University of Wollongong

Research Online

Faculty of Informatics - Papers (Archive)

Faculty of Engineering and Information Sciences

1-1-2009

A part-based template matching method for multi-view human detection

Duc Thanh Nguyen
University of Wollongong, dtn156@uow.edu.au

Wanqing Li
University of Wollongong, wanqing@uow.edu.au

Philip Ogunbona
University of Wollongong, philipo@uow.edu.au

Follow this and additional works at: https://ro.uow.edu.au/infopapers

 Part of the Physical Sciences and Mathematics Commons

Recommended Citation

Nguyen, Duc Thanh; Li, Wanqing; and Ogunbona, Philip: A part-based template matching method for multi-view human detection 2009, 357-362.
https://ro.uow.edu.au/infopapers/2125

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

A part-based template matching method for multi-view human detection

Abstract

This paper proposes a part-based template matching method for multi-view human detection. The proposed method includes two stages: matching and verification. In particular, the best individual matching parts given a detection window are determined using an improved template matching algorithm. The hypothesis of the matched parts forming a human is then verified by employing a Bayesian-based model. The verification is not only based on the matching costs of individual parts but also how well the combining the matched parts satisfying the configuration constraints of the human body. Experimental results have shown that the proposed method is robust for detecting humans at multiple views and outperforms other template matching-based methods.

Keywords

era2015

Disciplines

Physical Sciences and Mathematics

Publication Details

Nguyen, D., Li, W. & Ogunbona, P. (2009). A part-based template matching method for multi-view human detection. International Conference Image and Vision Computing New Zealand (pp. 357-362). Wellington, New Zealand: IEEE.

This conference paper is available at Research Online: https://ro.uow.edu.au/infopapers/2125

A Part-based Template Matching Method for Multi-view Human Detection

Duc Thanh Nguyen, Wanqing Li, and Philip Ogunbona
Advanced Multimedia Research Lab, ICT Research Institute
School of Computer Science and Software Engineering
University of Wollongong, Australia

***Abstract*—This paper proposes a part-based template matching method for multi-view human detection. The proposed method includes two stages: matching and verification. In particular, the best individual matching parts given a detection window are determined using an improved template matching algorithm. The hypothesis of the matched parts forming a human is then verified by employing a Bayesian-based model. The verification is not only based on the matching costs of individual parts but also how well the combining the matched parts satisfying the configuration constraints of the human body. Experimental results have shown that the proposed method is robust for detecting humans at multiple views and outperforms other template matching-based methods.**

I. Introduction

Human detection from images and videos is a crucial step in human motion analysis and activity recognition. The challenge of the task arises from the numerous variations that human postures can assume and the complexity of the surrounding environment (e.g. cluttered background, crowded scene, etc.).

A number of approaches employing temporal, appearance and depth information [1] have been proposed in the literature, however, shape-based approach holds several advantages. First, shape information constitutes a good descriptor for humans in images and videos. Second, shape plays an important role in discriminating human from other types of objects. It is a robust descriptor when there is no information about human appearance such as colour or texture.

Generally, shape can be described explicitly by contours or implicitly by features obtained through training. For example, authors in [2], [3], [4], [5], [6], [7] have explicitly employed full body or body part [8], [9] contours as templates for detecting humans. The explicit use of shape information has the advantage of allowing for the variations of human poses and viewpoints but the templates are required to be given in advance. In the implicit use of shape [10], [11], [12], [13], [14], [15], [16], [17], shape features are determined automatically through training human/non-human patterns with classifiers such as SVM, AdaBoost, etc. The problem is then often formulated as the binary classification. However, since the features must be common and shared by training patterns, this approach limits the possible variations of human poses and viewpoints.

Motivated by the advantage of using shape templates in describing humans at various postures and viewpoints, this paper introduces a part-based template matching method for multi-view human detection with the following contributions. First, we present an improved template matching method that combines both spatial and orientation information in a simple and effective way. We then propose a part-based human detection method in which a human body structure is decomposed into different parts including top, bottom, left, and right profile. Individual body parts are then detected using the improved template matching algorithm. A Bayesian model is then employed to verify the detection results. In this model, we introduce the use of prior to represent body configuration constraints to reduce invalid combinations of part templates while consolidate credible detected postures. Experiments have verified the effectiveness of the proposed method and demonstrated the potential ability of the method to detect humans in multiple views.

The rest of the paper is organized as follows. In section II, we briefly review the related works. Section III describes the improved template matching algorithm. Section IV presents the two-stage human detection method. Experimental results along with some comparative analysis are presented in Section V. Section VI concludes the paper with remarks.

II. Related Work

Explicit use of human shapes often requires templates as 2-dimensional contours in various postures and viewpoints. Gavrila et al. [2] clustered full body human templates into a hierarchical structure where the similarity between two templates was defined by the Chamfer distance. For each sliding detection window, the best matching template is found by traversing the tree from root to leaf in depth-first search strategy. At each node in the tree, the Chamfer distance between the template and the detection window is calculated and compared with a threshold to determine whether the detection window contains a human or not. The threshold is determined based on the level of the hierarchical structure and the density of sliding detection window. This work then has been extended in [3] in which the thresholds were computed automatically using a probabilistic method. Since the above works focus on detection of a full human body, we refer to them as global detection methods.

On contradictory, methods approaching the problem of human detection by detecting the partial human body or body parts are considered as local detection methods. For example, in [8], [9], Lin et al. decomposed a human body structure

978-1-4244-4698-8/09/$25.00 ©2009 IEEE

into a hierarchial tree of body parts including head-torso, upper legs, and lower legs. The detection was performed sequentially by detecting individual body parts from the root (head-torso) to the leaf (lower legs) in the hierarchical tree. In [6], [7], although using full human body templates, Thanh et al. proposed a local template matching method in which the points on the templates (contours of human body shape) were weighted. These weights were obtained through training and indicate the importance of the corresponding points. In addition, the credibility of the template was treated as the prior and then employed to verify the detection hypotheses.

Some methods combine both local and global detection [4], [5], [18]. The general idea can be summarized as follows. First, codebooks corresponding to local shapes are defined. The relationship between local and global shapes is learned through training. Codebooks are further used to vote for global shapes and the Chamfer matching is then applied to the global shape [4] to select the best fit of the joint global and local detection response. In [18], a segmentation step was employed to consolidate the detection results.

Implicit use of the shape in human detection algorithms often proceeds by learning shape features from training data. As mentioned, this approach does not require contours representing human templates but shapes are defined through training human/non-human patterns. For example, Mohan et al. [10] used Haar wavelets trained by SVM to describe body parts while Wu et al. [11] introduced a so-called "edgelet" feature selected by a real and nested cascade Adaboost algorithm. Edgelet features were then fed into a cluster boosted tree (CBT) classifier for multi-view pedestrian detection [19].

In the work of Dalal et al. [12] histograms of oriented gradients (HOG) were introduced. HOGs of overlapping areas were concatenated into a vector and used to train a linear SVM. Extensions of the HOG based shape description have been found in [13], [14], [15], [16], [17]. For example, in [13], an "Integral Image" was employed to speed up the computation of HOG. Moreover, blocks with variable sizes and cascade Adaboost were used. In [14], edge orientation histogram (EOH), instead of HOG, was proposed, in which each vector was a scalar number. In addition, a meta-stage was added to cascade Adaboost to exploit the inter-stage information. Orientation features together with logical frameworks were used to detect body parts as in [15]. In [16], [17], in addition to the HOG shape feature, body configuration was employed to improve the detection performance.

III. Template Matching with Orientation Map

An intrinsic problem of the conventional template matching based on Chamfer distance transform (DT) is that it is sensitive and fragile in cluttered images thus leads to high false positive rate. This is due to the conventional matching method focuses only on the spatial distance. However, in the cluttered images, the background and foreground edges contribute the same spatial distance value on the DT image. Notice that we call spatial distance transform as conventional distance transform hereafter. Fig. 1 represents an example where the spatial

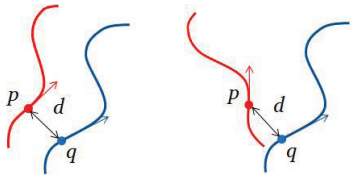


Fig. 1. The red curves and blue curves (best view in color) represent the image edges and template edges respectively while the (red/blue) arrows represent the corresponding edge orientations.

distances between the image point p and template point q in the left and the right image are same while the pair (p, q) in the left image actually presents the best match.

To reduce the affects of noisy edges, one can consider using the strength and orientation of edges. In this paper, the improved template matching proposed in [20] is employed (readers are referred to [20] for more details). In particular:

- A generalized distance transform (GDT) [21] is adopted so as to weight more on the strong edge points in computation of the distance transformed image.
- An orientation map (OM) is created simultaneously with the GDT image. Each value of the OM represents the edge direction of its nearest edge pixel.

A. Generalized Distance Transform (GDT)

Let $\mathcal{G}$ be a regular grid and $\Psi : \mathcal{G} \rightarrow \mathbb{R}$ a function on the grid. The GDT determined by Ψ can be defined as,

$$D_\Psi(p) = \min_{q \in \mathcal{G}} \{d(p, q) + \Psi(q)\} \tag{1}$$

where $d(p, q)$ is some measure of the distance between point p and q in the grid. Intuitively, for each point p we find a point q that is close to p, and for which $\Psi(q)$ is small. $\Psi(q)$ is defined as

$$\Psi(q) = \begin{cases} \frac{\eta}{\sqrt{I_x^2 + I_y^2}}, & \text{if } (q) \in e \\ \infty, & \text{otherwise} \end{cases} \tag{2}$$

where e represents the edge image, $I_x = \frac{\partial I}{\partial x}$ and $I_y = \frac{\partial I}{\partial y}$ are the horizontal and vertical gradients of the image I at position q. By defining $\Psi(\cdot)$ as in (2), we reduce the impact of weak edge points by placing more trust on strong edge points. In addition, using the algorithm proposed by Felzenszwalb and Huttenlocher [21], the GDT can be computed in $O(knm)$ time, where $n \times m$ is the image's size, k (= 2 in our case) indicates the number of dimensions.

B. Orientation Map (OM)

Let q^* be the closest edge point to the pixel p, that is,

$$q^* = \arg\min_{q \in \mathcal{G}} \{d(p, q) + \Psi(q)\}$$

and the orientation value at p is defined as,

$$O_\Psi(p) = \arctan(I_{x^*} / I_{y^*}) \tag{3}$$

where I_{x^*} and I_{y^*} are the gradients at q^*. In other words, the orientation of edge pixels will be propagated to their nearest

non-edge pixels. We can see that, $O_\Psi(p)$ and $D_\Psi(p)$ can be calculated simultaneously without increasing computational complexity.

C. Weighted Template Matching

Given a template T and a test image I, the matching cost or dissimilarity is defined as,

$$D(T, I) = \sum_{t \in T} w_t d_{T,I}(t) \tag{4}$$

where $d_{T,I}(t)$ is the dissimilarity between I and T at point t weighted by w_t. We define $d_{T,I}(t)$ as,

$$d_{T,I}(t) = \sqrt{\alpha D_\Psi^2(t) + (1-\alpha)\sin^2 |O_\Psi(t) - o(t)|} \tag{5}$$

where α is a parameter representing the importance of the spatial component relative to the orientation component. In our experiment, we set $\alpha = 0.5$. In (5), $o(t)$ is the orientation at point t in T and $D_\Psi(t)$ is normalized to $(0, 1)$ as,

$$D_\Psi(t) \leftarrow \exp(-\beta / max\{\varepsilon, \sqrt{D_\Psi(t)}\}) \tag{6}$$

where ε is a small positive number to avoid dividing by zero, $\beta = 1$ in our experiment. Notice that $0 < D(T, I) < 1$.

The values of w_t in (4) represent the importance of the point $t \in T$ and could be computed as in [6], [7]. As indicated in our previous works [6], [7], not all points of a template play the same role in matching. For instance, in case of human detection, the points along to the two curves of the head-shoulder template always appear in every head-shoulder pattern thus more discriminative compared with the points belonging to the arms, which are varied and dependent on human's postures. In this paper, the weight w_t of a point t is calculated simply based on the frequency that t appears given a set of similar templates. In particular, let Γ be a set of templates T representing a same part of human, w_t is then computed as,

$$w_t = \frac{\sum_{T' \in \Gamma}(1 - d_{T,T'}(t))}{\sum_{t' \in T'} \sum_{T' \in \Gamma}(1 - d_{T,T'}(t'))} \tag{7}$$

IV. Proposed Two-Stage Human Detection Method

One of the challenges of human detection is the articulation of humans in the scene. Global detection methods using template matching often employ 2D contours as templates to describe the shape of full human body. This leads to a trade-off problem between the robustness and efficiency in which the more templates are given, the more robustness the detection algorithm can afford but the more time consuming and less practical the detection method is. To solve this problem, one can decompose a full human body structure into a number of body parts in which individual parts are described by a small set of part templates. In addition, each individual part would contribute differently to recognition of the whole human body. We therefore adopt body part matching instead of full body matching. Each part detector will be assigned with a weight representing its importance to the overall matching cost. Furthermore, we employ the prior information in a Bayesian model to verify the combinations of individual parts.

The proposed method includes two processes: training and detection. In the training phase, presented in IV-A, a human body structure is modeled by a set of part templates and their combining relationship which is further called prior information. The proposed human detection method is then described in IV-B including two stages: matching and verification.

A. Learning Prior

A full human body is decomposed into 4 different parts including top (head-torso), bottom (legs), left, and right as shown in Fig. 2(b). The relationship between body parts is modeled by the joint probability of individual body parts being a human posture, i.e., $P(Human|t, b, l, r)$ where t, b, l, and r are individual body parts. To model this relationship, a number of full body human templates are first collected and labeled. Sample templates used in this step are shown in Fig. 2(a). Each template is centered in a 30×60 window and then divided into 4 different parts. For each type of parts, e.g. legs, sample templates are clustered and the mean templates are determined as the prototype part templates. Notice that Γ in (7) represents the clusters. Let T, B, L, R be the sets of prototypes for the top, bottom, left, and right parts respectively. In our implementation, $|T| = 5, |B| = 8, |L| = |R| = 6$. Fig. 2 shows the prototypes.

The next task of this procedure is to learn the relationship between part templates. This relationship represents the constraints on combinations of individual parts. It could also be described by the spatial constraints. In this paper, we simply use the co-occurrence of part templates to represent this relationship. Given a configuration $c = \{t, b, l, r\}$, where $t \in T, b \in B, l \in L, r \in R$, let $f(t, b, l, r)$ be the frequency that the parts $\{t, b, l, r\}$ co-occur. The conditional probability given a configuration c, $P(Human|c)$ is the prior and can be obtained by calculating the sigmoid function of the frequency as below,

$$\begin{aligned} P(Human|c) &= P(Human|t, b, l, r) \\ &= \frac{1}{1 + e^{-mf(t,b,l,r)}} \end{aligned} \tag{8}$$

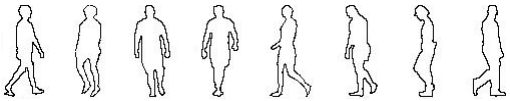
where m is an empirical parameter. The reason we do not compute the prior $P(Human|c)$ directly from the frequency is to avoid zero value for combinations of body parts not included in the training dataset. If $f(t, b, l, r) = 0$, $P(Human|t, b, l, r) = 0.5$.

B. Detection

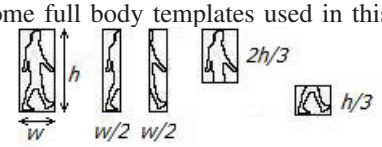
The problem of human detection is formulated as follows: Given the image I_W of a detection window W, determine whether the window contains a human or not by evaluate the following conditional probability.

$$P(Human|I_W) \geq \theta \tag{9}$$

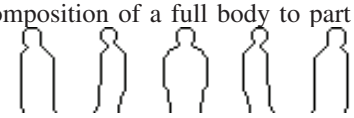
where θ is a threshold.



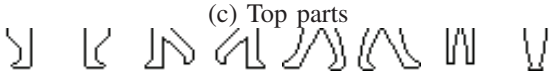
(a) Some full body templates used in this paper



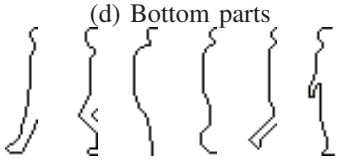
(b) Decomposition of a full body to part templates



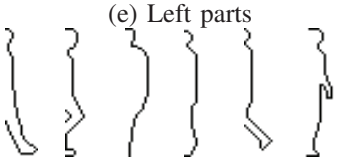
(c) Top parts



(d) Bottom parts



(e) Left parts



(f) Right parts

Fig. 2. Part templates used in the proposed human detection method.

We propose a two stage method consisting of matching and verification. In the matching step, given the image I_W, we find the best matching posture (configuration) $c^* = \{t^*, b^*, l^*, r^*\}$ as,

$$\begin{aligned} t^* &= \arg\min_{t \in T} D(t, I_W), \\ b^* &= \arg\min_{b \in B} D(b, I_W), \\ l^* &= \arg\min_{l \in L} D(l, I_W), \\ r^* &= \arg\min_{r \in R} D(r, I_W) \end{aligned} \tag{10}$$

where the matching cost is defined as in (4) and T, B, L, R are obtained as in IV-A.

For each detection window, the number of templates matched is $5{+}8{+}6{+}6 = 25$ (templates) to cover $5{\times}8{\times}6{\times}6 = 1440$ possible postures. Compared with full body detection approach, this is an advantage since the matching is performed on a small set of templates but can cover a variety of human postures. This fact will be represented in experimental results (see V) by comparing the performance of part-based detection and full body detection.

Once the best matching configuration, c^* is found, the set of its corresponding partial matching costs is denoted as $\Delta^* = \{D(t^*, I_W), D(b^*, I_W), D(l^*, I_W), D(r^*, I_W)\}$ and the verification is required to ascribe a degree of confidence on whether I_W contains a human. In the other words, we need to check the condition presented in (9). By replacing I_W by (c^*, Δ^*), $P(Human|I_W) = P(Human|c^*, \Delta^*)$ and the verification process can be stated as a conditional probability,

$$P(Human|c^*, \Delta^*) \geq \theta \tag{11}$$

The verification is conducted using the best matching configuration (c^*) and its corresponding partial matching costs (Δ^*). For the sake of simplicity of notation in the sequel, we drop the $*$ in the following discussion. Moreover, the set of partial matching costs Δ associated with each configuration c is considered as a random variable and also statistically independent of the configuration. Applying Bayes's theorem, we have,

$$\begin{aligned} P(Human|c, \Delta) &= \frac{P(c, \Delta|Human)P(Human)}{P(c, \Delta)} \\ &= \frac{P(c|Human)P(\Delta|Human)P(Human)}{P(c)P(\Delta)} \\ &= \frac{P(Human|c)P(Human|\Delta)}{P(Human)} \end{aligned} \tag{12}$$

In [2], [3], [8], the criterion for this verification is based on the matching cost. However, in this paper, we determine the confidence according to both the matching cost and prior information. The prior encodes the credibility we have in the best matching configuration. In (12), $P(Human|c)$ is identified as the prior since it indicates the degree of confidence that a given configuration c represents human posture and computed as in (8). $P(Human)$ can be assigned to a constant to control the margin between the accepted and rejected configurations. To evaluate $P(Human|c, \Delta)$, we assume that:

$$P(Human|\Delta) = \sum_{i=t,b,l,r} \omega_i P(Human|D(i, I_W)) \tag{13}$$

where $P(Human|D(i, I_W))$ is simply calculated as,

$$P(Human|D(i, I_W)) = 1 - D(i, I_W) \tag{14}$$

The values of $\omega_i, i = t, b, l, r$ represents the importance of each detected part i and $\sum \omega_i = 1$. Similarly to (7), in our implementation, $\omega_i, i = t, b, l, r$ are computed through training as follows. Let N be a set of positive samples containing one human centered inside. We define,

$$\omega_i = \frac{\sum_{I \in N} P(Human|D(i, I))}{\sum_{j=t,b,l,r} \sum_{I \in N} P(Human|D(j, I))} \tag{15}$$

V. Experimental Results

The proposed human detection method was evaluated on two multi-view pedestrian datasets: USC-C dataset [19] and Penn-Fudan dataset [18]. The USC-C dataset consists of 100 images with 232 pedestrians in multiple views including frontal/rear and profile. These 232 images are scanned at various scales (from 0.4 to 1.0). The Penn-Fudan dataset includes 170 images with 345 labeled pedestrians. The true and false positives are determined by comparing the detection results with true detections given in the ground truth using

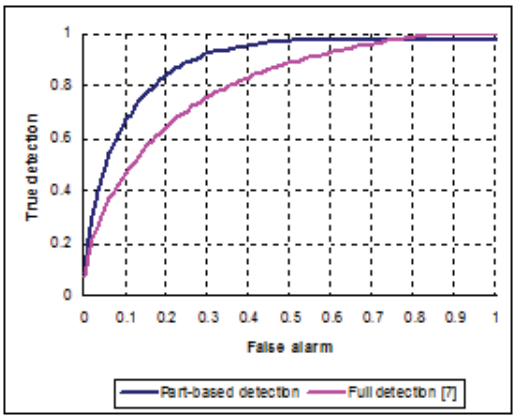


Fig. 3. ROC curves of the part-based human detection method and full body detection method [7] on the USC-C dataset. The result of [7] is generated from the original paper on the USC-C dataset.

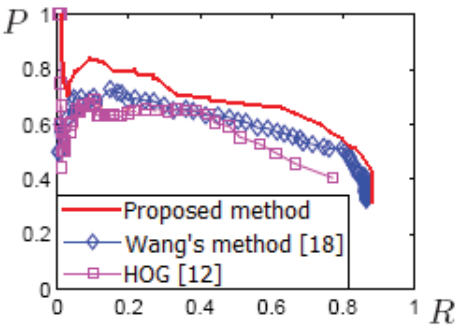


Fig. 4. PR Curves of the proposed method on the Penn-Fudan dataset. The results of Wang et al. and Dalal et al. are copied from [18].

the criteria proposed in [4]. The ROC (Receiver Operating Characteristic) and PR (Precision-Recall) curves achieved by the proposed method on the two datasets are shown in Fig. 3 and Fig. 4 respectively. Some detection results are shown in Fig. 5.

In addition to evaluation, we compare the proposed method with its variants and other state-of-the-art methods. On the USC-C dataset, we compare the performance of the part-based detection and full body detection proposed in the previous version [7] (see Fig. 3). For the full body detection, we selected 46 full body templates (Fig. 2(a) shows some of them) and applied the improved template matching. We also implemented and evaluated the hierarchical part-based template matching method proposed by Lin et al. [8]. Since the USC-C dataset contains un-occluded humans and for comparison of Lin's template matching model and ours, we ignored the occlusion processing. Although the effectiveness of the improved template matching has been proven for full body human detection [7] and general object detection [20], we also verified its robustness on the Lin's template matching model. In both cases (with and without using the improved template matching), the proposed method outperformed the hierarchical part-based template matching method proposed by Lin et al. For example, at the false positive rate of 0.2, the detection rate of Lin's method is ≈ 0.75 using the improved template matching and ≈ 0.42 using the conventional template matching. Similarly, at the false positive rate of 0.4, the detection rates of Lin's method corresponding to using and without using the improved template matching are ≈ 0.92 and ≈ 0.72 respectively. This fact once again represents the robustness of the improved template matching method.

On the Penn-Fudan dataset, we compare the proposed method with the work of Wang et al. [18] and Dalal et al. [12] (using HOG to encode human's shape). The PR curves of these methods are presented in Fig. 4.

VI. Conclusions and Future Works

This paper introduces a part-based human detection method using an improved template matching algorithm. The prosed human detection method is performed in a two-stage framework. Body parts are first determined using the improved template matching. Body configuration is then employed to verify the validation of detected parts. The proposed approach has some advantages. First, as shown by the experimental results, the improved template matching employing the orientation map and generalized distance transform weighting the distance transform on the strong edges improved the detection performance. In addition, the body part detection based approach is robust to detect humans at multi-viewpoints and actually outperforms full body detection approach while reducing the number of templates. The prior information might be extended to encode not only the constraints on view and pose validation but also the spatial locations of body parts. Finally, by approaching the problem of human detection using part templates, the proposed method is appropriate to detect multiple and partial occluded humans, on which detailed results will be reported in the near future.

References

[1] T. B. Moeslund, A. Hilton, and V. Krger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, pp. 90–126, 2006.

[2] D. M. Gavrila and V. Philomin, "Real-time object detection for smart vehicles," in *Proc IEEE International on Computer Vision*, vol. 1, 1999, pp. 87–93.

[3] D. M. Gavrila, "A Bayesian, exemplar-based approach to hierarchical shape matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1408–1421, 2007.

[4] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 878–885.

[5] E. Seemann, B. Leibe, and B. Schiele, "Multi-aspect detection of articulated objects," in *Proc IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1582–1588.

[6] N. D. Thanh, P. Ogunbona, and W. Li, "Human detection based on weighted template matching," in *Proc IEEE International Conference on Multimedia and Expo*, 2009.

[7] N. D. Thanh, W. Li, and P. Ogunbona, "A novel template matching method for human detection," in *Proc IEEE International Conference on Image Processing*, 2009.

[8] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, "Hierarchical part-template matching for human detection and segmentation," in *Proc IEEE International on Computer Vision*, 2007.

[9] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human detection and segmentation," in *Proc European Conference on Computer Vision*, vol. 4, 2008, pp. 423–436.

(a)

(b)

Fig. 5. Some results of human detection on USC-C (a) and Penn-Fudan dataset (b).

[10] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, pp. 349–361, 2001.

[11] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Proc IEEE International on Computer Vision*, 2005, pp. 90–97.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[13] Q. Zhu, S. Avidan, M. C. Yeh, and K. T. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1491–1498.

[14] Y. T. Chen and C. S. Chen, "A cascade of feed-forward classifiers for fast pedestrian detection," in *Proc Asian Conference on Computer Vision*, 2007, pp. 905–914.

[15] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis, "Bilattice-based logical reasoning for human detection," in *Proc IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[16] D. Tran and D. Forsyth, "Configuration estimates improve pedestrian finding," in *Proc Conference on Neural Information Processing Systems*, 2007.

[17] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[18] L. Wang, J. Shi, G. Song, and I. Shen, "Object detection combining recognition and segmentation," in *Proc Asian Conference on Computer Vision*, 2007, pp. 189–199.

[19] B. Wu and R. Nevatia, "Cluster boosted tree classifier for multi-view, multi-pose object detection," in *Proc IEEE International on Computer Vision*, 2007.

[20] N. D. Thanh, W. Li, and P. Ogunbona, "An improved template matching method for object detection," in *Proc Asian Conference on Computer Vision*, 2009.

[21] P. F. Felzenszwalb and D. P. Huttenlocher, "Distance transforms of sampled functions," Cornell Computing and Information Science, http://www.cs.cornell.edu/ dph/papers/dt.pdf, Tech. Rep., 2004.