**Indonesian Journal of Electrical Engineering and Computer Science**
Vol. 32, No. 3, December 2023, pp. 1805~1815
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v32.i3.pp1805-1815 

# Multiple object tracking using space-time adaptive correlation tracking

**Kusuma Sriram[1,2], Kiran Purushotham[3]**
[1]Department of Information Science and Engineering, M.S. Ramaiah Institute of Technology, Bengaluru, India
[2]Department of Computer Science and Engineering, R N S Institute of Technology Affiliated to Visvesvaraya Technological University, Bengaluru, India
[3]Department of Computer Science and Engineering, R N S Institute of Technology, Bengaluru, India

**Article Info**

***Article history:***

Received Jul 12, 2023
Revised Sep 26, 2023
Accepted Oct 18, 2023

***Keywords:***

DCCN
Deep learning
Multiple object tracking
Similarity map function
STACT

**ABSTRACT**

In application of tracking and detecting the suspicious activities, multiple object tracking (MOT) has been given fine attention due to its application as it provides the parallel task of identification and tracking of human. MOT ensures the identification and trajectory for each object frame as they interact, despite the changes in its appearance, occlusion and various other tasks involved. Recent adoption of deep learning has given a new perspective but still achieving high metrics remains a major issue to overcome such issues, this research work presents the integrated architecture of deep convolutional covariance networks (DCCNs) and space-time adaptive correlation tracking (STACT) algorithm with similarity map function (SMF). Moreover, in proposed work, DCCNs is utilized for feature extractions through each frame capturing the distinctive information, STACT is tracking approaches that utilizes the SMF for locating and tracking objects. SMFs are updated for any changes in human appearances and motion, also it deals with occlusion. Here the proposed model is evaluated on MOT17 and MOT20 dataset. Performance analysis is carried out through comparing the existing model and Integrated-DCCN achieves higher metrics.



***Corresponding Author:***

Kusuma Sriram
Department of Computer Science and Engineering, R N S Institute of Technology
Affiliated to Visvesvaraya Technological University
Bengaluru, India
Email: kusumas_12@rediffmail.com

## 1. INTRODUCTION

The task of tracking in video analysis presents a significant challenge due to the multitude of employment opportunities, this process entails the identification of a designated object, denoted by an enclosure, and allocating it to a distinct identifier that withstands across all frames within a particular sequence of images [1]. The significance of tracking is apparent in diverse fields, including but not limited to surveillance, self-driving technology, sophisticated driver assistance systems, behavioral analysis, motion forecasting, and particle transformation analysis. There are two fundamental categories of tracking research, specifically multiple object tracking (MOT) and single object tracking [2]. The MOT methodology considers object detection as prior knowledge, while the latter approach aims to detect and track an unfamiliar object based solely on the localization data acquired from the initial frame. In contrast to prior MOT methodologies, the kalman filter employs object motion and velocity states for surveillance objectives [3]–[5]. The intersection over union (IoU) metric enables the linkage of objects between successive frames. With the capacity to fulfill the prerequisites for real-time inference and demonstrate rapid inference durations, these two approaches face

challenges either in precisely monitoring objects that are obstructed or in motion through intricate patterns [6]. The integration of object appearance characteristics has led to notable improvements in modern tracking systems, effectively resolving the previously mentioned concerns. The proposed methodology initially involved an autonomous appearance embedding extractor that systematically performs operations on each bounding box that has been detected by an object detector [7].

The approach of tracking-by-detection has gained considerable attention in contemporary methods of MOT. The principal aim of this task is to identify and locate targets within individual frames, followed by the establishment of comprehensive trajectories by linking them. The effectiveness of these techniques has been proven in situations where there is a scarcity of target concentrations. The occurrence of detection failures is a prevalent concern in contemporary detectors, especially in scenarios characterized by dense target populations and occlusions, despite having undergone training on extensive datasets. Some research articles [5]–[7] have replaced crucial components, including detection modules and feature extraction modules, in the tracking-by-detection framework with convolutional neural networks (CNNs). The research activities have been primarily directed toward achieving comprehensive tracking tasks within a two-dimensional spatial context. The previously mentioned occurrence can be described as the effectiveness of CNNs in various computer vision implementations [8]. Although publicly available benchmark datasets have shown significant improvements in tracking performance, these methods demonstrate sub-optimal performance in indoor tracking scenarios, such as stage performances and scripted performances. In such situations, individuals, while encountering numerous obstacles amidst various objects, within the domain of deep learning, the task of MOT in the presence of occlusion entails the identification, surveillance, and forecasting of the trajectories of multiple entities across a sequence of frames, regardless of their total or partial occlusion [9].

Deep learning-based object detectors are frequently utilized in MOT scenarios involving occlusion to identify objects within each frame. The identified objects are monitored across frames by utilizing techniques that establish a correlation between identical entities in different frames based on their motion and visual characteristics. Wu *et al.* [10] the utilization of predictive models counters the estimation of an object's motion and subsequent prediction of its position in the event of occlusion, thus facilitating occlusion management. The restricted observations and advanced models exhibit the ability to recognize an entity or infer its presence through the utilization of additional contextual ones. The utilization of post-processing methods, such as trajectory refinement and false positive elimination, is a prevalent practice to improve the precision of tracking outcomes. The assessment of the effectiveness of such a system is typically carried out by employing metrics such as multiple objects tracking accuracy (MOTA) and multiple objects tracking precision (MOTP). The aforementioned metrics are utilized to assess the precision and accuracy of the object trajectories. The existing challenge and concerns [11] through the variability in the appearances of items, the unpredictability of object motion, the complexity of occlusion conditions, and the requirement for real-time analysis.

Numerous research gaps are present in this particular field, including the management of long-term occlusions, scalability improvement, real-time performance assurance, generalization enhancement to diverse scenarios, optimization of data association techniques, and closer integration of the detection and tracking phases. Numerous deep learning algorithms [12] that are presently accessible for MOT counter the challenges while managing occlusion scenarios, particularly those that are of a long-term nature. Object tracking is a challenging task for models as they often encounter difficulties in maintaining the tracking of an object or differentiating it from other entities when it is being occluded behind another object for a prolonged period. In situations characterized by a large number of entities, deep learning algorithms face challenges. As the number of items grows, the computational cost of these models tends to rise while their precision tends to decrease. Whilst some deep learning models demonstrate Zaech *et al.* [13] a notable level of precision in executing various tasks related to MOT, they frequently necessitate substantial computational resources and are incapable of operation in real time. The importance of real-time performance is not considered in various contexts such as video surveillance and autonomous vehicles. Several deep learning models have been devised for MOT. Nevertheless, their generalizability to various settings, objects, or surroundings might be constrained by their training and assessment of particular datasets. The issue described is a frequently encountered challenge through deep learning, and it is especially challenging when applied to MOT [14] due to the extensive range of possible scenarios that may be encountered. In current methodologies, the processes of identifying and tracking objects are frequently treated as separate stages that employ discrete models. The integration of detection and tracking functionalities in a system has the potential to yield significant advancements through mutual enhancement and refinement.

MOT, among other real-world uses, is essential to robotics, surveillance, autonomous vehicles, and human-computer interaction. Despite significant developments in this field, managing occlusions and ensuring real-time performance remain difficult. Numerous everyday technologies may be considerably improved to track numerous things in real time, especially when there are challenges. For instance, it may boost the efficiency of security systems, increase the safety and effectiveness of autonomous cars, and enable logical and

natural interactions between people and computers. As previously mentioned, there are several drawbacks to current deep learning architectures for MOT, including problems with occlusions, scalability problems, and a lack of real-time performance. Overcoming these constraints would be a big step forward for the discipline and may lead to new uses and opportunities. Many of the MOT models now in use do not adapt effectively to various situations, objects, or settings. MOT technologies would become more adaptable and broadly applicable if algorithms that are more resilient to these variances were to be created. Overcoming MOT's difficulties can improve deep learning altogether. For instance, it could result in novel neural network types, training strategies, or topologies. Beyond MOT, many other tasks and disciplines might benefit from the knowledge gained from this study. This research work develops an integrated architecture of deep learning based deep convolutional covariance networks (DCCNs) architecture and space-time adaptive correlation tracking (STACT) algorithm with similarity map function (SMF); proposed integrated architecture effectively tracks the multiple objects across the frames in given video sequence. Integrated-DCCNs (IDCNN) is evaluated considering two-challenged dataset i.e., MOT17 and MOT20 considering different metrics such as MOTA, higher-order tracking accuracy (HOTA) and so on to prove the model efficiency. MOT has recently been the subject of extensive research, whilst other MOT techniques, like multiple human tracking, are tailored for particular object types. Some tracking-by-detection techniques employ all of the object candidates in the sequence, whilst other techniques just use candidates up to that frame. While the latter are considered online techniques, the former are categorized as global approaches. Among the current trends in global MOT techniques are network flow optimization [9], graph-based clustering [15], multiple hypothesis tracking (MHT) [16], and bayesian filtering-based tracking [8]. To handle long-term fluctuations in track objects, some global MOT techniques [8] allocate detections to tracklets first (which are a mixture of matched detections in a few successive frames) and then assign tracklets to tracks. This is an alternative to directly associating detections of tracks.

Pinto *et al.* [17], employ a recurrent autoregressive network to calculate pair-wise costs using extensively learned individual re-identification scores and object localization data. To eliminate false negatives and calculate the pair-wise cost based on deep learning-based person re-identification score, Van-Nguyen *et al.* [18] suggested using the bounding boxes from both the object detector and the projected bounding box (from each track's history). Filtering of detections is done using maximum suppression with a cost definition dependent on track and detection confidence. Chen *et al.* [19] also employ a trained Siamese network for historical appearance-based matching in addition to motion and size-based matching, which special aids in lowering identity shifts during tracking are encountered. To calculate the total pair-wise cost between each detection and track, Song *et al.* [20] uses two deep networks: the spatial attention network (which uses a siamese architecture to compare detection-bounding boxes and track history) and the temporal attention network (which uses an-long-sort term memory (LSTM) architecture). Manually created version of [21] employing a histogram of gradients (HoG) and color names) for each track is unable to match a track or detection, the dual network is employed. To forecast the future location and apply the intersection over union and detection in the computation of association costs, Gao *et al.* [22] dynamically introduce sub-networks for each instance of a person. The detection is used as the positive sample and the areas around it as the negative sample by the authors of [23] when employing the discriminative appearance learning approach for each track. Additionally, they employ spatiotemporal matching based on item size and position, and they multiply these three measurements together along with the pair-wise cost

Liu *et al.* [24] to achieve high-performance online tracking, the suggested solution is based on the Gaussian mixture probability hypothesis density (GMPHD) filter, a hierarchical data association (HDA), and a mask-based affinity fusion (MAF) model. Segment-to-track and track-to-track linkages are the two types of associations found in the HDA. The GMPHD filter is used to calculate one affinity for position and motion, while the answers from single object trackers such as the kernelized correlation filter, SiamRPN, and DaSiamRPN are used to compute the other affinity for appearance. Simple score-level fusion techniques, such as MAF (min-max normalization), can be utilized to combine these two affinities. Upon creation of a graph, structure that can handle detection and track states simultaneously online. To do this, we use a fully trainable neural message-passing network for data association. This method radically increases track stability while offering a natural way to initialize the track and handle false positive detections.

You *et al.* [25] provide a unique inference-domain network evolution to improve the one-time MOT model's generalizability. To execute the one-time MOT task, we specifically create a spatial topology-based one-time network (STONet), where a self-supervision mechanism is used to encourage the feature extractor to learn the spatial contexts without any annotated input. A temporal identity aggregation (TIA) module is further suggested to help STONet lessen the negative impacts of noisy labels on the network's growth. Zaech *et al.* [13] based on linking each detection with an identity and a quantity characteristic, frame the MOT as a maximizing an identity-quantity posterior (MAIQP) problem, and then solve the two major issues that arise. The first step is the introduction of a local target quantification module, which counts the number of targets inside a single detection. Second, to settle the two properties, we provide an identity-quantity harmony

method. Based on this, we create a unique identity-quantity harmonic tracking (IQHAT) framework that enables multiple ID labels to be applied to detections that comprise multiple targets.

This research paper is organized into four sections. In the first section a brief description of the MOT methods, in the second section the related work described which discusses the existing approaches. In the third section, the proposed methodology is discussed and in the last section, the performance evaluation is discussed.

## 2. PROPOSED METHOD

MOT is a challenging task in computer vision, and researchers have developed various approaches to tackle this problem. One effective approach combines a DCCN with STACT and incorporates a SMF. The DCCN is a deep learning model that learns to estimate the spatial and temporal correlations between object features. It takes as input a sequence of frames from a video and extracts high-level representations using convolutional neural networks (CNNs). The DCCN then calculates the covariance matrix of the extracted features to capture the interdependencies between different objects in the scene. STACT is a tracking algorithm that uses SMF to locate and track objects over time. It utilizes the correlation between the target object and the search region in subsequent frames to estimate the object's position accurately. STACT adapts the SMF based on the appearance changes of the object and can handle variations in scale, rotation, and occlusion. To improve the tracking performance further, a SMF is introduced. This function computes a similarity map between the estimated object positions and the object candidates in the current frame. The similarity map assigns high values to the regions that are likely to contain the tracked objects based on appearance and motion cues. By incorporating the similarity map into the tracking framework, the tracker can better distinguish the tracked objects from the background and handle occlusion scenarios. A robust object-tracking algorithm based on the pooling network is proposed, the proposed framework consists of a generalized pooling to train the ImageNet by a wide range of image-grained datasets, upon offline training the priory trained networks are trained once again after the feature extraction process. During the tracking stage to enhance the tracking performance the various layers are used for target representation, in combination to enhance the performance.

### 2.1. Designing similarity map function

The tracking mechanism is responsible to segment into the generative and discriminant tracker, the discriminant algorithm based on the correlation mechanism has gained wide attention for greater discriminant ability and further enhance the speed. The fourier transform is introduced through the correlation filter, which gains tracking performance to gain higher computation efficiency. The correlation algorithm is regarded as shown in (1). Here, k denotes sample training, o depicts the filters, $\otimes$ is the correlation operator, j is the output and is focused on gaussian window operation, through which the element corresponds to the value of the label using the training sample, $\alpha$ is the regularization parameter.

$$\vartheta(o) = \min_{o} ||k \otimes o - j||^2 + \alpha||o||^2 \quad (1)$$

### 2.2. Deep convolutional covariance network

CNN is developed via efficient network architectures through stack convolutional layers, non-linear activation layer, and pooling layer, this is widely used in all frames of computer vision. The basic network requirements are to enhance the performance of the network by broadening and deepening the network. Less research is essential to enhance the feature depiction ability from the covariance information perspective. The discriminant features here propose a generalized pooling network for the target feature extractor. The forward and backward propagation are introduced. The output or the resultant of the convolution layer is processed after this the feature matrix is expressed in terms of $Y \in U^{m*N}$. Where m is the feature channel, $= a * b$ , a and b are the feature map size of the last convolutional layer, to estimate the co-variance of the feature map. Here $I = \frac{1}{Y}(I - \frac{1}{Y}kk^U)$, $K = [1, \dots, \dots \dots, 1]^U$is a $Y - th$ dimension vector. The eigenvalue decomposes the process to obtain the covariance matrix to obtain eigen value and vector.

$$L = YIY^U \quad (2)$$

$$L = J(p)J^U \quad (3)$$

However, $p = \text{diagonal}\ (\alpha_1 \dots \alpha_d)$ is the diagonal matrix where $\alpha_x$ is the eigen value decomposition mechanism to convert the matrix power to solve the eigen value. This can be mathematically represented in (4). Here $\beta = 0.5$, $JS(p)\text{diagonal}(s(\alpha_1) \dots . s(\alpha_d))$, $s(\alpha_x)$ depicts the exponential eigenvalues. The forward propagation for the covariance pooling layer is reduced. The chain-based rule is applied for J and p as in (6).

$$B = L^{\beta} = JS(p)J^{U} \quad (4)$$

$$s(\alpha_x) = a_x^{\beta} \quad (5)$$

$$\text{training}\left(\left(\frac{dv}{dj}\right)^{U} dj + \left(\frac{dv}{dp}\right)^{U} dp\right) = \text{training}\left(\left(\frac{dv}{dB}\right)^{U} dB\right) \quad (6)$$

$$\frac{dv}{dL} = \left(\frac{dv}{dB} + \left(\frac{dv}{dB}\right)^{U} JS\right) \quad (7)$$

$$\frac{dv}{dU} = \beta\left(\text{diagonal}\left(\alpha_1^{\beta-1}, \ldots\ldots, \alpha_v^{\beta-1}\right) J^{U} \frac{dv}{dB} J\right)_{\text{diagonal}} \quad (8)$$

$$\frac{dv}{dL} = J\left(\left(B^{U}\left(J^{U} \frac{dv}{dj}\right)\right) + \left(\frac{dv}{dB}\right)_{\text{diagonal}}\right) J^{U} \quad (9)$$

$$\frac{dv}{dN} = IN\left(\frac{dv}{dL} + \left(\frac{dv}{dL}\right)^{U}\right) \quad (10)$$

$$B_{xy} = \begin{cases} 1/(\alpha_x - \alpha_y), & x \text{ not} - \text{equals } y \\ 0, & x \text{ equals } y \end{cases} \quad (11)$$

### 2.3. Space-time adaptive correlation tracking with similarity map function

In the online classification model, a new instance is allocated in each iteration; this algorithm is capable -of predicting and then updating the classifier based on the newly based instance-label mechanism. The learning mechanism is passive to generate the classifier is similar to the previous manner. The learning mechanism is responsible to ensure a new instance, which is to be classified properly. A passive algorithm that introduces a time-series or sequential data, and extracts the cumulative loss irrespective of the fixed predictor. A regularization mechanism involving the term $||g\text{-}g_{x-1}||^2$ in the STACT algorithm. Here, $g_{x-1}$ denotes the utilization in the $(h-1)$th frame and β depicts the parameter used for regularization purposes. $\sum_{h=1}^{H} ||v.\, s^h||^2|$ Depicts the spatial regularized and $||g\text{-}g_{x-1}||^2$ depicts temporal regularized.

$$\underset{s}{\text{argmin}}\, 0.5\, \|\sum_{h=1}^{H} k_x^h * s^h - o\| + 0.5 \sum_{h=1}^{H} ||v.\, s^h\Big||^2 + \frac{\delta}{2}\Big| ||g\text{-}g_{x-1}||^2 \quad (12)$$

Algorithm 1: online classification model

```
Input: A sequence of n video frames V= {I1, I2, ..., In}, Initial bounding box B1 of the
target object in the first frame
Step1: Load a pre-trained DCCN.
Step2: Extract the region of interest (ROI) around the initial bounding box B1 in the first
frame I1.
Step3: Compute the initial feature vector.
Step4: Initialize the filter.
Step5: For each subsequent frame It, t ∈ {2, ..., n}.
Step6: Extract ROI around the previous bounding box Bt-1 in frame.
Step7: Apply the SMF to Xt to predict the new location ΔBt of the target object in frame.
Step8: Compute the new bounding box.
Step9: Adjust the learning rate λ.
Step10: A sequence of predicted bounding boxes.
```

This algorithm also acts as an extended extension via two steps, despite of classification it is a learning-based linear regression algorithm, instead of updating it. The sample at the batch level is collected in the STACT algorithm, and henceforth this naturally inherits the adaptive balance mechanism against the model learning that leads to a robust model for large variations. Similarly, the STACT algorithm implements simultaneous modeling and updating the temporal regularizer to serve the rational approximate via multiple samples, for instance, it suffers from over-fitting for the corruptness sample to alleviate the updating to keep it closer to the neighboring ones.

### 2.4. Integrated architecture of space-time adaptive correlation tracking and deep convolutional covariance network

To evaluate the efficiency of the deep feature in higher-order pooling while tracking, combining the spatiotemporal SMF through a multi-fusion environment. The generalized correlation framework proposed a spatial regularization that deals with the limitations and a temporal regularization that deals with the term for filter depreciation. The objective function is denoted as shown in (13). Here d depicts the index of the channel,

and D is the channel number. Then z is spatial penalty weight, $o_{p-1}$ is the temporal regularization co-efficient to denote the template of the previous frame. This method is responsible for effectively cope up with inclusion, movement, and other aspects.

$$A(o) = \left|\left|\sum_{d=1}^{D} s^d * o^d - c\right|\right|^2 + \beta \sum_{d=1}^{D} ||zo^d \Big|\Big|^2 + \boldsymbol{\tau}||o\text{-}o_{p-1}||^2 \quad (13)$$

## 3. RESULTS AND DISCUSSION

The proposed methodology was executed on a server equipped with four NVIDIA TITAN Xp GPUs, an Intel(R) Core (TM) i7-6800K CPU, and 32 GB of RAM. The implementation was carried out using PyTorch and subsequently trained. The performance of the proposed approach is evaluated by conducting experiments on the MOTChallenge, in the next section the details of the dataset, metrics involved, and implementation details are further aligned. The proposed method is evaluated with the existing state-of-art methods and the results are shown quantitatively based on the MOT challenge. A thorough analysis of the results is carried out that demonstrates the efficacy of the module through previous studies.

### 3.1. Dataset evaluation

The proposed approach is evaluated against the MOT17 dataset [26] and MOT20 [27] datasets accordingly to the detection mechanism. The dissimilarity between the two is simple and nearer to linear motion. These two datasets are benchmark datasets responsible for the MOT challenge whilst commonly using it for object tracking purposes.

### 3.2. MOT17

This collection comprises seven photographs depicting public areas, both indoor and outdoor, that are populated by pedestrians. Every situation in the video is divided into two segments for training and evaluation. The dataset comprises 14 distinct video sequences, half of which are allocated for training and the other half for testing purposes. The training data comprises 15,948 frames, 1,638 unique identifiers, and 336,891 cells that have been labeled. The test datasets consist of 564228-labeled compartments and 17,757 frames, encompassing 2,365 unique identifiers.

### 3.3. MOT20

The dataset comprises eight video sequences that are derived from three distinct contexts. The dataset is divided into two equal parts, with half of the sequences being used for training and the remaining half for testing. The dataset utilized for training purposes comprises 15,948 frames, 1,638 unique identifiers, and 1,336,920 cells that have been appropriately labeled. The dataset utilized for testing purposes comprises 765,465 compartments that have been tagged, along with 4,479 frames, and a cumulative count of 1,501 unique identifiers. The overall pedestrian density in the context of the MOT20 datasets is significantly greater than that of MOT17, as evidenced by an average ratio of 246 pedestrians per frame. All of the captured visuals were obtained from an aerial perspective

These two datasets are benchmark datasets responsible for the MOT challenge whilst commonly using it for object tracking purposes. Both of these have training and testing sets whereas validation is not observed. Henceforth in the experiments conducted each video in the MOT17 training set is segmented into two equal parts, the first part is used for training and the second section is used for validation purposes.

### 3.4. Metrics used

To evaluate the accuracy of MOT and HOTA metrics are quantitatively evaluated for overall tracking, the metrics are mentioned here. (a) MOTA (↑): multi-object tracking accuracy, (b) ID F1 (identification F1) (↑): ID F1-score, and (c) HOTA (↑): higher order tracking accuracy. Here (↑) means that the higher the score is, the better the performance.

### 3.5. Results FOR MOT17 dataset

Here the results are evaluated for the MOTA metric for the MOT17 dataset, the methods considered here for evaluation are semi-TCL (semi-track contrastive representation learning) [28] generates a value of 73.3 for the MOTA metric, STC (spatio-temporal context) [29] method generates a value of 75.8, SGT (sparse graph tracker) [30] method generates a value of 76.3. STDFormer-LMPH (sparse-to-dense former-linear motion prediction head) [31] method generates a value of 78.4, and STDFormer-EMPH (sparse-to-dense former-exponential motion prediction head) [31] generates a value of 78.8. Whereas the BOT-SORT (box object tracking with simple online real-time tracker) [26] method generates a value of 80.6 whereas the

proposed method generates a value of 82.36, which shows that, the proposed model works efficiently in comparison with the existing method for MOTA metric. In Figure 1, the result is shown in the form of the graph by comparing the existing state-of-art method with the proposed method.

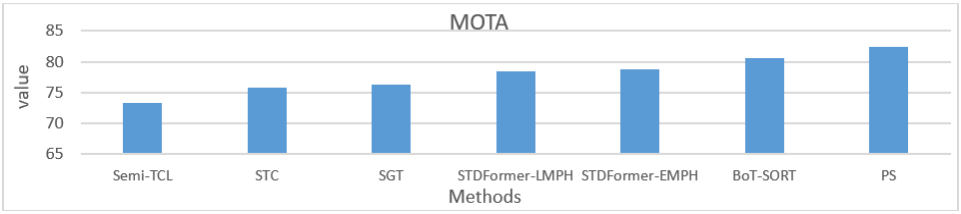


Figure 1. MOTA comparison of the existing state-of-art method with the proposed method

Here the results are evaluated for IDF1 metric for the MOT17 dataset, the methods considered here for evaluation are semi-TCL [26] generates a value of 73.2 for IDFF1 metric, STC [27] method generates a value of 70.9, SGT [28] method generates a value of 58.6 which shows the least value. STDFormer-LMPH [29] method generates a value of 73.1, and STDFormer-EMPH [29] generates a value of 71.5, which depicts an average value, whereas the bag of tricks simple online real tracking (BOT-SORT) [30] method generates a value of 79.5. Whereas the proposed method generates a value of 82.46, which shows that, the proposed model works efficiently in comparison with the existing method for IDF1 metric. In Figure 2, the result is shown in the form of the graph by comparing the existing state-of-art method with the proposed method for the IFD1 metric.

Here the results are evaluated for the HOTA metric for the MOT17 dataset, the methods considered here for evaluation are semi-TCL [26] generates a value of 59.8 for the HOTA metric, STC [27] method generates a value of 59.8, SGT [28] method generates a value of 60.6. STDFormer-LMPH [29] method generates a value of 60.9, and STDFormer-EMPH [29] generates a value of 59.9. Whereas the BOT-SORT [30] method generates a value of 64.6 whereas the proposed method generates a value of 67.85, which shows that the proposed model works efficiently in comparison with the existing method for HOTA metric. In Figure 3, the result is shown in the form of the graph by comparing the existing state-of-art method with the proposed method for the HOTA metric.

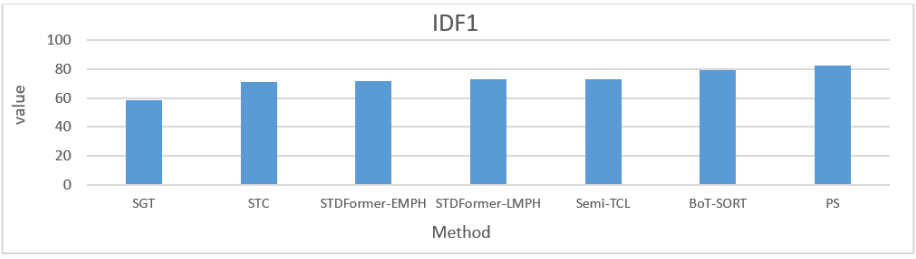


Figure 2. IDF1 comparison of the existing state-of-art method with the proposed method

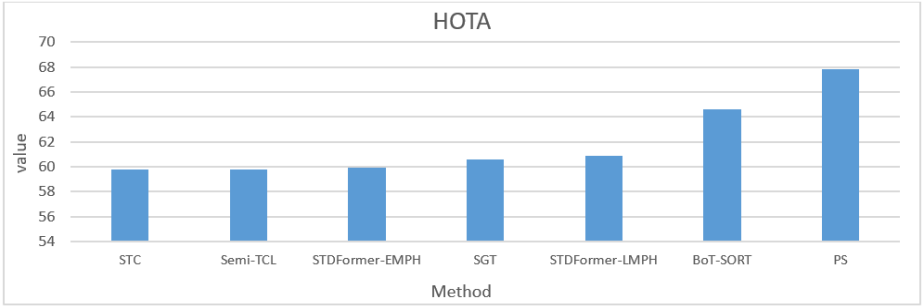


Figure 3. HOTA comparison of the existing state-of-art method with the proposed method for the MOT17 dataset

### 3.6. Results for MOT20 dataset

Here the results are evaluated for the MOTA metric for MOT20 dataset, the methods considered here for evaluation are semi-TCL [26] generates a value of 65.2 for the MOTA metric, STC [27] method generates a value of 73, SGT [28] method generates a value of 72.8. STDFormer-LMPH [29] method generates a value of 76.2, and STDFormer-EMPH [29] generates a value of 75.8. Whereas the BOT-SORT [30] method generates a value of 77.7 whereas the proposed method generates a value of 79.86, which shows that, the proposed model works efficiently in comparison with the existing method for MOTA metric. In Figure 4, the result is shown in the form of the graph by comparing the existing state-of-art method with the proposed method for the MOTA metric for MOT20 dataset.

Here the results are evaluated for IDF1 metric for MOT20 dataset, the methods considered here for evaluation are Semi-TCL [26] generates a value of 70.1 for IDF1 metric, STC [27] method generates a value of 67.5, SGT [28] method generates a value of 70.5. STDFormer-LMPH [29] method generates a value of 72.1, and STDFormer-EMPH [29] generates a value of 72.3. Whereas the BOT-SORT [30] method generates a value of 76.3 whereas the proposed method generates a value of 79.86, which shows that, the proposed model works efficiently in comparison with the existing method for MOTA metric. In Figure 5, the result is shown in the form of a graph by comparing the existing state-of-art method with the proposed method for IDF1 metric for MOT20 dataset.

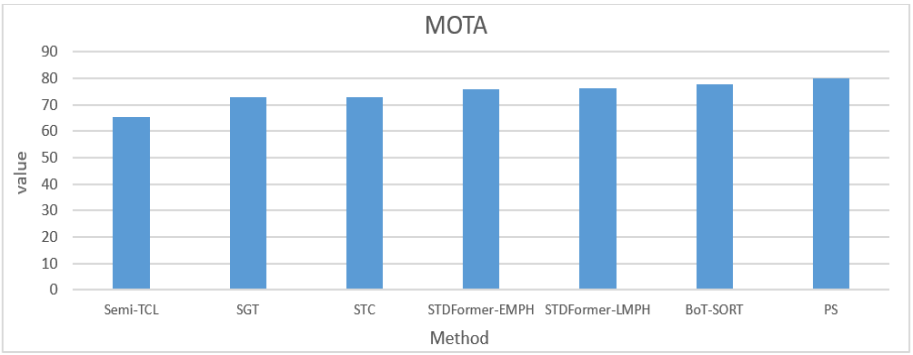


Figure 4. MOTA comparison of the existing state-of-art method with the proposed method for the MOT20 dataset

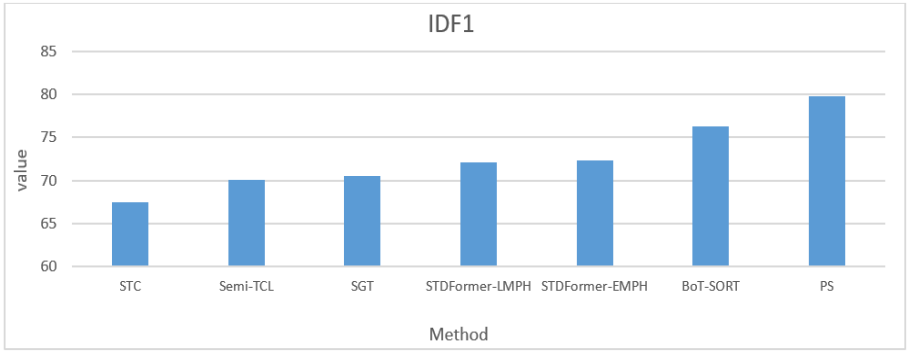


Figure 5. IDF1 comparison of the existing state-of-art method with the proposed method

Here the results are evaluated for the HOTA metric for the MOT20 dataset, the methods considered here for evaluation are semi-TCL [26] generates a value of 55.3 for the HOTA metric, STC [27] method generates a value of 56.3, SGT [28] method generates a value of 56.9. STDFormer-LMPH [29] method generates a value of 60.2, and STDFormer-EMPH [29] generates a value of 60. Whereas the BOT-SORT [30] method generates a value of 62.6 whereas the proposed method generates a value of 65.78, which shows that, the proposed model works efficiently in comparison with the existing method for MOTA metric. In Figure 6, the result is shown in the form of a graph by comparing the existing state-of-art method with the proposed method for the HOTA metric for MOT20 dataset.

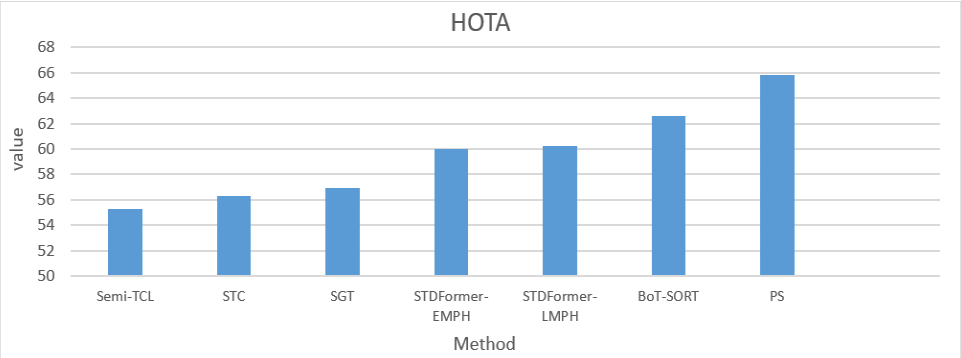


Figure 6. HOTA comparison of the existing state-of-art method with the proposed method

### 3.7. Comparative analysis

The comparative analysis here is carried out with the existing system and the proposed system. The results are shown below in the table for MOT17 dataset and MOT20 dataset for the metrics MOTA, IDF1, and HOTA for both the datasets. The comparison is carried out for the MOT17 dataset, which shows that for the MOTA metric the existing method STDFormer-LMPH shows a value of 78.4.

### 3.8. Comparative analysis for MOT17 dataset

Here the proposed model generates a value of 82.36 and the improvisation in % is 4.9266% and for STDFormer-EMPH shows a value of 78.8 whereas the proposed model generates a value of 82.36 and the improvisation in % is 4.41797. For IDF1 metric the existing method STDFormer-LMPH shows a value of 73.1 whereas the proposed model generates a value of 82.46 and the improvisation is 12.0339% and for STDFormer-EMPH shows a value of 71.5 whereas the proposed model generates a value of 82.46 and the improvisation is 14.2375%. For the HOTA metric the existing method STDFormer-LMPH shows a value of 60.9 whereas the proposed model generates a value of 67.85 and the improvisation is 10.7961%and for STDFormer-EMPH shows a value of 59.9 whereas the proposed model generates a value of 67.85 and the improvisation in (%) is 12.4462%.

### 3.9. Comparative analysis for MOT20 dataset

The comparison is carried out for the MOT20 dataset, which shows that for the MOTA metric the existing method STDFormer-LMPH shows a value of 76.2. Whereas the proposed model generates a value of 79.86 and the improvisation is 4.6905% and for STDFormer-EMPH shows a value of 75.8 whereas the proposed model generates a value of 79.86 and the improvisation is 5.2165%. For IDF1 metric the existing method STDFormer-LMPH shows a value of 72.1 whereas the proposed model generates a value of 79.86 and the improvisation is 10.2132% and for STDFormer-EMPH shows a value of 72.3 whereas the proposed model generates a value of 79.86 and the improvisation is 9.93691%. For the HOTA metric the existing method STDFormer-LMPH shows a value of 60.2 whereas the proposed model generates a value of 67.85 and the improvisation is 8.85855% and for STDFormer-EMPH shows a value of 60 whereas the proposed model generates a value of 67.85 and the improvisation is 9.19065%

## 4. CONCLUSION

The combination of the DCCN, STACT, and the similarity map function provides a robust and accurate solution for multiple object tracking. This approach leverages deep learning for feature representation, correlation-based tracking for localization, and a similarity map for refining the tracking results. The main idea of the paper lies in the utilization of an integrated network for target categorization. The present network is designed to represent the interconnections among multiple locations within a specified boundary area, to gather comprehensive contextual data. The network may extract more discriminative features and increase tracking accuracy by utilizing integrated architecture processes. In conclusion, the research presents a unique DCCNs-based visual tracking method. The proposed methodology exhibits more effective tracking capabilities compared to prior approaches, in gathering the contextual data and leveraging higher-order pooling operations. The present research contributes to the field of object tracking and presents potential applications across diverse industries such as surveillance, robotics, and autonomous vehicles. The comparative analysis shows that the proposed method performs better in comparison with the existing state-of-art methods.

**ACKNOWLEDGEMENT**

We would like to express our deepest gratitude to all parties who have supported and contributed to this research project. Most of all, we express our sincere thanks to our guide for his unwavering guidance, invaluable insight, and encouragement throughout the research process. No funds were collected for this research.

**REFERENCES**

[1] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, vol. 2017-Janua, pp. 3701–3710, doi: 10.1109/CVPR.2017.394.

[2] R. Henschel, L. Leal-Taixe, D. Cremers, and B. Rosenhahn, "Fusion of head and full-body detectors for multi-object tracking," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, vol. 2018-June, pp. 1509–150909, doi: 10.1109/CVPRW.2018.00192.

[3] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, vol. 2015 Inter, pp. 4696–4704, doi: 10.1109/ICCV.2015.533.

[4] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee, "Multi-class multi-object tracking using changing point detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9914 LNCS, 2016, pp. 68–83.

[5] K. Fang, Y. Xiang, X. Li, and S. Savarese, "Recurrent autoregressive networks for online multi-object tracking," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018, vol. 2018-Janua, pp. 466–475, doi: 10.1109/WACV.2018.00057.

[6] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2018, pp. 1–6, doi: 10.1109/ICME.2018.8486597.

[7] Y. Yoon, A. Boragule, Y. Song, K. Yoon, and M. Jeon, "Online multi-object tracking with historical appearance matching and scene adaptive detection filtering," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Nov. 2018, pp. 1–6, doi: 10.1109/AVSS.2018.8639078.

[8] S. H. Sreedhara, V. Kumar, and S. Salma, "Efficient big data clustering using adhoc fuzzy c means and auto-encoder CNN," in *Lecture Notes in Networks and Systems*, vol. 563, 2023, pp. 353–368.

[9] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: efficient convolution operators for tracking," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, vol. 2017-Janua, pp. 6931–6939, doi: 10.1109/CVPR.2017.733.

[10] H. Wu, Y. Hu, K. Wang, H. Li, L. Nie, and H. Cheng, "Instance-aware representation learning and association for online multi-person tracking," *Pattern Recognition*, vol. 94, pp. 25–34, Oct. 2019, doi: 10.1016/j.patcog.2019.04.018.

[11] S.-H. Lee, M.-Y. Kim, and S.-H. Bae, "Learning discriminative appearance models for online multi-object tracking with appearance discriminability measures," *IEEE Access*, vol. 6, pp. 67316–67328, 2018, doi: 10.1109/ACCESS.2018.2879535.

[12] H. Karunasekera, H. Wang, and H. Zhang, "Multiple object tracking with attention to appearance, structure, motion and size," *IEEE Access*, vol. 7, pp. 104423–104434, 2019, doi: 10.1109/ACCESS.2019.2932301.

[13] J.-N. Zaech, A. Liniger, D. Dai, M. Danelljan, and L. Van Gool, "Learnable online graph representations for 3D multi-object tracking," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5103–5110, Apr. 2022, doi: 10.1109/LRA.2022.3145952.

[14] R. Li, B. Zhang, J. Liu, W. Liu, and Z. Teng, "Inference-domain network evolution: a new perspective for one-shot multi-object tracking," *IEEE Transactions on Image Processing*, vol. 32, pp. 2147–2159, 2023, doi: 10.1109/TIP.2023.3263104.

[15] Y. He, X. Wei, X. Hong, W. Ke, and Y. Gong, "Identity-quantity harmonic multi-object tracking," *IEEE Transactions on Image Processing*, vol. 31, pp. 2201–2215, 2022, doi: 10.1109/TIP.2022.3154286.

[16] L. Chen, H. Ai, R. Chen, and Z. Zhuang, "Aggregate tracklet appearance features for multi-object tracking," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1613–1617, Nov. 2019, doi: 10.1109/LSP.2019.2940922.

[17] J. Pinto, Y. Xia, L. Svensson, and H. Wymeersch, "An uncertainty-aware performance measure for multi-object tracking," *IEEE Signal Processing Letters*, vol. 28, pp. 1689–1693, 2021, doi: 10.1109/LSP.2021.3103488.

[18] H. Van-Nguyen, H. Rezatofighi, B.-N. Vo, and D. C. Ranasinghe, "Distributed multi-object tracking under limited field of view sensors," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5329–5344, 2021, doi: 10.1109/TSP.2021.3103125.

[19] Y. Chen, J. Huang, H. Liu, M. Huang, and Z. Zou, "Appearance guidance attention for multi-object tracking," *IEEE Access*, vol. 9, pp. 103184–103193, 2021, doi: 10.1109/ACCESS.2021.3087168.

[20] Y.-M. Song, Y.-C. Yoon, K. Yoon, H. Jang, N. Ha, and M. Jeon, "Multi-object tracking and segmentation with embedding mask-based affinity fusion in hierarchical data association," *IEEE Access*, vol. 10, pp. 60643–60657, 2022, doi: 10.1109/ACCESS.2022.3171565.

[21] W. Ren, X. Wang, J. Tian, Y. Tang, and A. B. Chan, "Tracking-by-counting: using network flows on crowd density maps for tracking multiple targets," *IEEE Transactions on Image Processing*, vol. 30, pp. 1439–1452, 2021, doi: 10.1109/TIP.2020.3044219.

[22] Y. Gao, H. Xu, Y. Zheng, J. Li, and X. Gao, "An object point set inductive tracker for multi-object tracking and segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 6083–6096, 2022, doi: 10.1109/TIP.2022.3203607.

[23] X. Wan, J. Cao, S. Zhou, J. Wang, and N. Zheng, "Tracking beyond detection: learning a global response map for end-to-end multi-object tracking," *IEEE Transactions on Image Processing*, vol. 30, pp. 8222–8235, 2021, doi: 10.1109/TIP.2021.3113169.

[24] Q. Liu, B. Liu, Y. Wu, W. Li, and N. Yu, "Real-time online multi-object tracking in compressed domain," *IEEE Access*, vol. 7, pp. 76489–76499, 2019, doi: 10.1109/ACCESS.2019.2921975.

[25] S. You, H. Yao, and C. Xu, "Multi-object tracking with spatial-temporal topology-based detector," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3023–3035, May 2022, doi: 10.1109/TCSVT.2021.3096237.

[26] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: a benchmark for multi-object tracking," Mar. 2016.

[27] P. Dendorfer *et al.*, "MOT20: a benchmark for multi object tracking in crowded scenes," Mar. 2020.

[28] W. Li, Y. Xiong, S. Yang, M. Xu, Y. Wang, and W. Xia, "Semi-TCL: semi-supervised track contrastive representation learning," Jul. 2021.

[29] A. Galor, R. Orfaig, and B. Bobrovsky, "Strong-transcenter: improved multi-object tracking based on transformers with dense representations," Oct. 2022.

[30] J. Hyun, M. Kang, D. Wee, and D.-Y. Yeung, "Detection recovery in online multi-object tracking with sparse graph tracker," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 4839–4848, doi: 10.1109/WACV56688.2023.00483.
[31] M. Hu, X. Zhu, H. Wang, S. Cao, C. Liu, and Q. Song, "STDFormer: spatial-temporal motion transformer for multiple object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023, doi: 10.1109/TCSVT.2023.3263884.

## BIOGRAPHIES OF AUTHOR

**Mrs. Kusuma Sriram** received B.E. degree under VTU in 2004, and M.Tech. from RVCE in the year 2010. She is an active academician having 18+ years of teaching experience. Currently she is working at Ramaiah Institute of Technology, Bengaluru, as an Assistant Professor in the Department of Information Science and Engineering. She is Pursuing Ph.D. in the area of video processing/machine learning. She is a trained faculty of Infosys InfyTQ to motivate students for Infosys job opportunities. She has worked as trainer for infosys campus connect program. She can be contacted at email: kusumas_12@rediffmail.com.

**Dr. Kiran Purusotham** currently working as HOD, CSE Department with total experience of 20 years. He has done Research on Privacy Preserving Data Mining with focus on detection of sensitivity patterns and was awarded Ph.D. from VTU in 2014. His research interests include cryptography, randomization, anonymization methods in generalization, indexing techniques and design patterns. He has guided several M.Tech. and B.E. projects and Internships and is currently guiding four research scholars at VTU. He can be contacted at email: kiranpmys@gmail.com.