Multimedia Tools and Applications (2020) 79:26023–26043
https://doi.org/10.1007/s11042-020-09242-5

Efficient vehicle detection and tracking strategy in aerial videos by employing morphological operations and feature points motion analysis

Ahmed Gomaa[1,2,3] · Moataz M. Abdelwahab[2] · Mohammed Abo-Zahhad[2,4]

Received: 25 June 2019 / Revised: 1 May 2020 / Accepted: 24 June 2020 /
Published online: 9 July 2020

Abstract

Real-time automatic detection and tracking of moving vehicles in videos acquired by airborne cameras is a challenging problem due to vehicle occlusion, camera movement, and high computational cost. This paper presents an efficient and robust real-time approach for automatic vehicle detection and tracking in aerial videos that employ both detections and tracking features to enhance the final decision. The use of Top-hat and Bottom-hat transformation aided by the morphological operation in the detection phase has been adopted. After detection, background regions are eliminated by motion feature points' analysis of the obtained object regions using a combined technique between KLT tracker and K-means clustering. Obtained object features are clustered into separate objects based on their motion characteristic. Finally, an efficient connecting algorithm is introduced to assign the vehicle labels with their corresponding cluster trajectories. The proposed method was tested on videos taken in different scenarios. The experimental results showed that the recall, precision, and tracking accuracy of the proposed method were about 95.1 %, 97.5%, and 95.2%, respectively. The method also achieves a fast processing speed. Thus, the proposed approach has superior overall performance compared to newly published approaches.

Keywords Morphological operations · Aerial surveillance · Remote sensing · KLT tracker · K-means clustering · Vehicle detection and tracking.

1 Introduction

Traffic management technology using cameras mounted on drones or airplanes is becoming a hot topic [18, 29]. Recently, airborne surveillance provides more advantages than traditional monitoring techniques, as a stationary camera and bridge sensors. Such methods are providing more coverage with lessen expenses and are better in emergency response. Moreover, UAVs are highly portable to collect traffic data in an area with difficult geographic locations, when the conventional data-gathering techniques cannot be applied. However, the

✉ Ahmed Gomaa
ahmed.gomaa@ejust.edu.eg

Extended author information available on the last page of the article.

data collected by UAV cameras are needed to turn into useful resources. For traffic monitoring concerns, vehicle detection and tracking are considered as essential and challenging tasks. While there are many existing approaches designed for handling surveillance detection and tracking challenges such as shadows, occlusion, and reflections using cameras at a fixed location, the most challenging issue in the aerial video is the camera movement[20]. Both the background and foreground, in this case, are moving in the image due to the irregular motion of the camera that mounted on the UAV.

The traditional approaches to extract traffic information using static cameras based on blob detection or background subtraction cannot efficiently work with UAV-based videos. Vehicle detection and tracking are usually handled as two separate processes. Vehicle detection in images depends on spatial appearance features, while vehicle tracking depends on both spatial appearance and temporal motion features. Detection strategies of vehicle detectors based on visual and appearance features of vehicles' spatial domain have started to saturate [2, 27, 36], and the tracking strategies depend on manually vehicle detection initialization is not efficient in real applications [7, 8, 16]. In literature, various approaches have been introduced to tackle the detection and tracking problems for vehicles in images and videos. However, it is still an open issue due to the changes in appearance, scale, view of the objects, shadows, camera movement, illumination conditions, and partial occlusion. While significant progress has been made for the detection or tracking separately, most of them still need a high computation complexity with low accuracy.

Most vehicle surveillance strategies depend mainly on vehicle detection or vehicle tracking, with good accuracy, yet some false positive and false negative results have occurred. It is very difficult to achieve vehicle detection and tracking in real-time using airborne videos, so some researchers focus only on the detection [17, 24, 27, 29, 30, 32, 38] or tracking problem [7, 8, 11]. Moreover, tracking multi-vehicles with shadow and illumination changes due to day light and moving clouds at the same time is another challenge with different vehicles speed. Not much effort has done so far for using tracking information to improve the accuracy of object detection and to reduce the storage and training cost. Moreover, the object detector usually suffers from missing detection and false positives, which deteriorate the tracking process. Because of such issues, this paper presents an effective collaborative strategy between detection and tracking information to improve both the detection and tracking processes.

Detection and tracking of vehicles simultaneously are considered an efficient strategy for achieving accurate detection and tracking results. The morphological operations have been used in object detection with better foreground and background discrimination [38], but the performance is not perfect. The detection and tracking accuracy can be enhanced by reducing false-positive and false-negative results using a refining process to improve the foreground detection by the morphological operations. Hence, new vehicle detection and tracking strategy is suggested in this paper based on the collaboration process between collected detection information using Top-hat and Bottom-hat transformation and tracking information using KLT tracker. The work depends on the feature points information analysis between the fixed number of frames, and using the temporal information of the detection and tracking feature points between the frame sets to achieve better detection and tracking decisions.

In this work, we exploited the features of both the detection and tracking information to solve the problems associated with each of them separately. An efficient collaborative real-time approach is introduced that enhances the interaction between a non-trained object detector utilizing Top-hat and Bottom-hat transformation and KLT-based tracking methodology. This paper introduces a new simultaneous vehicle detection and tracking method by developing a three-step approach in each frameset. First, the power of Top-hat & Bottom-hat is exploited in the vehicle detection process before the vehicle refining and clustering process in the second step using the optical flow and k-means clustering. The Top-hat & Bottom-hat is used in the first frame, while the refining analysis considering the remaining frames in the frameset. Thus, a robust discrimination process between the foreground vehicles and noisy background regions is utilized. Thirdly an effective connecting strategy is offered to assign each vehicle with its corresponding trajectory based on the collected detection and tracking information. The detection and tracking accuracy is increased, and the algorithm works efficiently with different and challenging environments. Experiments carried on airplane videos confirm our superior performance in terms of precision, recall, tracking accuracy, and time performance competed to the existing techniques. Also, the method can be used in real-time applications with minimum execution time.

The paper is organized as follows: Related work is introduced in Section 2. The proposed vehicle detection and tracking method are described in Section 3. In Section 4, the experimental results are given and evaluated. Finally, the work is concluded in Section 5.

2 Related work

Most of the existing approaches handled one of the research topics, either vehicle detection or tracking. Vehicle detection can be categorized into two groups; vehicle detection methods based on machine learning techniques, and vehicle detection methods based on segmentation techniques. Vehicle detection methods based on the machine learning methods, including a sliding window [24, 28, 30] and appearance features method [2, 27, 36], always follow the processes of finding the vehicle features first and then categorize them into classification models that identify the group of the testing vehicle. These methods can only detect targets from pre-learned object classes. In [27], a support vector machine (SVM) was used for the detection process after pixel classification using AdaBoost classifier trained on Haar-like features. Deep learning strategies are suggested in [10, 12, 14, 22] for object detection, but the detection accuracy of these strategies depends on the trained images with more computation complexity. The authors of [37], provide a review and discussions on deep learning strategies for object detection. Vehicle detection methods based on segmentation techniques including morphological operations[38], thresholding [21, 35], and edge detection [31] in combination with connected component labeling. To further enhance those methods, several detection approaches were presented in[17] and [32]. In [17], the authors proposed a new technique that combined visual, temporal, and spatial features. Firstly, they divided the video frames into uniform regions before forming a region adjacency graph that assembles each of the frames. Then they proposed a multi-graph matching algorithm to match each of these regions with their corresponding through the frames. Finally, a graph coloring methodology was presented to efficiently labeling the objects in the background or foreground. Recently, authors in [32] used image stacking in a novel way, they employing image registration to small vehicles only, and the warping process has been used to blur all the stationary background in the neighborhood of the moving vehicles. The main goal of this algorithm is to remove disturbing image parts of the background, which can be smoothed to extract the observed vehicle from the background. However, previous strategies demand a high computational requirement.

Regarding vehicle tracking, various efforts have been made concerning multiple targets tracking in aerial videos [4, 7, 8, 23, 26, 34]. Existing methods can be classified into two categories: tracking through detection and tracking through prediction. In tracking through detection methods [23, 34], also called Data Association Tracking (DAT), first, the vehicles are localized in the scene using different detection strategies, then the identification of each one is assigned via proper data association schemes. The identity assignment process, in this case, is handled by various data association methods. Multi-Hypothesis Tracking (MHT) [26] and Joint Probabilistic Data Association (JPDAF) [4] are the two widely used basic algorithms but associated with some assumption that making these algorithms non-ideal in tracking a large number of targets. Moreover, they have a high computational cost. Tracking through prediction methods managed to track multiple targets by implementing several predictions based on single target trackers (STT). Each STT performs the prediction based tracking through two main steps. First, an estimated region of interest is predicted, then the localization of the target is conducted by the tracker in this region. In general, the prediction phase in the first step is established using various target motion models. The Kalman filter is used for tracking linear object movements, which means the object travel with constant movement parameters (acceleration or velocity) [25]. However, the linear motion model is not qualified as a non-linear model to handle the diversified target dynamics in the aerial surveillance videos. Non-linear motion model is handled using Extended Kalman filter (EKF) [5], Unscented Kalman Filter (UKF) [33], and Particle Filter (PF) [11, 15]. Although the effectiveness of such non-linear motion models in handling the tracking challenges, yet their performance in tracking a large number of targets is difficult due to the complexity of the model. Furthermore, any error that occurs in the prediction phase may lead to accumulated and irreversible tracking failure due to the absence of detection responses. Tracking strategy based on a dynamic grouping method was introduced in [8] where the authors used a collaborative framework, that involves a two-level tracking manner; namely, high, and low levels of tracking. The higher-level part forms a relevance network and divides vehicles into several groups based on the velocity magnitude and direction, where the relevance is estimated using the obtained information from the lower level with Kalman filtering and histogram matching help. However, grouping vehicles are difficult when their movement is not always uniform. A hierarchical layered structure for tracking is proposed in [7] to build an inter-object complementary assistance system for vehicle tracking. This system provided more reliable information about the motion model using a novel nonlinear motion and interaction model. However, the previous strategies have some limitations, including the manual detection initialization process, more complicated, and they cannot work efficiently with different motion models.

Recently, authors are working on simultaneous vehicle detection and tracking as in [1, 3, 16]. In [16], a hybrid framework has been developed for vehicle detection and tracking. The authors combined the frame difference technology with the background subtraction scheme in the detection step after the frame registration process. Then, they proposed a non-linear motion model, called adaptive velocity adaptive acceleration (AVAA), for the tracking purpose. However, the objects bound boxes are initialized drawing by hand in the first frames, to produce the initial working velocity and acceleration. Unfortunately, this strategy has a low processing speed. In [3], the authors proposed a multiscale target hypothesis in the detection part where the vehicles were characterized by a blob detection method as the Canny edge detector. Then the temporal consistencies of this hypothesis were tested to eliminate the outlier observations and enhance the detection results. Finally, the tracking was accomplished by reducing the spatial distance between associated vehicles in consecutive frames. Nevertheless, the disadvantage of this method is canny edge detector failures despite the good achievable precision result. Authors in [1], proposed a new vehicle detection and tracking strategy using Harries and Lucas-Tomasi feature tracker to detect good features that tracked in the image frames. The background points were removed by measuring the changes in the histogram of the pixels around each point with time to obtain the foreground features. Then they clustered them into separate trackable vehicles, according to their movement angles and displacement magnitudes. The algorithm achieved real-time performance for detection and tracking vehicles in airborne videos with low precision and recall accuracy. This paper introduces a new approach to achieve fast and accurate vehicle detection and tracking. The detected vehicles are extracted in the first frame of each frameset, by analyzing their feature characteristics between the remaining consecutive frames and the background regions were excluded. The vehicle's features are updated every fixed number of frames associated with an effective algorithm to assign each vehicle with its corresponding trajectory.

3 The proposed algorithm

The main steps of the proposed automatic vehicle detection and tracking scheme are described in the following subsections, as shown in Fig. 1.

3.1 Initial object detector

This step initially determines the object regions of interest in the first frame, every fixed number of frames, as shown in Fig. 2. These regions estimated using a mixture of a grayscale

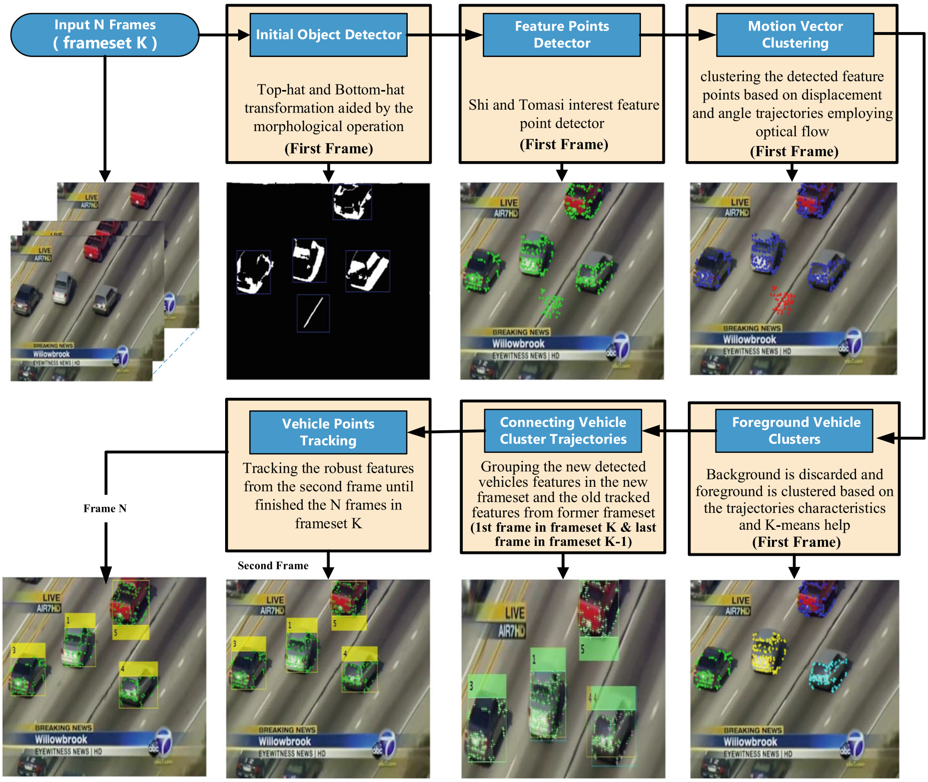

Input N Frames
(frameset K)
Initial Object Detector
Top-hat and Bottom-hat transformation aided by the morphological operation
(First Frame)
Feature Points Detector
Shi and Tomasi interest feature point detector
(First Frame)
Motion Vector Clustering
clustering the detected feature points based on displacement and angle trajectories employing optical flow
(First Frame)
Vehicle Points Tracking
Tracking the robust features from the second frame until finished the N frames in frameset K
Connecting Vehicle Cluster Trajectories
Grouping the new detected vehicles features in the new frameset and the old tracked features from former frameset
(1st frame in frameset K & last frame in frameset K-1)
Foreground Vehicle Clusters
Background is discarded and foreground is clustered based on the trajectories characteristics and K-means help
(First Frame)
Frame N
Second Frame

Fig. 1 Workflow chart of the proposed system for a specific frameset K

closing and bottom-hat transformation or grayscale opening and top-hat transformation [38], as described in equations 1, 2, 3 and 4.

$$\text{Opening} : I \circ ST = (I \ominus ST) \oplus ST \tag{1}$$

$$\text{Top-hat} : TOH = I - (I \circ ST) \tag{2}$$

$$\text{Closing} : I \bullet ST = (I \oplus ST) \ominus ST \tag{3}$$

$$\text{Bottom-hat} : BOH = (I \bullet ST) - I \tag{4}$$

Where I presents the original image, ST is the structuring element, $\circ$ represents the grayscale opening, and $\bullet$ represents the grayscale closing. $\ominus$ and $\oplus$ represent erosion and dilation operators, respectively. TOH presents the frame image after making the Top-hat transformation, and BOH represents it after making the Bottom-hat transformation.

The initial detection method includes the following major steps. First, the structure element of morphological transformation is identified based on the camera spatial resolution and the altitude from the ground. Then, the grayscale opening transformation, grayscale top-hat transformation, and Otsu partitioning method are employed to detect vehicles in a light background. Whereas the grayscale closing transformation, grayscale Bottom-hat transformation, and Otsu partitioning method, are utilized to detect vehicles in black background. In addition, vehicles have been detected from a light background and dark background. Then, the hypothesis vehicles detected from the two cases are overlaid, and the two identical hypothesis vehicles are amalgamated to a single one with the closing transformation. Bright objects are extracted from Top-hat transformation, while dark objects are extracted from Bot-hat transformation. More details can be found in [38].

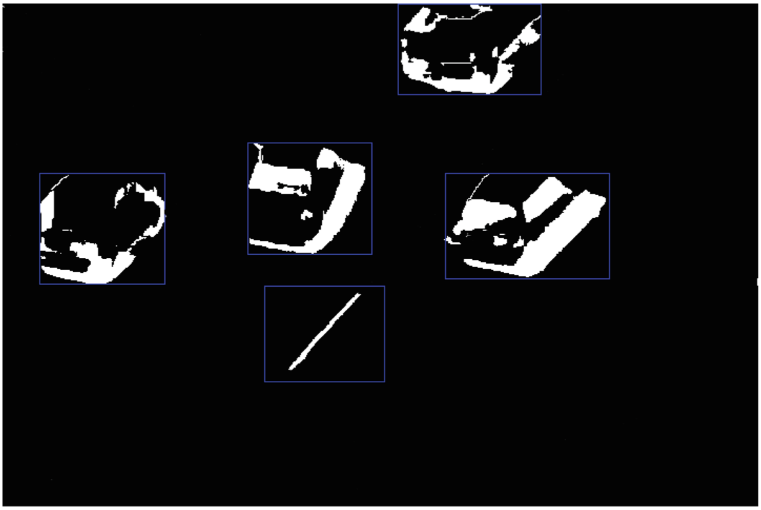


Fig. 2 Regions of detection resulted from Top-hat & Bot-hat transformation

Morphological operations are applied where the structure element is identified according to the spatial resolution of the camera sensor installed on UAV. The area covered by each image pixel is indicated through the camera spatial resolution that can be calculated by measuring the resolution cell, B (the area on the Earth's surface). The resolution cell is determined by multiply the camera height, C from the ground in its instantaneous field of view (IFOV), where IFOV was known as the camera angular cone of visibility, A, as illustrated in Fig. 3.

This detection algorithm has a low computational complexity since it can detect various vehicle types without a pre-trained process. Besides, it achieved a satisfactory recall accuracy. Nonetheless, this algorithm detects some of the false-positive objects that yield low precision accuracy. These false-positive results will be eliminated by employing K-means clustering and their tracking optical flow information, as described in the following subsections. The false-positive results obtained from the background regions have different motion characteristics compared to the foreground vehicles. Hence we exploit their feature points motion information to discard them before the final decision. Calculate the number of vehicles dynamically is another key factor in the detection phase. The detection process is implemented regularly to capture the diversity of moving objects number and to accomplish an automatic system for the detection and tracker updating.

3.2 Feature points detector

Selecting good features of the bounding boxes, that resulted from the detection step, is necessary for robustly tracking feature points across frames. While the Harris detector is the most famous corner detector, Shi and Tomasi's detector perform better than the Harris corner detector [20]. Here Tomasi detector has been implemented in each detection region based on the detection algorithm described in the previous section to extract the robust corner point inside each region as illustrated in Fig. 4.

The periodically updated vehicle feature points are very important to guarantee to track them for a long time because these features may disappear as a result of illumination change and out-of-plane rotation. So, the detection process is repeated regularly every fixed number of frames, N_{Frames}. The new extracted feature points from the detection step and the old tracked features in the latter frame of the former N_{Frames} are combined. Hence, the system is updated every N_{Frames} by integrating both the tracking and detection feature points.

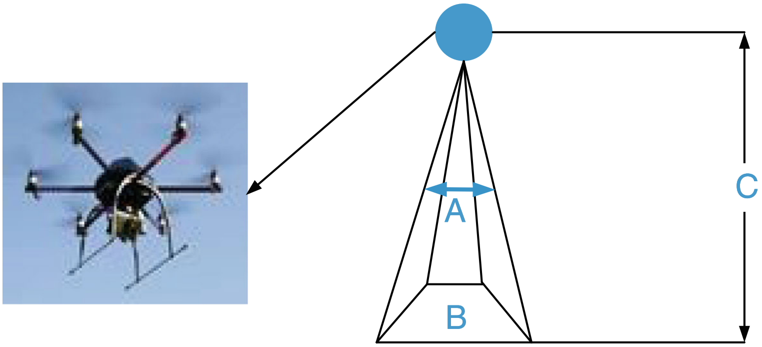


Fig. 3 Structuring element selection based on drone altitude and camera spatial resolution

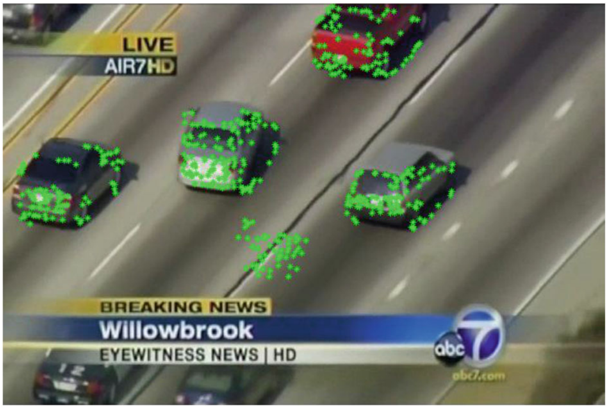


Fig. 4 Feature points extraction process using Shi-Tomasi detector in the detected regions of interest

3.3 Motion and foreground vehicles clustering

This step has two main purposes; removing the background regions to eliminate the false positive detected result, and clustering the extracted foreground vehicles. These functions are achieved by using the K-means clustering algorithm and the optical flow information. The K-means clustering algorithm and the methodology of using it with the optical flow are explained in the following subsections.

3.3.1 K-means clustering

K-means clustering is unsupervised learning, computationally efficient algorithm for large datasets. Initially, k samples, serving as the initial centroids, are chosen randomly to approximate the centroids of the initial clusters (K is a positive integer number). Simply K-means clustering is an algorithm to group the objects based on features into a K number of groups. The grouping process is done by minimizing the sum of squares of distances (Euclidean squared distance) between data and the corresponding cluster centroid. K-means algorithm will be carried out by executing the following three steps below until convergence (Iterate until stable) is obtained.

- Determine the centroid coordinates.
- Determine the distance between the centroids and each data feature.
- Group the data based on the minimum distance to find the closest centroid.

In the second and third steps, the data coordinates using the centroid and the nearest neighbor role are distributed. The data coordinates C_j are assigned to cluster $CL_K\ (m)$ if

$$\overset{|C_j-M_K(m)|<|C_j-M_c(m)|}{\text{For all } C = 1, 2, ..., Kand\ C \neq K} \tag{5}$$

Let $C = \{C_1, C_2, C_3, ..., C_N\}$ be the data sets (feature coordinates, angle or displacement) and k is the number of clusters; M_l is the centroid of cluster CL_l that defined by the following equation.

$$M_l\ (m) = \frac{1}{n_l} \sum_{C \in CL_l} C \qquad , l = 1, 2, 3, ..., K \tag{6}$$

Where n_l is the number of features in a cluster, l is the cluster number, m is the iteration number and N is the number of features. Convergence is obtained if none of the cluster centroids changed, that is

$$\text{For} \quad \overset{M_l(m+1)=M_l(m)}{l = 1, 2, 3, ..., k} \tag{7}$$

Where k is finally obtained.

3.3.2 Motion and vehicle clustering

To ensure the extraction between the vehicle clusters and the background regions, each of them has a different motion characteristic. The optical flow-based feature point tracking is considered a suitable approach to achieve our goal since it produces a perfect feature point matching accuracy and faster processing speed. After the detection step and extracting the robust points in frame $f(x, y, t)$, the detected corner point is tracked from frame t to frame $t+1$ using the Kanade-Lucas optical flow approach. The optical flow results in the first frame pairs are a set of vectors C e.g $C_i = (S_i, \theta_i)$, where S_i and θ_i are given by

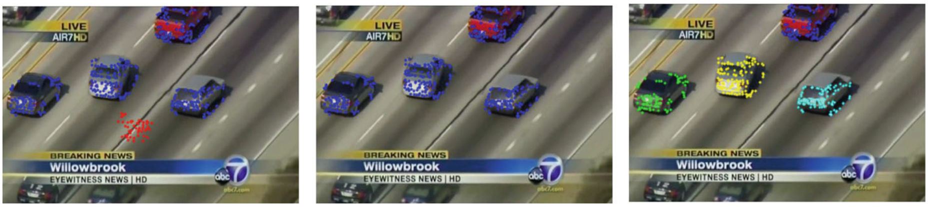
$$S_i = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \tag{8}$$

$$\theta_i = \arctan\left(\frac{Y_2 - Y_1}{X_2 - X_1}\right) \tag{9}$$

Each element in C matches a feature point P_i that tracked from frame t to frame $t+1$, where S_i and θ_i are two vectors comprising the displacement magnitudes and angles respectively for each feature point. In the initial stage before the background elimination, each feature point is a vertex $P_i = (S_i, \theta_i)$ and K-means clustering is used to cluster the feature points based only on their displacement magnitudes and angles. The k-means clustering algorithm is used to successfully separate the feature points of the vehicle regions from the background regions, as shown in Fig. 5 (a), by considering the trajectories of the points from multiple frames.

It should be noted that the trajectories behaviors of the background feature points through multiple frames have a small variation in S and θ or having too large variations have been considered [13]. On the other side, the trajectories behaviors of the vehicle feature points have a motion variation that should not be too small or too large. Hence, the background regions can be eliminated using k-means clustering approach. In this state, K is selected to be two; one is devoted to the foreground cluster and the other for the background cluster.

After removing the background cluster as shown in Fig. 5 (b), the number of detecting regions only contains the foreground vehicles which can be grouped using k-means clustering as shown in Fig. 5 (c). In this situation, the elements of the vector C_F contain S_F, θ_F, and points' coordinates where F relates to the foreground. K_F value is the number of



(a) Motion Vector Points **(b)** Foreground Vehicles **(c)** Foreground Vehicle Clusters

Fig. 5 Clustering the foreground vehicles using the K-means clustering and points information analysis

detected vehicles. Every feature point is a vertex $P_i = (X_i, Y_i, S_i, \theta_i)$, where X_i and Y_i represent the X and Y coordinates in the current frame, S_i and θ_i represent the displacement and angle of P_i from two consecutive frames, respectively, as shown in Fig. 5 (c) where only 4 vehicles were detected.

3.4 Connecting vehicle cluster trajectories

The detection vehicle regions with their most robust feature points inside each cluster will be accompanied by a bounding box that created according to feature points' coordinates in each cluster. Each one of the bounding boxes with their feature points will take a unique ID that tracked within the frameset. In order to judge the possibility to assign the same ID for the new detected vehicle or it is a new vehicle with a different ID, the intersection area of the new detected bounding box and the old tracking bounding boxes is calculated. For this purpose, the following two cases are considered.

1. Maintaining same vehicle cluster
   The intersection area, in this case, is greater than a predetermined α percentage. Hence the newly detected vehicle has the same label (ID) of the old matched one, as shown in Fig. 6 (b). Another case of maintaining the same ID, is when the detection algorithm cannot detect the tracked vehicle, so the result of the tracking will be used with the same ID for the next frameset.
2. Creating a new vehicle cluster
   This case established when there is no intersection area value or the estimated intersection area is less than or equal α, so the newly detected vehicle assigns a new label ID. This case is shown in Fig. 6 (a), when the new detected vehicle with red color in frame 121 has no intersection area with the existing tracks vehicles in frame 120, so it is assigned to a new vehicle.

3.5 Vehicle points tracking

The obtained vehicles features are tracked using KLT tracker [6]. First, the vehicle's corner points are extracted and clustered in the first frame of each frameset (N_{Frames}). Then KLT tracker is used to tracking them through the remaining frames starting from the second frame, as shown in Fig. 7 (a) to the N^{th} frame as shown in Fig. 7 (b).

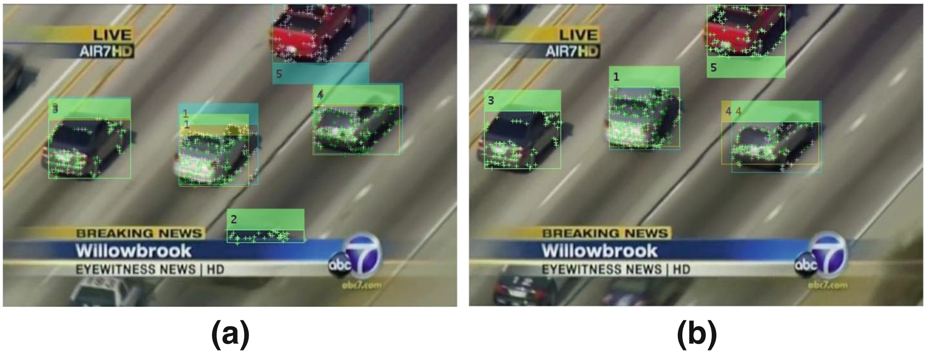


Fig. 6 Connecting vehicle trajectories based on the detected and tracked rectangular boxes

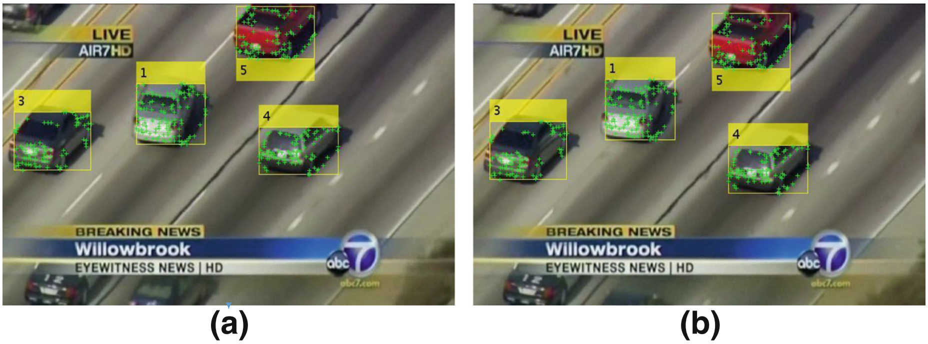


Fig. 7 Tracking Feature Points from the second frame **a** to the last frame **b** in the frame set

Vehicle bounding boxes are updated continuously in each frame using the updated points' coordinates resulted from the KLT tracker, where each vehicle trajectory is created between a pair of images by measuring the displacement vector. This vector added to the previous corresponding vector every frame to achieve a continuous trajectory for each vehicle.

To solve the problem of fixed bounding box size, we propose an adaptive bounding box using the output points from optical flow, every frame after getting the strongest points as in Fig. 8 (a), we can easily form the correct bounding box in relative with these points as in Fig. 8 (b). This process executed based on the maximum and minimum feature point coordinates X_{min}, Y_{min}, X_{max}, and Y_{max} for each vehicle, as shown in Fig. 9

$$Bounding\ Box = [X_{min} \quad Y_{min} \quad (X_{max} - X_{min}) \quad (Y_{max} - Y_{min})] \tag{10}$$

4 Experimental results

The proposed method is evaluated, compared with the existing methods [1, 3, 7, 8, 16, 17, 27, 32] and the experimental results are discussed in this section. For the sake of comparison with the recent detection and tracking strategies, we assume focusing on data sets containing vehicle objects only because of using Top-hat & Bottom-hat transformation and optical flow information cannot discriminate between the vehicles and other objects. DARPA VIVID [9]

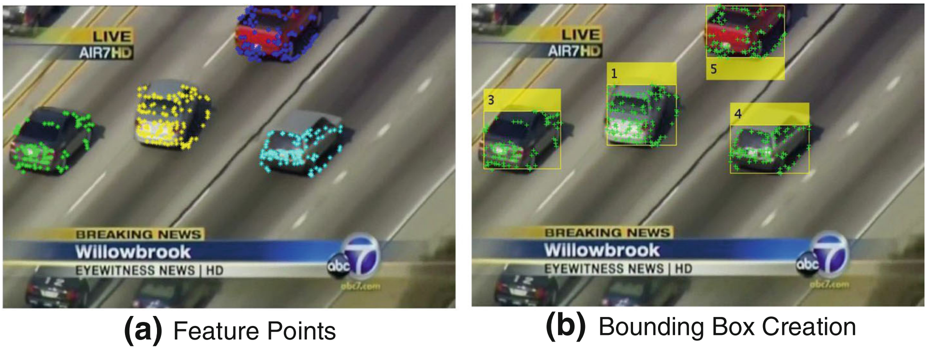


Fig. 8 Variable Bounding Box creation, the vehicles rectangular boxes that shown in figure (b) are created based on vehicles feature points coordinates that shown in figure (a)

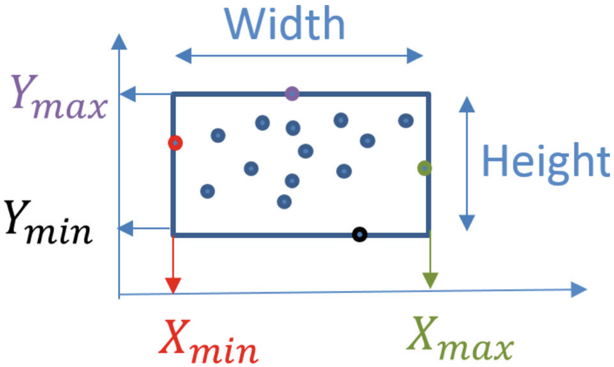


Fig. 9 Varying bounding box coordinates, each bounding box is created based on the minimum and maximum feature point coordinates

and police chase datasets are used to validate the proposed method. VIVID is widely used as a benchmark for evaluating moving object detection and tracking algorithms. It contains various challenges as the varying number of vehicles, occlusion, poses variation, blocked vehicles, and out of the plane rotation. Moreover, a real-world airborne video representing a police chase that contains 20 vehicles moving with high speed without following any pattern is used. It should be noted that the ground truth information is not available for vehicular targets in these data sets. Hence, the tracking accuracy was evaluated by manually labeling the results as shown in [3].

Here, the detection process was made every ten frames, $N_{Frames} = 10$, to reduce the computational complexity. As described in Fig. 3, the structure element was chosen to be a disc with radius 10 and 20 pixels for VIVID and police chase videos, respectively, based on the camera spatial resolution, vehicle size, and the altitude from the ground. The newly detected vehicle takes the same ID of the corresponding vehicle based on the history of trajectories when the overlapped intersection area between them is greater than 25 percentage, $\alpha = 25\%$. We examined the algorithm with different values of α, where we found that if α is too high, the same vehicle may be classified into the new one, which increases the Identity Switch error and consequently minimizes the tracking accuracy. The value of $\alpha = 25$ $\%$ yields to the best accuracy in our experiment. The proposed method offers a robust automatic multi-vehicles detection and tracking system. Quantitative evaluation of the detection will be discussed in Section 4.1, and compared with recent approaches [1, 3, 17, 27, 32]. In Section 4.2, a comparison with tracking approaches [1, 3, 7, 8, 16] was presented to examine the tracking performance of the proposed methodology.

4.1 Detection evaluation

The comparison with recent techniques was conducted to confirm the proposed method contribution, as illustrated in Tables 1 and 3. The detection accuracy is evaluated using quantitative performance metrics that have been used as a standard evaluation [19], known as Precision and Recall. The precision is calculated as the percentage of correctly detection vehicle pixels [true positive (TP)] over the total number of detecting object pixels, including TPs and false positive (FP).

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{11}$$

Table 1 Detection accuracy comparison of different methods on VIVID dataset

Dataset Videos	Accuracy %	Image Stacking Method [32]	TMRAG [17]	Rosenbaum et al.[27]	Proposed Method
(Egtest01)	Precission	99	94	86	99.4
	Recall	97.2	89	85	98.3
(Egtest01)	Precission	99	94	86	99.4
	Recall	97.2	89	85	98.3
(Egtest02)	Precission	N/A	94	86	97.99
	Recall	N/A	89	85	94
(Egtest03)	Precission	N/A	94	86	95
	Recall	N/A	89	85	94
(Egtest04)	Precission	N/A	94	86	98
	Recall	N/A	89	85	95
(Egtest05)	Precission	N/A	94	86	97
	Recall	N/A	89	85	94

Recall refers to the ratio of accurately detected vehicle pixels to the number of actual vehicle pixels that include the number of false negative pixels (FN).

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{12}$$

The precision and recall results for approaches focused on the detection or simultaneous detection and tracking are reported in Tables 1 and 3, respectively. The detection algorithm using a stacking technique [32] exceeds the accuracy of the state of art algorithms. However, this method has a declined precision accuracy as a result of the false positive detections that occur from the registration process. They used image stacking in a novel way by employing the image registration to small vehicles only, and the warping process has been used to blur all the stationary background in the neighborhood of the moving vehicles. The main goal of this algorithm is to remove disturbing image parts of the background, which can be smoothed to extract the observed vehicle from the background, such a strategy needs a huge computational requirement.

In [27], an AdaBoost classifier was trained on Haar-like features to classify pixels, and a support vector machine (SVM) was used for vehicle detection. However, this strategy fails to efficiently distinguish moving vehicles from the background. It is also can not detect the whole shape of the moving vehicles because the block regions may not be consistent in visual features and motion transformations. Image registration has been used as a basis for moving objects detection. Registration strategies attempt to discover correspondences between consecutive frame pairs based on image appearances transformations. However, spatial information is often ignored, and different motions from multiple moving objects cannot be efficiently modeled. In addition, image registration is not efficient to handle occlusion that can result in potential object misses. Authors in [17], presented a new approach to address these problems. Video frames are first divided into uniform regions, followed by the construction of region adjacent graphs to represent each frame. The corresponding regions are then matched with the consecutive frames by employing the multigraph matching algorithm. A graph coloring algorithm finally labels objects as being background or foreground region. However, this algorithm has a high computation complexity to choose the optimal frame sequence length for the best multi-graph matching performance. Although the results of [3] showed a good precision accuracy, the recall dramatically drops using the Canny edge detector that tends to fail in low contrast scenes. In addition, the major weakness of this algorithm is the incapability of occlusion handling because the distinctive objects were characterized using the Canny edge detector for blobs detection, which in turn generates multiple-scale target hypotheses with temporal consistencies and adaptive thresholding. It is asserted from Table 3 that the precision weakness for testing Egtest3 using the method in [3] results since this scene has more occlusion cases. The proposed algorithm achieves the best precision and recall, with average percentages of 97.5% and 95.1%, respectively with the higher precision and recall percentages. This result based on the false positive detections elimination process that yields from collecting information from the detection step and feature points motion analysis.

4.2 Tracking Evaluation

To evaluate the proposed method regarding the tracking perspective, we compare the vehicle tracking results of the proposed method with recent tracking techniques. Two tracking performance measures are used [19]; known as multiple object tracking accuracy (MOTA) and tracking speed, frame per second (FPS). MOTA which is the most widely used figures to evaluate a tracker's performance is defined as

$$MOTA = 1 - \frac{\sum_{f=1}^{ft} \left(FN_f + FP_f + IDSW_f\right)}{\sum_{f=1}^{ft} \left(GT_f\right)} \tag{13}$$

Where f is the frame number and ft are the total numbers of frames. It is expressed as it integrates three sources of errors, false negative, false positives, and identity switches. Frame-per-second (FPS) metric displays the processing speed of algorithms by calculating the computational time of the approach.

As shown in Tables 2 and 3, the proposed method achieved the highest accuracy. Different from the comparison algorithms, where vehicles in the first initial frames are tracked manually to initialize the tracking process [7], or where it needs to initialize the vehicle locations manually [8] and [16], the proposed approach automatically detects and track the vehicles. Moreover, the proposed algorithm achieves the minimum execution time. It is worth mentioning that, different from existing methods, the resulting execution time comprises both detections, and tracking process time. Moreover, our method was implemented using MATLAB and carried out on a single core of an Intel i7, 3.4 GHz Processor with 4 GB RAM. It is asserted that the average rate of 34.64 fps for the VIVID dataset achievable from the proposed approach is highly encouraging compared to 26.9 fps from the method presented in [3]. However, the result of [1] obtains the best computation processing speed compared with the state of arts, while the average detection and tracking results dramatically drop because the algorithm only depends on the feature points and histogram of pixels.

Detection and tracking results in the police chase video are shown in Table 4. It is asserted that the proposed method achieved the highest precision and recall accuracy of 98%, 91% respectively compared to 91%, and 78.7% respectively in [1]. Moreover, the proposed method increases the tracking accuracy by 12.2% from 79% to 91.2% compared to the method in [1].

Figure 10 shows some important visual evaluation of the experimental results. The first row illustrates the success in handling various motion patterns without existing known when

Table 2 Tracking (MOTA %) and time performance (FPS) comparison for the VIVID dataset

Tracking Method	Videos	Dynamic Group [8]	Jiang and Cao [16]	Cao et al [7]	Proposed Method
Tracking Performance MOTA %	(Egtest01)	68.3	79.2	92.8	99.94
	(Egtest02)	94.3	68.1	88.3	99.86
	(Egtest03)	N/A	75	N/A	91
	(Egtest04)	N/A	N/A	N/A	93
	(Egtest05)	N/A	N/A	N/A	92
Time Performance FPS	(Egtest01)	21.7	9.23	38.5	40.69
	(Egtest02)	33.2	9.19	27.4	39.54
	(Egtest03)	N/A	9.1	N/A	29
	(Egtest04)	N/A	N/A	N/A	31
	(Egtest05)	N/A	N/A	N/A	33

the tracking is persistent without a change in the vehicle's label. On the other side, the vehicle tracking accuracy will be deteriorated when a restrictive probabilistic motion model was employed, since some of the objects could not be tracked due to the various turning angles. The occlusion problem is solved, as illustrated in the second and third rows of Fig. 10 by utilized the combined detection and tracking results with minimum Euclidean distance selection between the centroid of their intersection area. The fourth row demonstrates the behavior of the proposed algorithm in dealing with vehicle detection failure that resulted from camera defocusing. Tracks are continued without fail, despite the missing observation. However, one should note that an identity switch error is generated due to camera defocussing since some feature points will be lost. The success of the proposed method in

Table 3 Simultaneous Detection and Tracking Accuracy Comparison of Different Methods on VIVID Dataset

Compared Algorithm	Evaluation Measure	(Egtest1)	(Egtest2)	(Egtest3)	(Egtest4)	(Egtest5)
Alkanat et al.[3]	Precision %	90	90	82	95	N/A
	Recall %	82	77.2	74	88	N/A
	Tracking performance MOTA %	99.1	96.2	93.2	92.3	N/A
	Time Performance FPS	25.33	26.9	26	25	N/A
Mohamed et. Al [1]	Precision %	N/A	N/A	N/A	N/A	N/A
	Recall %	89.65	61.37	N/A	N/A	N/A
	Tracking performance MOTA %	94.5	N/A	N/A	N/A	N/A
	Time Performance FPS	56.6	N/A	N/A	N/A	N/A
Proposed Method	Precision %	99.4	97.99	95	98	97
	Recall %	98.3	94	94	95	94
	Tracking performance MOTA %	99.94	99.86	91	93	92
	Time Performance FPS	40.69	39.54	29	31	33

Table 4 Simultaneous Detection and Tracking Accuracy Comparison of Different Methods on Police Chase Video

Detection and	Detection Performance %		Tracking Performance	Time Performance
Tracking Method	Precission	Recall	MOTA %	FPS
Mohamed [1]	91	78.7	79	N/A
Proposed Method	98	91	91.2	13.67

dealing with vehicles blocked by trees can be observed in the fifth row. The repeated detection process saved the blocked vehicle from being lost. The last row indicates the success in tracking high-speed vehicles without missing the trajectory. Furthermore, these sub-figures

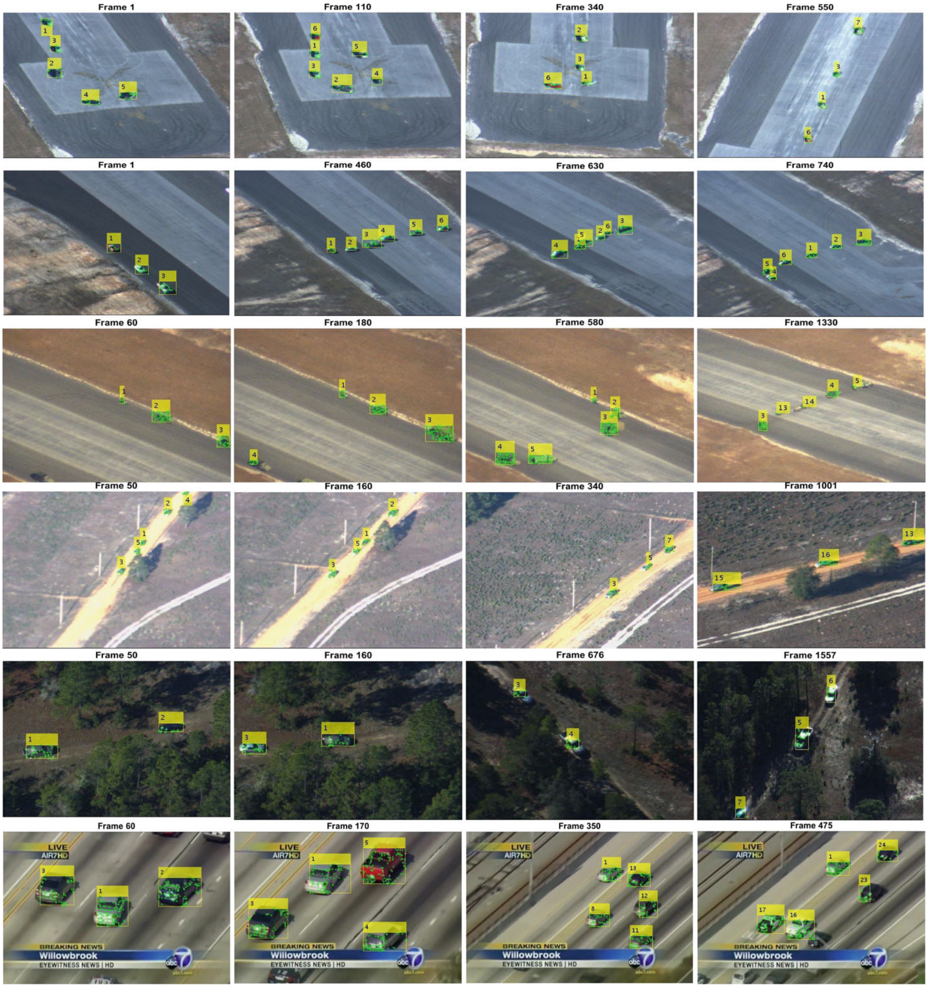


Fig. 10 Sample result on VIVID data set in the first five rows and police chase video in the last row

show the success in handling scale changes by using the proper structure elements in the detection phase.

5 Conclusion

This paper proposes a real-time and efficient approach for vehicle detection and tracking in aerial videos. In this method, both the detection and tracking procedures were linked together to get the robust feature points when they are updated regularly every fixed number of frames. These feature points are extracted and analyzed using a combined process between grayscale morphological-based detection, optical flow, and K-means clustering. Further, the feature points are clustered based on their motion properties to distinguish the foreground features from the background and to assign the detecting vehicles with its corresponding cluster based on the intersection area between the detected and tracked points' information. The proposed algorithm has been tested on videos taken in different scenarios to verify its efficiency over other algorithms. Excellent detection and tracking accuracy, and fast processing speed are the main factors, that nominate the proposed algorithm for real-time applications.

Acknowledgments This work was supported in part by the Egyptian Ministry of Higher Education (MoHE), Cairo, Egypt, and in part by the Egypt Japan University of Science and Technology (E-JUST), Alexandria, Egypt.

References

1. Abdelwahab MA, Abdelwahab MM (2015) A novel algorithm for vehicle detection and tracking in airborne videos. In: Multimedia (ISM), 2015 IEEE International Symposium on, Miami, FL, USA, pp 65–68
2. Al-Kaff A, de la Escalera A, Armingol JM (2015) Sift and surf performance evaluation and the effect of freak descriptor in the context of visual odometry for unmanned aerial vehicles. In: International Conference on Computer Aided Systems Theory, Las Palmas de Gran Canaria, Spain, pp 739–747
3. Alkanat T, Tunali E, Öz S (2015) Fully-automatic target detection and tracking for real-time, airborne imaging applications. In: International Joint Conference on Computer Vision, Imaging and Computer Graphics, Berlin, Germany, pp 240–255
4. Bar-Shalom Y, Fortmann T, Scheffe M, et al. (1980) Joint probabilistic data association for multiple targets in clutter. In: Prof. conf. on information sciences and systems, Albuquerque, pp 404–409
5. Barth A, Franke U (2009) Estimating the driving state of oncoming vehicles from a moving platform using stereo vision. IEEE Trans Intell Transp Syst 10(4):560–571
6. Bouguet J-Y (2001) Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. Intel Corporation 5(1-10):4
7. Cao X, Jiang X, Li X, Yan P (2018) Correlation-based tracking of multiple targets with hierarchical layered structure. IEEE transactions on cybernetics 48(1):90–102
8. Cao X, Shi Z, Yan P, Li X (2013) Tracking vehicles as groups in airborne videos. Neurocomputing 99:38–45
9. Collins R, Zhou X, Teh SK (2005) An open source tracking testbed and evaluation web site. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2005), January 2005, Breckenridge, Colorado
10. Gao Z, Gao L-S, Zhang H, Cheng Z, Hong R (2019) Deep spatial pyramid features collaborative reconstruction for partial person reid. In: Proceedings of the 27th ACM International Conference on Multimedia, pp 1879–1887

11. Gao T, Li G, Lian S, Zhang J (2012) Tracking video objects with feature points based particle filtering. Multimedia Tools and Applications 58(1):1–21
12. Gao Z, Xu C, Zhang H, Li S, de Albuquerque VHC (2020) Trustful internet of surveillance things based on deeply-represented visual co-saliency detection. IEEE Internet Things J 7(5):4092–4100
13. Gomaa A, Abdelwahab MM, Abo-Zahhad M (2018) Real-time algorithm for simultaneous vehicle detection and tracking in aerial view videos. In: 2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS), IEEE, pp 222–225
14. Gomaa A, Abdelwahab MM, Abo-Zahhad M, Minematsu T, Taniguchi R (2019) Robust vehicle detection and counting algorithm employing a convolution neural network and optical flow. Sensors 19(20):4588
15. Hess R, Fern A (2009) Discriminatively trained particle filters for complex multi-object tracking. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, Miami, FL, USA, pp 240–247
16. Jiang X, Cao X (2016) Surveillance from above: A detection-and-prediction based multiple target tracking method on aerial videos. In: Integrated Communications Navigation and Surveillance (ICNS), 2016, IEEE, pp 4D2–1
17. Kalantar B, Mansor SB, Halin AA, Shafri HZM, Zand M (2017) Multiple moving object detection from uav videos using trajectories of matched regional adjacency graphs. IEEE Trans Geosci Remote Sens 55(9):5198–5213
18. Kanistras K, Martins G, Rutherford MJ, Valavanis KP (2015) Survey of unmanned aerial vehicles (uavs) for traffic monitoring. In: Handbook of unmanned aerial vehicles. Springer, pp 2643–2666
19. Kasturi R, Goldgof D, Soundararajan P, Manohar V, Garofolo J, Bowers R, Boonstra M, Korzhova V, Zhang J (2008) Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2):319–336
20. Ke R, Li Z, Kim S, Ash J, Cui Z, Wang Y (2017) Real-time bidirectional traffic flow parameter estimation from aerial videos. IEEE Trans Intell Transp Syst 18(4):890–901
21. Kent P, Maskell S, Payne O, Richardson S, Scarff L (2012) Robust background subtraction for automated detection and tracking of targets in wide area motion imagery. In: Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII, vol 8546, International Society for Optics and Photonics, p 85460Q
22. Nguyen VD, Tran DT, Byun JY, Jeon JW (2018) Real-time vehicle detection using an effective region proposal-based depth and 3-channel pattern. IEEE Trans Intell Transp Syst 20(10):3634–3646
23. Niknejad HT, Takeuchi A, Mita S, McAllester D (2012) On-road multivehicle tracking using deformable object model and particle filter with improved likelihood estimation. IEEE Trans Intell Transp Syst 13(2):748–758
24. Noh S, Shim D, Jeon M (2016) Adaptive sliding-window strategy for vehicle detection in highway environments. IEEE Trans Intell Transp Syst 17(2):323–335
25. Rad R, Jamzad M (2005) Real time classification and tracking of multiple vehicles in highways. Pattern Recogn Lett 26(10):1597–1607
26. Reid D et al (1979) An algorithm for tracking multiple targets. IEEE transactions on Automatic Control 24(6):843–854
27. Rosenbaum D, Leitloff J, Kurz F, Meynberg O, Reize T (2010) Real-time image processing for road traffic data extraction from aerial images. Proc. of ISPRS TC VII Symposium IAPRS XXXVIII:469–474
28. Su A, Sun X, Liu H, Zhang X, Yu Q (2015) Online cascaded boosting with histogram of orient gradient features for car detection from unmanned aerial vehicle images. J Appl Remote Sens 9(1):096063
29. Tang Y, Zhang C, Gu R, Li P, Yang B (2017) Vehicle detection and recognition for intelligent traffic surveillance system. Multimedia tools and applications 76(4):5817–5832
30. Teutsch M, Kruger W (2015) Robust and fast detection of moving vehicles in aerial videos using sliding windows. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, pp 26–34
31. Teutsch M, Krüger W (2012) Spatio-temporal fusion of object segmentation approaches for moving distant targets. In: Information Fusion (FUSION), 2012 15th International Conference on, IEEE, pp 1988–1995
32. Teutsch M, Krüger W, Beyerer J (2017) Moving object detection in top-view aerial videos improved by image stacking. Opt Eng 56(8):083102
33. Wan EA, Van Der Merwe R (2001) The unscented Kalman filter. In: Haykin S. (ed) Kalman filtering and neural networks. Wiley, New York. ch. 7

34. Yang S, Xu J, Chen Y, Wang M (2014) On-road vehicle tracking using keypoint-based representation and online co-training. Multimedia tools and applications 72(2):1561–1583
35. Yao Y, Li J, Wu T, Zhang L (2017) Retracted article: Moving object surveillance using object proposals and background prior prediction. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-017-4820-9
36. Yu X, Shi Z (2015) Vehicle detection in remote sensing imagery based on salient information and local shape feature. Optik-International Journal for Light and Electron Optics 126(20):2485–2490
37. Zhao Z-Q, Zheng P, Xu S, Wu X (2019) Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems 30(11):3212–3232
38. Zheng Z, Zhou G, Wang Y, Liu Y, Li X, Wang X, Jiang L (2013) A novel vehicle detection method with high resolution highway aerial image. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6(6):2338–2343

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ahmed Gomaa received the B.Sc. degree (Hons.) in electronics and communications engineering, in 2009, the M.Sc. degree from Cairo University in Electronics and Communications Engineering, Cairo, Egypt, in 2015, and the Ph.D. degree from the Electronics and Communications Engineering Department with the School of Electronics, Communications, and Computer Engineering, Egypt-Japan University of Science and Technology, in 2020. From 2009 to 2015, he was a Teaching Assistant at Electronics and Communication Department, IAET. In 2015, he joined GPS laboratory, National Research Institute of Astronomy and Geophysics (NRIAG), Egypt. In February 2019, he has joined the Laboratory for Image and Media Understanding (LIMU), Kyushu University, Japan, as a special Research Student. He is currently an Assistant Professor with the National Research Institute of Astronomy and Geophysics (NRIAG). His current research interests include Image & Video Processing, Computer Vision, Remote Sensing, Wireless Sensor Network, and object detection & Tracking.

Moataz M. Abdelwahab received the B.S. and M.S. degrees in electrical engineering from the University of Alexandria, Egypt, in 1999 and 2001, respectively, and the Ph.D. degree from the University of Central Florida, USA, in 2007. He is currently an Assistant professor in School of Electronics, Communication and Computer Engineering, Egypt- Japan University of Science and Technology (E-JUST). He is the author of more than 40 articles, one US patent. His research interests include image processing, computer vision, biomedical signal processing: face recognition, vehicle detection and tracking, and human activity recognition. Dr. Moataz was a recipient of the International Association of Pattern Recognition MVA conference Best paper award 2012 and IEEE MWSCAS 2006.

Mohammed Abo-Zahhad (SM'00) received the B.S. and M.S. degrees in electrical engineering from the University of Assiut, Egypt, in 1979 and 1983, respectively, and the Ph.D. degree from the University of Kent, Canterbury, U.K., and Assiut University (channel system), in 1988. He was a member of the European Society of Circuit Theory and Applications in 1998, a member of the National Communication and Electronics Promotion Committee, and a Reviewer of the National Quality Assurance and Accreditation Authority,NAQQA, Egypt, since 2011. Since 1999, he has been a Professor of electronics and communication engineering with Assiut University. He is currently the Dean of the School of Electronics, Communication and Computer Engineering and has been a Professor of communication and electronics engineering with the Egypt- Japan University of Science and Technology (E-JUST) since 2017. He is also the General Director of the E-JUST Information and Communication Technology Centre. His research interests include biomedical and genomic signal processing, speech, image and video processing, optical and digital filters, switched-capacitor, data compression, wavelet-transforms, genetic algorithms, immune algorithms, wireless sensor networks, microwave, millimeter wave wireless communications, energy harvesting, and electronic circuits. He was a recipient of the Encouragement State Award in Engineering, from the Egyptian Research and Technology Academy, Ministry of Higher Education, Egypt, in 2005.

Affiliations

Ahmed Gomaa[1,2,3] · **Moataz M. Abdelwahab**[2] · **Mohammed Abo-Zahhad**[2,4]

Moataz M. Abdelwahab
moataz.abdelwahab@ejust.edu.eg

Mohammed Abo-Zahhad
mohammed.zahhad@ejust.edu.eg

[1] National Research Institute of Astronomy and Geophysics (NRIAG), Helwan 11731, Egypt

[2] Electronics and Communications Engineering Department, Egypt-Japan University of Science and Technology, Alexandria 21934, Egypt

[3] Center for Japan-Egypt Cooperation in Science and Technology, Kyushu University, Nishiku, Fukuoka 819-0395, Japan

[4] Electrical and Electronics Engineering Department, Faculty of Engineering, Assiut University, Assiut 71511, Egypt