to extract key features from an input image/video. These features are used to locate and classify objects within the same frame. Then, in object tracking, these detected objects are tracked based on the proximity of the object from frame to frame. Object detection refers to scanning and searching for objects of certain classes (eg human, car and building) within an image/video [1].

Object detection can be done using either image processing techniques or DL networks. Image processing techniques usually do not require historical data for training and are unsupervised. However, these techniques are limited by various factors such as complex scenarios, lighting effect, occlusion effect, and interference effect. All these problems are better solved in DL-based object detection. The working principle of DL networks is inherently supervised and limited by the huge amount of training data and GPU computing power. Many reference datasets are already developed in the field of object detection, such as Caltech, KITTI, ImageNet, PASCAL VOC, MS COCO, and V5. Due to the availability of such a huge amount of data and the development of GPUs, object detection based on DL networks is widely accepted by researchers. Object detection is followed by object tracking. The goal of object tracking is to locate the trajectory of the detected object and connect it to it. An efficient and robust system design is required for object tracking in a domain-specific or general scenario. Recently developed DL networks meet this goal. For example, consider the research on DL networks for image classification that was conducted in the ILSVRC 2012 competition. Here, the error rate is reduced by 10%compared to conventional methods. Then, new deeper learning networks are gradually developed for image classification. They are well accepted by the human vision community due to their effectiveness Multiple Object Tracking (MOT) is Deep Learning in detecting and tracking multiple objects which is more complex than single object tracking and more applicable in real-time scenario. That is why research at STK is overwhelmed by researchers. Although DL has been observed to be effective for MOT problems, the tracking performance is purely based on the success of correct image localization and classification [1].

II. LITERATURE SURVEY OF OBJECT DETECTION AND TRACKING ALGORITHMS

In, Qiang Ling et.al, developed feedback-based object detection algorithm. It adopts dual layer updating model to update the background and segment the foreground with an adaptive threshold method and object tracking is treated as an object matching algorithm. [BACKGROUND MODEL BASED DETECTION AND TRACKING] [2].

In computer vision systems, the basis of object detection and tracking is the background structure. The traditional background modeling method often requires complex calculations and is sensitive to lighting changes. Therefore, a new hierarchical method based on coarse to fine textures for background modeling was proposed in [3]. It has the following advantages:
(1) tolerance to illumination changes, (2) low computation, and (3) excellent description for each block when the multimode method is applied. This method is quite effective. [DETECTION BASED ON BACKGROUND MODELING].

Integrating a camera system into a mobile robot is very demanding and computationally intensive. High-quality image data only provides accurate information about the environment. However, it requires high computational demands. Hannes Bistry and Jianwei Zhang proposed an object detection algorithm based on SIFT. They created a distributed SIFT visionalgorithm, and the intelligent camera architecture can be used to integrate a complex vision algorithm. [SEGMENTATION BASED OBJECT DETECTION].

III. OBJECT DETECTION AND TRACKING

In this section, we briefly discuss various approaches, both conventional and DL-based, for the detection and tracking of multiple objects along with their characteristic features. As already mentioned, both object detection and tracking are important in the field of CV. In general, object detection is performed in two steps: finding foreground entities (using features) that are hypothesized to be an object, and then validating these candidates (using a classifier). We divide object detection into three board categories; i) based on appearance, ii) based on movement and iii) based on DL. Appearance-based approaches use image processing techniques to recognize objects directly from images/video [3]. However, these approaches usually fail to detect occluded objects. Whereas in motion-based approaches, a sequence of images is used for object recognition. These methods may not work well for object detection in complex scenarios. DL-based approaches use either appearance features or motion features or a combination of these to detect objects in images/videos. Due to recent technological breakthroughs, DL-based object detection approaches have received much attention compared to appearance- or motion-based approaches. Deep CNNs are used as the backbone in DL-based object detectors to extract features from the input image/video. These features are used to classify objects. DL-based approaches have two categories: i) two-stage detectors and ii) one-stage detectors. In two-stage detectors, approximate object regions are first designed using deep features, and then these features are used for classification as well as bounding box regression for the object candidate. On the other hand, for single-stage detectors, bounding boxes are predicted on images without a region design step. This process consumes less time and therefore can be used in real-time devices. Two-stage detectors achieve high detection accuracy, while single-stage detectors have high speed. Different backbone networks (feature generation networks) used in DL-based object detection are: i) AlexNet , ii) ResNet, and iii) VGG16, among others. With the development of backbone networks and the increasing capabilities of GPUs, remarkable progress has beenmade in the field of two-stage object detectors. Recently, the concept of granular computing has been incorporated into deep networks to greatly increase the computation speed and balance with the detection accuracy. Some such networks are granular CNNs and granular RCNN.[1]

IV. GENERIC OBJECT TRACKING

General object detectors aim to locate and classify objects in an image and mark them with rectangular bounding boxes to show certainty of existence. Generic object detectors are of two types: two-stage detectors and one-stage detectors. Two-stage detectors follow the traditional pipeline of object detection, i.e. object localization and its classification. While single-stage detectors treat the task of object detection as a regression/classificationproblem. For both detectors, the classification task is performed based on some features that are generated using a feature generation network, called a backbone network.

Recent advances in deep learning and the availability of computing power have revolutionized several fields such as computer vision and natural language processing (NLP). Object detection is a well-developed field in computer vision. Object tracking is typically the next process after object detection that receives an initial set of detected objects, inserts a unique identification (ID) for each of the initial detections, and then tracks the detected objects as they move between frames.[3]

Multiple Object Tracking (MOT) is a subset of object tracking that is designed to track multiple objects in a video and represent them as a set of high-accuracy trajectories. However, object tracking usually has one big problem when the same object does not have the same ID in all frames, which is usually caused by ID switching and occlusion. ID switching is a phenomenon where an object X with an existing ID A is assigned a different ID B, whichcan be caused by many scenarios, such as a tracker assigns another object Y an ID A because it resembles object X. Another problem is occlusion, which is when another object partially or completely covers one object for a short period of time.[3]

Figure 1 illustrates the MOT process. Initially, objects in the current frame are detected by the detector. The objects are then tracked as they are loaded into the MOT algorithm. Figure 2 then visualizes the process of tracking multiple tracked objects from the current frame to the next frame. Both figures show the MOT's tendency to accurately track a large number of objects [3]

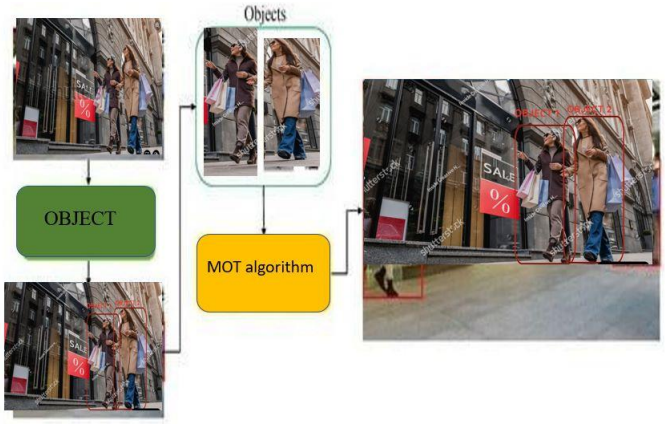

FIGURE 1

Figure 1. Using the MOT15 reference dataset, ID allocation method, one of the core concepts of MOT, is explained. Use the detector to detect objects in the current frame first. The detected results are passed to the MOT algorithm to assign an identifier to each object.[3]

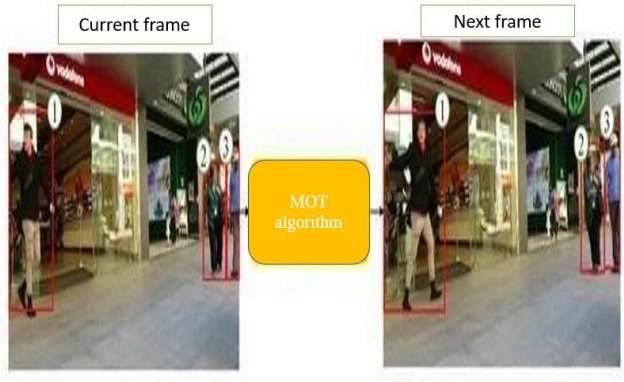

FIGURE 2

Figure 2. Object tracking visualization for the next frame using the MOT algorithm on the MOT15 test dataset. Objects in the next frame are detected first. The detected result is passed to the MOT algorithm to compare objects in the current frame with objects in the next frame. Finally, based on the current frame, each object in the next frame is assigned an ID as shown [3]

ID	Ref.	Year	Contributions
1	[9]	2021	• Offers a comprehensive review of object detection models. • Shows the development trends of both object detection and tracking. • Describes various comparative results for getting the best detector and tracker. • Categorizes deep learning based on object detection and tracking into three groups.
2	[10]	2020	• Reviews the previous deep learning-based MOT research in the past 3 years. • Divides previous papers into five main sections, which include detection, feature extraction and motion prediction, affinity computation, association/tracking, and other methods. • Shows the main MOT challenges.
3	[11]	2020	• Shows the key aspects in a multiple object tracking system. • Categorizes previous work according to various aspects, and explains the advances and drawbacks of each group. • Provides a discussion about the challenges of MOT research and some potential future directions.
4	[13]	2020	• Reviews the past five years of multi-object tracking systems. • Compares the results of online MOTs and public datasets environment in the deep learning model. • Focuses mainly on deep-learning-based approaches

TABLE 1

In recent years, many new MOT studies have been proposed to address existing tracking issues such as real-time tracking, ID switching, and occlusion. In addition, deep learning is increasingly applied to MOT to improve its performance and robustness. Table 1 details the contributions of some previous surveys on MOT. Overall, each survey focused on a specific MOT issue. Recently, Pal et al. focused on the deep learning method and explained detection and tracking separately so that readers can easily focus on their part of interest. However, dueto the description of many parts related to detection, the description of tracking is insufficient.[3]

On the other hand, Ciaparrone et al. reviewed deep learning-based MOT papers published in the last three years. They described online methods that work in real time and batch methods that can use global information and compare experimental results. However, they only focused on MOT benchmarks and did not provide any comparisons for other benchmarks. In another review, Luo et al. described the MOT methodology in two categories and the evaluation focused on the PETS2009-S2L1 sequence of the PETS benchmark. Finally, Kalake et al. reviewed MOT documents for the last 5 years. Although they covered many aspects of MOT, it was difficult to determine an exact rating for each monitoring method due to limited evaluation.[3]

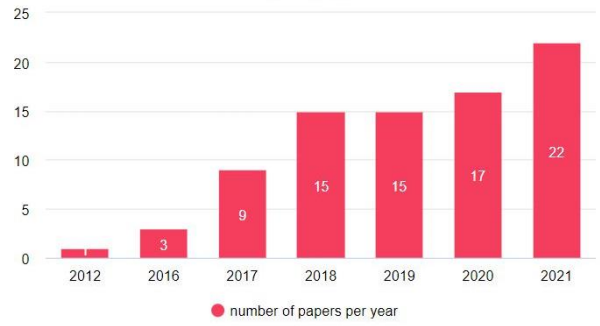

FIGURE 3

Figure 3: Figure 3: Shows the total number of articles reviewed in this review. In general, there is an increasing number of MOT articles presenting various deep learning-based MOT frameworks, new hypotheses, procedures, and applications. Previous surveys have partially addressed tracking, specifically the topic of MOT. However, some existing parts of MT were not covered in these reviews. For example (1) mostly

the surveys concentrated on the detection part rather than the tracking part, and (2) a limited number of benchmarks were mentioned. As a result, a comprehensive survey on recentMOT work is meaningful for stakeholders and researcherswho want to integrate MOT into the existing systems or start new MOT research. This survey summarizes the previous workand covers many aspects of MOT.

V. THE EMERGENCE OF DEEP MODULAR LEARNING

Compared to other machine learning methods, deep learning is remarkably modular. This modularity gives it unprecedented capabilities that put Deep Learning head and shoulders above any other conventional Machine Learning approach. However, recent research points to even more modularity than before. It is likely that monolithic deep learning systems will soon become a thing of the past.

Before I discuss what's coming in the future, let me firstdiscuss the concept of modularity. This is a concept familiar to software engineering, but the idea is not as commonlyfound in machine learning. In informatics, we build complex systems from modules. One module assembled from several simple modules. This allows us to build our digital world based only on NAND or NOR gates. Universal Boolean operators are necessary but not sufficient to build a complex system. Complex computing systems require modularity in order to have a manageable way to manage complexity.

Here are the six main design operators that must be supported in a modular system:

• Partitioning — Modules can be independent.

• Substituting — Modules can be substituted and swapped.

• Expansion – new modules can be added to create new solutions.

• Inverting — Hierarchical dependencies between modules can be rearranged.

• Porting — Modules can be used in different contexts.

• Excluding – Existing modules can be removed to create a usable solution. [4]

These operators are generic in nature and are inherent to any modular design. They allow the modification of an existing structure to new structures in well-defined ways In the context of software, this can mean refactoring statements at the source code level, language construction at specification, or component models at construction. These operators are perfect in that they can generate any structure in your computer design. The six-statement definition focuses on functional invariance in the presence of design variations. More clearly, allows you to use these operators and does not affect the function as a whole as shown in . In the context of deep learning, modularity operators are accepted as follows

• Splitting — Pretrained autoencoders can be split and reused as layers in another network.

• Substitution — Through transfer learning, student networks can serve as substitutes for teacher networks.

• Extensions — New meshes can be added later to increase accuracy. You can combine training networks to improve generalization. In addition, neural network outputs can be used as neural inputs, which can be used as representations forother neural networks.

• Porting — A neural network can be "ported" to another context by replacing the upper layers. This works in cases where the domains are similar enough. Further research on domain matching is needed to understand the limits of this method.

The two remaining modular operators are not available for current monolithic DL systems.

• Inverting — Layers in the network cannot be rearranged without catastrophic consequences. The layers of a monolithic DL system are too tightly coupled to allow this.

• Exclusion – There is no mechanism to “forget” or exclude functionality from a monolithic DL system. [4]

However, despite these two drawbacks, DL systems have an unrivaled advantage over competing machine learning techniques. One of the reasons for the tight coupling of layers in a monolithic DL system can be traced back to Stochastic Gradient Descent (SGD). SGD works in lock-step mode with training. It is a very highly synchronized mechanism that requires coordination of behavior across all layers. However, this monolithic structure is replaced by an even more modular system. Here are two interesting events related to this. DeepMind has explored a method called "Synthetic Gradients" that points the way to looser layers. The method essentially inserts a proxy neural network between the layers to approximate gradient descent

A second development that may lead to greater modularity is the concept of generative adversarial networks (GANs). Typical GANs have two competing neural networks that are essentially separate but contribute to the global objective function.However, now in research we are seeing the emergence of morecomplex configurations such as this Where you have a ladder network of separate encoders, generators anddiscriminators. The general pattern is that all normal functions of neural networks have also been replaced by neural networks. In particular, the SGD algorithm and objective function have been replaced by neural networks. All analysis features are gone! This is what happens when you have deep meta-learning as shown in .

Another very impressive result, also recently published with the same name (i.e. StackGAN), shows how efficient multiple separate GANs can be: The task here is to take a text description as input and generate an image matching the description. Here we have two GANs staged one after the other. The second GAN is able to refine the fuzzy image into one with a higher resolution. Modular networks have the ability to factor capabilities that would otherwise be entangledin an end-to-end network. In software engineering, we have the concept of API. That is, a restrictive language that communicates between different modules. In the above scenario, a neural network that "learns to communicate" acts as an API bridge between networks

We consider the problem of multiple agents sensing and acting in an environment to maximize their shared utility [5]. In these environments, agents must learn communication protocols to share information needed to solve tasks. By adopting deep neural networks, we are able to demonstrate end-to-end learning protocols in complex environments inspired by communication puzzles and multi-agent computer vision problems with partial observability.

Another recent paper titled "Generative Adversarial Parallelism" explores this further in relation to GANs. In this work, the authors attempt to address the difficulties in training GANs by extending the usual two-player generative adversarial games to a multiplayer game. They train many GAN-like variants in parallel while periodically swapping the discriminator and generator pairs. The motivation here is to achieve better separation between pairs. Much work remains to be done to determine whether separate interfaces between networks lead to better generalization

VI. COMPARISON

ImageNet images typically have one large object. Thus, our non-prediction-based methods, such as image-box, which treats the entire image as a bounding box, are suitable for ImageNet. To test whether our loss works with different distributions of multi-object images, we test it using the Conceptual Captions (CC) dataset. Even on this challenging dataset with multiple objects/labels per frame, Detic provides a gain of ∼2.6 points in new class detection over the best prediction-based methods. This suggests that our simpler Detic method can generalize to different types of image- tagged data. Overall, the results from Table 1 indicate that complex prediction-based methods that rely heavily on modelprediction scores do not perform well for open dictionary detection. Among our non-prediction-based variants, max size loss consistently performs best and is the default for Detic in our following experiments [6].

	IN-L		CC	
BOX SUPERVISED (BASELINE)	30.0±0 .4	16.3± 0.7	30.0±0. 4	16.3 ±0.7
YOLO9000	31.2±0 .3	20.4± 0.9	29.4±0. 1	15.9 ±0.6
WSDDN	29.8± 0.2	15.6 ±0.3	30.0±0 .1	16.5 ±0.8
Detic (Max-object-score)	32.2± 0.1	24.4 ±0.3	29.8±0 .1	18.2 ±0.6
Detic (Image-box)	32.4± 0.1	23.8 ±0.5	30.9±0 .1	19.5 ±0.5
Detic (Max-size)	32.4± 0.1	24.6 ±0.3	30.9±0 .2	19.5 ±0.3

VII. APPLICATIONS

Major One of the key safety concerns on construction sites is ensuring that workers wear the appropriate personal protective equipment (PPE), such as hard hats [7]. However, monitoring compliance with PPE requirements can be a challenging task, particularly on large construction sites with many workers.

This technology could be used to address this challenge by automating the detection of hard hats worn by construction workers. By training a THIS model on images and videos of workers wearing hard hats, the model could accurately detect when a worker is not wearing a hard hat or is wearing it improperly. This would enable safety managers to quickly identify and address compliance issues, reducing the risk of injuries and accidents on the worksite.

In addition, this technology can be used to monitor the compliance of individuals with safety regulations, such as the wearing of safety gear, in both school and office settings. For example, in schools, this could be used to detect whether students are wearing safety goggles or lab coats in a chemistry lab. In offices, it could be used to ensure that employees are wearing safety gear, such as helmets and safety shoes, in manufacturing and construction areas.

Another potential application of This technology in schools and offices is for environmental monitoring. By analyzing images and videos of classrooms and office spaces, This models could be trained to detect potential safety hazards, such as overcrowding, blocked exits, or fire hazards. This could help safety managers and facilities personnel identify and address potential safety hazards before they become serious problems.

VIII. CONCLUSION

Major advances in deep learning methods have been made in the areas of image recognition, object detection, and person re-identification, which also benefit from the development of multi-object tracking. In this article, we summarize the deep learning-based multi-object tracking methods that rank high in public benchmarks. The contribution of this article lies in three aspects. First, the use of deep learning for multi-object tracking is organized, and the mechanisms of deep feature transfer, neural network embedding, and end-to-end network training are analyzed based on existing methods, and the rules for designing a new tracking framework are analyzed. they are inspired. Second, we examine the roles of deep networks in surveillance and investigate the training issues of these networks. Third, comparisons between these multi-object tracking methods are

presented and reorganized according to common datasets and evaluations. Advantages and limitations of the methods are highlighted. From the analysis of the experimental evaluation, it can be seen that there is a lot of room for improving the tracking results using the deep learning paradigm. This document provides some useful insights. On the one hand, there are not enough labeled datasets to train satisfied tracking models under all conditions. A possible path can be paved by generative networks, which are excellent for supporting the generalization of deep learning models. On the other hand, to cope with adverse tracking results in a complex environment such as a moving platform,it is necessary for integrated network models to learn the features of these dynamic scenes. In addition, to further adapt to changing conditions, higher-order feature learning or online transfer features are expected for tracked object [8].

REFERENCES

1. S Pal, S.K., Pramanik, A., Maiti, J. et al. Deep learning in multi-object detection and tracking: state of the art. Appl Intell 51, 6400–6429 (2021).

2. Pawar, R. B., Phonde, V. S., & Patel, A. B. (n.d.). Real Time Object Detection. Real Time Object Detection. https://www.ijsrd.com/Article.php?manuscript=IJSRDV10I20113

3. Park Y, Dang LM, Lee S, Han D, Moon H. Multiple Object Tracking in Deep Learning Approaches: A Survey. *Electronics*. 2021; 10(19):2406. https://doi.org/10.3390/electronics10192406

4. Perez, C. E. (2017, October 13). The Emergence of Modular Deep Learning. Medium. Retrieved April 6, 2023, from https://medium.com/intuitionmachine/the-end-of-monolithic-deep-learning-86937c86bc1f

5. Assael, I. A. (2019). Deep learning for communication: emergence, recognition and synthesis (Doctoral dissertation, University of Oxford).

6. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., & Misra, I. (2022, November). Detecting twenty-thousand classes using image-level supervision. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX (pp. 350-368). Cham: Springer Nature Switzerland.

7. OSHA Strategic Partnership Program (OSPP) | Occupational Safety and Health Administration. (n.d.). OSHA Strategic Partnership Program (OSPP) | Occupational Safety and Health Administration. Retrieved April 6, 2023, from https://www.osha.gov/partnerships