OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21616058

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Articles

Volume-06|Issue-04|2026

Unsupervised Learning based Detection and Tracking of Objects in Videos

Ravina Rodrigues[1] and Terence Johnson[2]

[1]Research Scholar, Department of Computer Science and Engineering Research Centre, Mangalore Institute of Technology & Engineering, Visvesvaraya Technological University, Belagavi, India.
[2]Research Supervisor, Department of Computer Science and Engineering Research Centre, Mangalore Institute of Technology & Engineering, Visvesvaraya Technological University, Belagavi, India.

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Rodrigues, R. & Johnson, T. (2026) Unsupervised Learning based Detection and Tracking of Objects in Videos. *Indiana Journal of Multidisciplinary Research, 6*(4), 470-475.

Abstract: Over the last few years, a key area of study has been aimed towards gaining perception and monitoring of objects in the fields of machine vision and pattern detection. Object detection algorithms are adopted in multiple sectors, including surveillance, medical imaging, obstacle avoidance, retail, and navigation. The main purpose of object tracking is to assess the position of an object in images in a continuous manner. Different learning techniques including supervised and unsupervised methodologies can be used to track and monitor objects in videos. In spite of its significance, there exists limited research on unsupervised object detection techniques owing to the challenges presented by it. This paper explores the unsupervised learning approaches towards object tracking, especially focusing on clustering algorithms of K-Means and Mean Shift for detection and tracking of objects in videos.

Keywords: Object Detection, Video Object Tracking, Unsupervised Learning, K-Means, Mean Shift Algorithm.

INTRODUCTION

Object detection has emerged as a fundamental component of the computer **vision** ecosystem. Identifying objects in diverse scenes with no human intervention has been a longstanding objective of computer vision researchers. The concept of object monitoring in videos is rooted in the understanding of the detection of objects in **images**. Tracking objects in videos is the technique of detecting one or many objects in a frame and then monitoring their movements throughout the video. A video **comprises** many frames and techniques used in tracking objects in videos, examine every frame in order to detect an object or many objects. After detection, a bounded box is drawn around the objects of interest. Performing this process on every frame allows for good tracking of the objects throughout the video. From efficiency to **security** in industries and at home, object detection in videos is becoming prevalent in various applications [1]. Some of the applications include monitoring methodologies employed for monitoring suspicious activities or unauthorized access, detection and tracking leading to the discovery of correlations between buyer and goods which can be beneficial for analyzing shopping behavior, tracking inventory and improving customer experience [2], detection of pedestrians and obstacles, tracking of other vehicles for seamless and collision-free driving in case of autonomous vehicles [3], screening of objects for assistance in diagnosis for discovering and evaluating **anomalies** in medical scans [4].

Fig. 1. Applications of object detection

Innumerable methods for object detection in videos were put forward and a large amount of them exploit the supervised techniques. The conventional supervised learning algorithms are constantly used in areas such as computer vision, rely on human guidance to direct the machine by annotating inputs with desired outputs for learning [5]. Extensive labeled training data plays a vital role in the success of supervised learning models in visualization tasks; however their creation requires significant resources and time, which limits their availability. To overcome this problem, unsupervised learning techniques are required. While there has been expeditious advancement in supervised object detection, the unsupervised variant is still largely unexplored [6]. In this paper we aspire to demonstrate how clustering algorithms can be applied in videos for detection and tracking of objects.

RELATED WORKS

Algorithms in plenty have been devised in the past, for the screening and monitoring of objects in videos.

Supervised Object Detection and Tracking

Algorithms that work by learning under supervision are extensively used in object monitoring particularly to train classifiers in order to differentiate between objects and their background, as well as for detecting and discovering the exact location of objects in images and videos. These algorithms utilize the labeled training data to interpret the relationships between input features and the corresponding output, which comprises the detected object. The general steps involved for the detection and tracking of objects are as follows:

1. Collection of the dataset.
2. Preprocessing the dataset.
3. Training the model using supervised learning algorithms.
4. Evaluating the model.
5. Testing the model for detecting an object or an array of objects in the input video.
6. The output generated will have bounding boxes around the objects within the video.

Supervised object detection algorithms can be broadly classified as one-stage object detection algorithms and two-stage object detection algorithms. The one-stage object detection algorithm analyzes the whole image in a single pass to make predictions about the presence and location of objects in images thereby making them computationally efficient. The most viral and widely used one-stage object detection and tracking algorithm is You Only Look Once (YOLO), the latest version of which is YOLO v11. The YOLO algorithm takes an image as input and then splits that image into a grid of cells. If the object center lies inside a grid cell, then it is the responsibility of that grid cell for the detection of objects by putting a bounding box around it. YOLO

***Corresponding Author:** Ravina Rodrigues

examines the entire image at once, making it more competent than other algorithms in terms of speed and precision [7]. The network architecture of YOLO v11 comprises three primary modules: the backbone, neck and head. These modules work together to create a complete object detection pipeline from preprocessing the input to generating the final output [8]. To effectively detect and locate objects, a two-stage object detection algorithm performs two passes over the input image. During the first pass, a set of possible candidate regions of potential objects and its corresponding locations are generated. The second pass aims to refine these regions to determine the final output. Despite its greater accuracy as compared to one-stage object detection, it exhibits a higher computational overhead. A dominant two-phase object detection algorithm is Regions with Convolution Neural Networks (R-CNN). R-CNN is a family of two-phase object detection algorithms that utilize convolution neural networks to recognize and localize objects within images. Models such as R-CNN (Regions with Convolutional Neural Networks) and its successors—Fast R-CNN and Faster R-CNN—have significantly improved both the performance and efficiency of object detection [9]. In R-CNN, a selective search algorithm scans the input image for potential features, generating approximately 2,000 region proposals. Subsequently, we apply a convolutional neural network to these proposed regions. The result from every convolutional neural network goes as an input into a support vector machine, which employs linear regression to refine the bounding box of the feature [10], as shown in Figure 2.

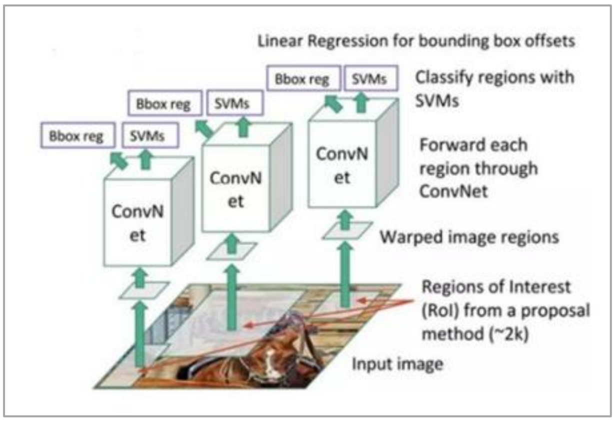

Fig. 2. R-CNN

Unsupervised Object Detection and Tracking

The objective of unsupervised object detection is inherently complex, often needing considerable efforts to rival the efficiency of its supervised counterparts. Existing unsupervised techniques for the detection of objects can be roughly divided into Contrastive Learning, Clustering, and Self-Supervised Learning. Contrastive learning is a form of unsupervised learning that focuses on instance-based learning. It identifies correlations among similar instances and separates instances that are not alike. Contrastive learning operates in feature space by reducing the proximity of analogous instances while increasing the separation among the contrasting ones [11]. One widely adopted unsupervised learning strategy is clustering which divides objects into different groups based on similarity. It groups objects in such a manner that the objects within each set are similar among one another than those in other groups. Clustering categorizes data into meaningful groups, known as clusters, with the purpose of capturing the intrinsic structure of the data [12]. Several clustering algorithms can be employed for segmenting analogous data points representing objects in an image. Function clustering, K-means clustering, hierarchical clustering etc. are among the clustering approaches currently employed in object detection tasks [13]. The fundamental difference between contrastive learning and clustering lies in their approach towards learning from unlabeled data. Contrastive learning learns by pulling similar data points together and pushing dissimilar ones apart, while clustering aims to assemble data into clusters on similarity. Another technique of unsupervised machine learning is self-supervised learning (SSL). The SSL paradigm is a technique that learns from unlabeled data by creating its own supervisory signals. SSL algorithms focus on learning distinguishing representations from extensive unlabeled datasets thereby bypassing the dependence on human annotations [14]. This allows the model to determine appropriate patterns which can then be transferred to other tasks, making it a powerful technique for various applications.

METHODOLOGY

Object detection algorithms comprise two core components: detection and tracking. Object detection involves the discovery and location of objects in an individual frame, while tracking extends this to subsequent frames, maintaining the entity's identity and position over time.

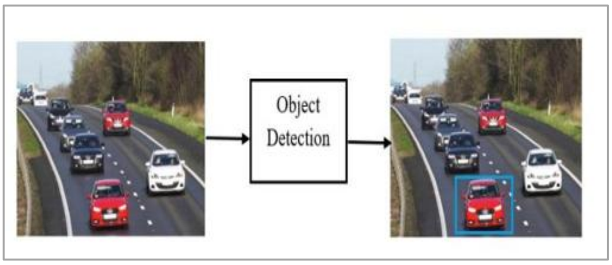

Fig. 3. Object Detection

Moving object detection typically utilizes time-associated information gathered from sequential frames either by detecting frame differences, establishing a reference background and comparing it with the present frame or identifying zones with frequent movement [15]. After the object has been detected, the next step is the tracking function, which locates the object correspondence in the subsequent frames. The entire section of the image occupied by the object can be provided using the object tracker. The approaches of object detection, and building consistency among its instances across frames can be processed independently or in a joint manner. In the initial scenario, an object detection algorithm is able to identify a set of possible entity regions in every frame, and the tracker then corresponds to these objects across the various frames. In the second case, the entity's region and correspondence are simultaneously estimated by iteratively updating its spatial location and boundaries obtained from previous frames. The block diagram for a general object detection and monitoring system is shown in Figure 4 below.

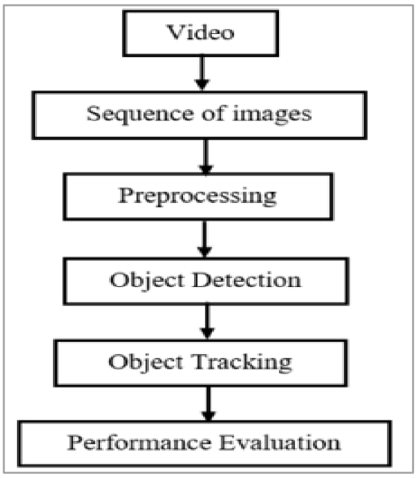

Fig. 4. General block diagram of Object Detection and Tracking System

CLUSTERING ALGORITHMS FOR VIDEO DETECTION AND TRACKING

In order to implement object detection and tracking, the algorithm examines a sequence of video frames and produces the motion of targets between the frames as output. There are a broad range of algorithms to perform these tasks. We have used the clustering algorithms that are both partitioning based and density based to perform detection and tracking of a single or multiple objects in videos.

Partitioning based Clustering

Partitioning based clustering is the methodology in machine learning that segments a dataset into sets of analogous data instances termed as clusters. Centroids, which are the core of every cluster, are employed to minimize the total distance between data instances and their corresponding cluster center. This ensures that the data instances are consistent with the center of the cluster thereby ensuring a maximum separation between different clusters. K-Means belongs to the family of partitioning-based clustering techniques, that is employed in several studies for comparative analysis with other algorithms [16]. It classifies a set of n-dimensional points into k clusters. k is predetermined which represents the number of clusters [17]. The K-Means algorithm could be viewed as a two-phase approach. The first phase involves the identification of k centroids that have been generally selected. The Euclidean distance metric is calculated as a distance measure. New centroids are computed during phase two of the algorithm [18]. This is reiterated until the clusters no longer change. For object detection, K-Means is used to cluster similar pixels together, which can then be applied to detect objects in an image. The algorithm for detection and tracking of objects using K-Means works as follows:

1. Input to the algorithm is a video.
2. A sequence of images is generated from the video.
3. Set the initial values for k number of clusters and centers.
4. Compute the Euclidean distance d from the center to each pixel p(x,y) in the image.
5. According to the calculated distance, assign all the image pixels to the closest center.
6. New centers are recalculated after assigning all the pixels.
7. Repeat the algorithm until the result falls within the predefined tolerance level.
8. Restore the image by reorganizing the clustered pixel data into its original spatial dimensions.
9. The sequence of images is then combined to detect and track moving entities in a video.
10. K-Means clustering is universally acknowledged for its efficiency and high speed, making it suitable for large-scale datasets [19]. However, it also has limitations such as sensitivity to outliers, assumption of only spherical and equal-sized clusters along with the correct choice of the number of clusters (k).

Density based clustering

Clusters are formed in density-based clustering by grouping data instances that fall inside dense regions. It detects clusters in tightly packed regions which are isolated by areas of low point concentration. These algorithms are useful for finding clusters of random shapes and managing outliers. The basic principle behind these algorithms is to identify areas where the data points are concentrated and where they are separated, empty or sparse. Instances not belonging to a cluster are labeled as noise. Among density-based clustering techniques, the most commonly adopted algorithm is the Mean Shift algorithm. The method iteratively shifts the data entries to converge at local peaks of the estimated density distribution. Fukunaga and Hostetler developed this approach in 1975 [20]. Mean Shift algorithm discovers clusters by continuously refining data instances to align towards the densest region in the feature space. After this iterative process, data instances assemble around the modes of the data distribution forming the clusters. In object detection, the Mean Shift algorithm detects the object by repeatedly adjusting a window to follow the object's color distribution. Mean Shift analysis derives the target item that is most similar to a given target model, and thus tracks entities with changing scale [21]. The algorithm for the detection and tracking of objects using Mean Shift works as follows:

1. Input to the algorithm is a video.
2. A sequence of images is generated from the video.
3. Consider each pixel p(x,y) of the image as a potential candidate for the cluster center.

4. A window also known as the radius is defined around every pixel. A mean of all the pixels within this radius is computed.
5. Each pixel is then shifted to this mean value thereby moving the pixels towards high-density regions.
6. Repeat Steps 4 and 5 of the algorithm until the result falls within the predefined tolerance level. Point stabilization occurs when no further shifts are detected, which indicates convergence to a local peak of the density function.

After the completion of Step 6, all the pixels will have been assigned to the clusters. Reshape the cluster pixels into an image. The sequence of images is then amalgamated together to detect and monitor moving entities in a video. Advantages of the Mean Shift algorithm may include low computational cost, rotation invariance and coping with partial occlusions, however it may present a dissatisfactory performance when the target moves too fast, or changes in size or pose.

IMPLEMENTATION AND RESULTS

The results of detecting and tracking objects in videos using the K-Means algorithm and the Mean Shift algorithm are shown below. Nine hundred and thirteen (913) frames were generated from the video. The Mean Shift algorithm performed significantly better, doubling the precision compared to the K-Means algorithm. Figure 5.a shows the original image. Figure 5.b shows object tracking using the K-Means algorithm at different frames while Figure 5.c shows object tracking using the Mean Shift algorithm at different frames.

Fig. 5. a Original Image frames extracted from videos

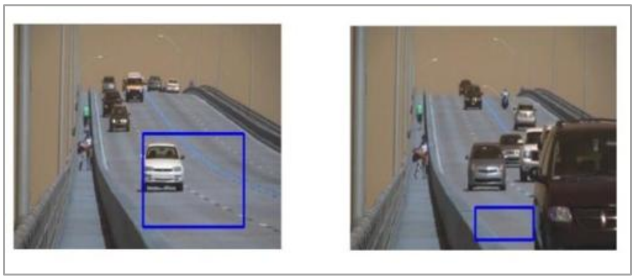

Fig. 5. b Entity monitoring using K-Means Algorithm

Fig. 5. c Entity monitoring using Mean Shift Algorithm

CONCLUSION

From our experiments, we conclude that the Mean Shift Algorithm performs better than the K-Means algorithm while detecting and tracking objects in videos.

REFERENCES

1. Vahab, A., Naik, M. S., Raikar, P. G., & R., P. S. (2019). Applications of object detection system. *International Research Journal of Engineering and Technology (IRJET), 6*(4).
2. Nguyen, K., Le, M., Martin, B., Cil, I., & Fookes, C. (n.d.). *When AI meets store layout design: A review*. *Artificial Intelligence Review*.
3. Zhao, D., Fu, H., Xiao, L., Wu, T., & Dai, B. (n.d.). Multi-object tracking with correlation filter for autonomous vehicle. *Sensors* (Special Issue: Signal Processing and Visual Computing).
4. Ritter, F., Boskamp, T., Homeyer, A., Laue, H., Schwier, M., Link, F., & Peitgen, H.-O. (2011). Medical image analysis. *IEEE Pulse*.
5. Audet, S. (2006). *Unsupervised learning for object recognition*. ECSE 626: Statistical Computer Vision.
6. Wang, Y., Chen, Y., & Zhang, Z. (2022). 4D unsupervised object discovery. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.
7. Devadkar, P., More, P., Gupta, R., Jaiswal, P., & Patil, S. (2023). Drowsiness detection system. *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 10*(6).
8. He, L.-H., Zhou, Y.-Z., Liu, L., Cao, W., & M., J.-H. (2025). Research on object detection and recognition in remote sensing images based on YOLOv11. *Scientific Reports*.
9. Patkia, Y., & Gawade, T. (n.d.). *Object detection using Faster R-CNN.*
10. Du, K., & Bobkov, A. (2022). An overview of object detection and tracking algorithms. In *Proceedings of the 15th International Conference on Intelligent Systems (INTELS'22)*.
11. Hu, H., Wang, X., Zhang, Y., Chen, Q., & Guan, Q. (2024). A comprehensive survey on contrastive learning. *Neurocomputing, 610*.
12. Talwar, P., Shubham, & Kaur, K. (n.d.). *Exploring clustering techniques in machine learning.*
13. Divakara, M. P., Trimal, K. V., Krishnan, A., & Karthik, V. (n.d.). Object detection using K-means

***Corresponding Author:** Ravina Rodrigues

clustering: A research. *International Journal of Recent Technology and Engineering.*
14. Gui, J., Chen, T., & Zhang, J. (n.d.). *A survey on self-supervised learning: Algorithms, applications, and future trends.*
15. Porikli, F., Shan, C., Xiang, T., & Gong, S. (n.d.). Object detection and tracking. In *Video Analytics for Business Intelligence* (Studies in Computational Intelligence).
16. Ju, & Xu, C. (2013). A new collaborative recommendation approach based on users clustering using artificial bee colony algorithm. *The Scientific World Journal, 2013.*
17. Farhang, Y. (2017). Face extraction from image based on K-means clustering algorithms. *International Journal of Advanced Computer Science and Applications, 8*(9).
18. Yin, H., Yi-xin, L., Jin-hui, L., & Shu-wei, X. (n.d.). An improved mean shift algorithm for vehicle tracking. *Advanced Materials Research.*
19. Wu, et al. (2013). Follow the leader: A centrality guided clustering and its application to social network analysis. *The Scientific World Journal, 2013.*
20. Fukunaga, K., & Hostetler, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory, 21*, 32–40.
21. Comaniciu, D., & Ramesh, V. (2000). Mean shift and optimal prediction for efficient object tracking. In A. Mojsilovic & J. Hu (Eds.), *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (pp. 70–73).

***Corresponding Author:** Ravina Rodrigues