Hierarchical Part-Template Matching for Human Detection and Segmentation

Zhe Lin, Larry S. Davis, David Doermann, and Daniel DeMenthon
Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742
{zhelin,lsd,doermann,daniel}@umiacs.umd.edu

Abstract

Local part-based human detectors are capable of handling partial occlusions efficiently and modeling shape articulations flexibly, while global shape template-based human detectors are capable of detecting and segmenting human shapes simultaneously. We describe a Bayesian approach to human detection and segmentation combining local part-based and global template-based schemes. The approach relies on the key ideas of matching a part-template tree to images hierarchically to generate a reliable set of detection hypotheses and optimizing it under a Bayesian MAP framework through global likelihood re-evaluation and fine occlusion analysis. In addition to detection, our approach is able to obtain human shapes and poses simultaneously. We applied the approach to human detection and segmentation in crowded scenes with and without background subtraction. Experimental results show that our approach achieves good performance on images and video sequences with severe occlusion.

1. Introduction

Human detection is a fundamental problem in video surveillance. It can provide an initialization for human segmentation. More importantly, robust human tracking and identification are highly dependent on reliable detection and segmentation in each frame, since better segmentation can be used to estimate more accurate and discriminative appearance models.

Previous approaches to human detection can be classified into two categories: shape-based approaches and blob-based approaches. **Shape-based approaches** are mostly developed for human detection in still images or moving videos. The shapes are modeled as local curve segments in [19, 11, 4, 20], modeled directly as a global shape model hierarchy in [5, 22], or implicitly represented by local or global descriptors in [10, 9, 13, 3, 21]. For highly articulated objects like humans, part-based representations have been shown to be very efficient for detection. For example, Mikolajczyk *et al*. [10] use local features for part detection and assemble the part detections probabilistically. Wu *et al*. [19] introduce edgelet part features for human detection. They extend this approach to a general object detection and segmentation approach by designing local shape-based classifiers [20]. One problem with these part-based detection approaches is that in very cluttered images too many detection hypotheses may be generated, and a robust assembly method (*e.g*. boosting) is thus needed to combine these detections. Recently, Shet *et al*. [14] propose a logical reasoning-based method for efficiently assembling part detections. On the other hand, Gavrila *et al*. [5] propose a more direct hierarchical template matching approach for global shape-based pedestrian detection. These shape-based detection methods can also be combined with appearance-based clustering for simultaneous detection and segmentation [8, 22, 18]. Shape-based approaches have the advantage that they do not require background subtraction, but they need to scan whole images and can generate many false alarms in cluttered regions.

In contrast, **blob-based approaches** are computationally more efficient but have a common problem that the results depend crucially on background subtraction. These approaches are mostly developed for detecting and tracking humans under occlusion. Some earlier methods [16, 7] model the human tracking problem by a multi-blob observation likelihood given a human configuration. Zhao *et al*. [24] introduce an MCMC-based optimization approach to human segmentation from foreground blobs. They detect heads by analyzing edges surrounding binary foreground blobs, formulate the segmentation problem in a Bayesian framework, and optimize by modeling jump and diffusion dynamics in MCMC to traverse the complex solution space. Following this work, Smith *et al*. [15] propose a similar trans-dimensional MCMC model to track multiple humans using particle filters. Later, an EM-based approach is proposed by Rittscher *et al*. [12] for foreground blob segmentation. Recently, Zhao *et al*. [23] use a part-based human body model to fit binary blobs and track humans.

We propose a hierarchical part-template matching approach for human detection and segmentation. The approach takes advantages of both local part-based and global

template-based human detectors by decomposing global shape models and constructing a part-template tree to model human shapes flexibly and efficiently. Edges are matched to the part-template tree efficiently to determine a reliable set of human detection hypotheses. Shape segmentations and poses are estimated automatically through synthesis of part detections. The set of detection hypotheses is optimized under a Bayesian MAP framework based on global likelihood re-evaluation and fine occlusion analysis. For meeting the requirement of real-time surveillance systems, we also combined the approach with background subtraction to increase efficiency, where region information provided by foreground blobs is combined with shape information from the original image in a generalized joint likelihood model.

2. Bayesian Problem Formulation

We model the detection and segmentation problem as a Bayesian MAP optimization:

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

where I denotes the image observation, $\mathbf{c} = \{\mathbf{h}_1, \mathbf{h}_2, ...\mathbf{h}_n\}$ denotes a human configuration (a set of human hypotheses), and n denotes the number of humans in the configuration. $\{\mathbf{h}_i = (\mathbf{x}_i, \theta_i)\}$ is an individual hypothesis which consists of foot position[1] $\mathbf{x}_i$ and corresponding human model parameter θ_i which are explained in Section 3. Using Bayes Rule, Equation 1 can be decomposed into a joint likelihood $P(I|\mathbf{c})$ and a prior $P(\mathbf{c})$ as follows:

$$P(\mathbf{c}|I) = \frac{P(I|\mathbf{c})P(\mathbf{c})}{P(I)} \propto P(I|\mathbf{c})P(\mathbf{c}). \tag{2}$$

We assume a uniform prior, hence the MAP problem reduces to maximizing the joint likelihood.

2.1. Joint Likelihood Model

Previous approaches [16, 7, 24] model the human detection and tracking problem by a multi-blob observation likelihood based on object-level and configuration-level likelihood. In [19], the joint likelihood is modeled as the probability of part-detection responses given a set of human hypotheses.

We decompose the image observation, I, into shape observation I_s (edge image) and region observation I_r (binary foreground image from background subtraction) assuming independence between the shape and region information. Then, the joint likelihood $P(I|\mathbf{c})$ is modeled as:

$$P(I|\mathbf{c}) = P(I_s|\mathbf{c})P(I_r|\mathbf{c}), \tag{3}$$

[1]Here, we choose the foot point as a reference to represent and search for human shapes. A foot point is defined as the bottom center point of a human bounding box.

where $P(I_s|\mathbf{c})$ and $P(I_r|\mathbf{c})$ denote shape likelihood and region likelihood respectively. The region observation is optional and we set $P(I_r|\mathbf{c}) = 1$ or equivalently $P(I|\mathbf{c}) = P(I_s|\mathbf{c})$ when background subtraction is not used.

3. Hierarchical Part-Template Matching

3.1. Tree-Structured Part-Template Hierarchy

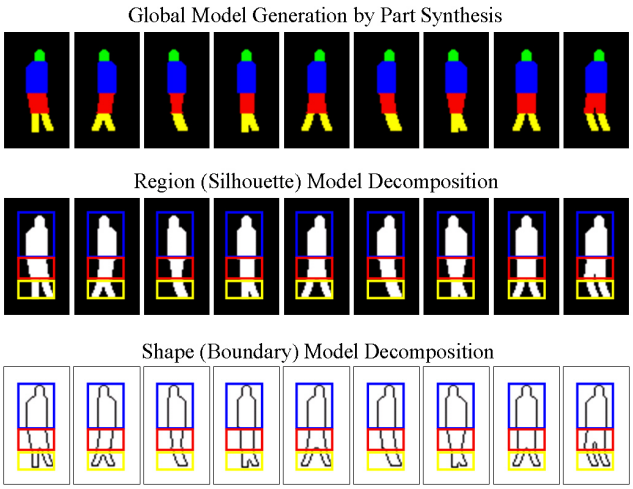
We take advantages of local part-based and global shape template-based approaches by combining them in a unified top-down and bottom-up search scheme. Specifically, we extend the hierarchical template matching method in [5] by decomposing the global shape models into parts and constructing a new part template-based hierarchical tree as shown in Figure 1(b).

We first generate a flexible set of global shape models by part synthesis (Figure 1(a)). For modeling human side views and front/back views individually, we represent the body with six part regions - (head, torso, pair of upper-legs, pair of lower-legs). Each part region is modeled by a horizontal parallelogram (five degrees of freedom) characterized by its position, size and orientation parameters. Thus, the total number of degrees of freedom is $5 \times 6 = 30$. For simplicity, we use only six degrees of freedom (head position, torso width, orientations of upper/lower legs) given the torso position as the reference. Heads and torsos are simplified to vertical rectangles (with rounded shapes at corners). The selected six parameters are discretized into $\{3, 2, 3, 3, 3, 3\}$ values. Finally, the part regions are independently collected and grouped to form $3 \times 2 \times 3 \times 3 \times 3 \times 3 = 486$ global shape models (Figure 1(a)).

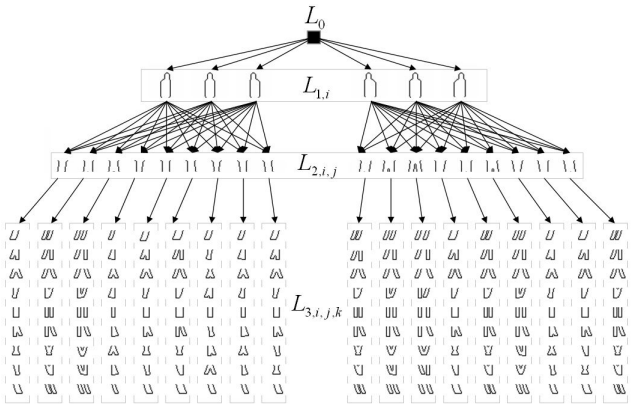
Next, silhouettes and boundaries are extracted from the set of generated global shape models and decomposed into three parts (head-torso, upper legs and lower legs) as shown in Figure 1(a). The parameters of the three parts ht, ul, ll are denoted as θ_{ht}, θ_{ul} and θ_{ll}, where each parameter represents the index of the corresponding part in the part-template tree. Then, the tree-structured part-template hierarchy is constructed by placing the decomposed part regions and boundary fragments into a tree as illustrated in Figure 1(b). The tree has four layers denoted as L_0, L_1, L_2, L_3, where L_0 is the root node which is set to be empty, L_1 consists of side-view head-torso templates $L_{1,i}, i = 1, 2, 3$ and front/back-view head-torso templates $L_{1,i}, i = 4, 5, 6$, and similarly, L_2 and L_3 consists of upper and lower leg poses for side and front/back views.

3.2. Part-Template Likelihood

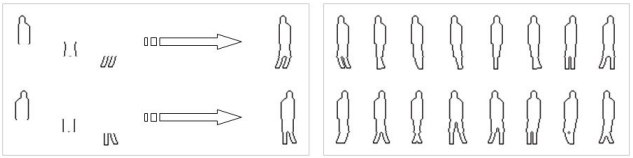
A part template T is characterized by its boundary and coverage region (Figure 1(a)). We match individual part-templates using both shape and region information (when region information is available from background subtrac-



(a) Generation of global shape models by part synthesis, decomposition of global silhouette and boundary models into region and shape part-templates



(b) Part-template tree characterized by both shape and region information



(c) Shape segmentation by synthesizing matched part-templates (designated by a path from L_0 to L_3)

Figure 1. An illustration of the part-template tree and its construction process.

tion). Shape information is measured by chamfer matching and region information is measured by part foreground coverage density.

For a foot candidate pixel $\mathbf{x}$ in the image, the likelihood $P(I|\mathbf{x},\theta_j)$ for a part template-T_{θ_j}, $j \in \{ht, ul, ll\}$ is decomposed into the part-shape likelihood $P(I_s|\mathbf{x},\theta_j)$ and the part-region likelihood $P(I_r|\mathbf{x},\theta_j)$ as follows:

$$P(I|\mathbf{x},\theta_j) = P(I_s, I_r|\mathbf{x},\theta_j) = P(I_s|\mathbf{x},\theta_j)P(I_r|\mathbf{x},\theta_j). \quad (4)$$

The part-shape likelihood is modeled by the chamfer score in an exponentially scaled distance transform image as follows:

$$P(I_s|\mathbf{x},\theta_j) = D_{chamfer}(\mathbf{x}, T_{\theta_j}), \quad (5)$$

$$D_{chamfer}(\mathbf{x}, T_{\theta_j}) = \frac{1}{|T_{\theta_j}|}\sum_{\mathbf{t}\in T_{\theta_j}} d_I^{'}(\mathbf{x}+\mathbf{t}), \quad (6)$$

where T_{θ_j} represents the part-template defined by parameter θ_j, $\mathbf{t} \in T_{\theta_j}$ represents the individual pixels in the template, and $D_{chamfer}(\mathbf{x}, T_{\theta_j})$ represents the average chamfer distance for foot candidate pixel $\mathbf{x}$. $d_I^{'}$ is a scaled distance transform image obtained by the following exponential transformation:

$$d_I^{'}(\mathbf{y}) = exp(-\beta d_I(\mathbf{y})), \quad (7)$$

where d_I is the Euclidean distance transform image and β is a constant factor.

When region information is not available, we set the part-region likelihood as $P(I_r|\mathbf{x},\theta_j) = 1$, otherwise, it is calculated by the part foreground coverage density $\gamma(\mathbf{x},\theta_j)$ which is defined as the proportion of the foreground pixels covered by the part-template T_{θ_j} at pixel $\mathbf{x}$.

We find the maximum likelihood estimate $\theta_j^*(\mathbf{x})$ as follows:

$$\theta_j^*(\mathbf{x}) = \arg\max_{\theta_j\in\Theta_j} P(I|\mathbf{x},\theta_j), \quad (8)$$

where Θ_j denotes the parameter space of part-template T_{θ_j}, and $P(I|\mathbf{x},\theta_j)$ denotes the part-template likelihood for pixel $\mathbf{x}$ and part-template T_{θ_j}.

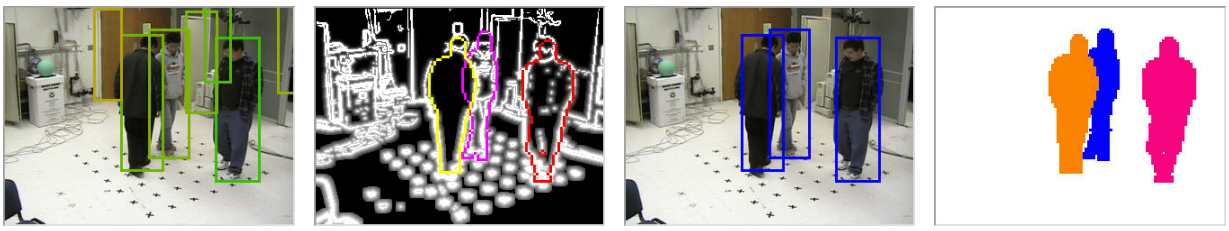
3.3. Hierarchical Part-Template Matching

Given ground plane homography information (see Section 5.1), we match the off-line constructed part-template tree to the edge map hierarchically.

Hierarchical part-template matching provides estimates for the model parameters $\theta^*(\mathbf{x})$ for every foot candidate pixel $\mathbf{x}$ in the image. We define a flexible likelihood function (a function of weight vector $\mathbf{w}$) here for evaluating likelihood for any parts or part combinations. The object-level likelihood function $P(I|\mathbf{x})$ for foot candidate pixel $\mathbf{x}$ is now expressed as follows:

$$P^{\mathbf{w}}(I|\mathbf{x}) = \sum_j w_j P(I|\mathbf{x},\theta_j^*(\mathbf{x})), \quad (9)$$

where $\mathbf{w} = \{w_j, j = ht, ul, ll\}$ is an importance weight vector to calculate a likelihood value for different parts or part combinations. For example, $\{w_{ht} = w_{ul} = w_{ll} = 1/3\}$ corresponds to a full body detector and $\{w_{ht} = 0, w_{ul} = w_{ll} = 1/2\}$ corresponds to a leg detector. The importance weights are normalized to satisfy $\sum_j w_j = 1$.



(a) (b) (c) (d)

Figure 2. An example of detection process without background subtraction. (a) Initial set of human detection hypotheses, (b) Human shape segmentations, (c) Detection result, (d) Segmentation result (final occlusion map).

Algorithm 1 Hierarchical Part-Template Matching

For each pixel $\mathbf{x}$ in the image, we adaptively search over scales distributed around the expected human size (w_0, h_0) estimated by foot-to-head plane homography and an average aspect ratio Δ.

1) We match the set of head-torso shape templates in layer L_1 with the image and estimate the maximum likelihood solution θ^*_{ht}.

2) Based on the part-template estimate θ^*_{ht} (either side or frontal view template), we match the upper leg template models and the lower leg template models to find the maximum likelihood solution for leg layers, the estimated leg part-template parameters are denoted as θ^*_{ul} and θ^*_{ll}.

3) We estimate human shapes by combining the maximum likelihood estimates of the local part-templates, and return the synthesized model parameters $\theta^* = \{\theta^*_{ht}, \theta^*_{ul}, \theta^*_{ll}\}$.

We have seven part or part-combination detectors, and if the head-torso is decomposed further into head-shoulder and torso, the number of detectors can be as high as 15.

Suppose we use K part detectors, $D_k, k = 1, 2...K$ corresponding to K weight vectors $\mathbf{w}_k, k = 1, 2...K$ for each foot candidate pixel $\mathbf{x}$ in the image. The likelihoods for these part detectors are calculated with the object-level likelihood function (Equation 9). We choose the final object-level likelihood $P(I|\mathbf{x})$ for foot candidate pixel $\mathbf{x}$ by maximizing the K detector responses:

$$P(I|\mathbf{x}) = \max_k P^{\mathbf{w}_k}(I|\mathbf{x}), \tag{10}$$

We threshold the final likelihood map by a detection threshold T and merge nearby weak responses to strong responses and adaptively select modes. This step can also be performed by local maximum selection after smoothing the likelihood image. Then, the generated set of human hypotheses is denoted as:

$$O = \{o_1, o_2, ...o_N\} = \{(\mathbf{x}_1, \theta^*(\mathbf{x}_1)), (\mathbf{x}_2, \theta^*(\mathbf{x}_2)), ...(\mathbf{x}_N, \theta^*(\mathbf{x}_N))\}, \tag{11}$$

and the corresponding likelihoods are denoted as $L(o_i), i = 1, 2...N$.

4. Optimization: Maximizing the Joint Likelihood

Suppose we have an initial set of human hypotheses $O = \{o_1, o_2, ...o_N\}$ obtained from hierarchical part template matching. The remaining task is to estimate its best subset through optimization. This is equivalent to maximize the joint likelihood $P(I|\mathbf{c})$ (Equation 3) with respect to the configuration $\mathbf{c}$.

4.1. Modeling the Joint Likelihood

If region information is not available, we set the region likelihood as $P(I_r|\mathbf{c}) = 1$, otherwise, it is calculated by the global coverage density of the binary foreground regions:

$$P(I_r|\mathbf{c}) = \frac{\Gamma(\mathbf{c})}{\Gamma_{fg}}, \tag{12}$$

where Γ_{fg} denotes the area of the foreground regions and $\Gamma(\mathbf{c})$ denotes the area of the foreground regions covered by the configuration $\mathbf{c}$. Intuitively, the more the foreground is covered by the configuration $\mathbf{c}$, the higher the probability $P(I_r|\mathbf{c})$. Areas covered by the hypotheses and located outside the foreground regions are not penalized here but considered in foot candidate region detection in Section 5.2. In fact, the region likelihood (Equation 12) has a bias towards more detections, but the bias is compensated for by the shape likelihood (Equation 13) (which involves a direct multiplication of individual likelihoods), since adding redundant hypotheses will decrease the shape likelihood.

The shape observation I_s now can be reduced to $o_1, o_2, ..., o_N$ since we only select the best subset from this initial set of hypotheses. This allows us to further decompose the shape likelihood as a product of likelihoods (assuming independence between each observation o_i given the configuration $\mathbf{c}$):

$$P(I_s|\mathbf{c}) = P(o_1, o_2, ..., o_N|\mathbf{c}) = \prod_{i=1}^{N} P(o_i|\mathbf{c}). \tag{13}$$

For evaluating the conditional probability $P(o_i|\mathbf{c})$, we need to model the occlusion status between different hypotheses in the configuration $\mathbf{c}$. For simplicity, we assume a known or fixed occlusion ordering for $\mathbf{c}$. Directly using the object-level likelihood $L(o_i)$ to model $P(o_i|\mathbf{c})$ will have problems since it only represents the strongest part response. We need to globally *re-evaluate* the object-level likelihood of each hypothesis o_i based on fine occlusion analysis; that is, we calculate the global shape likelihood only for the un-occluded parts when calculating the chamfer scores. This occlusion compensation-based likelihood re-evaluation scheme is effective in rejecting most false alarms while retaining the true detections.

Since we aim to select the best subset of O as our optimization solution $\mathbf{c}^*$, we assume $\mathbf{h}_j \in O, j = 1, 2...n$. We can treat the individual conditional probability $P(o_i|\mathbf{c})$ as a decision likelihood with o_i as the observation and $\mathbf{c}$ as the decision. Suppose the set of hypotheses O consists of n_{tp} true positives (tp), n_{tn} true negatives (tn), n_{fp} false positives (fp), and n_{fn} false negatives (fn). The decision rules (for the detection threshold T) for each observation o_i are defined as follows:
1. $P(o_i|\mathbf{c}) = p_{tp}$ if $o_i \in \mathbf{c}$ and $L(o_i|I_{occ}) \geq T$;
2. $P(o_i|\mathbf{c}) = p_{fp}$ if $o_i \in \mathbf{c}$ and $L(o_i|I_{occ}) < T$;
3. $P(o_i|\mathbf{c}) = p_{tn}$ if $o_i \notin \mathbf{c}$ and $L(o_i|I_{occ}) \geq T$;
4. $P(o_i|\mathbf{c}) = p_{fn}$ if $o_i \notin \mathbf{c}$ and $L(o_i|I_{occ}) < T$,
where I_{occ} denotes the occlusion map generated from the configuration $\mathbf{c}$ and $L(o_i|I_{occ})$ denotes the occlusion-compensated (re-evaluated) object-level likelihood. The probabilities p_{tp}, p_{fn}, p_{fp}, and p_{tn} are set to $p_{tp} = p_{fn} = \alpha$ and $p_{fp} = p_{tn} = 1 - \alpha$ (where $\alpha > 0.5$) for the current implementation. Finally, the shape likelihood (Equation 13) can be expressed as: $P(I_s|\mathbf{c}) = p_{tp}^{n_{tp}} p_{fp}^{n_{fp}} p_{tn}^{n_{tn}} p_{fn}^{n_{fn}} = \alpha^{(n_{tp}+n_{fn})}(1-\alpha)^{(n_{fp}+n_{tn})}$.

4.2. Optimization based on Likelihood Re-evaluation

We order the hypotheses in decreasing order of vertical (or y) coordinate as in [19]. This is valid for many surveillance videos with ground plane assumption, since the camera is typically looking obliquely down towards the scene. For simplicity, we assume $o_1, o_2, ..., o_N$ is such an ordered list. Starting from an empty scene, the optimization is performed based on iterative filling of humans based on occlusion compensation and likelihood re-evaluation.

An example of the detection and segmentation process is shown in Figure 2. Note that initial false detections are rejected in the final detection based on likelihood re-evaluation, and the occlusion map is accumulated to form the final segmentation.

Algorithm 2 Optimization algorithm

Given an ordered list of hypotheses $o_1, o_2, ..., o_N$,
initialize the configuration as $\mathbf{c} = \phi$, the occlusion map I_{occ} as empty (white image), and the joint likelihood as $P(I|\mathbf{c}) = 0$.
for $i = 1 : N$
1. re-evaluate the object-level likelihood of hypothesis o_i based on the current occlusion map I_{occ}, *i.e.* calculate $L(o_i|I_{occ})$.
2. if $L(o_i|I_{occ}) \geq T$ and $P(I|o_i \cup \mathbf{c}) > P(I|\mathbf{c})$, $o_i \mapsto \mathbf{c}$.
3. update the occlusion map I_{occ} using the current configuration $\mathbf{c}$.
endfor
return the configuration $\mathbf{c}$ and occlusion map I_{occ}.

5. Combining with Calibration and Background Subtraction

We can also combine the shape-based detector with background subtraction and calibration in a unified system.

5.1. Scene-to-Camera Calibration

If we assume that humans are moving on a ground plane, ground plane homography information can then be estimated off-line and used to efficiently control the search for humans instead of searching over all scales at all positions. A similar idea has been explored by Hoiem *et al.* [6] combining calibration and segmentation. To obtain a mapping between head points and foot points in the image, *i.e.* to estimate expected vertical axes of humans, we simplify the calibration process by estimating the homography between the head plane and the foot plane in the image [12]. We assume that humans are standing upright on an approximate ground plane viewed by distant camera relative to the scene scale, and that the camera is located higher than a typical person's height. We define the homography mapping as $\mathbf{f} = P_f^h : F \mapsto H$, where $F, H \in \mathbb{P}^2$. Under the above assumptions, the mapping $\mathbf{f}$ is one-to-one correspondence so that given an off-line estimated 3×3 matrix P_f^h, we can estimate the expected location of the corresponding head point $p_h = \mathbf{f}(p_f)$ given an arbitrary foot point p_f in the image. The homography matrix is estimated by the least squares method based on $L >> 4$ pairs of foot and head points pre-annotated in some frames. An example of the homography mapping is shown in Figure 3.

5.2. Combining with Background Subtraction

Given the calibration information and the binary foreground image from background subtraction, we estimate the binary foot candidate regions R_{foot} as follows: we first find all foot candidate pixels $\mathbf{x}$ with foreground coverage density $\gamma_{\mathbf{x}}$ larger than a threshold ξ. Given the estimated

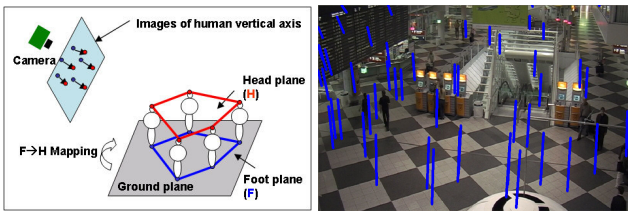


Figure 3. Simplified scene-to-camera calibration. Left: Interpretation of the foot-to-head plane homography mapping. Right: An example of the homography mapping. 50 sample foot points are chosen randomly and corresponding head points and human vertical axes are estimated and superimposed in the image.

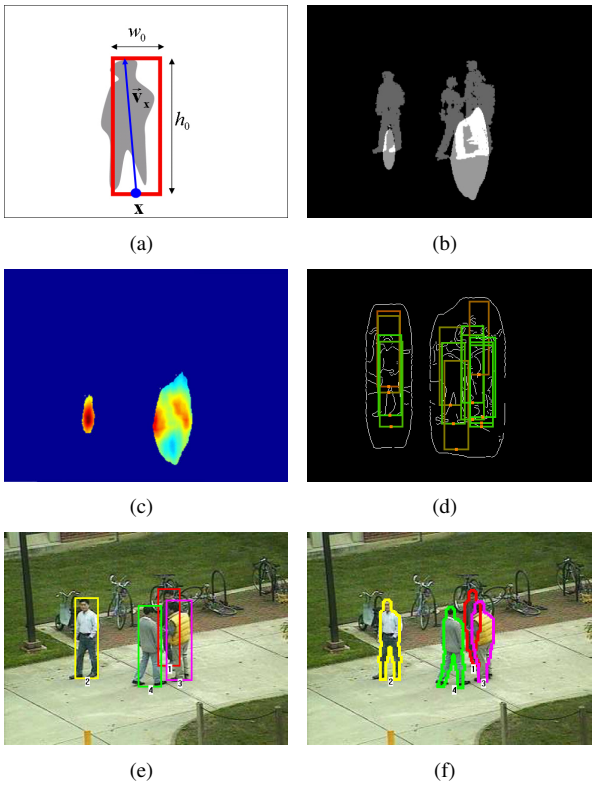


Figure 4. An example of the detection process with background subtraction. (a) Adaptive rectangular window, (b) Foot candidate regions R_{foot} (lighter regions), (c) Object-level (foot-candidate) likelihood map by the hierarchical part-template matching (where red color represents higher probabilities and blue color represents lower probabilities), (d) The set of human hypotheses overlaid on the Canny edge map in the augmented foreground region (green boxes represent higher likelihoods and red boxes represent lower likelihoods), (e) Final human detection result, (f) Final human segmentation result.

human vertical axis $\vec{\mathbf{v}}_{\mathbf{x}}$ at the foot candidate pixel $\mathbf{x}$, $\gamma_{\mathbf{x}}$ is defined as the proportion of foreground pixels in an adaptive rectangular window $W(\mathbf{x}, (w_0, h_0))$ determined by the foot candidate pixel $\mathbf{x}$. The foot candidate regions R_{foot} are defined as: $R_{foot} = \{\mathbf{x} | \gamma_{\mathbf{x}} \geq \xi\}$. The window coverage is efficiently calculated using integral images [17]. We detect edges in the augmented foreground regions R_{afg} which are generated from the foot candidate regions R_{foot} by taking the union of the rectangular regions determined by each foot candidate pixel $p_f \in R_{foot}$, adaptively based on the estimated human vertical axes. Figure 4 shows an example.

6. Experimental Results

In order to quantitatively evaluate the performance of our detector, we use the overlap measure defined in [9]. The overlap measure is calculated as the smaller value of the area ratios of the overlap region and the ground truth annotated region/detection region. If the overlap measure of a detection is larger than a certain threshold $\eta = 0.5$, we regard the detection as correct.

6.1. Results without Background Subtraction

We compared our human detector with Wu *et al*. [19] and Shet *et al*. [14] on USC pedestrian dataset-B [19] which contains 54 grayscale images with 271 humans. In these images, humans are heavily occluded by each other and partially out of the frame in some images. Note that no background subtraction is provided for these images. Figure 5 shows some example results of our detector and Figure 6 shows the comparison result as ROC curves. Our detector obtained better detection performance than the others when allowing more than 10 false alarms out of total of 271 humans, while detection rate decreased significantly when the number of false alarms was reduced to 6 out of 271. Proper handling of edge sharing problem would reduce the number of false alarms further while maintaining the detection rates. The running time of [19] for processing an 384×288 image is reported as about 1 frame/s on a Pentium 2.8GHz machine, while our current running time for a same sized image is 2 frames/s on a Pentium 2GHz machine.

6.2. Results with Background Subtraction

We also evaluated our detector on two challenging surveillance video sequences using background subtraction. The first test sequence (1590 frames) is selected from the Caviar Benchmark Dataset [1] and the second one (4836 frames) is selected from the Munich Airport Video Dataset [2]. The foreground regions detected from background subtraction are very noisy and inaccurate in many frames. From example results in Figure 7, we can see that our proposed approach achieves good performance in accurately detecting humans and segmenting the boundaries even under severe occlusion and very inaccurate background subtraction. Also, from the results, we can see that the shape estimates automatically obtained from our approach are quite

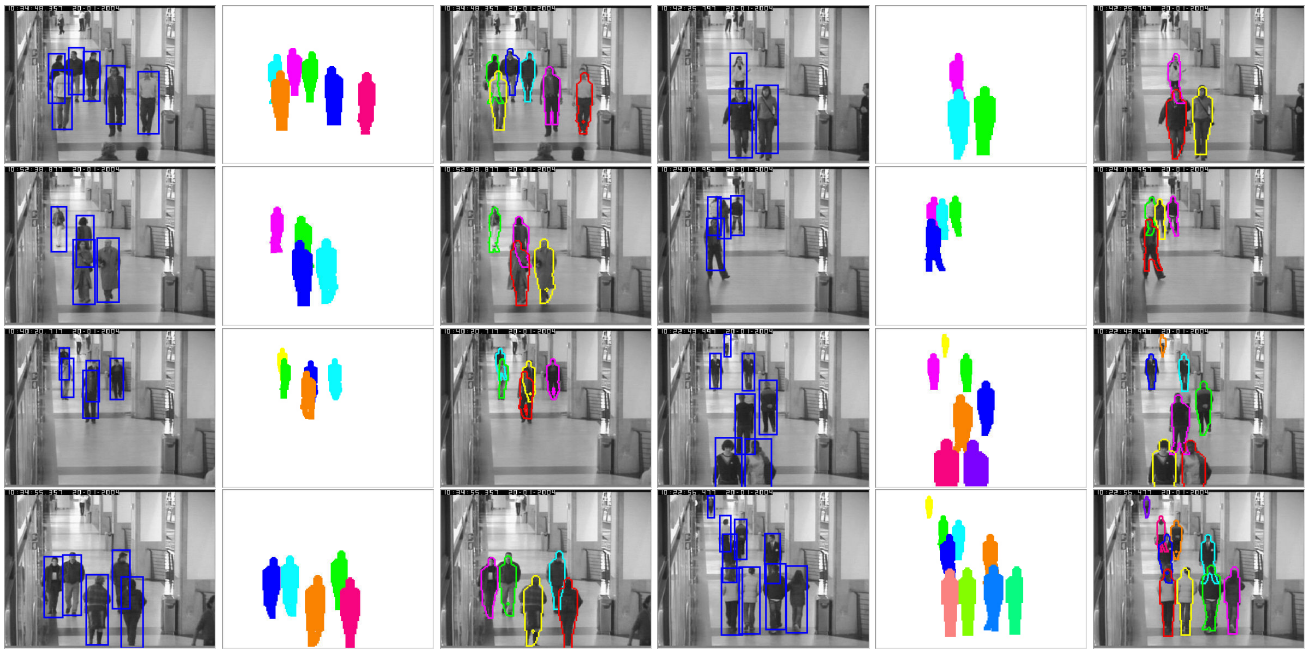


Figure 5. Detection and segmentation results (without background subtraction) for USC pedestrian dataset-B.

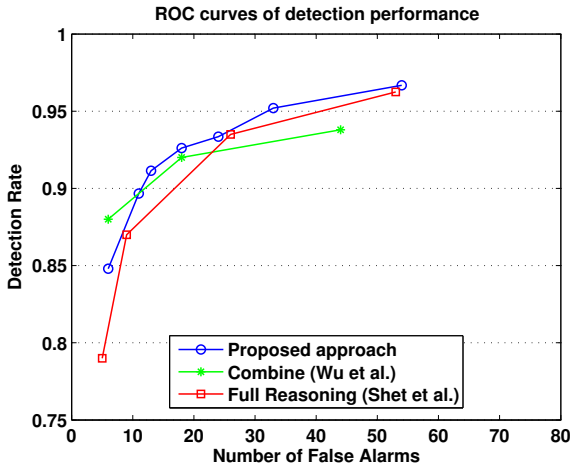


Figure 6. Evaluation of detection performance on USC pedestrian dataset-B (54 images with 271 humans). Results of [19] and [14] are copied for the comparison purpose.

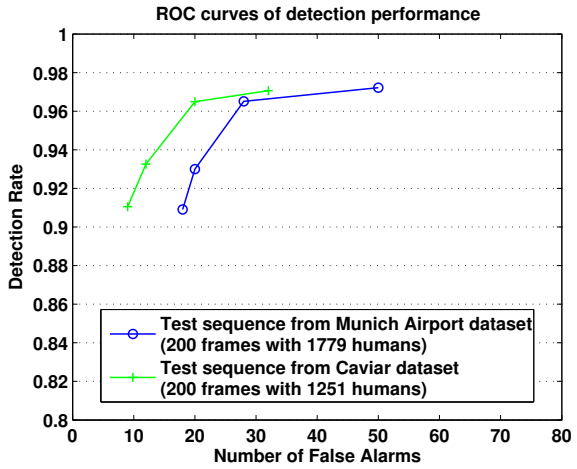


Figure 8. Evaluation of detection performance on two test sequences from Munich Airport dataset and Caviar dataset.

accurate. Some misaligned shape estimates are generated mainly due to low contrast and/or background clutter.

We evaluated the detection performance quantitatively on 200 selected frames from each video sequence. Figure 8 shows the ROC curves for the two sequences. Most false alarms are generated by cluttered background areas incorrectly detected as foreground by background subtraction. Misdetections (true negatives) are mostly due to the lack of edge segments in the augmented foreground region or complete occlusion between humans. Our system is implemented in C++ and currently runs at about 2 frames/s (without background subtraction) and 5 frames/s (with background subtraction) for 384×288 video frames on a Pentium-M 2GHz Machine.

7. Conclusions

A hierarchical part-template matching approach is employed to match human shapes with an edge map to de-

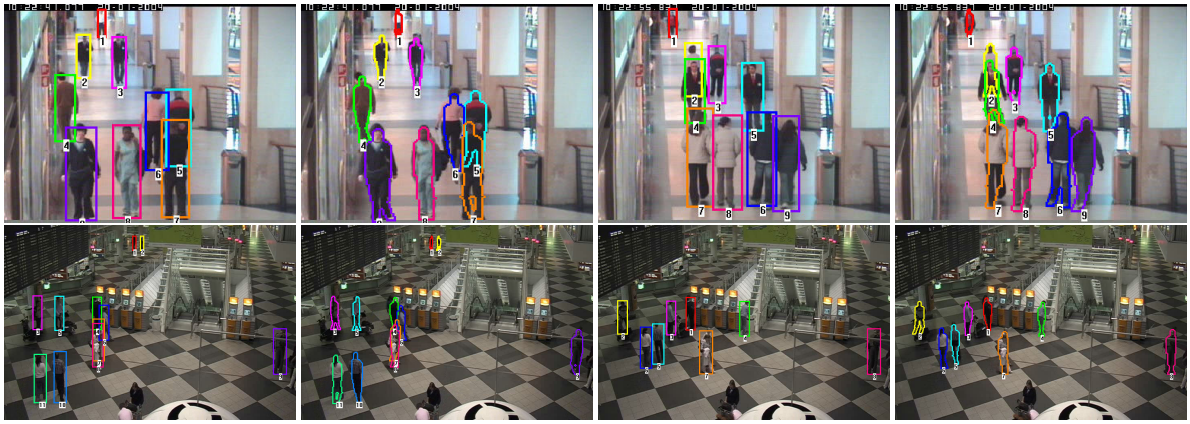


Figure 7. Detection and segmentation results (with background subtraction) for Caviar data [1] and Munich Airport data [2].

tect and segment humans simultaneously. Local part-based and global shape-template based approaches are combined to detect human hypotheses reliably and optimize them through likelihood re-evaluation and fine occlusion analysis under a unified Bayesian MAP framework. In addition to detection, human shapes and poses are segmented automatically through the detection process. The approach is applied to human detection and segmentation in crowded videos with and without background subtraction. The results demonstrate that the proposed part-template tree model captures the articulations of the human body, and detects humans robustly and efficiently. We are currently combing the approach with appearance-based segmentation to improve the result of shape segmentations.

Acknowledgement

This research was funded in part by the U.S. Government VACE program.

References

[1] Caviar Dataset: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[2] Munich Airport Video Dataset.

[3] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. *CVPR*, 2005.

[4] V. Ferrari, T. Tuytelaars, and L. V. Gool. Object Detection by Contour Segment Networks. *ECCV*, 2006.

[5] D. M. Gavrila and V. Philomin. Real-Time Object Detection for SMART Vehicles. *ICCV*, 1999.

[6] D. Hoiem, A. Efros, and M. Hebert. Putting Objects in Perspective. *CVPR*, 2006.

[7] M. Isard and J. MacCormick. BraMBLe: A Bayesian Multiple-Blob Tracker. *ICCV*, 2001.

[8] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj Cut. *CVPR*, 2005.

[9] B. Leibe, E. Seemann, and B. Schiele. Pedestrian Detection in Crowded Scenes. *CVPR*, 2005.

[10] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human Detection based on A Probabilistic Assembly of Robust Part Detectors. *ECCV*, 2004.

[11] A. Opelt, A. Pinz, and A. Zisserman. A Boundary-Fragment-Model for Object Detection. *ECCV*, 2006.

[12] J. Rittscher, P. H. Tu, and N. Krahnstoever. Simultaneous Estimation of Segmentation and Shape. *CVPR*, 2005.

[13] E. Seemann, B. Leibe, and B. Schiele. Multi-Aspect Detection of Articulated Objects. *CVPR*, 2006.

[14] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis. Bilattice-based Logical Reasoning for Human Detection. *CVPR*, 2007.

[15] K. Smith, D. G. Perez, and J. M. Odobez. Using Particles to Track Varying Numbers of Interacting People. *CVPR*, 2005.

[16] H. Tao, H. Sawhney, and R. Kumar. A Sampling Algorithm for Detecting and Tracking Multiple Objects. *ICCV Workshop on Vision Algorithms*, 1999.

[17] P. Viola and M. Jones. Robust Real-time Object Detection. *ICCV*, 2001.

[18] J. Winn and J. Shotton. The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects. *CVPR*, 2006.

[19] B. Wu and R. Nevatia. Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors. *ICCV*, 2005.

[20] B. Wu and R. Nevatia. Simultaneous Object Detection and Segmentation by Boosting Local Shape Feature based Classifier. *CVPR*, 2007.

[21] Y. Wu, T. Yu, and G. Hua. A Statistical Field Model for Pedestrian Detection. *CVPR*, 2005.

[22] L. Zhao and L. S. Davis. Closely Coupled Object Detection and Segmentation. *ICCV*, 2005.

[23] Q. Zhao, J. Kang, H. Tao, and W. Hua. Part Based Human Tracking in A Multiple Cues Fusion Framework. *ICPR*, 2006.

[24] T. Zhao and R. Nevatia. Tracking Multiple Humans in Crowded Environment. *CVPR*, 2004.