Segmentation of Multiple, Partially Occluded Objects by Grouping, Merging, Assigning Part Detection Responses

Bo Wu, Ram Nevatia, and Yuan Li
University of Southern California
Institute for Robotics and Intelligent Systems
Los Angeles, CA 90089-0273
$\{bowu|nevatia|yli8\}@usc.edu$

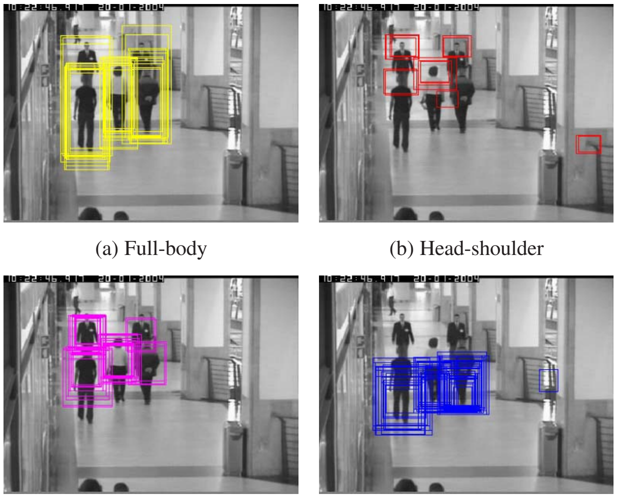
Abstract

We propose a method that detects and segments multiple, partially occluded objects in images. A part hierarchy is defined for the object class. Whole-object segmentor and part detectors are learned by boosting shape oriented local image features. During detection, the part detectors are applied to the input image. All the edge pixels in the image that positively contribute to part detection responses are extracted. A joint likelihood of multiple objects is defined based on the part detection responses and the object edges. Computing the joint likelihood includes an inter-object occlusion reasoning that is based on the object silhouettes extracted with the whole-object segmentor. By maximizing the joint likelihood, part detection responses are grouped, merged, and assigned to multiple object hypotheses. The proposed approach is applied to the pedestrian class, and evaluated on two public test sets. The experimental results show that our method outperforms the previous ones.

1. Introduction

Detection and segmentation of objects of a given class is a fundamental problem of computer vision. Recently, promising results have been achieved for several classes, including faces [18, 9], pedestrians [12, 8, 7, 10, 17, 11, 16], and cars [3, 19]. Many detection methods learn object classifiers from a labeled training set. Given a test image, the classifier is applied to the sub-windows with variable sizes at all positions. For detection of objects with partial occlusions, part based representations can be used. For each part, a detector is learned and the part detection responses are combined.

The part detectors are typically applied to overlapping windows and the windows are classified independently, hence one local feature may contribute to multiple overlapped responses for one object, see Fig.1. Some false detections may also occur, as local features may not be discriminative enough. Due to poor image cues or partial occlusions, some object parts may not be detected. To get a one-to-one mapping from part detection responses to object hypotheses, we need to group the responses and explain inconsistency between the observation and the hypotheses. When objects are close to each other, both the one-object-multiple-response problem and the part-object assignment problem require joint consideration of multiple objects, instead of treating them independently. We propose a unified framework for part response grouping, merging and assigning, and demonstrate that it outperforms the previous related methods.


(a) Full-body

(b) Head-shoulder

(c) Torso

(d) Legs

Figure 1. Examples of part detection responses for pedestrians.

1.1. Related work

Many previous efforts on detection, *i.e.* [7, 8], focus on the classification accuracy, measured by detection rate and false positive per window. However, the sub-window classification results are not the final outputs of a detection system. One main post-process is to “merge” the positive responses having large overlap and expect that each of the resulting clusters corresponds to one object, *e.g.* the aggregate clustering used in [21] and the adaptive bandwidth mean-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

shift used in [7]. Usually, thresholding on overlap is used to determine if two responses are from the same object. Setting this threshold can be tricky when objects are close to each other. Some other methods try to generate human hypotheses by directly grouping image features, *e.g*. the recent work by Sharma and Davis [4]. These methods treat individual objects independently. They do not consider the interaction between multiple objects.

Recently, part based representations have been used to detect objects with partial occlusions, *e.g*. [12, 6, 2, 15]. In these methods, several part detectors are learned by supervised learning. One way to build a set of part detectors is to train them independently, like in [12, 6]. However, this increases the time complexity of training linearly w.r.t the number of parts. Another way is to build one part detector as a true subset of the whole-object detector. For example, in [15] each sub-region detector use a subset of features of the whole-region detector and only the decision thresholds are different. The main limitation of this method is that a subset of features of the whole-object model may not be sufficient to construct a good part model.

For detection, the part detectors are applied to the input image and the detection responses are merged with some clustering method, as in the case of single object detector. When assigning part responses to object hypotheses, some joint analysis is done to cover partially occluded cases [12, 6, 2]. A joint image likelihood of multiple objects is computed by awarding successful part detection and penalizing missed detection of visible parts and false alarms. Different hypotheses configurations are tested, and the one with the highest likelihood is kept as the final interpretation of the image. The inputs of the part combination stage are the bounding boxes of parts. These are relatively coarse representations from which we can not get an accurate occlusion model. In addition, the errors from the overlap thresholding at the response merging stage are hard to correct at the part combination stage. Different from the part combination methods, Leibe *et al*. [14] propose an Implicit Shape Model based approach to detect multiple humans. Joint analysis is done to cover occluded objects.

1.2. Outline of our approach

Fig.2 shows a diagram of our approach. We define a part hierarchy for an object class, in which each part is a sub-region of its parent. As building part detectors independently is time consuming and building them as sub-sets of the whole-object detector may not be able to achieve a desirable accuracy, we choose a trade-off between these two approaches. For each part, a detector is learned by boosting local shape features. A child node in the hierarchy inherits its image features from its parent node and if a target performance can not be achieved from the inherited features, more features are selected and added to the child node. For whole-object, besides the detector a pixel-level segmentor is learned. Given a new image, the part detectors are applied. The image edge pixels that positively contribute to the detection responses are extracted. The part responses and object edges form an informative intermediate representation of the original image.

In our approach, we do not divide the tasks of merging responses and part combination into two separate stages; instead, we try to solve them under the same framework. From the part detection responses, multiple object hypotheses are proposed. For each hypothesis, a pixel-level segmentation is obtained by applying the whole-object segmentor, and the silhouettes are extracted. We apply occlusion reasoning to the object silhouettes. For joint image likelihood of multiple objects, besides the reward of successful detection, and penalties of missed detection and false alarm, we add a matching score between visible silhouettes and the object edges. Our joint analysis enforces the *exclusiveness* of low level features, *i.e*. one image feature can contribute to at most one hypothesis.

Our approach is a unified MAP framework that solves part merging, grouping, and assigning together. The main contributions include: 1) a part hierarchy design that enables efficient learning of part detectors by feature sharing; 2) an accurate occlusion reasoning approach based on silhouettes; 3) a joint image likelihood based on both detection responses and object edges that are assigned to object hypotheses exclusively. We demonstrate our approach on the class of pedestrians. Every module in our approach contributes to the robustness of the whole system. Though the situations where any single module may have an advantage are not frequent, together they result in a statistically significant improvement compared to the previous methods.

The rest of this paper is organized as follows: section 2 describes the part detector hierarchy; section 3 introduces our joint analysis algorithm for multiple objects; section 4 shows the experimental results; and some conclusions and discussions are given in the last section.

2. Hierarchical Body Part Detectors

We use the class of pedestrians to illustrate and validate our approach. We define a part hierarchy for human body, which consists of three levels including a full-body node and 11 body part nodes, see Fig.3.

2.1. Learning part detectors

For each node, a detector is learned. As we define the part hierarchy such that the region of one child node is a sub-region of its parent node, feature sharing between the parent and child nodes is possible. For each part node, a boosting algorithm is used to select good local shape features and construct a classifier as detector. The image features used are *edgelets* as in [12]. Before the regular boosting procedure, the detector of one node, except for the “full-body” node, inherits all the edgelet features overlapping

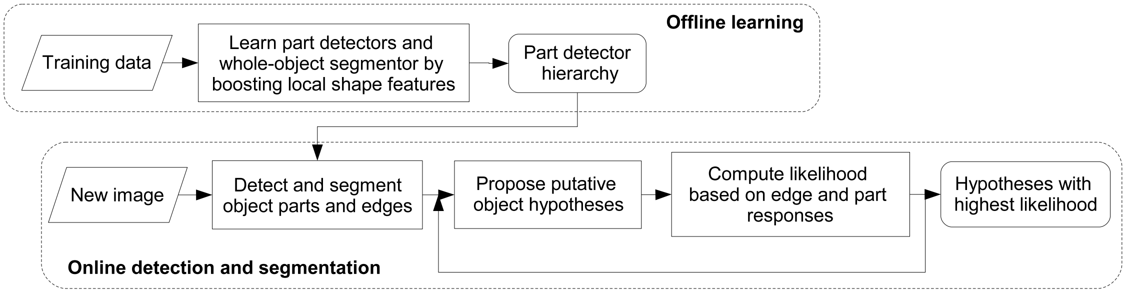

Offline learning
Training data
Learn part detectors and whole-object segmentor by boosting local shape features
Part detector hierarchy
New image
Detect and segment object parts and edges
Propose putative object hypotheses
Compute likelihood based on edge and part responses
Hypotheses with highest likelihood
Online detection and segmentation

Figure 2. Overall diagram of our approach.

with its sub-region from its parent node. For each inherited edgelet, the points that are out of the part's sub-region are removed, and the classification function is re-trained. Usually the detector can not achieve a high accuracy from the inherited features only. The regular boosting algorithm is then applied to add more features to the classifier. Fig.4 gives an illustration of feature sharing. The boosting algorithm used is the *Cluster Boosting Tree (CBT)* method in [3]. More details of the experimental setting are given later in section 4.

For the full-body node, we use the method in [5] to learn a pixel-level segmentor. Note, we do not learn segmentors for the other body parts. Because the full-body segmentor is based on local features, even when the object is partially occluded, the full-body segmentor can still segment the visible part well based on the visible features.

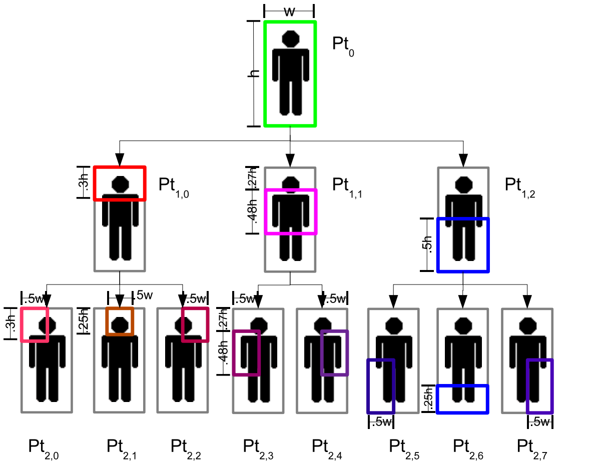


Figure 3. Hierarchy of human body parts. (Pt_0 is full-body; $Pt_{1,0}$ head-shoulder; $Pt_{1,1}$ torso; $Pt_{1,2}$ legs; $Pt_{2,0}$ left shoulder; $Pt_{2,1}$ head; $Pt_{2,2}$ right shoulder; $Pt_{2,3}$ left arm; $Pt_{2,4}$ right arm; $Pt_{2,5}$ left leg; $Pt_{2,6}$ feet; $Pt_{2,7}$ right leg. The left and right sides here are w.r.t. the 2-D image space.)

2.2. Detecting body parts and object edges

Given a new image, the part detectors are applied. Besides collecting part responses, we extract image edges that correspond to objects. For each edgelet feature f in the classifier, we call it a *positive feature* if it has higher average matching score on positive samples than on negative samples, *i.e.*

$$E\{f(\mathbf{x})|\mathbf{x} \in \mathcal{X}_+\} > E\{f(\mathbf{x})|\mathbf{x} \in \mathcal{X}_-\} \tag{1}$$

where $\mathcal{X}_\pm$ is positive/negative sample space. The computation of the matching score between an edgelet feature and an image is similar to edge template matching [12]. The average matching scores are evaluated during the off-line learning stage. For one sub-window that is classified as object, the positive features in the sub-window are ranked according to their matching scores. The positive features with top 5% scores are retained.

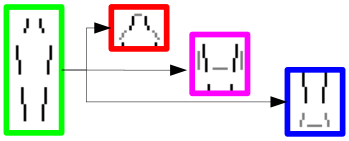


Figure 4. Illustration of feature sharing in part detector hierarchy. (The black ones are the inherited features, and the gray are the newly selected features.)

As one detector usually contains about one thousand positive features, a large number of edgelets are kept for one image. Some of these edgelets correspond to the same edge pixels. We apply a clustering algorithm to prune redundant edgelets. An edgelet consists of a chain of 2-D points. Denote the positions of the points in an edgelet by $\{\mathbf{u}_i\}_{i=1}^k$, where k is the length of the edgelet. Given two edgelets e_1 and e_2 with the same length, we define an affinity between them by

$$A\left(e_1, e_2\right) \triangleq \frac{1}{k}\sum_{i=1}^{k}\langle \mathbf{u}_{1,i} - \bar{\mathbf{u}}_1, \mathbf{u}_{2,i} - \bar{\mathbf{u}}_2\rangle \cdot e^{-\frac{1}{2}\|\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2\|^2} \tag{2}$$

where $\bar{\mathbf{u}}$ is the mean of $\{\mathbf{u}_i\}$. If the two features have different numbers of points, k_1 and k_2, first they are aligned by their center points, and then the longer feature is truncated to the length of the shorter one by removing points from the two ends. The affinity given by Equ.2 multiplied by a factor of $\frac{\min\{k_1,k_2\}}{\max\{k_1,k_2\}}$ is taken as the affinity for these edgelets.

The clustering algorithm is an iterative algorithm. First, we find the edgelet with the highest matching score, and then remove all edgelets with high affinity to it. This procedure is repeated until all object edgelets are examined. The remaining edgelets are the observations that support the putative object hypotheses, see Fig.5 for an example. Compared to general edge based image segmentation methods, where all edges are extracted, our edge extraction removes edges from background clutters and focuses on object shapes. These object edges, together with the bounding boxes, are input for the joint analysis of multiple objects.

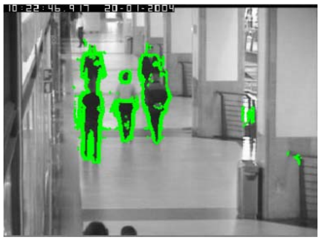


Figure 5. Extracted object edgelet pixels.

3. Joint Analysis for Multiple Objects

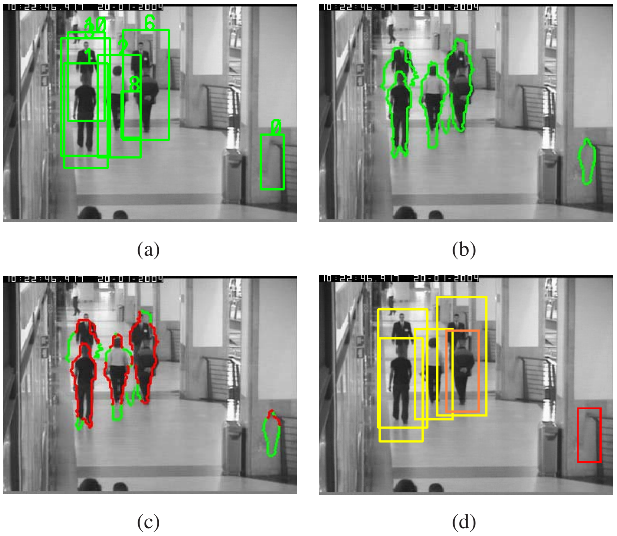
Similar to the previous methods [12, 6, 2], our joint analysis takes the detection results as input and searches for the multiple object configuration with the best image likelihood. The difference is that we enforce feature exclusiveness among multiple hypotheses, do occlusion reasoning to compute 1-D silhouette based *visibility score*, and add the object edge information into the likelihood definition. Fig.6 lists the main steps of the algorithm.

```
1. Propose initial object hypotheses sorted such that their $y$-coordinates are in a descending order.
2. Segment object hypotheses and extract their silhouettes.
3. Examine the hypotheses one by one, from front to back
    (a) For one hypothesis $H$, compute the joint occlusion maps for silhouettes of multiple objects, with and without $H$;
    (b) Match the detection responses and object edgelets with visible silhouettes;
    (c) Compute the image likelihood with $H$, $P_w(H)$, and the likelihood without $H$, $P_{w/o}(H)$;
    (d) If $P_w(H) > P_{w/o}(H)$, keep the hypothesis; otherwise remove it.
4. Output all remaining hypotheses.
```

Figure 6. Searching for the best multiple object configuration.

3.1. Proposing object hypotheses

Initially, object hypotheses are proposed from the detection responses of a subset of parts. For pedestrians, we use full-body, head-shoulder, left/right shoulder, and head to propose. During detection, only the part detectors for initial hypothesis proposal are applied to the whole image, while the others are applied to the local neighborhood around the initial hypotheses. The hypotheses with large overlap ratio, which is defined as the area of their intersection over the area of their union, are merged. Different from the traditional merging step [21], we use a high overlap threshold to obtain a set of "under-merged" responses, in which one object may have multiple hypotheses but hypotheses of different objects are unlikely to be merged. Although this under-merging reduces the search space, it can keep the responses of close by objects separate for further joint analysis. We sort the object hypotheses such that their y-coordinates are in a descending order, see Fig.7(a) for an example.



(a)

(b)

(c)

(d)

Figure 7. Computing joint image likelihood for multiple objects. a) the examined multiple object configuration. b) the visible silhouettes obtained by occlusion reasoning. c) the parts of the silhouettes that have matched edgelets (red points). d) result of matching full-body detection responses in Fig.1(a) with the current hypotheses (yellow: matched responses; orange: response not matched with any hypothesis; red: hypothesis without matched response).

3.2. Joint occlusion map of silhouettes

For each hypothesis, segmentation is computed by applying the whole-object segmentor and its silhouette is extracted. As in [12, 6, 2], we assume that objects are on a ground plane and the camera looks down towards the plane, so that the relative depths of objects can be inferred from their image coordinates. We render the segmentation masks of the ordered hypotheses by a z-buffer like method, and remove the invisible parts of the silhouettes that are out of image frame or occluded by other objects, see Fig.7(b). For each part of an object hypothesis, a visibility score is defined as the ratio between the length of the visible silhouette and the length of the whole silhouette. Compared to the 2-D region based visibility score in the previous work [12, 6, 2], the 1-D silhouette based visibility score is more accurate and meaningful for the shape based detectors. For example, when the region of a big object in the back is mostly occluded by a smaller object in the front, the silhouette based occlusion reasoning can retain the back one for further analysis as long as its contour is mostly visible.

3.3. Matching object edges with visible silhouettes

After getting the visible silhouettes, we assign the object edgelets extracted during part detection to the hypotheses by matching them with the visible silhouettes. For each edgelet, we find the closest silhouette to it and align the edgelet with the silhouette. Fig.8 gives the algorithm.

```
1. Compute distance transformation for all silhouettes;
2. For each object edgelet
    (a) Compute Chamfer matching scores to all the silhouettes, and assign the edgelet to the silhouette with the largest score;
    (b) Find the silhouette point $\mathbf{c}$ nearest to the edgelet and locally align the edgelet with the silhouette around $\mathbf{c}$;
    (c) Mark the part of the silhouette that is covered by the edgelet as "supported";
```

Figure 8. Matching and aligning edgelets with silhouettes.

To assign edgelets to silhouettes, first we compute the distance transformation for each visible silhouette. Then, we compute the Chamfer matching scores between all the edgelets and all the silhouettes through distance transformation. One edgelet is assigned to the silhouette that has the highest matching score with it. (If one edgelet has low scores with all the silhouettes, then it is not assigned to any.)

To align one edgelet with its corresponding silhouette, first we find the silhouette point $\mathbf{c}$ closest to the edgelet through distance transformation. Then we search a small neighborhood of $\mathbf{c}$ along the silhouette, ± 5 pixels. For each position, we cut a segment from the silhouette with the same length as the edgelet and compute its shape affinity to the edgelet by Equ.2. The position with the highest affinity is taken as the aligned position, and the corresponding segment of the silhouette is marked as "supported", see Fig.7(c). The ratio between the length of the supported segments and the overall length of the silhouette is called the *edge coverage* of the silhouette.

The above algorithm guarantees that one edgelet contributes to at most one hypothesis. If one silhouette can not get enough supporting edgelets, the corresponding hypothesis will be removed. This solves the one-object-multiple-hypotheses problem in a natural way and prune some false alarms. For example, the hypothesis 8 in Fig.7 is removed in this way.

3.4. Matching detection responses with visible parts

For each hypothesis, we remove the body parts whose visibility scores are smaller than a threshold θ_v (=0.7 in our experiments). The remaining parts are considered observable to the detectors. Matching part detection responses with the visible part hypotheses is a standard assignment problem, which we solve by the Hungarian algorithm [22]. For each response-hypothesis pair we compute their overlap ratio. If a pair's overlap ratio is larger than a threshold θ_o (=0.5 in our experiments), it is considered to be a potential match. After matching, we apply under-merging to the remaining part responses to remove redundant false alarms. Then we count the successful detections, false alarms, and missed detections, see Fig.7(d) for an example.

3.5. Computing joint image likelihood

Denote one visible part of an object hypothesis and one part detection response by $\mathbf{z}$ and $\mathbf{r}$ respectively. Denote the set of matched response-hypothesis pairs by SD (successful detection), the sets of false alarms and missed detections are defined by $FA = \{\mathbf{r}|\mathbf{r} \notin SD\}$ and $FN = \{\mathbf{z}|\mathbf{z} \notin SD\}$ (false negative) respectively. Denote the object edgelets from the response $\mathbf{r}$ by $E(\mathbf{r})$. The joint image likelihood of multiple objects is defined by

$$\begin{aligned} P(O|Z) = & \prod_{\{\mathbf{z},\mathbf{r}\} \in SD} P_{SD}(\mathbf{r}, E(\mathbf{r})|\mathbf{z}) \\ & \prod_{\mathbf{r} \in FA} P_{FA}(\mathbf{r}) \prod_{\mathbf{z} \in FN} P_{FN}(\mathbf{z}) \end{aligned} \quad (3)$$

where O packs all observations, and Z for all hypotheses. The first term in the right side of Equ.3 is the reward for successful detections. It is decomposed as

$$P_{SD}(\mathbf{r}, E(\mathbf{r})|\mathbf{z}) = P(\mathbf{r}|E(\mathbf{r}), \mathbf{z}) P(E(\mathbf{r})|\mathbf{z}) \quad (4)$$

To model $P(\mathbf{r}|E(\mathbf{r}), \mathbf{z})$, we evaluate the distribution of the part detector's true positive rate under different edge coverage of the silhouette. The distribution is represented as a histogram. Spatial error between the response and the hypothesis or poor contract reduces the edge coverage score. Lower edge coverage usually corresponds to lower true positive rate. We assumes that $P(E(\mathbf{r})|\mathbf{z})$ is an uniform distribution, hence it is ignored in practice. The second term of the right side of Equ.3 is the penalty for false alarms. It is computed by one minus the detector's precision. The third is the penalty for missed detection. It is computed by one minus the detection rate. These properties are evaluated for different part detectors independently.

3.6. Searching for the best configuration

To search for the best interpretation of the image, we examine the initial object hypotheses one by one, in the descending order of their y-coordinates, see Fig.9 for an example. If there are several hypotheses for one object, the algorithm will find the one that best aligns with the object edges and part responses. For example, the hypotheses h_1, h_3, h_4, h_5 in Fig.9 correspond to one human. Our algorithm chooses the best one (h_1) and removes the others. If there are inter-object occlusions, the algorithm will ignore the occluded parts. For example, the legs of hypothesis h_{12} are not detected, but this can be explained by occlusion from h_7. Therefore, h_{12} is kept.

4. Experimental Results

We demonstrate our approach on the problem of pedestrian detection. We evaluate our system on two public im-

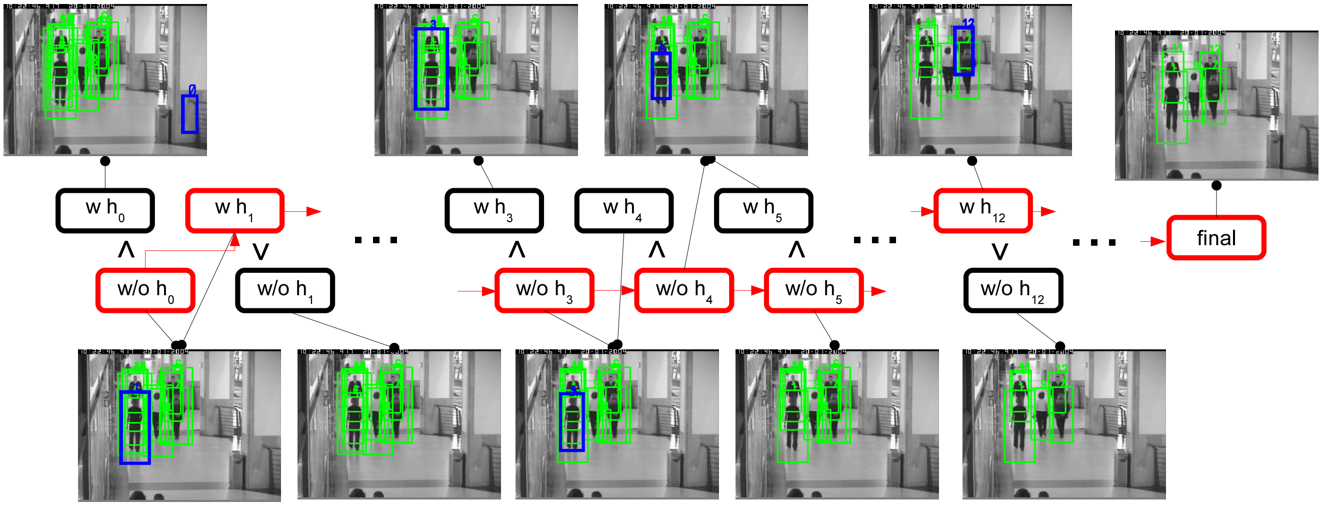


Figure 9. An example of searching for the best multi-object configuration. (The blue rectangles overlaid on the images are the hypotheses being examined. The red boxes are the states kept after comparing the image likelihoods with/without one hypothesis. When examining a hypothesis, one of the "with" and "without" likelihoods can be inherited from the previous round to reduce computational cost. For example "without h_0" and "with h_1" are the same state, as h_0 is removed.)

age sets, the "USC pedestrian set B" [12][1] and the "Zurich Mobile Pedestrian Sequences" [1][2]. Unlike the other popular test sets for pedestrian detection, *e.g*. the INRIA set [13] and the MIT set [20] which use segmented, separated human samples, these two sets contain images with multiple interacting humans. They are very challenging because of the frequent occlusions.

4.1. Training part detector hierarchy

To train the part detectors, we collect about 5,000 pedestrian samples covering different viewpoints, and 7,000 background images without humans from the Internet. The full-body samples are normalized to 24 × 58 pixels. The sizes of the other body parts can be derived based on their definitions in Fig.3. For training, the target false alarm rate of the overall classifier is set to 10^{-6}. The target detection rate and false alarm rate of each cascade layer are set to 99.8% and 0.5 respectively. Although feature sharing cuts training time by about a half, it requires about five days to train all the part detectors.

4.2. Results on the USC test set

The USC pedestrian set B contains 54 images with 271 humans from the CAVIAR corpus[3]. On this set, the performance of our individual part detectors is comparable to that in [12]. We compare the end-to-end performance of our system with some previous methods. Fig.10 shows the precision-recall curves. It can be seen that our method is significantly better than the other state-of-the-art methods. Here we do not use any scene structure or background subtraction to facilitate detection. The test image size is 384 × 288 pixels. We search humans from 24 to 80 pixels wide. We use four threads to run detection of different parts simultaneously. Our experimental machine is a dual-core dual-processor Intel Xeon 3.0GHz CPU. The average speed on this set is about 3.6 second per image. Fig.12(a) shows some example detection results on this set.

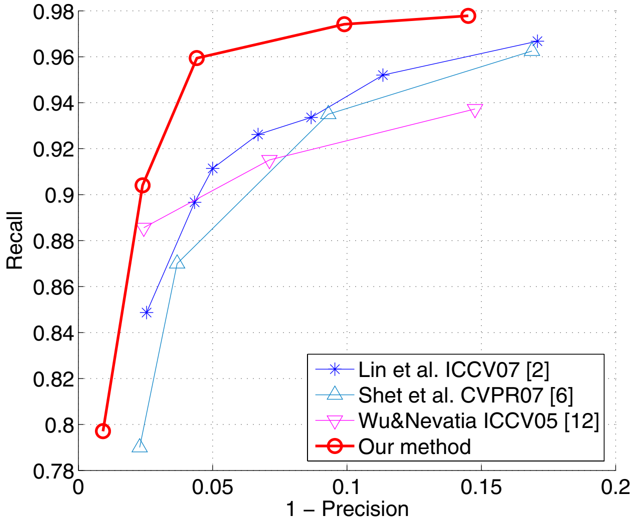


Figure 10. Precision-recall curves on the USC pedestrian test set B. (Our curve is obtained by changing the number of layers used for the part detectors. A detection response is counted as correct, if it overlaps with a ground-truth human by more than 50%.)

[1] http://iris.usc.edu/~bowu/DatasetWebpage/dataset.html

[2] http://www.vision.ee.ethz.ch/~aess/iccv2007/

[3] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

4.3. Results on the Zurich test set

The Zurich set contains three test sequences captured by a stereo pair of cameras mounted on a children's stroller. Same as [1], we only use the frames from the left camera for testing. The first test sequence contains 999 frames with 5,193 annotated humans; the second one contains 450 frames with 2,359 humans; the third one contains 354 frames with 1,828 humans. The frame size is 640×480. To compare with the results in [1], which combines scene analysis with the object detection method in [14], we develop a simple method to estimate the ground plane, which is used to facilitate detection. First we use the full-body detector to search for humans from 58 to 483 pixel high. Then from the full-body responses, we do a RANSAC style algorithm to estimate a linear mapping from the 2-D image position to the 2-D human height: $ax + by + c = h$, where x, y are the image position, h is the human height, and a, b, c are the unknowns. With ground plane, the other part detectors only search the valid regions in the position-scale space. This saves some computational cost and reduces the false alarm rate.

Fig.11 shows the precision-recall curves of our methods and those in [1]. It can be seen that on all the three sequences our method dominates. However, the efforts of this work and that in [1] focus on different aspects. Ess *et al.* [1] try to integrate scene structure analysis and object detection, while our approach tries to segment multiple, occluded objects jointly. These two complementary methods can be combined for further improvement. The average speed of our system on this set is about 2.5 second per image. Fig.12(b) shows some example results.

The performance on the USC set is much better than that on the Zurich set. This is mainly because the background of the Zurich set (outdoor) is much more cluttered than that of the USC set (indoor). At similar detection rate, the false alarm rate is much higher on the Zurich set. During testing the only parameter different on the two sets is the search range of human sizes.

5. Conclusion and Discussion

We described a method to group, merge, and assign part detection responses to segment multiple, possibly inter-occluded objects. Based on occlusion reasoning, joint likelihood of multiple objects is maximized to find the best interpretation of the input image. We demonstrated our approach on the class of pedestrians. The experimental results show that our method outperforms the previous ones.

To apply our approach to other object classes, some components may need to be modified according to the class of interest. First, the design of the part hierarchy is class dependent. Different object classes may need different partitions. Second, the ground plane assumption is valid for some objects in some applications, but not for all situations.

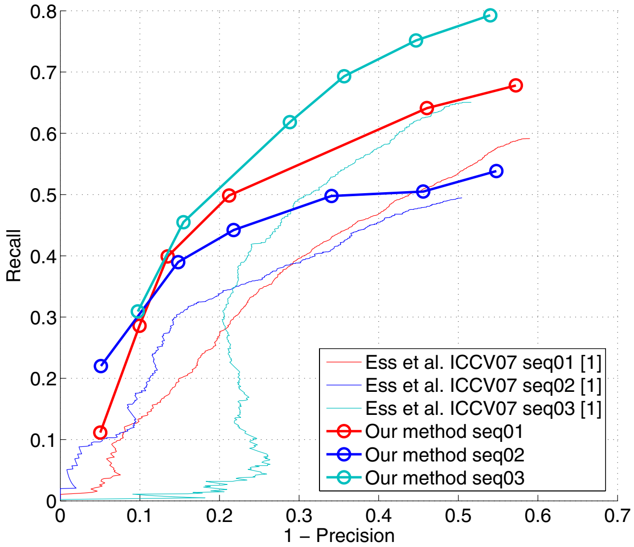


Figure 11. Precision-recall curves on the Zurich mobile pedestrian sequences. (Following [1]'s evaluation, only humans higher than 60 pixels are counted. The curves of [1] are for their full-system, *i.e.* with ground plane and stereo depth.)

When this is not true, we need to infer the objects' relative depths by other techniques. Third, though the feature exclusiveness idea should be helpful for any feature based detection, it may require different implementations for different features.

Acknowledgements: The authors would like to thank Mr. Ess and Dr. Leibe for kindly providing their test data and result curves for comparison. This research was funded, in part, by the U.S. Government VACE program.

References

[1] A. Ess, B. Leibe, and L. V. Gool. Depth and Appearance for Mobile Scene Analysis. ICCV 2007. 6, 7

[2] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon. Hierarchical Part-Template Matching for Human Detection and Segmentation. ICCV 2007 2, 4

[3] B. Wu, and R. Nevatia. Cluster Boosted Tree Classifier for Multi-View, Multi-Pose Object Detection. ICCV 2007. 1, 3

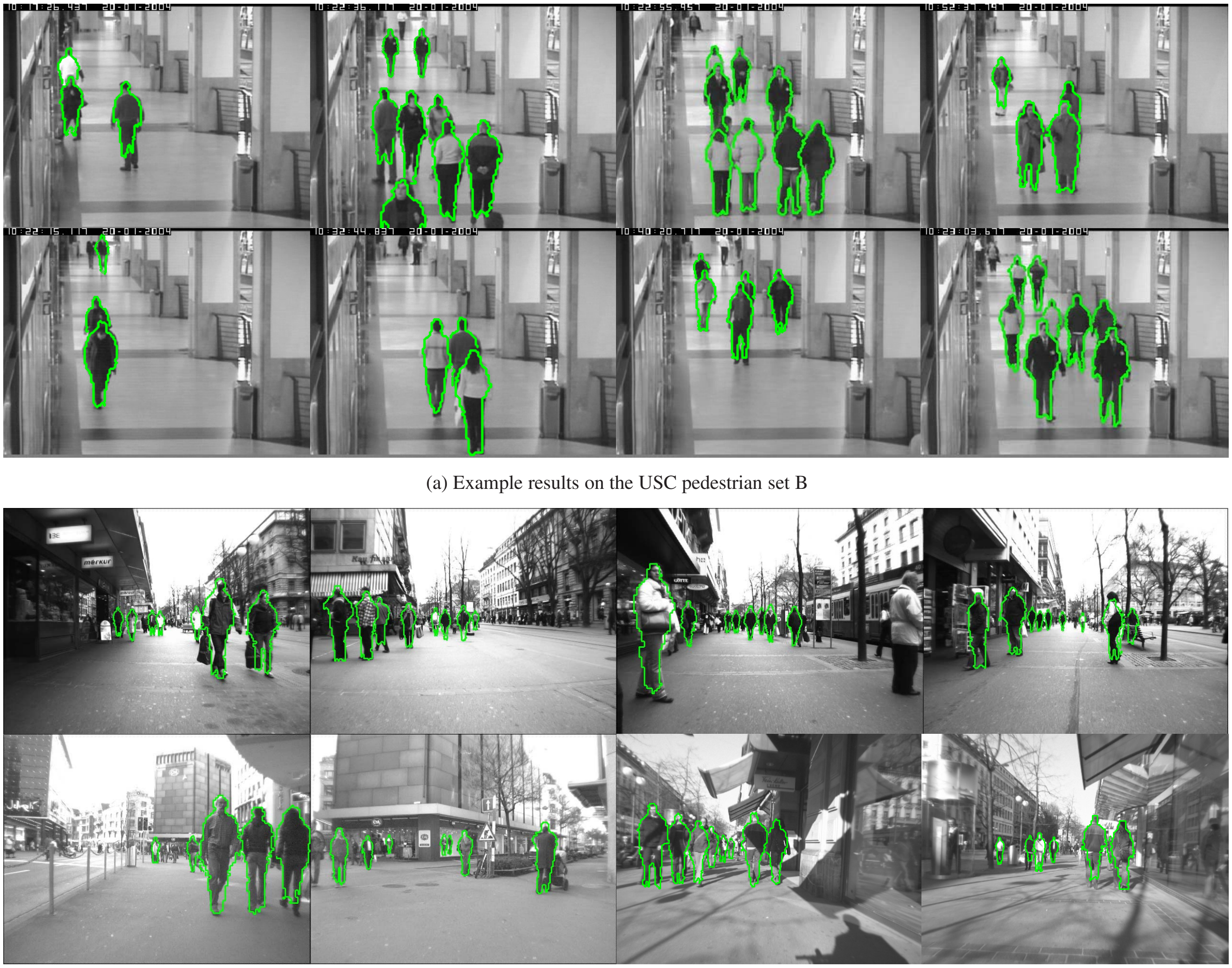
[4] V. Sharma, and J. W. Davis. Integrating Appearance and Motion Cues for Simultaneous Detection and Segmentation of Pedestrians. ICCV 2007. 2

[5] B. Wu, and R. Nevatia. Simultaneous Object Detection and Segmentation by Boosting Local Shape Feature based Classifier. CVPR 2007. 3

[6] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis. Bilattice-based Logical Reasoning for Human Detection. CVPR 2007. 2, 4

[7] O. Tuzel, F. Porikli, and P. Meer. Human Detection via Classification on Riemannian Manifolds. CVPR 2007. 1, 2

[8] P. Sabzmeydani and G. Mori. Detecting Pedestrians by Learning Shapelet Features. CVPR 2007. 1



(a) Example results on the USC pedestrian set B

(b) Example results on the Zurich mobile pedestrian sequences

Figure 12. Example detection and segmentation results.

[9] C. Huang, H. Ai, Y. Li, and S. Lao. High Performance Rotation Invariant Multi-View Face Detection. PAMI, 29(4): 671-686, 2007. 1

[10] D. M. Gavrila. A Bayesian, Exemplar-based Approach to Hierarchical Shape Matching. PAMI, 29(8): 1408-1421, 2007. 1

[11] Q. Zhu, S. Avidan, M.-C. Yeh, and K.-T. Cheng. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. CVPR 2006. 1

[12] B. Wu, and R. Nevatia. Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors. ICCV 2005. 1, 2, 3, 4, 6

[13] N. Dalal, and B. Triggs. Histograms of Oriented Gradients for Human Detection. CVPR 2005. 6

[14] B. Leibe, E. Seemann, and B. Schiele. Pedestrian Detection in Crowded Scenes. CVPR 2005. 2, 7

[15] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Fast Object Detection with Occlusion. ECCV 2004. 2

[16] C. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. ECCV 2004. 1

[17] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. ICCV 2003. 1

[18] P. Viola, and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. CVPR 2001. 1

[19] H. Schneiderman, T. Kanade. A Statistical Method for 3D Object Detection Applied to Faces and Cars. CVPR 2000. 1

[20] C. Papageorgiou, T. Evgeniou, and T. Poggio. A Trainable Pedestrian Detection System. In Proceedings of Intelligent Vehicles, 1998. 6

[21] H. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. PAMI, 20(1): 23-38, 1998. 1, 4

[22] H. W. Kuhn. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2:83-87, 1955. 5