Needles in a Haystack: Tracking City-Scale Moving Vehicles From Continuously Moving Satellite

Wei Ao, *Student Member, IEEE*, Yanwei Fu, Xiyue Hou, *Student Member, IEEE*, and Feng Xu, *Senior Member, IEEE*

***Abstract*— In recent years, the satellite videos have been captured by moving satellite platforms. In contrast to consumers, movies, and common surveillance videos, satellite videos can record the snapshots of city-scale scenes. In a broad field-of-view of satellite videos, each moving target would be very tiny and usually composed of several pixels in frames. Even worse, the noise signals also exist in the video frames, and the background of the video frames subpixel-level and uneven moving thanks to the motion of satellites. We argue that it is a novel type of computer vision task since previous technologies are unable to detect such tiny moving vehicles efficiently. This paper proposes a novel framework that can identify small moving vehicles in satellite videos. In particular, we offer a novel detecting algorithm based on the local noise modeling. We differentiate the potential vehicle targets from noise patterns by an exponential probability distribution. Subsequently, a multi-morphological-cue based discrimination strategy is designed to distinguish correct vehicle targets from the existing noises further. Another significant contribution is to introduce a series of evaluation protocols to measure the performance of tiny moving vehicle detection systematically. We annotate satellite videos manually to test our algorithms under different evaluation criterions. The proposed algorithm is also compared with the state-of-the-art baselines, which demonstrates the advantages of our framework over the benchmarks. Besides, the dataset would be downloaded from http://first.authour.github.com.**

***Index Terms*— Tiny object detection, probabilistic noise modeling, evaluation, vehicle detection.**

I. Introduction

With the recent advance in the earth observation (EO) technology, satellite videos could capture consecutive images from moving satellite platforms by utilizing the optical sensors to capture consecutive images from a moving satellite platform. Among the fruitful mechanism, the recent advanced compressive sensing techniques, just like Kronecker compressive sensing [1], make the satellite videos enable many potential applications, such as city-scale traffic surveillance, 3D reconstruction of urban buildings and quake-relief efforts, etc. For instance, Figure 1(a) shows a frame of a satellite video of Valencia city in Spain. As visualized in Figure 1(b), a corresponding areal map of each video frame is about 3×4 square kilometers. Satellite video can thus facilitate monitoring the dynamics city-scale scenes. In addition, to efficiently supervise the city-scale scene, one primary, and yet the critical task is to detect and track moving vehicles captured in the satellite videos. However, there is no previous technique for the detection of tiny moving vehicles in the satellite videos, due to the following challenges.

In satellite videos, vehicles are moving and very tiny. In satellite videos.[1] only several pixels represent each vehicle. Thus essentially, we have to detect tiny moving vehicles in satellite videos. Figure 3 shows two enlarged regions of the panorama in Figure 1(a). From these enlargements, a vehicle is only composed of several pixels without any distinctive color or texture. In such situation, the motion of these moving vehicles is the most robust feature. However, it is easily obscured and challenged by the background moving.

The frames of satellite videos cover a large-scale region and provide a dynamic scenario. In terms of the distances between a camera shot and observed objects, it is devided into near-field, medium-field, far-field surveillance videos, and extremely far-field satellite videos [2], [3]. As extremely far-field video, satellite video processing is more difficult than the "far-field" video. Not only a broad field-of-view does a satellite video provide but it also presents a very complex background. As shown in Figure 2, the visual content of satellite videos may include roads, buildings, vegetation and football field, etc. Furthermore, it also has varied traffic conditions as many as possible, e.g. straight arteries, intersections, and roundabouts, etc.

The background of satellite videos presents sub-pixel-level and uneven moving. The optical flow field [4], [5] of the above satellite videos is shown in Figure 4. It shows that the background is continuously moving, and the optical flow field is very uneven. Even worse, the relative motion of the satellite video is very complicated since intrinsically the satellite video frames are the 2D projection of a sophisticated 3D movement of the satellite platform. Moreover, since the satellites are very far away from the earth plane, we can only observe extremely

Manuscript received March 6, 2019; revised August 24, 2019; accepted September 19, 2019. Date of publication October 7, 2019; date of current version November 27, 2019. This work was supported in part by the Natural Science Foundation of China under Grant 61822107 and Grant 61571134. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel L. Lau. *(Corresponding author: Feng Xu.)*

W. Ao, X. Hou, and F. Xu are with the Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China (e-mail: wao16@fudan.edu.cn; 18210720056@fudan.edu.cn; fengxu@fudan.edu.cn).

Y. Fu is with the MoE Frontiers Center for Brain Science, Fudan University, Shanghai 200433, China, and also with the School of Data Science, Fudan University, Shanghai 200433, China (e-mail: yanweifu@fudan.edu.cn).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Digital Object Identifier 10.1109/TIP.2019.2944097

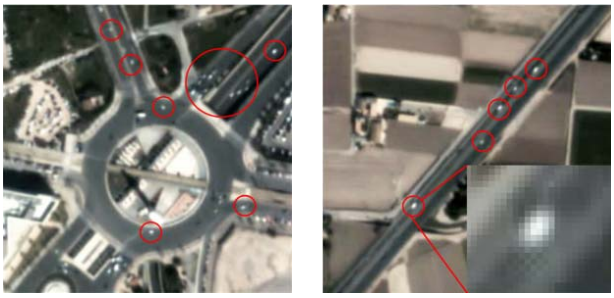
[1] Examples and results of our satellite videos are on https://drive.google.com/drive/folders/1eRUfv_l2mHzlyj6VJr3zpTn5a4NIoMF0?usp=sharing

(a) (b)

Fig. 1. An example of satellite videos. (a) A frame of a satellite video of Valencia, Spain. (b) The corresponding optical map downloaded from Google Earth.

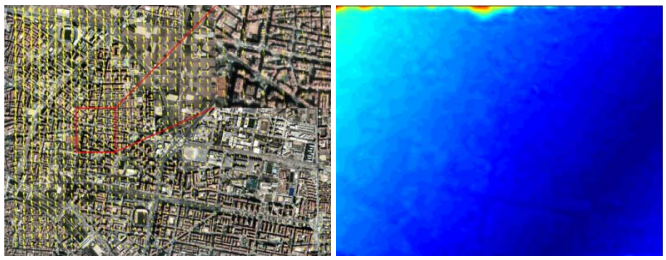
(a) (b) (c) (d)

Fig. 2. Varied parts of a frame in the satellite video. (a) Roads, (b) buildings, (c) vegetation, and (d) a football field.



(a) (b)

Fig. 3. The enlargements of two scenes in Figure 1 (a), where some vehicle targets are denoted with red circles. (a) intersection, (b) road.



(a) (b)

Fig. 4. The optical flow field which is obtained from frame 1 and 100 of the satellite video. (a) The magnitude and orientation of the optical flow field through vectors, (b) further illustrates the magnitude distribution of the optical flow field.

slow moving among the consecutive video frames. Such slow-moving will lead to small variants of the stationary pixels. It brings the benefits of avoiding the frame-by-frame video stabilization and the general registration. Critically, we notice that the moving of two successive satellite video frames is always sub-pixel-level. Overall, the key difficulty in detecting tiny moving vehicles is to differentiate the background motions from moving vehicles; otherwise, the moving background will negatively affect the detection of tiny moving vehicles.

The patterns of moving vehicles may also be confused with the noise patterns which are caused by the complex moving backgrounds. This paper focuses on detecting and tracking tiny vehicle in the satellite videos which are very hard to be identified and easily affected by noise. Such noise patterns may result in the regular moving of stationary corners or edges, and thus further hinder the detection of tiny moving vehicles.

To tackle the problems as mentioned above, we, for the first time, propose a framework of addressing the challenging tasks in the field of detecting tiny moving vehicles. The whole framework is built upon a series of statistical tools. In particular, we propose a motion-based detecting algorithm using novel local modeling. We decompose each frame into two parts, i.e., an original image and an additive random 2D noise signal map. A probability distribution is used to fit the noise patterns, which facilitates us to distinguish potential vehicle targets. A local tactic is applied to address intra-variants within a frame and discern inter-variants between frames, simultaneously. Then, a region growing method and a discrimination algorithm based on multiple-morphological-cues are proposed,

which can remove other noises. Kalman filter (KF) [6] is further used to track vehicles. Extensive experiments are conducted on the real-world satellite video dataset to evaluate the proposed models over the baselines.

The major contributions of this paper are fourfold: (1) To the best of our knowledge, the tasks of detecting tiny moving objects are, for the first time, studied. To further study this task, we contribute the satellite video dataset, which has the labeled ground-truth of tiny moving objects. (2) We propose a motion based detecting algorithm using novel local modeling. The noise pattern is modeled by probabilistic distributions. (3) A region growing algorithm is further designed and a discrimination algorithm based on multiple morphological cues is proposed. (4) We, for the first time, propose a set of evaluation protocols which can systematically measure the algorithms of analyzing tiny moving vehicles.

The remainder of the paper is organized as follows. First, Sec. II reviews some related work. Then, Sec. III details the proposed algorithms, including the overall framework and two major contributions, detecting and discrimination algorithms. Subsequently, the used evaluation metrics and the proposed evaluation algorithm are presented in Sec. IV. Sec. V shows experiments undertaken on the satellite video. Finally, Sec. VI concludes the paper.

II. Related Work

A. Earth Observation Technology and Satellite Videos

Nowadays many observation technologies have been developed and are of enormous significance, including optical satellite images, space-borne synthetic aperture radar (SAR) images and aerial videos. Such technology plays a critical role in both civil and military area, such as the city traffic system, maritime surveillance, aerial spy and battlefield monitor, etc. Both optical satellite imagery and space-borne SAR can observe a large region in a high resolution. However, an optical satellite is very susceptible to different illumination and various weather. Although SAR has the unique capability of earth imaging in all-whether regardless of day and night [7], the SAR images are difficult to be interpreted [8], [9]. Another weakness is that they cannot observe dynamics due to stationary imaging, which narrows down their applications.

The satellite videos have many advantages over the other conventional videos, such as aerial videos captured by the unmanned aerial vehicle (UAV). The aerial videos often suffer from undesirable dramatic motions of platforms and have to resort to complex stabilization preprocessing [10]. Thus to track objects, image registration has to be done at the first stage, which can then separate camera egomotion from object motion [11]–[13]. The aerial videos can only cover a small city scope, while our satellite videos can easily supervise a city-scale scene. Furthermore, some new legislation of the civil aviation safety has forbidden or restricted the usage of UAV.

Recently, some commercial companies are able to have satellites. For example, Satellite Imaging Corporation (SIC) successfully launched video satellites âŁ" SkySat-1 and SkySat-2, on November 21, 2013 and July 8, 2014, respectively. Chang Guang Satellite Technology CO., LTD (CGSTL) successfully launched two video satellites on October 7, 2015. Up to now, CGSTL has 8 on-orbit video satellites which are a part of the ongoing Jilin No. 1 satellite constellation. The time resolution of EO of Jilin No. 1 satellite constellation will be shorted to half an hour when the constellation is constructed by 2020.

Comparing with the conventional EO technologies, satellite videos can cover the largest scopes and is very stable. First, we can understand and forecast the dynamics of the earth. Second, a video satellite can turn its lens towards the region of interest (ROI) all the time through flying so long as this region is within the field-if-view of the satellite, which has very good image quality. So satellite videos are more stable compared with other videos. Third, a high altitude of a video satellite results in a broader field-of-view, which even covers a city-scale region. Another important issue should be taken into account is that a video satellite is a free platform which can record anywhere in the earth without any restriction.

B. Moving Target Detection and Tracking

Moving target detection can be taken as a special case of foreground segmentation. Such tasks can be solved by the Gaussian Mixture Model (GMM) [14]–[16], the state-of-the-art ViBe [17], and even deep learning method as Hierarchical Convolutional Features (HCF) [18], [19]. GMM is a representative of parametric models, the ViBe tries a non-parametric method to describe the dynamic patterns, while HCF extracts different features from each layer of deep learning for tracking. GMM utilizes a weighted mixture of Gaussian distributions to model the pixel value varying over time. However, the dynamics of pixel values may be not subjected to Gaussian distributions. In most cases, we cannot use a definite parametric model to represent the variance of pixel values. ViBe proposed a novel idea that some pixel values in different time steps are regarded as the samples of one space, in order to represent the patterns of those pixels. In addition, HCF costs much time resources and memory resources.

Those previous algorithms such as GMM, ViBe and HCF have several serious drawbacks if we apply them to our tasks. First, they require heavy computational cost and resources in processing the satellite videos, since the pixel-based modeling has heavy computational loadings. Second, they are relatively inefficient in distinguishing the differences between the moving targets and the ego-motion of the satellites. To this end, we propose a noise model in isolating the moving background and detecting the potential moving targets, simultaneously. Moreover, the proposed tiny moving vehicle detecting is implemented in the spatiotemporal domain. In terms of sub-pixel-level moving and the neighborhood similarity, we model pixels of inter-frame differences spatially rather than temporally. The most of visual tracking algorithms can be roughly categorized into two classes, i.e., generative and discrimination algorithms. Generative algorithm focuses on searching potential regions that are similar to the labeled target region within a neighborhood region, while the discrimination algorithm incorporates background information. Discrimination tracker works as a binary classifier which distinguishes the interested

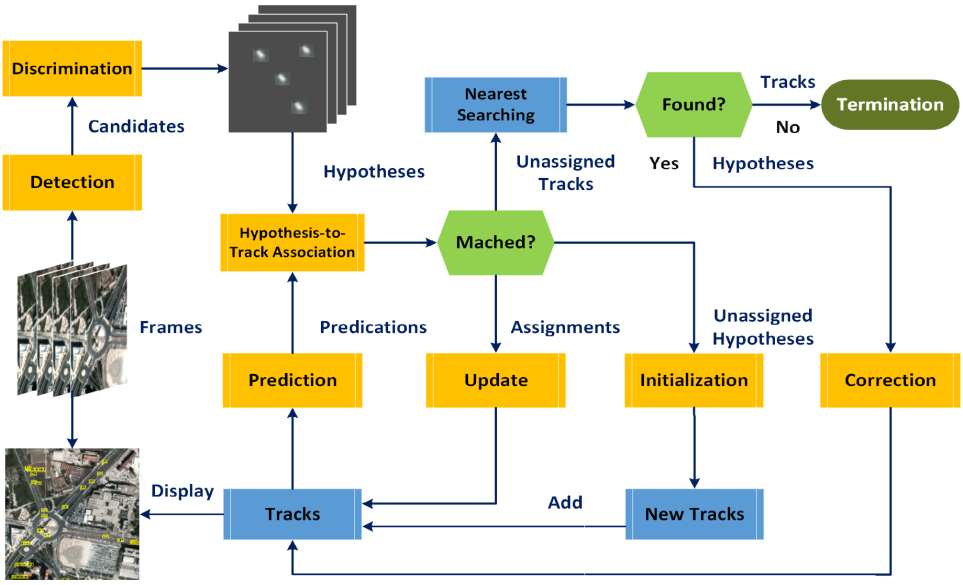


Fig. 5. The overall framework of tiny moving vehicles detection.

region from the background. Some fusion approaches try to integrate generative and discrimination algorithms [20]. Recently, in order to analyze underlying dynamics of the cellular procedure, Liang *et al.* [21] proposed a novel particle tracking algorithm based on multiple hypothesis tracking (MHT). A practical MHT is composed of three modules, i.e. detection, state estimation, prediction and linking [21], [22]. The cell tracking of fluorescence microscopy resembles vehicle tracking of satellite video to some extent.

III. Methodology of Detecting Tiny Moving Vehicles

The whole section is divided into three subsections. Firstly, we discuss the overall framework of the proposed tiny moving vehicles detection algorithms in Sec. III-A. Note that the proposed framework is under Kalman filter (KF) tracking framework as shown in Figure 5. These tiny moving vehicles are detected in each local region in Sec. III-B. We subsequently propose the discrimination algorithm to remove the false detected components in Sec. III-C.

A. Overview of the Proposed Algorithms

Key Concepts: Before fully developing our framework, some key concepts are explained here. *Detection* or a *detector* is a potential-vehicle detecting procedure that embodies the proposed local tactic and noise modeling algorithm. *Candidates* are the outputs of the detection that are composed of true vehicles and some noises. *Discrimination* or a *discriminator* is the distinguishing procedure between true vehicles and existing noises, including the proposed region growing and multi-morphological-cue based discrimination algorithms. *Hypotheses* are the outputs of the discrimination that are composed of true vehicles and a few noises. In addition, the final outputs of detecting and tracking framework are also defined as hypotheses. The *state* is a vector that includes position, velocity, and acceleration of a vehicle in a time step. The *track* is a sequence of states of a vehicle in the temporal domain. A track is marked with a unique ID and assigned with a KF. *Association* is a matching procedure that meets the minimum cost. *Prediction* is a current position of a track that is inferred by its KF.

The widely-used object tracking methods include Kalman filter, particle filter [6], and mean shift [23]. As shown in Figure 5, our whole vehicle tracking pipeline is based on Kalman filter, which is one of the most classical tracking algorithms. In Figure 5, the KF is the central module of the processing framework, which includes several interactive branches, as follows.

(1) Initialization. Initialization is to determine the initial state of a track. The hypotheses of the current frame and the previous frame are associated using Hungarian algorithm [24], [25]. So, we can derive their velocities and positions. Besides, their initial accelerations are regarded as zero.

(2) Prediction. The current state of one vehicle tracked can be inferred from the previous observation.

(3) Hypothesis-to-Track Association. The discriminator yields hypotheses, while the tracks yield predictions. In this stage, hypotheses are matched with predictions in order to meet the minimum cost. Here, the cost between a hypothesis and a prediction is defined as Euclidean distance. Hungary algorithm is employed to derive an optimal association between hypotheses and predictions. Hypothesis-to-track association yields assignments, unassigned tracks, and unassigned hypotheses. Assignments are optimal matches. Unassigned tracks are those tracks that do not successfully match with any hypothesis, likewise, unassigned hypotheses do. Then, assignments are utilized to update the stages; unassigned tracks are further processed in the nearest searching stage; unassigned hypotheses are used to initialize new tracks.

(4) Update. The state vectors of hypotheses of the assignments are used to update the state of their corresponding KFs.

(5) The Nearest Searching, correction and termination. We can not simply discard the unassigned tracks because their

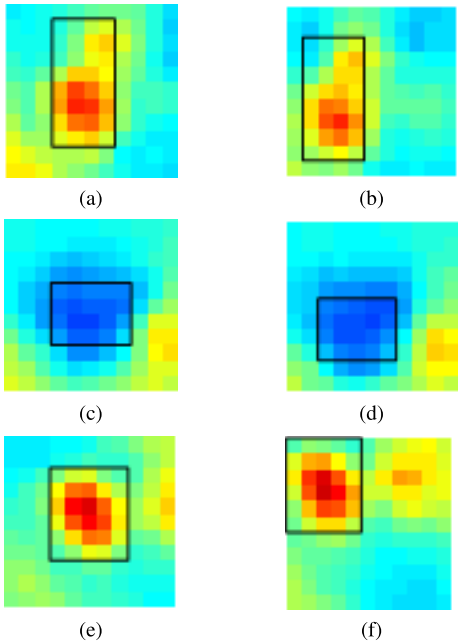


Fig. 6. Some cases of the nearest searching, wherein the black rectangles denote tracking vehicles. (a), (c), and (e) are the previous positions of vehicles, while (b), (d) and (f) are corresponding current positions. Please note that it is hard to recognize vehicles by naked eyes in original RGB images due to the low-contrast; so, the original images are converted to colorful pseudo color images just for view.

corresponding hypotheses may be missed by the detector or the discriminator. So, the nearest searching strategy is applied to find out whether there exists a connected region which resembles the tracking vehicle around the previous position of the vehicle. The matching resorts to structural similarity index (SSIM) [26]. If the similar region is found, the track is updated; otherwise, the track is terminated. The nearest searching using SSIM is illustrated in Figure 6. (a), (c) and (e) are from the previous frame where vehicles are marked by black rectangles, while (b), (d), and (f) are from the current positions where the black rectangles denote the results of the nearest searching. These experimental results demonstrate the efficiency of such a nearest searching strategy.

We will fully explain the above four steps of detecting the tiny moving objects in the next few sections. Once the objects are detected, Kalman filter is adopted to fit the motion of these vehicles, which is the optimal solution in the linear and Gaussian situations. Although the motions of vehicles in the real world are very complex, from an approximate viewpoint, this non-linear procedure can be decomposed into a series of linear procedures. So, KF is a simple but effective tool to measure, predict and track the motion of a moving vehicle. The evolution function of the system is defined as

$$\mathbf{x}_i = \mathbf{F}_i \cdot \mathbf{x}_{i-1} + \mathbf{v}_i \tag{1}$$

where $\mathbf{x}_i$, $\mathbf{F}_i$ and $\mathbf{v}_i$ denote the state vector, the evolution matrix and the procedure of noise vector, respectively, and the subscript i indicates the time step of a frame. Position, velocity, and acceleration of a vehicle constitute $\mathbf{x}_i$, namely

$$\mathbf{x}_i = [x, y, v_x, v_y, a_x, a_y]^T \tag{2}$$

Without loss of generality, we assume that vehicle targets move in constant acceleration and straight line during each fixed interval. So, the evolution matrix $\mathbf{F}_i$ can be written as

$$\mathbf{F}_i = \begin{bmatrix} 1 & 0 & \tau & 0 & \tau^2/2 & 0 \\ 0 & 1 & 0 & \tau & 0 & \tau^2/2 \\ 0 & 0 & 1 & 0 & \tau & 0 \\ 0 & 0 & 0 & 1 & 0 & \tau \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

The measurement function can be written as

$$\mathbf{y}_i = \mathbf{H}_i \cdot \mathbf{x}_i + \mathbf{n}_i \tag{4}$$

where $\mathbf{y}_i$, $\mathbf{H}_i$ and $\mathbf{n}_i$ denote measurement vector, measurement matrix, and measurement noise, respectively. The definition of $\mathbf{H}_i$ in this study is

$$\mathbf{H}_i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{5}$$

Assuming that a set of measurements is obtained, i.e. $\mathbf{y}_{1:i} = \{y_k | k = 1, 2, \cdots, i\}$, KF recursively derives posterior PDF of the state vector $\mathbf{x}_i$ via Bayesian theorem, i.e.

$$p(\mathbf{x}_i \mid \mathbf{y}_{1:i}) = \frac{p(\mathbf{y}_i \mid \mathbf{x}_i)\, p(\mathbf{x}_i \mid \mathbf{y}_{1:i-1})}{p(\mathbf{y}_i \mid \mathbf{y}_{1:i-1})} \tag{6}$$

B. Motion-Based Detection Using Local Noise Modeling

Inter-frame difference is a conventional but effective tool to discern changes between two frames. In contrast to pixel-based GMM, ViBe and HCF, it has two notable merits: high efficiency and low memory consuming. However, the traditional inter-frame difference [27] is based on a predefined threshold to separate moving pixels and background. Specifically, the grey-level image is converted to a binary image wherein ones denote moving pixels, while zeros denote stationary background pixels. This procedure is termed as *binarization* in our paper. Essentially, binarization differentiates moving pixels from the whole inter-frame difference image. But a fixed binarization threshold cannot be adapted to large-scale intro-variant scenarios of satellite videos.

In order to address the aforementioned challenges, we propose a novel detecting method based on local tactic. It is conceptualized as motion-based detection using local noise modeling. An adaptive binarization is derived by noise modeling in each local region, dealing with the variances of the local area by the neighborhood similarity.

1) Local Tactic: A local tactic is designed to tackle the dramatic intro-variance within a frame. Specifically, a 2D rasterizing is implemented along the vertical and horizontal directions in a frame. The original frame is converted into paved local areas. The size of a local is empirically set as 30×30 square pixels. For one thing, a local region has a much lower degree of heterogeneity than the whole frame. For another, it integrates local information to reduce the interference of the moving background. It greatly facilitates the following detecting.

2) Detecting Method: The proposed detecting method is composed of four stages as shown in Fig. 7, (1) deriving inter-frame difference images, (2) estimating noise distribution, (3) binarization and (4) logical AND operations to finally get the detection results. Our major contribution is to estimate

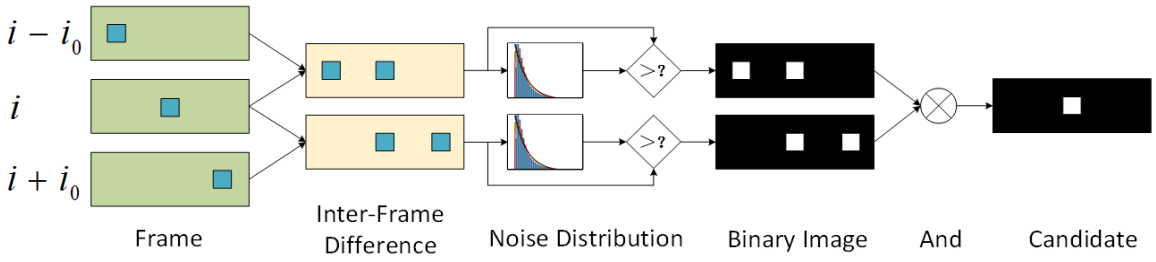


Fig. 7. The flow diagram of the motion-based detection algorithm via noise modeling.

noise distribution in the step (2), which can yield an adaptive binarization threshold of each local area in step (3). Each step details as follows.

(1) Deriving inter-frame difference images. Inter-frame difference is a trivial operation. Here, we provide a novel viewpoint on inter-frame difference images. By taking the frame as a 2D signal consisting of original optical signal and additive random noise signal, i.e.

$$G_i(x, y) = g_i(x, y) + n_i(x, y) \tag{7}$$

where $G_i(x, y)$ denotes the grey-level value of pixel (x, y) in frame i. Please note that in our study RGB frames are converted to grey-level images and all operations are implemented in grey-level images. $g_i(x, y)$ denotes the original amplitude of the pixel (x, y) in frame i, while $n_i(x, y)$ denotes the corresponding noise signal. Accordingly, the absolute inter-frame difference of two registered frames can be regarded as a set of random noise, i.e.,

$$\begin{aligned} D_{i,i+k}(x, y) &= |G_i(x, y) - G_{i+k}(x, y)| \\ &= |n_i(x, y) - n_{i,i+k}(x, y)| \end{aligned} \tag{8}$$

where $D(\cdot)$ denotes absolute inter-frame difference, k denotes the k frames interval. From Eq. 8, the inter-fame difference image signal is only corresponding with noises when two frames are registered. However, there still exists some outliers. These outliers are composed of tiny moving vehicles and non-vehicle targets. Thus the next issue is how to differentiate these outliers from random noises.

(2) Estimating noise distribution. Detecting the pattern of tiny moving vehicles is a challenge since noise patterns will blur the underlying patterns of tiny moving vehicles. Thus in this step, the key idea is to fit the noise patterns, namely $D_{i,i+k}(x, y)$ in Eq. 8, by the probabilistic distributions.

Intuitively, the value difference of the same pixel at two consecutive frames should approximate zero, while the value differences of the pixels of noise patterns, or moving vehicles should be larger than zero, while the values of outliers may be large. For the noise patterns, the noise pixels are inliers, while the other pixels are outliers. Figure 8 shows the histogram of the value differences of pixels of two consecutive frames. The amplitude histogram of noises exhibits notable regulations, like smooth decaying and a heavy tail. In the pattern of noise, true noise pixels are inliers, and the other pixels are outliers. Thus the heavy tail in Figure 8 should be corresponding to the

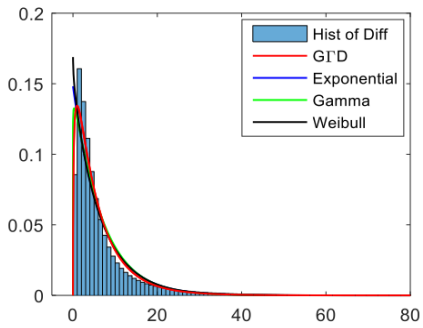


Fig. 8. Amplitude histogram of noises and some fitted probabilistic distributions. Note that "Histogram" and "Difference" are abbreviated as "Hist" and "Diff", respectively.

outliers. We further employ an adaptive threshold to split the outliers from inliers noises.

The probabilistic distribution is adopted to fit the histogram and derive a binary threshold given a probability. Thus several widely-used heavy-tail distributions, such as exponential distribution, Gamma distribution, Weibull distribution and generalized Gamma distribution (GΓD) are tested and compared in Figure 8. Quantitatively, Kullback-Leibler (KL) distance [28] and Kolmogorov-Smironv (KS) distance, (also known as Kolmogorove-Smirnov test [29]), are introduced to quantify the fitting performance of different distributions, as shown in Table I. The smaller scores of KL and KS distance indicate the better results for fitness.

The quantitative experiments further prove the above hypothesis. It also shows that the three-parameter distribution, GΓD with a higher degree of freedom (DoF) outperforms the other distributions. Alternatively, one-parameter distribution âŁ" exponential distribution also fits the noise distribution very well. Nevertheless, the parameter estimation of GΓD is more difficult, higher computational load and more time consuming than exponential distribution. To make a balance between accuracy and computational load, exponential distribution is adopted to fit noises; and the cumulative density function (CDF) is

$$c_E(x; \lambda) = \begin{cases} 1 - \exp(-\lambda x) & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{9}$$

(3) Binarization. Once we fit the distribution of pixel value differences of consecutive frames, we can utilize an adaptive threshold of binarization to determine the outliers.

TABLE I

KL AND KS DISTANCE OF SOME NOISE PROBABILITY MODELS. THE SMALLER VALUES, THE BETTER PERFORMANCE

Distance	Exponential	Gamma	Weibull	GΓD	Frame
KL	0.0959	0.0914	0.0919	0.0544	50
KS	0.0959	0.1018	0.0891	0.0813	
KL	0.0864	0.0812	0.0862	0.0579	100
KS	0.0875	0.0988	0.0896	0.0865	
KL	0.0846	0.0800	0.0845	0.0531	500
KS	0.0854	0.0964	0.0859	0.0816	

In particular, we introduce a predefined probability p_{fa} to derive the binarization threshold, namely [7]

$$th = c_E^{-1}(1 - p_{fa}; \lambda) \tag{10}$$

where $c_E^{-1}(\cdot)$ denotes the inverse function of the distribution in Eq. 9. We set p_{fa} as 5×10^{-2} here. If the pixel value difference is bigger than the binarization threshold th, this pixel would be taken as an outlier. By virtue of such a binarization algorithm, we turn the original image into a binary image: inliers have zero pixel values, outliers are ones.

(4) Logical AND. These outliers comprise vehicles and some noises. In the binary difference image, a true vehicle target exhibits as two symmetrical blobs. One blob indicates the current position, and another indicates the previous or future position. We derive the intersection of two binary inter-frame difference images, explicitly the difference between frame i and frame $i - i_0$ vs. the difference between frame i and the frame $i + i_0$, wherein i_0 is set as 10. to determine the current positions of vehicles. It is a Boolean operation that only one and one yield one, which is named as, "Logical AND". In addition to eliminating ambiguities, logical AND also reduces the existing noises due to their random appearing.

C. Region Growing and Multi-Morphological-Cue Based Discrimination

There still exists some noises in candidates. These noises include irregular noises and regular noises. Irregular noises result from dramatic illumination variants or slight deviations between frames. They may randomly appear in some consecutive frames. Generally, it is not necessary to design an algorithm of pruning the irregular noises since KF tracking can gradually eliminate this type of noises. Another noises derived from slight deviation between frames have to be addressed in particular. In contrast, we term the background moving patterns as regular noises. Particularly, these noises are caused by the slight deviation of satellite moving. Such a deviation may be appeared/detected as the edges or corners of some static objects in the frames. Even worse, these detected corners or edges exhibit a relative moving pattern with respect to the moving background. The regular noises have to be pruned by our algorithms. We visualize the regular noises in Figure 9. In order to differentiate this background moving from vehicle motions, we adopt a descriptive term, "pseudo motion", to indicate the moving patterns of regular noises, as shown in Figure 9. The background moving leads to the

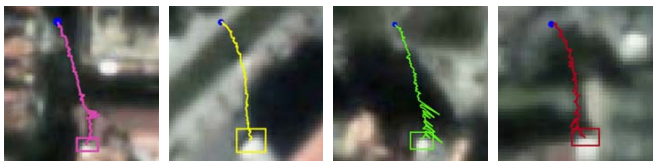


Fig. 9. The visualization of regular noises: the stationary corners or edges. Rectangles denotes the starting locations, while the filled blue points are their terminal positions.

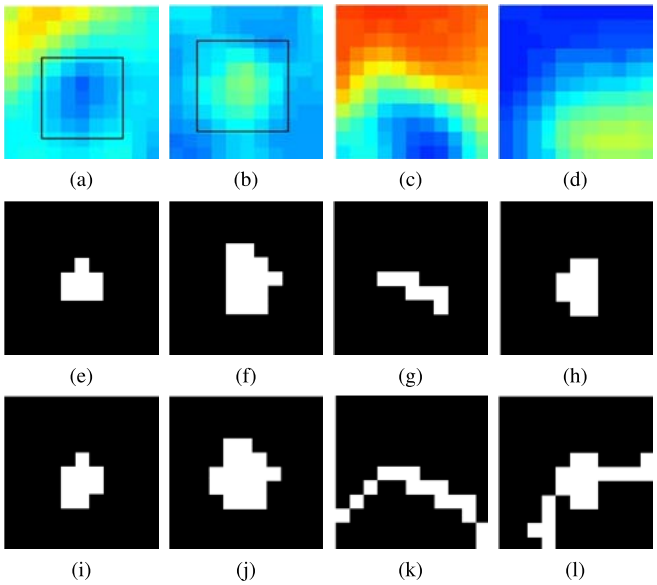


Fig. 10. Detecting and region growing results of vehicle targets and a noise. (a) - (d) are pseudocolor images converted from original RGB images for a view, where real vehicles are marked with black rectangles, and (c), (d) are falsely detected edges or corners of buildings. (e) - (h) are corresponding foregrounds generated by the detector. (i) - (l) are corresponding reconstructed geometries.

deviations between frames and regular noises. Furthermore, regular noises embody the background moving and represent pseudo-motion.

To this end, we propose a novel discrimination algorithm using the geometrical and neighborhood information. This discrimination algorithm includes two parts, i.e., region growing to reconstruct candidate geometry and multi-morphological-cue based discrimination to distinguish noises from vehicles. The key idea is that a vehicle target is a singular point in the 2D temporal domain. By contrast, these regular noises share similar temporal distributions as their neighborhood in frames. If the candidates can be connected with their similar neighborhood pixels, we can differentiate vehicles from regular noises in terms of shape: the vehicle targets approximate rectangles, while regular noises can be taken as arbitrary shapes.

1) Region Growing: A region growing algorithm is proposed to connect a candidate with its similar neighborhood pixels. This procedure is namely to reconstruct the whole geometry of the candidate. From Figure 10, the detector only yields a partial geometry of a candidate, because of the overlap of positions of a candidate in two adjacent frames. The region growing utilizes the detected partial geometry to restore the whole geometry of the candidate. Neighborhood area is defined as a 11×11 pixels window in the candidate. Gaussian distribution is employed to measure the similarity between neighbor pixels and the candidate. The CDF of Gaussian

distribution is

$$c_G(x\mu,\sigma) = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \tag{11}$$

where $erf(\cdot)$, μ and σ denote the related error function, the mean and the standard deviation, respectively. These parameters of a Gaussian distribution can be estimated using the values of those pixels of the candidate. A range can be obtained given the predefined the lower bound probability p_{fa}^{-} and the upper bound probability p_{fa}^{+}, i.e.

$$\begin{aligned} th_G^{-} &= c_G^{-1}({p_{fa}}^{-}\mu,\sigma) \\ th_G^{+} &= c_G^{-1}({p_{fa}}^{+}\mu,\sigma) \end{aligned} \tag{12}$$

where th_G^{-} and th_G^{+} represent the lower and upper bound threshold, respectively. p_{fa}^{-} and p_{fa}^{+} are set as 5×10^{-3} and $1 - 5 \times 10^{-3}$, symmetrically. If the grey-level value of a pixel inner the searching window is in $\left[th_G^{-}, th_G^{+}\right]$, the pixel will be re-classified as a candidate pixel. Those new candidate pixels connected with the original candidate are reserved. Finally, the results of region growing are shown in Figure 10 (i) - (l). From Figure 10, the detector captures only partial geometries of the candidates. Then, the proposed region growing algorithm reconstructs the whole geometries of the candidates. The region growing results also demonstrate the feasibility of discrimination in terms of shape.

2) Multi-Morphological-Cue-Based Discrimination: After region growing, we adopt a series of morphological properties to differentiate vehicle targets and noises. The employed morphological cues include area, extent, major axis length, and eccentricity as follows.

Area. The number of pixels of a candidate.

Extent. The ratio of pixels of a candidate to the area of candidate's bounding box.

Major Axis Length. If an ellipse has the same normalized second central moments as the connected region of the candidate, the major axis length of the ellipse is defined as the major axis length of the candidate.

Eccentricity. The eccentricity of a candidate is equal to the eccentricity of the above ellipse.

Area and major axis length cues represent the size of a candidate, while extent and eccentricity cues measure the similarity between a candidate and a rectangle. The spacing in the satellite videos represents about 1 meter in the real world. Thereby, these morphological cues indicate a real shape of a vehicle. They constitute a robust feature of vehicles because vehicles are rigid bodies without any deformation in satellite videos.

IV. Metrics and Protocols in Performance Evaluation

The widely-used evaluation metrics on object detection are precision/recall curve, and average precision (AP) [30]. These metrics are widely used in the traditional visual object benchmarks, i.e. PASCAL VOC [30] and MOT [31]. However, these metrics and protocols are not enought to evaluate the performance of our tasks. The key challenge is caused by the fact that each vehicle has only several pixels on each frame. Therefore, in order to evaluate the performance of tiny moving targets comprehensively, we systematically introduce a complete set of evaluation protocol in measuring the algorithm performance on our detection tasks in Sec. IV-B. Our evaluation protocol is built upon the existing evaluation metrics detailed in Sec. IV-A.

A. Evaluation Metrics

Generally, a single criterion cannot reckon the performance of detecting and tracking objectively and comprehensively. To the best of our knowledge, it is the first time to introduce a series of systematic evaluation metrics, including precision, recall, Jaccard similarity, F_1-score, MOTA, MOTP, etc.

Precision: With respect to detection performance evaluation, it is the most important to determine whether a hypothesis is a true positive (TP) that is an accurate target correctly covered by an output, or a false positive (FP) that is a non-target falsely covered by an output. Those missed accurate targets are called as false negatives (FNs). The ratio of the accurate targets to the detected targets is Precision, i.e.

$$P = \frac{TP}{TP + FP} \tag{13}$$

Recall: Recall measures the ability of a detector to capture true targets, which is equal to the ratio of TP to the number of all existing true targets, namely

$$R = \frac{TP}{TP + FN} \tag{14}$$

F_1-Score: F_1-score is a traditional criterion of binary classification between interest targets and non-targets, which is equal to the harmonic mean of Precision and Recall, i.e.

$$F_1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

Jaccard Similarity: Jaccard similarity is a criterion of evaluating tracking performance, which integrates TP, FP and FN as follows, [21]

$$J = \frac{TP}{TP + FP + FN} \tag{16}$$

MOTA: Multiple object tracking accuracy (MOTA) is a tracking performance metric to quantify multiple object tracking performance. The definition of MOTA is [31], [32]

$$MOTA = 1 - \frac{\sum_i (FN_i + FP_i + IDSW_i)}{\sum_i GT_i} \tag{17}$$

where FN_i, FP_i, $IDSW_i$ and GT_i represent the number of FN, FP, $IDSW$ and the ground truth, respectively. Frame i. $IDSW$ means identity switch of the trajectories associated to the ground truth, and please refer to [31] for more details. Obviously, MOTA score ranges from $-\infty$ to 1, and the bigger it is, the better detecting and tracking is.

MOTP: Multiple object tracking precision (MOTP) is adopted to measure the positioning precision of the detecting algorithms, which can be written as [25], [31], [32]

$$MOTP = \frac{\sum_i IoU_i}{\sum_i M_i} \tag{18}$$

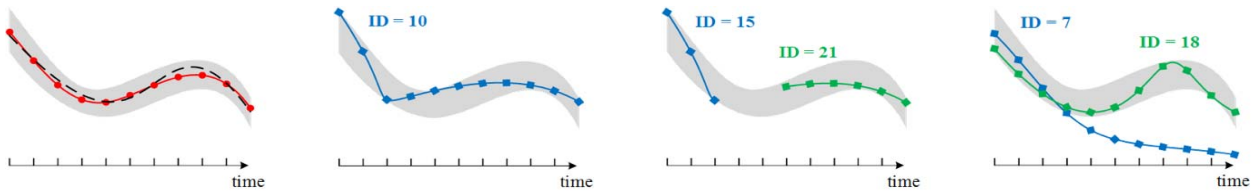


Fig. 11. Some cases of hypothesis-to-ground-truth associations. In panel (a), the black dash line, the red solid line with filled circles, and the grey filled polygon denote the accurate trajectory of a vehicle target, manually annotated ground truth of the trajectory, and the region where an output belongs to the trajectory, respectively. (a) illustrates that manually annotated ground truth cannot completely fit the accurate trajectory. (b) shows a hypothesis that the trajectory fits the ground truth in panel (a). (c) shows that two hypotheses cover the same ground truth, wherein IDSW is counted. From panel (d), the hypothesis whose ID equal to 7 outperforms the hypothesis whose ID equal to 18 during the first three frames. However, the former gradually loses the main pattern of the ground truth, the latter follows the ground truth more closely.

where IoU_i (intersection over union [30]) denotes the sum of overlap ratio of hypotheses to ground truths and M_i is the number of matches of ground truths and hypotheses in frame i. From above definition, MOTP score ranges from 0 to 1, and the bigger that is, the more precise the derived location is.

B. The Evaluation Protocol

The IoUs of bboxes between a hypothesis and a ground truth is adopted as the similarity between the hypothesis and the truth. Similar to the aforementioned hypothesis-to-track association, the matching of multiple hypotheses and ground truths also resorts to Hungary algorithm [24] in spatiotemporal domain not only in each frame.

Some cases of associations are listed in Figure 11 (b)-(d), and actual situation is more complicated than that. The protocol of performance evaluation of the proposed detection and tracking algorithm details as following:

1. The IoUs of each hypothesis and each ground truth can be obtained. Then, the reciprocal of a IoU is the distance between a hypothesis and a ground truth. Note that all IoUs are added with a very small value to avoid zero denominators. The threshold of matching distance is set as 50, empirically, which is equal to a very small IoU value, 0.02. It means that if the overlap of the hypothesis and the ground truth is bigger than 0.02, the hypothesis is regarded to cover the ground truth. In contrast, the IoU ratio is set as 0.5 for the detection of general objects, such as pedestrian, airplane, bicycle, etc., which cannot adapt to tiny vehicles in satellite videos. Obviously, the smaller the distance, the smaller is the cost of the association between hypothesis and ground truth. All distances constitute a cost matrix, given M hypotheses and N ground truths, i.e.

$$CM_t = \begin{bmatrix} \frac{1}{IoU_{1,1}} & \frac{1}{IoU_{1,2}} & \cdots & \frac{1}{IoU_{1,N}} \\ \frac{1}{IoU_{2,1}} & \frac{1}{IoU_{2,2}} & \cdots & \frac{1}{IoU_{2,N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{IoU_{M,1}} & \frac{1}{IoU_{M,2}} & \cdots & \frac{1}{IoU_{M,N}} \end{bmatrix} \tag{19}$$

where t indicates the frame number.

2. Repeat step 1 for K consecutive frames. Then, the cost matrices of these frames can constitute a cost tensor, namely

$$\mathbf{CT} = [CM_1, CM_2, \cdots, CM_K] \tag{20}$$

3. The optimal associations of hypotheses and ground truths can be obtained using Hungary algorithm. A time window is employed to reduce computational load and memory consumption. Explicitly, the association is implemented among ten consecutive frames rather than all frames, namely the K in Eq.20 is set as 10.

4. Repeat the step 1-3.

5. Finally, the metrics are calculated based on the associations.

V. Experiments and Discussion

A. Experimental Setups

1) Dataset: The experimental satellite video is provided by CGSTL. The videos are captured on March 7, 2017. A video satellite recorded a region in Valencia, Spain. The information of the satellite video is detailed in Table II. The study video is free provided by CGSTL for scientific research. Anyone can purchase the satellite videos from their official website http://mall.charmingglobe.com/videoIndex.html.

2) Competitors: We compare several other algorithms in detecting tiny moving vehicles, including GMM, ViBe and HCF. In order to fairly compare the proposed algorithm with baseline algorithms, GMM, ViBe and HCF are also linked with the same KF tracking framework.

3) Ground-Truth: We annotated the experimental satellite video to quantitatively evaluate the proposed algorithm. The annotation of satellite videos is arduous work, because vehicles are hard to be distinguished from the background in naked eyes for the lack of distinctive prominent features, which is illustrated in Figure 12. In Figure 12 (a), (b) and (c), there is no significant difference between real vehicle targets and the background, especially dark vehicles, like Figure 12 (b). Figure 12 (d) shows that a noise signal may be a stationary vehicle, and resembles the true positive, which results in ambiguities in the interpreting work. In order to tackle this problem, we inspect the previous and future frames to determine whether some candidate targets belongs to vehicles. In other words, we go through short-term consecutive frames to seek out moving vehicles. Besides the difficulty to

TABLE II
THE DETAILED INFORMATION ABOUT THE EXPERIMENTAL SATELLITE VIDEO

District	Frame Rate (fps)	Resolution (m)	Duration (s)	Height×Width (pixel^2)
Valencia, Spain	20	1.0	29	3072×4096

Corner	Top Left	Top Right	Bottom Left	Bottom Right
Lat. Lon.	39.4989N 0.3719W	39.4928N 0.3278W	29.4731N 0.3775W	39.4669N 0.3333W

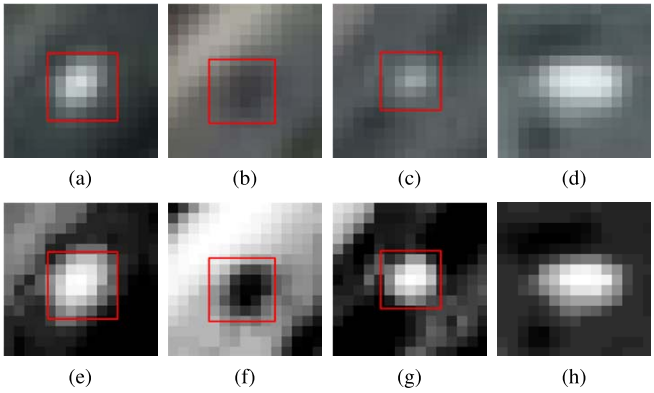


Fig. 12. Some vehicle samples and a noise sample. These images are scaled for viewing. (a), (b), (c), and (d) are RGB images where red rectangles represent real vehicles, while a noise signal exists in (d). Greyscale images (e), (f), (g), and (h) are corresponding enhanced images of (a), (b), (c), and (d), to improve contrast for viewing.

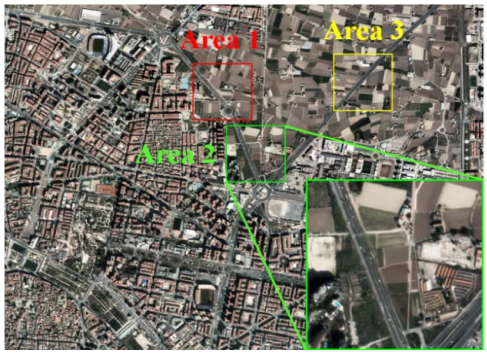


Fig. 13. The annotated regions of the experimental satellite video.

detect vehicles, we can hardly annotate all vehicles of the scope spanning about 3Ã-4 square kilometers frame by frame. We also seek the trade-off between workload and annotation accuracy here: first, three areas with 500 × 500 pixels of the video are randomly selected to be annotated, as illustrated in Figure 13; second, we manually annotate vehicles every 10 frames, while the ground-truth of other frames are obtained by linear interpolation. Figure 14 shows a representative scene of the annotation. The vehicle numbers of ground truth of area 1, area 2 and area 3 are 49, 41 and 29, respectively. Particularly, we utilized the Ground Truth Labeler App in MATLAB 2018a to help annotate the satellite video.

4) Complexity: Our model trained on CGSTL satellite videos dataset with 32G RAM, Core^{TM} i5-4590 CPU @ 3.30GHz, Geforce GTX 760. The operating system is wins7 - 64bits, and executive software is matlab 2018a. Algorithm complexity refers to the resources needed by the runtime after the algorithm is transformed into an executable program, including time resources and memory resources, namely time complexity and space complexity. The complexity includes the complexity of detecting and the complexity of tracking. As shown in Table III, N represents the number of pixels; M represents the number of vehicles, N>M; D represents the number of network layers and relevant parameters. GMM, ViBe, especially HCF have bigger redundant time consumption and poorer detecting results than our proposed method. While the complexity of HCF is related to the number of network layers and the number of nodes exponentially. Our proposed method achieves the optimal results in terms of time complexity and space complexity compared with other methods.

Fig. 14. One shot of the ground truth annotation. (a) shows the locations of vehicles, while (b) represents their corresponding IDs.

TABLE III
THE COMPLEXITY OF DIFFERENT ALGORITHMS

Complexity	Detecting	Time(hours)	Tracking	Time(hours)
Ours	O(N)	0.5	O(M)	0.34
GMM and ViBe	O(N)	>0.5	O(M)	(0.5,1)
HCF	O(N*D)	–	O(M*D)	19.43

B. Experimental Results and Discussion

1) Quantitative Results and Analysis: The quantitative results of comparison experiments are detailed in Table IV. Table IV shows that the proposed framework obviously outperforms ViBe, GMM and HCF in any criterion. Specifically, ViBe and GMM detect around 50% of vehicles, but yield about 90% of FPs. The high false positive extremely degrades the performance of ViBe and GMM. The results from ViBe

(a) Frame 50 (b) Frame 100

(c) Frame 150 (d) Frame 200

Fig. 15. The example detecting results of four frames. Note that the enlargements of scenes is Figure 16.

and GMM totally neglect the background moving and also fail to separate moving vehicles from noises. Furthermore, they are both sensitive to the varying of the pixels, so, they report a trivial Recall score: 50%. HCF tracks around 80% of vehicles. High recall and precision is fine, but it cost much time resources, which can only track targets not real-time tracking. On the contrary, our method has the unique capability of perceiving pixel moving not only varying both detecting and tracking. Leveraging the very high Precision, our method reports good scores of F_1-score, Jaccard Similarity, and MOTA, etc. These criteria are closely related to Precision. Moreover, the Recall score of our method is relatively low, although it is about 10% higher than ViBe and GMM. Some related criteria, Jaccard Similarity, and MOTA, is affected to some extent. The location precision metric MOTP shows that average overlap between ground truths and hypotheses is 0.52. The tiny size of vehicle targets leads to difficulty in precisely locating them. But we think the location precision and pixel-level deviation meet the application demand, especially in urban traffic surveillance.

2) Qualitative Results and Analysis: We give some visual results of detecting multiple tiny moving vehicles. Figure 15 shows four frames, frame 50, 100, 150, 200. Intuitively, many vehicles in the main arteries are detected by the proposed algorithms, and there are a few FPs because most of the annotated labels also exist in the main arteries. Besides, the number of detected vehicles is growing gradually frame by frame, which illustrates that tracking also facilitates detecting.

In order to clearly observe the detecting and tracking details, we provide four enlargements as shown in Figure 16. From Figure 16, we can obtain the dynamics of not only the moving of vehicles but also the detecting and tracking procedure. Once a vehicle has been repeatedly detected of two consecutive frames, a unique ID is assigned to it and simultaneously a KF is allotted to it. The position and velocity provided by the detector also are used to initialize the KF. Subsequently, according to the designed framework, the detector provides the current state frame by frame, while the KF continuously embodies the latest state of the tracking vehicle and further updates its systematic model. Therefore, the proposed processing workflow can detect tiny moving vehicles accurately and precisely.

3) Trajectories Analysis: Figure 17 shows four trajectories tracked by the proposed algorithms. These trajectories include linear tracks and curved tracks, which demonstrates the above theory that a series of linear procedures can approximate a non-linear procedure as accurate as possible. These four trajectories also cover different traffic scenarios, including a straight

TABLE IV

THE EVALUATION SCORES OF THE PROPOSED ALGORITHM AND BASELINE ALGORITHMS. NOTE THAT R REPRESENTS RECALL, P REPRESENTS PRECISION, F_1 REPRESENTS F_1-SCORE, J REPRESENTS JACCARD SIMILARITY, MOTA IS MULTIPLE OBJECT TRACKING ACCURACY, AND MOTP IS MULTIPLE OBJECT TRACKING PRECISION

Area	Method	R (%)	P (%)	F_1	J	MOTA	MOTP
1	Ours	64.15	81.71	0.72	0.56	0.46	0.50
	ViBe	51.72	15.10	0.23	0.13	-2.45	0.39
	GMM	43.82	12.29	0.19	0.11	-2.75	0.37
	HCF	52.08	51.40	0.52	0.35	0.03	0.49
2	Ours	62.80	82.23	0.71	0.55	0.47	0.52
	ViBe	61.70	9.14	0.16	0.09	-5.56	0.45
	GMM	61.83	7.5	0.13	0.07	-7.08	0.39
	HCF	45.95	44.27	0.45	X0.29	-0.11	0.52
3	Ours	60.42	77.26	0.68	0.51	0.41	0.56
	ViBe	41.53	6.76	0.12	0.06	-5.35	0.47
	GMM	46.10	6.34	0.11	0.06	-6.41	0.42
	HCF	51.31	41.97	0.46	0.30	-0.17	0.53
Avg.	Ours	63.06	81.04	0.71	0.55	0.46	0.52
	ViBe	52.86	10.74	0.18	0.10	-3.92	0.43
	GMM	49.66	8.79	0.15	0.08	-4.72	0.39
	HCF	59.84	58.48	0.59	0.42	0.18	0.49

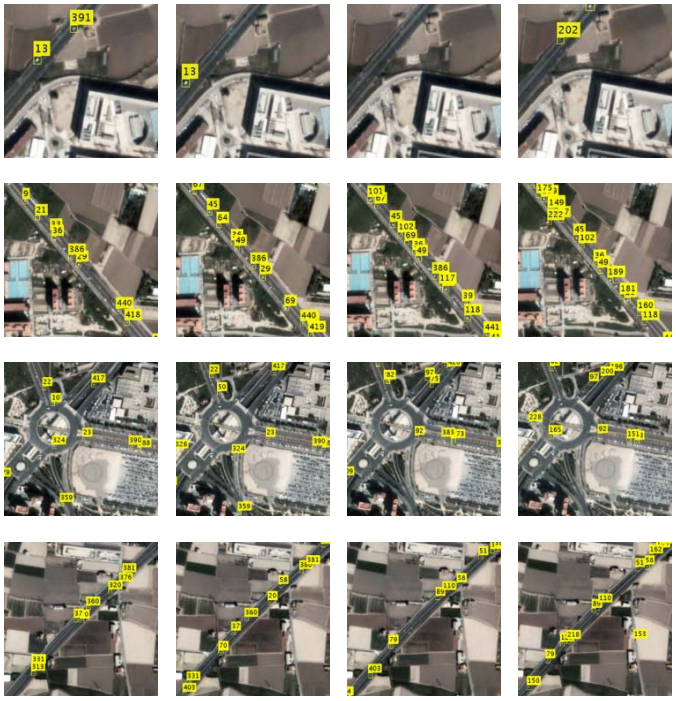


Fig. 16. The enlargements of four scenes in Figure 15.

artery in Figure 17(a), a right turn in Figure 17(b) and two roundabouts in Figure 17(c)-(d). It proves that the proposed algorithms can not only address simple traffic conditions but also adapt to complex traffic scenarios.

4) Detailed Comparative Analysis of Four Algorithms: Figure 18 shows the foreground segmentation results generated by our algorithms and baseline algorithms. From Figure 18 (a), our detector yields hypotheses composed of true vehicles and a few noises, while noises extremely outnumber true vehicles in Figure 18 (b) and (c). Candidate vehicles generated by the

Fig. 17. The tracks of four vehicles. Yellow lines indicate their moving tracks detected by our algorithms. Yellow rectangles mark their initial positions, while the filled red points denote their terminal locations. In contrast, the other methods, GMM and ViBe, cannot be used greatly to tracking vehicles.

(a) (b) (c)

Fig. 18. The foreground segmentation results of Frame 100. (a) the proposed algorithm, (b) ViBe, (c) GMM.

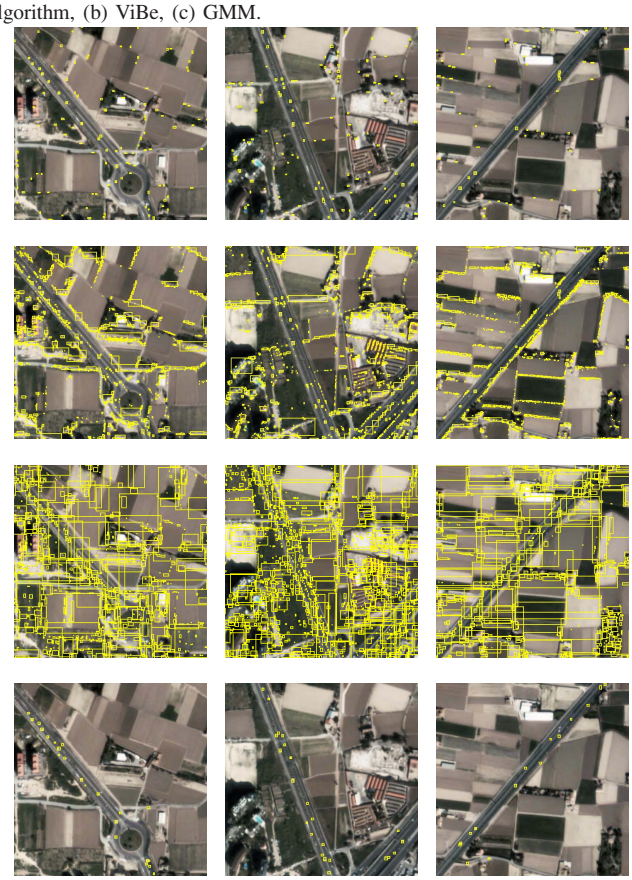
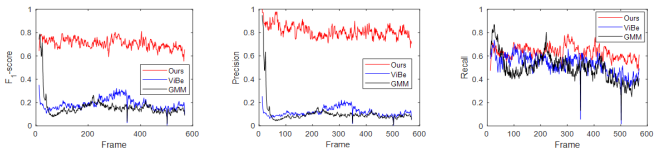


Fig. 19. Vehicle candidates yielded by our method (row 1), ViBe (row 2), GMM (row 3), and HCF (tracking-only) (row 4) in frame 100. The first, second and third column are corresponding to the vehicle candidates in Area 1, Area 2, Area 3, respectively.

proposed algorithm and competitors are shown in Figure 19. From Figure 19, our detector mainly perceives the moving vehicle pixels in the roads and yields limited false positives, while ViBe and GMM aimlessly detect varying of the pixels. It illustrates that ViBe and GMM are unable to separate the motions of vehicles from the slow and slight motions of background. ViBe and GMM try to estimate the pattern of each pixel using a non-parametric model or a Gaussian



(a) $F_1 - score$ (b) $Precision$ (c) $Recall$

Fig. 20. The comparative results of the proposed algorithm and baselines frame by frame of Table IV. Note that Jaccard Similarity, MOTA, and MOTP have the same results.

distribution, respectively. So, this strategy is very sensitive to the varying of the pixel and works well in common video processing, especially a stationary camera. However, they cannot address the satellite video processing because of neglecting local or neighborhood information. Our detection algorithms focus on a local area, not a single pixel, which can adapt to the moving background of satellite videos.However, HCF don not have foreground segmentation results, because it is not pixel-level tracking based. Besides, HCF is tracking-only method, but ours, Vibe and GMM is detection-tracking method. Although the tracking results of HCF look well, they can only tracking, that is, the initial position of vehicles are needed to be input, and the difficulty of this paper is tiny moving targets detection.

To mask a detailed contrast, we provide a detailed figure of detecting performance of the proposed algorithm and baselines, some scores frame by frame are presented in Figure 20. Figure 20 shows the most widely-used detecting metrics: F_1-score, Precision, and Recall. It further illustrates that the proposed algorithm outperforms baselines mainly by leveraging high Precision. From Figure 20, the F_1-score of ViBe and GMM rapidly decays at the beginning and then totally traps into the moving background. So, for ViBe and GMM, the background moving totally blurs the vehicle moving, leading to their failure. Likewise, this experiment demonstrates the outstanding performance of the proposed algorithm to separate the background moving and the vehicle moving.

VI. Conclusion

Satellite videos have the unique capability of observing city-scale regions. This paper addresses the tiny vehicle detecting algorithm in the satellite videos, and we design the practical detecting and tracking framework of tiny moving vehicles in satellite videos. It is the first time to adopt a probabilistic distribution to represent the pattern of noises in the spatiotemporal domain, which facilitates us to differentiate candidates from noises. We further propose the multi-morphological-cue based discrimination algorithm to distinguish true vehicle targets from a few existing noises. Another important issue is to introduce a series of evaluation metrics and to propose a complete evaluation protocol. The proposed algorithms are tested in three manual annotated areas of a satellite video, which are also compared with baseline algorithms. These experiments demonstrate the good performance of our algorithms.

Acknowledgment

The authors are grateful to CGSTL for providing the satellite video data used in this study. This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes processed satellite video. Contact fengxu@fudan.edu.cn for further questions about this work.

References

[1] L. T. Tan and L. B. Le, "Joint data compression and MAC protocol design for smartgrids with renewable energy," *Wireless Commun. Mobile Comput.*, vol. 16, no. 16, pp. 2590–2604, 2016.

[2] Y. Tian, R. Feris, H. Liu, A. Hampapur, and M.-T. Sun, "Robust detection of abandoned and removed objects in complex surveillance videos," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 5, pp. 565–576, Sep. 2011.

[3] C.-C. Chen and J. K. Aggarwal, "Recognizing human action from a far field of view," in *Proc. Workshop Motion Video Comput. (WMVC)*, 2009, pp. 1–7.

[4] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell. (IJCAI)*, 1981, pp. 674–679.

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 726–733.

[6] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[7] W. Ao, F. Xu, Y. Li, and H. Wang, "Detection and discrimination of ship targets in complex background from spaceborne ALOS-2 SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 2, pp. 536–550, Feb. 2018.

[8] Y. Zhou, H. Wang, F. Xu, and Y.-Q. Jin, "Polarimetric SAR image classification using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1935–1939, Dec. 2016.

[9] F. Xu, Y.-Q. Jin, and A. Moreira, "A preliminary study on SAR advanced information retrieval and scene reconstruction," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 10, pp. 1443–1447, Oct. 2016.

[10] A. Walha, A. Wali, and A. M. Alimi, "Video stabilization with moving object detecting and tracking for aerial video surveillance," *Multimedia Tools Appl.*, vol. 74, no. 17, pp. 6745–6767, 2015.

[11] B. P. Jackson and A. A. Goshtasby, "Registering aerial video images using the projective constraint," *IEEE Trans. Image Process.*, vol. 19, no. 3, pp. 795–804, Mar. 2010.

[12] H. Guo, S. Liu, T. He, S. Zhu, B. Zeng, and M. Gabbouj, "Joint video stitching and stabilization from moving cameras," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5491–5503, Nov. 2016.

[13] E. Molina and Z. Zhu, "Persistent aerial video registration and fast multi-view mosaicing," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2184–2192, May 2014.

[14] N. Jiang and W. Liu, "Data-driven spatially-adaptive metric adjustment for visual tracking," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1556–1568, Apr. 2014.

[15] P. Kaewtrakulpong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*. Boston, MA, USA: Springer, 2001.

[16] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 1999, pp. 246–252.

[17] O. Barnich and M. Van Droogenbroeck, "ViBE: A powerful random technique to estimate the background in video sequences," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., (ICASSP)*, Apr. 2009, pp. 945–948.

[18] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Robust visual tracking via hierarchical convolutional features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2709–2723, Nov. 2019.

[19] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3074–3082.

[20] L. Qin, H. Snoussi, and F. Abdallah, "Cascaded generative and discriminative learning for visual tracking," in *Proc. Int. Conf. Images Anal. Recognit. (ICIAR)*, 2013, pp. 397–406.

[21] L. Liang, H. Shen, P. De Camilli, and J. S. Duncan, "A novel multiple hypothesis based particle tracking method for clathrin mediated endocytosis analysis using fluorescence microscopy," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1844–1857, Apr. 2014.

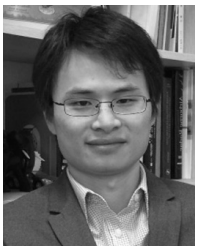
[22] S. T. Acton and N. Ray, *Biomedical Image Analysis: Tracking*. San Rafael, CA, USA: Morgan & Claypool, 2005.

[23] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 1, pp. 32–40, Jan. 1975.
[24] M. L. Miller, H. S. Stone, and I. J. Cox, "Optimizing Murty's ranked assignment method," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 33, no. 3, pp. 851–862, Jul. 1997.
[25] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," in *Proc. Int. Eval. Workshop Classification Events, Activities Relationships*. Berlin, Germany: Springer, 2006, pp. 1–44.
[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[27] J. Xiao, H. Cheng, H. Sawhney, and F. Han, "Vehicle detection and tracking in wide field-of-view aerial video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 679–684.
[28] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[29] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, vol. 2. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[30] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[31] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," 2016, *arXiv:1603.00831*. [Online]. Available: https://arxiv.org/abs/1603.00831
[32] R. Kasturi *et al.*, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 319–336, Feb. 2009.
[33] AFRL. (2009). *Wright-Patterson Air Force Base (WPAFB) Dataset*. [Online]. Available: http://sdms.afrl.af.mil/index.php?collection=wpafb2009

Wei Ao (S'17) received the B.E. degree (Hons.) from the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China, in 2016, and the M.S. degree from the School of Information Science and Technology, Fudan University, Shanghai, China, in 2019. He is currently pursuing the Ph.D. with Michigan State University. His research interests include deep learning, image processing, and target recognition.

Yanwei Fu received the Ph.D. degree from the Queen Mary University of London in 2014, and the M.Eng. degree from the Department of Computer Science and Technology, Nanjing University, China, in 2011. He held a postdoctoral position at Disney Research, Pittsburgh, PA, USA, from 2015 to 2016. He is currently a tenure-track Professor with Fudan University. His research interests are image and video understanding and life-long learning.

Xiyue Hou (S'18) received the B.E. degree (Hons.) from the School of Information, East China Normal University, Shanghai, China, in 2018. She is currently pursuing the M.S. degree at the School of Information Science and Technology, Fudan University, Shanghai. Her research interests include deep learning, image processing and target recognition.



Feng Xu (S'06–M'08–SM'14) received the B.E. degree (Hons.) in information engineering from Southeast University, Nanjing, China, in 2003, and the Ph.D. degree (Hons.) in electronic engineering from Fudan University, Shanghai, China, in 2008.

From 2008 to 2010, he was a Postdoctoral Fellow with the NOAA Center for Satellite Application and Research (STAR), Camp Springs, MD. From 2010 to 2013, he was with Intelligent Automation Inc., Rockville, while partly working for NASA Goddard Space Flight Center, Greenbelt, as a Research Scientist. In 2013, he joined Fudan University, where he is currently a Professor with the School of Information Science and Technology. He is the Vice Director of the MoE Key Laboratory for Information Science of Electromagnetic Waves and the Institute of Electromagnetic Big Data and Remote Sensing Intelligence. He has published over 50 articles in peer-reviewed journals and coauthored several books and patents, among many conference articles. His research interests include electromagnetic scattering modeling, SAR information retrieval, and radar system development. He was a recipient of the second-class National Nature Science Award of China in 2011, the 2014 Early Career Award of the IEEE Geoscience and Remote Sensing Society, and the 2007 SUMMA Graduate Fellowship in advanced electromagnetics area. He is the Founding Chair of the IEEE GRSS Shanghai Chapter. He currently serves as an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.