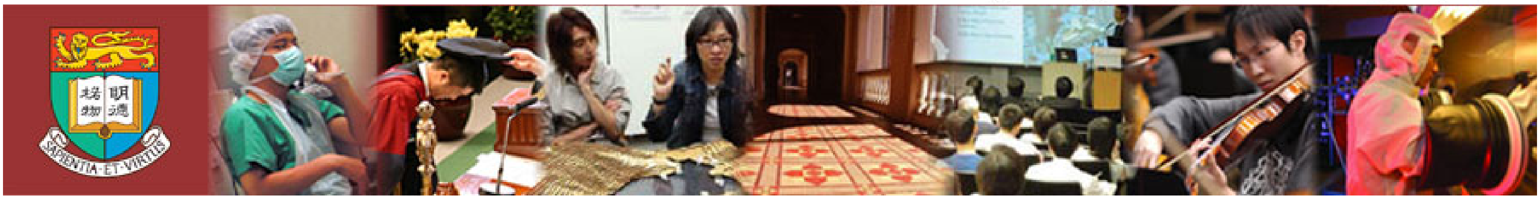


# Three-Dimensional Model-Based Human Detection in Crowded Scenes

Lu Wang and Nelson Hon Ching Yung, *Senior Member, IEEE*

***Abstract*—In this paper, the problem of human detection in crowded scenes is formulated as a maximum *a posteriori* problem, in which, given a set of candidates, predefined 3-D human shape models are matched with image evidence, provided by foreground extraction and probability of boundary, to estimate the human configuration. The optimal solution is obtained by decomposing the mutually related candidates into unoccluded and occluded ones in each iteration according to a graph description of the candidate relations and then only matching models for the unoccluded candidates. A candidate validation and rejection process based on minimum description length and local occlusion reasoning is carried out after each iteration of model matching. The advantage of the proposed optimization procedure is that its computational cost is much smaller than that of global optimization methods, while its performance is comparable to them. The proposed method achieves a detection rate of about 2% higher on a subset of images of the Caviar data set than the best result reported by previous works. We also demonstrate the performance of the proposed method using another challenging data set.**

***Index Terms*—Bayesian method, crowd segmentation, human detection, model-based method, video surveillance.**

## I. Introduction

AUTOMATED video surveillance of human objects has many applications in intelligent transportation systems. Monitoring pedestrian number and movement at road intersections provides useful information for the design of an adaptive signal control system [12], in which motor vehicle delay should be balanced with pedestrian delay in terms of their respective quantities. In addition, trajectory data obtained by human tracking are needed by the studies of pedestrian flows [31], which can be used for human traffic prediction, transportation infrastructure design, and evacuation control [29]. Furthermore, human behavior understanding would be helpful for fighting crime and terrorism in transit systems, such as airports, subway terminals, and bus stations [4].

Human detection, as a crucial step in the aforementioned applications, plays a vital role in automated human surveillance. However, human detection is not a trivial task. The appearance of human objects varies due to many factors, including viewpoint changes, lighting conditions, articulation, variations in clothing, poor figure-ground contrast, background clutter, etc. It becomes even more challenging in crowded scenarios where human objects visually occlude each other prevalently.

Recently, many methods have been proposed for crowd detection [2], [11], [14], [16], [26], [30], [34], [41], [42]. Most systems are based on 2-D template matching [2], [16] or 2-D discriminative training [11], [14], [26], [30], [34], [41]. Two-dimensional methods require a large amount of templates or training images to cover different postures and orientations. Furthermore, 2-D methods have the problem that they are not camera-angle invariant. If the camera parameters, e.g., swing angle and tilt angle, become significantly different from the assumed parameters, the system would fail, and new templates or training images need to be collected. On the contrary, a 3-D model-based approach does not have these problems. First, it is view invariant, i.e., given the camera parameters, the shape appearance of a human at any location within the image can be reasonably predicted. Second, as postures of 3-D models are easy to define, it does not need exemplar/training images. Third, a 3-D model-based approach can perform occlusion reasoning naturally. Given a model on the 2-D image, its position in the 3-D world can be obtained, and hence, its distance to the camera can be calculated. Then, based on the fact that the object nearer to the camera occludes the one farther away from the camera, the occlusion order between models can be determined. Reference [42] is based on 3-D human shape models. However, its proposed Markov chain Monte Carlo (MCMC) based optimization method requires a significant amount of computation. What is more, in [42], human shape information is not sufficiently utilized in the interior foreground region, resulting in a method that does not have enough discriminative power.

Considering the problems of the foregoing methods, this paper proposes a Bayesian 3-D model-based approach for human detection in crowded scenes, where computation and efficiency are balanced. In the proposed method, we assume that the camera is fixed and humans walk on a ground plane; therefore, moving areas can be extracted by background subtraction and camera calibration can be performed so that 2-D–3-D transformation relations can be obtained. In our approach, 3-D models representing human configurations are projected onto the image, and how good they are is measured by a posterior calculation, which is the product of a prior distribution and an image likelihood distribution. We use the prior distribution to model the distribution of each individual model's shape, restrict the different objects' mutual overlap, and require the configuration of object locations to be reasonable so that the real-world

Manuscript received November 29, 2010; revised August 13, 2011 and November 4, 2011; accepted November 13, 2011. Date of publication January 14, 2012; date of current version May 30, 2012. This work was supported in part by a Grant from the Research Grant Council of the Hong Kong Special Administrative Region, China, under Project HKU719608E and in part by the Postgraduate Studentship of the University of Hong Kong. The Associate Editor for this paper was J. Zhang.

L. Wang is with the College of Information Science and Engineering, Northeastern University, Shenyang 110004, China (e-mail: wanglu@ise.neu.edu.cn).

N. H. C. Yung is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: nyung@eee.hku.hk).

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

Digital Object Identifier 10.1109/TITS.2011.2179536

limitation is obeyed. The image likelihood distribution is used to measure how well the configurations are consistent with the foreground regions and the image's gradient information.

To deal with occlusion, we perform occlusion compensated model matching, which requires that the occluding objects of the object in consideration have already been known. To achieve this, we propose to estimate the human configuration by an iterative process of candidate selection, candidate model matching, and candidate validation and rejection. In candidate selection, we select those unoccluded candidates or the candidates whose occluding objects have been identified so that model matching can be correctly performed. To this end, the relationship among the multiple candidates is depicted using a directed graph, and candidates are selected based on this graph. For candidate validation and rejection, a minimum description length (MDL)-based method is first applied to reject those inferior candidates, and then model matching qualities and models' distances to the camera are compared to validate qualified candidates based on the argument that, generally, within a local neighborhood, true human objects have better model matching qualities than false objects, and unoccluded human objects have better model matching qualities than occluded objects.

The proposed human configuration estimation procedure balances between accuracy and computation. In each iteration, because candidates that are mutually dependent are considered simultaneously, wrong decisions that might be made by considering only one candidate at a time [2], [16], [41] can be avoided. On the other hand, as only a small portion of the candidates are considered, the computational cost is much lower than those methods that consider all the candidates at the same time [11], [26], [34], [42].

A problem with the 3-D model-based method is that the model parameter space is quite large. To solve the problem, we use a number of prototype 3-D models to approximate the whole model space, and for each scene, a 2-D model shape hierarchy is automatically constructed for efficient model matching.

The rest of this paper is organized as follows. In Section II, we review related works about human detection and crowd detection. Section III provides a theoretical formulation of the proposed method. In Section IV, we introduce the proposed optimization solution. In Section V, we demonstrate the performance of the system with experimental results on two data sets. Finally, we conclude this paper in Section VI.

## II. Related Work

To begin with, we briefly review techniques that aim at single human detection. Then, we review those methods that aim at a group of humans.

### *A. Single Human Detection*

From the feature detection point of view, there are global and local features that can be utilized for human detection. Global image features such as the image gradient, or the binarized edge, and image intensities are usually employed. References [10] and [16] proposed to use distance transform (DT) to match 2-D shape templates with image edges. In [22], image intensity is combined with shape to enrich the representation of pedestrians. A texture-based classifier, based on artificial neural networks, is then trained on the shape-normalized human foreground to help discrimination. Global features are computed quickly; however, 2-D template matching tends to produce false alarms in heavily cluttered areas, whereas the texture-based classifier is computationally demanding because the feature (image intensity) itself is not sparse enough.

More works use statistics of the basic image features in local image blocks. For instance, in [24], Haar wavelets are extracted to represent local intensity differences at various locations, scales, and orientations. In [36], use of AdaBoost cascades to automatically select the most discriminative Haar-like features was proposed, and the system is demonstrated to be quite efficient compared with some other popular methods [7]. In [5], densely calculated histograms of oriented gradients (HOGs) that are able to capture edge or gradient structures that are characteristic of the local shape and robust to location variability of body parts is introduced. HOG has been proven to be quite promising for human detection in many experimental studies [6], [7], [39], and many improvement works based on it have been proposed [15], [17], [25], [38], [43]. In [35], a new type of features that are based on the covariance of basic image features in blocks is proposed. Using LogitBoost classification on Riemannian manifolds, this method obtained a 5% higher detection rate on the INRIA data set than HOG. Some other shape-based features, such as edgelet [40], shaplet [27], local binary patterns (LBPs) [21], shape context (SC) [39], and adaptive contour features [9], have also been proposed for human detection. In general, local features have higher discriminative power and robustness than global features, which is paid for by higher computational complexity.

Some methods combine global information with local features to further improve detection performance and increase robustness. In [14], local appearance information from image patches is combined with global constraint from pedestrian's silhouette for robust pedestrian detection. In [25], global segmentation is used to verify object hypotheses generated by local feature-based classifiers. In [28], locally learned coarse shape information is combined with the global restriction of regularity and closure using Markov random field for simultaneous human detection and segmentation.

Most of the works described in the preceding paragraphs focus on holistic full body detection [5], [10], [14], [21], [22], [24], [25], [27], [28], [35], [36], [39]. To deal with posture variation and body part deformations, part-based methods have been proposed. Some methods [19], [20], [41] do the partition of the whole body based on semantic body parts, such as head, torso, and legs, and handle deformation by training part classifiers separately and assembling their responses. The drawback is that training data for each body part have to be manually labeled. Therefore, some approaches [8], [17] were proposed to select discriminative parts automatically through training.

However, without the explicit occlusion analysis, part-based methods are still sensitive to occlusion. References [41] and [38] proposed to use both full and part body detectors to cope with occlusions. In [38], full body detector based on HOG and LBP is first applied, and the classification score of each block is used to infer whether occlusion occurs and where it occurs. Then, Meanshift is used to segment the current scanning window into occluded and unoccluded regions. If occlusion is indicated with high likelihood, part detectors are applied on the unoccluded regions to do the final classification.

### B. Crowd Detection

For crowd detection, most methods use body part detectors to nominate a set of candidates and perform optimization on an objective function to select the best candidate subset as the final detection result. As the number of all the possible combinations of candidates is quite large, an efficient optimization method must be developed. In [2], [16], and [40], greedy methods for optimization are used. These methods assume an occlusion order of the candidates and decide to reject or accept a candidate sequentially from the candidate nearest to the camera to that farthest to the camera. They require the candidate nomination results to be very reliable; otherwise, the greedy methods tend to make wrong decisions because the assumed occlusion order may be incorrect. To alleviate the requirement for high quality candidate nomination, global optimization methods have been developed. In [26] and [34], optimization based on expectation maximization (EM) is used, whereas [42] and [11] use MCMC. The next paragraph will review some major works about crowd detection.

In [14], occlusion reasoning is performed among human hypotheses using MDL. However, as the detector is designed to detect full body only, it is unlikely to work well under crowded surveillance scenarios where only the upper body is visible for many human objects. In [41], the responses of part detectors are combined to form a joint likelihood model of human. In [16], a hierarchical part-template matching is proposed to handle partial occlusions. However, as we know that template matching is not as discriminative as learning-based detectors, the greedy optimization algorithm proposed in [16] may not be sufficient to give a satisfactory detection result. To improve the efficiency of template matching, [2] proposed to use contour integration, which is calculated from integral images constructed by oriented string scans, for human detection. To increase the reliability of candidate nomination, an SC-based human detector is also proposed. Combining the two detectors, the final configuration is obtained in a greedy manner. In [42], use of 3-D human shape models for crowd segmentation is proposed, and MCMC is used to search the solution space. In [26], use of EM to assign image features to human candidates, in which certainty is propagated from regions of low ambiguity to those of high ambiguity, is proposed. Akin to [26], image patches are assigned to candidates using EM in [34]. The difference is that occlusion reasoning is explicitly performed in the M-step in [34]. Given part detection results, [30] used bilattice-based logical reasoning to infer the optimal configuration.

## III. Problem Formulation

The goal of this paper is to find the optimal configuration of human objects given a set of candidates (see Section IV-A for details of candidate nomination), where occlusion may exist.

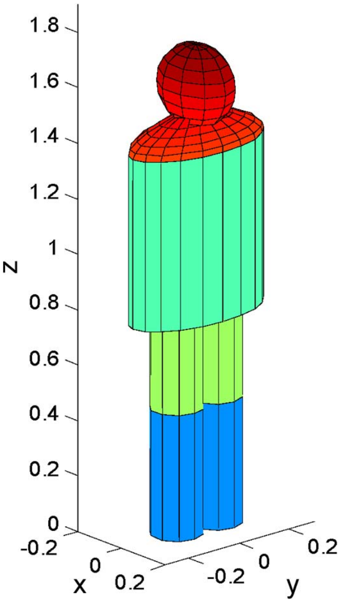


Fig. 1. Proposed 3-D human shape model.

We formulate it as a maximum *a posteriori* problem such that the optimal solution $\theta^*$ is given by

$$(\theta^*) = \arg\max_{\theta} P(\theta|I) \tag{1}$$

where $\theta$ consists of the number of human objects $n$ and their corresponding models $(m_i,\ i = 1,\ldots,n)$, and $I$ is the image observation. Each $m_i$ contains the shape information, such as height, posture, orientation, and position information. According to the Bayesian rule, (1) can be decomposed into a prior term and a joint likelihood term

$$P(\theta|I) = P(\theta)P(I|\theta)/P(I) \propto P(\theta)P(I|\theta). \tag{2}$$

In the following, we first define the 3-D human shape model and then define the prior $P(\theta)$ and the likelihood $P(I|\theta)$.

### A. Three-Dimensional Human Shape Model

The 3-D human shape model we propose consists of seven parts: the head (modeled by an ellipsoid), the shoulder (modeled by the upper half of an ellipsoid), the torso (modeled by a cylinder), and the left/right thigh/calf (each modeled by a cylinder), as depicted in Fig. 1. The dimension of the prototype model is of the average dimension of 50% man and 50% woman presented in [33], and it is scaled linearly to generate models of different heights. To restrict the search space, ten typical leg configurations of a walking cycle are selected for model matching according to the normal walking patterns of human beings [23]. The ten configurations correspond to the five typical walking postures shown in Fig. 2, whose number is doubled by differentiating the front leg in the left or right. To further consider different walking speeds, the average hip and knee rotation degrees for different postures are also increased and decreased 25%, respectively, by assuming local linearity in the model shape space. Therefore, the model has 30 postures in total.

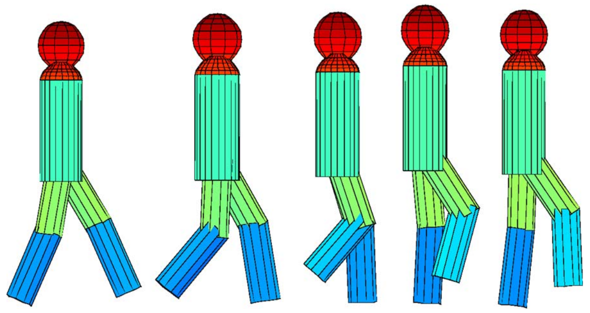


Fig. 2. Various postures of the 3-D human models.

In addition, the model is allowed to have 12 orientations (0°, ±30°, ±60°, ±90°, ±120°, ±150°, and 180°, with 0° corresponding to human facing the camera) and four scales (corresponding to heights of 1.55, 1.65, 1.75, and 1.85 m, respectively). The horizontal head torso deviation is defined in the image space, and the discretization step is set to be $\max(2, [W_{\text{head}}/6])$, where $W_{\text{head}}$ is the image head width.

### B. Prior Distribution

We assume that the prior term in (2) is the product of the prior probabilities of each individual model $m_i$ and is defined as

$$P(\theta) = \prod_{i=1}^{n} P_{\text{penal}}(m_i) P_{\text{pos}}(m_i) P_{\text{dev}}(m_i) P_{\text{height}}(m_i). \quad (3)$$

$P_{\text{penal}}(m_i)$ gives each model $m_i$ in $\theta$ a penalty according to its real world position $\mathbf{L}_i$, which in fact controls the minimum visible area of the model and hence avoiding the number of models $n$ to be unreasonably large. $P_{\text{penal}}(m_i)$ is defined as

$$P_{\text{penal}}(m_i) = \tau_1 \exp\left(-\alpha(\mathbf{L}_i)\right) \quad (4)$$

where $\alpha(\mathbf{L}_i)$ represents the minimum visible area of the model and is tunable: If it is set larger, then less false alarms would be produced whereas more missed detections may occur. We set its default value to be the head area of a standard human model located at position $\mathbf{L}_i$. $\tau_1$ is a normalization constant that makes $P_{\text{penal}}(m_i)$ a probabilistic distribution function.

$P_{\text{pos}}(m_i)$ is the prior probability about $m_i$'s real world position relative to the others (denoted as $-i$). It represents our prior knowledge that two persons must keep a certain distance away from each other in the real world and is given by

$$P_{\text{pos}}(m_i) = P(\mathbf{L}_i|\mathbf{L}_{-i}) = \tau_2 f\left(\min_{j\in 1,\ldots,n, j\neq i} |\mathbf{L}_i - \mathbf{L}_j|\right)$$

$$f(d) = \begin{cases} d/d_{\min}, & \text{if } d \leq d_{\min} \\ 1, & \text{if } d > d_{\min} \end{cases} \quad (5)$$

where $d_{\min}$ is the minimum distance required for any two human objects and is set to be 0.2 m in this paper.

The third and fourth terms are about $m_i$ itself. $P_{\text{dev}}(m_i)$ limits the head's horizontal deviation from the torso, describing our common sense that human head tends to lean forward, but not always leans left or right, and seldom leans backward. The best distribution of $P_{\text{dev}}(m_i)$ should be learned from ground truth data. However, because of the large amount of work required by manual labeling, we only approximate it in this paper. Suppose $(x_h, y_h, z_h)$ is the head centroid of $m_i$, and $(x_t, y_t, z_t)$ is the torso centroid. With reference to the 3-D coordinate system depicted in Fig. 1, $P_{\text{dev}}(m_i)$ is defined as in (6), shown at the bottom of the page, where $R_{\text{head}}$ is the radius of $m_i$'s head in the real world, and $\sigma_x = 2\sigma_y = 4R_{\text{head}}$.

The prior about the model height $P_{\text{height}}(m_i)$ is used to penalize very short or very tall heights and is approximated by

$$P_{\text{height}}(m_i) = \tau_4 \frac{1}{1 + \left|\frac{H_{m_i} - a_3}{a_1}\right|^{2a_2}} \quad (7)$$

where $H_{m_i}$ represents the real world height of model $m_i$. In our experiment, parameters of the bell function are selected such that $P_{\text{height}}(m_i)$ for $H_{m_i} = 1.7$ m is 1.0 and for $H_{m_i} =$ 1.5 m or 1.9 m is 0.95.

### C. Image Likelihood

Assuming the pixels are independent, the likelihood is defined as

$$P(I|\theta) = \prod_{p\in I_f} P(p|\theta) = \exp\left(-\sum_{p\in I_f}(1 - L_s(p)\right)$$
$$= \exp\left(\sum_{p\in I_f} L_s(p) - \text{area}(I_f)\right) \quad (8)$$

where $I_f$ is the foreground mask, and $L_s(p)$ is the shape likelihood obtained by matching the visible part of the boundary of $m_i$ with the foreground edge if $p$ belongs to the visible part of $m_i$; otherwise, $L_s\ (p) = 0$.

## IV. Proposed Solution

Given the problem formulation, in this section, we will introduce the details of the proposed solution for optimization. As shown in Fig. 3, given a video sequence, first, we calculate the camera parameters [13]. Then, for each frame, we extract the foreground [37]. After that, human candidates are nominated by a head detector and a foot detector, respectively. An iterative optimization procedure is then followed to find the optimal human

$$P_{\text{dev}}(m_i) = \begin{cases} \tau_3 \exp\left[-\frac{(x_h - x_t)^2}{\sigma_x^2}\right] \exp\left[-\frac{(y_h - y_t)^2}{\sigma_y^2}\right], & \text{if } 0 \leq x_h - x_t \leq 2R_{\text{head}} \\ & \text{and } |y_h - y_t| \leq R_{\text{head}} \\ \tau_3 \exp\left(-\frac{1}{2}\right), & \text{otherwise} \end{cases} \quad (6)$$

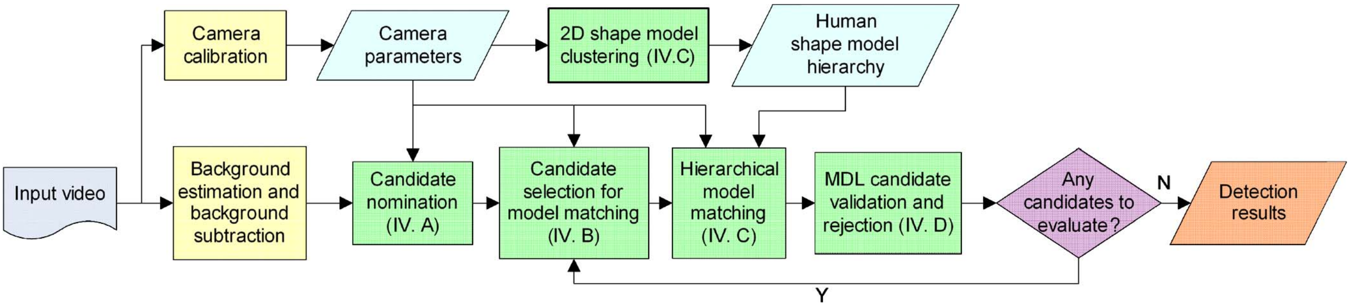

Camera calibration
Camera parameters
2D shape model clustering (IV.C)
Human shape model hierarchy
Input video
Background estimation and background subtraction
Candidate nomination (IV. A)
Candidate selection for model matching (IV. B)
Hierarchical model matching (IV. C)
MDL candidate validation and rejection (IV. D)
Any candidates to evaluate?
N
Detection results
Y

Fig. 3. Overview of the proposed solution.

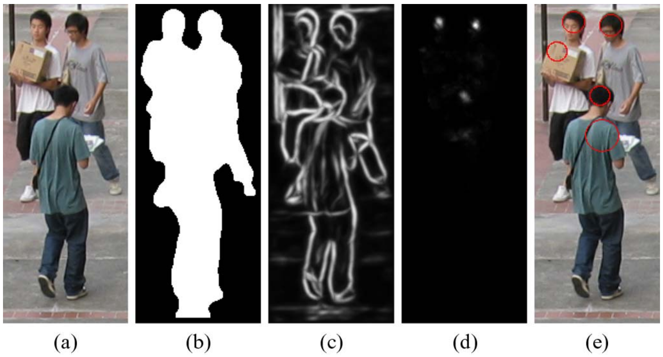


Fig. 4. HC detection. (a) Input image. (b) Foreground mask. (c) $pb$ map. (d) Head detection response $R(x, y)$. (e) Detected circles overlaid on the input image.

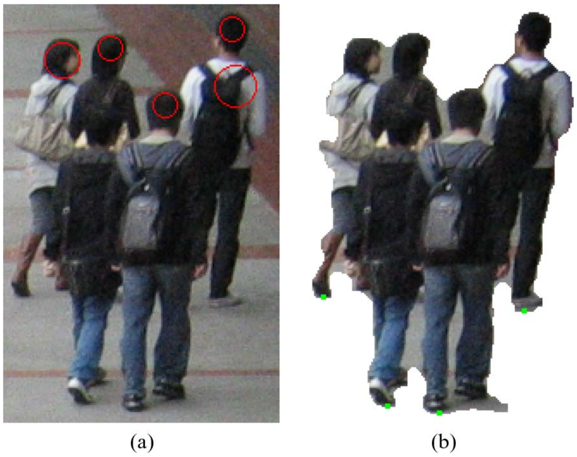


Fig. 5. Complementary characteristic of HCs and FCs. (a) HCs (red circles) and (b) FCs (green dots) of the same image. One human object is not nominated by the HC detection due to the low contrast, whereas the LE detection is able to nominate that human object's foot. Other body parts, i.e., shoulders, torso or legs, are not as prominent as head and feet in this image.

configuration. In each iteration, only a group of candidates that are either unoccluded or whose occluding objects have already been identified are selected for hierarchical model matching, and the results are fed into a candidate validation and rejection step. The iteration ends when all the candidates have been either validated or rejected.

### A. *Candidate Nomination*

From our observation, the most reliable feature of a human is the head. Therefore, we use an upper semicircle detector to nominate the head candidates (HC). The method used [3] is a Hough-like circle detector, in which each boundary element spreads its vote, modulated by the edge magnitude, into $(x_c, y_c, r)$ that represents the circle's center and radius. The directional filter we use is the probability of boundary $(pb)$ [18], which effectively removes the edge responses of textures and thus reduces the number of false positive detections. The radii set of the upper semicircle detection $Rad(x, y)$ for each image position $(x, y)$ is determined by projecting two spheres, representing the lower and upper bounds of real-world human head size, respectively, onto the position $(x, y)$ of the image and taking half of the projections' widths as the minimum radius and maximum radius, and the scale factor is set to be 1.1, which is sufficiently small for detecting all possible HCs.

Having the upper semicircle detection response of each radius in $Rad(x, y)$, the maximum of the responses of different radii forms the final response $R(x, y)$. Then, the local maxima of $R$ are thresholded to obtain the HC set. The threshold is conservatively set so as to avoid missed detections. Redundant candidates are then removed: if the center of one circle is inside another circle, then the one with the weaker response is discarded. An example of HC detection is shown in Fig. 4.

However, it is possible that head detection fails because the object being occluded has similar color as the head. To deal with this situation, we also detect the lower extrema (LE) on the boundary of $I_f$ as foot candidates (FCs). The complementary characteristic of HCs and FCs is depicted in Fig. 5. The HCs and FCs compose our candidate set $C_{\text{total}}$.

### B. *Candidate Selection for Model Matching*

We aim at detecting human objects in crowd by model matching. Due to the existence of occlusion, it is required that the object to be matched is either unoccluded or its occluding objects haven been detected and their corresponding image area has been identified, so that model matching can be performed only for the visible part of the object. To this end, we propose an iterative optimization process in which, in each iteration, candidates that are likely to be unoccluded or candidates whose occluding candidates are likely to have already been detected are selected, and then model matching and candidate validation/rejection are performed on them to make decisions.

*1) Mutual Dependency Description:* For any two candidates, they can have three relations: 1) They are far away

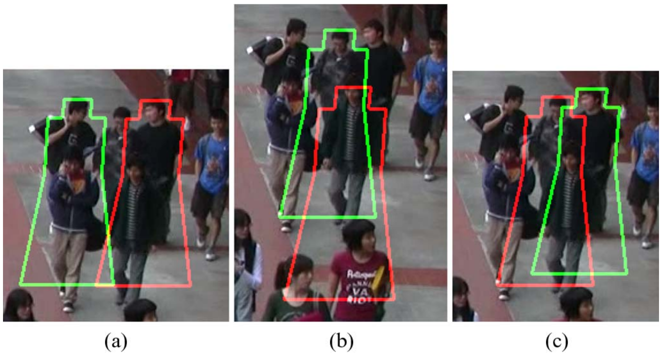


Fig. 6. Candidates' relationship. (a) Only lower bodies of their BPs intersect, insignificant overlap, and the two candidates are considered to be not related. (b) Significant overlap and significant vertical distance of their head tops; the lower one is the higher one's occluding HC. (c) Significant overlap with small vertical distance of their head tops, and the two candidates are required to do model matching simultaneously.

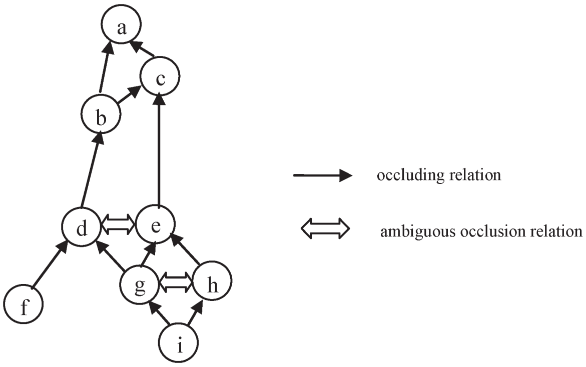


Fig. 7. Graph G, which depicts candidates' relationship. A candidate can do model matching only when all its occluding candidates have been matched. For ambiguous occlusion order, once one of the two candidates is triggered for model matching, the other candidate and all its occluding candidates are triggered as well.

from each other, i.e., they are not mutually dependent; 2) they are near to each other, and one candidate's image position is significantly lower than the other candidate, i.e., the lower candidate occludes the higher candidate; and 3) the two candidates are near each other, and they are of the left and right relationship, i.e., their occlusion order is ambiguous.

To evaluate the candidates' position relationship, we define bounding polygons (BP) for HCs. An HC's BP is a polygon that approximately defines the maximum spatial extents of human models with the head top position defined by the HC. A BP is composed of three parts: 1) head (rectangle, width: 0.3 m, height: 0.2 m); 2) torso (rectangle, width: 0.6 m, height: 0.6 m); and 3) lower body (trapezium, upper bottom: 0.6 m, lower bottom: 1.0 m, height: 1.2 m). The BP's projection on an image is obtained using the camera parameters by assuming that the BP faces the camera. For FCs, due to the wide range of possible foot positions relative to the human object's actual position, we can hardly define a BP for an FC from an LE. Therefore, we do model matching for every FC (see Section IV-C for details of model matching) and take the head of the best matched model as an HC. Then, we can define BPs for FCs similar to HCs.

For the BPs of two candidates $A$ and $B$: 1) If they do not intersect or only their lower body parts intersect [as depicted in Fig. 6(a)], i.e., the intersection is not significant, $A$ and $B$ are not related; 2) otherwise, if the torso of $A$ intersects with $B$, and $A$'s head top is either inside B [as depicted in Fig. 6(b)] or lower than B's torso top, $A$ is occluding $B$; 3) otherwise, both could be occluding the other [as depicted in Fig. 6(c)].

We describe candidates' dependency using a directed graph $G$, as illustrated in Fig. 7, where a node represents a candidate, and an edge represents the dependency between two candidates. The edge is defined as follows: 1) If $A$ is occluding $B$, then there is a single-directional edge that starts from $A$ and ends at $B$, and $A$ is called the occluding candidate of $B$, meaning that $B$ is eligible for matching only when $A$ has done the model matching; 2) if $A$ and $B$'s occlusion order is ambiguous, there is a bidirectional edge between $A$ and $B$, meaning $A$ and $B$ must do model matching simultaneously.

*2) Candidate Selection Based on Mutual Dependency and Distance to Camera:* If the bottom line of a candidate's BP does not intersect with any foreground pixel, it is likely that the human object that corresponds to this candidate is unoccluded. We call this kind of candidates unoccluded candidates.

In the first iteration, the candidate that is nearest to the camera is selected for model matching. Then, all the other unoccluded candidates whose BPs intersect with the matched candidates' models are also selected to do model matching. The selection repeats until there are no more candidates satisfying this requirement.

In the following iterations, the candidates intersected with the validated models and meanwhile with all their occluding candidates in $G$ having been matched are selected for model matching. Candidate $c_1$ and all its occluding candidates are selected if candidate $c_2$ with ambiguous occlusion relationship with $c_1$ has been selected. For any unmatched unoccluded candidate $c_i$, if there is one matched model whose distance to the camera is larger than $c_i$'s distance to the camera, $c_i$ is selected for matching. As an HC's distance to the camera is unknown, for an unoccluded HC, we take the lowest pixel of its BP's intersection with $I_f$ as its position. The selection ends when there are no more candidates satisfying the requirement.

In case that no candidates are selected in a new iteration and there are still unmatched candidates, the candidate that is nearest to the camera is selected.

### C. Model Matching

*1) Model Matching Likelihood:* Given a selected candidate $c_i$, if it is an HC, its head top position is assumed to be the corresponding human's head top position; if it is an FC, we search in the vicinity of the FC to find the most likely position.

The matching is measured by both the model's region coverage with the remaining mask $I_{\text{rem}}$, which is calculated by removing the region occupied by the validated models $I_{\text{occ}}$ from $I_f$, and the model boundary's matching with the *pb* map, which is not treated with the nonmaximum suppression and thus similar to the DT performed on an edge map. Formally, the matching likelihood $L(M_j)$ is calculated by

$$L(M_j) = L_r(M_j) L_s(M_j) \quad (9)$$

where $L_r(M_j)$ is the region matching likelihood, and $L_s(M_j)$ is the shape matching likelihood. $L_r(M_j)$ is defined as

$$L_r(M_j) = \frac{\text{area}(M_j \cap I_{\text{rem}}) - w \cdot \text{area}\left(M_j \cap (1 - I_f \cup I_{\text{occ}})\right)}{\text{area}(I_f)}. \tag{10}$$

In the dividend, the minuend encourages a larger foreground area to be explained by the model, whereas the subtrahend penalizes the model regions falling out of both $I_f$ and $I_{\text{occ}}$ and prevents unreasonable models to be selected. $w$, ranging from 0 to 1, is the penalty parameter whose value depends on the accuracy of the foreground extraction: the larger the false negative rate of the foreground extraction is, the smaller $w$ is, meaning that the foreground information is not reliable.

The shape matching likelihood $L_s(M_j)$ is defined as

$$L_s(M_j) = \frac{1}{|Mb_{j,\text{rem}}|} \sum_{k \in Mb_{j,\text{rem}}} pb(k) \left\langle \mathbf{O}_{pb}(k) \cdot \mathbf{O}_{Mb_j}(k) \right\rangle$$

$$Mb_{j,\text{rem}} = Mb_j \cap (1 - I_{\text{occ}}) \tag{11}$$

where $Mb_j$ is the boundary image of model $M_j$ and $Mb_{j,\text{rem}}$ is its visible part; $\mathbf{O}$ represents the orientation vector of the boundary point; and $L_s(M_j)$ is the average $pb$ value of points on $Mb_{j,\text{rem}}$ weighted by the consistency between the orientation of the model boundary and the orientation of the corresponding $pb$. Usually, the orientation's range is [0, 180°). To avoid some obvious false positives, such as the human-like shape formed by two persons walking together, we differentiate the orientations $o_1$ and $o_1 + 180°$ in the vicinity of the mask boundaries.

*2) Probabilistic Foreground Pixels Assignment to HCs:* It is possible that a model is attracted by strong edges that are not from the boundary of the corresponding human object, as demonstrated in Fig. 8(a). To avoid such situations, we introduce another measure, called coverage probability, during the model matching. Each foreground pixel is probabilistically assigned to an HC that may cover it. Then, a penalty would be given to a candidate model of the HC if the model does not cover the pixels the HC can cover.

As shown in Fig. 9, for each BP of an HC, a line, named central line, is drawn from the head top center to the bottom center of the upper body. Then, the distance $dc$ of each pixel $p$ inside the BP to the nearest pixel of the central line is calculated. The probability of each pixel $p$ within the BP being covered by the corresponding HC $c$ is calculated as

$$P_{\text{cover}}(p, c) = 1 - \sqrt{dc(p) / \max_{q \in BP(c)} (dc(q))} \tag{12}$$

which expresses the intuition that the point that is farther away from the central line has smaller probability to be covered by the HC. As a foreground pixel can be simultaneously covered by multiple HCs, its probability of being covered by a candidate $c_i$ is rectified as

$$\tilde{P}_{\text{cover}}(p, c_i) = \frac{P_{\text{cover}}(p, c_i)}{\max\left(1, \sum\limits_{\forall c_j, p \in BP(c_j)} P_{\text{cover}}(p, c_j)\right)}. \tag{13}$$

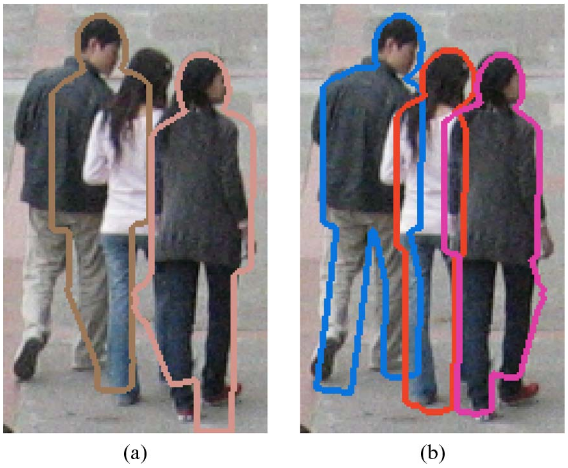


Fig. 8. Effect of foreground region assignment. (a) Without specifying which part of the foreground an HC should most likely explain, the model is attracted by strong boundaries that do not correspond to that human object. (b) After enforcing the coverage probability, the detection is more accurate.

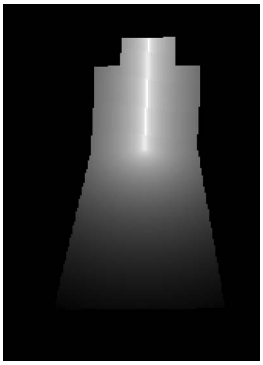


Fig. 9. Probability of the foreground pixels being covered by an HC (lighter means higher probability).

Then, the coverage probability of a model $M_j$ for a candidate $c_i$ can be calculated as

$$P_{\text{cov}}(c_i, M_j) = \frac{\sum\limits_{\mathbf{p} \in M_j} \tilde{P}_{\text{cover}}(\mathbf{p}, c_i)}{\sum\limits_{\mathbf{p} \in BP(c_i)} \tilde{P}_{\text{cover}}(\mathbf{p}, c_i)}. \tag{14}$$

With the coverage probability and given the already validated candidates $C_{\text{val}}$, the best matched model $m_i$ for $c_i$ is then selected as the one that results in the maximum increase of the posterior as follows:

$$m_i = \arg\max_{M_j} P(M_j|C_{\text{val}}) = \arg\max_{M_j} P_{\text{penal}}(M_j|C_{\text{val}})$$
$$P_{\text{pos}}(M_j|C_{\text{val}}) P_{\text{dev}}(M_j) P_{\text{height}}(M_j)$$
$$\times P_{\text{cov}}(c_i, M_j) \exp\left(L_r(M_j) L_s(M_j)\right). \tag{15}$$

*3) Hierarchical Model Matching:* To efficiently search a best model for a candidate among the large number of possible models, we refer to [10] and [32] to establish a template hierarchy. We divide the projected model shapes into seven groups according to their orientations: {0°, 180°}, {30°, 150°}, {60°, 120°}, {90°}, {−60°, −120°}, {−30°, −150°}, and {−90°}.

Then, for each group, we construct a model shape hierarchy based on shape dissimilarities measured by the chamfer distance *D* between 2-D model boundaries.

For each group, given the models and their associated dissimilarity matrix, the template hierarchy is established by agglomerative clustering, which stops when two clusters are left. For each nonleaf node of the hierarchy, if models of the leaf nodes below it are $M_1, \ldots, M_n$, the representative shape for this node is selected as

$$M = \arg\min_{M_j} \left( \max_{i=1\ldots n, i\neq j} D(M_i, M_j) \right). \tag{16}$$

The seven hierarchies constitute the final model shape hierarchical tree with the root being empty.

For hierarchical model matching, when matching the first level of the hierarchical tree, all the possible scales and horizontal head torso deviations are traversed, and the best matched scale and head torso deviation are fixed to that model, and only the adjacent scales and deviations are searched when matching models of lower levels. As in [32], at each level, the maximum and minimum of the posterior, i.e., $P_{\max}$ and $P_{\min}$, are computed, and a threshold is selected as

$$P_\tau = P_{\min} + \eta(P_{\max} - P_{\min}) \tag{17}$$

to discard the models that are not good enough. In our experiment, we set $\eta$ to be 0.3.

After the model matching, we finely tune the parameters for the three best matched models and evaluate the posterior using the foreground edge instead of *pb*. As the *pb* magnitudes vary over a wide range, which is caused by variations in illumination and contrast between human objects and their background, this procedure can be considered as contrast normalization. The shape matching likelihood $L'_s(M_j)$ calculated using foreground edge is defined as

$$L'_s(M_j) = 1 - \frac{1}{\tau}\sqrt{\frac{1}{N_{Mb_{j,\mathrm{rem}}}} \sum_{k\in Mb_{j,\mathrm{rem}}} \left(\min(d_{FE}(k), \tau)\right)^2} \tag{18}$$

where $d_{FE}(k)$ is point $k$'s nearest foreground edge point, and $\tau$ is an upper bound of the boundary point's distance to the edge point and is scale dependent, which is set to be the width of the model's head.

We also record the matched foreground edge points for each model $m_i$. Assuming that each edge point can only come from one object, candidates that share a large percentage of edge points cannot be validated at the same time. For the same reason, edge points matched to the validated candidates are not allowed to match with any other candidates in later iterations.

As the subsequent candidate validation and rejection step needs a value that indicates each candidate's model matching quality, we define the model matching score for each model as

$$S_m(m_i) = \log\left(P_{\mathrm{height}}(m_i) P_{\mathrm{dev}}(m_i)\right) + L'_s(m_i) \tag{19}$$

where the prior terms $P_{\mathrm{height}}(m_i)$ and $P_{\mathrm{dev}}(m_i)$ evaluate the quality of the shape model $m_i$, and $L'_s(m_i)$ evaluates how well $m_i$ matches with the foreground edge.

### *D. Candidate Validation and Rejection*

Given the model matching results of the selected candidates, we examine them for validation or rejection. To achieve this, we first reject the candidates that have unsatisfactory model matching qualities and the candidates whose corresponding image areas can be better explained by other candidates, and then confirm the candidates that are less likely to be occluded by any other candidates.

*a) Consider Single Candidate's Model Matching Quality:* For each candidate $c_i$ that is selected, if its model matching score $S_m(m_i)$ is smaller than a threshold $S_T$ (set to be 0.4 in our experiment), or adding $m_i$ into $\theta$ cannot increase the posterior $P(\theta|I)$, $c_i$ is rejected. This is to reject the model $m_i$ that is either poorly matched in shape or just explains a relatively small area of the foreground.

*b) Consider Other Candidates' Model Matching Quality:* For each remaining candidate $c_i$ and the corresponding model $m_i$, the MDL principle is applied to evaluate if it should be rejected. The evaluation is in terms of the savings that can be obtained by rejecting $c_i$ as follows:

$$\begin{aligned} Sav_i &= SE_i - SE_{-i} + SM_i \\ SE_i &= \text{area}(m_{i,\mathrm{rem}})\left(1 - S_m(m_i)\right) \\ SE_{-i} &= \max_{j,k\neq i} \sum_{p\in m_{i,\mathrm{rem}}} \left(1 - \max\left(S_m(m_j,p), S_m(m_k,p)\right)\right) \\ SM_i &= \alpha(\mathbf{L}_i) \cdot \frac{1 - \text{area}(m_i \cap I_{\mathrm{occ}})}{\text{area}(m_i)} \end{aligned} \tag{20}$$

where $m_{i,\mathrm{rem}}$ is $m_i$'s intersection with $I_{\mathrm{rem}}$, $SE_i$ is the error introduced by using $m_i$ to explain $m_{i,\mathrm{rem}}$, and $SE_{-i}$ is the error introduced by combining two other candidates matched in the current iteration to explain $m_{i,\mathrm{rem}}$. $S_m(m_j, p) = S_m(m_j)$ if $p \in m_{i,\mathrm{rem}}$ and $S_m(m_j, p) = 0$ otherwise. $SM_i$ is the cost of the model. According to the MDL principle, if $Sav_i$ is positive, $c_i$ is rejected. If two candidates can mutually explain each other, then the candidate with larger saving is rejected.

*c) Consider Candidates' Occlusion Order:* After rejecting the candidates that are not good enough, we exclude the candidates that are likely to be occluded in terms of the remaining mask $I_{\mathrm{rem}}$ and then validate the remaining candidates. Specifically, for any pair of intersected models, as they cannot be unoccluded at the same time, we exclude the model that is likely to be occluded according to the following rules.

1) If their distance to each other is smaller than $d_{\min}$, or their overlapping area is larger than 90% of the area of the smaller model, their occlusion order is ambiguous. To make the decision, we first compare their posterior: if one's posterior is significantly larger than the other, the one with the smaller posterior is excluded. The parameter that indicates "significantly larger" is learned through experiments and is fixed at 1.35 through all our experiments.

Otherwise, we compare their shape matching scores calculated by (19) and exclude that with the lower score.

2) Otherwise, the occlusion order is clear. The model that is farther away from the camera is excluded.

The remaining candidates are then temporarily validated. To ensure that edge points should not be shared among different human objects, for every pair of remaining candidates, the ratio of their shared edge points is calculated. If the ratio is higher than a threshold (10% is used in the experiment to tolerate the cases that would unlikely result in wrong decisions), only the candidate that is nearer to the camera is validated.

After the validation, the candidates whose head centers are inside the validated candidates are rejected, and all the related quantities are updated. As the human model does not contain arms and the items being carried, to avoid false alarms that try to explain these unmodeled areas, the dilated area of each validated model is used to update $I_{\text{rem}}$ and $I_{\text{occ}}$. The size of the dilation structuring element $r_{\text{se}}$ is tunable, and the larger $r_{\text{se}}$ is relative to the scale, the less the false alarms would be produced, whereas the more true human objects might be missed. We set its default value to be a quarter of the model's head width. The foreground edge map is updated by removing the edge points assigned to the validated candidates. Then, for the $pb$ map, the boundary response whose nearest foreground edge point that has been removed is set to be zero. The remaining foreground pixels are reassigned to the remaining HCs probabilistically.

The entire optimization procedure is summarized below.

**Algorithm:** Optimization Algorithm

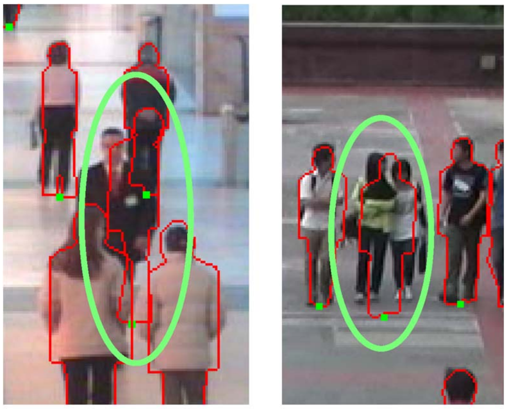
```
Given the candidate nomination $C_{\text{total}}$ and the foreground mask $I_f$,
initialize $\theta = \varnothing$, $I_{\text{occ}}$ as empty (black image), $I_{\text{rem}} = I_f$, the validated candidates set $C_{\text{val}} = \varnothing$, the rejected candidates set $C_{\text{rej}} = \varnothing$, and the posterior as $P(\theta|I) = \exp(-\text{area}(I_f))$.
Assign foreground pixels probabilistically to HCs.
Build the candidates' relation graph $G$.
while $C_{\text{val}} \bigcup C_{\text{rej}} \neq C_{\text{total}}$
do
    1. Select the candidates for model matching according to $I_f$, $I_{\text{rem}}$, and $G$.
    2. For each selected candidates in step 1, perform hierarchical model matching and select the best matched model as the one that results in the maximum posterior.
while at least one candidate is selected for model matching
    3. Validate and reject these matched candidates, and update $C_{\text{rej}}$, $C_{\text{val}}$, $\theta$, $I_{\text{rem}}$, $I_{\text{occ}}$, foreground edge map, $pb$ map and $\tilde{P}_{\text{cover}}(p, c_i)$.
end
return $\theta$.
```

## V. Experimental Result

We evaluated the proposed method using two data sets: the Caviar benchmark data set [1] and an outdoor scene video taken on our HKU campus. The camera parameters of the two data sets are illustrated in Table I. Due to the differences in the qualities of the two videos, the reliability of the extracted foreground is different: the foreground mask of the Caviar data set is more fragmented than the video taken by us and, hence, less reliable. Therefore, the parameter $w$ in (10) is set to 0 for the Caviar data and 0.8 for the HKU campus data. All the other parameters are set the same for the two data sets.

TABLE I
CAMERA PARAMETERS OF THE TWO TESTED DATA SETS

| | pan angle (deg) | tilt angle (deg) | swing angle (deg) | focal length (pixels) | Camera height (m) |
|---|---|---|---|---|---|
| Caviar | 86.05 | -6.33 | 0.5125 | 1680.5 | 3.948 |
| HKU | 96.21 | -15.34 | -0.0873 | 1810.3 | 5.927 |



(a) (b)

Fig. 10. Evaluation criterion 1). (a) The error within the green ellipse is counted as one missed detection and one false alarm. (b) The error within the green ellipse is counted as two missed detections and one false alarm.

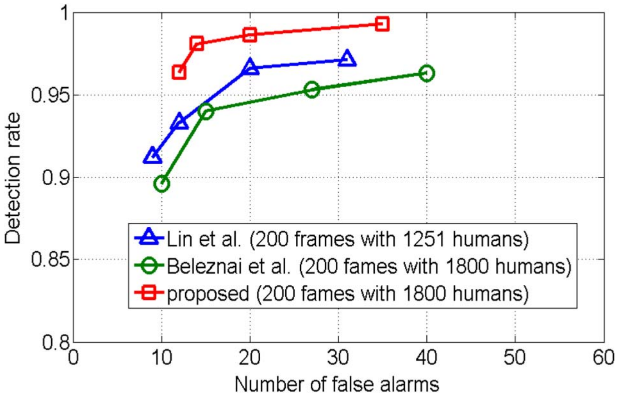
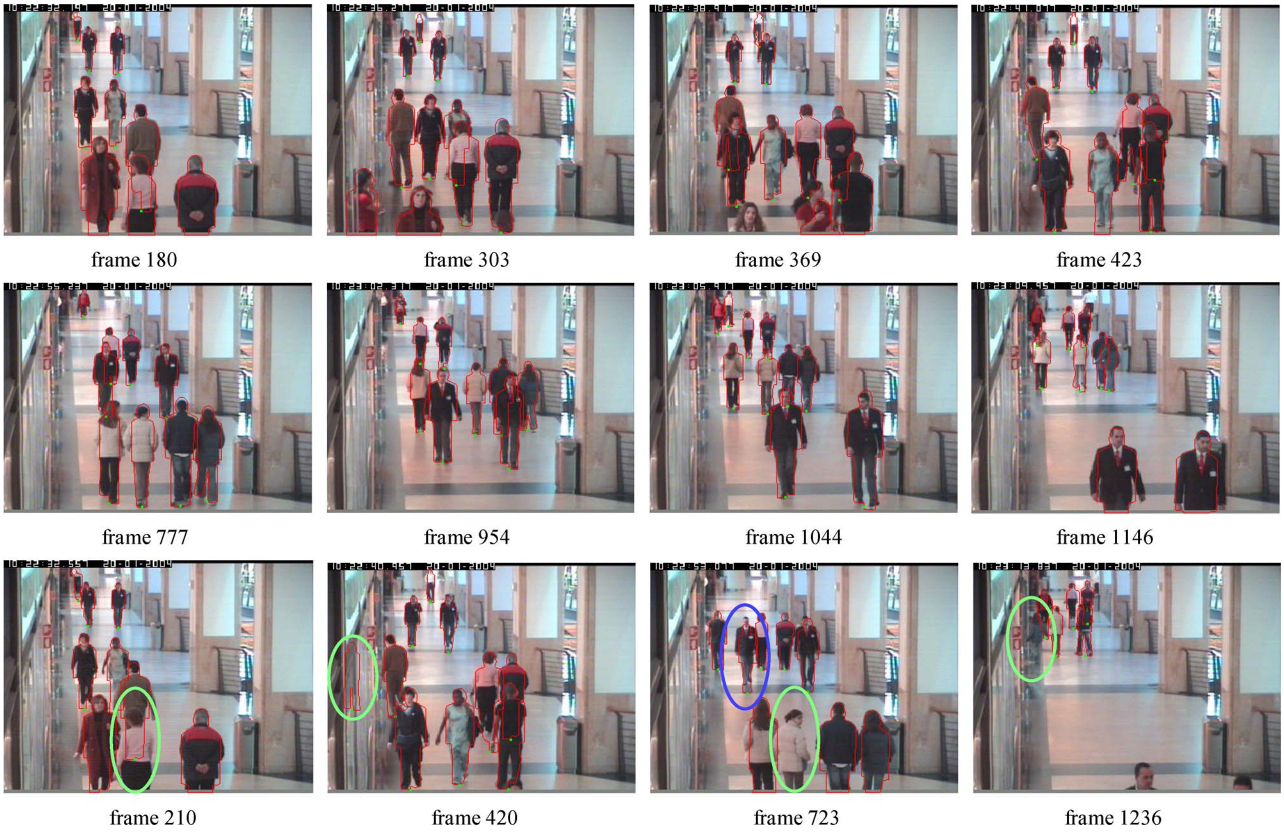


Fig. 11. ROC curves of evaluation on a subset of the Caviar data set.

The evaluation is based on the following criteria: 1) A correct detection is a detection $DT$ that has a one-to-one correspondent $GT$ in the ground truth human objects and satisfies

$$\text{Overlap}(GT, DT) = \frac{\text{area}(GT \cap DT)}{\text{area}(GT \cup DT)} > 0.5 \tag{21}$$

2) Human objects having less than 50% of the bodies inside the images are not evaluated. 3) Sitting and scene occluded (more than 20% occluded) human objects are not evaluated.



frame 180 frame 303 frame 369 frame 423

frame 777 frame 954 frame 1044 frame 1146

frame 210 frame 420 frame 723 frame 1236

Fig. 12. Detection results on Caviar data set.

4) Human objects staying in the scene for a relatively long time without significant movements are not evaluated and considered as scene objects. Fig. 10 shows the application of criterion 1) on two detection results.

### A. Detection Results on the Caviar Data Set

We evaluated the proposed method on the sequence *OneStopMoveEnter1Cor* (1590 frames with resolution being 384 × 288) of the Caviar data set. To compare the proposed method with previous works, e.g., [4], in which evaluation is done for 200 selected frames of this sequence, and, e.g., [2], in which evaluation is done for frames 801–1000, we first evaluated our method for frames 801–1000 by varying $\alpha(\mathbf{L}_i)$ and $r_{\text{se}}$. The receiver operating characteristic (ROC) curves for different methods are plotted in Fig. 11, from which we can see that the proposed method has a detection rate of around 98% with tolerable number of false alarms. However, as the frame rate of the video is 25 f/s, consecutive frames are highly correlated. Therefore, we also tested our approach on the whole sequence by sampling the first frame out of every three consecutive frames. By fixing $\alpha(\mathbf{L}_i)$ and $r_{\text{se}}$ at their default values, the obtained result is the following: among the 530 tested frames (containing 3705 human objects), the proposed method produced neither missed detections nor false alarms on 319 frames (containing 2166 human objects), and the overall detection rate is 94.3% with the false alarm rate being 1.62%. More detection results are depicted in Fig. 12, of which the first two rows show some successfully detected frames, whereas the third row illustrates some failed cases. It can be seen from the failed cases that a false alarm occurs in frame 420, which is caused by the motion of reflections on the glass wall, and frames 210, 723, and 1236 contain missed detections, which are caused by poor figure-ground contrast. We can also see from frame 723 that, although the man enclosed by the blue ellipse corresponds to a correct detection, the model does not fit the human object very well, which is actually caused by the shadows that are failed to be removed from the foreground. In the worst case, such inaccuracy introduced by shadows may lead to false alarms or missed detections.

### B. Detection Results on the HKU Data Set

The HKU campus data set is a 50-min video taken at 25 f/s with the resolution being 1280 × 720. The view is deep and wide, resulting in substantial scale changes (with the width of a normal human object varying from 10 pixels to 70 pixels), and the inclination varies as well (being ±7° inclined on the left and right border of the image compared with a vertical line). Unlike the Caviar data set where most people have a front/back view, the HKU data set contains humans walking in various orientations. They carried various items as well. In addition, on the right hand side of the scene, the illumination is weak, and the background is dark.

Due to the large number of frames, we subsampled the frames to 2.5 f/s, obtaining 7500 frames, on which the proposed method was tested. However, 7500 frames still represent a sizeable evaluation task. As such, we manually selected several portions of the sequence where occlusion occurs frequently

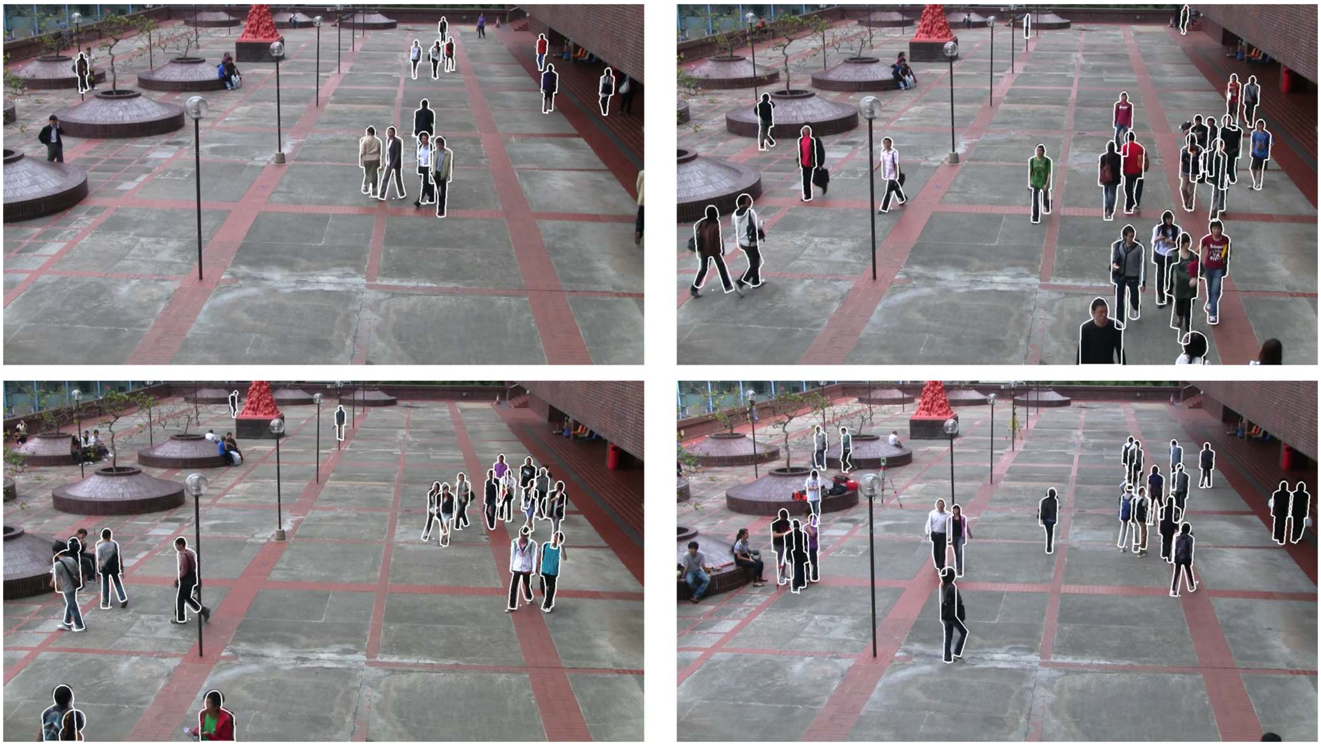


Fig. 13. Detection results on the HKU campus data set.

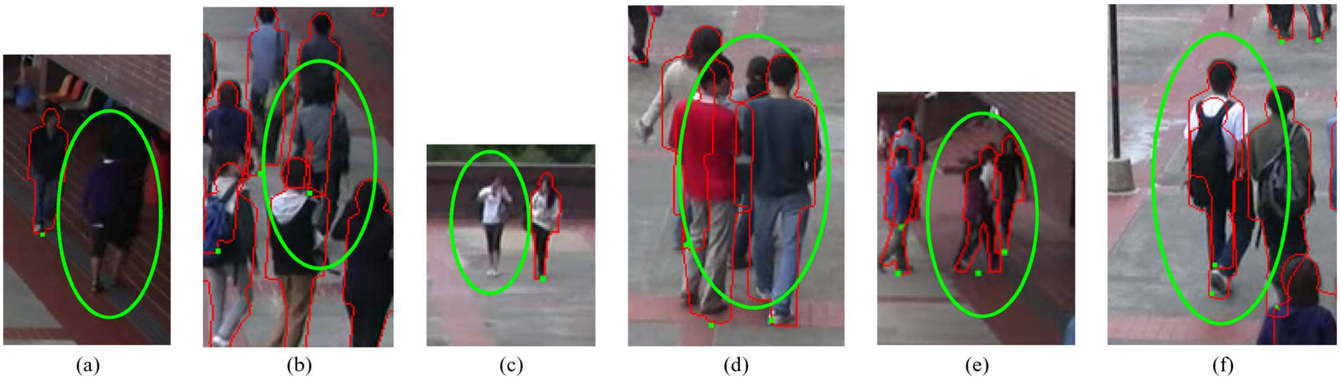


Fig. 14. Failed cases of the HKU campus data set.

and the numbers of humans are relatively large. There are in total 1105 such frames, containing 15 775 humans. The detection rate achieved is 90.74% when the false alarm rate is 1.88%. Fig. 13 shows some detection results, and Fig. 14 illustrates some typical failed cases. Among the errors, missed detections mainly come from poor figure-ground contrast [see Fig. 14(a) and (b)], low resolution [see Fig. 14(c)], and severe occlusion [see Fig. 14(d) and (e)], whereas false alarms are usually produced by texture rich regions [see Fig. 14(f)]. There are incorrect posture estimations as well, which are caused by shape ambiguities in 2-D and the rough approximation of various human shapes by a limited number of models.

Observing that missed detections mainly come from low resolution and poor contrast areas, where some human objects are hard to be identified by naked eyes, to fairly demonstrate the performance of our method, we also report the result by not considering these extreme cases. If the human objects whose feet appear beyond the line on the ground marked by the second farthest lamp pole are not counted, where a normal human object's width is less than 18 pixels (in [41], human width less than 24 pixels are not counted), the detection rate is 93.46%, and the false alarm rate is 1.91%. Further, if we also do not consider the dark red area on the right hand side of the scene, the detection rate goes up to 96.21% when the false alarm rate comes down to 1.77%.

### C. Computational Cost Analysis

Our detection method is currently implemented in Matlab. For each candidate, depending on the resolution of the hypothesized human object, model matching usually takes between 5 and 10 s. Therefore, if 50 times of model matching are needed for a frame, we need about 4–8 min to produce the detection result, not considering the computational time of the other

TABLE II
NUMBER OF TIMES VISITED FOR 8639 CANDIDATES

| Times visited | 1 | 2 | 3 | 4 | 5 | 6 | 7 | >=8 |
|---|---|---|---|---|---|---|---|---|
| No. of candidates | 5368 | 1886 | 908 | 346 | 109 | 19 | 3 | 0 |
| Percentage | 65.26% | 21.83% | 10.51% | 4.01% | 1.26% | 0.22% | 0.03% | 0 |

procedures, which can actually be neglected compared with the computational time of model matching. If the method is implemented in C++ with code optimization, we believe that the proposed method can run within several seconds for each frame.

We have also analyzed the efficiency of the optimization process experimentally. By checking the 8639 candidates that are nominated for frames 801–1000 of the tested Caviar sequence and counting the number of times they are selected for model matching during the optimization process, we obtained the result as shown in Table II. It can be seen that 62.1% of candidates are visited once and 87.2% of the candidates are visited no more than twice. The average times visited are 1.6122. This analysis demonstrates that the proposed method does not cost much more than the greedy method, in which each candidate is visited once.

## VI. CONCLUSION

A Bayesian approach for human detection in crowded scenarios has been proposed in this paper. Foreground regions and edges are used to provide image evidence for the 3-D model-based inference. Knowledge priors about human shape distribution and interhuman distance limitation are enforced during the model matching process. Foreground pixels are probabilistically assigned to candidates to avoid the model being attracted by incorrect edge points. Candidate validation and rejection based on the MDL principle and local occlusion reasoning are carried out after each iteration of model matching. The solution is obtained in a way that balances the computational cost and the performance. Detection rates of 94.3% and 90.7% with false positive rate of less than 2% are achieved on the Caviar data set and a data set taken by ourselves.

However, there are still missed detections, false alarms, and wrong posture estimation. These mistakes are not easy to deal with using current techniques. To improve the performance, the most important future work is to combine the detection results across consecutive frames, which can resolve the ambiguities of a single frame, to obtain a more reliable detection, counting, and posture estimation performance.

## REFERENCES

[1] *Caviar Test Case Scenarios*. [Online]. Available: homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

[2] C. Beleznai and H. Bischof, “Fast human detection in crowded scenes by contour integration and local shape estimation,” in *Proc. Comput. Vis. Pattern Recog.*, 2009, pp. 2246–2253.

[3] S. Bileschi and L. Wolf, “Image representation beyond histograms of gradients: The role of Gestalt descriptors,” in *Proc. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[4] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, “Understanding transit scenes: A survey on human behavior-recognition algorithms,” *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 206–224, Mar. 2010.

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

[6] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: A benchmark,” in *Proc. Comput. Vis. Pattern Recog.*, 2009, pp. 304–311.

[7] M. Enzweiler and D. Gavrila, “Monocular pedestrian detection: Survey and experiments,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *Proc. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[9] W. Gao, H. Ai, and S. Lao, “Adaptive contour features in oriented granular space for human detection and segmentation,” in *Proc. Comput. Vis. Pattern Recog.*, 2009, pp. 1786–1793.

[10] D. Gavrila, “Pedestrian detection from a moving vehicle,” in *Proc. Eur. Conf. Comput. Vis.*, 2000, pp. 37–49.

[11] W. Ge and R. Collins, “Marked point processes for crowd counting,” in *Proc. Comput. Vis. Pattern Recog.*, 2009, pp. 2913–2920.

[12] Z. Guohui and W. Yinhai, “Optimizing minimum and maximum green time settings for traffic actuated control at isolated intersections,” *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 164–173, Mar. 2011.

[13] X. C. He and N. H. C. Yung, “New method for overcoming ill-conditioning in vanishing-point-based camera calibration,” *Opt. Eng.*, vol. 46, no. 3, p. 037202, Mar. 2007.

[14] B. Leibe, E. Seemann, and B. Schiele, “Pedestrian detection in crowded scenes,” in *Proc. Comput. Vis. Pattern Recog.*, 2005, pp. 878–885.

[15] Z. Lin and L. Davis, “A pose-invariant descriptor for human detection and segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 423–436.

[16] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, “Hierarchical part-template matching for human detection and segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[17] Z. Lin, G. Hua, and L. Davis, “Multiple instance feature for robust part-based object detection,” in *Proc. Comput. Vis. Pattern Recog.*, 2009, pp. 405–412.

[18] D. R. Martin, C. C. Fowlkes, and J. Malik, “Learning to detect natural image boundaries using local brightness, color, and texture cues,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May 2004.

[19] K. Mikolajczyk, C. Schmid, and A. Zisserman, “Human detection based on a probabilistic assembly of robust part detectors,” in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 69–82.

[20] A. Mohan, C. Papageorgiou, and T. Poggio, “Example-based object detection in images by components,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, Apr. 2001.

[21] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, “Discriminative local binary patterns for human detection in personal album,” in *Proc. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[22] S. Munder, C. Schnorr, and D. Gavrila, “Pedestrian detection and tracking using a mixture of view-based shape-texture models,” *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 2, pp. 333–343, Jun. 2008.

[23] M. P. Murray, A. B. Drought, and R. C. Kory, “Walking patterns of normal men,” *J. Bone Joint Surg. Am*, vol. 46-A, no. 2, pp. 335–360, Mar. 1964.

[24] C. Papageorgiou and T. Poggio, “A trainable system for object detection,” *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, Jun. 2000.

[25] D. Ramanan, “Using segmentation to verify object hypotheses,” in *Proc. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[26] J. Rittscher, P. H. Tu, and N. Krahnstoever, “Simultaneous estimation of segmentation and shape,” in *Proc. Comput. Vis. Pattern Recog.*, 2005, pp. 486–493.

[27] P. Sabzmeydani and G. Mori, “Detecting pedestrians by learning shapelet features,” in *Proc. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[28] V. Sharma and J. Davis, “Simultaneous detection and segmentation of pedestrians using top-down and bottom-up processing,” in *Proc. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[29] A. Shende, M. P. Singh, and P. Kachroo, “Optimization-based feedback control for pedestrian evacuation from an exit corridor,” *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1167–1176, Dec. 2011.

[30] V. Shet, J. Neumann, V. Ramesh, and L. Davis, “Bilattice-based logical reasoning for human detection,” in *Proc. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[31] X. Song and H. B. L. Duh, "A simulation of bonding effects and their impacts on pedestrian dynamics," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 153–161, Mar. 2010.

[32] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical Bayesian filter," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1372–1384, Sep. 2006.

[33] A. R. Tilley, *The Measure of Man and Woman: Human Factors in Design*. New York: Whitney Library, 1993.

[34] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu, "Unified crowd segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 691–704.

[35] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.

[36] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, Jul. 2005.

[37] L. Wang and N. H. C. Yung, "Extraction of moving objects from their background based on multiple adaptive thresholds and boundary evaluation," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 40–51, Mar. 2010.

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

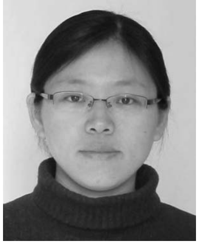
[39] C. Wojek and B. Schiele, "A performance evaluation of single and multifeature people detection," in *Proc. DAGM Symp. Pattern Recog.*, 2008, pp. 82–91.

[40] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. Comput. Vis. Pattern Recog.*, 2005, pp. 90–97.

[41] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, Nov. 2007.

[42] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *Proc. Comput. Vis. Pattern Recog.*, 2003, pp. 459–466.

[43] Q. Zhu, M. Yeh, K. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. Comput. Vis. Pattern Recog.*, 2006, pp. 1491–1498.



**Lu Wang** received the B.Eng. and M.Eng. degrees in computer science from Harbin Institute of Technology, Heilongjiang, China, in 2003 and 2005, respectively, and the Ph.D. degree from the University of Hong Kong, Hong Kong, in 2011.

She is currently an Assistant Professor with the College of Information Science and Engineering, Northeastern University, Liaoning, China. Her research interests include computer vision, image processing, and pattern recognition.

**Nelson Hon Ching Yung** (M'85–SM'96) received the B.Sc. and Ph.D. degrees from the University of Newcastle-Upon-Tyne, Tyne, U.K.

He was Lecturer with the University of Newcastle-Upon-Tyne from 1985 to 1990. From 1990 to 1993, he was a Senior Research Scientist with the Department of Defence, Australia. He joined Hong Kong University, Kowloon (HKU), Hong Kong, in late 1993, as an Associate Professor. He is the founding Director of the Laboratory for Intelligent Transportation Systems Research, HKU. He acts as a Consultant to government units and a number of local and international companies. He has coauthored five books and book chapters and has published over 150 journal and conference papers in the areas of digital image processing, parallel algorithms, visual traffic surveillance, autonomous vehicle navigation, and learning algorithms. He was a Guest Editor of the SPIE *Journal of Electronic Imaging*.

Dr. Yung is a member of the Hong Kong Institution of Engineers and the Institution of Electrical Engineers. He was the Regional Secretary of the IEEE Asia-Pacific Region, a Council Member and the Chairman of Standards Committee of Intelligent Transportation Systems-Hong Kong, and the Chair of the Computer Division, International Institute for Critical Infrastructures. He was a member of the Advisory Panel of the Intelligent Transportation Systems Strategy Review, Transport Department, Government of the Hong Kong Special Administrative Region. He serves as a Reviewer for a number of IEEE, Institution of Engineering and Technology (IET), and International Society for Optics and Photonics (SPIE) journals. He is a Chartered Electrical Engineer. His biography has been published in *Who's Who in the World* since 1998.