Simultaneous Detection and Segmentation of Pedestrians using Top-down and Bottom-up Processing *

Vinay Sharma James W. Davis
Dept. of Computer Science and Engineering
Ohio State University
Columbus OH 43210 USA
{sharmav,jwdavis}@cse.ohio-state.edu

Abstract

We present a method for the simultaneous detection and segmentation of people from static images. The proposed technique requires no manual segmentation during training, and exploits top-down and bottom-up processing within a single framework for both object localization and 2D shape estimation. First, the coarse shape of the object is learned from a simple training phase utilizing low-level edge features. Motivated by the observation that most object categories have regular shapes and closed boundaries, relations between these features are then exploited to derive mid-level cues, such as continuity and closure. A novel Markov random field defined on the edge features is presented that integrates the coarse shape information with our expectation that objects are likely to have boundaries that are regular and closed. The algorithm is evaluated on pedestrian datasets of varying difficulty, including a wide range of camera viewpoints, and person orientations. Quantitative results are presented for person detection and *segmentation, demonstrating the effectiveness of the proposed technique to simultaneously address both these tasks.*

1. Introduction

The ability of the human visual system to partition the world into distinct objects plays a crucial role in enabling us to successfully reason and interact with our environment. Not surprisingly, the tasks of object detection and segmentation have long been the subject of active research in the field of computer vision.

An object category of specific interest to many application consists of people in outdoor scenes. State-of-the-art pedestrian detection algorithms typically provide information regarding the location and scale of the detected person. However these approaches do not provide any information regarding the shape of the detected object, a cue critical for reasoning about most objects at a higher level. Especially when considering articulate entities such as people, the shape of the silhouette holds rich information that has proven to be useful in tasks such as action recognition, gait analysis, 3D pose reconstruction, etc.

Obtaining the shape or silhouette of an object is typically viewed as a segmentation problem and several object based segmentation techniques have been previously proposed. However these techniques extract object shapes assuming that the object has already been detected in the image, or, in other words, that the image region already contains a single instance of the object.

In this paper we propose a method for the simultaneous detection and segmentation of pedestrians from static images. The algorithm is capable of providing not only the location, but also the silhouette shape of the detected pedestrians using only weakly labeled training data. The required data is weakly labeled in the sense that, for each training image, the presence or absence of the person is labeled, but the person silhouette is not marked. Unlike other related approaches [7, 8, 6], our algorithm thus does not require any form of manual segmentation during training.

The tasks of object detection and shape recovery have been traditionally treated as distinct processes in computer vision. Object detection approaches mostly work in a top-down manner, while image segmentation methods predominantly use a bottom-up approach relying on lower-level cues such as edge continuity and grouping. The proposed algorithm aims to harmoniously integrate both top down and bottom up processing in order to simultaneously recover the location and shape of the target object. The role of low-level local and global perceptual cues in organizing visual information into percepts such as figure and ground has been argued since the Gestalt psychologists. Based on the simple observation that most interesting object categories (e.g.,

*Appears in *IEEE Workshop on Visual Surveillance*, Minneapolis, MN, June 22, 2007

people, vehicles) have regular shapes bounded by smooth boundaries, we aim to seamlessly integrate these powerful perceptual cues with top-down learning.

We first extract low-level contour-based features from the input image (or image patch) and obtain a coarse estimate of the person shape from weakly labeled data. Then, the relations between the contour features are exploited to extract mid-level cues such as contour continuity (local) and closure (global). We then employ a novel Markov Random Field (MRF), defined over the contour features, to integrate the priors expressed by the local and global mid-level cues with the likelihoods obtained from probabilistic learning.

Though the proposed method is applicable to other object categories, we restrict ourselves here to the task of detecting and segmenting people in outdoor scenes. We demonstrate our algorithm on three different datasets encompassing a wide range of different backgrounds, person orientations, and camera viewpoints. We present a detailed quantitative evaluation of the detection and segmentation performance of our algorithm using the well established MIT and INRIA pedestrian datasets.

2. Related Work

While numerous detection methods have been shown to be effective at detecting people in outdoor scenes [9, 3], here we briefly review only those methods that also provide the shape of the object. A template-based method for pedestrian detection was proposed in [5] that provided some shape information by matching the detected object with the most similar template used during training. Template-matching techniques, though, require large training sets with completely segmented object regions. The implicit shape model proposed in [7] addresses the tasks of detection and segmentation using prototypical image patches and their spatial distribution around the object centroid. However, this class of techniques also requires fully segmented object regions during training. In a related approach [8], discriminative boundary fragments (instead of image patches) were used to learn the object geometry. Another approach using object boundaries was proposed in [12]. In [6], an approach using a MRF defined over different object parts was proposed for detection and segmentation. These methods employ complex training schemes and require some form of manual segmentation during training (object centroids in [8], complete shape in [12, 6]). Other approaches that provide object shape can be categorized as strictly object segmentation methods. For instance, work such as [1] focus only on object segmentation, and discuss neither object detection, nor how the method works when the object is not present in the image. The approach described in this paper is most similar to that presented in [10]. While the algorithm in [10] primarily focuses on silhouette extraction, in this work we address both detection and segmentation, and present extensive evaluation of our algorithm utilizing standard person datasets.

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

Figure 1. Different processing stages. (a) Extracted contour features (weighted by E_{mag}). (b) Candidate contour features. (c) Final selected person contours. (d) Silhouette formed from selected contours.

3. Contour Features

We begin by describing the features that we extract from a given input image. We require features that capture the rough shape of the object and also enable us to extract perceptual cues that enforce the constraint that detected objects must have regular, closed boundaries. We exploit first-order gradient information by extracting edge-based features that capture the orientation, location, and magnitude of local image gradients, providing a simple representation to describe object shape. As we will show (Sect. 5), these features also enable us to derive useful perceptual cues.

We first find the edges in an image using the Canny edge detector. From these edges, we extract short, nearly linear, contour fragments composed of connected pixels with similar edge orientation. In order to ensure that the extracted contours are of reasonable size, the possible edge orientations are quantized into a smaller number of bins. We represent each contour fragment, c, by a feature vector $f = [p_1, p_2, E_{mag}]$, where p_1 and p_2 are the two contour endpoints and E_{mag} denotes the mean gradient magnitude along the contour. The feature vector f provides a compact representation of the location, orientation, extent, and magnitude of local edge information. The set of contour features $F = \{f_1, \ldots f_n\}$, extracted from along the edges of an object, forms a succinct representation that implicitly captures the shape of the object.

We show examples of extracted contour features overlaid on the input image (weighted by E_{mag}) in Fig. 1(a) and Fig. 2(b). Given all the contour features extracted from an image region our goal is to identify those contour features that belong to the object. We make use of both top-down and bottom-up processing to achieve this goal.

4. Top-Down Processing

Object familiarity is a top-down cue that is often used in the process of object detection. We exploit this cue by employing a simple training scheme to estimate the rough

shape of the target object given only weakly labeled data. The training data consists of cropped images divided into two sets, a positive set containing instances of the target object, and a larger negative set not containing the object. No manual annotation in the form of segmented foreground (object) pixels is required.

We first extract features, f, as described in Sect. 3, from each cropped image in the training set.These features populate a 5D space, where the dimensions represent the x and y coordinates of contour end-points (p_1 and p_2) and the edge magnitude (E_{mag}). In this 5D space, we create probability density functions (pdf) for the positive and negative features using normalized histograms. Other density estimation techniques could also be employed.

The modes of the positive pdf correspond to contour features characteristic of the target object class as seen in the training set. Given a new feature, the positive and negative pdfs are used to provide a likelihood measure of the feature belonging to the object or the background. Thus, using a simple likelihood ratio test, each contour feature extracted from a new image can be categorized as potentially belonging to the object or the background. Due to the simplicity of the learnining scheme, and the use of only weakly labeled data, the contour features determined to belong to the object at this stage are often not completely accurate. The selected contours can potentially correspond to edges belonging to background structure, leaving large portions of the object boundary untouched. We show in Fig. 1(b) the candidate contour features selected as belonging to the person for the example shown in Fig. 1(a).

5. Bottom-Up Processing

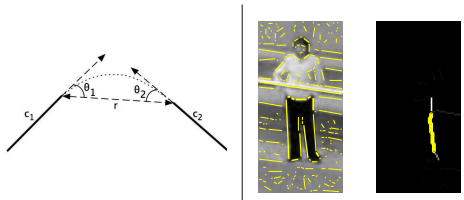
We use the high level, top-down information in conjunction with lower level, bottom-up processing. We attempt to exploit local and global perceptual cues such as continuity and closure to capture our expectation that objects have a natural structure with finite shapes and regular boundaries.

5.1. Contour Affinity

In order to capture the fact that objects have smooth boundaries, we make use of contour "affinity". Used in several computational figure completion methods, affinity measures how likely it is that two edge elements belong to the same underlying edge structure. We adapt the notion of affinity to deal with contours of finite size instead of the dimension-less edge elements used in the literature [11].

Given two contours c_1 and c_2, consider the simplest curve connecting an end-point of c_1 to an end-point of c_2 as shown in Fig. 2(a). Based on [11], we define the affinity for this curve joining c_1 and c_2 as

$$\mathcal{A} = e^{(-r/\sigma_r)} \cdot e^{(-\beta/\sigma_t)} \cdot e^{(-\Delta/\sigma_e)} \qquad (1)$$



(a) (b) (c)

Figure 2. Contour affinity (a) Affinity computation. (b) Extracted contour features (weighted by edge magnitude). (c). Relative affinity of marked (in circles) contour feature with other contours.

where r is distance between the end-points, and $\Delta = |E^{c_1}_{mag} - E^{c_2}_{mag}|$. The term $\beta = \theta_1^2 + \theta_2^2 - \theta_1 \cdot \theta_2$, where θ_1, as shown in Fig. 2(a), denotes the angle between the tangent vector at the end-point of c_1 and the line joining the end-points of c_1 and c_2. The angle θ_2, formed at the end-point of c_2, is analogous to θ_1. The normalization factors σ_r, σ_t, and σ_e are written as $\sigma_r = R/w_1$, $\sigma_t = T/w_2$, and $\sigma_e = E/w_3$, where R, T, and E equal the maximum possible value of r, β, and Δ, and (w_1, w_2, w_3) are weights that can be used to change the relative influence of each term in the affinity calculation. Since c_1 and c_2 have two end-points each, there are four curves connecting the contours depending on which pair of end-points are connected. We define the contour affinity, *Aff*(c_1, c_2), between contours c_1 and c_2 as the maximum affinity over the four possible curves.

We compute pairwise affinities between all the contour features extracted from an image. Features lying in close proximity along a common edge structure often align well and have similar intensities, and hence obtain high affinity values. In Fig. 2(c) we show the relative affinity values between the marked (in circles) contour and the rest of the contours extracted from the image shown in Fig. 2(b). The figure clearly shows that in spite of the large number of contours in close proximity, the neighboring contours *along* the person boundary have the highest affinity.

5.2. Contour Closure

While contour affinity provides a local measure of continuity, an important *global* cue used in perceptual grouping is closure. Objects of interest generally have a finite extent bounded by a closed boundary. Using affinity as a measure of similarity between contour features, we exploit this cue by requiring that object contours form a closed loop.

In order to compute closure, similar to [4], we treat the contour features as nodes in a weighted directed graph, where the weights on the arcs correspond to the affinity between the nodes. We limit the out-degree of each node so as to create a sparse graph. Considering each node in turn, we compute the mean and standard-deviation of its affinity values with every other node. We then preserve only those arcs that have affinity values with a Mahalanobis distance

greater than a threshold. We use a threshold value 1 standard deviation for all the results reported here. The arcs of the graph are then assigned weights equal to the negative log of the affinity values (high affinity corresponds to low arc weight). This enables us to find the most likely cycle passing through a pair of contours using standard and efficient shortest-path graph algorithms (e.g., Dijkstra's algorithm).

If a cycle C_{ij} exists between a pair of contours, c_i and c_j, it is assigned a score, S, equal to the product of the area of the cycle and the affinity of the arc with the maximum weight (minimum affinity) in the cycle. Thus, large cycles, formed by chains of high affinity contour features, are assigned higher scores.

6. Cue Integration

We now turn our attention to the issue of effectively integrating the top-down and bottom-up processes. While considering this problem, we note that the basis of our approach is the application of contextual constraints to enable interpretation of visual information. The contour features are primitives extracted in the context of image pixels and gradients. The notions of good continuity and closure are primitives extracted in the context of contour features. And finally, the object is detected in the context of these features, together with shape familiarty obtained from training data.

A method for modeling context dependent primitives, and the relations between them, in a consistent manner is provided by Markov random field theory. The MRF model provides a framework to maximize objective functions like the Maximum a posteriori (MAP) probability, where the likelihood models the data characteristics and the prior describes preferences between different hypotheses.

Given a rectangular image region we begin by extracting contour features $F = \{f_1, f_2, \ldots f_n\}$ from the input image, and aim to obtain a segmentation by assigning each feature a label from the set $\mathcal{L} = \{l_o, l_b\}$, corresponding to the "object" or "background" class.

Let B denote a configuration of labels such that $\{f_1 = b_1, f_2 = b_2, \ldots f_n = b_n\}$, where $b_i \in \mathcal{L}$. We formulate the search for the optimal label configuration B as a maximum a posteriori (MAP) problem. If we assume that the likelihood of a configuration of labels can be written as a product of the individual likelihoods, the MAP estimate is equivalent to minimizing the free energy [2]

$$E(B) = -\sum_i \log(p(f_i|b_i)) - \log(p(B)) \qquad (2)$$

The first term corresponds to the likelihood of each contour feature belonging to the positive (object) or the negative (background) class. These likelihoods are learned during the training procedure, and capture the coarse shape of the specific object category. The second term corresponds to the prior probability of a shape, as defined by a given configuration of contour labels. As described in Sect. 5.1 and 5.2, the object classes of interest have regular, smooth shapes, and are bounded by a closed contour. In what follows we describe a MRF used to enforce these mid-level cues, while minimizing Eqn. 2.

6.1. Structure and Neighborhood

We model the prior by employing a novel MRF defined over the set of contour features. In order to establish a neighborhood system for the MRF, we make use of contour affinity (see Eqn. 1) as a distance measure. There are different methods that could potentially be used to determine the neighborhood of an element. Two commonly used alternatives are to either restrict the neighborhood to a fixed number of elements, or to require all neighbors to be within a fixed radius. These methods are effective in the typical case, when the MRF is defined over a regular lattice such as image pixels. Under such conditions it is straightforward to determine the optimal size or radius of a neighborhood.

The proposed MRF is defined over an irregular, non-uniform set of elements, and it is not feasible to employ either method to establish a neighborhood system. We determine the neighborhood of a contour feature in an adaptive manner, after examining its relation to all the other contour features in the field. For each contour feature, c_p, we obtain the affinity value to all other features, and compute the mean and standard deviation of these values. Then, in a manner similar to as described in Sect. 5.2, the features having affinity values greater than t standard deviations from the mean are included in $\mathcal{N}_p$, the neighborhood of c_p.

6.2. Clique Potential

Following the Hammersley-Clifford theorem, we define the probability of a configuration $p(B) \propto \exp(-\sum_k V_k(B))$, where V_k denotes the clique potential defined over cliques k. We employ the generalized Potts model to define pairwise clique potentials

$$V_{(p,q)}(b_p, b_q) = u_{(p,q)}(1 - \delta(b_p - b_q)) \qquad (3)$$

where p and q are neighboring sites in the field, which in our case denote contour features. The quantity $u_{(p,q)}$ can be considered to be the cost of assigning different labels to p and q. In most applications, the MRF is defined over a regular lattice (e.g., pixels) and the neighborhood of a site is formed by its 4- or 8-connected neighbors. In such cases, $u_{(p,q)}$ is often defined as a constant (well potential) giving a homogeneous MRF with isotropic clique potentials.

However, in contrast to most previous applications, the MRF described here is defined on contour features (not

arranged in a regular grid). Further, the MRF is non-homogeneous, in that the clique potential across neighboring sites (contour features) depends on the properties of the sites. Instead of defining radially symmetric clique potentials, we wish to enforce a directional smoothness to the label configuration, such that if a contour feature has a positive (object) label, neighboring contours are assigned the same label only if they exhibit good continuity (high affinity) and closure (belong to a closed chain of contours).

We first identify a set F' of *candidate* contours that are likely to belong to the object class using the thresholded log-likelihood ratio

$$f_i \in F', \text{ if } \ln\left(\frac{p(f_i|l_o)}{p(f_i|l_b)}\right) > T \tag{4}$$

Then, using the technique described in Sect. 5.2, we search for cycles connecting pairs of contours taken from F'. As we find cycles, C_{ij}, connecting contour c_i with other contours c_j in F', we increment a pairwise interaction term, $\mathit{Cyc}(i,k)$, for all contours c_k included in those cycles

$$\mathit{Cyc}(i,k) = \mathit{Cyc}(i,k) + \begin{cases} S(C_{ij}) & c_k \in C_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The value of $\mathit{Cyc}(i,k)$ is normalized by the number of contours in F'. Thus, a high value of $\mathit{Cyc}(i,k)$ suggests that, among cycles computed between contours in F', many high scoring cycles passed through c_i and c_k.

We combine both *Aff* and *Cyc* in order to define the penalty term $u_{(p,q)}$ in Eqn. 3 as

$$u_{(p,q)} = \begin{cases} \mathit{Aff}(p,q) \cdot e^{(-\sigma_c/\mathit{Cyc}(p,q))} & c_p \in F' \\ \mathit{Aff}(p,q) & \text{otherwise} \end{cases} \tag{6}$$

where σ_c is a normalization constant. Thus, the cost of label discontinuity is greater for contour pairs with high affinity values. Furthermore, if a contour is in F', this cost is greater for those pairs that have a high affinity and are likely to belong to a closed contour cycle.

6.3. Energy Minimization

As shown in [2], minimizing the energy function $E(B)$ in Eqn. 2 is equivalent to solving the mincut problem on an appropriately constructed graph. Following [2], the graph is composed of two types of vertices, the c-vertices (contour features) and the l-vertices (labels, l_o and l_p). Among the c-vertices, if q is in the neighborhood of p, then p and q are connected by an arc with weight $w_{(p,q)} = 2u_{(p,q)}$. Each c-vertex also has an incoming directed arc from l_o (source) and an outgoing directed arc to l_b (sink) with a weight

$$w_p^l = (\ln(P(f_p|l)) + K) + \sum_{q \in \mathcal{N}_p} w_{(p,q)} \tag{7}$$

where $l \in \mathcal{L}$ and K is a constant ensuring that the weights are positive. The min-cut of this graph ensures that each contour feature is connected to only one of the l-vertices, l_o or l_b, and provides the required contour labeling. In Fig. 1(c) we show the final person contours determined by using the proposed MRF for the image shown in Fig. 1(a).

7. Detection and Silhouette Generation

The labels assigned to the contours in an image region directly provide us the means to classify the input image region as containing the target object or not. Typically, if the target object is not present, all the contour features are assigned the label l_b (corresponding to the background) and the classification of the image region follows trivially. However, depending on the structure in the scene, it is possible that some sporadic background contour features are incorrectly assigned label l_o. Hence, in order to classify an image patch as containing an object or not, we employ a simple measure of the "coherence" of the contours labeled l_o

Using the method described in Sect. 5.2, we search for cycles formed by the positively labeled contours, and score them based on the area enclosed by the cycle, and the minimum affinity arc in the cycle. This process does not impose large computational overhead since the pairwise contour affinities are already computed, and the graph is sparse, consisting only of the positively labeled contours. We compute the average of the best cycle score and the median positive likelihood of the object contours as the measure of coherence. The presence of the object is established by simply comparing this coherence value against a threshold.

We have described here a only very simple (yet effective) form of the coherence measure. Based on domain knowledge and prior expectations about the target object shape, several other factors could be incorporated into the computation of the coherence measure. Thus any falsely labeled background contours can be easily eliminated by relying on higher-level information, such as the shape characteristics of the best contour cycle formed, or the orientation of the axis of symmetry, etc. Traditional object detection schemes, that merely classify image patches without providing a segmentation of the object region, do not provide any such opportunity to reason about falsely detected image patches.

Computing the coherence measure also directly provides the silhouette of the object. The end-points of the contours belonging to the highest scoring cycle are simply joined using straight lines giving a complete, closed outline. This outline is flood-filled to generate the final silhouette. In Fig. 1(d) we show the final silhouette formed using this method from the selected person contours (shown in Fig. 1(c)).

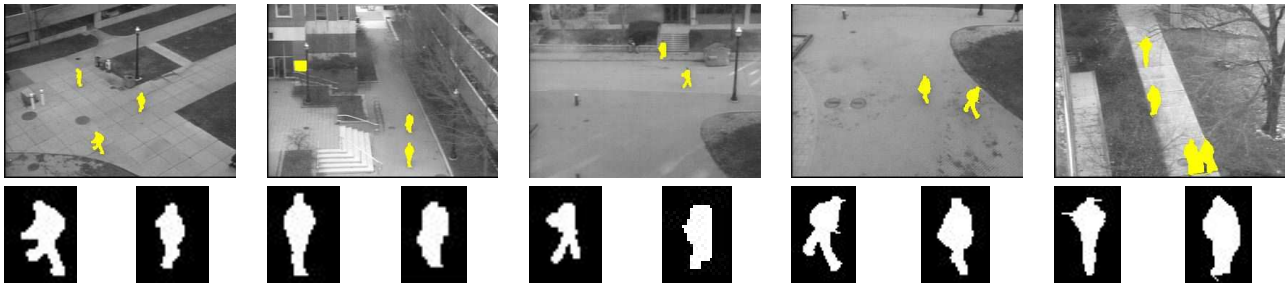


Figure 3. Top row: Example results of person detection and segmentation from surveillance images. Bottom row: Close-up view of silhouettes extracted from top row.

8. Experiments

We applied our algorithm for the detection and segmentation of person silhouettes from images encompassing a wide variety of different scenes and camera viewpoints. Three different datasets were employed to evaluate the efficacy of the proposed method for the tasks of detection and segmentation. The first consists of outdoor images of walkways and intersections on a University campus taken from typical roof-mounted surveillance cameras. The second dataset we used is the MIT pedestrian dataset consisting of front and back views of people in either standing or walking poses. And finally, we also evaluate our algorithm on the challenging INRIA person dataset, that consists of people in city scenes in various poses and orientations.

We use the surveillance dataset to provide a visual, qualitative assessment of the capability of the algorithm to accurately locate and segment people in outdoor scenes. In Fig. 3 we show the output of our algorithm on images from five different camera views. Marked in yellow are the person silhouettes detected by our algorithm. The bottom row of Fig. 3 shows a closer view of ten silhouettes extracted from these images. These results were obtained by training the algorithm on 2034 30×40 cropped person images and 5000 negative images cropped to the same size. No manual marking of object regions or centroids was required. Images were scanned using a three-level image pyramid. From the examples in Fig. 3 we see that the proposed method is effective at recovering both person location and shape without generating many false positives. The camera views shown in Fig. 3 contain several instances of confusing non-pedestrian objects like garbage cans and magazine stands. Further, the views shown second and fifth from the left also contain large image regions with significant amount of clutter due to tree branches and bushes. However, the effective combination of the top-down and bottom-up cues enables the algorithm to be unaffected by these challenging conditions. The second image in the figure shows an instance of a false detection (top left portion of the image).

In order to further evaluate our method quantitatively and to facilitate comparisons with other methods, we make use of the well established MIT and INRIA person datasets. While several approaches have used these datasets for person detection, we aim to employ these datasets to extract information regarding both person location and shape. The MIT pedestrian database [9] consists of 1018 training and 400 test examples of 64×128 cropped images of people in standing and walking poses against urban backgrounds. The INRIA person dataset also contains 64×128 cropped images of people, with 2478 images used for training and 1126 for testing (the number of images reported for both the datasets include left-right reflections). The INRIA dataset contains images of people in a wide variety of upright poses in many different orientation. Contrary to the surveillance dataset discussed earlier, both these datasets consist of images taken at eye-level (camera viewpoint parallel to the ground plane). For negative examples, the INRIA dataset consists of a collection of 1218 training and 453 testing images not containing people.

The respective positive contour likelihoods were learned using the positive training images from the two datasets. In order to obtain the negative contour likelihoods, a common set of 12180 64×128 image patches were sampled from the negative images provided in the INRIA dataset. The negative contour likelihoods computed from these image patches were used both for the MIT and INRIA datastes. A Canny edge detector was used to extract edges from all the training images. In order to remove any end-effects, a 4 pixel boundary was cropped from around the training images, reducing the image size to 56×120. Contours were extracted from the edges using an edge orientation bin size of 45 degrees. Contour likelihoods were obtained using normalized histograms. Different number of bins were experimented with to examine the effect of quantization in both edge-intensity and in the spatial dimensions. A fine quantization was obtained by binning the y, x, and E_{mag} dimensions into 30, 14, and 8 bins respectively (represented as $\{30, 14, 8\}$). Bin numbers of $\{15, 7, 8\}$ and $\{30, 14, 2\}$ were also employed to examine benefits of fine quantization in either the spatial or edge-magnitude dimensions alone.

To test our approach, we processed the positive test images from both datasets, and the 453 negative test images from the INRIA dataset. The negative images were scanned at three different resolutions, moving the detection window across every alternate row with a window stride of 2 pixels.

8.1. Detection

We used a simple threshold on the coherence measure to determine the presence or absence of a person within the input image region. The effect of varying this threshold on detection performance for the MIT and INRIA datasets is shown in Fig. 4(a) and (b) respectively using Detector Error Tradeoff (DET) curves (miss rate vs. False Positives Per Window (FPPW)). Examining the curves for the different bin sizes we see that the setting with the highest number of bins (highest resolution) $\{30, 14, 8\}$ yields better detection results than the others.

Examining Fig. 4(a) we see that, on the MIT dataset, the algorithm generates a miss rate of less than 0.2 at a FPPW rate of 1×10^{-4}. Perfect results (0 miss rate) are obtained when the FPPW rate is around 1.7×10^{-4}. Comparing the curves in Fig. 4(a) and (b) we see that, for identical parameter settings, detection performance on the INRIA dataset is lower than that obtained on the MIT dataset. This is expected since the INRIA dataset consists of people in various poses and orientations against more cluttered scenes . From Fig. 4(b) we see that, on the INRIA dataset, a FPPW rate of 1×10^{-4} corresponds to a miss rate of approximately 0.3. A at a slightly higher FPPW rate of 2×10^{-4}, however, our algorithm generates near perfect results.

In Fig. 4(b) we also show the DET curve obtained by the HOG person detector [3] on the same dataset. Compared to the HOG detector, the sharply dropping DET curves of our algorithm show that the proposed method generates lower miss rates at FPPW rates of 1.3×10^{-4} and higher (less than 8 false-positives per 320×240 image over a three-level pyramid). Our approach also relies on a simpler training procedure than the HOG detector. For the HOG detector, a procedure similar to that used by our algorithm is utilized to first train a preliminary detector. The results produced by this detector are then used to generate a much larger augmented training set. Further, as opposed to the SVM classifier used in the HOG detector, our algorithm employs a single threshold on the coherence measure for the purpose of detection. Additionally, our algorithm not only detects people, but also provides a segmentation of their silhouettes.

8.2. Segmentation

In order to quantify how well the person silhouettes are delineated we compare the segmentation results with manually marked silhouettes. We generated hand-drawn silhouettes for 200 images from the MIT dataset and 600 images from the INRIA dataset. The images were sampled randomly from the respective positive test images of both datasets. We show examples of silhouettes generated from the MIT and INRIA datasets in Fig. 5(a) and (b) respectively. The top row of Fig. 5 shows the input image, the middle row shows the silhouettes produced by the proposed method and the bottom row shows the corresponding ground truth images. We see from Fig. 5 that our method generates reasonable silhouette shapes of people in various poses against different backgrounds.

Number of bins	MIT	INRIA
$\{15, 7, 8\}$	0.84	0.78
$\{30, 14, 2\}$	0.77	0.75
$\{30, 14, 8\}$	0.75	0.72

Table 1. F-measure values for person segmentation results.

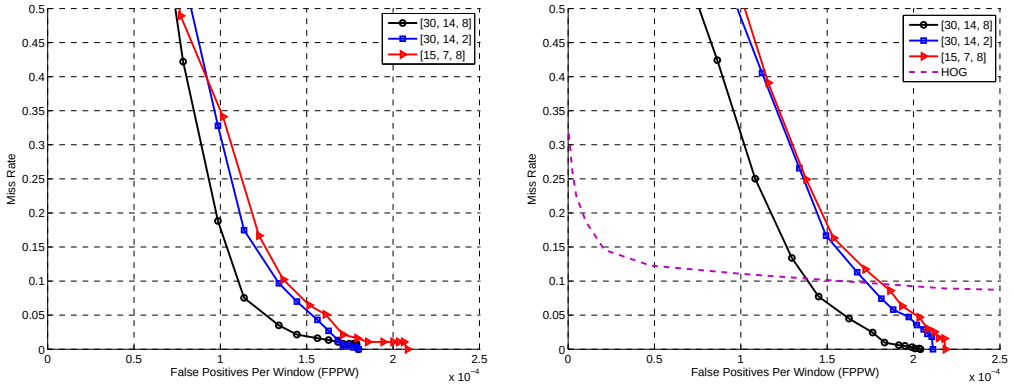
We compute the Precision and Recall of silhouettes generated by our algorithm in terms of the number of pixels overlapping with the manually marked silhouette. The segmentation results for the three quantization settings are summarized in Table 1 using F-measure values (harmonic mean of Precision and Recall). We see from the table that using a bin number setting of $\{15, 7, 8\}$ generates the best segmentation results for both datasets.

8.3. Discussion

The results obtained on the MIT and INRIA datasets demonstrate the ability of our algorithm to utilize comparatively small, weakly labeled datasets to simultaneously generate competitive detection and segmentation results. Comparing our algorithm to the HOG detector (see Fig.4(b)), we see the benefit of employing low-level, perceptual cues. Incorporating bottom-up cues into the detection process, enables our algorithm to successfully label most image regions not containing the object as background. This has the effect of reducing the number of false positives generated, and enables the use of a simple threshold (on the coherence measure) to determine the final detection. While the detection results were similar on both datasets, the segmentation performance was distinctly better on the MIT dataset. The poorer performance on the INRIA dataset can be attributed to the large pose variations included in the dataset, and the presence of several images with weak edge structure.

9. Conclusion

We presented a method for the simultaneous detection and segmentation of people from static images using only weakly labeled data. The method integrates both bottom-up and top-down processing in a coherent manner. We employ low-level contour features and a simple training phase to estimate the rough shape of the target object. Based on the fundamental observation that the person shape has regular,



(a) (b)

Figure 4. DET plots for person detection results (a) MIT dataset. (b) INRIA dataset.

(a) (b)

Figure 5. Examples of segmented silhouette regions from (a) MIT dataset, and (b) INRIA dataset. Top row: Input image. Middle row: Extracted person silhouette. Bottom row: Manually marked silhouettes (ground-truth).

closed boundaries, we then make use of the relationships between the contour features to extract mid-level cues such as contour continuity and closure. An MRF defined over the contour features is then used to integrate these different sources of information to obtain a final contour labeling. We evaluated our method on three different datasets of varying difficulty, encompassing a wide variety of backgrounds and camera viewpoints. Quantitative analysis of detection and segmentation results demonstrate the ability of the approach to effectively perform both tasks simultaneously. In the future, we plan on exploring stronger models of top-down processing in order to further improve detection performance. We also plan on exploring the incorporation of motion information into our framework.

10. Acknowledgements

This research was supported in part by the National Science Foundation under grant No. 0428249.

References

[1] E. Borenstein and J. Malik. Shape guided object segmentation. In *Proc. Comp. Vis. and Patt. Rec.*, volume 1, pages 969–976, 2006.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 23(11):1222–1239, 2001.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. Int. Conf. Comp. Vis.*, pages 886–893, 2005.

[4] J. Elder and S. Zucker. Computing contour closure. In *Proc. Eur. Conf. Comp. Vis.*, pages 399–412, 1996.

[5] D. Gavrila. Pedestrian detection from a moving vehicle. In *Proc. Eur. Conf. Comp. Vis.*, pages 37–49, 2000.

[6] M. Kumar, P. Torr, and A. Zisserman. Objcut. In *Proc. Comp. Vis. and Patt. Rec.*, pages 18–25, 2005.

[7] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Proc. European Conf. Comp. Vis. Workshop*, 2004.

[8] A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment-model for object detection. In *Proc. Eur. Conf. Comp. Vis.*,

2006.

[9] M. Oren et al. Pedestrian detection using wavelet templates. In *Proc. Comp. Vis. and Patt. Rec.*, pages 193–199. IEEE, 1997.

[10] V. Sharma, J. Davis, and A. Tyagi. Extraction of person silhouettes from surveillance imagery using MRFs. In *Proc. Wkshp. Applications of Comp. Vis.*, 2007.

[11] E. Sharon, A. Brandt, and R. Basri. Completion energies and scale. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 22(10):1117–1131, 2000.

[12] J. Shotton, A. Blake, and R. Cipolla. Contour-based learning for object detection. In *Proc. Int. Conf. Comp. Vis.*, 2005.