Optik - International Journal for Light and Electron Optics 262 (2022) 169279

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.com/locate/ijleo

Multi-view data fusion in multi-object tracking with probability density-based ordered weighted aggregation

Alireza Dadgar, Yasser Baleghi *, Mehdi Ezoji

Department of Electrical & Computer Engineering, Babol Noshirvani University of Technology, Babol, Mazandaran, Iran

ARTICLE INFO

Keywords:
Multi-object tracking
Mask R-CNN
Fusion
Probability density-based order weighted aggregation

ABSTRACT

In this paper, a method is presented for Multi-Object Tracking (MOT) in presence of partial or complete occlusions. This work focuses on improved object detection and data association in a single view, and also fuses data from multiple views using the Ordered Weighted Aggregation (OWA) algorithm. Hence, a deep learning model was proposed to detect objects more accurately in a tracking-by-detection framework. This paper aims to enhance object detection, data association, and the trajectories of the objects in the MOT algorithm respectively by applying Mask R-CNN, Zernike Moments and combination of several similarity metrics, and fusion of multi-camera information by probability density-based OWA (PD-OWA). The inter-frame detected objects are matched based on the appropriate similarity metrics. In the fusion part, the Kernel Density Estimation (KDE) is utilized to assign weight coefficient to each camera view and determine the descending order of data in the OWA algorithm. Finally, the positions of each object coming from different views are weighted and aggregated. The results show that the proposed method improves object detection, association performance, and tracking trajectory in the "PETS09-S2L1″ and the "EPFL Terrace" video sequences and achieved 81.6% and 79.6% multiple objects tracking accuracy (MOTA), respectively.

1. Introduction

Multi-object tracking (MOT) is an important topic in computer vision. It is typically one of the first steps of video analysis in surveillance, sports, and industrial applications. The number of objects in this approach will vary over time and may merge, split, appear, or disappear in the scene. Due to object shape deformations, brightness changes, and occlusion and distraction issues, the current techniques continue to fail in some cases [1–7]. The increased number of objects introduces new and significant challenges in detection, data association, and tracking. The algorithms that have been proposed have focused on these issues and provided solutions to them [1,8–10].

Human detection is the process of determining whether an image or video contains a person, identifying his/her position, and tracking the movements. Human identification is challenging due to the different positions people take, occlusion, and the presence of humanoid-looking items in the backdrop, leading to false detections [3,11,12].

Deep learning is a subclass of machine learning methods that employ artificial neural networks. It is utilized in visual recognition, natural language processing, and speech recognition to replicate human brain processing. This concept arose from adding additional layers to neural networks to deal with more complex challenges. On the other hand, adding more layers introduced the issue of

* Correspondence to: Electrical & Computer Engineering Department, Babol Noshirvani University of Technology, Iran.
E-mail address: y.baleghi@nit.ac.ir (Y. Baleghi).

https://doi.org/10.1016/j.ijleo.2022.169279
Received 2 February 2022; Received in revised form 8 May 2022; Accepted 8 May 2022
Available online 12 May 2022

vanishing gradients, in which the gradient appears to get smaller as the process proceeds backward, resulting in poor results. The incorporation of novel activation functions appears to solve this difficulty, allowing more layers to be added to the network and contributing to what is now known as deep learning [9,13].

Convolutional neural networks (CNN) require the configuration of hyper-parameters for the number, shape, stride, and filters. Hence, the key contribution in CNN is to extract the optimum hyper-parameters from the deep learning model to increase human detection accuracy. By adjusting the hyper-parameters, it is also possible to lower the computing resource required for execution [13,14].

After correct object detection in consecutive frames by the CNN model, special attention should be paid to determining each object's identification and more accurate assignment at any time, maintaining the consistency of object identities while tracking and solving multi-identity matching challenges. As a result, validation criteria can be examined alongside each object during the video sequence to accomplish accurate object matching. In this context, the proposed MOT method will handle issues such as implementing and updating a robust CNN model for object detection and data association based on orthogonal Zernike Moments feature and multiple similarity metrics to improve object matching [15,16].

In order to overcome the drawbacks mentioned above in the case of MOT, the fusion [17] approach is presented to eliminate weaknesses of MOT and thus associated objects more accurate. The OWA (Ordered Weighted Aggregation) operator applies the weights assigned to the ordered input values, from different camera views [18], rather than the specific criteria. This allows one to model various aggregation preferences, preserving the simultaneously impartiality with respect to the individual attributes [19,20].

In machine vision disciplines, the issue of aggregating the data obtained from multiple camera views in order to form the entire objective functions is very important [18,21]. The most frequently used aggregation acts on the basis of the weighted sum. One of the critical issues in employing the OWA operator for decision-making is determining the weights of the OWA operator. In order to determine the OWA operator weights [22], a probability density technique is presented. In order to determine the probability density function (PDF) to fit the entire input values, a robust mathematical tool is presented, which is called the kernel density estimation (KDE), [23]. The KDE may effectively catch the complex data distribution feature of input values without the need for applying complicated techniques of classification. In this way, to deal with the relatively dense distribution of people and their occlusions, the output of several cameras is used to identify the objects accurately. Due to the variety of appearances between different viewing angles and by estimating the coordinates of the object in the same reference (ground plane/top plane), higher accuracy and less error can be expected.

This paper presents a multi-object tracking method based on the multi-view data fusion. This study aims to improve object detection and object matching in intra tracking, respectively by employing Mask R-CNN and similarity-based object matching. The superiority of this method lies on more accurate estimation of the position of objects in the scene by fusion of multi-view information based on the PD-OWA aggregation algorithm.

The remainder of the paper is structured as follows: The second section will include the background of the literature, followed by preliminaries and proposed method in the third and fourth sections, respectively. In the fifth section, the results and evaluation will be extended, and the conclusion will be presented in the sixth section.

2. Background

The MOT comprises two parts: the first is responsible for object detection, and the second is to associate the corresponding detected objects. As object detection is an extensive research field, many MOT methods focus on the association step, where different cues (position, motion, visual appearance, pose, etc.) are considered for linking detections to tracks [24]. Occlusion challenges, for example, may produce unfavourable results throughout the tracking process. The relevant investigations on tracking-by-detection framework, object detection, tracking, data association and multi-camera approaches will be discussed in the following.

2.1. Tracking-by-detection framework

Tracking-by-detection is a popular method for tracking that initially detects objects [25], and then associates them in consecutive frames. This strategy can be classified into two parts: local and global tracking methods. For data association, the local approaches examine only two frames. This makes them computationally economical, but their performance is vulnerable to tracking-irrelevant factors, including camera motion and poses variation, among others.

In contrast with local techniques, a more considerable number of frames are employed in global techniques, in order to carry out data association. Data association is turned into a network flow problem in novel techniques in this field. For instance, a constrained flow optimization problem was solved by Berclaz et al. [26] for multiple object tracking. They employed the k-shortest paths technique in order to associate the tracks. The min-cost network flow framework was also extended by Chari et al. [27] via a pairwise cost. They suggested a convex relaxation approach along with a rounding heuristic for tracking. Also, Shitrit et al. [28] used multi-commodity network flows for MOT. Even though this category of algorithms is popular, and it highly depends on object detectors.

2.2. Object detection approaches

In the literature, object detection and pattern recognition method can be divided into four general categories; background modelling (e.g. Gaussian mixture model, frame differencing, hidden Markov model), optical flow, segmentation (e.g. mean shift, active contour) [29,30], and point detection (e.g. deep learning detectors, traditional detectors) [31,32]. In the case of pedestrian detectors,

Table 1
Different types of detectors, their approach, and their drawbacks.

Detector	*Approach*	*Drawback*
Model-based detectors	A background model created, and then a pixel-wise or block-wise comparison of a new image against the background model is performed to detect regions that do not fit the model (Gaussian Mixer Model, etc.)	Very sensitive to changes in illumination and occlusions.
Template-based detectors	The detectors use a pre-learned set of templates	Significantly affected by background clutter and occlusions
Part-based detectors	A template for each body part is learned separately	Computational cost, Data set
Block-based detectors	The objective is to learn the appearance of blocks inside the bounding box of a detection. (Based on HOG features or SIFT features, etc.)	Occlusions
Convolutional Neural Network	Train objects based on input data from data sets and learn special features based.	Computational cost

the performance of each commonly approach and its drawbacks are presented in Table 1 [32].

Given the emergence of the new methods, including Convolutional Neural Networks (CNNs), these studies have received quite noticeable success. One can categorize these enhancements as structural reformulation, regularization, optimization, etc [33]. In recent years, the object detectors acting based on bounding boxes have been steadily enhanced. In this concept, R-CNNs (Region-based Convolutional Neural Networks), are successful classes of two-stage methods [34]. In R-CNNs, the candidate ROIs (Regions of Interest) are proposed in the first stage, while the object classification is carried out in the second stage. Through an ROI pooling operation, one can rapidly extract region-wise characteristics from shared feature maps [35]. The speed of instance-level detection is increased by feature sharing, which makes recognition of higher-order interactions possible that could not be computed otherwise.

Uçar et al. [36] proposed a novel hybrid local multiple system with feature extraction and robust classification based on CNNs and Support Vector Machines (SVMs). Using several CNNs, they partitioned the entire image into local regions in the proposed method. They used principal component analysis (PCA) to extract discriminating features, then integrated into multiple SVMs utilizing empirical and structural risk minimizations. Finally, they attempted to combine SVM outputs. They applied the pre-trained AlexNet model, and the Caltech-101 Pedestrian datasets to conduct object recognition and pedestrian detection studies [37]. Their proposed approach produced better outcomes with a low miss rate and enhanced object recognition and detection with increased accuracy. Zhou et al. [38] presented the architecture and deep learning techniques in an object detection task application. They built their dataset and demonstrated that utilizing CNN algorithm for object detection yielded good results. Experimental results show that deep learning technics effectively passes the artificial feature with extensive qualitative data.

Using SORT (Simple Online and Realtime Tracking) [39], the position of tracks is propagated to the following frame with a Kalman filter, and detection boxes are associated to tracks on the basis of the overlap measured via Intersection over Union (IoU). Deep visual features extracted through a convolutional neural network have been integrated into the association procedure via further developments in DeepSORT, while human poses have been incorporated as well in [40]. Numerous other tracking frameworks have also pursued the idea of combining various cues to obtain high-quality similarity measures [41]. One of the deficiencies of these techniques emanates from the necessity of training and designing separate networks, which results in additional computational costs. In contrast, the authors in [42] trained an appearance embedding model along with the detection model in a shared network to enhance the efficiency of the entire MOT system.

2.3. Tracking approaches

The targeted objects must be tracked using the best method to extract the absolute path travelled by the objects independently. Depending on the situation, object tracking employs two types of algorithms: non-predictive (e.g., Mean shift and CAM shift) and predictive (e.g., Kalman filter and particle filter) [43,44]. The tracking in the first group is based on matching and the second group is employed the object position in frame k to estimate the object position in frame $k + 1$. The same approach, like the second group method, serves as the foundation for the tracker in this paper.

2.4. Data association approaches

Detected objects in the MOT process, in the k^{th} frame, must be associated with the objects in the prior frames. The Nearest Neighbour (NN) and General Nearest Neighbour (GNN) methods are popular data association approaches. When the object areas are close together or the number of incorrect measurements increases, the mentioned approaches may also be inaccurate. To improve the GNN, the Joint Probabilistic Data Association (JPDA) approach has been developed. The PDA approach encounters several targets at the same time [45].

Multiple Hypothesis Tracking (MHT) is a statistical association algorithm that can even postpone data association to the following repetitions to reduce ambiguities [46]. The MHT technique is generally divided into hypothesis matrix generation, hypothesis generation, hypothesis probability calculation, Kalman filter calculation associated with the target, and hypothesis management. As a result, several hypotheses will be the output of a single hypothesis in occlusion and noisy situations. However, depending on the

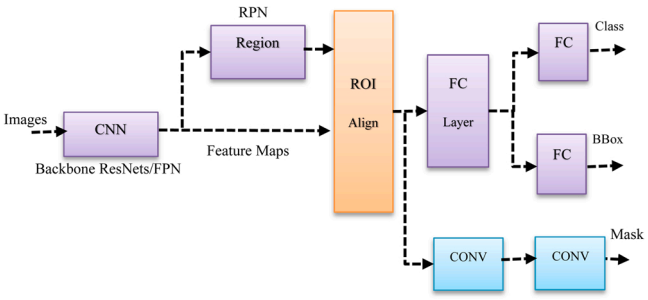

Fig. 1. In the Mask R-CNN diagram, Faster R-CNN is extended by the addition of a branch used to predict segmentation masks on every ROI. The mask branch is in parallel with bounding box regression and classification branches.

application complexity, the computing cost can be substantial. The Markov Chain Monte Carlo data association approach is an approximation that treats the problem as a hybrid optimization problem and investigates it through random space exploration rather than enumerating all association possibilities [47].

2.5. Multi-camera approaches

In order to remove some limitations of single view in MOT and to help track multiple objects more accurately, we will use the multi-camera data fusion. The fusion method like OWA can be used by the incorporation of multiple cameras, and integrate received data from different views. Since its presentation, a wide range of fields, such as expert systems, decision-making, neural networks, database systems, mathematical programming, and fuzzy logic controllers, have utilized the OWA operator [19]. One of the hot research topics in the field of decision-making analysis is how to determine the OWA operator weights, and many techniques have been presented so far, including quantifier functions, distribution assumption, and constraint optimization models. These techniques determine the OWA operator weights objectively. This paper presents an MOT method based on multi-camera data fusion with the help of the Mask R-CNN framework [48], which is trained to extract humans. On the other hand, an object matching method is performed by applying a combination of similarity criteria, e.g. Zernike Moments [15,16,49], Hausdorff distance [50], and EMD [51].

3. The preliminaries

In this paper, R-CNN was enhanced to allow ROIs on feature maps to be attended to utilizing ROI Pool, resulting in faster speed and higher accuracy. Also, the OWA method is employed to fuse data from different camera views. In the following, the basics of the Mask R-CNN network and the OWA fusion technique will be explained.

3.1. Main frame (Mask R-CNN)

Instance segmentation methods are based on segment proposals due to the efficacy of R-CNN. Mask R-CNN is easy to develop and train due to the Faster R-CNN framework, which allows for various configurable architecture designs. Furthermore, the mask branch adds only a minor computational overhead, allowing for a fast system and rapid experimentation. In theory, Mask R-CNN is an intuitive extension of Faster R-CNN, but adequately designing the mask branch is vital for good results [48]. More details about the Mask R-CNN subsections will be provided in the following.

3.1.1. Backbone

To show the generality of Mask R-CNN, it is applied by employing several architectures. According to Fig. 1, for more clarity, the convolutional network can be presented in the following two distinguishable sections: (i) the convolutional backbone architecture is employed to extract the features from an entire image, and (ii) the network head used for bounding box recognition (including regression and classification) and prediction of the mask of each target, which is separately applied to each ROI [48].

The network-depth-features nomenclature is used to define the backbone architecture. It evaluates ResNet (Deep residual networks) and ResNeXt networks of depth 50 or 101 layers. The initial Faster R-CNN with ResNets implementation extracted features from the last convolutional layer [52]. It is also experimented with a Feature Pyramid Network (FPN), a more effective backbone recently proposed by Lin et al. [53]. Using a ResNet-FPN backbone for feature extraction yields excellent accuracy and speed gains. The mask branch is a small FCN (Fully Connected Network) applied to each ROI and predicted a segmentation mask pixel-by-pixel. ResNet50 was utilized in the suggested approach. The model then identifies ROI in the network architecture's head, and classification and detection are made based on the resulting ROI classification and detection.

- RoIAlign

  A quantization-free layer known as RoIAlign is used to correct the misalignment, which reliably preserves exact spatial locations. RoIAlign utilizes bilinear interpolation rather than nearest-neighbour interpolation to compute each position's pixel value. It

traverses region suggestions first, then splits each region proposal into k × k units. The coordinate values are then computed for each unit, and the pixel values of locations are generated using bilinear interpolation before the max-pooling operation is done [48].

- Mask

  A mask encodes the spatial layout of an input object. The pixel-to-pixel correlation offered by convolutions makes it easy to extract the spatial pattern of masks. An m×m mask is projected from each RoI utilizing an FCN. This enables each layer in the mask branch to keep the specific m×m object spatial layout without reducing it into a vector representation with no spatial dimensions.
- RPN

  The RPN receives an image as input and returns a collection of rectangular object suggestions, each with an objectness score. It detects the anchor in the foreground or background and conducts the initial coordinate adjustment for foreground anchors. The RPN generates k object boxes (k = 15 in this study) with a fixed aspect ratio and scale for each pixel using sliding windows on shared convolutional feature maps [52].
- Loss Function

With the addition of a mask branch, the multi-task loss function of Mask R-CNN may be stated as follows [48]:

$$L_{final} = \quad L_{RPN-cls} + \quad L_{RPN-bbox} + \quad L_{cls} + \quad L_{bbox} + \quad L_{mask} \tag{1}$$

where $L_{RPN-cls}$ indicates the classification loss function in the RPN, $L_{RPN-bbox}$ indicates the position regression loss function in the RPN, L_{cls} is the classification loss function, L_{bbox} represents the position regression loss function, and L_{mask} is the average binary cross-entropy.

3.2. OWA (ordered weighted aggregation)

Yager [19] presented the OWA operator, which is a parameterized category of the mean type aggregation operators through the assignment of weights to every input data. In case the weights of the OWA operator are determined, the particular aggregation operators, including the arithmetic average, min, and max operators will be determined. As a result of its practicality, it is a widely used solution in different fields, including EEG signal improvement, machine learning, risk assessment, and time-series data fusion. The process of applying an OWA operator comprises of the following three steps: (1) rearrangement of the input values in descending order, (2) determining the weights of the values of rearranged inputs via a robust technique, and (3) ultimately aggregating the same rearranged input values into a single value in accordance with the derived weights [54]. Determining the weights of rearranged input values is an important factor for the OWA operator.

Various aggregation operators, including ordered weighted averaging (OWA) operator, weighted averaging operator, and weighted geometric operator, have been presented in order to enhance the ranking results for multi-criteria decision-making problems. In fact, an ordered weighted averaging is a mapping function $F_w : R^n \quad \rightarrow R$, which is related to a weight vector $W = [w_1, w_2, \ldots, w_n]$, which satisfies $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_j \in [0,1]$ (Eq. (2)) [20].

$$F_w(I_1, \quad I_2, \ldots, I_n) = \quad \sum_{j=1}^{n} w_j v_j \tag{2}$$

In which $I_1, I_2, \ldots, I_n$ stand for the input values to be aggregated, and v_j represents the jth largest input amongst them (i.e., input values I_1,I_2,…,I_n). In order to characterize the ordered weighted aggregation operators, dispersion and orness, two important measures have defined. The latter measure, which is also known as attitudinal character, is described as $O(w) =)1/)$n-1(($\sum_{j=1}^{n}(n-j) \quad w_j$, in which $O(w)$ stands for the orness of the ordered weighted averaging operator (Andness acts as the complementary concept for orness and vice versa, $andness(w) = 1 - orness(w)$). It is noteworthy that the orness measure of the OWA operator ranges within the [0,1] unit interval and characterizes its similarity degree to the max operator [20].

A fundamental aspect of an OWA operator is the reordering step, in particular, an aggregate I_i is not associated with a particular weight w_i but rather a weight is associated with a particular ordered position of an argument. If $w_1 = 1$, then F_w is the pure "or" operator and, if $w_n = 1$, then F_w is the pure "and" operator. In order to classify OWA operators with respect to their location between "and" and "or", a measure of orness is associated with any weight vector. It is easy to see that for any w, the $O(w)$ always lies in the unit interval. Note that, if

$w_1 = 1$, then $O(w) = 1$; if $w_n = 1$, then $O(w) = 0$, for any other choice of w, $O(w)$ lies in $[0,1]$. Maximum aggregation representing the fuzzy "or" operator is obtained with $w = (1, 0, \ldots, 0)^T$. Here, one gets $O(w) = 1$. Minimum aggregation representing the fuzzy "and" operator is obtained with $w = (0, 0, \ldots, 1)^T$. Here, one gets $O(w) = 0$. The aggregations with orness greater than or equal to 1/2 are considered or-like whereas the aggregations with orness smaller than or equal to 1/2 are treated as and-like. The former corresponds to rather an optimistic preference, while the latter represents a pessimistic preference.

One can define the andness/orness as the desired risk level to be added to the aggregation procedure. The former measure, which is termed entropy as well, is described by $E(w) = -\sum_{j=1}^{n} w_j \ln(w_j)$, in which $E(w)$ stands for the OWA operator dispersion and determines how uniformly the input values are being utilized. From another viewpoint, one can call entropy the tradeoff level between criteria. Given the obtained PDF, larger weights are assigned to the input values with higher probability densities. Also, smaller weights are assigned to the input values with lower probability densities. The present paper suggests a new probability density-based ordered weighted averaging (PD-OWA) operator on the basis of the probability density technique and investigates its favourable character-

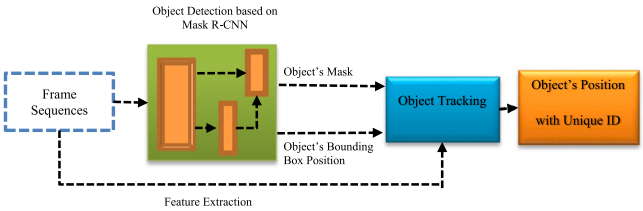

Fig. 2. Tracking-by-detection paradigm in each camera view. Primarily, an independent detector is applied in each frame to obtain likely human detections. Secondly, a tracker is run on the set of detections to perform data association based on the object's features, and lastly a unique ID is assigned to each object.

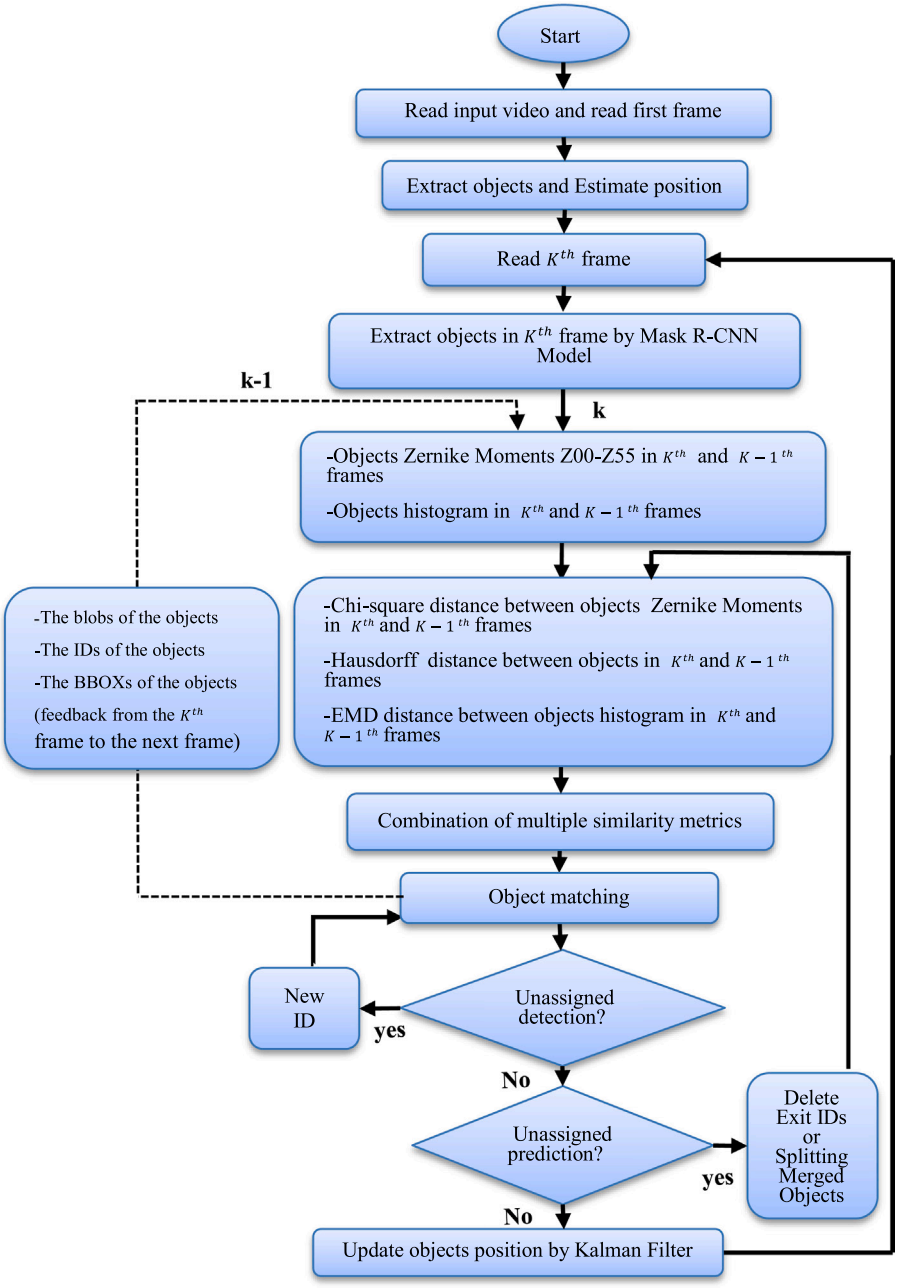

Fig. 3. Multiple-Object Tracking in a single view flow diagram (Intra tracking).

istics. In addition, in the procedures of weighting, aggregating, and reordering the input values of the suggested PD-OWA operator, the positions of the objects are employed as inputs [55].

4. Proposed method

Object detection and data association are the two most common processes in MOT. Objects are first detected in each sequence frame, and then the detections are matched to form complete trajectories. This is known as the tracking-by-detection paradigm [25], and it serves as the framework for the rest of the study (Fig. 2).

MOT challenges via a novel approach. The objects masks are obtained by Mask R-CNN. Sparse and semi-crowded scenarios are focused on throughout this research. In this article, the tracking-by-detection paradigm is performed in each camera view, according to the block diagram illustrated in Fig. 2. It involves two distinct steps: (i) detecting objects on all individual frames and (ii) tracking and associating those detections across frames.

This paper investigates a practical method to MOT in which the emphasis is on high accuracy object detection and fast/precise object association. To that aim, detection quality has been identified as a critical element affecting tracking success. The Mask R-CNN architecture produces new bounding box detections at each time step [48]. The states of all existing objects up to the current frame are predicted. This is accomplished by running the appropriate Kalman filter's prediction step. The new detections are then associated with existing objects. Following that, the Kalman filters update step is run for all objects with associated detections. This includes the new measurements. However, there may be predictions with no assigned detections or detections with no assigned predictions. For unassigned detections, new tracks are constructed but only labelled as stable after three consecutive observations; otherwise, they were discarded.

The association heuristic employs various similarity metrics to determine whether or not a detection belongs to an existing track. These costs are calculated for each pair of existing tracks and new detections. As a result of the new detections, the predictors are updated, and the predicted positions are returned to the tracking process. The same method is repeated until the last frame. Meanwhile, a new tracker begins tracking objects that do not meet any of the predictions. The previously stopped trackers are removed. Missed detections, false alarms, similar appearance, groups, object fragmentation during object detection, and the association of related IDs due to numerous occlusions, and other unique behaviours are all challenges addressed in this study.

After objects are matched in a single view, then the same objects matching should be done between the camera views (For instance, the object which has been identified by ID=2, in Cam1 must be labeled ID = 2, in all views). In fusion part, first the probability distribution of each object's intensity histogram is obtained based on KDE, in each view. In the next step, the appropriate weight confidence is assigned to each camera due to the KDE. Thus, the weighted positions of the objects obtained from different views are aggregated by PD-OWA. In this study the MOT in multi-camera is developed in two parts: (i) Intra Tracking: Tracking objects in each camera view, (ii) Inter Tracking: Associate objects observed in multiple views.

4.1. Intra tracking

According to the flow diagram in Fig. 3, objects are primarily detected in a single view based on the Mask R-CNN model. For assigning the same ID to each object during the time step, several similarity criteria are evaluated as a feature. The detected objects in the previous and current frames are matched based on a variety of similarity metrics (SM) [1]: Hausdorff distance between objects, EMD distance between their colour histograms, and Chi-square distance between their Zernike Moments. Therefore, with the help of hard voting, the criteria are combined, and the appropriate ID is assigned to each object. The different steps and challenges observed in flow diagram Fig. 3, which we confront in intra tracking are as follows:

- At first, the pairs of bounding boxes which are candidate for matching in the K^{th} and $(K-1)^{th}$ frames must be overlapped more than 70 percentages.
- The object step length in the last five frames is a convenient indicator for rejecting or accepting an object as a candidate of label i, in K^{th} frame. It should be noted that the object step length is calculated according to the distance from the center of the objects in the K^{th} and $(K-1)^{th}$ frames.
- After, in two consecutive frames, an object cannot meet the threshold values due to the similarity criteria, then a new ID considered.
- Even after a brief occlusion, the declared ID must stay consistent throughout the tracking procedure. In order to distinguish the objects, Zernike Moments with order n = 10 will model the totality as well as the details of each object in 55 moments values (SM_{ZM}).
- The EMD similarity distance (SM_{EMD}) of two object's histograms and the Haussdorf similarity distance (SM_{HD}) of two object's bounding boxes can be obtained for the objects, which are candidate for matching.
- The Kalman filter is run for each object. While, the motion vector of each object will be predicted based on the object's centroid $[C_{x.i} \quad C_{y.i}]$.

4.2. Inter tracking

Of course, a set of cameras can provide additional information. So, the merged information can be better and more consistently interpreted about the scene [56]. In other words, to solve the problem of dense distribution of people and their occlusion, the fusion of

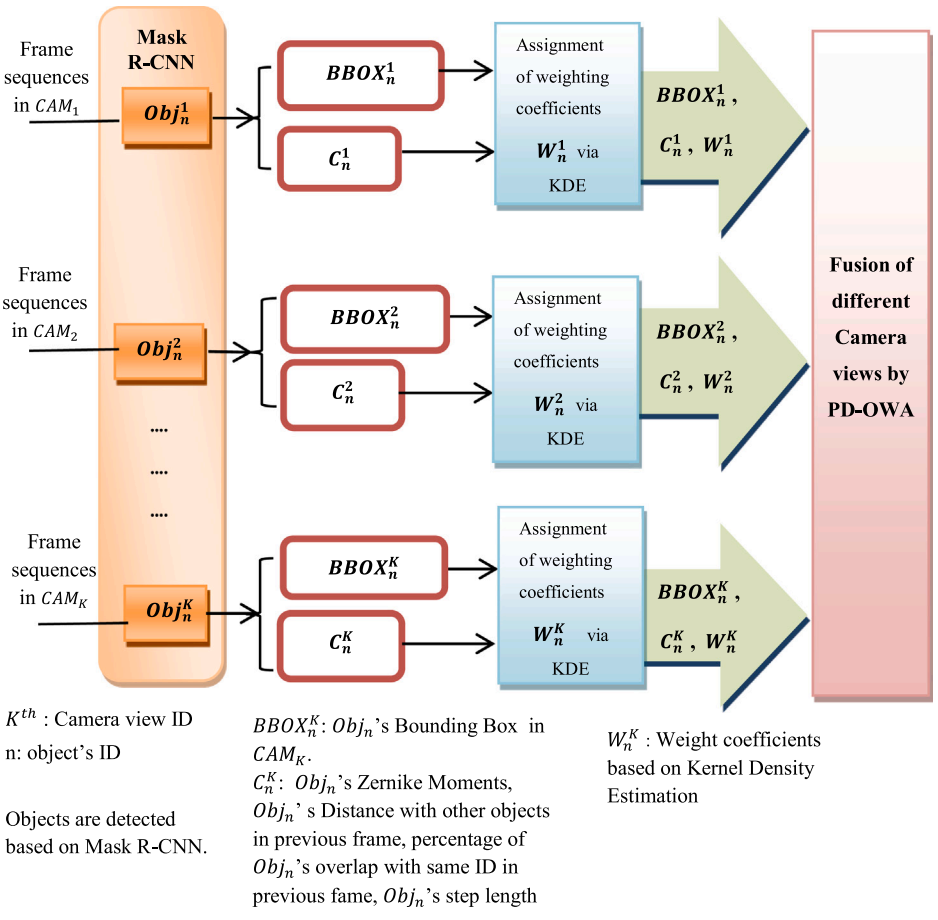

Fig. 4. The flow diagram of the proposed method in the stage of multi-object multi-camera tracking. It illustrates each camera view tracking process and at last, all information from different views will be fused.

multiple cameras is used to identify the occluded objects. In this regard, due to the variety of viewing angles, the object's coordinates should be estimated in a common reference (ground plane/top view). The tracking is carried out in each camera view, as shown in the flow diagram Fig. 4.

The inter tracking can be expressed in detail as follows:

- Based on the Mask R-CNN model, the objects detection and the initial mask extraction is done in any view.
- A new person who enters the scene will be identified and similarity characteristics will be calculated for him/her (such as Zernike Moments, etc.) in all views. Hence, they will be compared with the characteristics of the objects in the previous frames; in this regard, a new ID will be assigned or is matched to prior objects.
- A tracker is activated for each object so; each tracker will provide predictions for the subsequence frames. Some of them may not be precise due to a person leaving the scene or an occlusion occurring while tracking.
- Improved object matching will be done in different cameras and between different views by the same object identification along with tracking.
- We assign a reliability coefficient to each tracker based on KDE.
- Decide to choose the right view or to fuse all information via PD-OWA for right object tracking.

The inter tracking goal is to associate between objects information, across all views. Our MOT proposed method is based on the fusion of multiple cameras information using PD-OWA.

4.3. Similarity metrics

Since the colour feature, e.g. colour histogram, offers poor performance when the background colour is similar to the objects, in this research the Zernike Moments [49] are used as effective object's feature. The Zernike Moments are able to determine the overall object shape in low orders and represent the object details in high orders. The Zernike Moments are robust descriptors with orthogonality

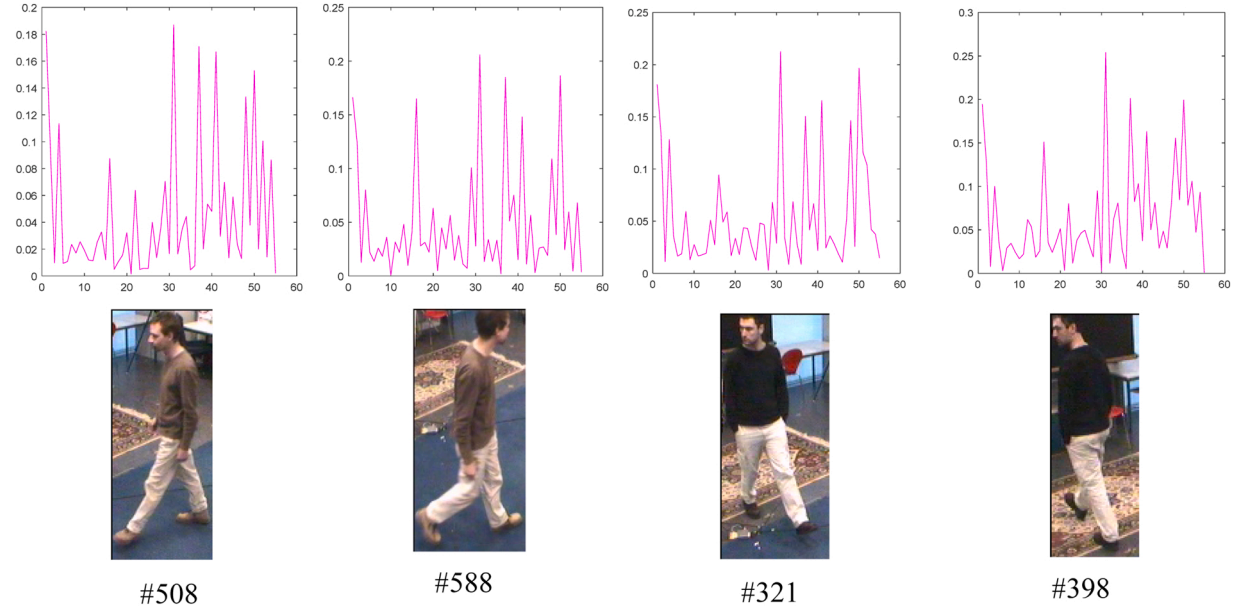

Fig. 5. Changes in Zernike Moments of order$n = 10$. It presents the detected object's Zernike Moments magnitudes change in different camera views for different people.

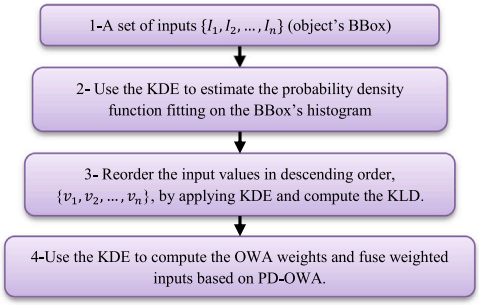

Fig. 6. The fusion part of proposed method for multi-object multi-camera tracking flow diagram.

feature, low noise sensitivity, and rotation insensitivity. Therefore, it is good for distinguishing objects in inter and intra tracking in different views. Thus, an occlusion predicted based on abrupt changes along with the magnitude of the Zernike Moments in an object path.

Among the criteria that are effectively improve the object matching results of this article are Hausdorff distance [50] and Earth Mover's Distance (EMD) [57]. The Hausdorff distance is calculated between two finite sets of points, including objects bounding boxes. The EMD method answers this question: What is the lowest cost to convert one distribution to another, assuming that two histograms have the same number of columns and frequencies [1]. The EMD can be stated in terms of a linear programming problem. The objective function denotes the set of all feasible amounts, f_{ij}, flows, d_{ij}, between bins. The linear programming is the solution of the optimal flow determination between the source and destination. For further study, one can refer to [1]. According to Fig. 5, the Zernike Moments change due to the different objects in multiple views.

4.4. Fusion by probability density-based OWA

One can choose different decision techniques throughout multi-criteria decision-making processes in order to manage the information of alternatives to rank them. The data of alternatives are fused into the overall values via the aggregation operators. In accordance with the overall values of alternatives, one can rank the whole alternatives and then obtain the optimum one.

The key point in determining the probability density on the basis of the OWA (PD-OWA) operator is how to determine the characteristics of data distribution for the input values [55]. The input values in the ordered weighted averaging operator are assumed to be identically distributed and independent while following a normal distribution. However, in most cases, the input values have an irregular distribution and may not follow the ideal normal distribution in real circumstances. As a result, in order to prevent

unreasonable assumptions, a new probability density technique is suggested, which employs KDE in order to evaluate the PDF that determines the distribution of the input values. Fig. 6 shows the fusion part of Fig. 4.

The implementation procedure of the suggested PD-OWA operator is presented as below:

1) In order to determine the input values probability distributions, the KDE is employed. The KDE is conducted on the histogram of each input, and Kullback Leibler (KLD) is used to compare objects distributions.
2) On the basis of the input's KDE, a certain set of input values $\{I_1, I_2, \ldots, I_n\}$ are reordered in descending order as $\{v_1, v_2, \ldots, v_n\}$.
3) The KDE function is also employed to estimate the weights of the reordered inputs on the basis of the probability density of each object.
4) The reordered input values related to their weights are aggregated in the form of a single value.

The KDE, as a nonparametric technique, can determine the PDF of the given input values without the need for prior knowledge of the properties of data distribution. In accordance with the KDE, the PDF of the reordered input values, $\{v_1, v_2, \ldots, v_n\}$, is estimable as Eq. (3) [55]:

$$\widehat{P}(v) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{v - v_i}{h}\right) \tag{3}$$

Where, n indicates the number of input values, $\widehat{P}(v)$ stands for the calculated PDF of the random variable v, $K()$ denotes a kernel, and h stands for a smoothing parameter, also known as bandwidth. One can cite the following three features for the kernel $K(.)$: (1) $K(x)$ is symmetric; (2) $\int_{-\infty}^{\infty} K(x)dx = 1$; and (3) for the entire values of x, $K(x) \geq 0$.

Based on the research and review performed on the different kernels, such as; the Gaussian Kernel (GK), Logistic Kernel (LK) and Uniform Kernel (UK) in [58], we decided to use the Gaussian kernel for density estimation. The choice of LK was motivated by the criticism of GK that its use might lead to a continuous distribution that does not preserve the higher moments of the original discrete distribution. The choice of UK was motivated by its similarity to the linear interpolation that is widely used in practice. In [58], it is suggested that the three kernels (with the various versions due to rescaling) provide very similar equating results and that despite the criticism, GK does well in preserving the higher order cumulants, the PRE, and the level of accuracy. The main differences between the three kernels seem to be their out-of-range characteristics (i.e., GK and LK have strictly positive density on the whole real line, but UK does not). So, in this study the Gaussian kernel is described as $K(x) = (1/(\sqrt{2\pi}))e^{-(1/2)x^2}$, in which $x = (v - v_i)/h$ [55].

Eq. (3) shows that if the number of input values is large enough, the accuracy of the calculated PDF is dependent on the kernel function $K(.)$ and the smoothing parameter h. According to statistical experiments, if the smoothing parameter remains constant, various kinds of kernel functions slightly affect the accuracy. Nonetheless, various smoothing parameter values greatly affect the accuracy. According to this rule, one should assign large weights to the input values with higher probability densities, while comparatively small weights should be assigned to lower probability densities. As a result, one can determine the weights of the ordered weighted averaging operator, on the basis of the estimated PDF, as Eq. (4) [55]:

$$w_j = \frac{\widehat{P}(v_j)}{\sum_{l=1}^{n} \widehat{P}(v_l)} = \frac{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{v_j - v_i}{h}\right)}{\sum_{l=1}^{n}\left(\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{v_l - v_i}{h}\right)\right)} = \frac{\sum_{i=1}^{n} K\left(\frac{v_j - v_i}{h}\right)}{\sum_{l=1}^{n}\sum_{i=1}^{n} K\left(\frac{v_l - v_i}{h}\right)} = \frac{\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{v_j - v_i}{h}\right)^2}}{\sum_{l=1}^{n}\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{v_l - v_i}{h}\right)^2}} \tag{4}$$

Where, $W = [w_1, w_2, \ldots, w_n]$ stands for the weight vector of the ordered weighted averaging operator satisfying $w_j \in [0, 1]$ and $\sum_{j=1}^{n} w_j \in [0, 1]$. One can define the PD-OWA operator on the basis of the above-cited weight determining technique as Eq. (5):

$$PD - OWA(I_1, I_2, \ldots, I_n) = \sum_{j=1}^{n} w_j v_j = \frac{\sum_{j=1}^{n}\left(\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{v_j - v_i}{h}\right)^2} v_j\right)}{\sum_{l=1}^{n}\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{v_l - v_i}{h}\right)^2}} \tag{5}$$

According to the orness, $O(W)$, and dispersion, $E(w)$, of the PD-OWA operator are computed as Eq. (6) and Eq. (7) [55]:

$$O(W) = \frac{1}{n-1} \sum_{j=1}^{n} (n - j)\left(\frac{\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{v_j - v_i}{h}\right)^2}}{\sum_{l=1}^{n}\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{v_l - v_i}{h}\right)^2}}\right) \tag{6}$$

$$E(w) = -\sum_{j=1}^{n}\left(\frac{\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{v_j - v_i}{h}\right)^2}}{\sum_{l=1}^{n}\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{v_l - v_i}{h}\right)^2}}\right) \ln\left(\frac{\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{v_j - v_i}{h}\right)^2}}{\sum_{l=1}^{n}\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{v_l - v_i}{h}\right)^2}}\right) \tag{7}$$

When $I_1 = I_2 = \ldots = I_n$, then we have $O(W) = 1/2$ and $E(w) = \ln n$. As shown in Fig. 7, the probability distributions of the objects were different from each other. Thus, the KDE is a proper tool to rearrange the inputs data in descending order, with the help of KLD.

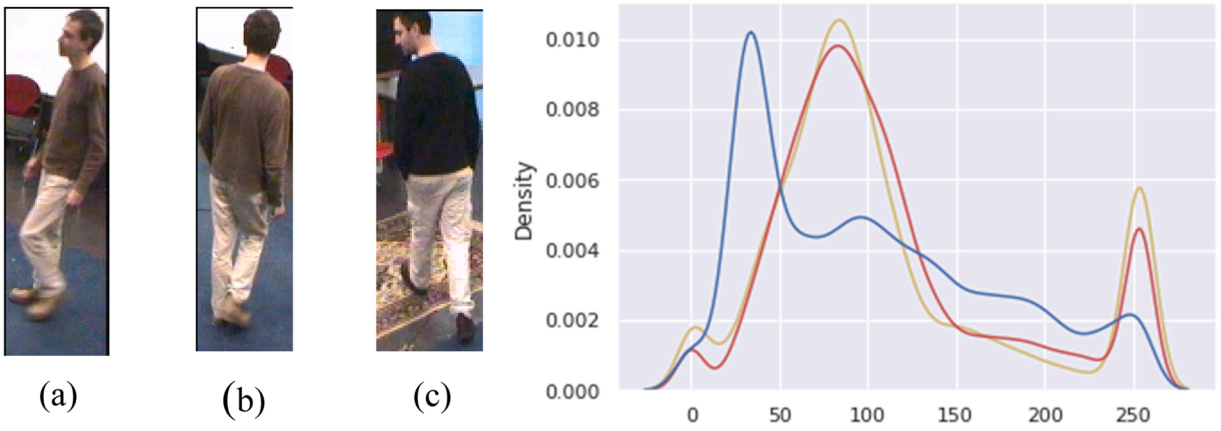

Fig. 7. Three detected objects and their kernel density estimations. Red, green and blue curves illustrate the KDE curves corresponding to objects (a), (b) and (c), respectively. By calculating the KLD between these three distributions it can be concluded that objects (a) and (b) are the same, due to very similar KDEs.

4.5. Data association

Data association and objects matching in the video sequence are influential topics in the MOT, both in intra and inter tracking. Association in the tracking process, is searching for the best assignment between foreground and the predicted objects at time t. In a single view, objects matching has been done according to the combination of similarity distances metrics, e.g. colour histogram distance calculated by EMD, images' Hausdorff distance related to the detected objects at time t and prediction of objects in time $t-1$. Thus, to improve the data association, the magnitude of the Zernike Moments of objects are compared by Chi-square distance according to Eq. (8) [1].

$$D(O_t, O_{t+1}) = \sum_{i=1}^{n} \frac{(O_{t_n} - O_{t_{n+1}})^2}{O_{t_n} + O_{t_{n+1}}} \tag{8}$$

Where; O_{t_n} and $O_{t_{n+1}}$ are the feature values of the objects. Subsequently, object matching is conducted based on the thresholded values. Since the number of detected objects is less than the predictions, a possibility of occlusion is predicted. Hence, the merging and splitting of the bounding boxes must be evaluated. So, the shape of the bounding box's metric is defined as R ratio, according to Eq. (9) [1].

$$R_i = \frac{Height}{Width} \quad , \quad \tau_{RatioDown} < R_i < \tau_{RatioUp} \tag{9}$$

Where; *Height* and *Width* are the height and the width of the bounding box of the object, respectively. The R criterion is limited by the upper bound threshold of $\tau_{RatioUp}$, and the lower bound threshold of $\tau_{RatioDown}$, in Eq. (9). Because of lighting, occlusion, shade merging, or the intersection of individuals' limbs, several objects in the scene may merge and be detected as a single object. If the value of R is greater than the predefined thresholds, it is assumed that several objects have been merged and must be separated.

4.6. Tracking

In this study, object tracking is performed using the Kalman filter. A tracker is launched for each new object that enters the scene in each camera view. Subsequently, the data association and object matching are performed based on the features of the objects. It should be noted that Kalman Filter's equations are not contextualized in this paper due to conciseness, but one can refer to [1,44,59] for more details.

5. Result and evaluation

In this section we present datasets, evaluation metrics, simulations and evaluation results of the proposed method. Our experimental results are propounded in detail in the case of Mask R-CNN detector, the MOT in single view, and fusion of multiple cameras information. This research aims to maintain the tracking of each object consistent throughout the MOT process in multi-camera datasets, therefore, each object is given a unique and new ID.

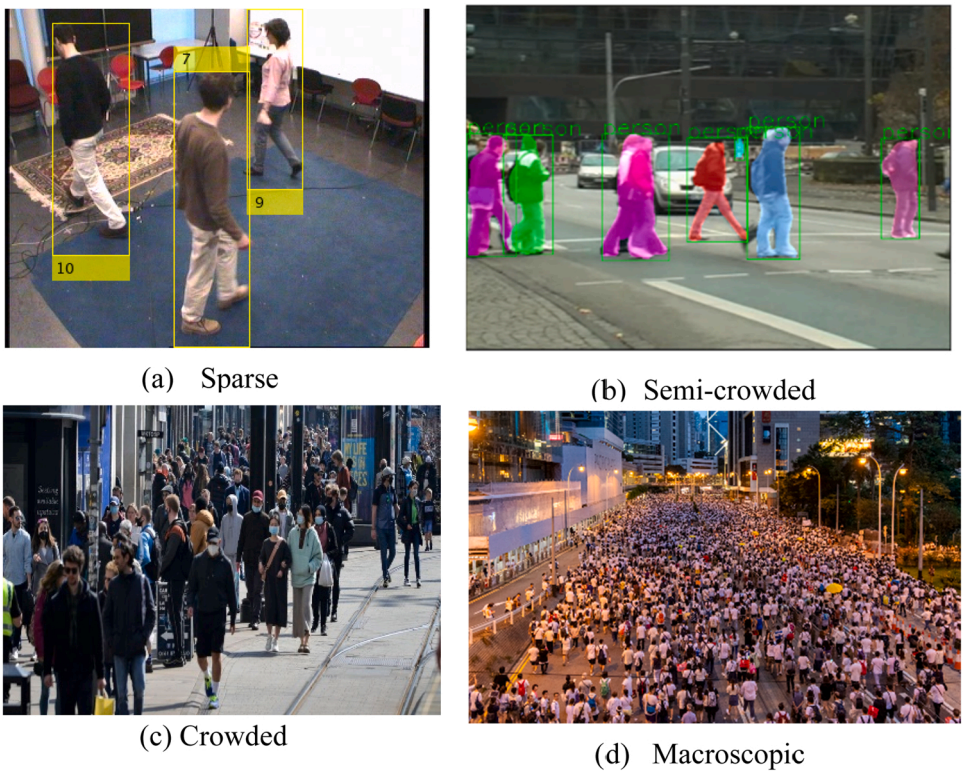

Fig. 8. Scenarios with different crowdness levels.

5.1. Dataset and evaluation metrics

The performance of the tracking implementation is simulated and assessed on a varied collection of testing sequences defined by the "MOT15"[1] benchmark database and video sequences from the multi-camera pedestrian video "EPFL"[2] and "PETS09-S2L1".[3] Several videos were captured concurrently from a particular area utilizing static cameras at different angles in these databases. Calibration information and homography matrices are given for each camera, as well as the homography matrix, H, which is provided in the calibration file in the "EPFL" database, so projects all points to the corresponding locations in the top view, Eq. (10).

$$H * X_{image} = X\_topview \tag{10}$$

It should be pointed that four scenarios with varying levels of crowdness can be assumed, depending on the number of people in the scene (Fig. 8) [60]: (a) Sparse: people are detected and tracked across the scene. (b) Semi-crowded: people can still be detected, but occlusions and missed detections are widespread, making tracking difficult. (c) Crowded: full-body pedestrian tracking is no longer viable. However, detection and tracking of heads are still possible. Person counting is a regular task for videos with this level of crowding. (d) Macroscopic scenario: As people cannot be detected well in these scenarios, the aim is often to determine the overall flow of the crowd.

Sparse and semi-crowded scenarios are focused throughout this research, as seen in Fig. 8.a and Fig. 8.b. Hence, due to the consecutive occlusion, it is basically impossible to detect the whole body in other scenarios, and it is recommended to rely on other measures in crowded videos, such as head detection. Of course, our proposed method will have low efficiency in crowded scenes, so it is considered as a limitation.

Accuracy and precision and the three major metrics of False Negative, False Positive, and Identity switch (IDs) are the usual criteria for assessing the MOT's quality. The criteria specified in this research were used to assess the tracking quality. Eq. (11) expresses Multi-Object Tracking Accuracy (MOTA). The missing objects, false positives, and identity switches at t are denoted by FN, FP, and ID_{sw}, respectively. $\overline{FN}$is the percentage of missing objects computed over all objects and frames [61]. $\overline{FP}$ denotes the rate of false positives, and $\overline{ID_s}$ denotes the rate of mismatches. Let also g_t, be the number of objects present at time t.

[1] https://motchallenge.net/data/MOT15/
[2] https://www.epfl.ch/labs/cvlab/data/data-pom-index-php/
[3] https://motchallenge.net/vis/PETS09-S2L1

Fig. 9. Results of object detection in EPFL dataset; Laboratory, Passage and Terrace video sequences.

$$MOTA = 1 - \frac{\sum_t (FN + FP + ID_{sw})}{\sum_t g_t},\ \overline{FN} = \frac{\sum_t FN}{\sum_t g_t},\ \overline{FP} = \frac{\sum_t FP}{\sum_t g_t},\ \overline{ID_{sw}} = \frac{\sum_t ID_{sw}}{\sum_t g_t} \tag{11}$$

Another criterion is Multi-Object Tracking Precision (MOTP). This criterion reflects the total object position error for the "object-prediction" pair over all frames. It demonstrates the tracker's capacity to estimate the exact location of an object. which is computed using Eq. (12) [61].

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_t C_t} \tag{12}$$

Where, C_t, is the number of matches found for time t. For each of those matches the distance $d_{i,t}$ is computed between i^{th} object and its corresponding hypothesis.

In the object detection, standard metrics apply to evaluate the results, such as AP (average precision), AP50, and IoU (intersection over union), to confirm result validity. They are a widely used and authoritative indicators for evaluating the performance of a deep network model in object detection and instance segmentation. AP50 is the IoU threshold (Eq. (13)), ranging from 0.50 to 0.95 with a 0.05 step [62].

$$IOU = \frac{Area\ \ of\ \ Overlap}{Area\ \ of\ \ Union} \tag{13}$$

5.2. Simulation and results

This section provides a report of the MOT simulations results, utilizing the multi-camera pedestrian video "EPFL" (Terrace sequences), PETS09-S2L1 and the MOT15 (TUD-Crossing) datasets. Throughout the MOT, each detected object is given an ID, meanwhile some objects may exit the scene while being tracked, or their lifetime may be shorter than the threshold level, or they may be occluded. Object matching is enhanced using a combination of similarity metrics e.g. EMD, Hausdorff, and Chi-square distances which respectively, measure histogram similarity distance, the distance between the blobs of objects, and the distance of Zernike Moments magnitudes of objects. Also, matching the same label through different views is done.

5.2.1. Experiment 1

Instance segmentation results were presented in this section. On the other hand, if there are objects that merged, they are split into

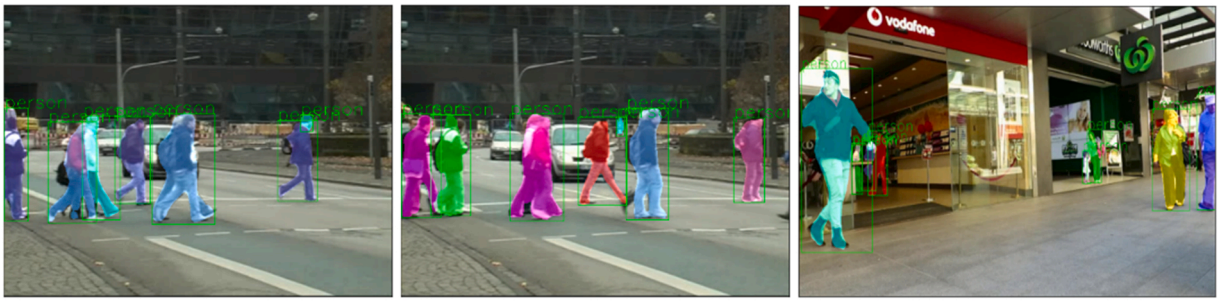

Fig. 10. Results of object detection in MOT15 dataset; MOT15: TUD-Crossing and ADL-Rundle-6.

Table 2
Object matching in EPFL video stream up to frame 1500, where the same IDs are matched based on similarity metrics.

	Matched IDs			
First object labelled as ID1	ID1	ID5	ID9	ID13
Second object labelled as ID2	ID2	ID6	ID12	
Third object labelled as ID3	ID8	ID11		

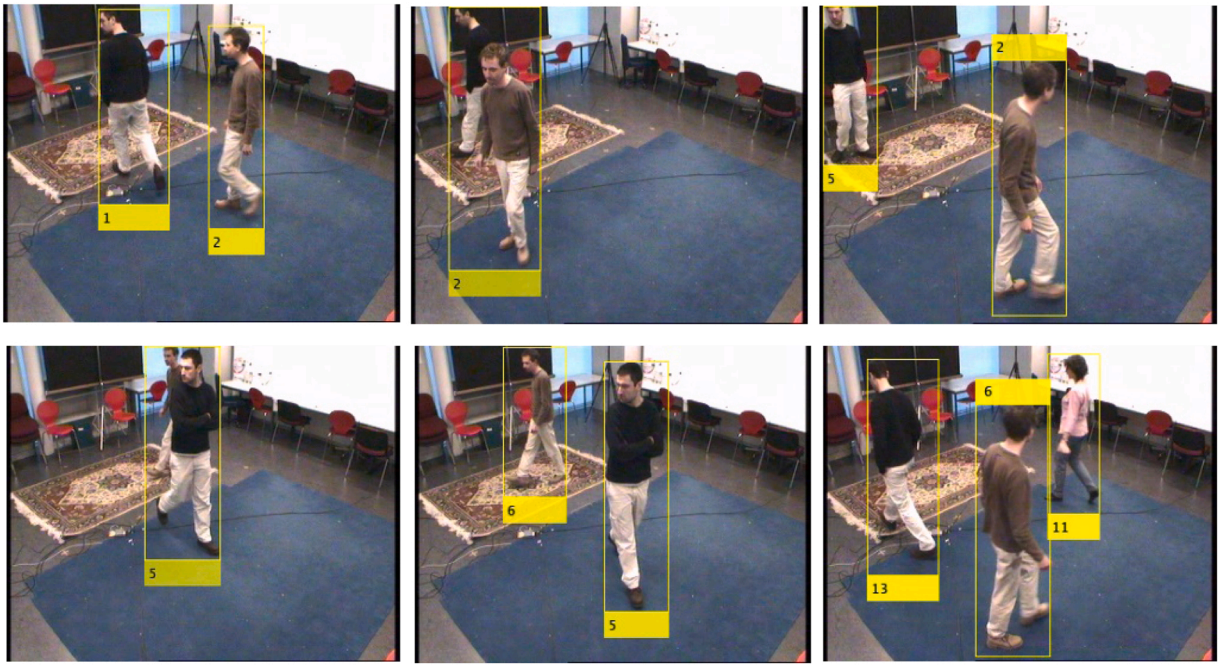

Fig. 11. Tracking process in consecutive frames by generating multiple IDs for the same object.

separate objects based on the R criterion. If the merged objects are detected, they will be divided into columnar patches, and their Zernike Moments will be compared employing the Chi-Square distance. The patches with the nearest distance in the magnitudes of the moments will be merged to form an object, and considered to the same object [1]. The COCO dataset[4] is used to train the suggested Mask R-CNN detector. The object detection was developed in GoogleColab, which was tested on a PC (Intel Core i7-7700 CPU, 16 GB of RAM, and an Nvidia Gforce 930x GPU), whilst Pytorch was used to develop the tracking component, which was tested on the aforementioned PC. Fig. 9 and Fig. 10, respectively illustrate the Mask R-CNN object detection result on the EPFL and the MOT15

[4] https://cocodataset.org/

Table 3
Comparing single view tracking results between the proposed method and state-of-art MOT techniques, based on TUD-crossing MOT15 sequences.

MOT15: TUD-Crossing					
	MOTA	MOTP	IDs	FP	FN
Two-granularity tracking [65]	58.3	73.1	13	180	267
Joint tracking [66]	53.9	72.8	15	37	456
Subgraph decomposition [67]	80.9	78.0	**1**	11	198
Tracklet MC + traj [63]	82.9	76.9	5	22	161
Minimum cost multicuts [68]	72.7	77.2	14	204	83
Joint multicut [63]	83.3	77.3	2	22	160
KCF [69]	31.5	74	2	53	634
MDP_SubCNN [70]	78.9	76.7	6	32	195
RFTrack [71]	80.8	70.2	7	42	163
DMT [72]	70	73.3	29	73	229
TSML_CDE [73]	75.7	75.8	11	18	239
Social Force Model [64]	87.1	73.6	3	39	100
Proposed method (Intra tracking)	**87.5**	**79.6**	11	**9**	**29**

datasets (TUD-Crossing and ADL-Rundle-6).

In MOT in each camera view, an ID is assigned to each object, but if occlusion occurs, the process of using previous IDs will stop in the next frame, and new IDs are assigned to the new merged objects. New IDs will be matched and the object tracking trajectories will be continued based on the matching to the initial IDs. It should be noted that the association of related IDs is an important challenge in MOT due to numerous occlusions. Hence, the number of IDs may temporarily increase during the occlusion. However, after object matching in consecutive frames, the number of IDs will certainly be the same as the number of objects in the scene.

Another factor that will lead to an increase in the number of IDs during tracking is that, due to lighting or partial occlusion, only a part of the object such as the head, arm or leg may be identified as object and an ID is assigned to an incomplete object. Therefore, this type of incomplete detected object does not match any of the objects in the previous frames, so these IDs will be removed during tracking.

The IDs assigned to the objects in each frame are matched to the IDs generated in the previous frames based on similarity metrics such as Husdorff distance, Earth Movers' distance and Zernike Moments. Hence, the trajectories of the same objects will be merged and connected after object matching, despite having different IDs. Table 2 shows the simulation results in matching steps based on similarity metrics, which are obtained from the dataset video up to 1500th frame. Hence, ID1 is the same as ID5 and ID9. ID2 is the same as ID6 and ID12. ID8 is also matched to ID11. Also Fig. 11 illustrates frames in which multiple IDs are generated for the same objects in the tracking process.

5.2.2. Experiment 2

The results of proposed method in MOT in the single view (flow diagram in Fig. 3), are compared with existing state-of-the-art MOT approaches [63,64] (Table 3). The suggested tracking technique was tested on the MOT15 benchmark dataset, which consists of 11 test videos and 11 training videos. Each video has a different duration, number of recordings, and quality and includes sparse and dense crowds, dynamic and static cameras, outdoor and indoor scenes of public areas. Most of the annotations are precise. Thus, we are reporting the tests results [63,64] on the 2D MOT15 TUD-crossing benchmark and compare with our proposed intra tracking method, which are given in Table 3. TUD Crossing contains 201 individual frames taken with a static camera about 2 m above the ground. In this dataset people walking on the street. Pedestrians continue to wander in front of the store, where all the pedestrians look alike after being occluded by others due to poor lighting. It is shown that our method achieves comparable results with the existing state-of-the-art algorithms. Our method has the advantages of employing Mask-RCNN to detect people and utilizing similarity metrics for object matching and data association. Based on the results, utilizing the proposed method the accuracy and the precision are increased due to the less FP and FN.

5.2.3. Experiment 3

Our suggested framework comprises of three major phases: (i) detection and extraction of feature, (ii) multi-object single-camera (MTSC) tracking, and (iii) multi-object multi-camera (MTMC) tracking. The tracking-by-detection was the paradigm followed in our technique. We employed Mask R-CNN in order to detect pedestrians and discover their bounding boxes. In the next step, the appearance features and similarity metrics are extracted from each detected object. These features are used throughout the association and object matching procedure of MTMC and MTSC by comparing their pairwise distance. Object matching, for each person, in the same camera is done through a feature-based similarity matching algorithm.

In this research, the final positions of the objects are obtained through the fusion of multiple cameras information based on PD-OWA. The OWA is used to aggregate input data from multiple cameras. For this purpose, KDE, object's pairwise similarity, and PDF are calculated for each object. Then the positions of the objects are reordered in descending form based on KLD by comparing their KDE. Hence, the coefficient weights are assigned to each object's coordinate obtained from each view and then are aggregated via the PD-OWA technique (Fig. 12).

Our proposed method focuses on tracking multiple objects in multiple cameras. It leads to an important reduction of identity

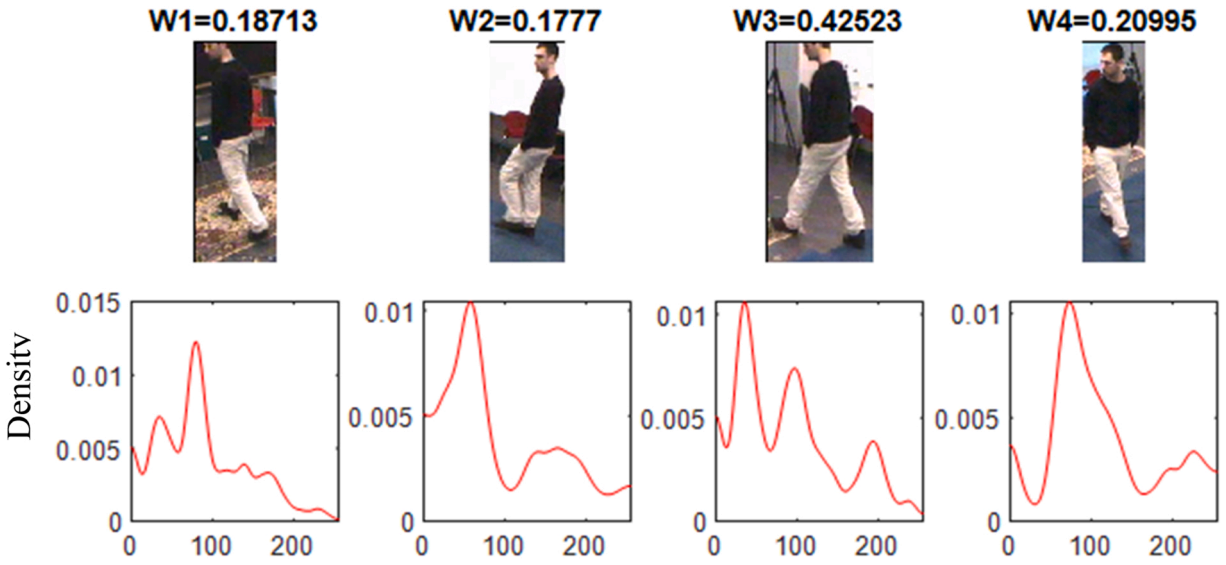

Fig. 12. Dedication of weight coefficients to each view in inter tracking process.

Table 4
Comparing MOT evaluation metric on MDP multi-view method with proposed method on "PETS09-S2L1" sequence.

Method	*FP*	*FN*	*IDs*	*MOTA*	*MOTP*
MDP multi-view [74]	1233	3379	**240**	69.8	68.9
Proposed method (Inter tracking)	**845**	**1642**	684	**81.6**	**79.4**

Table 5
Comparing MOT evaluation metric on MDP multi-view method with proposed method on "EPFL Terrace" sequence.

Method	*FP*	*FN*	*IDs*	*MOTA*	*MOTP*
MDP multi-view [74]	741	14,278	689	33.5	72.6
Proposed method (Inter tracking)	**448**	**1752**	**136**	**79.6**	**77.1**

Table 6
Ablation study of fusion task in multi-object tracking on "PETS09-S2L1" and "EPFL Terrace" sequences.

Fusion	Method	Data set	MOTA	MOTP
×	**Tracking by single view (Intra tracking)**	Cam.1 on "PETS09-S2L1" sequence	74.2	68.6
		CAM.0 on "EPFL Terrace" sequence	69.5	65.3
√	**Orness**	"PETS09-S2L1" sequence	71.6	64.1
		"EPFL Terrace" sequence	65.1	71.2
	Andness	"PETS09-S2L1" sequence	66.5	64.8
		"EPFL Terrace" sequence	62.5	63.7
	Aggregation (Proposed)	"PETS09-S2L1" sequence	**81.6**	**79.4**
		"EPFL Terrace" sequence	**79.6**	**77.1**

switches and a significant improvement in ID measures. Tables 4 and 5 present our proposed method evaluation metrics in the inter tracking MOT and compare the results with MDP (Markov Decision Process) multi-view [74], in detail. The results are released based on the "PETS09-S2L1" and "EPFL Terrace" datasets. On EPFL and PETS09-S2L1 sequences, our method effectively improved the FP, FN. But ID switch issue in the PETS dataset was not as improved as in the EPFL dataset. On the other hand, our proposed method has been able to significantly reduce FP and FN in the PETS dataset, and also these improvements are more evident in the EPL database.

5.2.4. Ablation study

In order to demonstrate the contribution of each module of our tracking framework, we disable or replace parts of the algorithm. In Table 6, the impact of fusion is addressed. In this case, object detection will be still based on the Mask R-CNN method. First, tracking by the single view is done on the Cam.1 on "PETS09-S2L1" sequence and on the CAM.0 on "EPFL Terrace" sequence. Hence, the MOTA has

Table 7
Ablation study of object detector in multi-object tracking on "PETS09-S2L1" and "EPFL Terrace" sequences.

Object detection	Data set	FP	FN	IDs	MOTA	MOTP
GMM	"PETS09-S2L1" sequence	3325	4829	1523	45.3	39.7
	"EPFL Terrace" sequence	3102	4125	1268	40.2	38.1
Faster R-CNN	"PETS09-S2L1" sequence	1428	2385	884	68.5	58.4
	"EPFL Terrace" sequence	759.2	2596	542	66.4	63.7
Mask R-CNN (Proposed)	"PETS09-S2L1" sequence	**845**	**1642**	**684**	**81.6**	**79.4**
	"EPFL Terrace" sequence	**448**	**1752**	**136**	**79.6**	**77.1**

significantly decreased. Now, by adding different views, new and more useful information is received from other views. In the orness row of Table 6, fusion is based solely on selecting the largest weight as the final weight determining the object position, and the rest of the weight coefficients will be zero. Once again, in the andness row, the fusion is based on selection of the smallest weight coefficient as the final weight determining the position of the object, and the rest of the weight coefficients are zero. The MOTA of these two fusion methods, has improved from that without fusion. Nevertheless, in our proposed method, a weight coefficient is assigned to each object position in each camera view, and finally the sum of the product of the weight coefficients and the object position in each view will be determined for the final object position. It is clear that MOTA of our method has shown much higher accuracy and efficiency.

Object detector is one of the main and effective components in the MOT. Table 7 examines the different object detectors and evaluates their impact on the tracking process. As it turns out, first a general-purpose detector such as Gaussian Mixture Model (GMM) was used, which due to the successive occlusion between the objects, and the lack of correct detection, the correct matching will not occur during the tracking and consequently the false positive and false negative errors increase. Consequently, tracking accuracy has been significantly reduced. By replacing the Faster R-CNN, as can be seen, the tracking accuracy, MOTA, will increase. However, applying the Mask R-CNN object detector, we have been able to achieve 81.6% and 79.6% accuracy on the "PETS09-S2L1" sequence and the "EPFL Terrace" sequence, respectively.

6. Conclusion

In this paper, we present a novel multi-object tracking method based on the multi-view data fusion. This study aimed to improve object detection and object matching in intra tracking, respectively by employing Mask R-CNN and similarity-based object matching. The superiority of our method is in estimating the position of objects in the scene more accurately by fusion of multi-view information based on the PD-OWA aggregation algorithm. Accordingly, the KDE is employed for weight assignment to the input data from each view. The test on two datasets, EPFL and PETS09, has verified the robustness of our multi-view fusion approach dealing with the problem of re-identification and occlusion. Moreover, a technique is utilized based on similarity-based hard voting for object matching in intra tracking and Zernike Moments features for data association in inter tracking. Several criteria based on various characteristics of detected objects, in consecutive frames, were evaluated. In this case, Hausdorff and EMD distance criteria were utilized as distance metrics rather than Chi-square distance. Additionally, in this research, the separation of the merged objects reduces false negatives and can cope with people's dense distribution and mutual occlusion in the tracking process. Experimental results show a solid improvement over state-of-art algorithms.

CRediT authorship contribution statement

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

Declaration of Competing Interest

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Acknowledgement

The authors acknowledge the funding support of Babol Noshirvani University of Technology through Grant program No. BNUT/370123/00. The authors would also like to thank the editor and anonymous reviewers for the useful and constructive comments which have significantly improved the article.

References

[1] A. Dadgar, Y. Baleghi, M. Ezoji, Improved object matching in multi-objects tracking based on zernike moments and combination of multiple similarity metrics, Int. J. Eng. 34 (6) (2021) 1445–1454.
[2] Asvadi, A., M. Karami, Y. Baleghi. Object tracking using adaptive object color modeling, in: Proceedings of the Fouth Conference on Information and Knowledge Technology, 2012.

[3] S.S.A. Rajjak, A. Kureshi, Multiple-object detection and segmentation based on deep learning in high-resolution video using mask-RCNN, Int. J. Pattern Recognit. Artif. Intell. (2021), 2150038.
[4] Asvadi, A., et al. Incremental discriminative color object tracking, in: Proceedings of the International Symposium on Artificial Intelligence and Signal Processing, 2013, Springer.
[5] Asvadi, A., et al. Improved object tracking using radial basis function neural networks, in: Proceedings of the 2011 Seventh Iranian Conference on Machine Vision and Image Processing, IEEE, 2011.
[6] A. Asvadi, et al., Online visual object tracking using incremental discriminative color learning, CSI J. Comput. Sci. Engine 12 (24) (2014) 16–28.
[7] A. Asvadi, M. Karami, Y. Baleghi, Efficient object tracking using optimized K-means segmentation and radial basis function neural networks, Int. J. Inf. Commun. Technol. Res. 4 (1) (2012) 29–39.
[8] S. Zhou, et al., A Survey of Multi-object Video Tracking Algorithms, Springer International Publishing, Cham, 2019.
[9] Y. Xu, et al., Deep learning for multiple object tracking: a survey, IET Comput. Vis. 13 (4) (2019) 355–368.
[10] J. Chen, et al., Online multiple object tracking using a novel discriminative module for autonomous driving, Electronics 10 (20) (2021) 2479.
[11] Luo, W., et al., Multiple object tracking: a literature review, arXiv preprint arXiv: 14097618, 2014.
[12] A. Asvadi, et al., Incremental Discriminative Color Object Tracking, Springer International Publishing, Cham, 2014.
[13] S. Pouyanfar, et al., A survey on deep learning: Algorithms, techniques, and applications, ACM Comput. Surv. (CSUR) 51 (5) (2018) 1–36.
[14] A. Khan, et al., A survey of the recent architectures of deep convolutional neural networks, Artif. Intell. Rev. 53 (8) (2020) 5455–5516.
[15] M. Khare, N.T. Binh, R.K. Srivastava, Human object classification using dual tree complex wavelet transform and zernike moment. in: Transactions on Large-scale Data-and Knowledge-centered Systems xvi, Springer, 2014, pp. 87–101.
[16] Górniak, A. ,E. Skubalska-Rafajłowicz. Object classification using sequences of zernike moments, in: Proceedings of the IFIP International Conference on Computer Information Systems and Industrial Management, 2017, Springer.
[17] M. Kakooei, Y. Baleghi, A two-level fusion for building irregularity detection in post-disaster VHR oblique images, Earth Sci. Inform. 13 (2) (2020) 459–477.
[18] Zhu, C., Multi-Camera People Detection and Tracking. 2019.
[19] R.R. Yager, Families of OWA operators, Fuzzy Sets Syst. 59 (2) (1993) 125–148.
[20] E. Cables Pérez, et al., On OWA linear operators for decision making, Fuzzy Inf. Eng. 10 (1) (2018) 80–90.
[21] R. Iguernaissi, et al., People tracking in multi-camera systems: a review, Multimed. Tools Appl. 78 (8) (2019) 10773–10793.
[22] M. Lenormand, Generating OWA weights using truncated distributions, Int. J. Intell. Syst. 33 (4) (2018) 791–801.
[23] X. Yang, et al., Probability interval prediction of wind power based on KDE method with rough sets and weighted Markov chain, IEEE Access 6 (2018) 51556–51565.
[24] M. Manafifard, H. Ebadi, H.A. Moghaddam, A survey on player tracking in soccer videos, Comput. Vis. Image Underst. 159 (2017) 19–46.
[25] Z. Sun, et al., A survey of multiple pedestrian tracking based on tracking-by-detection framework, IEEE Trans. Circuits Syst. Video Technol. 31 (5) (2020) 1819–1833.
[26] Berclaz, J., F. Fleuret, P. Fua. Multiple object tracking using flow linear programming, in: Proceedings of the 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2009, IEEE.
[27] Chari, V., et al. On pairwise costs for network flow multi-object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
[28] H.B. Shitrit, et al., Multi-commodity network flow for tracking multiple people, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2013) 1614–1627.
[29] N. Nikbakhsh, Y. Baleghi, H. Agahi, Maximum mutual information and Tsallis entropy for unsupervised segmentation of tree leaves in natural scenes, Comput. Electron. Agric. 162 (2019) 440–449.
[30] N. Nikbakhsh, Y. Baleghi, H. Agahi, A novel approach for unsupervised image segmentation fusion of plant leaves based on G-mutual information, Mach. Vis. Appl. 32 (1) (2021) 1–12.
[31] Khemmar, R., et al., Real time pedestrian and object detection and tracking-based deep learning, Application to Drone Visual Tracking. 2019.
[32] C.B. Murthy, et al., Investigations of object detection in images/videos using various deep learning techniques and embedded platforms—a comprehensive review, Appl. Sci. 10 (9) (2020) 3280.
[33] P. Bharati, A. Pramanik, Deep learning techniques—R-CNN to mask R-CNN: a survey. in: Computational Intelligence in Pattern Recognition, Springer, 2020, pp. 657–668.
[34] L. Liu, et al., Deep learning for generic object detection: a survey, Int. J. Comput. Vis. 128 (2) (2020) 261–318.
[35] X. Zhang, et al., Research on object detection model based on feature network optimization, Processes 9 (9) (2021) 1654.
[36] A. Uçar, Y. Demir, C. Güzeliş, Object recognition and detection with deep learning for autonomous driving applications, Simulation 93 (9) (2017) 759–769.
[37] Z.-Q. Zhao, et al., Object detection with deep learning: a review, IEEE Trans. Neural Netw. Learn. Syst. 30 (11) (2019) 3212–3232.
[38] Zhou, X., et al. Application of deep learning in object detection, in: Proceedings of the IEEE/ACIS Sixteenth International Conference on Computer and Information Science (ICIS), IEEE, 2017.
[39] Wojke, N., A. Bewley, D. Paulus. Simple online and realtime tracking with a deep association metric, in: Proceedings of the IEEE International Conference on Image Processing (ICIP, IEEE, 2017.
[40] Azhar, M.I.H., et al. People tracking system using DeepSORT, in: Proceedings of the Tenth IEEE International Conference on Control System, Computing and Engineering (ICCSCE), IEEE, 2020.
[41] G. Ciaparrone, et al., Deep learning in video multi-object tracking: a survey, Neurocomputing 381 (2020) 61–88.
[42] Wang, Z., et al. Towards real-time multi-object tracking, in: Proceedings of the Sixteenth European Conference, Computer Vision–ECCV 2020, Glasgow, UK, August 23–28, 2020, Part XI 16. 2020, Springer.
[43] Li, X., et al. A multiple object tracking method using Kalman filter, in: Proceedings of the IEEE International Conference on Information and Automation, IEEE, 2010. 2010.
[44] M. Soleh, G. Jati, M.H. Hilman, Multi object detection and tracking using optical flow density–Hungarian Kalman Filter (Ofd-Hkf) algorithm for vehicle counting, J. Ilmu Komput. Dan. Inf. 11 (1) (2018) 17–26.
[45] L. Svensson, et al., Set JPDA filter for multitarget tracking, IEEE Trans. Signal Process. 59 (10) (2011) 4677–4691.
[46] L. Ying, T. Zhang, C. Xu, Multi-object tracking via MHT with multiple information fusion in surveillance video, Multimed. Syst. 21 (3) (2015) 313–326.
[47] S. Oh, S. Russell, S. Sastry, Markov chain Monte Carlo data association for multi-target tracking, IEEE Trans. Autom. Control 54 (3) (2009) 481–497.
[48] He, K., et al. Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision. 2017.
[49] Shams-Baboli, A. , M. Ezoji. A Zernike moment based method for classification of Alzheimer's disease from structural MRI. in: Proceedings of the Third International Conference on Pattern Recognition and Image Analysis (IPRIA), IEEE, 2017.
[50] A.A. Taha, A. Hanbury, An efficient algorithm for calculating the exact Hausdorff distance, IEEE Trans. Pattern Anal. Mach. Intell. 37 (11) (2015) 2153–2163.
[51] Y. Rubner, C. Tomasi, L.J. Guibas, The earth mover's distance as a metric for image retrieval, Int. J. Comput. Vis. 40 (2) (2000) 99–121.
[52] S. Ren, et al., Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015) 91–99.
[53] Lin, T.-Y., et al. Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[54] L.-G. Zhou, H.-Y. Chen, Continuous generalized OWA operator and its application to decision making, Fuzzy Sets Syst. 168 (1) (2011) 18–34.
[55] M. Lin, et al., Determine OWA operator weights using kernel density estimation, Econ. Res. Èkon. istraživanja 33 (1) (2020) 1441–1464.
[56] Liu, W., O. Camps, M. Sznaier, Multi-camera multi-object tracking, arXiv preprint, arXiv:1709.07065, 2017.
[57] Z. Zuo, et al., Cross-modality earth mover's distance-driven convolutional neural network for different-modality data, Neural Comput. Appl. (2019) 1–12.
[58] Lee, Y.H. , A.A. von Davier, Comparing alternative kernels for the kernel method of test equating: gaussian, logistic, and uniform kernels. ETS Research Report Series, 2008. 2008(1), i-26.

[59] Azari, M., A. Seyfi, A.H. Rezaie. Real time multiple object tracking and occlusion reasoning using adaptive kalman filters, in: Proceedings of the Seventh Iranian Conference on Machine Vision and Image Processing, IEEE, 2011.
[60] Leal-Taixé, L., Multiple object tracking with context awareness, ArXiv, 2014. abs/1411.7935.
[61] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the clear mot metrics, EURASIP J. Image Video Process. 2008 (2008) 1–10.
[62] He, K., et al., Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, 2017, 2980–2988.
[63] Keuper, M., et al., A multi-cut formulation for joint segmentation and tracking of multiple objects. arXiv preprint, arXiv:1607.06317, 2016.
[64] Y. Xue, Z. Ju, Multiple pedestrian tracking under first-person perspective using deep neural network and social force optimization, Optik 240 (2021), 166981.
[65] Fragkiadaki, K., et al. Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions, in: Proceedings of the European Conference on Computer Vision, 2012, Springer.
[66] Milan, A., et al. Joint tracking and segmentation of multiple targets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[67] Tang, S., et al. Subgraph decomposition for multi-target tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[68] Keuper, M., B. Andres, T. Brox. Motion trajectory segmentation via minimum cost multicuts, in: Proceedings of the IEEE International Conference on Computer Vision, 2015.
[69] Chu, P., et al. Online multi-object tracking with instance-aware tracker and dynamic model refreshment, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019 2019.
[70] Xiang, Y., A. Alahi, S. Savarese. Learning to track: Online multi-object tracking by decision making, in: Proceedings of the IEEE International Conference on Computer Vision, 2015.
[71] J. Xiang, et al., End-to-end learning deep CRF models for multi-object tracking deep CRF models, IEEE Trans. Circuits Syst. Video Technol. 31 (1) (2020) 275–288.
[72] Kim, H.-U. , C.-S. Kim. CDT: Cooperative detection and tracking for tracing multiple objects in video sequences, in: Proceedings of the European Conference on Computer Vision, 2016, Springer.
[73] B. Wang, et al., Tracklet association by online target-specific metric learning and coherent dynamics estimation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (3) (2016) 589–602.
[74] Le, Q.C., D. Conte, M. Hidane. Online multiple view tracking: Targets association across cameras, in:Proceedings of the Sixth Workshop on Activity Monitoring by Multiple Distributed Sensing (AMMDS 2018), 2018.