Impact of Different Distance Metrics for Deep Learning-based multiple Object Detection and Tracking

Praful V. Barekar[1], Kavita R. Singh[2*], Gajanan Tikhe[3], Chandu Vaidya[4], Roshni S. Khedgaonkar[5]

[1]Assistant Professor, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur
[2]Professor, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, singhkavita19@gmail.com
[3]Assistant Professor, Department of Computer Engineering, Bajaj Institute of Technology, Wardha
[4]Assistant Professor, Department of Computer Science and Engineering, S. B. Jain Institute of Technology, Management & Research, Nagpur
[5]Assistant Professor, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur

Object detection and tracking are fundamental tasks in computer vision with numerous applications ranging from surveillance to autonomous vehicles. In recent years, deep learning methods have revolutionized these tasks, enabling accurate detection and tracking of objects in complex scenes. Central to these methods is the notion of measuring similarity or distance between objects for tracking purposes. In this study, we investigate the impact of different distance metrics on the performance of deep learning-based multiple object detection and tracking systems.By understanding the trade-offs associated with each metric, researchers and practitioners can design more effective and reliable tracking systems for real-world applications. Ultimately, our study contributes to advancing the state-of-the-art in computer vision by providing valuable guidance on the selection of distance metrics for multiple object detection and tracking tasks.

Keywords: Object detection, Object tracking, Computer vision, Deep learning, Distance metrics, and Similarity measures, Surveillance.

1. Introduction

Detecting and tracking multiple objects are considered to be the most ubiquitous and challenging tasks in computer vision[1]. The increase in the need for video surveillance allowed developers to create more efficient object tracking algorithms in real-time. Multiple

object detection and tracking has an extensive range of applications include pedestrian detection for self-driving cars, agricultural crop monitoring, ball tracking in sports, and other similar uses.

Multiple object detection and tracking applications have played a significant role in the recent times. Object detection and tracking are not only important in security-related systems, but also, in many other areas like crowd control, anomaly detection, automated surveillance systems, and human computer interaction[2][3]. The wide range of applications aims at the development of novel and efficient algorithms to detect and track one or more effective targets from a particular frame or video data. This system can be used in various applications like traffic checking, robot vision and activity, anomaly detection, crowd counting, automatic visual surveillance, and video correspondence.

The main objective of a multiple object detection and tracking system is to find all object instances of one or more specified object classes, regardless of scale, position, pose, camera view, partial occlusions, and lighting conditions. It can detect multiple objects of different classes or a single class can be detected from multiple viewpoints. Object detection and tracking are used in a variety of applications like robotics for service robots, human-computer interaction, security in terms of recognition and tracking, and transportation, in specific[4]. The processing speed, occlusion, rotations, and identification under posture changes are the criteria for each of these applications. Many artificial vision applications, such as object tracking and action detection, are proven to be reliable features extracted from a pretrained deep network. As the need for multiple object tracking increases, many challenges arise beyond the necessity for image classification. The problem of occlusion is not effectively treated since it leads to incorrect tracking by detecting the initially tracked object with an identity that changes in the next few frames due to the occluded view[5]. A prominent requirement in object detection and tracking systems is that they should detect and track the objects in the shortest period of time with accuracy. It is critical to make improvements in the training and tracking speed of the system in real-time object tracking methods. Localization of objects in a real-time video is the true challenge behind object detection[6].

Real-time video processing using object and facial recognition is claimed as a vital technology in surveillance systems. It involves a series of steps that include decoding, computation, and encoding. Decoding is required to bring back a compressed form of video into its original format. Computation is the process of executing a particular operation on the original video frame[7]. Reverting the processed frame to its original compressed state is the encoding process. The goal of implementing a video processing system is to complete all these steps as quickly and as accurately as possible. This series of steps is enhanced by using techniques like computer vision, machine learning, and deep learning in the domain of video processing

Background Subtraction Background subtraction (BS) refers to the process of creating a foreground mask, which is a binary image that includes the pixels of moving objects in a scene, utilising static cameras. To determine the foreground mask, BS subtracts the current frame from a reference frame called the "Background Image" or "Background Model." This reference frame contains the static scene elements or, more broadly, anything that can be considered background based on the scene's characteristics[8]. The main purpose of this method is to find what's in the forefront of a video frame, which is called foreground

identification[9]. Furthermore, since the particular objects in the photographs' foregrounds are of interest, further analysis is focused on a segment of the sequence; as a result, knowledge of the complete contents is not necessary for many applications. Currently, every detection approach relies on first establishing the image's backdrop and then identifying any changes that occur therein[10]. It could be somewhat difficult to provide the correct background when there are shapes, shadows, and moving parts in it. The underlying premise of all background definition schemes is that the stationary items' colour and intensity can alter throughout time[11].

Preprocessing The majority of computer vision systems relied on basic spatial and temporal smoothing to decrease camera noise during initial processing stages[12]. Images captured in harsh weather, including snow or rain, may have random background noise removed using smoothing. To reduce the processing speed of data, real-time systems often use frame-size and frame-rate reduction. If the cameras are in motion or if there are several cameras positioned in different areas, picture registration between successive frames or between multiple cameras is necessary before background modeling[13]. Another crucial aspect of preprocessing to consider is the data format used by the chosen background removal method. A single scalar integer represents each pixel's luminance intensity, which is the primary focus of most algorithms[14][15]. A single scalar value represents each pixel's luminance intensity, which is the primary focus of most algorithms. Nevertheless, background removal using colour photos is becoming more common, regardless of whether the image is in the RGB or HSV colour system.

Background Modeling The fundamental purpose of any method for background removal is to model the backdrop. Developing a backdrop model that can detect all moving items of interest while being unaffected by changes in the surrounding environment has been the primary focus of most studies. There were two categories of background modelling techniques: recursive and non-recursive. Using a sliding window technique, a non-recursive algorithm may estimate the background[16].

Each pixel's temporal fluctuation was used to estimate the backdrop image, which was then stored in a buffer of the previous L video frames. A non-recursive approach's adaptability stems from the fact that it is not reliant on history beyond the frames stored in the buffer. A few examples of algorithms that do not recurse include the non-parametric model, linear predictive filter, median filter, and frame difference. For recursive methods, it does not maintain a buffer for background estimate. No, instead, they update a single background model repeatedly using each input frame. Consequently, the current background model could be affected by input frames from a long time ago[17]. Despite recursive processes' lower storage requirements, any error in the background model might remain for a much longer period as compared to non-recursive approaches.

Foreground Detection In order to generate a binary candidate foreground mask, it searches the video frame for pixels that the background model cannot completely explain. forefront detection checks the input video frame against the backdrop model to find pixels that might be in the forefront. The most common way to identify foreground objects is to compare the input pixel to the equivalent background estimate and see whether there is a substantial difference.

2. Object Classification

Classification of the objects should follow feature extraction. Making ensuring all photos are categorised according to their specific sectors or groups is the main objective of image classification. Classification is easy for humans but quite difficult for computers owing to differences in size, shape, colour, and other environmental factors. There are two major varieties of object classification methodologies, namely;

a. Shape based object classification

b. Motion based object classification.

Object Classification Based on Shapes Many approaches have relied on shapes to identify and locate items in real-world photos because of how useful they are. Model building and object detection and identification are the two main components of shape-based approaches. An object's form may be efficiently and effectively captured by wrapping the shapes of a continuum manipulator around it, according to the primary notion. Objects may be categorised into three main groups: humans, animals, and vehicles. Each frame underwent a classification process to ensure accurate temporal analysis.

Object Classification using Motion Data A quality is shown by the motion of non-rigid articulated items; hence, this has been used as a cue for item categorization. Optical flow may also be very useful for object categorization. Stiffness and periodicity may be studied using moving residual flow. People and other non-rigid objects had periodic components and larger average residual flows than rigid objects. In order to achieve classification, motion-based classification offers a computationally efficient and dependable alternative to relying on the objects' spatial primitives.

Combining picture categorization with picture localization is what object detection is all about. The method involves accurately anticipating the locations of several things in a picture and then categorising each object. It finds the object's category and where it is in the provided picture. Bounding boxes around the detected and expected items are added to a picture when object detection is applied to it[18]. Finding all instances of one or more specified object classes in any given scene, independent of size, location, pose, camera view, partial occlusions, or lighting conditions, is the primary goal of a multiple object recognition and tracking system. Either it can identify things of distinct classes simultaneously or it can detect objects of the same class from different perspectives. Numerous fields make use of object detection and tracking algorithms, including transportation, human-computer interaction, robotics (service robots), and security (identification and tracking). These uses all have specific requirements for processing speed, occlusion, rotations, and recognition in the face of changes in posture. Object tracking and action detection are just two examples of the many successful AI vision applications that have relied on characteristics retrieved from a pre-trained deep network. Object detection algorithms that use a two-stage process first propose regions to be classified and then place themselves. The two-stage method is often more accurate. In terms of accuracy, the two-stage algorithm is superior, but it is also more complex[7][8]. There are essentially two parts to it. After running initial tests, eliminating any positive samples, and creating regions of interest (RoIs), the algorithm moves on to stage two, where it refines the locations

and classifications of the RoIs it established in stage one. Many popular algorithms include of two stages: Faster R-CNN, R-FCN, FPN, and so on. Combining Fast R-CNN with the Region Proposal Network (RPN) suggested by the R-CNN algorithm creates an end-to-end learning network that is both faster and more accurate than Fast R-CNN alone[10].

3. Moving Multi-Object Detection And Tracking In A Congested Environment

The tracking-by-detection approach establishes a connection between the current recognition assumptions and the previously evaluated item trajectories by comparing the objects' appearance or motion. A Pearson Similarity centred Kuhn-Munkres (PS-KM) method was suggested for effective item recognition and tracking in a complicated setting. We were able to filter out the incorrect frame data after the conversion stage by using the Entropy-like Divergence-based Kernel K-Means Algorithm (EDK2MA) for background removal[11]. In order to remove unnecessary data, the EDK2MA first sorts the frames into k clusters, groups them based on common properties, and then removes extraneous features. In order to confirm that the EDK2MA was accurate, it was compared to the KMA, K-Medoids, and Fuzzy C-Means (FCM). Attention was given to precision, f-measure, accuracy, and recall[12]. The frame's properties were further reduced after extraction using Information Gain centred Singular Value Decomposition (IG-SVD), which simultaneously reduced the dimensionality of the features from higher to lower. Information Gain (IG) calculates the crucial class label and evaluates the information gain of each attribute separately. A method called MRNN (Modified Recurrent Neural Network) was then used to categorise the entities[13].

4. Proposed Methodology

The video-centered OD and target re-identification are strongly related to multiple object tracking MOT. With the representational capacity that deep learning brings, the best OD and tracking approaches revolve on deep neural networks. However, improving MOT's performance in challenging real-world scenarios remains a challenge. Here, tracking techniques and moving multi-OD were suggested using a DL model in conjunction with PS-KM. The suggested solution has the potential to handle occlusion, shadows, and camera jitter related to objects. The process begins with collecting the input data (videos) and converting it to VF from the publically available dataset. The next step is to remove the frame's backdrop using the Entropy-like Divergence-based Kernel K-Means Algorithm (EDK2MA). The characteristics are then retrieved from the frames with the backgrounds removed[14]. The next step is to use IG-SVD to reduce the features, which will convert the higher-dimensional characteristics to lower-dimensional ones. Afterwards, MRNN is used for OD and classification. Consequently, the object tracking system is provided with the output of the classifier, which includes the identification of different object types. Next, the items that have been detected are tracked using the PS-KM method.

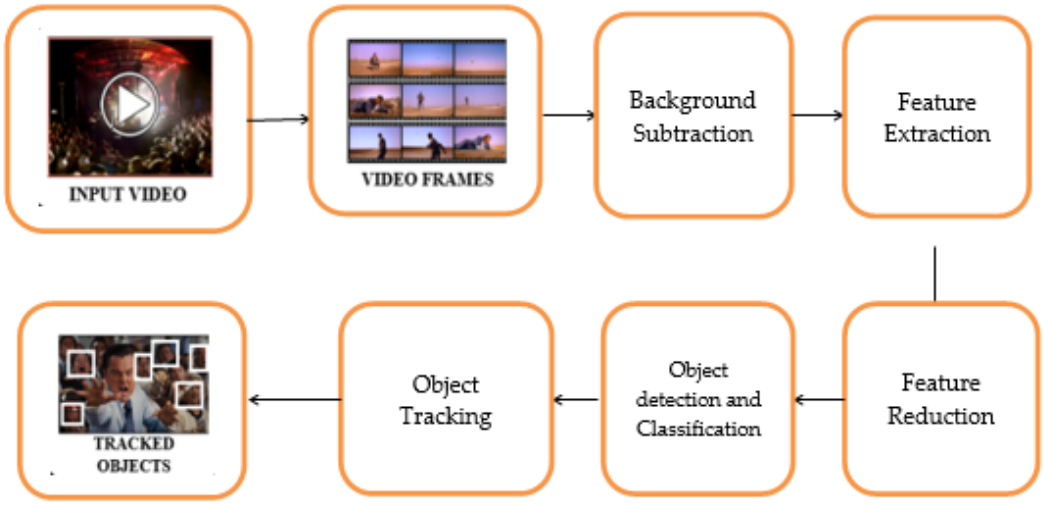

Figure 1. Proposed system architecture

From the publically accessible dataset, the input videos (V) are initially gathered. Next, the amassed data is converted into VF (Vf) which are presented by,

$$Vf = \{V1\ ,2\ , \cdots\cdots\ , VN\}$$

The number of VF is represented by N. The VF's background subtraction is then carried out using an EDK2MA. By dividing the frames into k clusters and then grouping the clusters based on identical characteristics, the EDK2MA filters out the useless information while simultaneously eliminating the unrelated features. The generic K-Means Algorithm (KMA) was defined by the Euclidean distance (ED), and it makes use of the Gaussian Kernel in cluster construction. Because of its sensitivity to outliers and noise, ED makes it difficult to get satisfactory results with complicated non-convex data. To get around this, entropy-like divergence is introduced into traditional KMA by combining the Jenson-Shannon/Bregman divergence with a convex function and its mercer kernel function, which is known as the entropy-like divergence-based kernel (EDK). The algorithm's anti-noise resilience is better, and the segmentation accuracy of the clustering process is improved. A combination of EDK and KMA is known as EDK2MA. The stages involved in EDK2MA are explained in further detail[15].

Iterative feature reduction is followed by object detection and classification using MRNN. The RNN-based method makes use of sequential data and the previous output to forecast the upcoming output; the method is also known as RNN. This network has a memory that stores all the data it has seen so far. Accordingly, the selection of the Activation Function (AF) is crucial to the generalisation performance and training stability of the neural network. Even though sigmoid functions are widely used and accepted in RNNs, they may have significant problems with gradient diffusion. The saturation problem in AFs is the main cause of this phenomenon; inputs close to or in the saturation area have very little effect on the neurons' outputs[16]. To circumvent this drawback of conventional RNNs, an AF is included into the

RNN. In order to provide more stable training, a new AF is designed to recognise the need of having a limited output range. A new AF has been integrated into the common RNN, and the result is known as MRNN[19].

The choice of distance metrics significantly impacts the performance of deep learning-based multiple object detection and tracking systems:

When retrieving object identities after long-term occlusions, the DeepSORT algorithm's usage of the cosine distance measure shines. When objects are obscured for long periods of time, the cosine distance takes appearance into account, making it more discriminative than depending just on motion signals. Important metrics for gauging tracking performance as a whole are MOTP (Multiple Object Tracking Precision) and MOTA (Multiple Object Tracking Accuracy). In contrast to MOTP, which evaluates the tracker's localization accuracy, MOTA evaluates its object detection, false positive/negative avoidance, and object identity consistency capabilities. Enhancing tracking accuracy, particularly in difficult situations with frequent occlusions, is achieved by the use of enhanced association metrics, such as DeepSORT, which integrate motion and appearance data. For deep learning-based multi-object tracking and detection systems to function at their best, distance metrics must be carefully chosen and evaluated.

5. Conclusion

When it comes to tracking algorithms that use deep learning for multiple object recognition and tracking, the effect of various distance measures is substantial. To maximise the efficiency of deep learning-based tracking systems for multiple objects, it is crucial to choose suitable distance metrics (e.g., DeepSORT's cosine distance) and evaluation metrics (e.g., MOTA and MOTP) to guarantee accurate and dependable tracking outcomes.

Conflicts of Interest

The authors declare that they have no competing interests.

References

1. Duan K S., Bai L., Xie H., Qi Q., Huang & Q. Tian, 2019, ‘CenterNet: Keypoint Triplets for Object Detection,’ 2019 IEEE/CVF International Conference on Computer Vision (ICCV).pp.6568-6577
2. Elhoseny M, 2019, ‘Multi-object detection and tracking (MODT) machine learning model for real-time video surveillance systems’, Circuits, Systems and Signal Processing, vol. 39, no. 3, pp. 611-630
3. Eltantawy A and Shehata M S., 2019, ‘An Accelerated Sequential PCP-Based Method for Ground-Moving Objects Detection from Aerial Videos,’ in IEEE Transactions on Image Processing, vol. 28, no. 12, pp. 5991-6006
4. HyochangAhn& Han-Jin Cho,2019, ‘Research of multi-object detection and tracking using machine learning based on knowledge for video surveillance system’, Personal and Ubiquitous Computing, vol. 26, no. 2, pp. 385-394
5. Kalsotra R. &Arora S. 2022, ‘Performance analysis of U-Net with hybrid loss for foreground

detection', Multimedia Systems' pp.1-18.
6. .Upschulte E., Harmeling S., Amunts K., Dickscheid T., 2022. 'Contour proposal networks for biomedical instance segmentation', Med Image Anal. Vol. 77
7. AzzedineBoukerche&ZhijunHou, 2021, 'Object detection using deep learning methods in traffic scenarios' , ACM Computing Surveys, vol. 54, no. 2, pp. 1-35.
8. Chen P Y., Chang M C., Hsieh J W & Chen Y S, 2021, 'Parallel Residual Bi-Fusion Feature Pyramid Network for Accurate Single-Shot Object Detection,' in IEEE Transactions on Image Processing, vol. 30, pp. 9099-9111.
9. Li W., Chen Z., Li B., Zhang D & Yuan Y, 2021, 'HTD: Heterogeneous Task Decoupling for Two-Stage Object Detection,' in IEEE Transactions on Image Processing, vol. 30, pp. 9456-9469
10. Fang J, X., Tan & Y Wang, 2021, 'ACRM: Attention Cascade RCNNK with Mix-NMS for Metallic Surface Defect Detection,' 2020 25th International Conference on Pattern Recognition (ICPR).
11. Law H., Deng J. 2020, 'CornerNet: Detecting Objects as Paired Keypoints' Int J Comput Vis 128, 642–656.
12. Liu C., Yao R., Rezatofighi S H., Reid I & Shi. Q, 2020, 'Model-Free Tracker for Multiple Objects Using Joint Appearance and Motion Inference,' in IEEE Transactions on Image Processing, vol. 29, pp. 277-288.
13. Mustafa R. J., Younis YM., Hussein HI.&Atto M., 2020, 'A New Video Steganography Scheme Based on Shi-Tomasi Corner Detector', in IEEE Access, vol. 8, pp. 161825- 161837.
14. Tan M., Pang R., & Le Q V., 2020 'EfficientDet: Scalable and Efficient Object Detection', IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
15. Xiong F., Zhou J and Qian Y., 2020, 'Material Based Object Tracking in Hyperspectral Videos,' in IEEE Transactions on Image Processing, vol. 29, pp. 3719-3733.
16. Ingole, K., &Padole, D. (2023). Design Approaches for Internet of Things Based System Model for Agricultural Applications. In 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP) (pp. 1-5). Nagpur, India
17. Pattnaik, Sonali, Hota, SwetaLeena, Kumar, Arya, Hota, Anish Raj &Kiran, Prabha(2024) Awaking deep emotional needs of consumers based on big data and emotional analysis, Journal of Statistics and Management Systems , 27:1, 105–119, DOI: 10.47974/JSMS-1196
18. Pujari, Purvi, Arora, Monika, Kumar, Anuj&Pandey, Anoop(2024) Understanding factors influencing technical inertia in family-run SMEs : A study on technology adoption challenges, Journal of Statistics and Management Systems , 27:1, 155–168, DOI: 10.47974/JSMS-1208
19. Johri, P., Khatri, S.K., Al-Taani, A.T., Sabharwal, M., Suvanov, S., Kumar, A. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. In: Abraham, A., Castillo, O., Virmani, D. (eds) Proceedings of 3rd International Conference on Computing Informatics and Networks. Lecture Notes in Networks and Systems, vol 167. Springer, Singapore. https://doi.org/10.1007/978-981-15-9712-1_31