Tracking as a Whole: Multi-Target Tracking by Modeling Group Behavior With Sequential Detection

Yuan Yuan, *Senior Member, IEEE*, Yuwei Lu, and Qi Wang, *Senior Member, IEEE*

***Abstract*— Video-based vehicle detection and tracking is one of the most important components for intelligent transportation systems. When it comes to road junctions, the problem becomes even more difficult due to the occlusions and complex interactions among vehicles. In order to get a precise detection and tracking result, in this paper we propose a novel tracking-by-detection framework. In the detection stage, we present a sequential detection model to deal with serious occlusions. In the tracking stage, we model group behavior to treat complex interactions with overlaps and ambiguities. The main contributions of this paper are twofold: 1) shape prior is exploited in the sequential detection model to tackle occlusions in crowded scene and 2) traffic force is defined in the traffic scene to model group behavior, and it can assist to handle complex interactions among vehicles. We evaluate the proposed approach on real surveillance videos at road junctions and the performance has demonstrated the effectiveness of our method.**

***Index Terms*— Multi-target, vehicle detection, tracking, road junction, sequential detection, group behavior.**

I. Introduction

INTELLIGENT Transportation System (ITS) will be the development direction of the future traffic system. With the popularity of monitoring equipment, more and more traffic videos and images have to be analyzed. Faced with the large amount of data, traditional manual management has become unrealistic and automatic analysis is incrementally necessary as a consequence.

Among the various techniques that enable ITS, detecting and tracking vehicles are fundamentally important. They can help get the information of the primary traffic occupations and infer the quantitative statistics of traffic status. The goal of this work is to automatically detect and track each individual vehicle in the surveillance video of a road junction, where the traffic condition is more complex and achieving an effective control is more essential.

Several challenges render this problem very difficult. First, vehicles have complex dynamics in the field of camera view. Second, occlusion is very serious between vehicles in the real traffic scene. Third, road intersections have much more kinds of objects and complex environment that will lead to ambiguities during detecting vehicles. Under these difficulties, the first important thing is to detect most appeared vehicles and then track their movements. Thus an excellent method is supposed to detect and track targets as many as possible. However, traditional trackers, such as [1], [2], either ignore detection or many of them [3]–[5] annotate the target by hand in the first frame of video sequence. These manual methods are impractical for real traffic videos, because too many targets exit in the field of view and new ones keep emerging. As a result, discriminative tracking methods with online learning [6], [7] are proposed. In such approaches, a specific detector is trained in a semi-supervised fashion and then used to find out the object in continuous frames. However, these algorithms only focus on single target without considering the multi-target situation. Several techniques [8]–[11] are consequently dished to deal with multi-target tracking by optimizing detection assignments over a temporal window. Such approaches apply off-line trained detectors to locate the targets and associate them with their tracks. Although they can overcome several difficulties such as the uncertainty in the number of targets and template drift, they are still inadequate when facing occlusion. Particularly, when tracking a crowd of vehicles in the traffic surveillance video, the data association often fails in the aforementioned methods because of partial occlusions and complex interactions with overlaps and ambiguities. Similar to our approach, some methods use Social Force Model [12] to improve tracking results, e.g. [13], [14]. Our approach is different than [13], [14] in that we model group behavior based on traffic force. The difference between them will be discussed in III-B.1.

In this work, we deal with such difficulties by proposing a sequential detection model, which explores shape prior segmentation, and integrates tracking with group behavior context. While the deformable part-based model (DPM) [15] has outstanding performance in VOC challenges [16], yet it still has poor performance in crowded scenes. Since there are more complex background and targets in actual environment, it is hard to detect all the targets only with one detector. Our sequential detection model utilizes the deformable part-based model whose threshold value (η) is much smaller as a sieve to obtain more candidates of targets. Meanwhile, to decrease the inevitable false detections, a shape prior based segmentation is put forward to refine the results of DPM.

Manuscript received July 4, 2016; revised November 15, 2016 and January 27, 2017; accepted March 15, 2017. Date of publication April 12, 2017; date of current version December 7, 2017. This work was supported by the National Natural Science Foundation of China under Grant 61379094. The Associate Editor for this paper was L. M. Bergasa. *(Corresponding author: Qi Wang.)*

The authors are with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@nwpu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2017.2686871

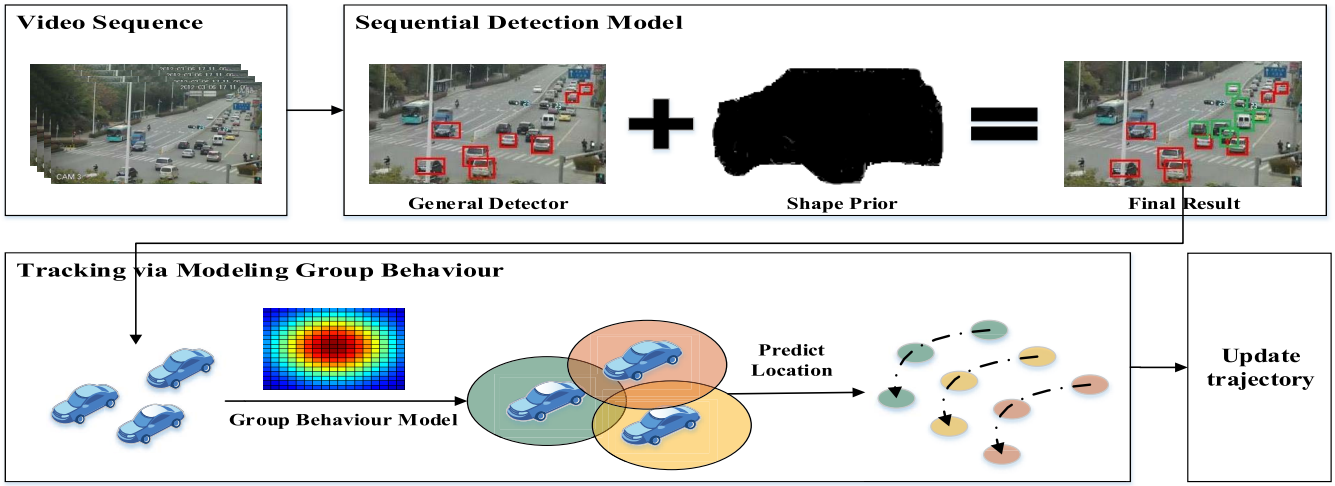


Fig. 1. Overview of the proposed tracking-by-detection framework.

Due to the lower threshold DPM, the partial occlusion can be handled more robustly. However, complex dynamics still bothers us. In the actual environment, each single vehicle is impacted by the others and the speed of one vehicle will be affected by its neighbors. In this case, traditional trackers which track vehicles without taking their neighbors into account will be more likely to drift because they do not capture the intrinsic properties of the vehicle movements. Therefore, we model group behavior that simultaneously considers both the individual vehicle and its neighboring context to improve the performance of the tracker. Especially, our group model acts on vehicles with the same direction. The group behavior model is based on the distance between two vehicles. The closer two vehicles are, the greater repulsive force between them will be. Since the contextual information is added, the group behavior model (GBM) will assist to predict possible locations of target more reasonably.

The proposed tracking-by-detection framework consists of the steps illustrated in Fig. 1. First, a general vehicle detector is used. The detector is supposed to provide candidates of vehicles as many as possible on each frame, no matter whether including false detections or not. In this work, we utilize DPM as an example to implement vehicle detection. Second, the shape prior information is employed to refine the detection results. An energy function is formulated in the graph-cut segmentation by incorporating the vehicle shape prior. Through minimizing the energy function, vehicles can be distinguished from false detections. After that, a GBM-based tracker contributes to predicting locations of the detected vehicles with taking the influences of the neighboring vehicles into account. Owing to modeling group behavior, we can obtain more reasonable vehicle locations. Finally, these predicted locations will be assigned to trajectories and the tracked vehicles will be updated.

In summary, faced with the actual scene of road junctions, we extend traditional approaches so as to adapt to a crowded scene. The main contributions of this paper are as follows. First, we adopt a novel sequential detection model which exploits a shape prior segmentation to tackle occlusion in crowded scene. Second, the complex group behavior in traffic scene is modeled by traffic force to handle influence of the neighboring vehicles.

The rest of this paper is organized as follows. Section II introduces the related work while Section III describes the proposed approach. Experimental results are discussed in Section IV and conclusion and future direction are presented in Section V.

II. Related Work

Object detection and tracking have a long history in computer vision. Much progress has been made in recent years. Since this work is mainly about multi-target detection and tracking, we review existing works in terms of two main categories: vehicle detection and multi-target tracking.

A. Vehicle Detection

Object detection has a very wide range of applications. In this paper, we mainly discuss vehicles in the traffic scene, so we will just review the vehicle detection methods instead of general object detection. Vision-based vehicle detection for traffic surveillance video has received considerable attention. As a rigid target, vehicles have significant structural characteristics, which is more stable than flexible objects. In this paper, we follow the common two steps in vehicle detection [17]: Hypothesis Generation (HG) and Hypothesis Verification (HV).

1) Hypothesis Generation: The goal of HG step is to find candidate vehicle locations in an image quickly for further exploration. HG approaches can be mainly classified into the following three types: knowledge-based, stereo-based and motion-based.

Knowledge-based methods make use of a priori knowledge to hypothesize target locations. Different kinds of priori information are used to identify vehicles. Teoh and Braunl [18] study symmetry. However, symmetry estimations are sensitive to noise in homogeneous areas. Therefore, shadow information is applied in [19] to hypothesize vehicle locations. Inevitably, the intensity of the shadow is influenced by light conditions.

Shadow prior is destined not to have an excellent performance for vehicle detection. Apart from shadow, the structural information is utilized in a large amount. Wu *et al.* [20] use edge detection to find moving vehicles on the road. Moreover, color feature becomes popular after aggregate channel features (ACF) are proposed in [21]. The ACF proposes 10 channel features including 3 LUV color channels, and this feature is outstanding in vehicle detection. Ohn-Bar and Trivedi [22] employ ACF and clustering method to narrow down the search area for detecting vehicles.

Stereo vision-based methods [23]–[25] apply stereo information for vehicle detection in two ways. One is disparity map, while the other is Inverse Perspective Mapping (IPM), an anti-perspective transformation. Lefebvre and Ambellouis [23] convert the disparity map into a 3D map to extract 3D points, while Bertozzi *et al.* [24] employ IPM to acquire stereo vision. Both the disparity and IPM are employed to get contours of targets. With the help of contour information, vehicles can be detected on images.

Motion-based methods employ movement information of vehicles to distinguish vehicles from background. Normally, velocity is the most useful cue to take motion into consideration. Since vehicles keep moving and background is always static, objects and background can be separated according to the difference of velocity. Martinez *et al.* [26] use optical flow method to estimate velocity of each pixel in the image. Afterwards, pixels which have the similar velocity will be clustered together, and these clusters of pixels is the hypothesis of vehicles. However, generating a displacement vector for each pixel is time consuming. In contrast to pixel-based optical flow, feature based methods, such as color features [27], extract features from an image. And then optical flow of feature points will be computed. By clustering optical flow of feature points like pixels-based methods, vehicle hypotheses are generated. Since not all the velocities of pixels need to be estimated, this makes feature based methods faster than pixel-based ones.

2) Hypothesis Verification: Compared with HG step, the input of HV step is the set of hypothesized locations from HG step. For this procedure, tests are performed to verify the correctness of a hypothesis. There are mainly two types of HV approaches: template-based and appearance-based.

Template-based approaches apply predefined patterns from the vehicle class and perform correlation between the image and the template. Li *et al.* [28] propose an And-Or model that integrates context and occlusion for verifying hypotheses. Felzenszwalb *et al.* [15] propose deformable part models (DPM) to structure template model. Each model is composed of different parts of the object. The system detects objects by scoring each hypothesis according to the similarity between hypothesis and DPM model and thresholding scores. León and Hirata [29] put forward a template-based approach using mixture of deformable parts models. They expand the original DPM [15] to adapt to crowded scenes. Wang *et al.* [30] also propose a probabilistic inference framework based on part models for improving detection performance.

Appearance-based approaches learn the features of the vehicles from a set of training images which should capture the variability in vehicle appearance. Usually, appearance models treat a two-class pattern classification problem: vehicle and nonvehicle. Wu and Zhang [31] apply standard Principal Components Analysis (PCA) for extracting global features to detect vehicles. Owing to small training data set, it is difficult to draw any meaningful conclusions. Li *et al.* [32] employ segmentation and neural network classifier for distinguishing vehicles from background. Khammari *et al.* [33] add depth image to set up their appearance models. Apart from the observed features, Zheng and Liang [34] design image strip features based on the vehicle structure for vehicle detection. Since features come from the side view of the vehicle, their detector is sensitive to the viewpoint.

B. Multi-Target Tracking

A significant amount of work has been reported for multi-target tracking. There are two main representative approaches, detection-based data association and energy minimization.

1) Detection-Based Data Association: Detection-based data association regards multi-target tracking as a data association problem. Longer tracklets can be formed by detections between two continuous frames. The most classic framework of this approach is proposed by Huang *et al.* [35]. They handle data association in three levels. In the low-level, they connect detection responses in continuous frames into short tracklets. A threshold value would be used to exclude unreliable ones. In the mid-level, they compute an affinity score for each reliable tracklet obtained from low-level and connect short tracklets into longer tracklets. In the high-level, a scene structure model is estimated based on the tracklets provided by the middle level. Afterward, with the help of scene knowledge, the long-range trajectory association is performed.

Some work follows this basic framework. Zhang *et al.* [10] define data association as a maximum-a-posteriori problem given a set of object detection results as input observations, while Brendel *et al.* [36] formulate data association problem as finding the maximum-weight independent set of graph that is built by pairs of detection responses from every two consecutive frames.

2) Tracking via Energy Minimization: Many problems can be transformed into an energy minimization problem. This is true for multi-target tracking. In recent years, several energy minimization-based tracking methods [8], [37] are proposed. In these methods, detection responses are known and solution space is the combination of tracklets that are composed of these responses, which is different from common data association methods whose current frame is inferred by previous ones. Milan *et al.* [8] construct an energy function that depends on the locations and motions of all targets in all frames for obtaining globally optimal solution considering physical constraints, such as target dynamics. By minimizing the energy function, they get the final tracking result. Leibe *et al.* [37] present a multi-object tracking approach which considers object detection and space-time trajectory estimation as an optimization problem. And the successful trajectory hypotheses are fed back to guide detection in the future frames. Since minimizing energy function is time-consuming,

energy minimization-based methods for multi-object tracking suffer from low computational efficiency.

III. Our Approach

In this section, we will give a detailed explanation of our tracking-by-detection framework. As mentioned before, in the surveillance video of road intersections, occlusions and complex interactions with overlapping and ambiguities are the main difficulties. Hence, in the detection stage, we present a sequential detection model that explores shape prior segmentation to improve the detector in crowded scene. In the tracking stage, on the other hand, traffic force is proposed to model group behavior. Interactions between individual vehicles are taken into consideration to tackle nonlinear dynamics in vehicle tracking.

A. Sequential Detection Model

Though object detection has made much progress, existing detection approaches are still not well tailored to crowded scenes. Our motivation comes from boosting algorithm, in which the single classifier does not work well but combining several weak classifiers to a cascade classifier can achieve an outstanding performance. In the same way, a single vehicle detection algorithm cannot find out all the targets. Therefore, we combine several basic algorithms having distinctive superiorities to produce a sequential detector. The Sequential Detection Model consists of two main parts, DPM and shape prior segmentation.

1) DPM: As shown in Fig. 1, a general detector is utilized in our framework to get enough possible locations of targets. The reason why we choose DPM are mainly as follows. First, DPM can easily get shape templates with various viewpoints and describe the target with abundant information. Second, DPM is well suited for occlusions that are serious in the scene of road intersections.

In DPM, the detection score of a hypothesis, $score(h_{obj})$, is given by the filter score at the examined location minus a deformation cost that depends on the relative position of each part with respect to the root filter plus the bias, as is shown in Eq.1:

$$score(h_{obj}) = \sum_{i=0}^{n} F_i - \sum_{i=1}^{n} D_i + b, \tag{1}$$

where b is a bias term, n is the number of parts, D_i is the deformation cost of part i, F_i is the score of each part and F_0 represents the root part.

A score represents the similarity between pre-trained model and a hypothesis. DPM get final detections by thresholding score. However, different from the original DPM, we set a low threshold value instead of self-generated one for DPM in detection procedure. As a result, reducing threshold can obtain vehicle candidates as many as possible. Due to the significant low threshold value, the results of DPM have both vehicle and nonvehicle targets. Though we improve the recall rate of detecting vehicles, more false detections appear inevitably. In order to exclude the false detections, a shape prior segmentation is further applied.

2) Shape Prior Segmentation: As is implied by the name, the shape prior segmentation takes shape information into consideration so as to exclude the false detections. Each detection window obtained from the DPM is processed independently.

Image segmentation can be regarded as a pixel labeling problem actually. The label of the pixel depends on whether it is in object or background and this process can be achieved by minimizing the energy function through minimum graph cut. Let $L = \{l_1, l_2, \ldots, l_i, \ldots, l_m\}$ be the label set of pixels, where l_i is the label of the pixel i in the image. The pixel is assigned label $l_i = 1$ if it belongs to object and $l_i = 0$ if it belongs to background. The energy function for the shape prior based graph cut is usually defined as the following equation [38], [39]:

$$E(L) = R(L) + B(L) + E_{shape}, \tag{2}$$

where, $R(L)$ is the regional term, $B(L)$ is the boundary term and E_{shape} is the shape prior term. Compared to traditional graph cut based segmentation, the shape prior term is added. The goal of shape prior term is to segment targets with similar shape to the template.

In sequential detection model of our framework, shape prior segmentation is applied to distinguish vehicles from other targets. Our shape prior segmentation is just like a refinement. It can remove those false detections and extract vehicles. We define the energy function of graph cut segmentation with shape prior in the following way:

$$E(L) = \sum_{p \in P} D_p(l_p) + \sum_{(i,j) \in N_p : l_i \neq l_j} V_{i,j}(l_i, l_j) + \sum_{(i,j) \in N_p : l_i \neq l_j} E_{i,j}(l_i, l_j), \tag{3}$$

where P is the set of all pixels and N_p is the set of pixels in the neighborhood of p. $D_p(l_p)$ is the penalty of assigning label $l_p \in L$ to p, and $V_{i,j}(l_i, l_j)$ is the penalty of labelling the pair i, j with labels $l_i, l_j \in L$, respectively. $E_{i,j}(l_i, l_j)$ represents a pairwise shape constraint term that penalizes the difference between the shape template and the target.

To be specific, the region term $D_p(l_p)$ is:

$$D_p(l_p = 1) = -log P_r(I_p | obj), \tag{4}$$

$$D_p(l_p = 0) = -log P_r(I_p | back), \tag{5}$$

where $P_r(I_p|obj)$ and $P_r(I_p|back)$ are the probability distributions that can be learned beforehand for both the object and the background, and I represents the pixel intensity.

And the edge term which punishes those pixels with similar intensities is defined as:

$$V_{i,j}(l_i, l_j) = \exp(-\frac{(I_i - I_j)^2}{2\alpha^2}) \frac{1}{dis(i, j)}, \tag{6}$$

where α can be seen as camera noise, and $dis(i, j)$ is the Euclidean distance between pixels i and j.

When applying adaptive shape prior, we employ the idea of level-set template, and define shape energy term $E_{i,j}(l_i, l_j)$ in Eq.(3) in the following way:

$$E_{i,j}(l_i, l_j) = \phi(\frac{pos_i + pos_j}{2}), \tag{7}$$

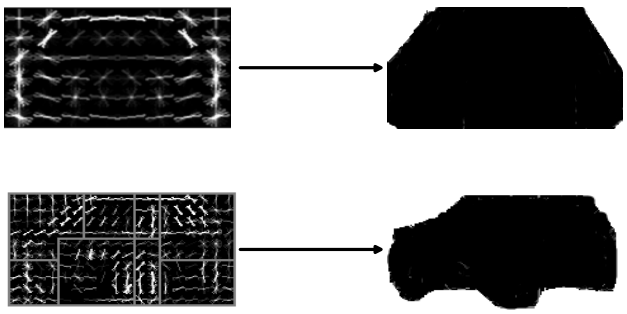


Fig. 2. Examplar transformation from HOG template to our binary template.

where ϕ is a regular, unsigned distance function whose zero level set corresponds to the shape template. pos_i and pos_j are the locations of pixels. $\phi(pos_i)$ will be zero if $l_i = 1$, and $\phi(pos_i)$ will be the shortest distance to the shape boundary if $l_i = 0$. Since they are neighboring pixels, they will be along the contour of shape, when minimizing $E_{i,j}$.

As for the definition of shape template, we employ binary figures of the vehicles. Since we incorporate shape prior by zero level set, a binary figure will speed up the computation. For this purpose, the shape templates are generally supposed to be trained by samples. Fortunately, DPM has trained multi-parts from various viewpoints of the vehicle. In this work, we adopt vehicle models trained by DPM and transform them into binary templates. These models are trained by VOC-2007 [16] dataset and our own data. As a result, when DPM detects a vehicle by one of the parts, we can utilize the corresponding binary figure as the shape template. Namely, different shapes of the vehicle are applied depending on the maximum response part used in the DPM detection stage. Fig. 2 shows some examples of our shape templates.

By minimizing the energy function of Eq. 3, false detections are removed and vehicles will be detected finally.

B. Tracking via Group Behavior Model

After obtaining the final detections, we employ these detection windows as our observation targets for tracking. As mentioned before, occlusions and complex motions will be the challenges. Owing to part-based model, we can deal with occlusions. However, complex motion model still troubles us.

Under the dynamics, it is hard for us to treat the motion of a vehicle separately. Each vehicle is affected by its surroundings and regarding the vehicles as a group in the scene of road intersections is more reasonable. In view of the above fact, we attempt to model the group behavior of vehicles to assist tracking procedure. When tracking a vehicle in such a group, we should consider not only the state of the vehicle itself, but also the influence of other vehicles in the group.

1) Group Behavior Model: We first explain the difference between the proposed tracking model and existing ones. Fig. 3(a) represents the traditional state-space model employed for the generic object tracking. Formally, this state-space model is defined as follows:

$$\begin{aligned} \theta_{t+1} &= f_t(\theta_t), \\ X_t &= g_t(\theta_t), \end{aligned} \tag{8}$$

where θ_t is the state of the target at time t, and X_t is the observation. Thus, current state is only determined by the previous state, and f_t and g_t are nonlinear unknown functions. Fig. 3(b) graphically describes the link from θ_t to θ_{t+1} and the link from θ_t to X_t in Eq. 8. Along this line of consideration, we model group behavior that includes surrounding information for tracking. In our group behavior model, the movement of each target is influenced by the neighboring ones. Fig. 3(b) illustrates our group behavior model and in this case Eq. 8 can be rewritten as follows:

$$\begin{aligned} \theta_{t+1} &= f_t(\theta_t, S_{t+1}), \\ S_t &= (P_t, V_t), \\ X_t &= g_t(\theta_t), \end{aligned} \tag{9}$$

where S_t is the surrounding information, including the location P_t and velocity V_t information.

As we all know, vehicles have to keep the minimum safe distance between each other according to the traffic regulations. In other words, when two vehicles become too close, the back one will have the tendency that it will keep away from the front one. We assume that a potential repulsive force exists among vehicles in this situation and this potential repulsive force is named as traffic force (TF). The traffic force makes each individual in the traffic scene hold a minimum distance from others and avoid collision. We regard this behavior caused by the traffic force as GBM. By building the GBM, we try to simulate the behaviors among vehicles and improve our tracker.

The traffic force between individuals is inversely related to their distance. If the distance decreases, this force increases. With this in mind, the distance between the predicted locations of targets can be used to calculate TF. Let $Tf_t = [tf_t^1, tf_t^2, \ldots, tf_t^i, \ldots]$ be the vector of traffic force. We use tf_t^i to represent the TF of the i_{th} target from its neighbors. And the surrounding information $s_t^{ij} \in S_t$ is the j_{th} target around the target i. The overall force is defined as:

$$tf_t^i = \sum_{i \neq j} \mu_{ij} w(s_t^{ij}), \tag{10}$$

where $w(s_t^{ij})$ is the force between the target i and its neighboring target j. Each TF between the two vehicles is computed as:

$$w(s_t^{ij}) = \exp\left(\frac{-d_{ij}^2(t)}{2\sigma_d^2}\right), \tag{11}$$

where σ_d controls the distances of a vehicle to be avoided, and d_{ij} is the Euclidean distance between the two targets. We assume that target i has the predicted position $prep_t^i$ and its neighboring target j has the predicted position $prep_t^j$:

$$d_{ij}(t) = \left\| prep_t^i - prep_t^j \right\|, \tag{12}$$

Moreover, μ_{ij} in Eq. (10) represents the influence of target j in the overall TF on target i. It guarantees that different vehicles will have different influences. The definition of μ_{ij} is:

$$\mu_{ij} = \exp\left(\frac{-\left\| prep_{t-1}^i - prep_{t-1}^j \right\|^2}{2\sigma_w^2}\right), \tag{13}$$

where σ_w is the radius of targets influence.

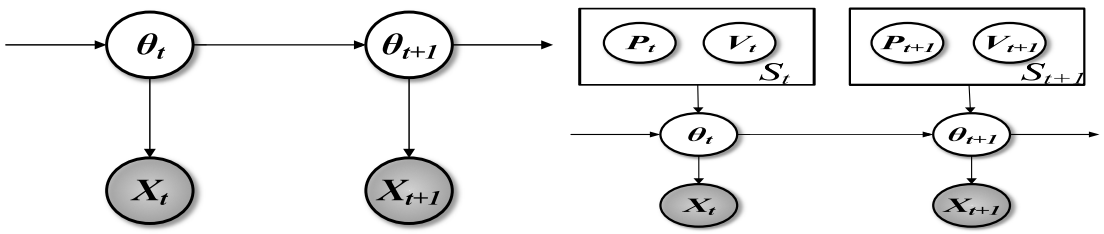


Fig. 3. Group Behavior Model. (a) The state-space model of traditional tracking. (b) Our approach for modeling group behavior.

By utilizing the traffic force Tf_t, we can take group behavior cue into consideration for tracking multiple vehicles. However, GBM here is different from social force model (SFM) [12] in theory. There are three main differences between GBM and SFM:

- GBM is calculated based on the whole object, while SFM is calculated based on pixels.
- GBM is defined by the distance, while SFM is based on the velocity.
- GBM cares more about the movement of the individual target affected by the whole group. SFM, on the other hand, focuses on the movement tendency of the whole group.

2) GBM-Based Tracking: Our tracking model utilizes GBM in the Kalman filter [40] to predict the locations of vehicles. For the state prediction, our constraint of group behavior model is applied in the following form:

$$\theta_t = Tf_t \cdot F_t \cdot \theta_{t-1} + B_t \cdot u_t, \tag{14}$$

where F_t is the process transition matrix, u_t is the control vector, and B_t converts control vector u_t to state space. The state θ_t will be predicted by θ_{t-1} under the constraint of Tf_t. Recalling the definition of $w(s_t^{ij})$ in III-B.1, it is easy to know that $0 < w < 1$. The physical significance of w can be seen as decelerating vehicles that are affected in the group. And this constraint is reasonable due to the traffic regulations, in which to keep the minimum safe distance between vehicles and avoid collisions, vehicles have to slow down. From this perspective, the modification to the Kalman filter is promising. After obtaining predicted locations, these locations and new detected ones will be assigned to existing trajectories. We assign locations to trajectories by judging the motion tendency of a vehicle. If one location is consistent with the motion tendency and it is near the trajectory, the location will be allocated to the trajectory and the trajectory will be updated. A simple greedy algorithm will be employed to assign locations to trajectories. If one location doesn't belong to any trajectory. It will be the start of a new trajectory.

The reasons why the Kalman is chosen are as follows: First, Kalman filter has less computation price than other methods. Second, vehicle velocity will not change violently in this traffic context. Because of traffic light, most vehicles have similar velocities. Thus, velocity has a Gaussian distribution that is the necessary condition for Kalman filter. Therefore, Kalman filter is suitable for this situation.

The whole tracking-by-detection procedure is summarized in Algorithm 1.

Algorithm 1 GBM-Based Tracking Algorithm

procedure GBMTRACKER($videoseq$)
1. Initialize tracker and detector
2. Get each $frame$ from $videoseq$
3. Obtain possible locations as many as possible via DPM with $\eta = -0.78$
4. Extract vehicles from possible locations via shape segmentation Eq. 3-Eq. 7
5. Modeling group behavior via detection information Eq. 9-Eq. 13
6. Predict locations Eq. 14
7. **if** *new location belongs to existing trajectory*
   update assigned tracks
   else
   generate new tracks
   end if
8. Display result

end procedure

IV. Experiment and Discussion

To demonstrate the capabilities of the presented approach, extensive experiments are conducted and evaluated. In this section, the experiments will be introduced from the following aspects: data set, evaluation measure, parameter selection, experimental results and analysis.

A. Data Set

Multiple object tracking has many public data sets. However, there are no surveillance videos in the road junctions. For this reason, our experiments are performed on videos that we collected. The data set contains one short video (Seq1) and two long videos (Seq2 and Seq3). Seq1 involves 748 frames (688×384) and 25 frames per second. It describes only two opposite directions. Seq2 includes 6200 frames (1280×720), while Seq3 includes 7908 frames (1280×720). Both of them are 30 frames per second and describes all the possible directions in road junctions.

B. Evaluation Measure

Diverse evaluation measures are employed for different stages in our approach.

1) Evaluating Detection: Precision-recall measure is adopted to evaluate the detection performance. Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. For classification tasks, the terms true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) compare the classifier results under test with trusted external judgements. Based on these definitions, the two metrics are calculated as:

$$precision = \frac{TP}{TP + FP}, \tag{15}$$

$$recall = \frac{TP}{TP + FN}, \tag{16}$$

In our vehicle detection task, TP is the number of vehicles that are correctly detected on all frames. FP is the ones that are incorrectly detected as positives and FN is the ones that are not detected but should have been detected.

2) Evaluating Tracking: We evaluate our tracking results using the standard CLEAR MOT metrics [41]. The indexes of CLEAR MOT metrics are MOTP (multiple object tracking precision) and MOTA (multiple object tracking accuracy). MOTP shows the ability of the tracker to estimate precise object positions, independent of its skill at recognizing object configurations, keeping consistent trajectories, and so forth. Briefly, MOTP embodies the precision of the target locations. It is defined as:

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}, \tag{17}$$

where $\sum_{i,t} d_t^i$ is the total error in estimated position for matched object-hypothesis pairs over all frames, and $\sum_t c_t$ is the total number of matches made.

Moreover, MOTA accounts for all object configuration errors made by the tracker, false positives, misses, mismatches, over all frames. Compared with MOTP, MOTA cares more about the accuracy of the number of targets. MOTA can be seen as derived from three error ratios:

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}, \tag{18}$$

where $\sum_t m_t / \sum_t g_t$ is the ratio of misses in the sequence, $\sum_t fp_t / \sum_t g_t$ is the ratio of false positives and $\sum_t mme_t / \sum_t g_t$ is the ratio of mismatches. We count all tracker hypotheses for which no real object exists as false positives and count all occurrences where the tracking hypothesis for an object changed compared to previous frames as mismatch errors. All of them are computed over the total number of objects present in all frames.

C. Parameter Selection

In our framework, some parameters play an important role in the experiments. In order to obtain a significant performance, the value of those parameters should be selected meticulously.

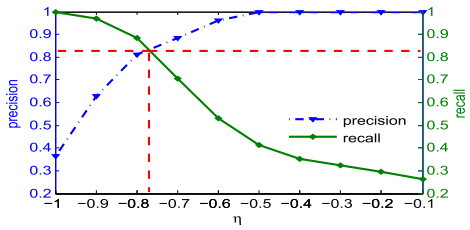


Fig. 4. Effect of varied η on the precision and recall results.

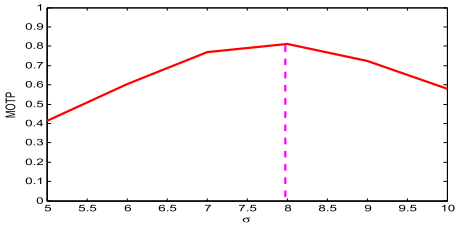


Fig. 5. Effect of varied σ_w on the MOTP results.

1) Threshold η in Detection: As mentioned before, DPM in our framework is used to obtain vehicle hypotheses as many as possible. To this end, we are supposed to choose a much smaller threshold η that is used to determine whether the pixel area is object or not. However, the default η equals to -0.5. This threshold cannot provide enough possible hypothesis in road junctions. Therefore, we select another proper threshold through experiment.

Fig. 4 displays how η is selected. As we all know that the precision is inversely related to the recall. It is impossible to make precision and recall the biggest simultaneously. In Fig. 4, we can see that when η is smaller than -0.5, recall of our detector increases significantly, but precision decreases slowly. With the recall improving, false detections should have increased and the precision should have decreased greatly. However, the reason why the precision decreases slowly is that our shape prior segmentation that is added into the sequential detection model prevent it from being in steep decline. According to the shape difference between vehicles and other target, shape prior segmentation takes out many false detections which is produced by a significant low threshold value η. Proper exploring the shape prior segmentation makes us seek out a point that has relatively high precision and recall at the same time. Therefore, we finally choose $\eta = -0.78$.

2) Influence Radius σ_w of Each Target: In Section III-B, each vehicle in our tracking model has its own influence radius. In theory, each vehicle will be affected by its neighboring ones from all directions. However, due to the limitation of viewpoint, the influence from various direction seems different. Given this actual situation, our influence radius of vehicles is selected by experiment.

Fig. 5 shows our experimental curve of selecting σ_w. With the increase of σ_w, the MOTP of our tracker improves first but decreases after the peak value. We attempt to explain the result through the phenomenon of the realistic traffic scene: When the influence area of a vehicle is too small, the relationship between vehicles become weak. Our tracker will degenerate

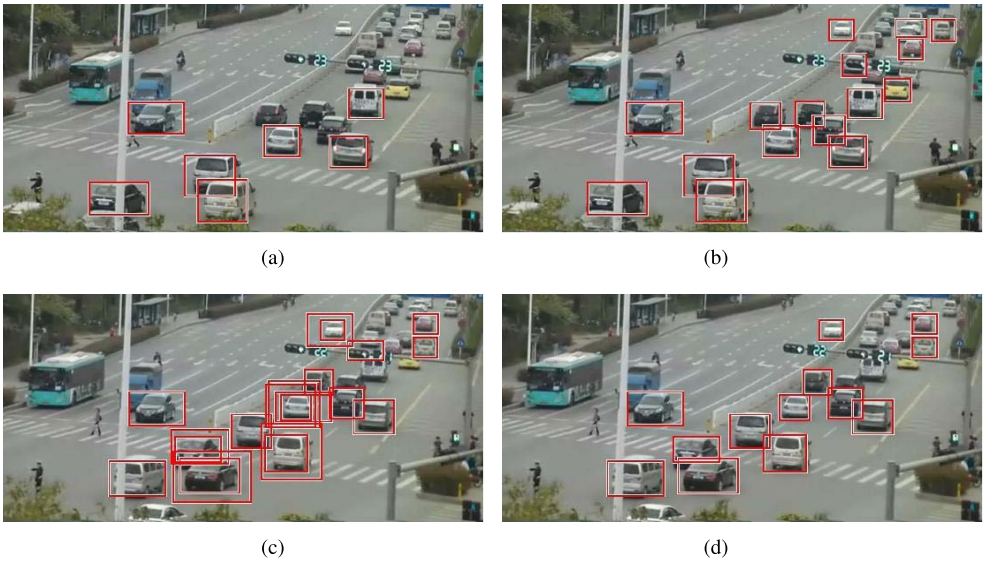


Fig. 6. The first column is the detection result with original DPM detector, while the second column shows the result of the proposed sequential detector with shape prior. (a) DPM detector. (b) Detecting with shape prior. (c) DPM detector. (d) Detecting with shape prior.

into a conventional tracker that only consider the vehicle itself. This will lead to a poor performance. On the contrary, if a vehicle influences much large areas, more vehicles will be affected. That is to say, not only the neighboring ones are affected, the others are also influenced. This will be contradictory to our assumption that vehicles can only affect their neighboring ones. The performance is also unsatisfying. Hence, we finally select $\sigma_w = 8.0$ for our group behavior model.

D. Result Exhibition

With the parameters presented at section IV-C, a more detailed analysis of our experiment will be presented in the following sections respectively.

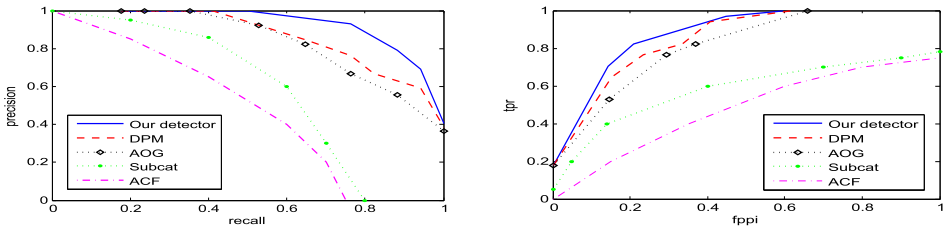
1) Sequential Detection With Shape Prior: Employing shape prior segmentation in detection makes our detector unique. Fig. 6 reveals our experiment results of sequential detection. Fig. 6(a) employs original DPM with its common threshold $\eta = -0.5$, while Fig. 6(b), Fig. 6(c) and Fig. 6(d) utilize our detector that has a lower threshold $\eta = -0.78$ with shape information of vehicles. When not applying shape information of vehicles, we will get fewer targets in Fig. 6(a) or more overlapping bounding boxes in Fig. 6(c). After making use of shape prior, Fig. 6(b) demonstrates we can obtain more targets, while Fig. 6(d) shows another superiority of shape prior that can remove false detections.

The advantages of our detecting method are as follows: First, we use a significant low threshold value. In original DPM, this low threshold value will lead to lots of false detections and duplicate bounding boxes. However, we don't have this trouble. We just regard the output of DPM as proposals, while they are final detection results in original DPM. With this low threshold value, we can get candidates of vehicles as many as possible without considering the false detection. Second, shape prior is added to the segmentation. The significant low threshold value allows us to detect more targets in theory, while the shape prior segmentation guarantees the feasibility of detecting more targets in practice. We employ this shape prior segmentation to deal with the proposals obtained by DPM. Owing to applying shape information of the vehicle, we can easily distinguish correct vehicles from false detections.

We compare our detector with DPM [15], AOG [28] and ACF-based detectors (Subcat [22] and ACF [21]) . Both DPM and AOG deal with occlusion very well, while ACF (Aggregate Channel Features) is popular in recent years because of its speediness and high efficiency. Subcat make vehicle detection much faster, due to classifying vehicles into subcategories. Fig. 7(a) and Fig. 7(b) shows the result of comparison. ROC (receiver operating characteristic) is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. FPPI represents false positives per image and TPR (true positive rate) has the same value with recall in Eq. 16.

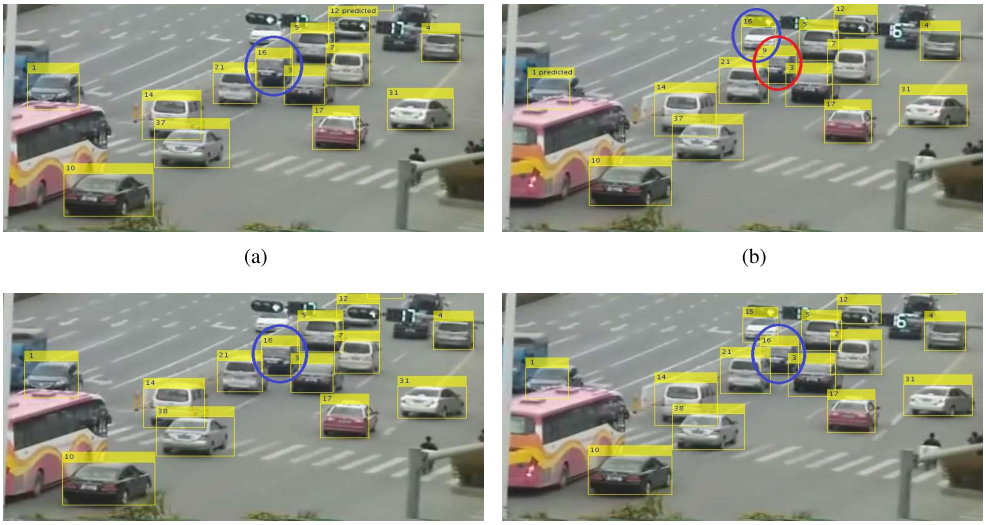
Nevertheless, different from results in public data set, ACF-based detectors [21], [22] have a poor performance in road intersections. The reason is that ACF detectors doesn't specially aim at occlusion that is quite serious in our traffic surveillance video. On the other hand, AOG and DPM has similar performance taking occlusions into account. And our detector, which divide detection problem into detecting and segmentation, own the best performance in such a real scene, since we have more candidates and make full use of shape information of the vehicle. The result also proves our thought that aggregating several basic techniques can reach a significant performance.

Despite that the performance of our sequential detection is outstanding. There are still some problems to be discussed. One of the controversial problems is that whether we should detect all the vehicles in a frame or not. That is to say, for those distant targets whose size is quite small, is it necessary to detect them all? Early in our research, we tried to detect



(a) (b)

Fig. 7. Comparison between five detection methods, including our sequential detector, DPM detector, AOG detector, original ACF detector and Subcat detector. (a) Comparison with ROC curve. (b) Comparison with FPPI curve.



(a) (b)

(c) (d)

Fig. 8. The first row is the tracking result without modeling social behavior, while the second row shows the result with GBM. Different number of the bounding box indicates the different vehicles. (a) frame#161. (b) frame#171. (c) frame#161. (d) frame#171.

all the vehicles regardless of their size. Unfortunately, it is really difficult to detect all the vehicle, especially for tiny ones. They are even hard to identify the structure. Therefore, we only focus on those vehicles near the intersection. When those distant vehicles come to near, we can also detect them. On the other hand, the traffic junction is an accident black spot. Detecting those distant vehicles seems to be meaningless. Moreover, those distant vehicles may be near ones compared to other camera. In consequence, we only detect those near target in this work. It is sufficient for surveillance video of road junction. Another controversial problem is that since the templates of shape prior segmentation are from the training of DPM, will the results from DPM and shape prior segmentation be correlated and detection errors be reinforced? Though the DPM models are correlated to the templates of shape prior segmentation, the detection errors cannot be reinforced. DPM models are more concerned about the relationship between the various parts of the object, however, shape templates only care about the similarity of the whole object. They influence detection results in different aspects. Fig. 6(c) and Fig. 6(d) also prove this point of view. The traffic light detected in Fig. 6(c) is removed by shape prior segmentation in Fig. 6(d).

2) GBM-Based Tracking: Modeling group behavior of vehicles contributes to tracking in road junctions and taking interplay of targets into consideration makes our result more reasonable. Fig. 8 reveals one of the advantages of modeling group behavior, which can relieve the drift of targets. In Fig. 8(a) and Fig. 8(b), the tracker regards bounding boxes in blue circle as the same vehicle. Obviously, they are different targets and the vehicle's number in the red circle changes from 16 to 9. That means drift happens during the tracking course. After modeling group behavior in our tracking model, we no longer track a vehicle separately. Each vehicle's state is influenced by its surroundings which is relatively stable in the actual situation of real world. Therefore, in Fig. 8(c) and Fig. 8(d), we can see that the vehicle in blue circle keep the same number. In our GBM, when interaction occurs between vehicles, GBM restricts each target in the group to follow its trajectory. Since each vehicle is restrained by its surroundings, drifting becomes difficult.

By the same token, another advantage of GBM is that our tracker is better suited to complex movement of vehicles. Interplay between vehicles can be ignored in some simple scene. But for the traffic scene, it is impossible. When a vehicle crosses the view of camera, the movement of its surroundings will also be taken into account. Thus, we treat these vehicles as a group. They can not only affect each other, but also have their own trajectory.

TABLE I
THE MOTP IN DIFFERENT TEST SEQUENCES

Tracking Algorithm	Seq1	Seq2	Seq3
Our approach	80.6 %	82.1 %	81.2 %
CEM [8]	72.8 %	76.0 %	76.8 %
DCO [42]	72.4 %	74.3 %	74.7 %
Baseline(Kalman Filter)	59.8 %	62.4 %	62.0%

TABLE II
THE MOTA IN DIFFERENT TEST SEQUENCES

Tracking Algorithm	Seq1	Seq2	Seq3
Our approach	63.8 %	65.1 %	64.6 %
CEM [8]	60.5 %	62.1 %	61.6 %
DCO [42]	58.7 %	60.2 %	59.9 %
Baseline(Kalman Filter)	48.6 %	50.1 %	50.7 %

TABLE III
COMPARISON OF TRACKING PERFORMANCE WITH MOTA AND MOTP

Tracking Algorithm	MOTA	MOTP
Our approach	64.5 %	81.3 %
CEM [8]	61.4 %	75.2 %
DCO [42]	59.6 %	73.8 %
Baseline(Kalman Filter)	49.8 %	61.4 %

Since methods of multi-target tracking by minimizing energy function are popular in recent years, we compare our result with some typical methods CEM [8] and DCO [42]. The baseline tracking results are from Kalman Filter(without the traffic force). CEM focus on designing an energy that corresponds to a more complete representation of the problem, rather than one that is amenable to global optimization. It takes into account physical constraints, such as target dynamics, mutual exclusion, and track persistence. In addition, partial image evidence is handled with explicit occlusion reasoning, and different targets are disambiguated with an appearance model in CEM. On the other hand, DCO proposes a discrete-continuous optimization method for minimizing energy function. In DCO, data association is performed using discrete optimization with label costs, yielding near optimality. And trajectory estimation is posed as a continuous fitting problem with a simple closed-form solution, which is used in turn to update the label costs. Due to not need to pre-compute trajectories, the accuracy of estimating trajectories improves.

Although CEM and DCO have great performance in some public data sets, they still handle targets individually. The relationship between targets is ignored. Compared with these trackers, our GBM model group behavior to handle interactions among targets. GBM will help to predict locations more accurately in traffic video due to considering both vehicles and their surroundings. TABLE III shows the final results of the experiments, while TABLE I and TABLE II exhibit the MOTP and MOTA in three test sequences, respectively. In GBM, each target belongs to a group. It will be affected by group members. The influence between vehicles prevents them from drifting, and it makes targets follow regular motion model. However, both CEM and DCO haven't applied group information. They just tracking vehicles individually. Therefore, our MOTP value is obviously higher. We have to admit that our MOTA is not superior. Since there are many targets in traffic video sequence, improving accuracy without any other modifying of tracker is difficult. This problem maybe remit in our future work.

V. CONCLUSION AND FUTURE WORK

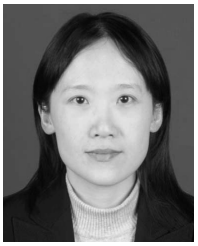
In summary, a novel tracking-by-detection framework is proposed in this paper. Our approach captures rich information about road junctions, such as vehicle shape and motion priors. As a consequence, the proposed approach has higher efficiency than traditional tracking algorithms in crowded scenes. Though our approach is tested in road intersections, by applying pedestrian detectors the proposed method is also suitable for other crowded situations, such as supermarket and subway station. The main contributions of this work are as follows. First, we exploit shape prior in the sequential detection model to tackle occlusions in crowded scene. Second, Traffic force is defined to model group behavior in the traffic scene. With GBM, we can handle the influence of neighboring vehicles and obtain more precise localizations. The proposed framework is evaluated on real traffic videos and has shown its significant performance through intensive comparisons and analyses.

However, the proposed tracking-by-detection framework still can be improved. A faster vehicle detector particularly designed for the traffic scene is expected in the future work. Besides, we also plan to judge whether the vehicles violate the traffic rules on the basis of this work in the future.

REFERENCES

[1] M. D. Jenkins, P. Barrie, T. Buggy, and G. Morison, "Selective sampling importance resampling particle filter tracking with multi-bag subspace restoration," *IEEE Trans. Cybern.* [Online]. Available: http://ieeexplore.ieee.org/document/7778232/
[2] Q. Wang, J. Fang, and Y. Yuan, "Multi-cue based tracking," *Neurocomputing*, vol. 131, pp. 227–236, May 2014.
[3] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2259–2272, Nov. 2011.
[4] L. Wang, H. Yan, H.-Y. Wu, and C. Pan, "Forward–backward mean-shift for visual tracking with local-background-weighted histogram," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1480–1489, Sep. 2013.
[5] Y. Yuan, J. Fang, and Q. Wang, "Robust superpixel tracking via depth fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 15–26, Jan. 2014.
[6] S. Avidan, "Ensemble tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 261–271, Feb. 2007.
[7] S. Sivaraman and M. M. Trivedi, "A general active-learning framework for on-road vehicle recognition and tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 267–276, Jun. 2010.
[8] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 58–72, Jan. 2014.
[9] Q. Gu, T. Tang, and F. Ma, "Energy-efficient train tracking operation based on multiple optimization models," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 3, pp. 882–892, Mar. 2016.
[10] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[11] B. Wang, G. Wang, K. L. Chan, and L. Wang, "Tracklet association by online target-specific metric learning and coherent dynamics estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 589–602, Mar. 2016.

[12] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 935–942.
[13] Z. Qin and C. R. Shelton, "Improving multi-target tracking via social grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1972–1978.
[14] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 261–268.
[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[16] M. Everingham *et al. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*, accessed on Apr. 07, 2007. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
[17] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 694–711, May 2006.
[18] S. S. Teoh and T. Braunl, "Symmetry-based monocular vehicle detection system," *Mach. Vis. Appl.*, vol. 23, no. 5, pp. 831–842, 2012.
[19] Y.-M. Chan, S.-S. Huang, L.-C. Fu, and P.-Y. Hsiao, "Vehicle detection under various lighting conditions by incorporating particle filter," in *Proc. IEEE Intell. Transp. Syst. Conf.*, Sep. 2007, pp. 534–539.
[20] B.-F. Wu, C.-C. Kao, J.-H. Juang, and Y.-S. Huang, "A new approach to video-based traffic surveillance using fuzzy hybrid information inference mechanism," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 485–491, Mar. 2013.
[21] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.
[22] E. Ohn-Bar and M. M. Trivedi, "Learning to detect vehicles by clustering appearance patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2511–2521, Oct. 2015.
[23] S. Lefebvre and S. Ambellouis, "Vehicle detection and tracking using Mean Shift segmentation on semi-dense disparity maps," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2012, pp. 855–860.
[24] M. Bertozzi, L. Bombini, P. Cerri, P. Medici, P. C. Antonello, and M. Miglietta, "Obstacle detection and classification fusing radar and vision," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2008, pp. 608–613.
[25] Q. Wang, J. Fang, and Y. Yuan, "Adaptive road detection via context-aware label transfer," *Neurocomputing*, vol. 158, pp. 174–183, Jun. 2015.
[26] E. Martinez, M. Diaz, J. Melenchon, J. A. Montero, I. Iriondo, and J. C. Socoro, "Driving assistance system based on the detection of head-on collisions," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2008, pp. 913–918.
[27] P. Dave, N. M. Gella, N. Saboo, and A. Das, "A novel algorithm for night time vehicle detection even with one non-functional taillight by CIOF (color inherited optical flow)," in *Proc. Int. Conf. Pattern Recognit. Syst.*, Apr. 2016, pp. 1–6.
[28] B. Li, T. Wu, and S.-C. Zhu, "Integrating context and occlusion for car detection by hierarchical and-or model," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 652–667.
[29] L. C. León and R. Hirata, Jr., "Vehicle detection using mixture of deformable parts models: Static and dynamic camera," in *Proc. SIBGRAPI Conf. Graph., Patterns Images*, Aug. 2012, pp. 237–244.
[30] C. Wang, Y. Fang, H. Zhao, C. Guo, S. Mita, and H. Zha, "Probabilistic inference for occluded and multiview on-road vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 215–229, Jan. 2016.
[31] J. Wu and X. Zhang, "A PCA classifier and its application in vehicle detection," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2001, pp. 600–604.
[32] X. Li, X. Yao, Y. L. Murphey, R. Karlsen, and G. Gerhart, "A real-time vehicle detection and tracking system in outdoor traffic scenes," in *Proc. 17th Int. Conf. Pattern Recognit.*, Aug. 2004, pp. 761–764.
[33] A. Khammari, F. Nashashibi, Y. Abramson, and C. Laurgeau, "Vehicle detection combining gradient analysis and adaboost classification," in *Proc. IEEE Intell. Transp. Syst.*, Sep. 2005, pp. 66–71.
[34] W. Zheng and L. Liang, "Fast car detection using image strip features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2703–2710.
[35] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 788–801.
[36] W. Brendel, M. Amer, and S. Todorovic, "Multiobject tracking as maximum weight independent set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1273–1280.
[37] B. Leibe, K. Schindler, and L. van Gool, "Coupled detection and trajectory estimation for multi-object tracking," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[38] H. Wang, H. Zhang, and N. Ray, "Adaptive shape prior in graph cut image segmentation," *Pattern Recognit.*, vol. 46, no. 5, pp. 1409–1414, 2013.
[39] J. Zhou, M. Ye, and X. Zhang, "Graph Cut segmentation with automatic editing for industrial images," in *Proc. Int. Conf. Intell. Control Inf. Process.*, Aug. 2010, pp. 633–637.
[40] E. Neuburger and V. Krebs, "Einführung in die theorie des linearen optimalfilters (Kalman-Filter) (introduction to linear optimal filtering theory (Kalman Filter))," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 6, no. 11, p. 796, Nov. 1976.
[41] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, no. 1, pp. 1–10, 2008.
[42] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1926–1933.



Yuan Yuan (M'05–SM'09) is a Full Professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as IEEE TRANSACTIONS AND PATTERN RECOGNITION, and conference papers in CVPR, BMVC, ICIP, and ICASSP. Her research interests include visual information processing and image/video content analysis.

Yuwei Lu received the B.E. degree in software engineering from Northwestern Polytechnical University, Xi'an, China, in 2015, where he is currently working toward the Ph.D. degree with the School of Computer Science and the Center for Optical Imagery Analysis and Learning. His research interests include computer vision and pattern recognition.

Qi Wang (M'15–SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent system from University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently an Associate Professor with the School of Computer Science and the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.