Vehicle Detection Based on Semantic Component Analysis

Guosheng Cui[1,3], Qi Wang[2,*], Yuan Yuan[1]

[1]Center for OPTical IMagery Analysis and Learning (OPTIMAL),
State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics,
Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China
[2]Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University,
Xi'an 710072, Shaanxi, P. R. China
[3]University of the Chinese Academy of Sciences, 19A Yuquanlu, Beijing, 100049, P. R. China
cuiguosheng@opt.cn; crabwq@nwpu.edu.cn; yuany@opt.ac.cn

ABSTRACT

Vehicle detection is a hot topic in traffic monitoring applications. Though many researchers has done a lot work towards this direction, the detection in occluded conditions is rarely explored and it still remains a challenge. In this work, we focus on the occlusion problem in vehicle detection and propose a novel method based on semantic component analysis and scale consideration. Two contributions are claimed in this procedure: 1) Tackling vehicle detection by semantic component detection and synthesis. 2) Addressing the scale variation of vehicles by simple yet effective standard component definition. The experimental results on two typical surveillance videos show that the proposed method can effectively detect the vehicles in the crowded traffic conditions with occlusion.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

General Terms

Algorithms, Experimentation, Performance

Keywords

Traffic monitoring, vehicle detection, occlusion, deformable part based model, semantic component

1. INTRODUCTION

Traffic monitoring systems can extract rich information for the human understanding of the traffic status [11]. The obtained statistics can be substantially useful for various purposes such as abnormal behavior detection, road user (pedestrians, vehicles or motorbikes) recognition, and congestion prediction. To this end, many approaches are feasible but the vision based one has been recognized as one of the most popular techniques because of its invasive property, low cost, and convenience. Therefore, designing a robust and effective method for the processing of camera captured video is fundamentally important. Vehicle detection is a such a significant one because many other applications rely on this preliminary step.

In the recent few decades, lots of methods have been proposed for the vehicle detection. These methods can be roughly divided into two categories, motion-based [2][3][12][14] and feature based [1][4][8][9][15][16]. However, most of them do not try to solve the problem of vehicle detection in the crowded environment. When the traffic becomes heavy, occlusion may cause great impact on the performance of these methods. Therefore, the vehicle detection in the crowded traffic condition is still a problem that needs more research.

1.1 Related work

Vehicle detection has been extensively discussed by previous literatures. However, works toward the problem of occlusion in the crowded traffic conditions are very limited. There are only a few pertinent methods available. For example, Yin *et al.* [17] proposed to handle the occlusion problem by using Hidden Markov Model (HMM). They first exact features from the input images using Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA) and then use HMM to classify the input image into three different areas (road, vehicle head and body). Unfortunately, when the vehicle head is occluded in the congested condition, the performance of the method will be severely impaired.

Robert [13] proposed a framework to extract image features and then to fuse them to detect the vehicle features (headlights or windshields). After that, they are organized to identify a whole vehicle. This method can relieve the problem of occlusion through the usage of partial information. However, it has the similar problem with [17]. Headlights can easily be invisible in the crowded roads.

Kanhere *et al.* [7] proposed to employ the key point extraction for the detection of vehicles on the high way. The feature points are first detected and tracked in the image sequence. Then the 3D coordinates are estimated and grouped together to segment and track the individual vehicles. This method is reported to handle the occlusion problem on the highway. However, it solely depends on the number of feature points which is very limited. So this method needs more other information to make it robust.

More recently, Felzenszwalb *et al.* [6] proposed an object detection system based on the mixture of multiscale

ICIMCS'14, July 10–12, 2014, Xiamen, Fujian, China.

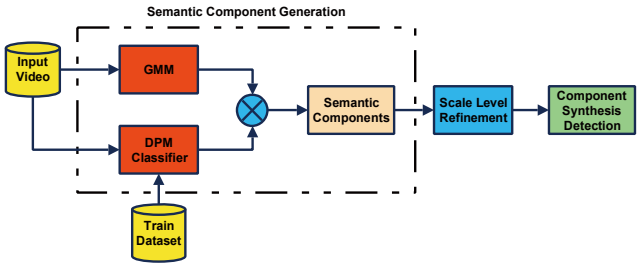


Figure 1: Flow chart of the proposed method.

deformable part-based models. This model can be used to detect vehicles with partial occlusion and its performance is promising. However, when the traffic becomes heavy, the accuracy of this method is not satisfying.

1.2 Overview of the proposed method

In this article, we propose a new method to detect the vehicles in the crowded scene. Our method is based on semantic component detection and analysis, as well as the scale consideration. The flowchart is illustrated in Fig.1 and the detailed introduction is as follows.

- Semantic component generation. Semantic components are generated in this step. To reduce false detection, the foreground estimation is first conducted to extract the moving foreground. Then the deformable part based model (DPM)[6] [5] is applied to recognize the real semantic components contained in the moving blobs.

- Scale level refinement. The obtained semantic components still have false positives after the aforementioned step. To refine the results, the examined scene is uniformly divided into several scale regions, each of which defines a reasonable component size. With this constraint, the true components can be correctly screened.

- Component synthesis detection. The acquired vehicle components cannot make an accurate decision individually. Therefore, they are synthesized according to their geometric relationships. The generated joint components are more robust for the determination of vehicle existence.

1.3 Contributions

The contributions of the proposed method could be summarized as follows:

- Semantic feature analysis in vehicle detection. Traditional methods for vehicle detection is mainly based on the low-level features. But these features have limited ability to describe the object and they are hard to be used for scene understanding. Semantic features are thought to be promising for this problem but no reports have been seen to tackle vehicle detection from this point. In this work, we represent each semantic component by the deformable part based model (DPM), whose inner structure is learned by a latent *SVM* (LSVM) [6]. After that, those detected semantic features are synthesized to jointly recognize the vehicles in the crowded traffic environment. Comparative experiments show that the proposed semantic feature can effectively increase the true detection rate.

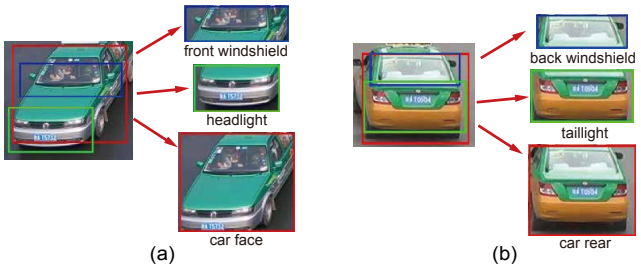


Figure 2: Semantic component definition for the front and back view of the vehicle.

- Simple yet robust scale level refinement. Most existing methods tackle the scale variation by pyramid related approaches, which are often associated with high computational complexity or need special support of hardware [10]. In this work, we design a simple method of defining the reasonable vehicle size by a learned prior. This treatment avoids complex computation of different scales. Experiments demonstrate that this strategy is very useful for ruling out the false detection.

The rest of this paper is arranged as follows. Section 2 introduces the proposed method, including three steps: semantic component generation, scale level refinement, and component synthesis detection. Section 3 shows the experimental results proving the effectiveness of the proposed method. Section 4 concludes the paper.

2. OUR METHOD

In this section, the proposed method is introduced in detail. There are mainly three steps: semantic component generation, scale level refinement, and component synthesis detection.

2.1 Semantic component generation

Part-based model is a widely used model, which can help to tackle the occlusion problem by involving partial information of the object. However, existing works have not yet make full use of the relationship of the parts well. In this paper we introduce the concept of the semantic components, which are specific areas of the vehicle and have an intrinsic connection among them.

Since the front view and back view are the common cases for traffic surveillance, the semantic components are accordingly defined as Fig.2 illustrates. With this definition, the semantic components are generated through three steps. For a given input video, the moving foreground is firstly estimated by Gaussian mixture model (GMM). Then the candidate components are extracted with the DPM model. The obtained foreground and candidate components may both contain the false detections. Therefore, they are finally combined to determine the semantic components.

1) Foreground estimation

The aim of this step is to extract the moving foreground and exclude the interference of background. To this end, the state-of-the-art GMM [18] is adopted as it can timely update its parameters and is adaptive to the changing scene. To be specific, the intensity probability of each pixel in the frame is represented by M Gaussian distributions

$$\hat{p}(x) = \sum_{m=1}^{M} w_m N(x; \mu_m, \Sigma_m), \quad (1)$$

where μ_m and Σ_m are the estimate of the mean and variance of the m^{th} gaussian component, and w_m is its weight. The first B components of the GMM are then chosen to describe the background

$$B = arg\min_{b}(\sum_{m=1}^{b} w_m) > T, \tag{2}$$

where T is a preset threshold indicating the minimum portion of the background in the scene. After that, the Mahalanobis distance between the pixel x and the background distribution is computed, according to which its belonging to the background or foreground can be determined. All the parameters involved in this procedure are learned from previous observations and will be re-calculated once this frame is processed [18].

2) Candidate component generation

In this step, we aim to extract the physical components of the vehicle using the deformable part based model (DPM), and to learn the structure of the components using the latent SVM [6]. The model is trained in a fully supervised framework in which the positive samples are annotated with bounding boxes, and the parameters representing the layout of the parts are learned as latent information.

In the processing, the input image is repeatedly smoothed and down sampled to form an image pyramid, and a variant HOG feature is then computed on each level of the image pyramid to form a feature pyramid. One DPM includes several linear filters, one root filter and a set of its corresponding part filters. All these filters will be applied to the feature maps to check whether one specific location is actually the desired part. This is achieved by

$$f_\beta = \max_{z} \beta \cdot \psi(y, z), \tag{3}$$

where β denotes the vector of model parameters, y denotes the root location in the feature pyramid, z denotes latent variables which means the locations of the parts relative to the root, and $\psi(y, z)$ denotes the feature vector. The output f_β is a confidence score indicating the possibility of the detected part. More detailed explanation can be found in [6].

3) Final component identification

The results from the above two steps are combined here to rule out the false detections of the semantic components. Suppose the bounding boxes of the foreground and the candidate component are denoted as BBF and BBC, respectively. Our assumption is that the BBCs whose overlap with BBFs are large enough should be maintained as the true detections. To be specific, we keep the BBCs which conform to the following formula as the final detected components

$$\frac{BBC \bigcap BBF}{BBC} \geq \theta, \tag{4}$$

where θ denotes the threshold.

2.2 Scale level refinement

After performing the previous steps, most of the semantic components are detected. But there still exists a problem: the sizes of the components are not necessarily correct. This is mainly caused by the scale variation of the objects because its size changes uniformly from the near to the distant. But the detector cannot well model this difference. Existing methods tackle this problem by pyramid representation of the objects, but it will bring tremendous computational complexity. In this work, we take a simple strategy to tackle the scale problem to further refine the detection results, without the need of computing the precise scale for each kind of component. In the implementation, the scene is uniformly divided into several horizontal regions. In each region, a standard component (SC) of every kind is defined, as illustrated in Fig. 3. With the help of SC, the false detections whose sizes are much smaller or bigger than the size of SC can be discarded and the detection results will be be further refined. Fig. 4 shows the effect of this scale level refinement through a typical example.

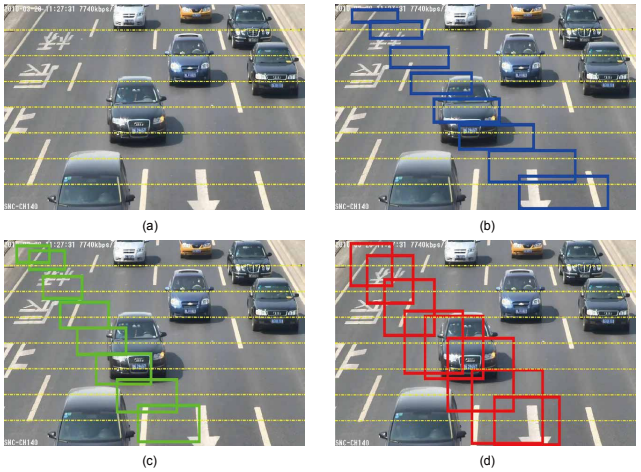


Figure 3: Illustration of standard component definition. (a) The divided horizontal regions. The standard components of front windshield (b), headlight (c), and car face (d) in each horizontal region.

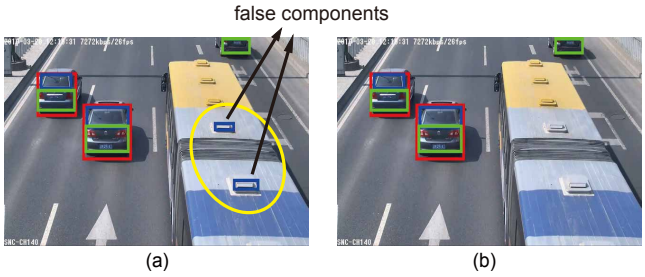
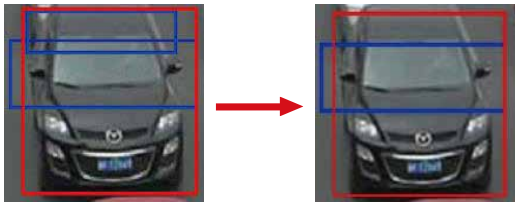


Figure 4: Illustration before (a) and after (b) scale level refinement. The false detections of the semantic components are effectively eliminated by the comparisons with SCs.

As for the selection of SCs, they are determined by a voting strategy of maximum likelyhood. Since each component is identified with a probability, the component with the highest probability of one kind is specified as the SC for this type of component. The SCs keep updating frame by frame, if there are new components with higher probabilities.

2.3 Component synthesis detection

With the obtained semantic components, a final decision should be made to decide whether there exists a vehicle. In order to synthesize these components for an accurate identification, one question needs to be answered first: how the different components are associated together to one single



(a) (b)

Figure 5: When several components of the same kind belong to one individual vehicle (a), only the one with the highest confidence score will be kept (b).

vehicle. Since the semantic components are all physically defined, they maintain certain kinds of relationships.

Take the front view for example. The first restriction is the headlight and front windshield should have large overlaps with the car face (Because the car face includes the two components as shown in Fig. 2). Accordingly, we have

$$\frac{BBC_f \bigcap BBC_h}{BBC_h} \geq \theta_h, \tag{5}$$

$$\frac{BBC_f \bigcap BBC_w}{BBC_w} \geq \theta_w. \tag{6}$$

where BBC_f, BBC_h and BBC_w respectively denote the car face, headlight and front windshield, θ_h and θ_w are two threshholds.

Furthermore, three constraints should also be satisfied for a properly organized vehicle. They are

- Headlight locates below front windshield;
- Headlight locates on the lower half of a car face;
- Front windshield locates on the upper half of a car face.

These three restrictions can be obviously inferred from Fig. 2. But in the implementation, there may occur two components of the same kind for one single vehicle. Fig. 5 illustrates such an example containing two detected front windshields that both satisfy all the constraints. In this case, the one with the higher probability is selected as the true match.

After associating the components to one single vehicle, we should make the final decision. Two criterions are intuitively defined as

- If two or three components have been detected for a vehicle, we will define it as a true detection.
- If only one component has been detected and its confidence value is bigger than a threshold λ, we will define it as a true detection.

The above processing is all introduced in the context of front view condition. As for the back view situation, the procedure follows a similar manner.

3. EXPERIMENTS

3.1 Dataset

Three traffic monitoring video sequences are employed in the experiments. The first one is used for selecting the training components. For the front view and back view situations, 400 samples are manually chosen for each kind of component. Then they are all applied in the DPM training process. For the second and third videos, there are respectively 450 front view and 1800 back view frames for testing. All the three videos have many occlusions between two adjacent vehicles.

3.2 Implementation details

There are several parameters to be set in this work and they are intuitively set according to our experimental results. We only introduce the choices in the front view condition. Their settings in the back view are the same. The first one is the threshold θ in Section 2.1. For the car face, front windshield, and headlights, they are set as 0.75, 0.75, and 0.8 respectively. For the θ_h and θ_w in Section 2.3, they are both set as 0.75. The confidence score λ is set to -0.5. Besides, the GMM and DPM implementations use the default parameters provided by the authors [18][6].

As for the evaluation metrics, precision and the recall are employed. They are defined as

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, \tag{7}$$

where TP denotes the true positives, FP denotes the false positives, and FN denotes the false negatives.

3.3 Results

The proposed method is made up of several parts, including GMM, DPM, scale level refinement (SLR) and component synthesis detection (CSD). In order to prove the presented method is most effective, we conduct the experiments on different part combinations and then compare their performance. These experiments are denoted as DPM + SLR + CSD (DSC for short), GMM + DPM + CSD (GDC for short), GMM + DPM + SLR + CSD (GDSC for short). Besides, we also compare the proposed method with the original DPM and DPM + SRC (DPMS).

The experimental results are shown in Table 1. We can see clearly that, most of the time, the proposed method GDSC has a higher precision and recall at the same time. Though some other combinations may outperform GDSC with respect to a single precision or recall, the other metric is usually very low for them. In fact, we cannot sacrifice one metric to make the other metric more satisfying. Both of them are very critical for a method. From this aspect, the proposed method is more useful. Besides, the fact that the proposed GDSC is superior to DSC proves the foreground extraction is necessary in our processing. GDSC also outperforms GDC, which indicates the scale level refinement plays an important part in excluding the false detections. DPM is obviously inferior to GDSC and DPMS, which verifies the claimed semantic representation and synthesis are more effective than the original part based model.

The reason behind the success of the proposed method mainly lies in two points. The first one is the informative descriptive ability of the semantic features and the robust manner to jointly use them. The semantic feature is a high level understanding of the scene, enabling more discriminative capability for vehicle recognition compared with tradi-

Methods	GDSC		GDC		DSC		DPMS		DPM	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Back view	0.9933	0.9490	0.8841	0.9809	0.9510	0.9490	0.9975	0.8544	0.9865	0.8544
Front view	0.9964	0.9325	0.9286	0.9635	0.9796	0.9292	0.9987	0.8406	0.9941	0.8396

Table 1: Experimental comparison of different combinational methods.

tional low level features. The second one is the tactful multiscale treatment of the vehicle size. Though the presented strategy is simple, but it is very efficient and robust. These two factors together ensure the outstanding performance of the proposed method.

4. CONCLUSION

In this paper, we propose a novel method to tackle the occlusion problem in vehicle detection. Firstly, the semantic components are defined and extracted from the input video sequences. Then the obtained components are refined from a scale level consideration. In the end, the semantic components of different kinds are synthesized to make the final decision of vehicle existence. Two contributions are claimed in the procedure: 1) Tackling vehicle detection by semantic component detection and synthesis; 2) Addressing the scale variation of vehicles by simple yet effective standard component definition. The experimental results show that the synthesis of the semantic components can detect much more occluded vehicles, and the scale constraint can reduce the number of false positives.

In our work, we do not define the side view components of the vehicles. Only the most frequent front and back view components are discussed. In the future, we plan to extend this implementation to random viewpoints and make the method more adaptive.

5. ACKNOWLEDGMENTS

This work is supported by the State Key Program of National Natural Science of China (Grant No. 61232010), the National Natural Science Foundation of China (Grant No. 61172143, 61379094 and 61105012), and the Fundamental Research Funds for the Central Universities (Grant No. 3102014JC02020G07).

6. REFERENCES

[1] N. Buch, J. Orwell, and S. A. Velastin. 3d extended histogram of oriented gradients (3dhog) for classification of road users in urban scenes. In *Proc. British Machine Vision Conference*, pages 1–11, 2009.

[2] N. Buch, J. Orwell, and S. A. Velastin. Urban road user detection and classification using 3-d wireframe models. *IET Computer Vision*, 4(2):105–116, 2010.

[3] Z. Chen, T. Ellis, and S. A. Velastin. Vehicle detection, tracking and classification in urban traffic. In *Proc. IEEE Conference on Intelligent Transportation Systems*, pages 951–956, 2012.

[4] J.-Y. Choi, K.-S. Sung, and Y.-K. Yang. Multiple vehicles detection and tracking based on scale-invariant feature transform. In *Proc. IEEE Conference on Intelligent Transportation Systems*, pages 528–533, 2007.

[5] J. Fang, Q. Wang, and Y. Yuan. Part-based online tracking with geometry constraint and attention selection. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):854–864, 2014.

[6] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[7] N. K. Kanhere, S. T. Birchfield, and W. A. Sarasua. Vehicle segmentation and tracking in the presence of occlusions. *Transportation Research Record: Journal of the Transportation Research Board*, 1944(1):89–97, 2006.

[8] W. Lu, S. Wang, and X. Ding. Vehicle detection and tracking in relatively crowded conditions. In *Proc. IEEE Conference on Systems, Man and Cybernetics*, pages 4136–4141, 2009.

[9] X. Ma and W. E. L. Grimson. Edge-based rich representation for vehicle classification. In *Proc. IEEE International Conference on Computer Vision*, pages 1185–1192, 2005.

[10] Y. Ma, J. Kosecka, and S. Sastry. Optimization criteria and geometric algorithms for motion and structure estimation. *International Journal of Computer Vision*, 44(3):219–249, 2001.

[11] V. D. Nguyen, T. T. Nguyen, D. D. Nguyen, S. Lee, and J. W. Jeon. A fast evolutionary algorithm for real-time vehicle detection. *IEEE Transactions on Vehicular Technology*, 62(6):2453–2468, 2013.

[12] A. Ottlik and H.-H. Nagel. Initialization of model-based vehicle tracking in video sequences of inner-city intersections. *International Journal of Computer Vision*, 80(2):211–225, 2008.

[13] K. Robert. Video-based traffic monitoring at day and night vehicle features detection tracking. In *Proc. IEEE Conference on Intelligent Transportation Systems*, pages 1–6, 2009.

[14] X. Song and R. Nevatia. Detection and tracking of moving vehicles in crowded scenes. pages 4–4, 2005.

[15] Q. Wang, Y. Yuan, P. Yan, and X. Li. Saliency detection by multiple-instance learning. *IEEE Transactions on Cybernetics*, 43(2):660–672, 2013.

[16] S. Wang, L. Cui, D. Liu, R. C. Huck, P. K. Verma, J. J. S. Jr., and S. Cheng. Vehicle identification via sparse representation. *IEEE Transactions on Intelligent Transportation Systems*, 13(2):955–962, 2012.

[17] M. Yin, H. Zhang, H. Meng, and X. Wang. An hmm-based algorithm for vehicle detection in congested traffic situations. In *Proc. IEEE Conference on Intelligent Transportation Systems*, pages 736–741, 2007.

[18] Z. Zivkovic and F. van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, 2006.