

Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors

Bo Wu Ram Nevatia
University of Southern California
Institute for Robotics and Intelligent System
Los Angeles, CA 90089-0273
{bowu|nevatia}@usc.edu

Abstract

This paper proposes a method for human detection in crowded scene from static images. An individual human is modeled as an assembly of natural body parts. We introduce edgelet features, which are a new type of silhouette oriented features. Part detectors, based on these features, are learned by a boosting method. Responses of part detectors are combined to form a joint likelihood model that includes cases of multiple, possibly inter-occluded humans. The human detection problem is formulated as maximum a posteriori (MAP) estimation. We show results on a commonly used previous dataset as well as new datasets that could not be processed by earlier methods.

1. Introduction

Robust detection of humans in images is important for many applications, such as visual surveillance, smart rooms, and driver assistance systems. Considerable progress has been made in detection and tracking of humans in video sequences, e.g. see [16]. However, such methods rely on segmentation of a *foreground* motion blob assumed to contain one or more of the walking humans. Motion blob segmentation by background modeling is simple and effective when the camera is stationary and changes in illumination are gradual, but for many applications the camera may be in motion and sudden illumination changes occur, particularly in indoor scenes. In such cases, motion blob segmentation is challenging; direct detection of human forms may actually help solve the motion segmentation problem. Also, there are static humans in videos and the problem of finding humans in single images has its own applications.

The problem of human detection in single images has also received attention in the computer vision literature but most systems rely either on a high

resolution face to be visible or the entire human shape to be visible. In this paper we focus on the problem of locating multiple, possibly inter-occluded humans from static images where faces are not necessarily visible or have enough resolution to be recognized easily. We limit the view to be approximately frontal or rear (the range of left-right out-of-plane rotation angle is within $[-30^\circ, +30^\circ]$) and with upright standing or walking pose with camera looking down with a tilt angle not exceeding about 45° . Recent work on the similar problem, such as [1] and [4] consider similar views ([1] also includes a profile view), but do not address inter-human occlusions. We show results on the common datasets (the “MIT dataset [5]”) with these efforts but also on other datasets we have assembled from images available on the Internet and on the CAVIAR dataset [15] that contains occluded humans.

Detection of human body is more complicated than for objects, such as cars and faces, as human body is highly articulated; although we constrain the pose to walking and standing, the possible variations are still very large. Projected images then depend on the view point, illumination and occlusion. Image appearance is also strongly influenced by the clothing people wear which has a wide diversity of colors, textures and styles. We propose to overcome these difficulties, at least partially, by using a part-based representation for humans which helps account for some of the occlusion and changes in perceived 2-D shape. We use an appearance based learning method to detect parts but use of silhouette-based features makes our method less sensitive to changes caused by clothing. We combine the outputs of part detectors to compute likelihood of presence of multiple humans, considering possible occlusions. Our method assumes that humans stand or walk on a ground plane; this enables reasoning in 3-D to account for possible occlusions. Ideally, one would

make inferences of 3-D part shapes for a robust system but this continues to be a difficult problem to solve.

1.1. Related work

Many of the earlier methods for human detection represent the human body as an integral whole, e.g. Papageorgiou et al.'s SVMs detectors [5], Felzenszwalb's shape models [6], and Gavrila et al.'s edge templates [7][8] locate humans by recognizing the full-body pattern. The positive sample set in [5] is known as the *MIT pedestrian dataset* which is available online. The framework for object detection proposed by Viola and Jones [3] has proved very efficient for the face detection problem. The basic idea of this method is to select weak features, Haar wavelets, by AdaBoost [12] to build a cascade structured detector. Viola et al. [2] report that applied to human detection, this approach does not work well using the static Haar features. They augment their system to use simple motion features to achieve much better performance.

Some part-based representations have also been developed. Mohan et al. [4] divide human body into four parts, head-shoulder, legs and left/right arms. They learned SVM detectors using Haar wavelet features. Mikolajczyk et al. [1] divide human body into seven parts, face/head for frontal view, face/head for profile view, head-shoulder for frontal and rear view, head-shoulder for profile view, and legs. For each part, a detector is learned by following the Viola-Jones approach applied to SIFT [13] like orientation features. Their method achieved better results on the MIT dataset than [4]. Neither of these papers, however, addresses the situation of crowded scenes where humans may be occluded by others.

1.2 Outline of our approach

Our approach uses a part-based representation. The advantages of this approach are: 1) it can deal with partial occlusion, e.g. when the legs are occluded, the human can still be detected from the upper-body; 2) final decision is based on multiple evidence so that the false alarms can be reduced; and 3) it's more tolerant to the view point changes and the pose variation of articulated objects.

There are two levels for the part based method: detection of parts and then their combination. For the first level, we use detectors learned from a novel set of silhouette oriented features that we call edgelet features; these consist of small connected chains of edges. These features are more suitable for human detection as they are relatively invariant to clothing

differences, unlike gray level or color features used commonly for face detection. We learn part detectors by a boosting approach which is an enhanced version of Viola and Jones's framework [3] proposed by Chang et al. [10]. The main improvements of this enhanced version include: use of the real-valued AdaBoost [11] instead of the discrete-valued AdaBoost [12] algorithm; histogram weak classifiers replacing the threshold weak classifiers; and upgrading the cascade structured detector to the nesting structured detector.

In the second level, we combine the results of various part detectors. In previous such approaches, Mohan et al. [4] trained their combined detector from the output of the part detectors with SVMs; and Mikolajczyk et al. [1] combined the responses of part detectors by defining a joint likelihood for an individual human. Both of these combined detectors consider humans independent from each other and do not model inter-object occlusion. In our approach, we define a joint image likelihood function for multiple, possibly inter-occluded humans. A missing part is explained as the missing detection of the part detector or occlusion by other objects. We assume that the humans walk on a plane and that the image is captured by a camera looking down. Thus the relative depth of humans can be inferred from their image y -coordinates. The shape of a human is approximated by a 2D ellipse and the visibility of body parts is calculated according to the relative depth order of human objects. We formulate multiple human detection problem as a MAP estimation problem and search the solution space to find the best interpretation of the image observation.

The rest of this paper is organized as follows: Section 2 describes the edgelet features; Section 3 introduces the learning method for part detectors; Section 4 derives our joint likelihood model for multiple humans; Section 5 shows the experiment results; and Section 6 sums up.

2. Edgelet Features

An edgelet is a short segment of line or curve. Denote the positions and normal vectors of the points in an edgelet by $\{\mathbf{u}_i\}_{i=1}^k$ and $\{\mathbf{n}_i^E\}_{i=1}^k$, where k is the length of the edgelet. Given an input image I , denote by $M^I(\mathbf{p})$ and $\mathbf{n}^I(\mathbf{p})$ the edge intensity and normal at position \mathbf{p} of I . The affinity between the edgelet and the image I at position \mathbf{w} is calculated by

$$S(\mathbf{w}) = (1/k) \sum_{i=1}^k M^I(\mathbf{u}_i + \mathbf{w}) \left| \langle \mathbf{n}^I(\mathbf{u}_i + \mathbf{w}), \mathbf{n}_i^E \rangle \right|.$$

The above edgelet affinity function captures both the intensity and the shape information of the edge. In our

experiments, the intensity $M^l(\mathbf{p})$ and normal vector $\mathbf{n}^l(\mathbf{p})$ are calculated by 3×3 Sobel kernel convolutions. Since we want to use the edgelet features only as weak features for a boosting algorithm, we simplify them for computational efficiency. First, we quantize the orientation of the normal vector into six discrete values, see Figure 1. The range $[0^\circ, 180^\circ)$ is divided into six bins evenly, which correspond to the integers from 0 to 5 respectively. An angle θ within range $[180^\circ, 360^\circ)$ has the same quantized value as its symmetry $360^\circ - \theta$. Second, the dot product between two normal vectors is approximated by the following function:

$$l[x] = \begin{cases} 1 & x = 0 \\ 4/5 & x = \pm 1, \pm 5 \\ 1/2 & x = \pm 2, \pm 4 \\ 0 & x = \pm 3 \end{cases}$$

where the input x is the difference between two quantized orientations. Denote by $\{V_i^E\}_{i=1}^k$ and $V^l(\mathbf{p})$ the quantized edge orientations of the edgelet and the input image I respectively. The simplified affinity function is

$$S(\mathbf{w}) = (1/k) \sum_{i=1}^k M^l(\mathbf{u}_i + \mathbf{w}) l[V^l(\mathbf{u}_i + \mathbf{w}) - V_i^E]$$

Thus the computation of edgelet feature only includes short integer operations.

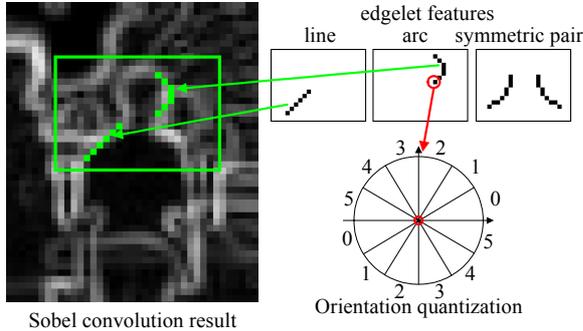


Figure 1. Edgelet features.

In our experiments, the possible length of one single edgelet is from 4 pixels to 12 pixels. The edgelet features we used consist of single edgelets, including lines, $1/8$ circles, $1/4$ circles, and $1/2$ circles, and their symmetric pairs. A symmetric pair is the union of a single edgelet and its mirror. Figure 1 illustrates the definition of our edgelet features. When the size of the reference window is 24×58 , the overall number of possible edgelet features is 857,604.

3. Learning Part Detectors

Human body parts used in this work are head-shoulder, torso, and legs. Besides the three part

detectors, a full-body detector is also learned. Figure 2 shows the definition of the body parts. We use an enhanced version [10] of the original boosting method of Viola and Jones [3] to learn the part detectors. An edgelet feature can be seen as a function from the image space to the feature space. Denote by f_{edgelet} an edgelet feature and f_{edgelet} has been normalized to $[0, 1]$, divide the range of f_{edgelet} into n sub-ranges:

$$\text{bin}_j = [(j-1)/n, j/n), j=1, \dots, n$$

In our experiments, $n = 16$. This even partition on the feature space corresponds to a partition on the image space. For object detection problem, a sample is represented as $\{\mathbf{x}, y\}$, where \mathbf{x} is the image patch and y is the class label whose value can be $+1$ (object) or -1 (non-object). According to the real-valued version of AdaBoost algorithm [11], the weak classifier h based on f_{edgelet} can be defined as

$$\text{If } f_{\text{edgelet}}(\mathbf{x}) \in \text{bin}_j \text{ then } h(\mathbf{x}) = \frac{1}{2} \ln \left(\frac{\bar{W}_{+1}^j + \varepsilon}{\bar{W}_{-1}^j + \varepsilon} \right)$$

where ε is a smoothing factor [11], and

$$\bar{W}_i^j = P(f_{\text{edgelet}}(\mathbf{x}) \in \text{bin}_j, y = i), i = \pm 1, j = 1, \dots, n.$$

Given the characteristic function

$$B_n^j(u) = \begin{cases} 1 & u \in [j-1/n, j/n) \\ 0 & u \notin [j-1/n, j/n) \end{cases}, j = 1, \dots, n$$

the weak classifier can be formulated as:

$$h(\mathbf{x}) = \frac{1}{2} \sum_{j=1}^n \ln \left(\frac{\bar{W}_{+1}^j + \varepsilon}{\bar{W}_{-1}^j + \varepsilon} \right) B_n^j(f_{\text{edgelet}}(\mathbf{x})).$$

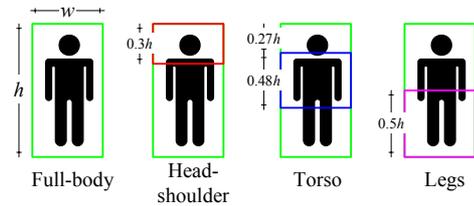


Figure 2. The definition of body parts.

For each edgelet feature, one weak classifier is built. Then the real AdaBoost algorithm [11] is used to learn strong classifiers, called layers, from the weak classifier pool. The strong classifier H is a linear combination of a series of weak classifiers:

$$H(\mathbf{x}) = \sum_{i=1}^T h_i(\mathbf{x}) - b,$$

where T is the number of weak classifiers in H , and b is a threshold. The learning procedure of one layer is referred to as a boosting stage. At the end of each boosting stage, the threshold b is tuned so that H has a high detection rate and new negative samples for the next stage are collected in a bootstrap way. Finally nesting structured detectors [10] are constructed from these layers. One of the main advantages of this

enhanced version [10] over the original method [3] is that the number of features needed to achieve a reasonable performance is reduced greatly.

4. Joint Likelihood for Multiple Humans

To combine the results of part detectors, we compute the likelihood of presence of multiple humans at the hypothesized locations. If inter-object occlusion is present, the assumption of conditional independence between individual human appearances given the state, as in [1], is not valid and a more complex formulation is necessary.

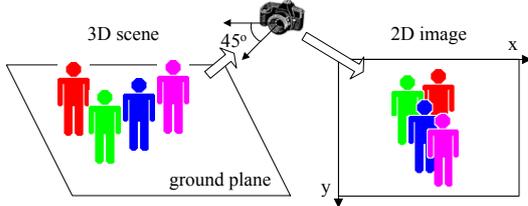


Figure 3. 3D assumption.

We begin by formulating the state and the observation variables. The state of multiple humans is defined as $\mathbf{S} = \{\mathbf{X}_1, \dots, \mathbf{X}_m\}$, where m is the number of humans, which is unknown, and \mathbf{X}_i is the state of the i -th human. $\mathbf{X} = \{\mathbf{p}, s\}$, where \mathbf{p} is the image position and s is the size. To model inter-object occlusion, we assume that the humans walk on a ground plane and the image is captured by a camera looking down to the ground, see Figure 3. This assumption is valid for common surveillance systems and many hand-held photos. This configuration brings two observations: 1) if a human in the image is visible then at least his/her head is visible and 2) the further the human from the camera, the smaller his/her y -coordinate. With the second observation, we can find the relative depth of humans by comparing their y -coordinates and build an occupancy map, see Figure 4.a and b. The shape of an individual human is modeled as an ellipse which is tighter than the box obtained by part detectors. From the occlusion map, the ratio of the visible area to the overall area of the part is calculated. If the ratio is above a threshold (set to 0.7 in our experiments), then the part is classified as visible, otherwise occluded. The state of an individual part is also defined as $\mathbf{P} = \{\mathbf{p}, s\}$. We represent the state of all visible parts as

$$\tilde{\mathbf{S}} = \left\{ \left\{ \mathbf{P}_i^{FB} \right\}_{i=1}^{m^{FB}}, \left\{ \mathbf{P}_i^{HS} \right\}_{i=1}^{m^{HS}}, \left\{ \mathbf{P}_i^T \right\}_{i=1}^{m^T}, \left\{ \mathbf{P}_i^L \right\}_{i=1}^{m^L} \right\},$$

where the superscripts FB , HS , T , and L stand for full-body, head-shoulder, torso and legs respectively, and m^{FB} , m^{HS} , m^T , m^L are the numbers of the visible parts.

$\tilde{\mathbf{S}}$ is a reduced version of the state vector \mathbf{S} by removing all occluded body parts. We can assume the

likelihoods of the visible parts in $\tilde{\mathbf{S}}$ are conditional independent. Denote by

$$\mathbf{Z} = \left\{ \left\{ \mathbf{Z}_i^{FB} \right\}_{i=1}^{n^{FB}}, \left\{ \mathbf{Z}_i^{HS} \right\}_{i=1}^{n^{HS}}, \left\{ \mathbf{Z}_i^T \right\}_{i=1}^{n^T}, \left\{ \mathbf{Z}_i^L \right\}_{i=1}^{n^L} \right\}$$

the responses of the full-body, head-shoulder, torso, and leg detectors respectively, where n^{FB} , n^{HS} , n^T , n^L are the numbers of the responses, and \mathbf{Z}_i is a single response and in the same space as \mathbf{P} . In order to detect body parts at different scales the input image is re-sampled to build a scale pyramid with a scale factor of 1.2, then the image at each scale is scanned by the detector with a step of 2 pixels. With \mathbf{Z} as the observation and $\tilde{\mathbf{S}}$ as the state, we define the following likelihood to interpret the outcome of the part detectors on an input image I :

$$\begin{aligned} p(I | \mathbf{S}) &= p(\mathbf{Z} | \tilde{\mathbf{S}}) \\ &= p(\mathbf{Z}^{FB} | \mathbf{P}^{FB}) p(\mathbf{Z}^{HS} | \mathbf{P}^{HS}) p(\mathbf{Z}^T | \mathbf{P}^T) p(\mathbf{Z}^L | \mathbf{P}^L) \end{aligned}$$

where $\mathbf{Z}^P = \{\mathbf{Z}_i^P\}$ and $\mathbf{P}^P = \{\mathbf{P}_i^P\}$, $P=FB, HS, T, L$. The

detector responses in \mathbf{Z} and the parts in $\tilde{\mathbf{S}}$ can be classified into three classes: successful detections (SD), false alarms (FA), and false negatives (FN), i.e. missing detections, denoted by T_{SD} , T_{FA} , and T_{FN} respectively. The likelihood for one part, take head-shoulder as an example, is calculated by

$$\begin{aligned} p(\mathbf{Z}^{HS} | \mathbf{P}^{HS}) &\propto \prod_{\mathbf{z}_i^{HS} \in T_{SD}^{HS}} p_{SD}^{HS} p(\mathbf{z}_i^{HS} | \mathbf{P}(\mathbf{z}_i^{HS})) \cdot \\ &\quad \prod_{\mathbf{z}_i^{HS} \in T_{FA}^{HS}} p_{FA}^{HS} \cdot \prod_{\mathbf{z}_i^{HS} \in T_{FN}^{HS}} p_{FN}^{HS} \end{aligned}$$

where $\mathbf{P}(\mathbf{z}_i)$ maps a response to its correspond part, p_{SD} is the reward of successful detection, p_{FA} and p_{FN} are the penalties of false alarm and false negative respectively, and $p(\mathbf{z}_i | \mathbf{P}(\mathbf{z}_i))$ is the conditional probability of a detector response given its matched part. Denote by N_{FA} , N_{SD} and N_G the number of false alarms, the number of successful detections, and the ground-truth number of target objects respectively, p_{FA} and p_{SD} are calculated by

$$p_{FA} = \frac{1}{\alpha} e^{-\beta} \frac{N_{FA}}{N_{FA} + N_{SD}}, \quad p_{SD} = \frac{1}{\alpha} e^{\beta} \frac{N_{SD}}{N_{FA} + N_{SD}},$$

where α is a normalization factor so that $p_{FA} + p_{SD} = 1$ and β is a factor to control the relative importance of detection rate vs. false alarms. p_{FN} is calculated by

$$p_{FN} = (N_G - N_{SD}) / N_G.$$

$p(\mathbf{Z} | \mathbf{P}(\mathbf{Z}))$ is a Gaussian distribution. N_{FA} , N_{SD} , N_G and $p(\mathbf{Z} | \mathbf{P}(\mathbf{Z}))$ are all learned from a verification set. Note for different detectors, p_{SD} , p_{FA} , p_{FN} and $p(\mathbf{Z} | \mathbf{P}(\mathbf{Z}))$ may be different.

The match between the responses and parts, i.e. $\mathbf{P}(\mathbf{Z})$, is done in a greedy way. First the distance matrix

\mathbf{D} of all possible response-part pairs is calculated, i.e. $\mathbf{D}(i,j)$ is the distance between the i -th response and the j -th part. Then in each step, the pair, denote by (i^*, j^*) , with the smallest distance is taken and the i^* -th row and the j^* -th column of \mathbf{D} are deleted. This selection is done iteratively until no more valid pair is available. With the match $\mathbf{P}(\mathbf{Z})$, the sets T_{SD} , T_{FA} and T_{FN} are easy to obtain.

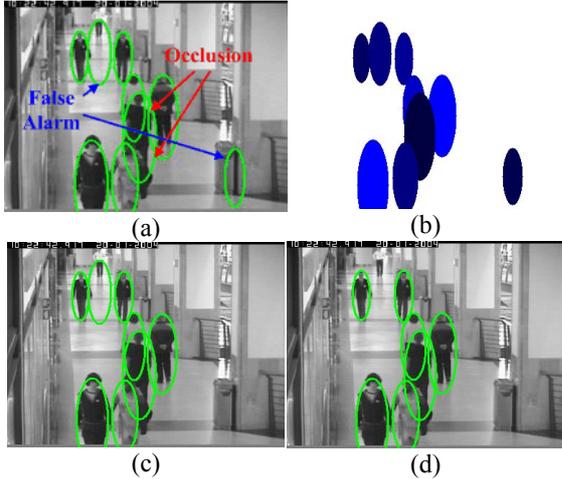


Figure 4. Search for the best interpretation of the image: a) initial state; b) occupancy map of the initial state; c) an intermediate state; and d) final state. This test picture comes from the CAVIAR sequences [15].

1. Scan the image with the part and body detectors.
2. Propose initial state vector \mathbf{S} from the responses of the head-shoulder and the full-body detectors.
3. Sort the humans according to their y -coordinate in a descending order.
4. Match the detector responses to the visible parts.
5. For $i = 1$ to m
 - a) Calculate the image likelihood $p(\mathbf{Z}|\mathbf{S})$ and $p(\mathbf{Z}|\mathbf{S}-\{\mathbf{X}_i\})$.
 - b) If $p(\mathbf{Z}|\mathbf{S}-\{\mathbf{X}_i\}) > p(\mathbf{Z}|\mathbf{S})$ then $\mathbf{S} \leftarrow \mathbf{S}-\{\mathbf{X}_i\}$.
6. Output \mathbf{S} as the final result.

Figure 5. Searching algorithm.

Finally we need a method to propose the candidate state \mathbf{S} and search the solution space to maximize the posterior probability $p(\mathbf{S}|I)$. According to Bayes' rule

$$p(\mathbf{S}|I) \propto p(I|\mathbf{S})p(\mathbf{S}) = p(\mathbf{Z}|\tilde{\mathbf{S}})p(\mathbf{S})$$

Assuming a uniform distribution of the prior $p(\mathbf{S})$, the above MAP estimation is equal to maximizing the joint likelihood $p(\mathbf{Z}|\tilde{\mathbf{S}})$. In our method the initial candidate state \mathbf{S} is proposed based on the responses of the head-shoulder and full-body detectors. Then each human is verified with the above likelihood model in

their depth order. The details of this procedure are listed in Figure 5. Figure 4 gives an example of the combined algorithm. At the initial state, there are two false alarms which get not enough evidence and are discarded later. The legs of the human in the middle is occluded by another human and missed by the leg detector, but this missing part can be explained by inter-object occlusion, so that no penalty is put on it.

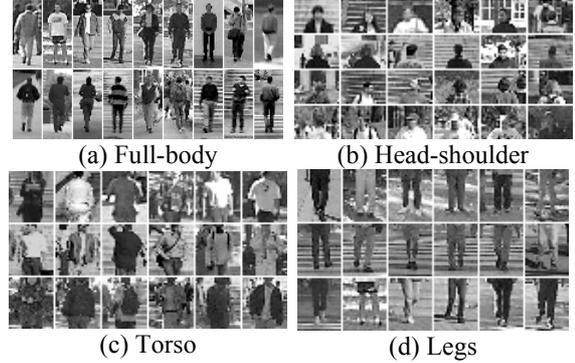


Figure 6. Positive training samples.

5. Experimental Results

We describe a number of tests conducted to evaluate our method. We first describe our results on examples of unoccluded humans, and then present the results on occluded examples taken from the CAVIAR video dataset [15]. We also provide some comparison with results presented in earlier papers using the MIT dataset [5]; however, this comparison is not direct as we do not have knowledge of the exact samples used for training and testing in their experiments, nor access to their code.

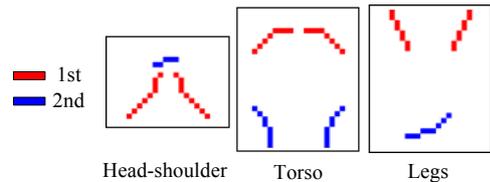


Figure 7. The first two edgelet features learned.

5.1 Evaluation for unoccluded examples

We evaluated our system for detection of humans, as opposed to mere classification in [1] and [4] where segmentation has already been performed. As there is no satisfactory benchmark data set for this task yet, we created one of our own. We included examples from the MIT dataset but enhanced it by images acquired from the Internet to meet our viewing requirements but without any other bias.

Our training set contains 1,742 human samples, 925 of which are from the MIT pedestrian set [5] and the rest are collected from the Internet. The samples are aligned according to the positions of head and feet. The size of full-body samples is 24×58 pixels and the sizes of the other parts can be derived from Figure 2. Figure 6 shows some examples from our training set. Our negative image set contains 7,000 negative images without humans.

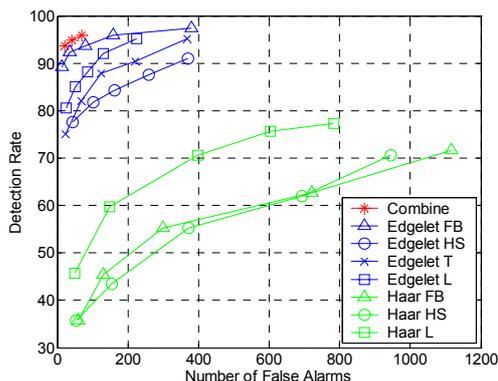


Figure 8. ROC curves of evaluation as detector on our Internet test set (205 images with 313 humans).

Table 1. Complexity of the part detectors.

| | FB | HS | T | L |
|-----------|-----|-------|-----|-----|
| Layer # | 11 | 22 | 18 | 18 |
| Feature # | 277 | 1,157 | 767 | 753 |

We collected a test set from the Internet containing 205 real-life photos and 313 different humans with frontal/rear view. This test set does not have heavy inter-object occlusion and is not included in the training set. In this experiment, the part and full-body detectors are learned with 1,742 positive and 6,000 negative samples for each boosting stage. (The negative samples are 24×58 patches cut from the negative images.) Figure 7 shows the first two features learned. They are quite meaningful. Table 1 lists the complexity of our part and full-body detectors. The head-shoulder detector needs more features than the other detectors. And the full-body detector needs many fewer features than any individual part detector. We evaluated our edgelet detectors and the Haar detectors from OpenCV4.0b [14] on our Internet test set. Figure 8 shows the ROC curves of the part, full-body and combined detectors and Figure 11 shows some examples of successful detections and interesting false alarms, where locally the image looks like the target parts. Figure 12.a show some image results of the combined detector. The sizes of the humans considered vary from 24×58 to 128×309 . The running speed of the combined detector is about 1 FPS with 384×288 pixel images on a 2.8GHz CPU.

It can be seen that, in examples without occlusion, the detection rate of the combined detector is not much higher than that obtained by the full body detector, but this rate is achieved with fewer false alarms. From Figure 12.a we can see that even though the individual part detectors may have false alarms, they do not coincide with the geometric structure of human body and are removed by the combined detector.

Some observations on the part detectors are: 1) the edgelet features are more powerful for human detection than Haar features; 2) full-body detector is more discriminative than individual parts; and 3) head-shoulder is less discriminating than the other parts. The last observation is consistent with that reported in [4], but inconsistent with that in [1]. Mohan et al. [4] gave an explanation for the superiority of leg detector: the background of legs is usually road or grassland and so on, which is relatively clutter-free compared to the background of head-shoulder. However, the leg detector of Mikolajczyk et al. [1] is slightly inferior to their head-shoulder detector. This may be due to the fact that their leg detector covers all frontal, rear, and profile views.

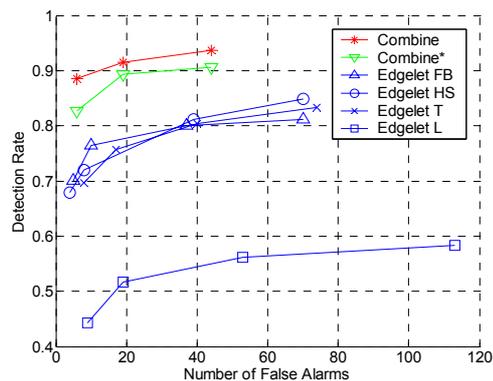


Figure 9. ROC curves of evaluation on our CAVIAR test set (54 images with 271 humans). Combine* is the detection rate on the 75 partially occluded humans.

Table 2. Detection rates on different degrees of occlusion (with 19 false alarms).

| Occlusion degree (%) | >70 | 70~50 | 50~25 |
|----------------------|-----|-------|-------|
| Human # | 10 | 31 | 34 |
| Detection rate (%) | 80 | 90.3 | 91.2 |

5.2 Results with occluded humans

To evaluate our method with occlusion, we used 54 frames with 271 humans from the CAVIAR sequences [15]. In this set, 75 humans are partially occluded by others, and 18 humans are partially out of the scene. The CAVIAR data is not included in our training set. Figure 9 shows the ROC curves of our part, full-body

and the combined detectors on this set. The curve labeled “Combine*” in Figure 9 shows the overall detection rate on the 75 occluded humans and Table 2 lists the detection rates on different degrees of occlusion. Figure 12.b shows some image results on the CAVIAR test set. It can be seen that for the crowded scene: 1) the performance of full-body and leg detectors decreases greatly, as lower-body is more likely to be occluded; 2) the combined detector outperforms the individual detectors; 3) the detection rate on partially occluded humans is only slightly lower than the overall detection rate and declines slowly with the degree of occlusion. In the first example of Figure 12.b, the occluded person is detected just from the head-shoulder detector output. Note that even though the head-shoulder detector by itself may create several false alarms, this will result in a false alarm for the combined result only if the head-shoulder is found in the right relation to another human.

5.3 Comparison with other systems

It is difficult to compare our method with previous ones due to variability in datasets and lack of access to the earlier method’s code. Nonetheless, we show a comparison with methods in [1] and [4] as they both use the MIT dataset that was available to us. Note that this dataset contains un-occluded examples only. Also, these methods report classification rather than detection results; for a proper comparison, we also use classification results in this section. Mohan et al. [4] used 856/866 positive and 9,315/9,260 negative samples to train their head-shoulder/leg detectors. The detection and false alarm rates were evaluated on a test set with 123 positive samples and 50 negative images. Mikolajczyk et al. [1] trained their head-shoulder/leg detector with 250/300 positive and 4,000 negative samples for each boosting stage, and evaluation was done with 400 positive samples and 200 negative images. The positive samples of [1] and [4] are all from the MIT set. As mentioned before a direct comparison with [1] and [4] is difficult, so we do the comparison in a less direct way. We trained our head-shoulder and leg detectors with 6/7 of the MIT set and 3,000 negative samples for each boosting stage, and evaluated with the remaining 1/7 of the MIT set and 200 negative images. Our experimental setup is comparable to that of Mohan et al. [4]. When training with only 300 positive samples, like in [1], our method suffered from over-fitting. Figure 10 shows the ROC curves. Although our sample sets are not exactly the same with those in [1] and [4], the positive samples are all from the MIT set. It can be seen our method achieves higher accuracy than the previous methods.

6. Conclusions and Future Work

We described a static human detection method based on part detectors using novel edgelet features to describe silhouette patterns. We defined a joint likelihood for multiple humans based on the responses of part detectors and explicit modeling of inter-object occlusion. Our combined detection method results in better performance for individual human detection and furthermore can deal with crowded scenes. Our method can be generalized in several ways. One is to use of other viewpoints and poses. Other is to make use of multiple views available in a video sequence. This will allow us to combine detection and tracking and make use of color and texture properties.

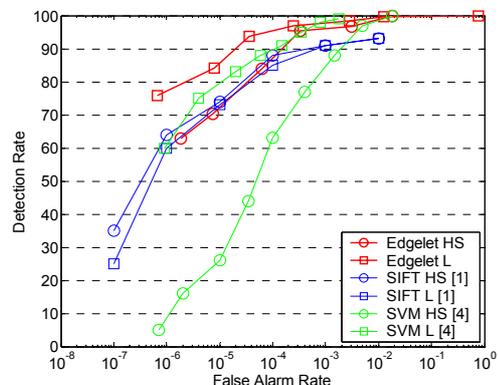


Figure 10. ROC curves of evaluation as classifier on MIT set. The results of [1] and [4] are copied from the original papers.

Acknowledgements: This research was partially funded by the Advanced Research and Development Activity of the U.S. Government under contract MDA-904-03-C-1786.

7. References

- [1]. K. Mikolajczyk, C. Schmid, and A. Zisserman, “Human Detection Based on a Probabilistic Assembly of Robust Part Detector”, ECCV 2004. Vol I: 69-82
- [2]. P. Viola, M. Jones, and D. Snow, “Detecting Pedestrians Using Pattern of Motion and Appearance”, ICCV 2003. pp. 734-741
- [3]. P. Viola and M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features”, CVPR 2001. Vol I: 511-518
- [4]. A. Mohan, C. Papageorgiou, and T. Poggio, “Example-Based Object Detection in images by Components”, IEEE Trans on PAMI, 23(4): 394-361, 2001.

- [5]. C. Papageorgiou, T. Evgeniou, and T. Poggio, "A Trainable Pedestrian Detection System", In: Proc. of Intelligent Vehicles, 1998. pp. 241-246
- [6]. P. Felzenszwalb, "Learning Models for Object Recognition", CVPR 2001. Vol I: 1056-1062
- [7]. D. M. Gavrilu and V. Philomin. "Real-Time Object Detection for 'Smart' Vehicles", ICCV 1999. pp. 87-93
- [8]. D. M. Gavrilu, "Pedestrian Detection from a Moving Vehicle", ECCV 2000. Vol II: 37-49
- [9]. D. M. Gavrilu, "The Visual Analysis of Human Movement: A Survey", CVIU, 73(1): 82-98, 1999.
- [10]. C. Huang, H. Ai, B. Wu, and S. Lao, "Boosting Nested Cascade Detector for Multi-View Face Detection", ICPR 2004. Vol II: 415-418
- [11]. R. E. Schapire and Y. Singer, "Improved Boosting Algorithms Using Confidence-rated Predictions", Machine Learning, 37: 297-336, 1999.
- [12]. Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm", The 13-th Conf. on Machine Learning, 1996, 148-156.
- [13]. D. G. Lowe, "Object recognition from local scale-invariant features", ICCV 1999. pp. 1150-1157
- [14]. H. Kruppa, M. Castrillon-Santana and B. Schiele, "Fast and Robust Face Finding via Local Context", Joint IEEE Int'l Workshop on VS-PETS, 2003.
- [15]. <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>
- [16]. T. Zhao and R. Nevatia, "Bayesian Human Segmentation in Crowded Situations", CVPR 2003. Vol II: 459-466
- [17]. B. Leibe, E. Seemann, and B. Schiele, "Pedestrian Detection in Crowded Scenes", CVPR 2005. Vol I: 878-885

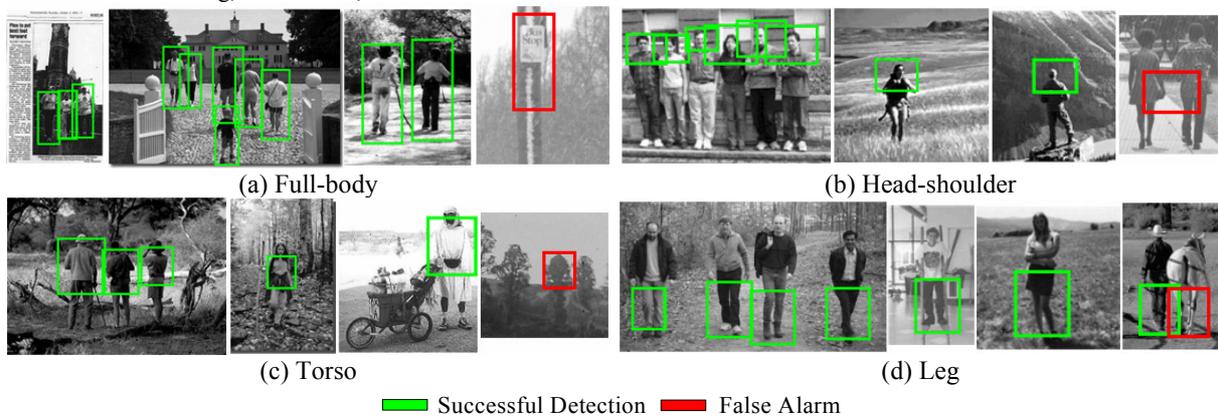


Figure 11. Some detection results of edgelet part detectors on images from our Internet test set.

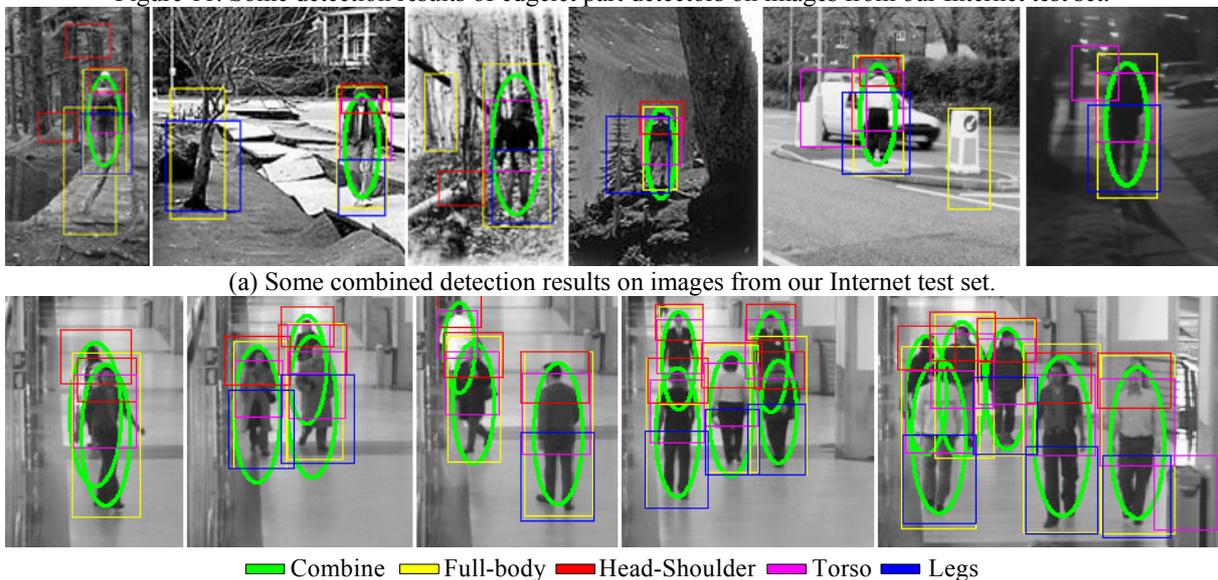


Figure 12. Combined detection results.