Adaptive Model for Robust Pedestrian Counting

Jingjing Liu, Jinqiao Wang, and Hanqing Lu

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{jjliu,jqwang,luhq}@nlpr.ia.ac.cn

Abstract. Toward robust pedestrian counting with partly occlusion, we put forward a novel model-based approach for pedestrian detection. Our approach consists of two stages: pre-detection and verification. Firstly, based on a whole pedestrian model built up in advance, adaptive models are dynamically determined by the occlusion conditions of corresponding body parts. Thus, a heuristic approach with grid masks is proposed to examine visibility of certain body part. Using part models for template matching, we adopt an approximate branch structure for preliminary detection. Secondly, Bayesian framework is utilized to verify and optimize the pre-detection results. Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm is used to solve such problem of high dimensions. Experiments and comparison demonstrate promising application of the proposed approach.

Keywords: pedestrian counting, adaptive model, grid mask, RJMCMC.

1 Introduction

Pedestrian counting is sometimes required and essential in realms of multimedia and computer vision because of its wide-spread applications, such as video surveillance, driver assistance and scene analysis. However, it is a challenging task due to various clothing, body articulation and spatial occlusion and so on. For decades, research works and realistic applications on pedestrian counting have employed a myriad of methods, many of which rely on human detection. Hence it is indispensable to discuss the methods of pedestrian counting as well as the relevant ones for detection. Traditional methods pay more attention to static appearance of human, for instance texture, shape or silhouette. In the following, we will review these methods in detail.

Model-based methods build human models via shape or silhouette representation with prior knowledge. Lin *et al.* [1] proposed a hierarchical matching method based on artificially constructed models made up by part-templates to deal with occlusion. Zhao *et al.* [2] used 14 artificial 3D models to simulate specific poses of pedestrian, capturing side and front characters to handle body flexibility. Wu *et al.* [3] presented an approach which learned a series of certain part detectors for body parts through edgelet so as to improve detection rate.

The other methods train low-level feature based classifiers for human detection or counting whereas without premise of human body models. The low-level features reflect local details yet do not belong to any corporal concepts. Viola *et al.*

K.-T. Lee et al. (Eds.): MMM 2011, Part I, LNCS 6523, pp. 481–491, 2011.

[4] learned a cascade structure classifier using textures of pedestrian appearance. Leibe *et al.* [5] trained visual codebooks to estimate spatial occurrence distribution of pedestrians. The well-known HOG descriptor proposed by Dalal *et al.* in [6] made use of local gradient information to depict appearance of human. A more recent method appears in [7]. Gao *et al.* proposed a feature representation (ACF) similar to HOG descriptor and formed a classifier exploring the co-occurrence of discriminative features.

Besides the static appearance, motion information from video sequences is employed for pedestrian counting or relevant tasks. Here, relationships among sequential frames in video are all taken as motion information. Viola *et al.* [4] learned feature filters for final classifier, including filters based on different images of sequential images. Given camera is fixed, others used motion information for pre- or post- process, for example motion segmentation. Zhao *et al.* [2], [8] verified and optimized preliminary detection results with foreground mask under Baysian framework. A.B. Chan *et al.* [9], [10] introduced Dynamic Texture (DT) into the field of crowd monitoring. As a pre-process, DT segmentation can not only capture the moving regions for the following features extraction and regression but also distinguish objects with different velocities.

Toward robust pedestrian counting with partly occlusion, we propose an adaptive model based approach for pre-detection. Three part models: head-shoulder (HS), right torso (RT) and left torso (LT) are established according to a pre-constructed pedestrian model. Assuming that there are only pedestrians standing or walking in the motion regions, head and shoulder are seldom occluded otherwise we do not take the sample as a valid pedestrian. Thus head-shoulder model is always contained in the adaptive model. In contrast, whether the model of left/right torso is reserved or discarded relies on visibility of the relevant body part in the image. Here, we propose a heuristic approach using low-level feature to determine the co-occurrence between head-shoulder and torso sides. With the adaptive model, we build up an approximate branch structure for integrate detection. In order to avoid penalty caused by deformation or articulation, we place template matching of torso sides before final classification using the adaptive model. Then, we verify and optimize the detection results under Bayesian framework on the basis of motion segmentation.

2 Adaptive Model

Model-based methods with prior knowledge try to link concept 'human'/'body part' to low-level features. Yet current methods with permanent part models usually have less flexibility thus bringing about improper punishment, in condition that some part detection get low scores due to occlusion or deformation. On the other hand, although some classifiers, for example the one learned by boosting algorithm, rely on training set and are sensitive to parameter changes and occlusion to some degree, low-level features based methods are expert in explaining local details. So it makes sense to take advantage of both model-based methods and low low-level feature based ones.

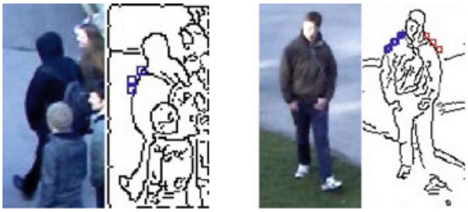
2.1 Part Models

In crowded scenes, pedestrians often occlude each other which adds difficulties for detection. Part detectors have been proved efficient in pedestrian detection meanwhile dealing with occlusion [1], [3]. In common sense, the head-shoulder, the left torso and right torso are seldom occluded simultaneously. Nevertheless, it is still unknown whether either torso side of pedestrian is occluded or which side is not. So it is not proper to construct a classifier of cascade structure [11] using part detectors or a classifier composed of part-templates with different weights [1]. Besides, it is not wisdom to ignore some persuasive evidence such as torso region which indicates existence of pedestrian.

The proposed adaptive model is defined as a set of part models corresponding to upper body parts of pedestrian. All the part models as well as legs model compose the whole pedestrian model. Dealing with occlusion, the adaptive model could tolerate some part models' missing but the others must be consisted. Therefore, distinct situations of occlusion are supposed to refer to different pedestrian adaptive models. As adaptive, the model ties to utilize as much convincing information as possible thus pursuing robust result. See Fig. 1(a), to detect such a person, adaptive model should only contain model of left torso (from perspective of picture viewers) while discarding the unconvincing information of right torso for it being vague. However, in Fig. 1(b), as being visible in scene, models of both torso sides should be adopted in final adaptive model.

Compared with color and texture, shape information is more confident to capture the characteristic of a pedestrian. However, body articulations and some equipment such backpack might bring about some mistakes or false alarm into the appearance. Thereby, we just consider the relatively invariant shape of human body with prior knowledge, neglecting the parts of high DOF (degree of freedom). A low dimensional shape model is constituted by 3 ellipses whose positions are relatively fixed: one indicates head, the others torso and legs. Fig. 2(a) illustrates integral shape model of a pedestrian with normal height, body proportion and fatness, similar to 2D model in [8]. We find that such a simple model is competent in practice.

Under the help of integral shape model, we separate the model into 3 regions (see Fig. 2(b)), which is similar to the division in [3] : head-shoulder, torso and legs. Our part models are different from [3], which include head-shoulder, left



(a) (b)

Fig. 1. Different occlusion situations ought to correspond to different adaptive models. (a) right torso part is vague; (b) both of the torso sides are visible in image.

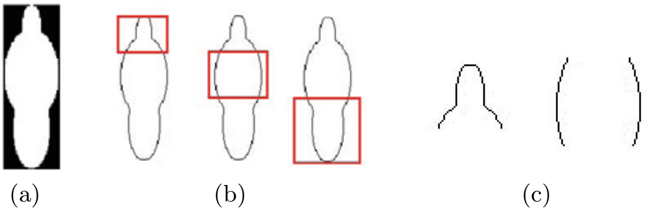


Fig. 2. Pedestrian model. (a) Full-body shape model; (b) regions of body parts: head-shoulder, torso and legs; (c) our part models: head-shoulder, left torso and right torso.

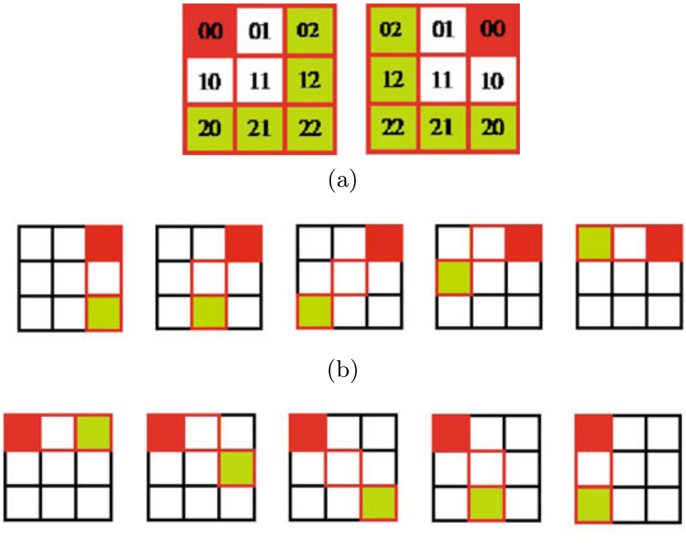
torso and right torso, as shown in Fig. 2(c). Each part model is represented by two parameter sets: positions of silhouette points and the unit image gradient vector set of the silhouette points. The introduce of gradient vector can partially handle problem of limited size change since unique model lacks scale variance.

2.2 Grid Mask for Torso Detection Using Consistent Contour

Because of size change, body articulations or occlusion, original part-template matching for torso detection is not robust. Consequently, it is crucial to avoid bringing improper punishment to final determination. To escape from this trap, we propose a heuristic approach based on contour consistence which searches existence of torso sides. Two grid-structure masks (as illustration in Fig. 3(a)) each containing 9 squares (in practice we use 3×3 or 5×5 size) are adopted to undertake this commission. For each mask, the red square marked by '00' is a start point. The mask with a top-right start detects left torso side whereas the other right side. Squares in green are the terminals. A path being found under defined consistent criterion between the start and the terminal square means consistence of torso sides edges in the squares' effective field.

The low-level features have shown good performance in explaining the local contour. So, we use a gradient feature similar to HOG descriptor [6] as the description of a square: gradient vectors of all points in the square's effective field merge into a unit vector, which represents the direction of the edge in the square. Since these features rely on a relative broad area but not pixels, this approach with grid masks is capable of tolerating with fat/thin, actions of upper body and contour deformation. See Fig. 3(b), all the possible paths covers a series of torso side edge. The algorithm through which we find a path in the masks is illustrated in Alg. 1.

In practice, first of all, we use the head-shoulder model to get some candidate positions of head. Then for each candidate, we place the the grid masks at the torso regions relative to the head candidates. If a terminal is found in the mask, a torso side is supposed to be visual in the image and relevant part model should be contained in the adaptive model. See Fig. 2(a), left torso is detected while in Fig. 2(b) both torso sides are found.



(a)

(b)

(c)

Fig. 3. (a) Grid masks for left and right detection; (b) consistent paths of left torso; (c) consistent paths of right paths

Algorithm 1. Torso side detection

```
1. Initialize the start coordinate and terminal coordinate:
   $(x_{start}, y_{start}) = (0,0)$, $(x_{terminal}, y_{terminal}) = (0,0)$
   angle between the unit gradient vector of start square and left/right shoulder
   vector of the constructed model $D_{start}$ ought to be smaller than a threshold
   $T_{torso}$
2. While $x_{terminal} \neq 2$ and $y_{terminal} \neq 2$
     – three coordinates of candidate squares :
       $(x_{start}+1, y_{start})$, $(x_{start}, y_{start}+1)$, $(x_{start}+1, y_{start}+1)$
     – select the best candidate square as the current end as well as the next
       mining start point
       $((x_{start}, y_{start}) = (x_{candidate}, y_{candidate})$
       $(x_{terminal}, y_{terminal}) = (x_{candidate}, y_{candidate})$
   If there is no such square, the iteration stops.
3. Either $x_{terminal}$ or $y_{terminal}$ if equal to 2, the left/right torso side is detected.
```

As a ***best candidate*** [7], we mean angle between unit gradient vector of the current start square and the best candidate square's direction defined above is the smallest among the three ones, besides lower than a threshold $T_{between}$. Besides, angle between the line connecting the two candidate's centers and the best candidate's direction ought to be larger than the sum $T_{consist}$.

3 Pedestrian Detection Based on Branch Structure

Inspired by the head candidate detection in [2], we adopt the part-template matching method along with contour gradients for body parts detection. Besides, motivated by efficiency of the cascade structure [11], we construct an approximate branch structure for pedestrian detection (see Fig. 4). Considering invariant shape and impossibility of occlusion of head-shoulder body part in this task, firstly, head-shoulder part-template matching is used to filter out lots of negatives. We set a relatively low threshold for this detection, pushing more candidates to pass through this access.

In the second step, the grid trigger decides which branch the candidates from the first step should go along. In practice, we find that even if the grid mask finds existence of a torso side, sometimes the score of torso side part-template matching is low because of deformation or poses of pedestrians, undermining the overall score of the adaptive model. This phenomenon obeys our expectation that is to employ warranted information as much as we could. To avoid such phenomenon, we place preliminary part-template matching of torso side on the branches before the adaptive model classification process. If some candidates could not get high score in the right/left torso template matching, the part model would not appear in the adaptive model. Although lose some information, these candidates could also possess a high possibility as a pedestrian because the only contained head-shoulder model might get a relative high score. Delicate thresholds achieve a good balance between grade of detect rate and decrease of false positive. Different branches result in different adaptive models.

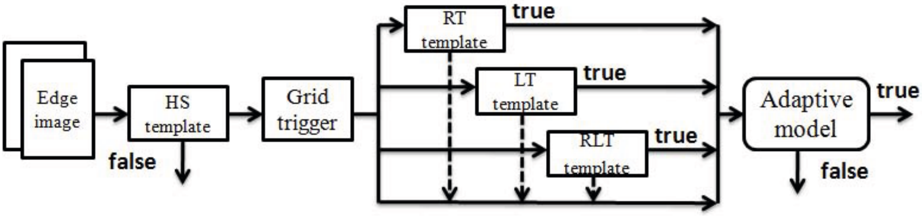


Fig. 4. The Approximate Branch Structure

Final classification in each branch relies on matching scores of part models contained in the adaptive model. Instead of fixed weights, we dispatch parts' weights according to their current scores . The higher is the score of the certain part, the bigger weight of this part, similar to making choices in realistic decision that human apt to accept information from more compromising channels.

4 Pedestrian Verification and Optimization

Since people sometimes share homogeneous velocities, it is hard to count pedestrians in crowded scene only with velocity. But motion information is no doubt effective complement for people counting. Some methods have shown good performance in obtaining exact shape of motion regions. Additionally, we make use of motion information in Bayesian framework to verify and optimize the pre-detection instead of estimating overlapping rate.

4.1 The Bayesian Framework

Given a well segmented foreground image, an optimization problem is defined to find the object mask best covering the foreground image with the guide of the results from pre-detection. Achieve the solution equals to maximum a posterior (MAP) estimation:

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

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

where O denotes the objects mask with multiple models while I is the foreground image. Likelihood and prior are discussed in the following sections.

4.1.1 Prior Distribution

We use the integral shape of pedestrian model as mask of one object. In practice, an object has only two parameters: position and height (the ratio of height to width is fixed). In the object mask, objects model can be resized on the basis of the object's height. Hence, the prior probability of an object i contains the probability of its parameters and the punishing item of adding an object:

$$P(O) = \prod_{i=1}^{n} P(s_i)P(x_i, y_i)P(h_i) \tag{3}$$

$$P(s_i) = e^{\lambda_{punish} A_i} \tag{4}$$

The first probability relying on the object's size is a punishing item. A_i is area size of the object. λ_{punish} is set to handle different scenes of variant pedestrian density. More crowded the image, smaller the λ_{punish}. Suppose that pedestrian could appear at every position of the scene, $P(x_i, y_i)$ is uniform distribution while $P(h_i)$ is a Gaussian distribution $N(\mu_h, \sigma_h)$ whose expectation and variance refer to height of average people and real height fluctuation coming from statistics. Besides, the range of the height distribution h_{rang} depends on statistic parameters.

4.1.2 Likelihood

We accept the assumption that there are only pedestrians in the foreground scene and pixels in image are independent. Following the inference of [2], we easily reach likelihood of the MAP problem:

$$P(I|O) = \alpha e^{-(\lambda_{10} N_{10} + \lambda_{01} N_{01})} \tag{5}$$

where N_{10} is the number of pixels in foreground image which are not covered by the object mask. Comparatively, N_{01} is the number of pixels in object mask which covers background. Incremental computation [2] to compute N_{10} and N_{01}

is also adopted in our experiments. λ_{10} and λ_{01} are also set to compromise between two kinds of wrong overlapping mentioned above.

Using the prior Equ. 3 and the likelihood Equ. 5, the posterior probability is:

$$P(O|I) \propto (\prod_{i=1}^{n} e^{\lambda_{punish} A_i} e^{-(\frac{h_i - \mu_h}{\sigma_h})}) \cdot e^{-(\lambda_{10} N_{10} + \lambda_{01} N_{01})}; h_i \in [1.5, 1.9] \qquad (6)$$

4.2 RJMCMC for Pedestrian Counting

The solution of the MAP estimation includes the number of pedestrians, the positions and heights. There is no doubt that the solution space is of high dimensions. RJMCMC has demonstrated strong ability in searching for a solution in such a space [2]. We also utilize this algorithm to solve such MAP estimation. During iterations, a sampled candidate state in the solution space is accepted according to value of the Metropolis-Hasting acceptance ratio:

$$p(x, x') = \min(1, \frac{p(x')q(x', x)}{p(x)q(x, x')}) \qquad (7)$$

where x and x' are the current and the candidate state, $p(\cdot)$ is the posterior distribution and $q()$ is the proposal probability. Through iterations, the sampled states become dense and ideally the states converge to the solution. The number of crowd could be inferred from the solution. In pre-detection, the results offer probable pedestrian positions, which provide domain knowledge to the proposal probability. Actually, parameter (x_i, y_i) of states are sampled at these detections with little drifting, resulting in a fast convergence. Other parameters are sample through iterations according to their distributions.

5 Experiments

Pedestrian counting experiments are carried on three video sequences from PETS 2009, which is a considerably challenging open dataset: S1.L1: 13-57-001 (seq.1), S2.L1: 12-34-001 (seq.2) and S2.L1: 12-34-008 (seq.3). Here, Seq.1 is a 221 frame sequence in which average number of pedestrian per frame is over 20, so people usually have severe occlusion. Seq. 2 contains 795 frames; average pedestrian number per frame is about 6. Seq.3 records the same scene as seq.2 but from a different viewpoint.

We take every pedestrian entering the scene as a valid pedestrian, until he leaves, even if he is totally occluded in the frame. In addition, we define what a good count is: we use position of head as a reference; the head position of a detection nearing the head top of a valid pedestrian with limited tolerance is seemed to be a true positive; however, if a correct count has been related to a pedestrian, other counts even if in the area of a good count, are seemed as false positives.

For motion segmentation, we employ Gaussian Mixture Models (GMMs) [12]. In every sequence, standard sizes of pedestrian model are re-estimated by size of height of a normal pedestrian in the scene and associated with μ_h. We multiply

Table 1. Compared results of performance evaluations

		Seq.1	Seq.2	Seq.3
Valid pedestrians		4719	4651	3856
Adaptive Model	Detection rate	65.75%	78.11%	67.66%
	False positive rate	5.11%	11.81%	17.17%
Approach in [2]	Detection rate	56.58%	79.54%	40.54%
	False positive rate	10.76%	11.08%	17.05%

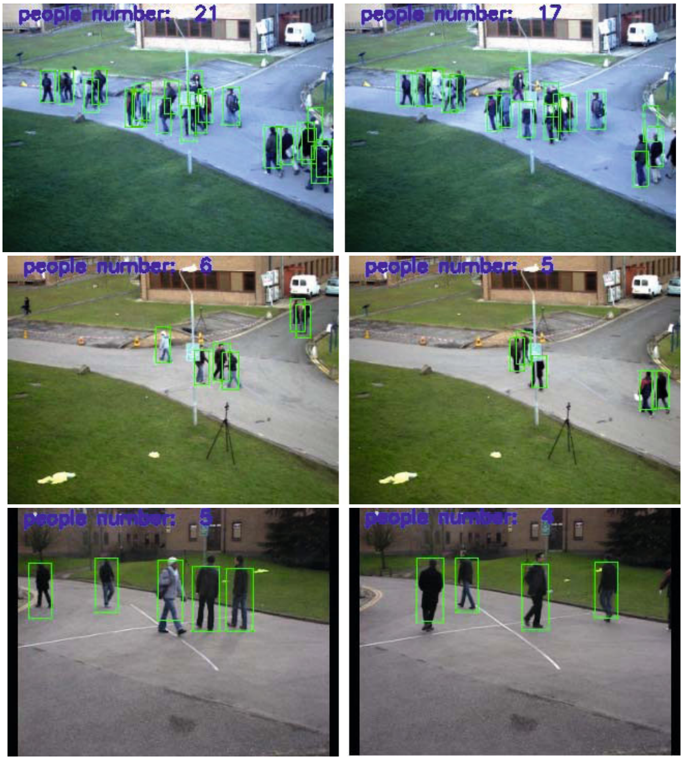


Fig. 5. Some experiment results: from top to bottom related to seq.1, seq.2 and seq.3

a coefficient λ_{torso} to scores of torso sides template so as to compensate lose due to body articulation. In our experiments, crucial parameters are fixed as follow: $D_{start} = (0.707, 0.707)$, $T_{torso} = 0.9$, $T_{between} = 0.8$, $T_{consist} = 0.707$, $\lambda_{10} = 0.8$, $\lambda_{01} = 1.0$ and $\lambda_{torso} = 1.2$ (the three thresholds is recalculated by cosine operation according to angle). 1000 iteration of RJMCMC runs for each frame which means 4000 solution states are sampled. Our approach makes use of Bayesian framework with motion segmentation as [2] has done, however, we utilize the adaptive model for pre-detection, reaching a robust result toward pedestrian counting in crowd. Experimental results of our approach and comparison with the approach in [2] are illustrated in Table. 1. Similar work in [13] is also under Bayesian framework whereas ours is more concise for without training process.

Our approach represents an enhanced performance towards previous approach [2] in seq.1 and seq.3. Frames of seq.1 and seq.3 contain some occluded people and are selected to prove better ability of our approach in dealing with occlusion. Consider that the verification process can handle some occlusion and approach in [2] uses residue foreground analysis, the improvement on the database is precious. In seq.2, because of occlusion seldom happening, detection rates of the two approaches are close; ours even has a little bit decrease compared with approach in [2]. As discussed above, detected torso sides might get low score in template matching, causing improper punishment to total matching score, thus detection rate declines.

Some experiment results of our approach are illustrated in Fig. 5. Several factors prevent our method from better performance: (1) Ground truths are strictly labeled by people; (2) The strict constrain of a true positive leads to so many false positives. (3) The size of pedestrians usually flux so fiercely that a unique model varies in a limited scale range is not omnipotent. This phenomenon explains the high false positive rate. (4) GMMs used in our experiment just obtain the area which has apparent moving, the standing or slowly moving pedestrians are not reflected in the foreground mask.

For each 320 × 240 frame, our approach takes 0.2s with c++ code, satisfying a real-time application.

6 Conclusion

In this paper, a novel method named adaptive model have been proposed. Part models are not fixed or weighted depending on probability but on the inkling of body part existence. We put forward a heuristic algorithm with grid masks to detect torso sides using contour consistence. Then pre-detection are implemented with a classifier of branch structure. Detection results are than verified and optimized with RJMCMC. Experiments on an open database show promising results indicating our approach is robust in pedestrian counting in crowd scenes.

In the future, we are planning to make use of the prior knowledge in some method of feature training. So convert the original adaptive model to a new adaptive model, part models of which is semi-supervised learning.

Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No.60905008, 60833006) and National Basic Research Program (973) of China under contract No.2010CB327905.

References

1. Lin, Z., Davis, L.S., Doermann, D., DeMenthon, D.: Hierarchical part-template matching for human detection and segmentation. In: 11th IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, pp. 1–8 (2007)

2. Zhao, T., Nevatia, R.: Bayesian Human Segmentation in Crowded Situations. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, pp. 459–466 (2003)
3. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detection. In: 11th IEEE International Conference on Computer Vision, Beijing, China, pp. 90–97 (2005)
4. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: 9th IEEE International Conference on Computer Vision, Nice, France, pp. 734–741 (2003)
5. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, pp. 878–885 (2005)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, pp. 454–461 (2005)
7. Gao, W., Ai, H.Z., Lao, S.H.: Adaptive contour features in oriented granular space for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, USA, pp. 1786–1793 (2009)
8. Zhao, T., Nevatia, R.: Stochastic human segmentation from a static camera. In: Workshop on Motion and Video Computing, Orlando, USA, pp. 9–14 (2002)
9. Chan, A.B., John Liang, Z.S., Vasconcelos, N.: Privacy preserving crowd monitoring: counting people without people models or tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, USA, pp. 1–7 (2008)
10. Chan, A.B., Morrow, M., Vasconcelos, N.: Analysis of Crowd Scenes using Holistic Properties. In: 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Miami, USA (2009)
11. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai Marriott, USA, pp. 511–518 (2001)
12. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Ft. Collins, USA, pp. 246–252 (1999)
13. Ge, W., Collins, R.T.: Marked Point Processes for Crowd Counting. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA (2009)