Pedestrian Detection based on Clustered Poselet Models and Hierarchical And-Or Grammar

Bo Li, Yaobin Chen, *Senior Member, IEEE*, Fei-Yue Wang, *Fellow, IEEE*

***Abstract*—In this paper, a novel part-based pedestrian detection algorithm is proposed for complex traffic surveillance environments. In order to capture posture and articulation variations of pedestrians, we define a hierarchical grammar model with the And-Or graphical structure to represent the decomposition of pedestrians. Thus, pedestrian detection is converted to a parsing problem. Next, we propose clustered poselet models, which use the affinity propagation (AP) clustering algorithm to automatically select representative pedestrian part patterns in the keypoint space. Trained clustered poselets are utilized as the terminal part models in the grammar model. Finally, after all clustered poselet activations in the input image are detected, one bottom-up inference is performed to effectively search maximum a posterior (MAP) solutions in grammar model. Thus, consistent poselet activations are combined into pedestrian hypotheses and their bounding boxes are predicted. Both appearance scores and geometry constraints among pedestrian parts are considered in inference. A series of experiments are conducted on images both from the public TUD-Pedestrian dataset and collected in the real traffic crossing scenarios. The experimental results demonstrate that our algorithm outperforms other successful approaches with high reliability and robustness in complex environments.**

***Index Terms*—And-Or graph, clustered poselet, computer vision, pedestrian detection**

I. Introduction

VISION-BASED pedestrian detection has become one hot topic in intelligent transportation systems (ITS). It can collect pedestrian data for traffic management and analysis in artificial transportation systems [1]. Besides, it is a key module in advanced driver assistance systems (ADAS) of intelligent vehicles [2]. Detection results provide important data for robust vehicle tracking control [3], [4].

In natural traffic surveillance environments, pedestrian detection is one challenging task. Firstly, pedestrians are non-rigid and highly-articulated objects. Their intra-class differences are extremely obvious because of varieties in clothing, poses, appearances, and so on. Secondly, many environmental disturbances may deteriorate the detection performance, such as cluttered background, various illuminations, and severe occlusions. Especially in urban mixed traffic scenarios, pedestrians are visually occluded by vehicles or other moving objects prevalently.

Bo Li and Fei-Yue Wang are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (Email: bo.li@ia.ac.cn).

Yaobin Chen is with the Department of Electrical and Computer Engineering and also with the Transportation Active Safety Institute, Indiana University-Purdue University Indianapolis, Indianapolis, IN46202, USA (E-mail: ychen@iupui.edu).

In recent years, more and more part-based models have been studied and achieve success for human modeling in computer vision [5]–[7]. Our work is motivated by pose estimation algorithms that utilize the context among human parts to capture various human articulations. However, without considering specific characteristics of pedestrians in traffic environments, their human decomposition and part selection patterns may not be the optimal for pedestrians. For example, general pictorial models in human modeling can hardly cover all possible appearances of pedestrians. Moreover, most algorithms finely decompose human into arms and legs, which are not discriminative for detection tasks with cluttered background and unknown pedestrian number and size. Numerous part detection false alarms will be generated by objects with similar shapes. Therefore, we hope to investigate reasonable pedestrian decomposition structure and part models considering both pedestrian articulation varieties and discriminative abilities of parts.

In this pape, we present one novel part-based pedestrian detection algorithm for complex traffic surveillance environments. The algorithm is under the grammar-based framework, which is a novel structural model in computer vision. With a set of production rules, grammar models have strong and flexible representation ability for complex compositional structures, such as various pedestrian articulations. Main contributions of our work are following: i) According to inherent characteristics of pedestrians, one hierarchical And-Or grammar model is proposed. With the coarse-to-fine decomposition structure, traditional holistic pedestrian detection and part-based detection are simultaneously combined in one uniform framework. ii) Based on poselets [8], [9] that are originally proposed for person detection and attribute recognition, we improve them and propose clustered poselet models. An unsupervised learning algorithm is introduced to discover the representative part forms for specific part types. Some pedestrian characteristics are integrated into clustered poselets to make them suitable for traffic environments. iii) In detection, one effective bottom-up inference algorithm is presented to select consistent part activations and combine them into pedestrians. Except for the excellent detection performance in complex traffic environments, the main advantage of our algorithm is that we not only detect pedestrian locations, but also estimate detailed part configurations and attributes of pedestrians parts.

The remainder of this paper is organized as follows. Previous pedestrian detection algorithms are generally reviewed in Section II. Next, our hierarchical And-Or grammar for pedestrians is explained in Section III. Then, the detailed pedestrian detection algorithm is presented in Section IV. In

This is the author's manuscript of the article published in final edited form as:
Li, B., Chen, Y., & Wang, F. Y. (2015). Pedestrian Detection Based on Clustered Poselet Models and Hierarchical and #x2013;or Grammar. IEEE Transactions on Vehicular Technology, 64(4), 1435–1444. http://doi.org/10.1109/TVT.2014.2331314

Section V, the performance of our algorithm is evaluated by both public dataset and real traffic surveillance environments. Experimental results and their comprehensive discussions are also included in the same section. Finally, we make a conclusion of this paper and present the future work.

II. Related Works

Pedestrian detection is a long-standing problem in both computer vision and ITS. In literatures, large amount of algorithms and techniques have been proposed on pedestrian detection [10]. Some of them take advantage of multiple sensors such as stereo cameras and infrared cameras [11], [12]. In this section, we mainly concentrate on algorithms for monocular cameras. Conventional pedestrian detection algorithms are roughly classified into two categories: template-based and feature-based, which respectively correspond to generative and discriminative models.

For template-based methods, exact mathematical models for pedestrians are defined firstly. Then, the image is searched for template matching under the Bayesian framework. Generally, human shapes are modeled for detection. Discrete shape models denote a set of contour exemplars for edge matching [13], [14]. Continuous shape models are parametric contours, which can represent all likely poses theoretically. In [15] and [16], 3D models are defined to segment pedestrians given the foreground crowd regions. [17] uses mixtures of Bernoulli distribution and marked point process to represent pedestrians.

The most popular pedestrian detection approaches are feature-based methods, which rely on discriminative feature descriptors and classification models. Feature is the key factor. Earlier works extract statistics in local image blocks as shape features. [18] and [19] extract Haar wavelets to compute the local intensity differences. [20] proposes histogram of oriented gradient (HOG), which is one of the most popular feature descriptor for pedestrian detection. Many subsequent works are based on HOG and its variants [21], [22]. Some new shape-based features in high-level forms are proposed afterwards. [23] defines edgelet features utilizing a segment of lines or curves. [24] proposes shapelet automatically selecting gradients to form mid-level shape features.

Feature fusion is a straightforward strategy to provide complementary information and outperforms the performance of singular feature. [25] extracts feature set consisting of HOG, co-occurrence matrix and color frequency. In [26], HOG and local receptive field (LRF) is used to train classifiers respectively and classification results are combined finally. [27] presents the combination of Harr-like features, shapelets, shape context, and HOG. Then, [28] extends this idea and combines color self-similarity with motion features. Wang et al. [29] combines HOG with local binary patterns (LBP) texture feature. In [30], Haar-like features are computed over multiple channels, such as color, grayscale, and gradient channels. Thus, multiple feature types are integrated.

Compared with various feature strategies, classification models are relatively fixed. Most of work utilize support vector machine (SVM), boosting, or their variants as the learning framework.

Most of the above-mentioned methods mainly focus on holistic pedestrian detection. Recently, part-based approaches have been studied to deal with human articulation and posture variations. [5] defines the human parts as head-shoulder, torso, and legs. Their detection responses are combined with simple geometric relations. Many works extend pedestrians to humans in a broad sense for pose estimation. [6] uses basic semantic human parts, such as the torso and left upper arm. These parts are considered as nodes in a graphical model with pictorial structure. [7] continues this framework and improves the part appearance model with more discriminative features. Distinguished from these natural human partition ways, in [8] and [9], poselets are proposed as a kind of novel parts. They defines parts clustered in joint configuration space and appearance space.

In this paper, we improve the notion of poselets with a more compact representation. Then, it is embedded into the hierarchical grammar-based human parsing framework to detect pedestrians with various appearances and postures.

III. Hierarchical And-Or Grammar Model for Pedestrians

In this section, a hierarchical grammar model is defined to determine the pedestrian decomposition pattern. The proposed model is formulated as an And-Or graph [31], as shown in Fig. 1. The And-Or graph contains two kinds of nodes: and-nodes and or-nodes. And-nodes denote compositional relationships, which uniquely identify the specific combination of child nodes. Or-nodes denote reconfigurable relationships, which mean the current node can be in any state of the children. Children of one or-node are interchangeable with each other. Production rules of the grammar model are represented by node configurations in the graph. Owing to various combinations of two kinds of nodes, the grammar model can generate several objects with different appearances.

The And-Or graphical structure in Fig. 1 is designed according to general characteristics of pedestrians. From the root, pedestrians are decomposed hierarchically in a coarse-to-fine manner. In the second layer, pedestrians are classified by their body orientations from the standpoint of cameras. In this paper, we define four common viewpoint categories in traffic environments: front, back, left, and right. For each pedestrian that is specified one viewpoint, we semantically divided the full-body into three constituent part types: head-shoulder, torso, and legs. There are certain geometry constraints among these human parts.

In order to capture human articulations and simultaneously ensure the parts are discriminative for detection, we assign multiple forms for each part to represent the local deformation, rather than using small components, such as upper arms and lower legs. The relations are implemented with or-nodes, as shown in the fourth layer of the grammar model. The local articulation pattern is denoted by configurations of significant kinetic joints of human, which are called as keypoints in this paper. For head-shoulder part, we directly use one uniform configuration since mostly walking pedestrians keep their upper-body straight and head-shoulder variations are

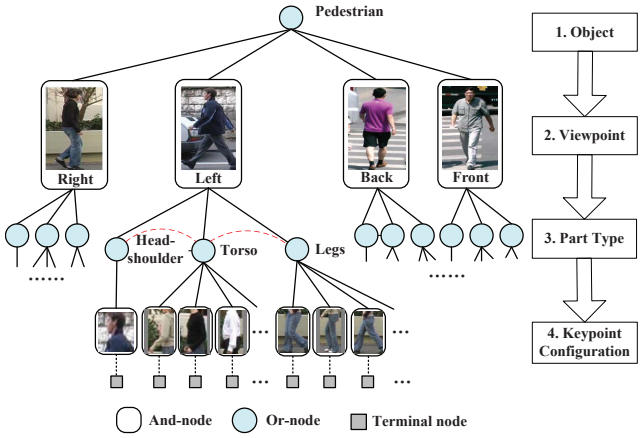

Fig. 1. The And-Or grammar for pedestrians. Rounded rectangles denote and-nodes, and rounded ones denotes or-nodes. Red dotted lines among human parts mean the geometry constraints. The figure on the right shows the decomposition hierarchy of pedestrians.

relatively small. However, for torso and legs, their local pose variations are obvious caused by arms and legs swing. Thus, we select certain typical keypoint configurations for parts instead of enumerating all their possible states. These part forms correspond to terminal nodes in the graphical structure.

The dictionary of grammar model consists of both terminal and non-terminal and-nodes. Appearance models for all items in the dictionary are trained, including holistic pedestrian models in specific viewpoints and their part models with certain part forms. Thus, both holistic-based and part-based pedestrian detection are combined in the unified framework to improve robustness of detection.

As a result, the pedestrian And-Or grammar model is defined with the 5-tuple in (1).

$$g = (V_g, E_{and}, E_{or}, R_g, T_S) \tag{1}$$

V_g denotes the set of nodes, E_{and} and E_{or} are edges for composition and selection respectively. R_g means geometry relations in the part composition, and T_S is the root. Through selecting certain forms from or-nodes in the bottom-up inference, one specific pedestrian example is obtained. Meanwhile, a unique parse graph pg is constructed with and-nodes encountered in inference.

According to [32], the And-Or grammar represents the probability distribution on the pg in a Bayesian framework, as shown in (2).

$$P(pg|I) \propto P(I|pg)P(pg) \tag{2}$$

I denotes the input image. The likelihood distribution $P(I|pg)$ is considered as the appearance model, which is related to the detection responses of parse nodes. Prior model $P(pg)$ evaluates the probability for a certain parse graph. We mainly consider geometric relations among parts as the prior. Thus, the detection problem is converted to human parsing as a maximum a posterior (MAP) estimation task.

IV. Proposed Method

A. Overview

Flowchart of the pedestrian detection algorithm is illustrated in Fig. 2. At first, the grammar dictionary is constructed by proposed clustered poselet models, which select representative keypoint configurations of human parts and use HOG-SVM to learn their appearance models (in Section III-B). Next, given the test image, pyramid HOG features are extracted at multiple scales. Activations for all clustered poselet filters are computed by convolution (in Section III-C). Finally, the bottom-up inference is performed to search the MAP solution and assemble these activations into the pedestrian full-body (in Section III-D). Through inference, pedestrian locations as well as part configurations are estimated. We give the detailed implementation and explanation for each procedure below.

B. Clustered poselet Models

As new notions of parts, poselets [8] show several advantages in person detection. They learns numerous common patterns for various human postures in the keypoint space. Poselet detectors embody keypoint locations, which help the subsequent pose estimation and human segmentation. Moreover, since related poselet image patches are extracted by keypoint alignments, this reduces the impact of unaligned training samples in detector learning.

However, limitations also exist if we directly apply poselets to pedestrian detection. For example, it is difficult to determine the appropriate number of poselets. Then, the randomness in seed window selection may lead to improper or various window locations at every runs. Especially when the pedestrian training set is relatively small, pattern variations become more obvious. Thus, acquired poselet models are not representative enough to cover all articulation changes.

Therefore, we propose clustered poselets, which are considered as one new poselet-based model for our pedestrian detection grammar. The main improvement is utilizing the unsupervised clustering algorithm to automatically discover the typical articulation patterns. Compared with original poselets, our clustered poselets have more compact representation, convenient implementation, definite semantic meaning, and sufficient discriminative ability.

Significant modifications on poselets are briefly summarized as three aspects. Firstly, the range of operated keypoints is explicitly defined according to the specific part type. This avoids the randomness in seed window selection. Secondly, the affinity propagation (AP) clustering algorithm [33] is introduced. It is the key operation to find clusters of local postures. Through AP clustering, the image patch matching and search strategy for original poselets is replaced with the segmentation in keypoint configuration space. This ensures that one image patch only belongs to a unique poselet category. Meanwhile, redundant and rate patterns will not be generated. Thus, extra postprocess of poselet selection can be removed. Thirdly, considering the straight characteristics of pedestrians, the rotation constraint is added to compute the similarity metric between two image patches.

The detailed training algorithm contains following steps:

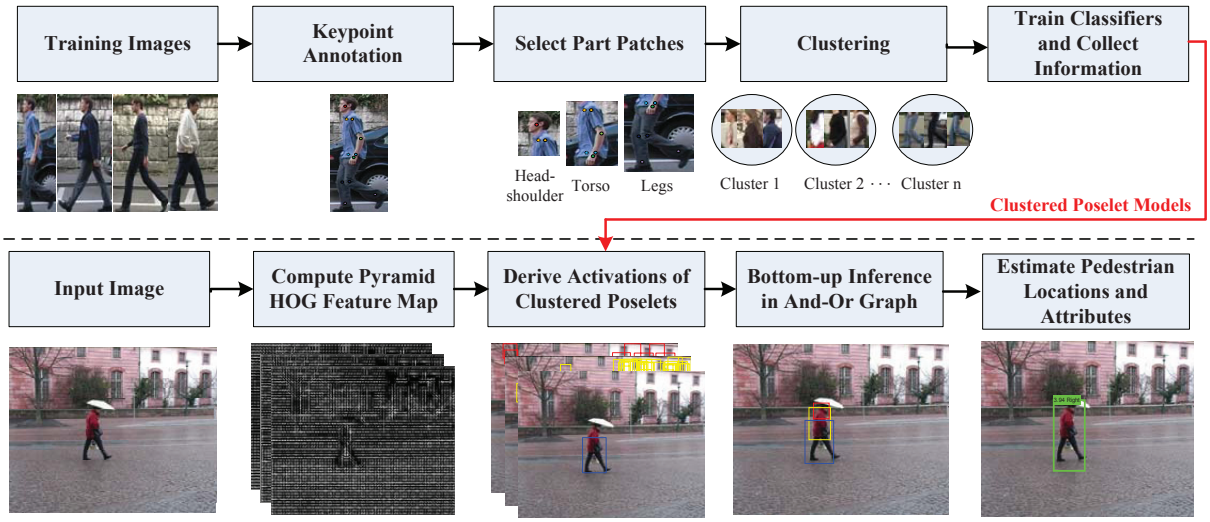

Fig. 2. The flowchart of the grammar-based pedestrian detection algorithm.

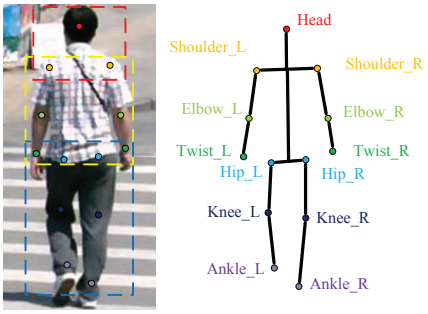

Fig. 3. The pedestrian keypoint definition. Keypoints with different types are denoted by different colors. Dotted rectangles in the left figure mean pedestrian parts (Red: head-shoulder, Yellow: torso, Blue: legs).

TABLE I
RANGE OF ACTIVE KEYPOINTS FOR DIFFERENT PART TYPES

Part Type	Active Keypoints
Head-shoulder	Head, Shoulder_L\R
Torso	Shoulder_L\R, Elbow_L\R, Wrist_L\R
Legs	Hip_L\R, Knee_L\R, Ankle_L\R

1. Keypoint annotation. At first, human keypoints are manually marked in training pedestrian images. In this paper, totally 13 keypoints are defined as the head and main joints in arms and legs. Fig. 3 shows the intuition of all keypoints. These keypoints are sufficient to determine the general posture of one pedestrian. We annotate the image coordination as well as the visible attribute for each keypoint. If the keypoint position can be roughly inferred while it is occluded, we define it as "nonvisible". Otherwise, it is annotated as "visible".

2. Determine the range of active keypoints. For a certain part type, we should specify the range of keypoints to be operated, which is called active keypoints. The detailed configuration of active keypoints for each part type is shown in Table I.

3. Compute the distance matrix. For two pedestrian images that are specified the same range of active keypoints, we measure their similarity according to keypoint attributes. The distance metric refers to [9], as in (3). P_1 and P_2 denote two active keypoint configurations, including the location, type, and the visible attribute. d_{proc} is the Procrustes distance, which evaluates the average displacement of keypoints after keypoint alignment by linear least square transformation $T(\theta)$. It is noted that the rotation constraint is added in the transformation to ensure that the alignment is made in an approximate upright precondition, as in (4). d_{vis} is the visibility distance, which means the intersection over visible keypoints. Thus, we can build a distance matrix for the training set to save the pairwise distance between keypoint configurations.

$$d(P_1, P_2) = d_{proc}(P_1, P_2) + \lambda d_{vis}(P_1, P_2) \quad (3)$$

$$d_{proc}(P_1, P_2) = \min \|P_1 - T(\theta) \cdot P_2\|, \text{s.t. } \theta \leq \theta_{max} \quad (4)$$

4. Clustering in keypoint configuration space. Based on the distance matrix, the AP clustering is performed to segment the keypoint configuration space to several representative and salient patterns.

AP is an exemplar-based clustering algorithm that takes similarity between pairs of data points as the input. AP clustering has some particular advantages for this task. Firstly, the cluster number is not required. Thus, keypoint configurations are clustered in a self-organizing way. Secondly, the cluster center is chosen as the most representative data point in the cluster. The exemplar just corresponds to the seed patch in our application. Thirdly, without random operations, AP clustering results keep stable with numerous operations. This removes the randomness of original poselets.

In our implementation, we define the affinity metric sim as the function of the distance value, as in (5). Larger distance means less affinity. The parameter γ controls the compact degree of the clustering. The number of clusters reduces with γ increases.

$$sim(P_1, P_2) = -[d(P_1, P_2)]^{\gamma} \quad (5)$$

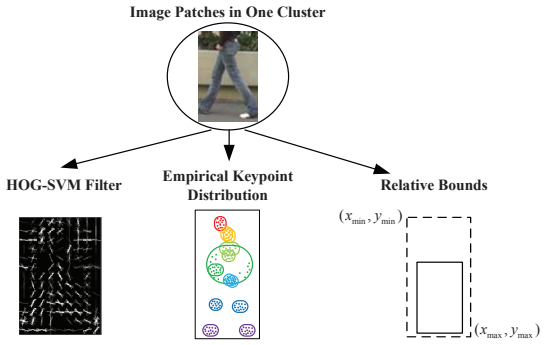

Fig. 4. The content of clustered poselet models.

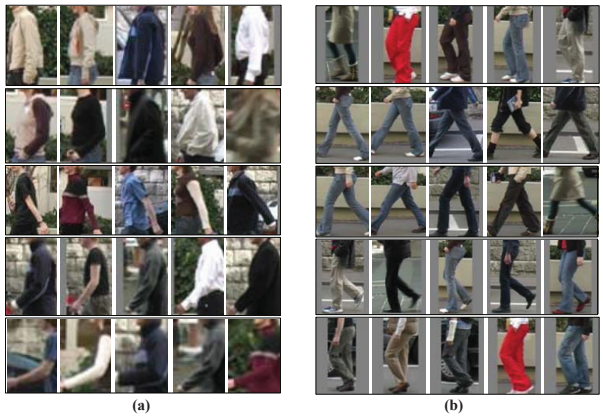

Fig. 5. Some AP clustering results for left-view pedestrians in TUD dataset. Each row denotes one cluster that represents a clustered poselet model for the part (a) torso and (b) legs.

Based on pairwise similarities that are computed from the distance matrix, several clusters are obtained through AP clustering. We remove small clusters and select each cluster center as the seed to collect other image patches in the same cluster. Thus, one cluster contains aligned image patches with similar local posture. Each cluster generates one clustered poselet model.

5. Train classifiers and collect necessary information on each cluster. We train one simple but effective discriminative HOG-SVM detector for each clustered poselet according to [20]. Except for SVM filters, other additional information are necessary to complete the detection, as shown in Fig. 4. Operations for the information collection are listed below:

(a) A logistic transformation on SVM classification responses is trained to obtain parameters A and b in Platt's scaling. The transformation converts the SVM score s_c into the detection probability s_p, as in (6).

$$s_p = \frac{1}{1 + \exp(A \cdot s_c + b)} \tag{6}$$

(b) Given each clustered poselet, the position statistics for each keypoint is modelled by a Gaussian distribution. The mean and variance of each keypoint distribution are saved.

(c) We collect the holistic pedestrian bounding boxes relative to each clustered poselet patch. The relative bounding box is represented by coordinations of two vertexes as $bbox = (x_{min}, y_{min}, x_{max}, y_{max})$. Statistics of the 4 variables are also fitted by Gaussian distributions.

Through above procedures, several clustered poselet models for different part types are automatically learned from the pedestrian training set. Taking the TUD-Pedestrian dataset as the example, AP clustering results for torso and legs are presented in Fig. 5. It is observed that image patches belong to the same category have the similar local articulation pattern. Thus, pedestrian posture varieties in continuous space are discretized with mixtures of clustered poselets. They are utilized as nodes that represent the part forms. Moreover, to complete the inference framework, full-body pedestrians are also considered as a special kind of clustered poselets that skip the AP clustering procedure. Thus, clustered poselets mentioned below also include holistic pedestrian models.

C. poselet Activation Derivation

Given the learned clustered poselets, the beginning of our detection approach is to find their strong activations in the input image. The classical sliding window paradigm is utilized to compute the convolution responses of clustered poselet filters on the multi-scale pyramid-HOG feature map. For each SVM classification score that is computed with linear weighted sum, the Platt's transformation [34] is conducted to normalize the score within $[0, 1]$. After the thresholding and non-maximal suppression (NMS) strategy, we derive activations for all clustered poselets with their locations and sizes. Next, we will rely on the bottom-up inference to combine these activations.

D. Bottom-up Inference

Based on the And-Or grammar, pedestrian detection is viewed as a human parsing problem. We aim to determine the optimal parse graph pg^* with MAP criterion in multiple locations. As in (7), given the hierarchical structure of the grammar model, the log-posterior can be formulated as the recursive scoring function of the root node v_0 [32]. Thus, we propose a bottom-up inference algorithm to effectively search the solution that maximize $s(v_0)$ and locate pedestrians with optimal part composition.

$$pg^* = \arg\max_{pg} P(pg|I) = \arg\max_{pg} s(v_0) \tag{7}$$

The algorithm is performed by selecting the optimal node configuration from bottom to top in the grammar model, which is shown in Fig. 6. According to different node types that encountered, the inference mainly comprises two significant operations. The or-node corresponds to the NMS, which selects the most reliable candidates in the neighborhood by scores. The and-node corresponds to the part combination, which finds consistent child nodes and aggregates their compositional scores. The part combination is the essence of the inference algorithm. Thus, we mainly present this algorithm in detail, which is given in Algorithm 1. The algorithm consists of three major processes: component clustering, score aggregation, and bounds prediction.

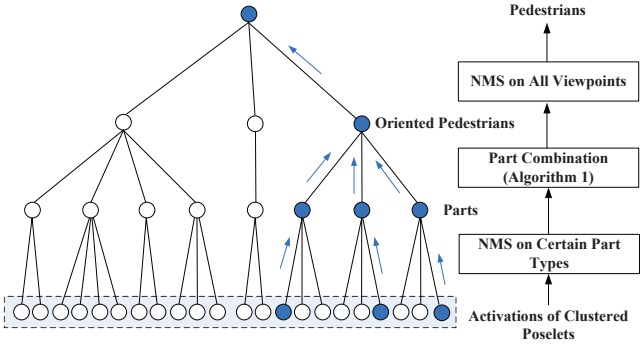

Fig. 6. The bottom-up inference framework. Dark nodes denote encountered nodes to reach the root. Operations for different layers are shown in the right diagram.

Algorithm 1 Part combination in the bottom-up inference.

Input:

The set of poselet activations, P_n;

The distance threshold for activation to cluster, D_{th};

Output:

Bounding boxes of holistic pedestrians, B_m;

1: Sort activations in P_n in a descending order with their detection scores s_p in (6), denoted by $P_n = \{a_1, a_2, ..., a_n\}$;
2: Initial the first cluster $C_1 = \{a_1\}$;
3: **for** each a_i in P_n **do**
4: Compute the distance between a_i and each cluster, and find the closest cluster C_j with distance $D(a_i, C_j)$;
5: **if** $D(a_i, C_j) > D_{th}$ **then**
6: Add a new cluster with activation a_i;
7: **else**
8: **if** $a_k \in C_j$ and a_k has the same part type with a_i **then**
9: Compute the new aggregation score $S^*(C_j)$ with (11) if a_k is replaced by a_i;
10: **if** $S^*(C_j) > S(C_j)$ **then**
11: Replace a_k with a_i in the cluster C_j;
12: **end if**
13: **else**
14: Merge a_i in the cluster C_j;
15: Compute the new cluster score $S(C_j)$ with (11).
16: **end if**
17: **end if**
18: **end for**
19: **for** each cluster C_i in all clusters **do**
20: Predict the bounding box $B_m(i)$ of the cluster C_i;
21: **end for**
22: **return** B_m;

Component clustering: Component clustering is to group poselet activations that belong to the same pedestrian. The approach is similar with that in [9], which is a form of greedy clustering starting from the poselet activation with the highest score. The metric for evaluating whether two poselet activations are consistent is to measure the similarity between their empirical keypoint distributions, which are obtained in clustered poselet training. If two activations correspond to the same person in the image, they should have uniform keypoint positions with small variance. The KL-divergence is introduced to measure the consistency between two poselet activations. Let N_i^k denotes the distributions of the k-th keypoint in poselet activation a_i, the distance between two poselet activations is formulated in (8). D_{KL} denotes the KL-divergence between Gaussian distribution N_i^k and N_j^k. K is the number of all keypoints.

$$D(a_i, a_j) = \frac{1}{K} \sum_{k=1}^{K} [D_{KL}(N_i^k || N_j^k) + D_{KL}(N_j^k || N_i^k)] \quad (8)$$

From the step 3 in Algorithm 1, the poselet activation is successively taken to compute the distance to each cluster. The distance of the activation a_i to the cluster C_j is estimated by averaging distances of a_i to all samples in C_j, as in (9). $|C_j|$ denotes the number of activations in C_j.

$$D(a_i, C_j) = \frac{1}{|C_j|} \sum_{a_j \in C_j} D(a_i, a_j) \quad (9)$$

If this distance is larger than a certain threshold, then we form a new cluster. Otherwise, the activation is merged in the cluster. We make the constraint that each poselet activation in the cluster should has the unique part type. Thus, if the new poselet activation has the same part type with anyone in the cluster, we will discuss whether to replace the old poselet by evaluating the effects that the new activation can bring.

Score aggregation: We define a score function to measure the compositional consistency of the poselet activations in the cluster. In the grammar model, the score is expressed as the log-posterior of a parse node v that represents the holistic pedestrian, as in (10).

$$s(v|I) = s_a(v|I) + s_g(v) + \sum_{v_i \in C(v)} s(v_i|I) \quad (10)$$

The aggregation score includes the appearance score $s_a(v|I)$, the geometry score $s_g(v)$, and scores from all children. $s_a(v|I)$ denotes the detection response of the holistic pedestrian model. $C(v)$ is the set of children of node v, which means clustered poselet activations with different part types. Thus, $s(v_i|I)$ only remains the appearance term and represents the detection response. The geometry score $s_g(v)$ is represented by displacement costs among parts. It is computed by the minus logarithm format of pairwise KL-divergence distance in (8). In this way, the score of cluster C_i can be expressed in (11), where w_g denotes the geometry cost weight. The score is computed by assembling appearance scores and measuring geometry costs. It measures how likely the cluster corresponds to a pedestrian compared with to be a false alarm.

$$S(C_i) = \sum_{a_i \in C_i} s_p(a_i) - w_g \sum_{a_i, a_j \in C_i} \log D(a_i, a_j) \quad (11)$$

The main circulation (step 3 to 18 in Algorithm 1) terminates when the last poselet activation has been treated to be included into one existing cluster or be a new cluster. Thus, we can get many clusters. Each cluster is a pedestrian hypothesis

Fig. 7. The example of the bottom-up inference results. (a) Component clustering and score aggregation results. (b) Bounds prediction results.

consisting of several poselet activations.

Bounds prediction: Finally, the exact bounding box of one pedestrian is predicted given all clustered poselet activations in the merged cluster. For each activation, we have fitted its Gaussian distribution for the relative holistic bounding box in training. Thus, the weighted average of mean values for pedestrian bounds distributions is computed from all poselet activations. The weights correspond to appearance scores of activations and statistical variance. The response of pedestrian hypothesis is defined as the cluster score.

After all viewpoint-specified holistic pedestrians are located through above part combination, the NMS is performed to select the optimal viewpoint at last. Thus, the root node is reached. The part configuration of each pedestrian can be recovered by the backtrack. Fig. 7 (a) shows results for poselet activation clustering and score aggregation. Then, after NMS and thresholding, the predicted bounding box is illustrated in Fig. 7(b).

V. Experiments

A. Experiments on TUD-Pedestrian

At first, we evaluate our approach on the public TUD-Pedestrian dataset, whose pedestrian images have relatively high resolution comparing with most of publicly available datasets. The dataset contains 250 images with 311 pedestrians. Corresponding training set contains 400 pedestrian images with resolution 200×100. Pedestrians in this dataset are mostly captured in a side view, where pedestrians are most likely to perform various articulation patterns in arms and legs.

In experiments, we use all training images to learn clustered poselet models. According to major viewpoints in the dataset, nodes in the second layer of the grammar model are limited as two types: left-side and right-side. For each viewpoint category, we set $\gamma = 5$ to obtain 5 clustered poselets in torso and 5 in legs. With the single model for holistic pedestrian and head-shoulder, totally 22 HOG-SVM filters are utilized in the grammar dictionary. Then, we use these filters to locate their activations in images and perform the bottom-up inference to evaluate our detection performance.

The receiver operating characteristic (ROC) curve is utilized as the evaluation criterion. ROC curves plot the true positive rate versus the false positive rate at various threshold settings. Fig. 8 shows the ROC curve for our algorithm in TUD-Pedestrian, as well as other results from literatures on the same dataset for comparison. It is observed that our approach significantly outperforms the classical HOG-SVM holistic pedestrian detector. Meanwhile, comparative performance is achieved comparing with detectors in [7] and [35], which finely decompose human into arms and legs to detect them in multiple orientations for articulation estimation. In contrast, our approach needs less computation complexity on part likelihood computation and inference. Fig. 9 shows some detection results of our approach. It is observed that not only precious bounding boxes of pedestrians are obtained, but also detailed pose attributes and part configurations are estimated.

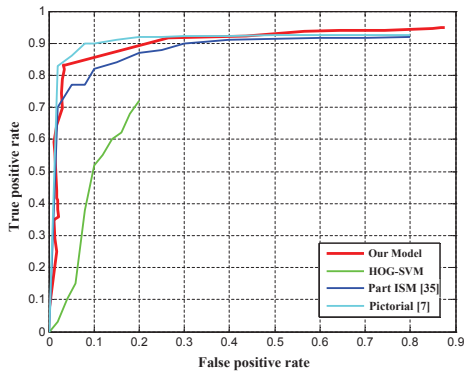

Fig. 8. The detection performance on the TUD-Pedestrian dataset.

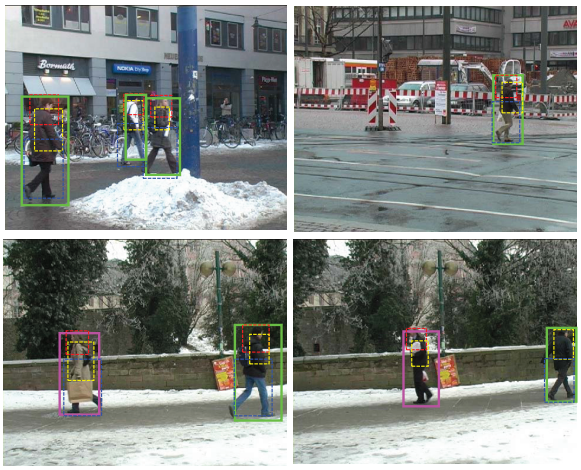

Fig. 9. Detection result examples on the TUD-Pedestrian dataset. Green and pink bounding boxes denote pedestrians with right-view and left-view respectively. Detected human parts are illustrated with dotted rectangles in different colors.

B. Experiments on Practical Traffic Environments

In order to evaluate our algorithm in practical traffic scenarios, we create a dataset collecting from real traffic environments with a high-resolution CCD camera for surveillance. The dataset is built from two videos that are respectively captured in two busy crossings. Both videos are taken at 8 fps with the resolution being 2592×1936. We subsample 300 frames from each video. Besides, the rectangular region of interest (ROI) is defined in test images to reduce false alarms and improve the time efficiency. Detailed information of each dataset is shown in Table II. Since the camera view is deep

and wide, object scales greatly change and pedestrian heights vary from 150 to 700 pixels. Unlike the TUD-Pedestrian dataset where most pedestrians have the view in left or right, pedestrians in our dataset mainly walking in front or back view. This dataset is much more challenging than other public dataset, since it contains a large number of pedestrians with various appearances and carrying different items. The occlusion occurs frequently, too. Meanwhile, the background is quite cluttered and complex with several disturbances, like vehicles, buildings, and plants.

In experiments, we randomly select 400 pedestrian images for each viewpoint from the captured video as the training samples. A software tool is designed to extract pedestrian images in video frames with fixed height-width ratio of 2. Then, pedestrian images are uniformly normalized into the resolution of 300×150. Since clustered poselets can deal with the alignment problem in detector learning, pedestrian positions in training images need not be strictly restricted. In implementation, the parameter γ is set to be 7. Then, we get 4 clusters in torso and 5 in legs for front-view pedestrians, as well as 5 torso patterns and 6 legs patterns for back-view pedestrians. Totally 22 HOG-SVM models are utilized. The prior of scene geometry is utilized in the post-process of detection to remove detection results that do not consistent with height constraints.

TABLE II
DETAILED INFORMATION OF OUR COLLECTED DATASET

Set	Image Number	Pedestrian Number	ROI Area	Pedestrian Height
Set 1	300	1870	1500×1500	200∼600 px
Set 2	300	1177	1500×1400	150∼700 px

We use the same evaluation criterion as the TUD-Pedestrian. ROC curves for both dataset are respectively illustrated in Fig. 11. Deformable part models (DPM) which achieve state-of-the-art accuracy on general object detection are brought in for comparison. We train DPM on the same training set with 8 components. The ROC comparison results clearly show that with the same false positive rate over 0.05, our approach achieves remarkable advantages on the detection true positive rate for both dataset. This indicates that clustered poselets and hierarchical grammar models are more flexible to capture diverse articulations of pedestrians compared with general DPM, which uniformly learn components without considering specific pedestrian characteristics. Fig. 10 shows some detection result examples and Fig. 12 shows detection details. It is observed that our approach can successfully detect pedestrians in this challenging traffic scenarios, even for some occluded pedestrians. Viewpoint attribute for each pedestrian is decided by the bottom-up inference. These attributes can be used for pedestrian behavior recognition in the future.

C. Discussions

Wrong and missed detection samples are collected to analyze limitations of our approach. We observe that most

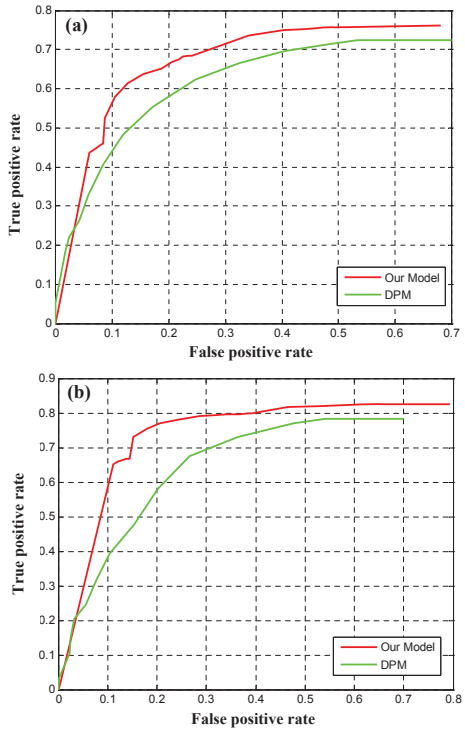

Fig. 11. The ROC curves for dataset of (a) Set 1 and (b) Set 2 respectively.

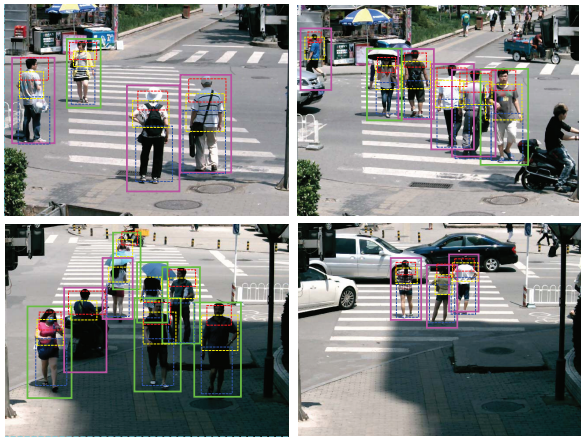

Fig. 12. Some detection details of our own collected dataset.

failed cases mainly correspond to three situations. Firstly, poor contrast areas caused by shadows or strong light make pedestrians hard to be identified. Secondly, pedestrians in distance have low resolution in images. Thus, they are obscure and generally cannot generate strong responses of poselet filters. Thirdly, pedestrians with severe occlusion, such as only one part is visible, are still difficult to be detected. In this case, the aggregation score of the pedestrian hypothesis is quite inferior being opposed to false alarms.

False alarms of our approach are usually produced by texture rich regions in the background. However, compared with holistic detectors, our false alarm rate are significantly reduced due to the part-based detection strategy. There are incorrect viewpoint estimations as well, which are usually caused by ambiguities in the image or some rare posture

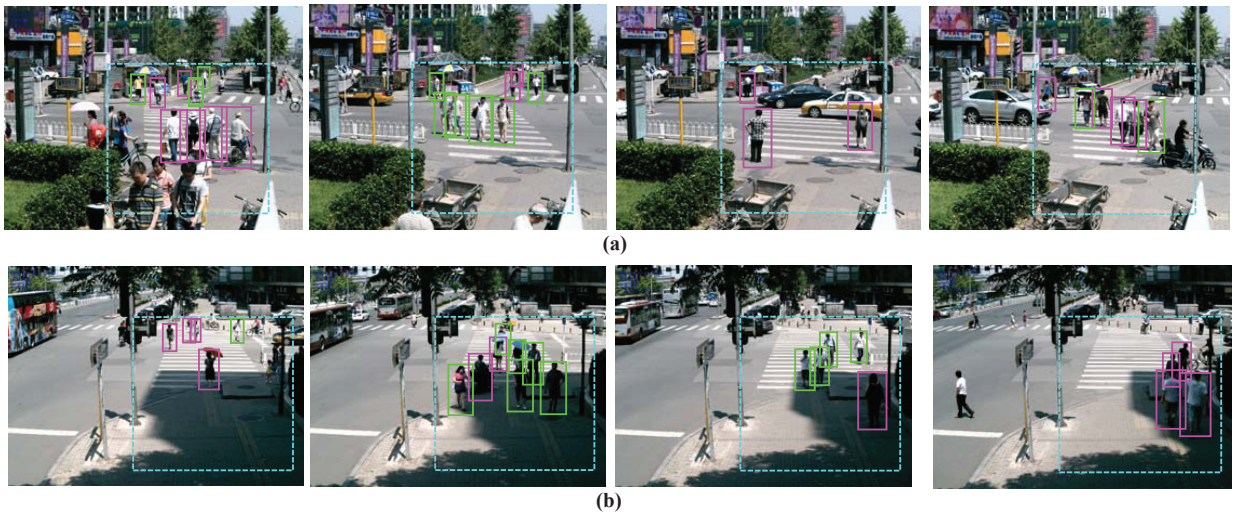

Fig. 10. Examples of detection results for our own collected dataset. Green and pink bounding boxes denote pedestrians with front-view and back-view respectively. Cyan dotted rectangles denote the detection ROI. (a) Results for Set 1. (b) Results for Set 2.

patterns that cannot be captured by the grammar model.

The shortcoming of our proposed approach is that the detection accuracy highly depends on the number of activated poselets for one pedestrian hypothesis. Generally, at least two parts should be detected to ensure the successful detection. We believe that our approach is suitable for detecting pedestrians in high-resolution images. More advanced imaging equipment and techniques are utilized to improve the resolution of pedestrians, better performance we could achieve.

Finally, we evaluate the processing time of our algorithm. Our experiments are conducted in an Intel Core i5-3210M CPU at 2.50GHz. The code has parts in C++ and others in Matlab. No parallel implementation or algorithm optimization strategies are used in experiments. With these settings, our model consumes much less time to learn DPM. The training of all clustered poselet filters costs about 1 hours in average, while DPM learning nearly need 10 hours. For testing, our algorithm runs in about 3 seconds per frame for our own collected dataset. Most of the time are consumed in computing the activations of clustered poselets.

VI. Conclusion

A novel part-based pedestrian detection algorithm is proposed in this paper. Pedestrians are uniformly decomposed with a hierarchical And-Or grammar. To acquire representative articulation patterns for human parts, we propose clustered poselet models, which combine poselets with AP clustering algorithm to generate terminal filters in And-Or grammar. After computing activations of these clustered poselets in the input image, an effective bottom-up inference algorithm is proposed to combine part activations to holistic pedestrians. Due to the detection framework, not only pedestrians are detected, but also detailed pose types and part configurations are determined. Experimental results demonstrate that our approach can achieve reliable and robust detection performance in the complex traffic environment. Compared with other pedestrian detection algorithms, our approach contains following properties: (1) Pedestrians with numerous appearances and postures can be captured in a uniform hierarchical decomposition model. (2) Clustered poselet models inherit advantages of poselets and automatically specify the pedestrian part forms from large amount of samples. (3) With the score aggregation strategy from multiple part detectors in inference, false alarms in complex traffic environments are significantly reduced. Meanwhile, some occluded pedestrians can be successfully located.

However, our approach still has limitations. In the future, more efforts can be made to improve the performance of constituent component detection. For example, we plan to combine more discriminative feature to increase the detection accuracy of pedestrian parts. Additionally, we can attempt to conduct parallel computing strategies with hardware support to enhance our execution efficiency.

Acknowledgment

This work is supported in part by key projects from National Natural Science Foundation of China (Grant No. 71232006, 61174172).

References

[1] F.-Y. Wang, "Parallel control and management for intelligent transportation systems:concepts, architectures, and applications," *IEEE Trans. Intelligent Transportation Systems*, 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 and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2010.

[3] H. Zhang, Y. Shi, and M. Liu, "h_∞ step tracking control for networked discrete-time nonlinear systems with integral and predictive actions," *IEEE Trans. Industrial Informatics*, vol. 9, no. 1, pp. 337–345, 2013.

[4] H. Zhang, X. Zhang, and J. Wang, "Robust gain-scheduling energy-to-peak control of vehicle lateral dynamics stabilisation," *Int. J. Vehicle Mechanics and Mobility*, vol. 52, no. 3, pp. 309–340, 2014.

[5] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proc. European Conf. Computer Vision*, 2004.

[6] P. Felzenszwalb, D. McAllester, and D. Ramannan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

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

[8] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Proc. IEEE Int. Conf. Computer Vision*, 2009.

[9] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *Proc. European Conf. Computer Vision*, 2010.

[10] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: survey and expriments," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[11] Y. Fang, K. Yamada, Y. Ninomiya, B. K. P. Horn, and I. Masaki, "A shape-independent method for pedestrian detection with far-infrared images," *IEEE Trans. Vehicular Technology*, vol. 53, no. 6, pp. 1679–1697, 2004.

[12] G. D. Nicolao, A. Ferrara, and L. Giacomini, "Onboard sensor-based collision risk assessment to improve pedestrians' safety," *IEEE Trans. Vehicular Technology*, vol. 56, no. 5, pp. 2405–2413, 2007.

[13] M. Enzweiler and D. M. Gavrila, "Integrated pedestrian classification and orientation estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[14] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human detection and segmentation," in *Proc. European Conf. Computer Vision*, 2008.

[15] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded envrionments," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1198–1211, 2008.

[16] L. Wang and N. H. C. Yung, "Three-dimensional model-based human detection in crowded scenes," *IEEE Trans. Intelligent Transportation Systems*, vol. 13, no. 2, pp. 691–713, 2012.

[17] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[18] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[19] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[20] N. Dalal and B. Triggs, "Histogram of oriented gradients for human detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[21] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006.

[22] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is effecient," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[23] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Proc. IEEE Int. Conf. Computer Vision*, 2005.

[24] G. M. P. Sabzmeydani, "Detecting pedestrians by learning shapelet features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[25] W. R. Schwarts, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *Proc. IEEE 12th Int. Conf. Computer Vision*, 2009.

[26] L. Oliveira, U. Nunes, and P. Peixoto, "On exploration of classifier ensemble synergism in pedestrian detection," *IEEE Trans. Intelligent Transportation Systems*, vol. 11, no. 12, pp. 16–27, 2010.

[27] C. Wojek and B. Schiele, "A performance evaluation of single and multi-feature people detection," in *Proc. DAGM Symp. Pattern Recognition*, 2008.

[28] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[29] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *Proc. 12th IEEE Int. Conf. Computer Vision*, 2009.

[30] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. British Machine Vision Conf.*, 2009.

[31] S. C. Zhu and D. Mumford, "A stochastic grammar of images," *Foundations and Trends in Computer, Graphics and Vision*, vol. 2, no. 4, pp. 259–362, 2006.

[32] B. Rothrock and S.-C. Zhu, "Human parsing using stochastic and-or grammars and rich appearances," in *Proc. IEEE Int. Conf. Computer Vision*, 2011.

[33] J. F. Brendan and D. Delbert, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[34] P. John C, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, 1999.

[35] M. Andriluka, S. Roth, and B. Schiele, "People tracking-by-detection and people-detection-by-tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

Bo Li received the B.S. degree in School of Automation, Southeast University, Nanjing, China, in 2009. She is currently working toward the Ph.D. degree in Control Theory and Control Engineering at the State Key Laboratory of Management and Control for Complex Systems, Chinese Academy of Sciences, Beijing, China. Her research interests include image processing, computer vision and their applications in intelligent transportation systems.

Yaobin Chen received his Ph.D. degree in Electrical Engineering from Rensselaer Polytechnic Institute, Troy, New York, in 1988. Dr. Chen is a senior member of IEEE, a member of SAE and ASEE. He is currently Professor and Chair of Electrical and Computer Engineering, and Director of the Transportation Active Safety Institute in the Purdue School of Engineering and Technology, Indiana University-Purdue University Indianapolis (IUPUI). Dr. Chen's current research interests include modeling, control, optimization, and simulation of advanced transportation and automotive systems, energy and power systems, computational intelligence and its applications.

Fei-Yue Wang (S'87-M'89-SM'94-F'03) received his Ph.D. in Computer and Systems Engineering from Rensselaer Polytechnic Institute, Troy, New York in 1990.

He joined the University of Arizona in 1990 and became a Professor and Director of the Robotics and Automation Lab and Program in Advanced Research for Complex Systems. In 1999, he found the Intelligent Control and Systems Engineering Center at the Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Oversea Chinese Talents Program, and in 2002, was appointed as the Director of the CAS Key Lab for Complex Systems and Intelligence Science. From 2006 to 2010, he was Vice President for research, education, and academic exchanges at the Institute of Automation, Chinese Academy of Sciences. Since 2005, he is the Dean of the School of Software Engineering, Xi'an Jiaotong University. In 2011, he became the State Specially Appointed Expert and the Founding Director of the State Key Laboratory of Management and Control for Complex Systems. His research is focused in social computing and parallel systems, and has published over 10 books and 300 papers in related areas over the past three decades.

Dr. Wang was the Editor-in-Chief of the IEEE Intelligent Systems in 2009 to 2012. He is currently the EiC of the IEEE Transactions on Intelligent Transportation Systems. He has served as General or Program Chair of more than 20 IEEE, INFORMS, ACM, ASME conferences. He was the President of IEEE ITS Society from 2005 to 2007, Chinese Association for Science and Technology (CAST, USA) in 2005, and the American Zhu Kezhen Education Foundation from 2007-2008. Dr. Wang is member of Sigma Xi, an Outstanding Scientist of ACM, and Fellow of IFAC, IEEE, INCOSE, ASME, and AAAS. Currently, he is the Vice President and Secretary General of Chinese Association of Automation. In 2007, he received the 2nd Class National Prize in Natural Sciences of China for his work in intelligent control and social computing. He received IEEE ITS Outstanding Application and Research Awards, IEEE Intelligence and Security Informatics Outstanding Research Award, and ASME MESA Achievement Award for his cumulative contribution to the field of mechatronic/embedded systems and applications, in 2009, 2012, and 2013, respectively.