International Journal of Computer Vision 75(2), 247–266, 2007

DOI: 10.1007/s11263-006-0027-7

Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors

BO WU AND RAM NEVATIA
University of Southern California, Institute for Robotics and Intelligent Systems, Los Angeles, CA 90089-0273
bowu@usc.edu
nevatia@usc.edu

Received August 18, 2006; Accepted December 13, 2006

First online version published in January, 2007

Abstract. Detection and tracking of humans in video streams is important for many applications. We present an approach to automatically detect and track multiple, possibly partially occluded humans in a walking or standing pose from a single camera, which may be stationary or moving. A human body is represented as an assembly of body parts. Part detectors are learned by boosting a number of weak classifiers which are based on *edgelet* features. Responses of part detectors are combined to form a joint likelihood model that includes an analysis of possible occlusions. The combined detection responses and the part detection responses provide the observations used for tracking. Trajectory initialization and termination are both automatic and rely on the confidences computed from the detection responses. An object is tracked by data association and meanshift methods. Our system can track humans with both inter-object and scene occlusions with static or non-static backgrounds. Evaluation results on a number of images and videos and comparisons with some previous methods are given.

Keywords: human detection, human tracking, AdaBoost

1. Introduction

Detection and tracking of humans is important for many applications, such as visual surveillance, human computer interaction, and driving assistance systems. For this task, we need to detect the objects of interest first (i.e., find the image regions corresponding to the objects) and then track them across different frames while maintaining the correct identities. The two principle sources of difficulty in performing this task are: (a) change in appearance of the objects with viewpoint, illumination and clothing and (b) partial occlusion of objects of interest by other objects (occlusion relations also change in a dynamic scene). There are additional difficulties in tracking humans after initial detection. The image appearance of humans changes not only with the changing viewpoint but even more strongly with the visible parts of the body and clothing. Also, it is hard to maintain the identities of objects during tracking when humans are close to each other.

Most of the previous efforts in human detection in videos have relied on detection by changes caused in subsequent image frames due to human motion. A model of the background is learned and pixels departing from this model are considered to be due to object motion; nearby pixels are then grouped into motion blobs. This approach is quite effective for detecting isolated moving objects when the camera is stationary, illumination is constant or varies slowly, and humans are the only moving objects; an early example is given in Wren et al. (1997). For a moving camera, there is apparent background motion which can be compensated for, in some cases, but errors in registration are likely in presence of parallax. In any case, for more complex situations where multiple humans and other objects move in a scene, possibly occluding each other to some extent, the motion blobs do not necessarily correspond to single humans; multiple moving objects may merge into a single blob with only some parts visible for the occluded objects, and a single human may appear split into multiple blobs. Figure 1 shows two

Electronic Supplementary Material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s11263-006-0027-7

Figure 1. Sample frames: (a) is from the CAVIAR set (http:// homepages.inf.ed.ac.uk/rbf/CAVIAR/), and (b) is from data we have collected.

examples where such difficulties can be expected to be present.

A number of systems have been developed in recent years, e.g. (Isard and MacCormick, 2001; Zhao and Nevatia, 2004a; Smith et al., 2005), to segment multiple humans from motion blobs. While these systems demonstrate impressive results, they typically assume that all of a motion region belongs to one or more person but real motion blobs may contain multiple categories of objects, shadows, reflection regions and blobs created because of illumination changes or camera motion parallax.

We describe a method to automatically track multiple, possibly partially occluded humans in a walking or standing pose. Our system does not rely on motion for detection, instead it detects humans based on their shape properties alone. We use a part based representation. We learn detectors for each part and combine the part detection results for more robust human detection. For occluded humans, we can not expect to find all the parts; our system explicitly reasons about occlusion of parts by considering joint detection of all objects. The part detectors are view-based, hence our system has some limitations on the viewpoint. The viewpoint is assumed to be such that the camera has a tilt angle not exceeding 45°; the humans may be seen in any orientation but in a relatively upright pose. Also, shape analysis requires adequate resolution; we require that the human width in image is 24 pixels or more.

Tracking in our system is based on detection of humans and their parts, as a holistic body representation can not adapt to the changing inter-human occlusion relations. Figure 2 gives an example which shows the necessity of part based tracking. We use a multi-level approach. Humans are tracked based on complete detection where possible. In presence of occlusion, only some parts can

Figure 2. Example of changing occlusion relations.

be seen; in such cases, our system tracks the visible parts and combines the results of part associations for human tracking. When no reliable detection is available, a meanshift tracker is applied. For complete occlusion, by other humans or scene objects, the tracks are inferred by observations before and after such occlusion. Our method does not require manual initialization (as does a meanshift tracker for example); instead, trajectories are initiated and terminated automatically based on detection outputs.

Our method has been applied to a number of complex static images and video sequences. Considerable and persistent occlusion is present and the scene background can be highly cluttered. We show results on stationary and moving camera examples. Environment can be indoors or outdoors with possibly changing illumination. Quantitative evaluation results on both standard data sets and data set we have collected are reported. The results show that our approach outperforms the previous methods for both detection and tracking.

The main contributions of this work include: (1) a Boosting based method to learn body part detectors based on a novel type of shape features, edgelet features; (2) a Baysian method to combine body part detection responses to detect multiple partially occluded humans; and (3) a fullly automatic hypotheses tracking framework to track multiple humans through occlusions. Parts of our system have been previously described in Wu and Nevatia (2006a,b); this paper presents several enhancements, and provides a unified and detailed presentation and additional results.

The rest of this paper is organized as follows: Section 2 introduces some related works; Section 3 gives an outline of our approach; Section 4 describes our body part detection system; Section 5 gives the algorithm that combines the body part detectors; Section 6 presents the part detection based human tracking algorithm; Section 7 provides the experimental results; and conclusions and discussions are in the last section.

2. Related Work

The literature on human detection in static images and on human tracking in videos is abundant. Many methods for static human detection represent a human as an integral whole, e.g. Papageorgiou et al.'s SVMs detectors (Papageorgiou et al., 1998) (the positive sample set in Papageorgiou et al. (1998) is known as the MIT pedestrian sample set which is available online[1]), Felzenszwalb's shape models (Felzenszwalb, 2001), Wu et al.'s Markov Random Field based representation (Wu et al., 2005), and Gavrila et al.'s edge templates (Gavrila and Philomin, 1999; Gavrila, 2000). The object detection framework proposed by Viola and Jones (2001) has

proved very efficient for the face detection problem. The basic idea of this method is to select weak classifiers which are based on simple features, e.g. Haar wavelets, by AdaBoost (Freund and Schapire, 1996) to build a cascade structured detector. Viola et al. (2003) report that applied to human detection, this approach does not work very well using the static Haar features. They augment their system by using local motion features to achieve much better performance. Overall, holistic representation based methods do not work well with large spatial occlusion, as they need evidence for most parts of the whole body.

Some methods for representation as an assembly of body parts have also been developed. Mohan et al. (2001) divide human body into four parts: head-shoulder, legs, left arm, and right arm. They learn SVM detectors using Haar wavelet features. The results reported in Mohan et al. (2001) show that the part based human model is much better than the holistic model in Papageorgiou et al. (1998) for detection task. Shashua et al. (2004) divide human body into nine regions, for each of which a classifier is learned based on features of orientation histograms. Mikolajczyk et al. (2004) 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 (Lowe, 1999) like orientation features. The methods of Shashua et al. (2004) and Mikolajczyk et al. (2004) both achieved better results than that of Mohan et al. (2001), but there is no direct comparison between (Shashua et al., 2004) and (Mikolajczyk et al., 2004). However these part-based systems do not use the parts for tracking nor consider occlusions. In Zhao and Nevatia (2004a), a part-based representation is used for segmenting motion blobs by considering various articulations and their appearances but parts are not tracked explicitly.

Several types of features have been applied to capture the pattern of humans. Some methods use spatially global features as in Gavrila (2000), Felzenszwalb (2001) and Leibe et al. (2005); others use spatially local features as in Papageorgiou et al. (1998), Mohan et al. (2001), Viola et al. (2003), Mikolajczyk et al. (2004), Wu et al. (2005), Leibe et al. (2005), and Dalal and Triggs (2005). The local feature based methods are less sensitive to occlusions as only some of the features are affected by occlusions. Dalal and Triggs (2005) compared several local features, including SIFT, wavelets, and Histogram of Oriented Gradient (HOG) descriptors for pedestrian detection. Their experiments show that the HOG descriptors outperform the other types of features on this task. However, of these only Leibe et al. (2005) incorporates explicit inter-object occlusion reasoning. The method of Leibe et al. (2005) has two main steps: the first generates hypotheses by evidence from local features, while the second verifies the hypotheses by constraints from the global features. These two steps are applied iteratively to compute a local maximum of the image likelihood. The global verification step greatly improves the performance, but it does not deal with partial occlusion well. They achieved reasonable accuracy, an equal error rate of 71.3%, on their own test set of side view pedestrians.

For tracking of human, some early methods, e.g. (Zhao and Nevatia, 2004b) track motion blobs and assume that each individual blob corresponds to one human. These early methods usually do not consider multiple objects jointly and tend to fail when blobs merge or split. Some of the recent methods (Isard and MacCormick, 2001; Zhao and Nevatia, 2004a; Smith et al., 2005; Peter et al., 2005) try to fit multiple object hypotheses to explain the foreground or motion blobs. These methods deal with occlusions by computing joint image likelihood of multiple objects. Because the joint hypotheses space is usually of high dimension, an efficient optimization algorithm, such as a particle filter (Isard and MacCormick, 2001), MCMC (Zhao and Nevatia, 2004a; Smith et al., 2005) or EM (Peter et al., 2005) is used. All of these methods have shown experiments with a stationary camera only, where the background subtraction provides relatively robust object motion blobs. The foreground blob based methods are not discriminative. They assume all moving pixels are from humans. Although this is true in some environments, it is not in more general situations. Some discriminative methods, e.g. (Davis et al., 2000) build deformable silhouette models for pedestrians and track the models from edge features. The silhouette matching is done frame by frame. These methods are less dependent on the camera motion. However they have no explicit occlusion reasoning. None of the above tracking methods deal with occlusion by scene objects explicitly.

Part tracking has been used to track the *pose* of humans (Sigal et al., 2004; Ramanan et al., 2005; Lee and Nevatia, 2006). However the objectives of pose tracking methods and multiple human tracking methods are different. The methodologies of the two problems are also different. The existing pose tracking methods do not consider multiple humans jointly. Although they can work with temporary or slight partial occlusions, because of the use of part representation and temporal consistency, they do not work well with persistent and significant occlusions as they do not model occlusions explicitly and the part models used are not very discriminative. The automatic initialization and termination strategies in the existing pose tracking methods are not general. In Ramanan et al. (2005) a human track is started only when a side view walking pose human is detected, and no termination strategy is mentioned.

Figure 3. Examples of tracking results.

3. Outline of Our Approach

Our approach uses a part-based representation. The advantages of this approach are: (1) it can deal with partial occlusions, e.g. when the legs are occluded, the human can still be detected and tracked from the upper-body; (2) final decision is based on multiple evidence which reduces false alarms; and (3) it is more tolerant to view point changes and pose variations of articulated objects. Figure 3 shows some tracking examples.

Figure 4 gives a schematic diagram of the system. Human detection is done frame by frame. The detection module consists of two stages: detection of parts and then their combination. The tracking module has three stages: trajectory initialization, growth, and termination.

In the first stage of detection, we use detectors learned from a novel set of silhouette oriented features that we call *edgelet* features. These features are 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 tree structured multi-view part detectors by a boosting approach proposed by Huang et al. (2004, 2005) which is an enhanced version of Viola and Jones' framework (Viola and Jones, 2001).

In the second stage of detection, we combine the results of various part detectors. We define a joint image likelihood function for multiple, possibly inter-occluded humans. We formulate the multiple human detection problem as a MAP estimation problem and search the solution space to find the best interpretation of the image observation. Performance of the combined detector is better than that of any individual part detector in terms of the false alarm rate. However the combined detector does explicit reasoning only for inter-object occlusion, while the part detectors can work in the presence of both inter-object and scene occlusions. The previous such approaches, e.g. (Mohan et al., 2001; Mikolajczyk et al., 2004; Shashua et al., 2004), consider humans independently from each other and do not model inter-object occlusion.

Our tracking method is based on tracking parts of the human body. The detection responses from the part detectors and the combined detector are taken as inputs for the tracker. We track humans by data association, i.e., matching the object hypotheses with the detection responses, whenever corresponding detection responses can be found. We match the hypotheses with the combined detection responses first, as they are more reliable than the responses of the individual parts. If for a hypothesis no combined response with similar appearance and close to the predicted position is found, then we try to associate it with part detection responses. If this fails again, a meanshift tracker (Comaniciu et al., 2001) is used to follow the object. Most of the time objects are tracked successfully by data association; the meanshift tracker gets utilized only occasionally and then for short periods. Since our method is based on part detection, it can work under both scene and inter-object occlusion conditions. Also, as the cues for tracking are strong, we do not utilize statistical sampling techniques as in some of the previous work, e.g. (Isard and MacCormick, 2001; Zhao and Nevatia, 2004a; Smith et al., 2005). A trajectory is initialized when evidence from new observations can not be explained by the current hypotheses, as also in many previous methods (Davis et al., 2000; Isard and MacCormick, 2001; Zhao and Nevatia, 2004a; Smith et al., 2005; Peter et al., 2005). Similarly, a trajectory is terminated when it is lost by the detectors for a certain period.

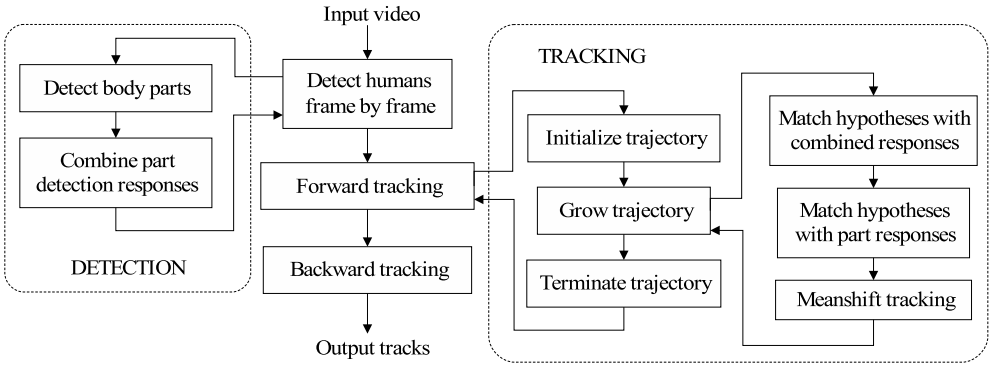

Figure 4. A schematic diagram of our human detection and tracking system.

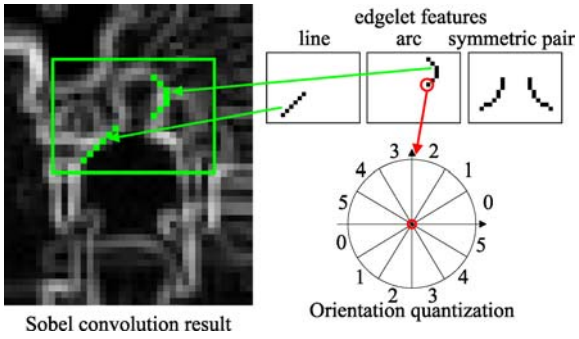

Figure 5. Edgelet features.

4. Detection of Human Body Parts

We detect humans by combining responses from a set of body part detectors that are learned from local shape features.

4.1. Edgelet Features

Based on the observation that silhouettes are one of the most salient patterns of humans, we developed a new class of local shape features that we call *edgelet* features. An edgelet is a short segment of a line or a curve. Denote the positions and normal vectors of the points in an edgelet, E, by $\{\mathbf{u}_i\}_{i=1}^k$ and $\{\mathbf{n}_i^E\}_{i=1}^k$, where k is the length of the edgelet, see Fig. 5 for an illustration. 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 E and the image I at position $\mathbf{w}$ is calculated by

$$f(E; I, \mathbf{w}) = \frac{1}{k}\sum_{i=1}^{k} M^I(\mathbf{u}_i + \mathbf{w}) \left|\left\langle \mathbf{n}^I(\mathbf{u}_i + \mathbf{w}), \mathbf{n}_i^E \right\rangle\right| \tag{1}$$

Note, $\mathbf{u}_i$ in the above equation is in the coordinate frame of the sub-window, and $\mathbf{w}$ is the offset of the sub-window in the image frame. The edgelet affinity function captures both intensity and shape information of the edge; it could be considered a variation of the standard Chamfer matching (Barrow et al., 1977).

In our experiments, the edge intensity $M^I(\mathbf{p})$ and normal vector $\mathbf{n}^I(\mathbf{p})$ are calculated by 3×3 Sobel kernel convolutions applied to gray level images. We do not use color information for detection. Since we use the edgelet features only as weak features in a boosting algorithm, we simplify them for computational efficiency. First, we quantize the orientation of the normal vector into six discrete values, see Fig. 5. 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 $360^\circ - \theta$. Second, the dot product between two normal vectors is approximated by the following function:

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

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

$$\tilde{f}(E; I, \mathbf{w}) = \frac{1}{k}\sum_{i=1}^{k} M^I(\mathbf{u}_i + \mathbf{w}) \cdot l\left[V^I(\mathbf{u}_i + \mathbf{w}) - V_i^E\right] \tag{3}$$

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

In our experiments, the possible length of one single edgelet is from 4 pixels to 12 pixels. The edgelet features we use consist of single edgelets, including lines, $\frac{1}{8}$ circles, $\frac{1}{4}$ circles, and $\frac{1}{2}$ circles, and their symmetric pairs. A symmetric pair is the union of a single edgelet and its mirror. Figure 5 illustrates the definition of our edgelet features. For a sample size of 24×58, the overall number of possible edgelet features is 857,604.

4.2. Boosting Edgelet based Weak Classifiers

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 6 shows the spatial relations of the body parts. We use an enhanced version (Huang et al., 2004) of the original boosting method of Viola and Jones (2001) to learn the part detectors. Suppose the feature value calculated by Eq. (3) has been normalized to [0, 1]. Divide the range into n sub-ranges:

$$bin_j = \left[\frac{j-1}{n}, \frac{j}{n}\right), \quad j = 1 \ldots n \tag{4}$$

In our experiments, $n = 16$. This even partition of the feature space corresponds to a partition of the image space. For object detection, a sample is represented as a tuple $\{\mathbf{x}, y\}$, where $\mathbf{x}$ is the normalized image patch and y is

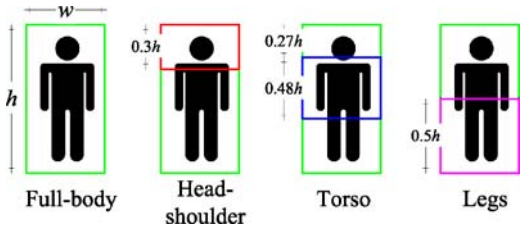

Figure 6. Spatial relations of body parts.

the class label whose value can be $+1$ (object) or -1 (non-object). According to the real-valued version of AdaBoost algorithm (Schapire and Singer, 1999), the weak classifier $h^{(w)}$ based on an edgelet feature E is defined as

$$\text{if } \tilde{f}(E;\mathbf{x},\mathbf{O}) \in bin_j \text{ then } h^{(w)}(\mathbf{x}) = \frac{1}{2}\ln\left(\frac{\bar{W}_{+1}^{j}+\varepsilon}{\bar{W}_{-1}^{j}+\varepsilon}\right) \tag{5}$$

where $\mathbf{O}$ is the origin of the patch $\mathbf{x}$, ε is a smoothing factor (Schapire and Singer, 1999), and

$$\bar{W}_c^j = P\left(\tilde{f}(E;\mathbf{x},\mathbf{O}) \in bin_j, y = c\right), \quad c = \pm 1,\ j = 1\ldots n \tag{6}$$

Given the characteristic function

$$B_n^j(u) = \begin{cases} 1, & u \in \left[\frac{j-1}{n}, \frac{j}{n}\right) \\ 0, & \text{otherwise} \end{cases}, \quad j = 1\ldots n \tag{7}$$

the weak classifier based on the edgelet feature E can be formulated as:

$$h^{(w)}(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{n}\ln\left(\frac{\bar{W}_{+1}^{j}+\varepsilon}{\bar{W}_{-1}^{j}+\varepsilon}\right)B_n^j\left(\tilde{f}(E;\mathbf{x},\mathbf{O})\right) \tag{8}$$

For each edgelet feature, one weak classifier is built. Then the real AdaBoost algorithm (Schapire and Singer, 1999) is used to learn strong classifiers, called layers, from the weak classifier pool. The strong classifier $h^{(s)}$ is a linear combination of a series of weak classifiers selected:

$$h^{(s)}(\mathbf{x}) = \sum\nolimits_{i=1}^{T} h_i^{(w)}(\mathbf{x}) - b \tag{9}$$

where T is the number of weak classifiers in $h^{(s)}$, 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 set so that $h^{(s)}$ has a high detection rate (99.8% in our experiments) and reject as many negative samples as possible. The accepted positive samples are used as the positive set for the training of the next boosting stage; the false alarms obtained by scanning the negative images with the current detector are used as the negative set for the next boosting stage. Finally, nested structured detectors (Huang et al., 2004) are constructed from these layers. Training is stopped when the false alarm rate on the training set reaches 10^{-6}. A nested structure differs from a cascade structure (Viola and Jones, 2001); in a nested structure, each layer is used as the first weak classifier of its succeeding layer so that the information of classification is inherited efficiently. Figure 7 illustrates a nested structure. The main advantage of the nested structure is that the number of features needed to achieve a level of performance is reduced greatly, compared to that needed for a cascade detector.

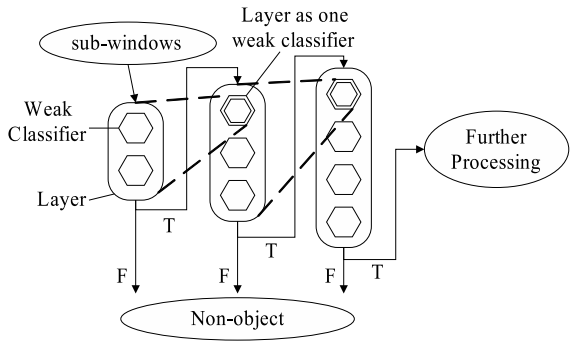

Figure 7. Nested structure.

4.3. Multi-View Part Detectors

To cover all left-right out-of-plane rotation angles, we divide the human samples into three categories, left profile, frontal/rear, and right profile, according to their view points. For each part, a tree structured detector is trained. Figure 8 illustrates the structure of the multi-view detector. The root node of the tree is learned by the vector boosting algorithm proposed in Huang et al. (2005). The main advantage of this algorithm is that the features selected are shared among different view point categories of the same object type. This is much more efficient than learning detectors for individual view points separately. We make one detector cover a range of camera tilt angle, about $[0^\circ, 45^\circ]$ which is common for most surveillance systems, by including samples captured with different tilt angles in our training set. If we want to cover a larger range of tilt angle, some view point categorization along the tilt angle would be necessary.

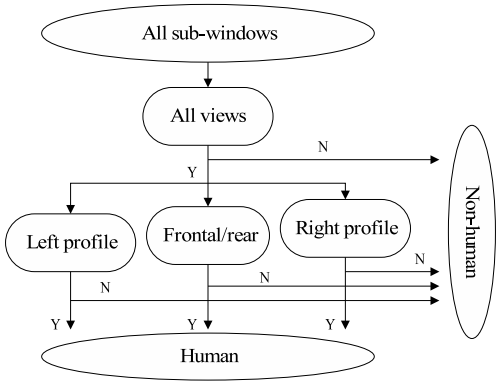

Figure 8. Tree structured multi-view part detector.

Figure 9. Part detection responses (yellow for full-body; red for head-shoulder; purple for torso; blue for legs).

During detection, an image patch is first sent to the root node whose output is a three-channel vector corresponding to the three view categories. If all the channels are negative then the patch is classified as non-human directly; otherwise, the patch is sent to the leaf nodes corresponding to the positive channels for further processing. If any of the three leaf nodes gives a positive output, the patch is classified as a human; otherwise it is discarded. There could be more than one positive channel for one input patch. 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. The outputs of the part detectors are called *part responses*. Figure 9 shows an example of part detection result.

We collect a large set of human samples, from which nested structured detectors for frontal/rear view humans and tree structured detectors for multi-view humans are learned. Figure 10 shows the first two learned features for head-shoulder, torso, and legs of frontal/rear view point. They are quite meaningful. Table 1 lists the complexities, i.e., the number of features used, of our part and full-body detectors of frontal/rear view and multi-view. 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. More details of the experimental setup and the detection performance are given later in Section 7.1.

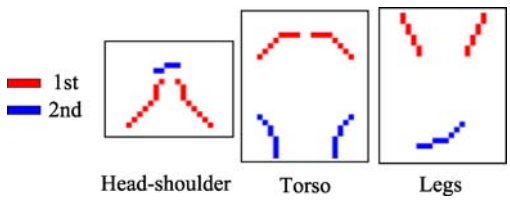

Figure 10. The first two edgelet features learned for each part.

Table 1. Numbers of features used in the detectors. (The nested structured detectors are for frontal/rear view; the tree structured detectors are for multi-view; FB, HS, T, and L stand for full-body, headshoulder, torso, and legs respectively.)

	FB	*HS*	*T*	*L*
Nested detector	227	1,157	767	753
Tree detector	1,059	3,047	2,546	2,256

5. Bayesian Combination of Part Detectors

To combine the results of the part detectors, we compute the likelihood of the 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 Mikolajczyk et al. (2004), is not valid and a more complex formulation is necessary.

We begin by formulating the state and the observation variables. To model inter-object occlusion, besides the assumption that humans are on a plane, we also assume that the camera looks down to the plane, see Fig. 11. This assumption is valid for common surveillance systems. 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 farther the human is from the camera, the smaller is the y-coordinate of his/her feet's image position. With the second observation, we can find the relative depth of humans by comparing their y-coordinates and build an occupancy map, which defines which pixel comes from which human, see Fig. 12(b). The overall image shape of an individual human is modeled as an ellipse which is tighter than the box obtained by part detectors. From the occupancy map, the ratio of the visible area to the overall area of the part is calculated as a visibility score v. If v is larger than a threshold, θ_v (set to 0.7 in our experiments), then the part is classified as visible, otherwise occluded.

A part hypothesis is represented as a 4-tuple $\mathbf{sp} = \{l, \mathbf{p}, s, v\}$, where l is a label indicating the part type, $\mathbf{p}$ is the image position, s is the size, and v is the visibility score. A human hypothesis in one image frame, $H^{(f)}$, consists of four parts, $H^{(f)} = \{\mathbf{sp}_i | l_i = FB, HS, T, L\}$, where FB, HS, T, and L stand for full-body, head-shoulder, torso, and legs respectively. The set of all human

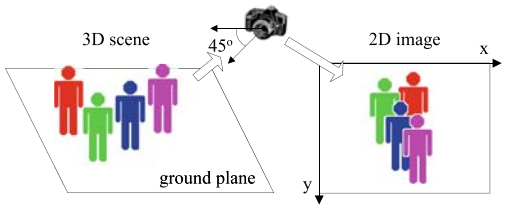

Figure 11. 3D assumption.

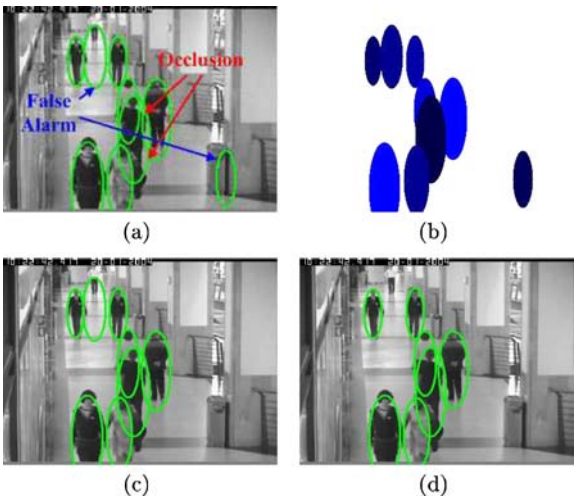

Figure 12. Search for the best interpretation of the image: (a) initial state; (b) occupancy map of the initial state; (c) an intermediate state; and (d) final state.

hypotheses in one frame is $S = \{H_i^{(f)}\}_{i=1}^m$, where m is the number of humans, which is unknown. We represent the set of all visible part hypotheses as

$$\tilde{S} = \{\mathbf{sp}_i \in S | v_i > \theta_v\} \tag{10}$$

$\tilde{S}$ is a subset of S by removing all occluded part hypotheses. We assume that the likelihoods of the visible part hypotheses in $\tilde{S}$ are conditional independent. Let

$$RP = \{\mathbf{rp}_i\}_{i=1}^n \tag{11}$$

be the set of all part detection responses, where n is the overall number of the responses, and $\mathbf{rp}_i$ is a single response, which is in the same space as $\mathbf{sp}_i$. With RP as the observation and $\tilde{S}$ as the state, we define the following likelihood to interpret the outcome of the part detectors for an image I:

$$P(I|S) = P(RP|\tilde{S}) = \prod_{p \in PT} P\left(RP^{(p)}|\tilde{S}^{(p)}\right) \tag{12}$$

where $PT = \{FB, HS, T, L\}$, $RP^{(p)} = \{\boldsymbol{rp}_i \in RP | l_i = p\}$, and $\tilde{S}^{(p)} = \{\mathbf{sp}_i \in \tilde{S} | l_i = p\}$.

To match the responses and hypotheses, a "Hungarian" algorithm (Kuhn, 1955) could be used for an optimal solution, but it is complex. As the response-hypothesis ambiguity is limited in our examples, we chose to implement a greedy algorithm instead. First the distance matrix $\mathbf{B}$ of all possible response-part pairs is calculated, i.e. $\mathbf{B}(i, j)$ is the Euclidean distance between the i-th response and the j-th part hypothesis. Then in each step, the pair, denoted by $(i^\star, j^\star)$, with the smallest distance is taken and the $i^\star$-th row and the $j^\star$-th column of $\mathbf{B}$ are deleted. This selection is done iteratively until no more valid pair is available.

For a match, the responses in RP and the hypotheses in $\tilde{S}$ are classified into three categories: successful detections (SD, responses that have matched hypotheses), false alarms (FA, responses that do not have matched hypotheses), and false negative (FN, hypotheses that do not have matched responses), i.e. missing detections, denoted by T_{SD}, T_{FA}, and T_{FN} respectively. The likelihood for one part type is calculated by

$$P\left(RP^{(p)}|\tilde{S}^{(p)}\right) \propto \prod_{\mathbf{rp}_i \in T_{SD}^{(p)}} P_{SD}^{(p)} P(\mathbf{rp}_i|\bar{\mathbf{sp}}_i) \cdot \prod_{\mathbf{rp}_i \in T_{FA}^{(p)}} P_{FA}^{(p)} \cdot \prod_{\mathbf{rp}_i \in T_{FN}^{(p)}} P_{FN}^{(p)} \tag{13}$$

where $\bar{\mathbf{sp}}_i$ is the corresponding hypothesis of the response $\mathbf{rp}_i$, P_{SD} is the reward of a successful detection, P_{FA} and P_{FN} are the penalties of a false alarm and a false negative respectively, and $P(\mathbf{rp}_i|\bar{\mathbf{sp}}_i) = P(\mathbf{p}_{\mathbf{rp}}|\mathbf{p}_{\bar{\mathbf{sp}}})P(s_{\mathbf{rp}}|s_{\bar{\mathbf{sp}}})$ is the conditional probability of a detection response given its matched part hypothesis. $P(\mathbf{p}_{\mathbf{rp}}|\mathbf{p}_{\bar{\mathbf{sp}}})$ and $P(s_{\mathbf{rp}}|s_{\bar{\mathbf{sp}}})$ are Gaussian distribution. Denote by N_{FA}, N_{SD} and N_G the number of false alarms, the number of successful detections, and the number of ground-truth objects respectively, P_{FA}, P_{SD} are calculated by

$$P_{FA} = \frac{1}{\alpha} e^{-\beta} \frac{N_{FA}}{N_{FA} + N_{SD}}, \; P_{SD} = \frac{1}{\alpha} e^{\beta} \frac{N_{SD}}{N_{FA} + N_{SD}}, \tag{14}$$

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 (set to 0.5 in our experiments). P_{FN} is calculated by

$$P_{FN} = \frac{N_G - N_{SD}}{N_G} \tag{15}$$

N_{FA}, N_{SD}, N_G, $P(\mathbf{p}_{\mathbf{rp}}|\mathbf{p}_{\bar{\mathbf{sp}}})$ and $P(s_{\mathbf{rp}}|s_{\bar{\mathbf{sp}}})$ are all learned from a verification set. For different detectors, P_{SD}, P_{FA}, P_{FN} and $P(\mathbf{rp}|\bar{\mathbf{sp}})$ may be different.

Finally we need a method to propose the hypotheses to form the candidate state S and search the solution space to maximize the posterior probability $P(S|I)$. According to Bayes' rule

$$P(S|I) \propto P(I|S)P(S) = P(RP|\tilde{S})P(S) \tag{16}$$

Assuming a uniform distribution of the prior $P(S)$, the above MAP estimation is equal to maximizing the joint likelihood $P(RP|\tilde{S})$. In our method, the initial set of hypotheses S is proposed from the responses of the head-shoulder and full-body detectors. Each full-body or head-shoulder response generates one human hypothesis. Then the hypotheses are verified with the above likelihood model in their depth order. The steps of this procedure are listed in Fig. 13. Figure 12 gives an example of the results of the combination algorithm. At the initial state,

1. Scan the image with the part and body detectors
2. Propose the initial state vector S from the responses of the head-shoulder and the full-body detectors
3. Sort the humans according to their y-coordinates in a descending order
4. for i=1 to m do
   (a) Match the detector responses to the visible parts
   (b) Calculate the image likelihood $P(RP|S)$ and $P(RP|S - H_i^{(f)})$
   (c) if $P(RP|S - H_i^{(f)}) > P(RP|S)$, then $S \leftarrow S - H_i^{(f)}$
5. Output S as the result

Figure 13. Searching algorithm for combining part detection responses.

there are two false alarms which do not get enough evidence and are discarded later. The legs of the human in the middle are occluded by another human and missed by the legs detector, but this missing part can be explained by inter-object occlusion, so no penalty is put on it. In our combination algorithm, the detectors of torso and legs are not used to propose human hypotheses. This is because the detectors used for initialization have to scan the whole image while the detectors for verification only need to scan the neighborhood of the proposed hypotheses. So if we use all the four part detectors, the system will be at least two times slower. Also we found that the union of the full-body and head-shoulder detection responses already gives very high detection rate and that most of the time, the part that is occluded is the lower body. We call the above Bayesian combination algorithm a *combined detector*, whose outputs are *combined responses*.

The outputs of the detection system have three levels. The first level is a set of the *original responses* of the detectors. In this set, one object may have multiple corresponding responses, see Fig. 14(a). The second level is that of the *merged responses*, which are results of applying a clustering algorithm to the original responses. The clustering algorithm randomly select one original response as a seed and merges the responses having large overlap with it; this procedure is applied iteratively until all original responses are processed. In the set of merged responses, one object has at most one corresponding response, see Fig. 14(b). The third level is that of the combined responses. One combined response has several matched part responses, see Fig. 14(c) for an example. The detection response may not be highly accurate spatially, because the training samples include some parts of the background regions in order to cover some position and size variations.

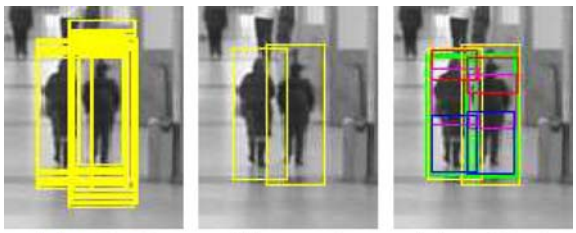

Figure 14. Detection responses. (a) and (b) are from the full-body detector; (c) is from the combined detector (green for combined; yellow for full-body; red for head-shoulder; purple for torso; blue for legs).

6. Human Tracking based on Part Detection

The human tracking algorithm takes the part detection and the combined detection responses as the observations of human hypotheses.

6.1. *Affinity for Detection Responses*

Both the original and the merged detection responses are part responses. For tracking we add two more elements to the representation of the part responses, $\mathbf{rp} = \{l, \mathbf{p}, s, v, f, \mathbf{c}\}$, where the new element f is a real-valued detection confidence, and $\mathbf{c}$ is an appearance model. The first five elements, l, $\mathbf{p}$, s, v and f, are obtained from the detection process directly. The appearance model, $\mathbf{c}$, is implemented as a color histogram; computation and update of $\mathbf{c}$ is described later, in detail, in Section 6.3. Representation of a combined response is the union of the representations of its parts, $\mathbf{rc} = \{\mathbf{rp}_i | l_i = FB, HS, T, L\}$.

Humans are detected frame by frame. In order to decide whether two responses, $\mathbf{rp}_1$ and $\mathbf{rp}_2$, of the same part type from different frames belong to one object, an affinity measure is defined

$$A(\mathbf{rp}_1, \mathbf{rp}_2) = A_{pos}(\mathbf{p}_1, \mathbf{p}_2) A_{size}(s_1, s_2) A_{appr}(\mathbf{c}_1, \mathbf{c}_2) \tag{17}$$

where A_{pos}, A_{size}, and A_{appr} are affinities based on position, size, and appearance respectively. Their definitions are

$$\begin{aligned} A_{pos}(\mathbf{p}_1, \mathbf{p}_2) &= \gamma_{pos} \exp\left[-\frac{(x_1 - x_2)^2}{\sigma_x^2}\right] \exp\left[-\frac{(y_1 - y_2)^2}{\sigma_y^2}\right] \\ A_{size}(s_1, s_2) &= \gamma_{size} \exp\left[-\frac{(s_1 - s_2)^2}{\sigma_s^2}\right] \\ A_{appr}(\mathbf{c}_1, \mathbf{c}_2) &= B(\mathbf{c}_1, \mathbf{c}_2) \end{aligned} \tag{18}$$

where $B(\mathbf{c}_1, \mathbf{c}_2)$ is the Bhattachayya distance between two histograms and γ_{pos} and γ_{pos} are normalizing factors. The affinity between two combined responses, $\mathbf{rc}_1$ and $\mathbf{rc}_2$, is the average of the affinity between their common

visible parts

$$A(\mathbf{rc}_1, \mathbf{rc}_2) = \frac{\sum_{l_i \in PT} A(Pt_i(\mathbf{rc}_1), Pt_i(\mathbf{rc}_2)) I(v_{i1}, v_{i2} > \theta_v)}{\sum_{l_i \in PT} I(v_{i1}, v_{i2} > \theta_v)} \tag{19}$$

where $Pt_i(\mathbf{rc})$ returns the response of the part i of the combined response $\mathbf{rc}$, v_{ij} is the visibility score of $Pt_i(\mathbf{rc}_j)$, $j = 1, 2$, and I is an indicator function. The above affinity functions encode the position, size, and appearance information.

Given the affinity, we match the detection responses with the human hypotheses in a similar way to that of matching part responses to human hypotheses described in Section 5. Suppose at time t of an input video, we have n human hypotheses $H_1^{(v)}, \ldots, H_n^{(v)}$, whose predictions at time $t+1$ are $\widehat{\mathbf{rc}}_{t+1,1}, \ldots, \widehat{\mathbf{rc}}_{t+1,n}$, and at time $t+1$ we have m responses $\mathbf{rc}_{t+1,1}, \ldots, \mathbf{rc}_{t+1,m}$. First we compute the $m \times n$ affinity matrix $\mathbf{A}$ of all $(\widehat{\mathbf{rc}}_{t+1,i}, \mathbf{rc}_{t+1,j})$ pairs, i.e. $\mathbf{A}(i, j) = A(\widehat{\mathbf{rc}}_{t+1,i}, \mathbf{rc}_{t+1,j})$. Then in each step, the pair, denoted by $(i^\star, j^\star)$, with the largest affinity is taken as a match and the $i^\star$-th row and the $j^\star$-th column of $\mathbf{A}$ are deleted. This procedure is repeated until no more valid pairs are available.

6.2. Trajectory Initialization

The basic idea of the initialization strategy is to start a trajectory when enough evidence is collected from the detection responses. Define the precision, pr, of a detector as the ratio between the number of successful detections and the number of all responses. If pr is constant between frames, and the detection in one frame is independent of the neighboring frames, then during consecutive T time steps, the probability that the detector outputs T consecutive false alarms is $P_{FA} = (1 - pr)^T$. However, this inference is not accurate for real videos, where the inter-frame dependence is large. If the detector outputs a false alarm at a certain position in the first frame, the probability is high that a false alarm will appear around the same position in the next frame. We call this the *persistent false alarm* problem. Even here, the real P_{FA} should be an exponentially decreasing function of T, we model it as $e^{-\lambda_{init}\sqrt{T}}$.

Suppose we have found $T(>1)$ consecutive responses, $\{\mathbf{rc}_1, \ldots, \mathbf{rc}_T\}$ corresponding to one human hypothesis $H^{(v)}$ by data association. The confidence of initializing a trajectory for $H^{(v)}$ is then defined by

$$InitConf\left(H^{(v)}; \mathbf{rc}_{1..T}\right) = \underbrace{\frac{1}{T-1}\sum_{t=1}^{T-1} A(\widehat{\mathbf{rc}}_{t+1}, \mathbf{rc}_{t+1})}_{(1)} \cdot \underbrace{\left(1 - e^{-\lambda_{init}\sqrt{T}}\right)}_{(2)} \tag{20}$$

The first term in the left side of Eq. (20) is the average affinity of the T responses, and the second term is based on the detector's accuracy. The more accurate the detector is, the larger should the parameter λ_{init} be. Our trajectory initialization strategy is: if $InitConf(H^{(v)})$ is larger than a threshold, θ_{init}, a trajectory is started from $H^{(v)}$, and $H^{(v)}$ is considered to be a *confident trajectory*; otherwise $H^{(v)}$ is considered to be a *potential trajectory*. In our experiments, $\lambda_{init} = 1.2, \theta_{init} = 0.83$. A trajectory hypothesis $H^{(v)}$ is represented as a triple, $\{\{\mathbf{rc}_t\}_{t=1,\ldots,T}, \mathbf{D}, \{\mathbf{C}_i\}_{i=FB,HS,TS,L}\}$, where $\{\mathbf{rc}_t\}$ is a series of responses, $\{\mathbf{C}_i\}$ is the appearance model of the parts, and $\mathbf{D}$ is a dynamic model. In practice, $\mathbf{C}_i$ is the average of the appearance models of all detection responses, and $\mathbf{D}$ is modeled by a Kalman filter for constant speed motion.

6.3. Trajectory Growth

After a trajectory is initialized, an object is tracked by two strategies: data association and meanshift tracking. For a new frame, for all existing hypotheses, we first look for their corresponding detection responses in this frame. If there is a new detection response matched with a hypothesis $H^{(v)}$, then $H^{(v)}$ grows based on data association, otherwise a meanshift tracker is applied. The data association itself has two steps. First, all hypotheses are matched with the combined responses by the method described in Section 6.1. Second, all hypotheses which are not matched in the first step are associated with the remaining part responses which do not belong to any combined response. Matching part responses with hypotheses is a simplified version of the method for matching combined responses with hypotheses. At least one part must be detected for an object to be tracked by data association. We do not associate the part responses with the tracks directly, because occlusion reasoning, which is done before association, from the detection responses in the current frame is more robust than from the predicted hypotheses, which are not very reliable.

Whenever data association fails (the detectors can not find the object or the affinity is low), a meanshift tracker (Comaniciu et al., 2001) is applied to track the parts individually. The results are combined to form the final estimation. The basic idea of meanshift is to track a probability distribution. Although the typical way to use meanshift tracking is to track a color distribution, there is no constraint on the distribution to be used. In our method we combine the appearance model, $\mathbf{C}$, the dynamic model, $\mathbf{D}$, and the detection confidence, f, to build a likelihood map which is then fed into the meanshift tracker. A dynamic probability map, $P_{dyn}(\mathbf{u})$, where $\mathbf{u}$ represents the image coordinates, is calculated from the dynamic model $\mathbf{D}$, see Fig. 15(d). Denote the original responses of one

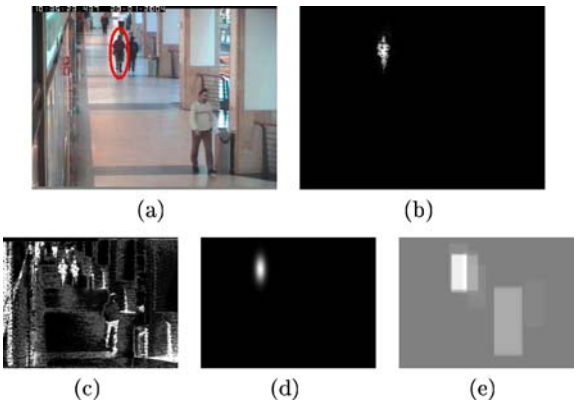

Figure 15. Probability map for meanshift: (a) original frame; (b) final probability map; (c), (d) and (e) probability maps for appearance, dynamic and detection respectively. (The object concerned is marked by a red ellipse.)

part detector at the frame j by $\{\mathbf{rp}_j\}$, the detection probability map $P_{det}(\mathbf{u})$ is defined by

$$P_{det}(\mathbf{u}) = \sum_{j:\mathbf{u}\in Reg(\mathbf{rp}_j)} f_j + ms \tag{21}$$

where $Reg(\mathbf{rp}_j)$ is the image region, a rectangle, corresponding to $\mathbf{rp}_j$, f_j is a real-valued detection confidence of $\mathbf{rp}_j$, and ms is a constant corresponding to the missing rate (the ratio between the number of missed objects and the total number of objects). ms is calculated after the detectors are learned. If one pixel belongs to multiple positive detection responses, then we set the detection score of this pixel as the sum of the confidences of all these responses. Otherwise we set the detection score as the average missing rate, which is a positive number. This detection score reflects the object salience based on shape cues. Note, the original responses are used here to avoid effects of errors in the clustering algorithm (see Fig. 15(e)).

Let $P_{appr}(\mathbf{u})$ be the appearance probability map. As $\mathbf{C}$ is a color histogram (the dimension is $32 \times 32 \times 32$ for r,g,b channels), $P_{appr}(\mathbf{u})$ is the bit value of $\mathbf{C}$ (see Fig. 15(c)). To estimate $\mathbf{C}$, we need the object to be segmented so that we know which pixels belong to the object; the detection response rectangle is not accurate enough for this purpose. Also, as a human is a highly articulated object, it is difficult to build a constant segmentation mask. Zhao and Davis (2005) proposed an iterative method for upper body segmentation to verify the detected human hypotheses. Here, we propose a simple PCA based approach. At the training stage, examples are collected and the object regions are labeled by hand, see Fig. 16(a). Then a PCA model is learned from this data, see Fig. 16(b). Suppose we have an initial appearance model $\mathbf{C}_0$, Given a new sample (Fig. 16(c)), first we calculate its color probability map from $\mathbf{C}_0$ (Fig. 16(d)),

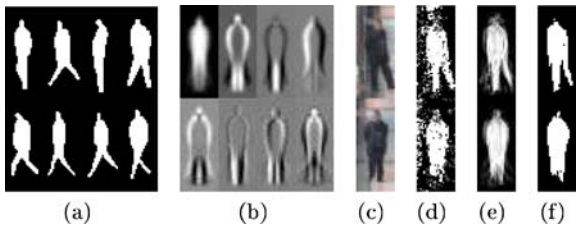

Figure 16. PCA based body part segmentation: (a) training samples; (b) eigenvectors. The left top one is the mean vector; (c) original human samples; (d) color probability map; (e) PCA reconstruction; (f) thresholded segmentation map.

then use the PCA model as a global shape constraint by reconstructing the probability map (Fig. 16(e)). The thresholded reconstruction map (Fig. 16(f)) is taken as the final object segmentation, which is used to update $\mathbf{C}_0$. The mean vector, the first one of Fig. 16(b), is used to compute $\mathbf{C}_0$ the first time. For each part, we learn a PCA model. This segmentation method is far from perfect, but very fast and adequate to update the appearance model.

Combining $P_{appr}(\mathbf{u})$, $P_{dyn}(\mathbf{u})$, and $P_{det}(\mathbf{u})$, we define the image likelihood for a part at pixel $\mathbf{u}$ by

$$L(\mathbf{u}) = P_{appr}(\mathbf{u})P_{dyn}(\mathbf{u})P_{det}(\mathbf{u}) \tag{22}$$

Figure 15 shows an example of probability map computation. Before the meanshift tracker is activated, inter-object occlusion reasoning is applied. Only the visible parts which were detected in the last successful data association, are tracked. Finally only the models of the parts which are detected and not occluded are updated. Meanshift tracking is not always performed and fused with association results, because the shape based detectors are much more reliable than the color based meanshift.

6.4. *Trajectory Termination*

The strategy of terminating a trajectory is similar to that of initializing it. If no detection responses are found for an object $H^{(v)}$ for consecutive T time steps, we compute a termination confidence of $H^{(v)}$ by

$$\begin{aligned} &EndConf\left(H^{(v)}; \mathbf{rc}_{1..T}\right) \\ &= \left(1 - \frac{1}{T-1}\sum_{t=1}^{T-1} A(\widehat{\mathbf{rc}}_{t+1}, \mathbf{rc}_{t+1})\right)\left(1 - e^{-\lambda_{end}\sqrt{T}}\right) \end{aligned} \tag{23}$$

Note that the combined responses $\mathbf{rc}_t$ are obtained from the meanshift tracker, not from the combined detector. If $EndConf(H^{(v)})$ is larger than a threshold, θ_{end}, hypothesis $H^{(v)}$ is terminated; we call it a *dead trajectory*, otherwise we call it an *alive trajectory*. In our experiments, $\lambda_{end} = 0.5$, $\theta_{end} = 0.8$.

Forward Human Tracking
Let the set of hypotheses be S, initially $S = \Phi$.
For each time step t (denote by S_t the set of all alive trajectories in S at time t)

1. Static detection:
   (a) Detect parts. Let the result set be RP_t.
   (b) Combine part detection responses, including inter-object occlusion reasoning. Let the result set be RC_t.
   (c) Subtract the parts used in RC_t from RP_t.
2. Data association:
   (a) Associate hypotheses in S_t with combined responses in RC_t. Let the set of matched hypotheses be S_{t1}.
   (b) Associate hypotheses in $S_t - S_{t1}$ with part responses in RP_t. Let the set of matched hypotheses be S_{t2}.
   (c) Build a new hypothesis $H^{(v)}$ from each unmatched response in RC_t, and add $H^{(v)}$ into S and S_t.
3. Pure tracking: For each confident trajectory in $S_t - S_{t1} - S_{t2}$, grow it by meanshift tracking.
4. Model update:
   (a) For each hypothesis in $S_{t1} + S_{t2}$, update its appearance model and dynamic model.
   (b) For each potential trajectory in S_{t1}, update its initialization confidence.
   (c) For each trajectory in $S_{t1} + S_{t2}$, reset its termination confidence to 0.
   (d) For each trajectory in $S_t - S_{t1} - S_{t2}$, update its termination confidence.

Output all confident trajectories in S as the final results.

Figure 17. Forward human tracking algorithm.

6.5. The Combined Tracker

Now we put the above three modules, trajectory initialization, tracking, and termination, together. Figure 17 shows the full *forward* tracking algorithm (it only looks ahead). Trajectory initialization has a delay; to compensate we also apply a *backward* tracking procedure which is the exact reverse of forward tracking. After a trajectory is initialized, it may grow in both forward and backward directions. Note that this is not the same as forward-backward filtering, as each detection is processed only once, either in the forward or in the backward direction. In the case where no image observations are available, and the dynamic model itself is not strong enough to track the object, we keep the hypothesis at the last seen position until either the hypothesis is terminated or some part of it is found again. When full occlusion is of short duration, the person could be reacquired by data association. However, if full occlusion persists, the track may terminate prematurely; such broken tracks could be combined at a higher level of analysis; we have not implemented this feature.

A simplified version of the combined tracking method is to track only a single part, e.g. the full-body. In the results in Section 7.2.3, we show that the combined tracking outperforms single part tracking. The combined tracking method is robust because:

1. The combined tracker uses combined detection responses, which have high precision, to start trajectories. This results in a very low false alarm rate at the trajectory initialization stage.
2. The combined tracker tries to find the corresponding part responses of an object hypothesis. The probability that at least one part detector matches is relatively high.
3. The combined tracker tries to follow the objects by tracking their parts, either by data association or by meanshift. This enables the tracker to work with both scene and inter-object occlusions.
4. The combined tracker takes the average of the part tracking results as the final human position. Hence even if the tracking of one part drifts, the position of the human can still be tracked accurately.

7. Experimental Results

We now present some experimental results. We note that our focus is on detection and tracking of humans where occlusions may be present and the camera may not necessarily be stationary. There are not many public data sets with these characteristics on which many results have been reported. Thus, we collected our own data set. We also include results on some data sets from earlier work, even though they consist largely of un-occluded humans in the center of the image, to facilitate comparision with earlier work. We separate the evaluation of detection and tracking modules. There are more reported systems for detection so we can provide more comparisons for detection than for tracking.

7.1. Detection Evaluation

We train our detectors by a large set of labeled samples and evaluate them on a number of test sets. First, in Section 7.1.2, we evaluate our body part detectors. Second, in Section 7.1.3, we evaluate our method with two public data sets, on which many previous papers report quantitative results (Mohan et al., 2001; Mikolajczyk et al., 2004; Dalal and Triggs, 2005); the samples in these

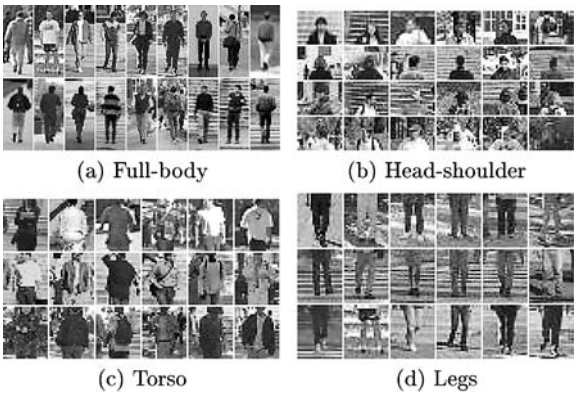

Figure 18. Examples of positive training samples.

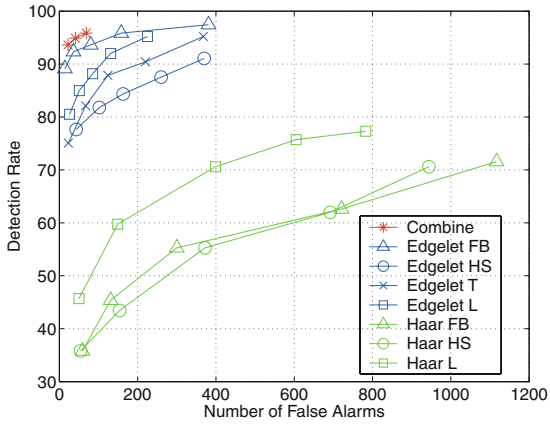

Figure 19. ROC curves of evaluation as detector on our test set (205 images with 313 humans).

two experiments are un-occluded ones. Third, in Section 7.1.4, we evaluate our method on images with occluded humans, where none of the above methods work. Before giving the evaluation results, we first describe our training set.

7.1.1. Training Set. Our training set contains 1,742 humans of frontal/rear view and 1,120 side view. Among these samples, 924 frontal/rear view ones are from the MIT pedestrian set (Papageorgiou et al., 1998) and the rest are 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. Figure 18 shows some examples from our training set. The negative image set contains 7,000 negative images without humans. During learning of the part and full-body detectors, 6,000 negative samples are used for each boosting stage. (The negative samples are patches cut from the negative images.) Note that this training set is used for all experiments in this work, except for that in Section 7.1.3.a, which is designed to compare with previous methods only on the MIT set.

7.1.2. Comparison of Part Detectors. We evaluate our edgelet based part detectors and compare with those based on Haar features (Kruppa et al., 2003). As there is no satisfactory benchmark data set for pedestrian detection task, we created one of our own. We collected a test set from the Internet containing 205 real-life photos and 313 different humans of frontal/rear view.[2] This set does not have heavy inter-object occlusion and is independent of the training set. We evaluated our edgelet detectors and the Haar feature based human detectors provided by OpenCV4.0b (Kruppa et al., 2003) on this test set. As the OpenCV detectors are only for frontal/rear view, we use the nested detector for frontal/rear view here for comparison. When the intersection between a detection response and a ground-truth box is larger than 50% of their union, we consider it to be a successful detection. Figure 19 shows the ROC curves of the part, full-body and combined detectors. Figure 20 shows some examples of successful detections and interesting false alarms, where locally the images look like the target parts. Figure 21 shows some image results of the combined detector. The

Figure 20. Examples of part detection results on images from our Internet test set. (Green: successful detection; Red: false alarm).

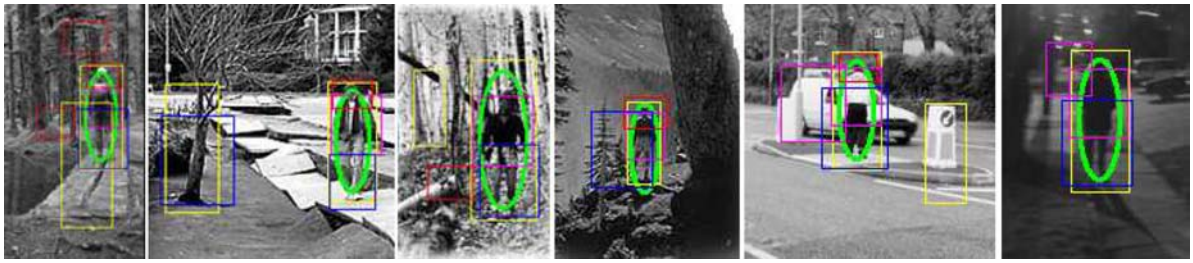

Figure 21. Examples of combined detection results on the Internet test set. (Green: combined response; yellow: full-body; red: head-shoulder; purple: torso; blue: legs).

sizes of the humans considered vary from 24 × 58 to 128 × 309.

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. 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 other part detectors; and (3) head-shoulder part detector is the least discriminative. The last observation is consistent with that reported in Mohan et al. (2001), but inconsistent with that in Mikolajczyk et al. (2004). Mohan et al. (2001) gave an explanation for the superiority of legs detector: the background of legs is usually road or grassland, which is relatively clutter-free compared to the background for head-shoulder. However, the legs detector of Mikolajczyk et al. (2004) is slightly inferior to their head-shoulder detector. This may be due to the fact that their legs detector covers all frontal, rear, and profile views.

7.1.3. Comparison of Classification Models. It is difficult to compare our method with previous ones due to variability in data sets and lack of access to the earlier methods' code. We show a comparison with other methods that report results on two public data sets, the MIT set and the INRIA set.[3] Note that these data sets contain un-occluded examples only. Also, these methods report classification (given a bounding box, predict the label of the sample) results rather than detection results; for a proper comparison, we also use classification results in this section.

7.1.3.a Comparison on the MIT Set. In Mikolajczyk et al. (2004), Dalal and Triggs (2005) and Mohan et al. (2001), the MIT pedestrian set is used to evaluate the methods. Mohan et al. (2001) used 856/866 positive and 9,315/9,260 negative samples to train their head-shoulder/legs detectors. The detection and false alarm rates were evaluated on a test set with 123 positive samples and 50 negative images. Mikolajczyk et al. (2004) trained their head-shoulder/legs 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. Dalal and Triggs (2005) trained a full-body detector with 509 positive samples and test with 200 images.

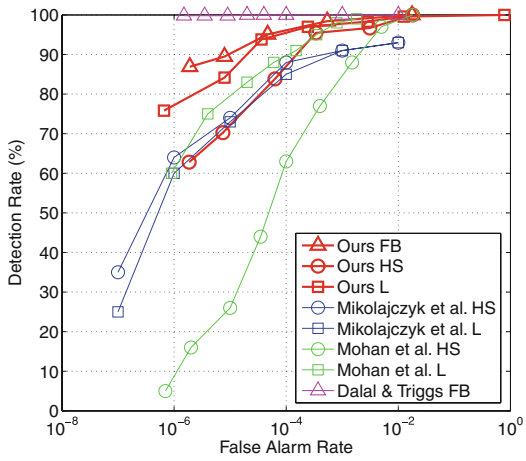

Figure 22. ROC curves of evaluation as classifier on MIT set. The results of Mikolajczyk et al. (2004), Dalal and Triggs (2005), and Mohan et al. (2001) are copied from the original papers.)

As mentioned before a direct comparison is difficult, so we compare in a less direct way. We trained our part detectors with 6/7 of the MIT set, and evaluated with the remaining 1/7 of the MIT set and 200 negative images. As all the samples in this set are for frontal/rear view point, we learn the nested structured detector here. Our experimental setup is comparable to that of Mohan et al. (2001), and Dalal and Triggs (2005). When training with only 300 positive samples, like in Mikolajczyk et al. (2004), our method suffered from over-fitting. Figure 22 shows the ROC curves. It can be seen that the full-body detector of Dalal and Triggs (2005) achieved the highest accuracy, almost perfect, on this set, and our full-body detector is the second best one.

7.1.3.b Comparison on the INRIA Set. As near-ideal results were achieved on the MIT data set, Dalal and Triggs (2005) concluded that the MIT set is too easy and they collected their own data set, called the INRIA data set. The INRIA set contains a training set, which has 614 positive samples and 1,218 negative images, and a test set, which has 564 positive samples and 453 negative images. The positive samples are spatially aligned and cover frontal, rear, and side views. Dalal and Triggs (2005) trained their classifiers on the INRIA training set and evaluated them on the INRIA test set. They report that with a false alarm rate of 10^{-4}, the HOG based classifier got a detection rate of about 90%.

We evaluate our tree structured multi-view full-body detector on it. Note that the tree detector is learned from our own training set described in Section 7.1.1. We do not use any training data from the INRIA set in this experiment. On the INRIA test set, our detector has a detection rate of about 93% with a false alarm rate of 10^{-4}. Again this is not a direct comparison, as the training sets are different. However it can be seen that our method is comparable to that in Dalal and Triggs (2005) in terms of classification accuracy, while the boosted cascade classifier is much more efficient computationally than the SVM classifier used in Dalal and Triggs (2005).

Note that (Mohan et al., 2001; Mikolajczyk et al., 2004; Dalal and Triggs, 2005) did experiments on 64 pixel wide samples, while our method requires samples to be 24 pixel wide only and still have comparable performance. This allows our method to be applicable for humans observed at farther distances.

7.1.4. Evaluation on Occluded Examples. To evaluate our combined detector with occlusion, we use 54 frames with 271 humans from the CAVIAR sequences (http://homepages.inf.ed.ac.uk/rbf/CAVIAR/). 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. We do not evaluate our method on all frames of the CAVIAR set, because the frames in video sequences have large correlation. Figure 23 shows the ROC curves of our part, full-body and the combined detectors on this set. The curve labeled "Combine*" in Fig. 23 shows the overall detection rate on the 75 occluded humans and Table 2 lists the detection rates on different degrees of occlusion. Figure 24 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 legs 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 Fig. 24, 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 results in a false alarm for the combined result only if the head-shoulder is found in the right relation to another human.

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

Occlusion degree (%)	25–50	50–75	>70
Human no.	34	31	10
Detection reate (%)	91.2	90.3	80

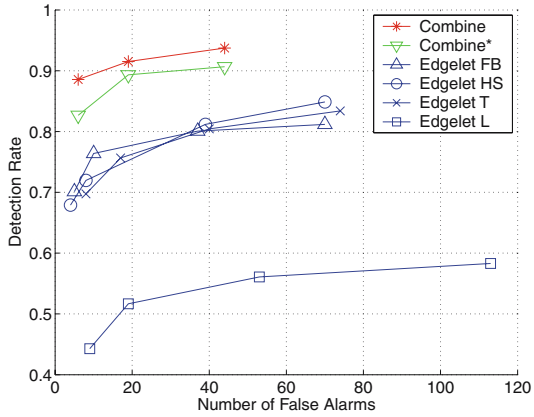

Figure 23. ROC curves of evaluation on our CAVIAR test set (54 images with 271 humans). Combine* is the detection rate on the 75 partially occluded humans.

7.2. *Tracking Evaluation*

We evaluated our human tracker on three video sets. The first set is a selection from the CAVIAR video corpus (http://homepages.inf.ed.ac.uk/rbf/CAVIAR/), which is captured with a stationary camera, mounted a few meters above the ground and looking down towards a corridor. The frame size is 384 × 288 and the sampling rate is 25 FPS. The second set, called the "skate board set", is captured from a camera held by a person standing on a moving skate board. The third set, called the "building top set", is captured from a camera held by a person standing on top of a 4-story building looking down towards the ground. The camera motions in the skate board set include both translation and panning, while those of the building top set are mainly panning and zooming. The frame size of these two sets is 720 × 480 and the sampling rate is 30 FPS. As the humans in the test videos include both frontal/rear and profile views, we use the

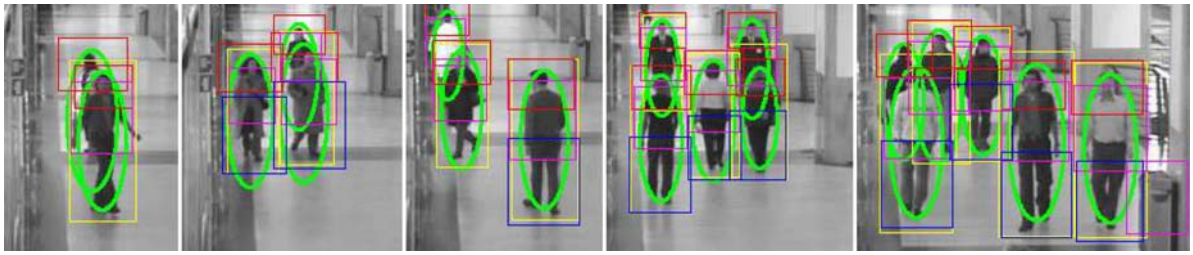

Figure 24. Examples of combined detection results on the CAVIAR test set. (Green: combined response; yellow: full-body; red: head-shoulder; purple: torso; blue: legs).

tree structured detectors for multi-view object detection in the tracking experiments. We compare our results on the CAVIAR set with a previous system from our group (Zhao and Nevatia, 2004a). We are unable to compare with others as we are unaware of published, quantitative results for tracking on this set by other researchers.

7.2.1. Tracking Performance Evaluation Criteria. To evaluate the performance of our system quantitatively, we define five criteria for tracking:

1. number of "mostly tracked" trajectories (more than 80% of the trajectory is tracked),
2. number of "mostly lost" trajectories (more than 80% of the trajectory is lost),
3. number of "fragments" of trajectories (a result trajectory which is less than 80% of a ground-truth trajectory),
4. number of false trajectories (a result trajectory corresponding to no real object), and
5. the frequency of identity switches (identity exchanges between a pair of result trajectories).

Figure 25 illustrates these definitions. These five categories are by no means a complete classification, however they cover most of the typical errors observed in our experiments.

7.2.2. Results on CAVIAR Set. The only previous tracker for which we have an implementation in hand is that of Zhao and Nevatia (2004a). In this experiment, we compared our method with that in Zhao and Nevatia (2004a). This method is based on background subtraction, and requires a calibrated stationary camera.

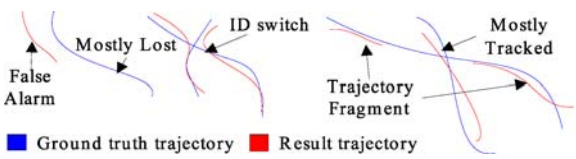

Figure 25. Tracking evaluation criteria.

Table 3. Tracking level comparison with (Zhao and Nevatia, 2004a) on CAVIAR set, 26 sequences.

	GT	MT	ML	Fgmt	FAT	IDS
Zhao-Nevatia	189	121	8	73	27	20
This Method		140	8	40	4	19

GT: ground-truth; MT: mostly tracked; ML: mostly lost; Fgmt: trajectory fragment; FAT: false alarm trajectory; IDS: ID switch.

For comparison, we build the first test set from the CAVIAR video corpus (http://homepages.inf.ed.ac.uk/rbf/CAVIAR/). Our test set consists of the 26 sequences for the "shopping center corridor view", overall 36,292 frames. The scene is relatively uncluttered, however the inter-object occlusion is intensive. Frequent interactions between humans, such as talking, and shaking hands, make this set very difficult for tracking. Our detectors require the width of humans to be larger than 24 pixels. In the CAVIAR set there are 40 humans, which are smaller than 24 pixels most of the time, and 6 humans, which are mostly out of the scene. We mark these small humans and out-of-sight humans in the ground-truth as "do not care". Table 3 gives the comparative results at tracking level.[4] It can be seen that our method outperforms the method of Zhao and Nevatia (2004a) when the resolution is good. This comes from the low false alarm rate of the combined detector. Some sample frames and results are shown in Fig. 26. However, on the small humans, our shape based method does not work (the combined tracker only gets only 1 out of the 40 small humans tracked) while the motion based tracker gets 21 small humans mostly tracked. This great superiority of the motion based tracker at low resolution is because the motion based method does not rely on a discriminative model of humans.

The comparison with the method in Zhao and Nevatia (2004a) is done on cases where both methods work. However, each has different limitations. The method of Zhao and Nevatia (2004a), which is based on 3D model and motion segmentation, is less view dependent and can work on lower resolution videos, while our method, which is based on 2D shape, requires higher resolution and does not work with large camera tilt angles. On the

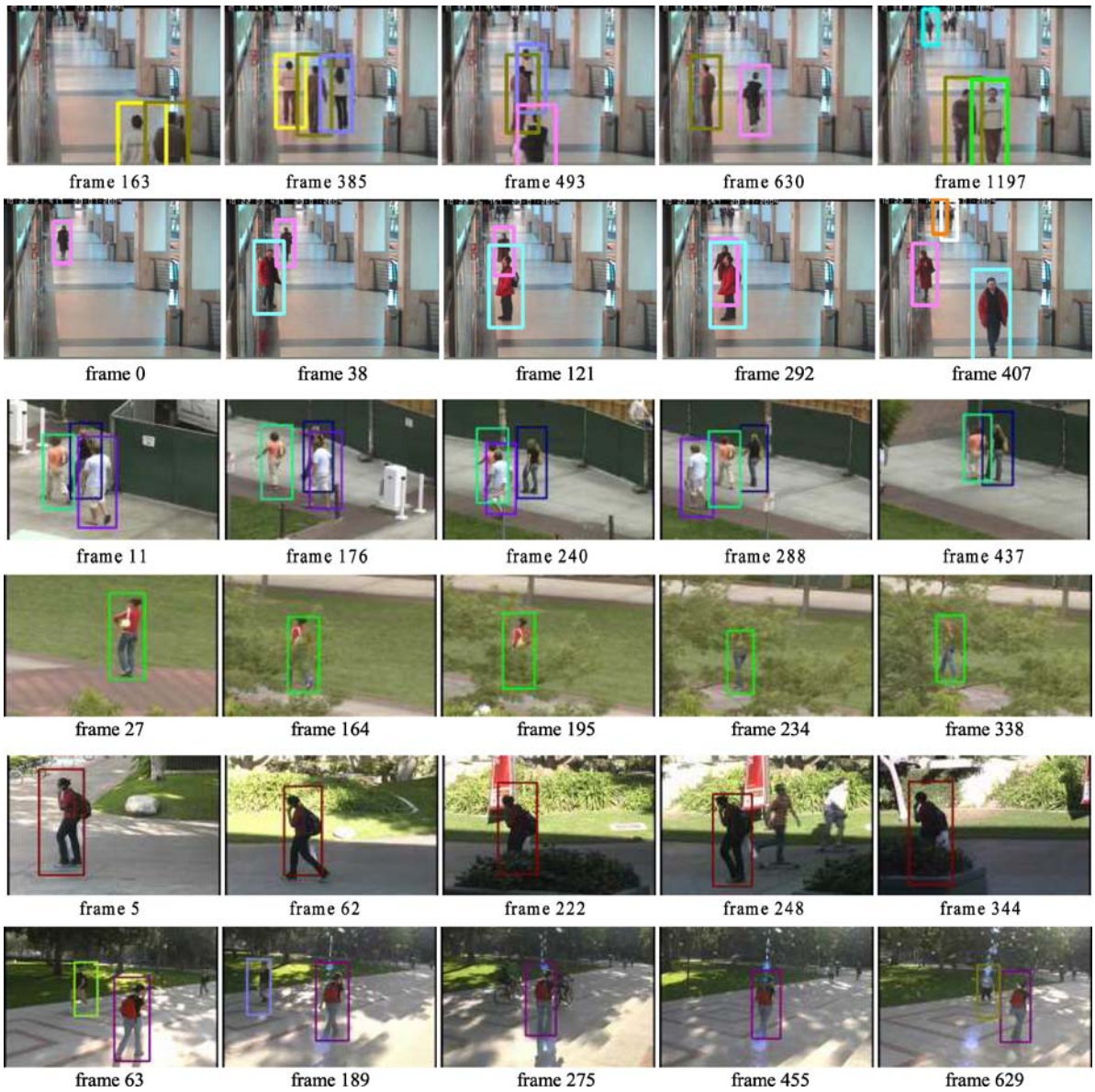

Figure 26. Sample tracking results. The 1st and the 2nd rows are from the CAVIAR set; the 3rd and the 4th rows are from the skate board set; the 5th and the 6th rows are from the building top set.

other hand, our method, which is based on frame by frame detection, can work with moving and/or zooming cameras, while the method of Zhao and Nevatia (2004a) can not.

The tracking method also greatly improves the detection performance (without considering the identity consistency). Table 4 gives the detection scores before and after tracking. We set the detection parameters to get a low false alarm rate.

Table 4. Detection performance before and after tracking.

		DR (%)	FAR (# PF)
Before tracking	Full-body detector	70.32	0.28
	Combined detector	57.91	0.05
After tracking		94.11	0.02

DR: detection rate; FAR: false alarm rate; PF: per frame.

7.2.3. Results on Skate Board Set. The main difficulties of the skate board set are small abrupt motions due to the uneven ground, and some occlusions. This set contains 29 sequences, overall 9,537 frames. Only 13 out of them have no occlusion at all. Some sample frames and results are shown in Fig. 26. The combined tracking method is applied. Table 5 gives the tracking performance

Table 5. Performance on skate board set, 29 sequences.

GT	MT	ML	Fgmt	FAT	IDS
50	39	1	16	2	3

See Table 3 for abbreviations.

Table 6. Comparison between part tracker and combined tracker on skate board set, 13 sequences.

	GT	MT	ML	Fgmt	FAT	IDS
Part tracking	21	14	2	7	13	3
Combined tracking		19	1	5	2	2

See Table 3 for the abbreviations.

of the system. It can be seen that our method works reasonably well on this set.

For comparison, a single part (full-body) tracker, which is a simplified version of the combined tracker, is applied on the 13 videos that have no occlusions. Because the part detection does not deal with occlusion explicitly, it is not expected to work on the other 16 sequences. Table 6 shows the comparison results. It can be seen that the combined tracker gives many fewer false alarms than the single part tracker. This is because the full-body detector has more persistent false alarms than the combined detector. Also the combined tracker has more fully tracked objects, because it makes use of cues from all parts.

7.2.4. Results on Building Top Set. The building top set contains 14 sequences, overall 6,038 frames. The main difficulty of this set is due to frequency of occlusions, both scene and object, see Table 8. No single part tracker works well on this set. The combined tracker is applied to this data set. Table 7 gives the tracking performance. It can be seen that the combined tracker obtains very few false alarms and a reasonable success rate. Some sample frames and results are shown in Fig. 26.

Table 7. Performance on building top set, 14 sequences.

GT	MT	ML	Fgmt	FAT	IDS
40	34	3	3	2	2

See Table 3 for the abbreviations.

Table 8. Frequencies of and performance on occlusion events. n/m: n successful tracked among m occlusion events.

Video set		SS	LS	SO	LO	Overall
CAVIAR	Zhao-Nevatia	0/0	0/0	40/81	6/15	46/96
	This method	0/0	0/0	47/81	10/15	57/96
Skate board		6/7	2/2	11/16	0/0	19/25
Building top		4/7	11/13	15/18	4/4	34/42

SS: short scene; LS: long scene; SO: short object; LO: long object.

7.2.5. Tracking Performance with Occlusions. We characterize the occlusion events in these three sets with two criteria: if the occlusion is by a target object, i.e. a human, we call it an object occlusion, otherwise a scene occlusion. If the period of the occlusion is longer than 50 frames, it's considered to be a long term occlusion; otherwise a short term one. So we have four categories: short term scene, long term scene, short term object, and long term object occlusions. Table 8 gives the tracking performance on occlusion events. Tracking success of an occlusion event means that no object is lost, no trajectory is broken, and no ID switches occur during the occlusion. It can be seen that our method can work reasonably well in the presence of scene or object partial occlusion, even long term ones. The performance on the CAVIAR set is not as good as those on the other two sets. This is because 19 out of 96 occlusion events in the CAVIAR set are fully occluded ones (more than 90% of the object is occluded) while the occlusions in the other two sets are all partial ones.

For tracking, on average, about 50% of the successful tracking is due to the data association with combined responses, i.e. the object is "seen" by the combined detector; about 35% is due to the data association with part responses; the remaining 15% is from the meanshift tracker. Although the detection rate of any individual part detector is not high, the tracking level performance of the combined tracker is much better. The speed of the entire system is about 1 FPS. The machine used is a 2.8 GHz 32-bit Pentium PC. The program is coded in C++ using OpenCV functions. Most of the computation cost is in the static detection component. We do not tune the system parameters for different sequences. Basically, we have three sets of parameters for the three video sets. The main different parameters are the searching range of the 2D human size, as the image size of humans in the CAVIAR set is much smaller than those in the other two sets, and the parameters for the Kalman filter, as the image motion of humans with moving/zooming camera is much more noisy than that with stationary camera.

8. Conclusion and Discussion

We have described a human detection and tracking method based on body part detection. Body part detectors are learned by boosting edgelet feature based weak classifiers. We defined a joint likelihood for multiple humans based on the responses of part detectors and explicit modeling of inter-object occlusion.

The responses of the combined human detector and the body part detectors are taken as the observations of the human hypotheses and fed into the tracker. Both the trajectory initialization and termination are based on the evidence collected from the detection responses. To track

the objects, most of the time data association works, while a meanshift tracker fills in the gaps between data association. From the experimental results, it can be seen that the proposed system has low false alarm rate and achieves a high tracking accuracy. It can work under both partial scene and inter-object occlusion conditions reasonably well. We have also applied this framework to other applications, e.g. speaker tracking in seminar videos (Wu et al., 2006) and conferee tracking in meeting videos (Wu and Nevatia, 2006c), and have achieved good scores in the VACE (http://www.ic-arda.org/InfoExploit/vace/) and CHIL (http://chil.server.de/servlet/is/101/) evaluations.

We learn our detectors with a sample size of 24 × 58 pixels, as this is common for real applications, such as visual surveillance. However at such a small scale, some body parts are not very distinguishable, e.g. head-shoulder. Learning part detectors with different scales could be a better choice.

Currently our system does not make use of any cues from motion segmentation. When motion information is available, it should help improve the tracking performance. For example, recently Brostow and Cipolla (2006) proposed a method to detect independent motions in crowds. The outputs are *tracklets* of independently moving entities, which may facilitate object level tracking. Conversely, shape-based tracking can help improve motion segmentation.

We have not explored the interaction between detection and tracking. The current system works in a sequential way: tracking takes the results of detection as input. However, tracking can be used to facilitate detection. One of the most straightforward ways is to speedup detection by restrainting the searching in the neighborhood of prediction by tracking. We plan to study such interactions in future work.

In our current system, four general human part detectors, which are learned off-line, are used. However during tracking, if these general detectors are somehow adapted to a specific environment, we could achieve both higher accuracy and better efficiency. There is some existing work on online learning of classifiers for object detection and tracking, (e.g., Avidan, 2005; Grabner and Bischof, 2006). We plan to investigate improving our detectors by online learning in future work.

Acknowledgments

The authors would like to thank Mr. Tae Eun Choe and Mr. Qian Yu for their help for capturing the videos, and Dr. Naveneet Dalal and Dr. Bill Triggs for kindly providing the program to generate the ROC curves of their method. This research was partially funded by the Advanced Research and Development Activity of the U.S. Government under contract MDA-904-03-C-1786 and the Disruptive Technology Office of the U.S. Government under contract DOI-NBC-#NBCHC060152.

Notes

1. http://cbcl.mit.edu/software−datasets/PedestrianData.html.
2. http://iris.usc.edu/ bowu/DatasetWebpage/dataset.html.
3. http://pascal.inrialpes.fr/data/human/.
4. In our previous paper (Wu and Nevatia, 2006a), we show results on a subset, 23 sequences, only, as ground-truth for three sequences was not available at that time.

References

Avidan, S. 2005. Ensemble tracking. *CVPR*, vol. II, pp. 494–501.

Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., and Wolf, H.C. 1977. Parametric correspondence and chamfer matching: Two new techniques for image matching. *IJCAI*, pp. 659–663.

Brostow, G.J. and Cipolla, R. 2006. Unsupervised bayesian detection of independent motion in crowds. *CVPR*, vol. I, pp. 594–601.

Comaniciu, D., Ramesh, V. and Meer, P. 2001. The variable bandwidth mean shift and data-driven scale selection. *ICCV*, vol. I, pp. 438–445.

Dalal, N. and Triggs, B. 2005. Histograms of oriented gradients for human detection. *CVPR*, vol. I, pp. 886–893.

Davis, L., Philomin, V. and Duraiswami, R. 2000. Tracking humans from a moving platform. *ICPR*, vol. IV, pp. 171–178.

Felzenszwalb, P. 2001. Learning models for object recognition. *CVPR*, vol. I, pp. 56–62.

Freund, Y. and Schapire R.E. 1996. Experiments with a New Boosting Algorithm. *The 13th Conf. on Machine Learning*, pp. 148–156.

Gavrila, D. and Philomin, V. 1999. Real-time object detection for "Smart" Vehicles. *ICCV*, vol. I, pp. 87–93.

Gavrila, D. 2000. Pedestrian detection from a moving vehicle. *ECCV*, vol. II, pp. 37–49.

Grabner, H. and Bischof, H. 2006. Online boosting and vision. *CVPR*, vol. I, pp. 260–267.

Huang, C., Ai, H., Wu, B., and Lao. S. 2004. Boosting nested cascade detector for multi-view face detection. *ICPR*, vol. II, pp. 415–418.

Huang, C., Ai, H., Li, Y., and Lao, S. 2005. Vector boosting for rotation invariant multi-view face detection. *ICCV*, vol. I, pp. 446–453.

http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

http://www.ic-arda.org/InfoExploit/vace/

http://chil.server.de/servlet/is/101/

Isard, M. and MacCormick, J. 2001. BraMBLe: A bayesian multiple-blob tracker. *ICCV*, vol. II, pp. 34–41.

Kruppa, H., Castrillon-Santana, M., and Schiele, B. 2003. Fast and robust face finding via local context. *Joint IEEE Int'l Workshop on VS-PETS*.

Kuhn, H.W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–87.

Lee, M. and Nevatia, R. 2006. Human pose tracking using multi-level structured models. *ECCV*, vol. III, pp. 368–381.

Leibe, B., Seemann, E. and Schiele B. 2005. Pedestrian detection in crowded scenes. *CVPR*, vol. I, pp. 878–885.

Lowe, D.G. 1999. Object recognition from local scale-invariant features. *ICCV*, vol. II, pp. 1150–1157.

Mikolajczyk, C., Schmid, C., and Zisserman, A. 2004. Human detection based on a probabilistic assembly of robust part detectors. *ECCV*, vol. I, pp. 69–82.

Mohan, A., Papageorgiou, C., and Poggio, T. 2001. Example-based object detection in images by components. *Trans. PAMI*, 23(4):349.

Papageorgiou, C., Evgeniou, T., and Poggio, T. 1998. A trainable pedestrian detection system. In *Proceeding of Intelligent Vehicles*, pp. 241–246.

Peter, J.R., Tu, H., and Krahnstoever, N. 2005. Simultaneous estimation of segmentation and shape. *CVPR*, vol. II, pp. 486–493.

Ramanan, D., Forsyth, D.A., and Zisserman, A. 2005. Strike a pose: Tracking people by finding stylized poses. *CVPR*, vol. I, pp. 271–278.

Schapire, R.E. and Singer, Y. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336.

Shashua, A., Gdalyahu, Y., and Hayun, G. 2004. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. *IEEE Intelligent Vehicles Symposium*, Parma, Italy, pp. 1–6.

Sigal, L., Bhatia, S., Roth, S., Black, M.J., and Isard M. 2004. Tracking loose-limbed people. *CVPR*, vol. I, pp. 421–428.

Smith, K., G.-Perez, D., and Odobez, J.-M. 2005. Using particles to track varying numbers of interacting people. *CVPR*, vol. I, pp. 962–969.

Viola, P. and Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. *CVPR*, vol. I, pp. 511–518.

Viola, P., Jones, M., and Snow, D. 2003. Detecting pedestrians using patterns of motion and appearance. *ICCV*, pp. 734–741.

Wren, C.R., Azarbayejani, A., Darrell, T., and Pentland, A.P. 1997. Pfinder: Real-time tracking of human body. *IEEE Trans. PAMI*, vol. 19, no. 7.

Wu Y., Yu, T., and Hua. G. 2005. A statistical field model for pedestrian detection. *CVPR*, vol. I, pp. 1023–1030.

Wu, B. and Nevatia, R. 2006a. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. *ICCV*, vol. I, pp. 90–97.

Wu, B. and Nevatia, R. 2006b. Tracking of multiple, partially occluded humans based on static body part detection. *CVPR*, vol. II, pp. 951–958.

Wu, B. and Nevatia, R. 2006c. Tracking of multiple humans in meetings. In V4HCI'06 workshop, *in conjunction with CVPR*, pp. 143–150.

Wu, B., Singh, V.K., Nevatia, R., and Chu, C.-W. (2006). Speaker tss-racking in seminars by human body detection. In *CLEAR 2006 Evaluation Campaign and Workshop, in conjunction with FG*.

Zhao, T. and Nevatia, R. 2004a. Tracking multiple humans in crowded environment. *CVPR*, vol. II, pp. 406–413.

Zhao, T. and Nevatia, R. 2004b. Tracking multiple humans in complex situations. *IEEE trans. on PAMI*, 26(9):1208–1221.

Zhao, L. and Davis, L. 2005. Closely coupled object detection and segmentation. *ICCV*, vol. I, pp. 454–461.