Part-based Pedestrian Detection using Grammar Model and ABM-HoG Features

Bo Li*, Ye Li*, Bin Tian*, Fenghua Zhu*†, Gang Xiong*†, and Kunfeng Wang*†
*The State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences
Beijing, China
Email: bo.li@ia.ac.cn
†Dongguan Research Institute of CASIA
Cloud Computing Center, Chinese Academy of Sciences
Dongguan, China

***Abstract*—To handle the pedestrian appearance and pose variations in complex traffic environments, we present one part-based pedestrian detection approach using a stochastic grammar model in this paper. The And-Or graph model is introduced to represent the human body as an assembly of compositional and reconfigurable parts. Thus, the task of detection is converted into the human parsing problem, which is a Bayesian inference process. We model the appearance of pedestrian parts in a rich feature representation. This appearance model enhances the Histogram of Gradients (HoG) map with Active Basis Model (ABM), which is a sparse deformable template depicting salient structures of objects. Then, geometry constraints among parts are described by Gaussian distributions. Finally, the bottom-up parsing inference is conducted by aggregating scores to get the pedestrian detection responses. In experiments, we show the superiority of our appearance model, as well as the reliable pedestrian detection results of our approach in complex traffic scenes.**

***Keywords*—pedestrian detection; grammar model; part-based object detection**

I. Introduction

Pedestrians are important participants in transportation systems. With the popular application of image sensors, vision-based pedestrian detection has become one hot topic in the research of Intelligent Transportation Systems (ITS). It can be used in video surveillance to collect pedestrian data for traffic management and analysis in artificial transportation systems. [1]. Meanwhile, pedestrian detection is also a key module in the Advanced Driver Assistance System (ADAS) for intelligent vehicles to realize automatic collision avoidance [2].

Nowadays, pedestrian detection is still one challenging task in computer vision and ITS research. Firstly, the pedestrian is a highly-articulated object. The intra-class differences in pedestrian category are extremely obvious because of varieties of clothing, poses, appearances, and so on. Secondly, lots of disturbances from the real-world traffic environment make this problem more difficult, such as background clutter, severe self-occlusion, and various illumination conditions and weather. These difficulties make pedestrian detection be a long-standing topic.

This work is supported in part by NSFC project 71232006,61233001, 90920305; and CAS 2F11N01, 2F11N07.

Large amount of work has been done on pedestrian detection in past years [3]. Most of the traditional methods are holistic-based, which take the entire human body as the detection target [4]–[9]. These methods mainly focus on utilizing better combination of features and classifiers. Lots of effective feature descriptors on shape, texture, and color are proposed to map pedestrian images into feature vector space. Histogram of Gradient (HoG) [5] is one of the most widely used shape features in pedestrian detection. HoG greatly promotes the development of pedestrian detection. Many later works are developed by fusing multiple types of features with HoG to enrich the pedestrians representation [7]–[9]. This strategy can improve the detection results to some extent.

However, holistic-based detection methods have their limitations. For example, they can not handle the occlusion situations, and are hard to capture large variations of pedestrians. Thus, part-based detection models under various formalisms have been proposed and achieve great success in recent years [10]–[15]. Compared with holistic-based detection models, part-based models are more suitable for detecting articulated objects. They divide pedestrians into several parts and utilize contextual information among these constituent parts. The structure and appearance features are simultaneously considered in detection. Therefore, it is more robust with occlusion and appearance variations in pedestrian detection problems.

In this paper, we present a novel part-based pedestrian detection algorithm to address some difficulties in real traffic application. Our algorithm is motivated by the part-based models for human parsing and pose estimation [14]–[16]. A new stochastic grammar model, And-Or graph [17], [18], is utilized in this paper under the similar framework of [16]. We construct a specific And-Or graph model to represents the human body as one assembly of compositional and reconfigurable parts. These parts in our model are described by their geometric configurations as well as type attributes. Based on the grammar model, large variations in human poses and appearances can be easily captured by a set of production rules. We define the constraints among parts as one probability model. Thus, the detection task can be considered as a parsing problem through bottom-up inference and dynamic programming.

Another contribution of this paper is that a new pedestrian appearance model is proposed. Under the framework of

978-1-4799-0380-1/13/$31.00 ©2013 IEEE

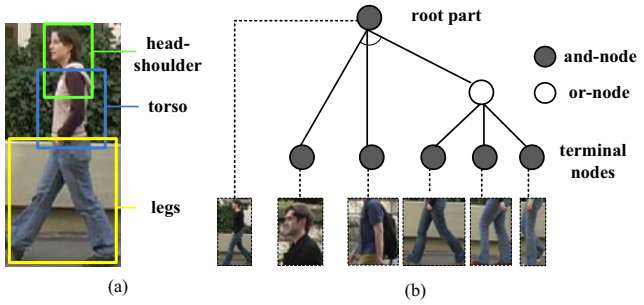

Fig. 1. The part-based representation of pedestrians. (a) Definition of pedestrian parts. (b) And-Or graph grammar model.

multi-feature fusion, we combine two kinds of heterogeneous features to describe both shape and texture characteristics of an object simultaneously. Combination of the two complementary features increases the discriminability for object detection. Better pedestrian part detector also makes contributions to the final detection accuracy.

This paper is organized as follows. In the next section, we explain the proposed And-Or graph model for pedestrian detection. Then, our pedestrian detection algorithm based on the grammar model is presented in Section III, including pedestrian appearance model, geometric model, and complete inference process for human parsing. In Section IV, we give the experiments to show the performance of our proposed algorithm. Finally, we briefly make the conclusion in Section V.

II. And-Or Graph Grammar

In this section, an And-Or graph grammar model is utilized to represent the appearance variations of pedestrians, as in [16]. There are two kinds of nodes in And-Or graph: and-nodes and or-nodes. And-nodes mean the part forms, which uniquely identify the specific compositional parts. Or-nodes denote the part types, which consist of alternative part forms. Different part forms of the same part type are replaceable. The production rules in grammar for two kinds of nodes denote part compositions and part form selections respectively. Owing to the various combinations of the production rules in the graphical structure, the grammar model can generate a series of objects with different appearance.

Taking the general characteristics of pedestrians into consideration, we hope to construct a specific And-Or grammar model for pedestrian detection. Poses of most walking pedestrians are generally known or can be partially inferred. Hence, we divide the whole pedestrian body into three parts: head-shoulder, torso, and legs. The partition of a pedestrian is illustrated in Fig. 1(a).

Since walking pedestrians keep their upper-body straight in most time, the appearance variations in head, shoulder, and torso is relatively small. Although the arm poses may be changed because of the arm swing during walking, these differences can be ignored because the torso has more appearance evidences. Thus, for simplicity, we do not assign multiple part forms for head-shoulder and torso. They are directly configured as terminal nodes.

However, the motion of legs is quite obvious during people walking. It is difficult to use a uniform model to capture the apparently different leg poses. Thus, we aim to separate legs into several part forms. Based on several collected examples of leg parts, we try to cluster them by aspect ratio using K-means. After some trials, it is observed that three forms are sufficient to cover most of the leg poses during people walking.

As a result, we build a new And-Or grammar model in this paper, as shown in Fig. 1(b). We can see that the full body of a pedestrian is considered as the root part, which is designed as the start of grammar. The whole body is composed by three constituent part types. Of all these part types, the node for legs is an or-node, which contains three compatible part forms. These part forms are represented by legs-1, legs-2, and legs-3 respectively. In the And-Or graph, both terminal and non-terminal parts are included in the grammar dictionary. Thus, the holistic appearance model is also utilized in detection. In this way, holistic-based detection and part-based detection are combined in the And-Or grammar model.

Through selecting certain part forms from or-nodes in the bottom-up inference, one specific pedestrian example is obtained. For each and-node encountered in inference, a unique parse node is determined to form a parse graph at last. According to [16],the parse tree is denoted as $pg = (V_{pg}, E_{pg}, R_{pg})$, where V_{pg} are instantiations of parse nodes, E_{pg} are the corresponding edges from part form compositions, and R_{pg} denote the constraints between parts. Each parse node is denoted by $pn = (f, x, y, s)$, which contains the part form f, coordinates (x, y) in the image , and scale s that represents the part size.

The And-Or model represent a distribution on the parse graph in a Bayesian framework, as shown in (1).

$$P(pg|I) \propto P(I|pg)P(pg) \tag{1}$$

I denotes the input image and pg denotes the parse graph. The likelihood distribution $P(I|pg)$ can be called as an appearance model, which is related to the part detection responses in the image. Prior model $P(pg)$ evaluates the probability for a certain parse graph. It considers the geometric relations between parts. We will discuss these two models in the Section III. Thus, the detection problem is converted to a human parsing task to maximize the posterior probability.

III. Pedestrian Detection By Grammar Model

A. Part Appearance Model

It is assumed that the appearance for each parse node is conditionally independent, as in (2)

$$P(I|pg) = q(I_{\overline{\Lambda_{pg}}}) \prod_{v_i \in V_{pg}} p(I_{\Lambda_{v_i}}|v_i) \tag{2}$$

where $I_{\Lambda_{v_i}}$ and $q(I_{\overline{\Lambda_{pg}}})$ mean the distribution for image patches occupied by part v_i and without parse nodes respectively. Ignoring the constant background distribution, the likelihood can be decomposed as the factor of all the part likelihood. The likelihood evaluates the probability of the part presence in the image patch given the appearance model. This probability can be reflected by the part detection responses. Therefore, we hope to train more discriminative detector for each part form in the grammar dictionary.

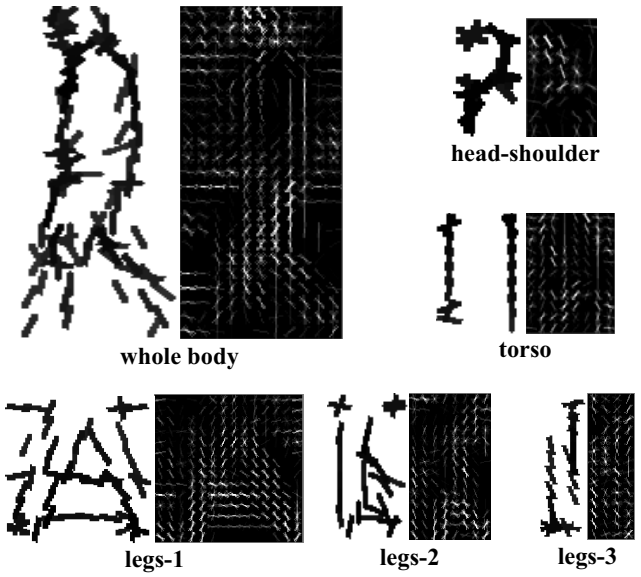

Fig. 2. Trained ABM and HoG templates for all part forms. For each pair of templates, the left one is ABM template and the right one is HoG template.

Generally, pedestrians have the recognizable shape and texture features for distinguishing them from other objects. However, constituent parts of pedestrians usually have less discriminative information. It is likely to detect many false alarms of parts, which may influent the overall detection performance. In this paper, a multi-feature fusion strategy is applied to combine two heterogeneous and reliable features, HoG and Active Basis Mode (ABM) [19]. A more comprehensive appearance model is obtained to increase the discriminability of the part detectors.

HoG is a widely used feature descriptor for object detection and classification in recent years. HoG densely extract gradient orientation histogram from overlapping blocks in the image. Hence, it can effectively describe the local shape and appearance of objects. However, HoG feature simply give the orientation field of objects. While important category information is reflected by the HoG feature map, the statistical histogram is more appropriate to capture the texture characteristic of an object. It is hard to observe the salient structures or contour attributes, which are even more discriminative for detecting some human parts like head-shoulder. Therefore, we introduce ABM to enhance the representation for pedestrian parts.

Compared with HoG feature map, ABM gives a more sparse representation for explicit object shape. It can be considered as a deformable image template of sketch elements, which are learned from a few image examples. Elements of the sketch template are small Gabor filters, which have strong responses to edges at a given orientation. Thus, the elements represent several short edgelets at different orientations. Each element is allowed to perturb locally to fit the small variations of objects. From Fig. 2, we observe that the learned ABM templates can depict the primal sketch features, which describe the rough contour of objects. Thus, their combination with HoG features will be a complementary representation on some local details of objects.

The templates for the two features are trained separately in the same training set. HoG template is trained by linear SVM, which assign a weight for each bin in the histogram. In order to capture the appearance details of constituent parts, we use cells with smaller size than that for root part in HoG feature extraction. The matching score of a HoG template is defined in (3), where $\mathbf{h}$ denotes the HoG feature of image patch I_Λ, and $\mathbf{w}$ is the learned weight vector. The ABM is learned by projection pursuit. The optimal Gabor bar is selected in each training step to the model until it is sufficiently approximate the target distribution of filter responses. The template can be learned by fully generative way, which only needs a small number of positive examples without negative ones. Therefore, ABM can represent the essential characteristic of objects and will robust to the background clutter. The learned ABM template includes the position, orientation, and weights for each Gabor bases. The matching score of ABM template is defined in (4), where n is the number of Gabor basis in the template and $r(I_\Lambda)$ means the Gabor response in a local image patch. λ_j and $\log z_j$ are the model parameters and normalizing constants respectively. We can consider them as the Gabor weights. As a result, two templates are trained for each part forms. Fig. 2 shows the learned ABM and HoG templates for all parts in grammar dictionary.

$$S_{HoG}(I_\Lambda) = \mathbf{w}^T \mathbf{h}(I_\Lambda). \tag{3}$$

$$S_{ABM}(I_\Lambda) = \sum_{j=1}^{n} [\lambda_j r_j(I_{\Lambda_j}) - \log z_j]. \tag{4}$$

Given part templates and an input image, the process for appearance likelihood computation is illustrated in Fig. 3. Firstly, the gradient operator and a series of Gabor filters in different orientations are applied to the image. Thus, we obtain the related gradient map and Gabor filter responses. Next, HoG and ABM template matching scores are computed by (3) and (4). The sliding window strategy can be conducted to perform template matching in each image coordination. We combine both scores by linear weighted methods, as in (5). A sigmoid transformation $r(.)$ is then employed to suppress the weighted score sum too large. In this way, we acquire the result of image likelihood $p(I_{\Lambda_v}|v)$ for a part node v by the rich appearance representation.

$$p(I_{\Lambda_v}|v) = \exp\{r(k_h S_{HoG} + k_a S_{ABM})\} \tag{5}$$

The detection performance evaluation for this ABM-HoG appearance model is detailed in the Section IV. An example of appearance likelihood for each part is shown in Fig. 4. We can see that high detection scores may appear in the wrong places. In addition, because of obvious appearance variations in clothing, the right pedestrian in the image does not achieve peak value for the whole body detection responses. Hence, we aim to correctly locate the pedestrian and inhibit false alarms through a combination of evidences from other parts.

B. Geometry Model

In this paper, we consider prior model as the geometric relationship between parse nodes. The geometry model describes

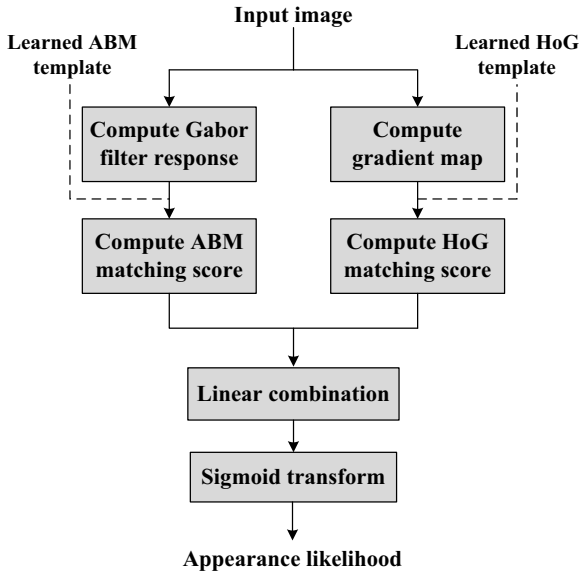

Fig. 3. Flowchart of the appearance likelihood computation by fusing ABM and HoG.

the probability distribution on the geometric relationship for all parts. In order to realize an efficient bottom-up inference, we restrict that the geometric relations only exist in pairwise edges linking parent and child node. Thus, there are no cycles in our grammar model. The parts of the human body construct a tree structure, which is the topology of relation set R_{pg}. As a result, the entire prior model defined on the relation set of parse graph uses the tree factorization format, which is written as (6).

$$P_g(pg) = \prod_{(i,j)\in R_{pg}} p(v_i, v_j) \tag{6}$$

We define Gaussian relation for each pair of parent-child parts. Since pedestrian parts in our grammar model are not likely to behave strongly articulated feature, we do not make special transformation to those part coordinates. Using $\mathbf{x} = (x, y, s)$ as the shorthand for the geometric state of a parse node, probability distribution between part pairs is expressed in (7). This distribution computes the deformation score for two parts. Reasonable spatial position relation or slight deformation between a pair of parts will contribute a high probability to the prior model. Meanwhile, two parts with severe deformation will also reduce the prior, which makes the parse graph tend to an irregular combination of parts.

$$p(v_i, v_j) \propto \mathcal{N}(\mathbf{x}_j - \mathbf{x}_i, \mu_{ij}, \Sigma_{ij}). \tag{7}$$

C. Inference

Based on the And-Or grammar model, the pedestrian detection can be viewed as human parsing problems in multiple locations. The inference process for human parsing is similar with that in [16]. Parsing is to find the parse graph with maximal posterior probability. The log-posterior of a parse node v in this grammar model can be expressed by a recursive scoring function.

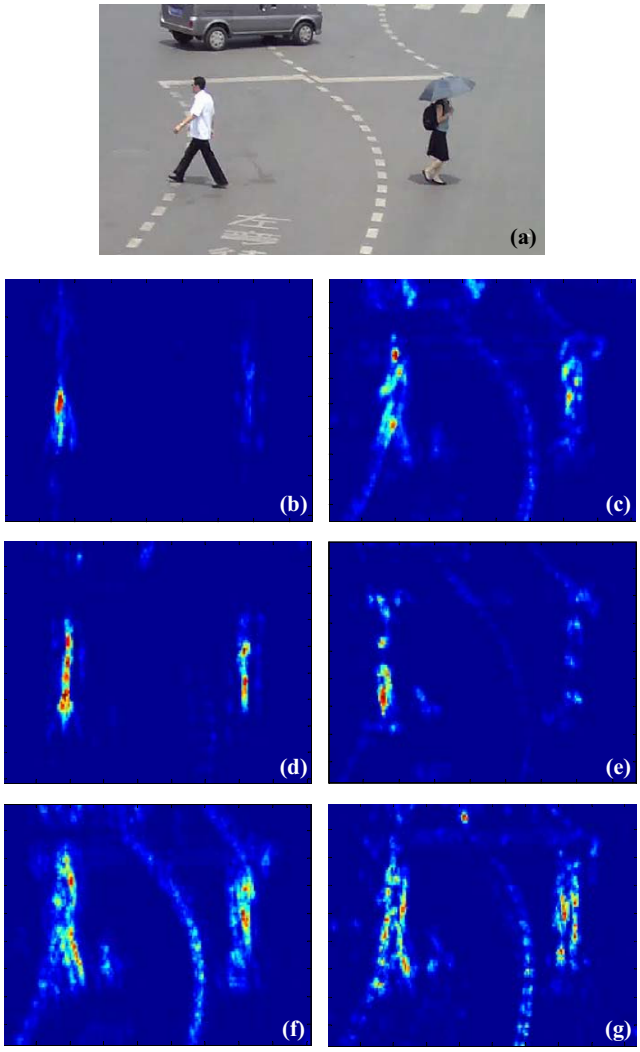

Fig. 4. An example of appearance likelihood probability map. (a) Input image. (b)-(g) are maps for each part form. Red regions indicate high values, and blue regions indicate low values. (b) Whole body. (c) Head-shoulder. (d) Torso. (e) Legs-1. (f) Legs-2. (g) Legs-3.

For pedestrian detection, we aim to find the parse graph pg^*, which maximizes the score function $s(v_0|I)$ with root part v_0. The final score for each pedestrian hypothesis is recursively calculated by collecting scores from bottom to upper levels. The aggregating process from child node v_j to parent v_i is achieved by performing the following three steps:

- Incremental aggregation score computation: We compute the incremental aggregation score $S(v_i, v_j)$, which includes appearance score, geometry score, and the final aggregation scores from all children. If v_j is a terminal node, $S(v_i, v_j)$ only remains the appearance term.
- Location maximization: The above incremental aggregation score is maximized over all the geometric location of v_j. The optimal location of v_j is determined for each form.
- Part form maximization: Optimal part form is selected by maximizing the result of last step over different

(a) (b)

Fig. 5. Inference results. (a) Final aggregation scores for the root part. Red regions indicate high scores, and blue regions indicate low scores. (b) Pedestrian detection results. Bounding boxes for both pedestrians and their parts are marked in the image.

part forms of v_j.Thus, with the contribution of v_i, the final aggregation score of v_j is obtained. When further inference is needed, it will substitute the result of part form maximization in the process of incremental aggregation score computation.

In order to make the inference more effective, we use lookup tables to save some intermediate results on maximal score maps during the recursion. The bottom-up inference is conducted recursively by repeating above procedures until the algorithm reaches the root node. Thus, detection responses of a test image are obtained. Then, we apply thresholding and Non-Maximal Suppression (NMS) to locate the pedestrian hypotheses with local maximal score. For each pedestrian hypothesis, the optimal locations of all the parts can also be determined by backtracking from the root node to the terminal nodes. This is done by simply replacing the max with an argmax in (13). This top-down inference recovers the full parse graph. We can search the detailed position and part forms of the detected pedestrian.

A visualization of the final aggregation score map through the bottom-up inference is shown in Fig. 5(a). The bounding boxes of pedestrians as well as the inferred part forms are determined, as shown in Fig. 5(b). Through the evidence aggregation of child nodes, detection score of the right pedestrian is enhanced, which leads to more reliable detection performance.

IV. Experiments

A. Evaluation for ABM-HoG Appearance Model

At first, we conduct some experiments to evaluate the discriminability of our rich appearance representation model on fusion of ABM and HoG. In this paper, implementation of HoG extraction refers to [10], which project the gradient orientation histogram in a cell to one 31-dimensional feature vector. The cell size in HoG computation is set to 8. In the training and matching of ABM templates, we choose a dictionary of Gabor filters at 12 orientations. The number of sketch elements is limited to 60. We use the INRIA person dataset [5] to train templates and evaluate the classification performance. Our ABM-HoG feature is compared with ABM matching and HoG-SVM algorithm respectively in the same parameter configuration. We use the metric of Average Precision (AP) and Area Under ROC Curve (AUC) to evaluate the discriminability of feature models in simple holistic pedestrian detection. The comparison results are shown in TABLE I. It is observed that the fusion of ABM and HoG exceed any of these single feature descriptions.

TABLE I. Comparison of Appearance Models on INRIA Dataset

Appearance Models	AP	AUC
HoG+SVM	0.869	0.975
ABM	0.879	0.976
ABM+HoG	0.902	0.985

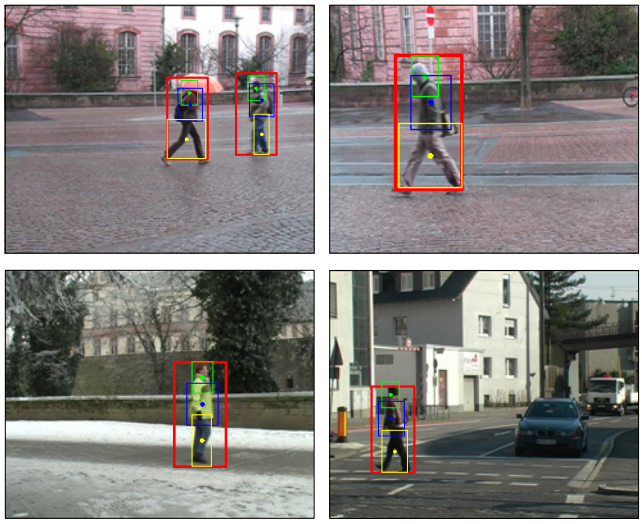

Fig. 6. Examples of pedestrian detection results.

B. Experiments on Traffic Scenes

We choose the public TUD-Pedestrian dataset [20] to show our pedestrian detection performance on complex traffic scenes. Test images in this dataset are captured in various outdoor natural environment from side-view. Thus, all the pedestrians exhibit significant appearance variations in clothing and articulation with serious background clutter. The training dataset contains 400 pedestrian images that are captured in the similar scene and configuration. The training images are fully annotated with part position, part types and part forms. We train our multi-feature appearance model as well as Gaussian geometry model for each part on these images. The detection results of some examples in test dataset are given in Fig. 6. We can see that our part-based pedestrian detection approach successfully locates the pedestrians with various poses and appearances in different and uncontrolled traffic scenes. Besides, after a pedestrian is detected, the specific position and part form of each part type can be obtained through the top-down inference. These extra part information is useful for some future high-level applications, such as action recognition and scene understanding.

V. Conclusions

In this paper, we present one part-based pedestrian detection approach using a grammar model to represent the human body as an assembly of compositional and reconfigurable parts. The pedestrian is decomposed and described by the And-Or grammar model, where the and-nodes denote the part

composition and the or-nodes denote alternative parts with different attributes.

According to the characteristics of pedestrians in traffic scene, we construct a simple And-Or graph model. Thus, pedestrian detection is converted into human parsing problem to find the optimal parse graph that maximizes the posterior probability. In order to obtain better part detection performance, we use a rich feature representation to model the appearance of pedestrian parts. This new appearance model enhances HoG feature with a sparse deformable template ABM to depict the salient part structures. Besides, the geometry constraints among parts are described by Gaussian distributions. Finally, the bottom-up and top-down inference are respectively conducted to get the pedestrian detection response and recover the specific parse tree. In the experimental section, we show the superiority of our rich appearance representation as well as our reliable pedestrian detection results in typical traffic scenarios. Based on this part-based detection model, our approach can capture the variations in pedestrian appearance and pose. Meanwhile, the aggregation of detection evidences from separate parts also helps to reduce the false alarms. In the future work, we aim to build more complex grammar models to represent the pedestrian. In addition, some hardware acceleration strategies can be considered to improve the detection efficiency.

ACKNOWLEDGMENT

We would like to thank professor Fei-Yue Wang for his instruction and encouragement to our work.

REFERENCES

[1] F.-Y. Wang, "Parallel control and management for intelligent transportation systems:concepts, architectures, and applications," *IEEE Trans. on Intell. Transp. Syst.*, vol. 11, no. 3, pp. 630–638, 2010.

[2] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, 2010.

[3] B. S. P.Dollar, C. Wojek and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Trans. on Pattern Analysis and Mach. Intell.*, vol. 34, no. 4, pp. 743–761, 2012.

[4] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int J. Comput. Vis.*, vol. 65, no. 2, pp. 153–161, 2005.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[6] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[7] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 32–39.

[8] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 91.1–91.11.

[9] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1030–1037.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[11] L. Zhu, Y. Chen, A. Torralba, W. Freeman, and A. Yuille, "Part and appearance sharing: recursive composional models for multi-view multi-object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1919–1926.

[12] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *Proc. Europ. Conf. Comput. Vis.*, 2010, pp. 168–181.

[13] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnorr, "A study of parts-based object class detection using complete graphs," *Int. J. Comput. Vis.*, vol. 87, no. 1-2, pp. 93–117, 2010.

[14] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2011, pp. 723–730.

[15] M. Andriluka, S. Roth, and B. Schiele, "Discriminative appearance models for pictorial structures," *Int. J. Comput. Vis.*, vol. 99, no. 3, pp. 259–280, 2012.

[16] B. Rothrock and S.-C. Zhu, "Human parsing using stochastic and-or grammars and rich appearances," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 640–647.

[17] S.-C. Zhu and D. Mumford, "A stochastic grammar of images," *Found. Trends Comput. Graph. Vis.*, vol. 2, no. 4, pp. 259–362, 2006.

[18] T. Wu and S.-C. Zhu, "A numeric study of the bottom-up and top-down inference processes in and-or graphs," *Int. J. Comput. Vis.*, vol. 93, no. 2, pp. 226–252, 2011.

[19] Y. N. Wu, Z. Si, H. Gong, and S.-C. Zhu, "Learning active basis model for object detection and recognition," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 198–235, 2010.

[20] S. R. M. Andriluka and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1919–1926.