Hough Regions for Joining Instance Localization and Segmentation*

Hayko Riemenschneider[1,2], Sabine Sternig[1],
Michael Donoser[1], Peter M. Roth[1], and Horst Bischof[1]

[1] Institute for Computer Graphics and Vision, Graz University of Technology, Austria
[2] Computer Vision Laboratory, ETH Zurich, Switzerland

Abstract. Object detection and segmentation are two challenging tasks in computer vision, which are usually considered as independent steps. In this paper, we propose a framework which jointly optimizes for both tasks and implicitly provides detection hypotheses and corresponding segmentations. Our novel approach is attachable to any of the available generalized Hough voting methods. We introduce Hough Regions by formulating the problem of Hough space analysis as Bayesian labeling of a random field. This exploits provided classifier responses, object center votes and low-level cues like color consistency, which are combined into a global energy term. We further propose a greedy approach to solve this energy minimization problem providing a pixel-wise assignment to background or to a specific category instance. This way we bypass the parameter sensitive non-maximum suppression that is required in related methods. The experimental evaluation demonstrates that state-of-the-art detection and segmentation results are achieved and that our method is inherently able to handle overlapping instances and an increased range of articulations, aspect ratios and scales.

1 Introduction

Detecting instances of object categories in cluttered scenes is one of the main challenges in computer vision. Standard recognition systems define this task as localizing category instances in test images up to a bounding box representation. In contrast, in semantic segmentation each pixel is uniquely assigned to one of a set of pre-defined categories, where overlapping instances of the same category are indistinguishable. Obviously, these tasks are quite interrelated, since the segmentation of an instance directly delivers its localization and a bounding box detection significantly eases the segmentation of the instance.

* This work was supported by the Austrian Research Promotion Agency (FFG) under the projects CityFit (815971/14472-GLE/ROD) and MobiTrick (8258408) in the FIT-IT program and SHARE (831717) in the IV2Splus program and the Austrian Science Fund (FWF) under the projects MASA (P22299) and Advanced Learning for Tracking and Detection in Medical Workflow Analysis (I535-N23).

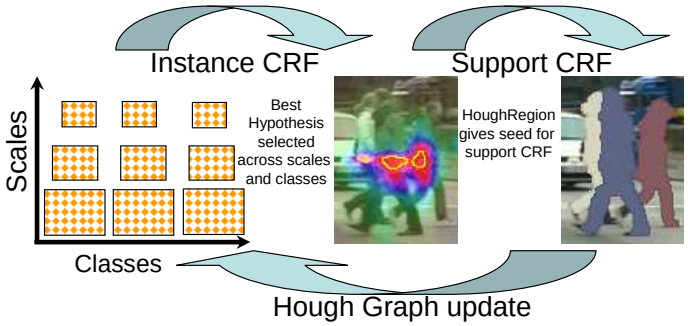

Fig. 1. *Hough regions* (yellow) find reasonable instance hypotheses across scales and classes left) despite strongly smeared vote maps (middle). By jointly optimizing for both segmentation and detection, we identify even strongly overlapping instances (right) compared to standard bounding box non-maximum suppression.

Both tasks have recently enjoyed vast improvements and we are now able to accurately recognize, localize and segment objects in separate stages using complex processing pipelines. Problems still arise where recognition is confused by background clutter or occlusions, localization is confused by proximate features of other classes, and segmentation is confused by ambiguous assignments to foreground or background.

In this work we propose an object detection method which combines instance localization with its segmentation in an implicit manner for jointly finding optimal solutions. Related work in this field, as will be discussed in detail in the next section, either tries to combine detection and segmentation in separate subsequent stages [1] or aims at full scene understanding by jointly estimating a complete scene segmentation of an image for all object categories [2].

We place our method in between these two approaches, since we combine instance localization and segmentation in a joint framework. We localize object instances by extracting so-called *Hough regions* which exploit available information from a generalized Hough voting method, as it is illustrated in Figure 1. *Hough regions* are an alternative to the de-facto standard of bounding box non-maximum suppression or vote clustering, as they are directly optimized in the Hough vote space. Since we are considering maxima regions instead of single pixel maxima, articulation and scale become far less important. Our method is thus more tolerant against diverging scales and the number of overall scales in testing is reduced. Further, we implicitly provide instance segmentations and to increase the recall in scenes containing heavily overlapping objects.

2 Related Work

In this work we consider the problem of localizing and accurately segmenting category instances. Thus, in this section we mainly discuss the related work in

the three most related research fields: non-maximum suppression, segmentation and scene understanding.

One of the most prominent approaches for improving object detection performance is non-maximum suppression (NMS) and its variants. NMS aims for suppressing all hypotheses (i.e. bounding boxes) within a certain distance (e.g. the widespread 50% PASCAL criterion) and localization certainty with respect to each other. Barinova et al. [3] view the Hough voting step as an iterative procedure, where each bounding box of an object instance is greedily considered. Desai and Ramaman [4] see the bounding box suppression as a problem of context evaluation. In their work they learn pairwise context features, which determine the acceptable bounding box overlap per category. For example, a couch may overlap with a person, yet not with an airplane.

The second approach for improving object detection is to use the support of segmentations. The work of Leibe et al. [5] introduced an implicit shape model which captures the structure of an object in a generalized Hough voting manner. They additionally provide segmentations per detected category instance, but require ground truth segmentations for every positive training sample. To recover from overlapping detections, they introduce a minimum description length (MDL) criterion to combine detection hypotheses based on their costs as separate or grouped hypotheses. Borenstein and Ullman [6] generate class-specific segmentations by a combination of object patches, yet this approach is decoupled from the recognition process. Yu and Shi [7] show a parallel segmentation and recognition system in a graph theoretic framework, but are limited to a set of a priori known objects. Amit et al. [8] are treating parts as competing interpretations of the same object instances. Larlus and Jurie [9] showed how to combine appearance models in a Markov Random Field (MRF) setup for category level object segmentation. They used detection results to perform segmentation in the areas of hypothesized object locations. Such an approach implicitly assumes that the final detection bounding box contains the object-of-interest and cannot recover from examples not sticking to this assumption, which is also the case for methods in full scene understanding. Gu et al. [10] use regions as underlying reasoning for object detection, however they rely on a single over-segmentation of the image, which cannot recover from initial segmentation errors. Tu et al. [11] propose the unification of segmentation, detection and recognition in a Bayesian inference framework where bottom-up grouping and top-down recognition are combined for text and faces.

The third approach for improving object detection is to strive for a full scene understanding to explain every object instance and all segmentations in an image. Gould et al. [12] jointly estimate detection and segmentation in a unified optimization framework, however with an approximation of the inference step, since their cost formulation is intractable otherwise. Such an approach cannot find the global optimal solution. Wojek and Schiele [13] couple scene and detector information, but due to the inherent complexity the problem is not solvable in an exact manner. Winn and Shotton [14] propose a layout consistent Conditional Random Field (CRF) which splits the object into parts and enforces a coherent

layout when assigning the labels and connects each Hough transform with a part to extract multiple objects. Yang et al. [15] propose a layered object model for image segmentation, which defines shape masks and explains the appearance, depth ordering, and labels of all pixels in an image. Ladicky et al. [2] combine semantic segmentation and object detection into a framework for full scene understanding. Their method trains various classifiers for *"stuff"* and *"things"* and incorporates them into a coherent cooperating scene explanation. So far, only Ladicky et al. managed to incorporate information about object instances, their location and spatial extent as important cues for a complete scene understanding, which allows to answer a question like what, where and how many object instances can be found in an image. However, their system is designed for difficult full scene understanding and not for an object detection task, as they only integrate detector responses and infer from the corresponding bounding boxes. This limits the ability to increase the detection recall or to improve the accuracy of instance localization.

Our method improves the accuracy of object detectors by using the object's supporting features for joint segmentation and reasoning between object instances. We achieve a performance increase in recall and precision, through the ability to better handle articulations and diverging scales. Additionally, we do not require a parameter sensitive non-maximum suppression since our method delivers unique segmentations per object instance. Please note, in contrast to related methods [5, 16] these segmentation masks are provided without learning from ground truth segmentation masks for each category.

3 Joint Reasoning of Instances and Support

The main goal of this work is the joint reasoning about object instances and their segmentations. Our starting point is any generalized Hough voting method like the Implicit Shape Model (ISM) [5], Hough Forests [17] or the max-margin Hough transform [18]. We formulate our problem in terms of a Bayesian labeling of a first-order Conditional Random Field (CRF) aiming at the minimization of a global energy term. Since global inference in the required full random field is intractable, we propose a greedy method which couples two stages iteratively solving each stage in a global optimal manner. Our model inherently links classifier probabilities, corresponding Hough votes and low-level cues such as color consistency and centroid proximity.

We first describe in Section 3.1 the global energy term for providing pixel-wise assignments to category instances, analyzing unary and pairwise potentials. In Section 3.2 we introduce our approach for minimizing the energy term by a greedy approach. Finally, in Section 3.3 we directly compare the properties of our method to related work.

3.1 Probabilistic Global Energy

We assume that we are given any generalized Hough voting method like [17, 5, 18, 19], which provides our N voting elements X_i within the test image and

corresponding object center votes $\mathcal{H}_i$ into the Hough space. We further assume that we are given $p\left(C|X_i, D_i\right)$ per voting element, which measures the likelihood that a voting element X_i belongs to category C by analyzing a local descriptor D_i. This could be feature channel differences between randomly drawn pixel pairs as in [17]. All this information can be directly obtained from the generalized Hough methods; see experimental section for implementation details.

The goal of our method is to use the provided data of the generalized Hough voting method to explain a test image by classifying each pixel as background or as belonging to a specific instance of an object category. We formulate this problem as Bayesian labeling of a first-order Conditional Random Field (CRF) by minimizing a global energy term to derive an optimal labelling.

Such random field formulations are widespread, especially in the related field of semantic segmentation, e.g. [20]. In a standard random field $\mathcal{X}$ each pixel is represented as a random variable $X_i \in \mathcal{X}$, which takes a label from the set $\mathcal{L} = \{l_1, l_2, \ldots, l_K\}$ indicating one of K pre-defined object classes as e.g. grass, people, cars, etc. Additionally, an arbitrary neighborhood relationship $\mathcal{N}$ is provided, which defines cliques c. A clique c is a subset of the random variables $\mathbf{X}_c$, where $\forall i, j \in c$ holds $i \in \mathcal{N}_j$ and $j \in \mathcal{N}_i$, i.e. they are mutually neighboring concerning the defined neighborhood system $\mathcal{N}$.

In our case, we not only want to assign each pixel to an object category, we additionally aim at providing a unique assignment to a specific instance of a category in the image, which is a difficult problem if category instances are highly overlapping. For the sake of simplicity, we define our method for the single class case, but the method is easily extendable to a multi-class setting.

We also represent each pixel as random variable X_i with $i = 1 \ldots N$, where N is the number of pixels in the test image, and aim at assigning each pixel a category-specific *instance* label from the set $\mathcal{L} = \{l_0, l_1, \ldots, l_L\}$, where L is the number of instances. We use label l_0 for assigning a pixel to the background. We seek for the optimal *labeling* $\mathbf{x}$ of the image which is taken from the set $\mathbf{L} = \mathcal{L}^N$. The optimal labeling minimizes the general energy term $E\left(\mathbf{x}\right)$ defined as

$$E\left(\mathbf{x}\right) = -log\ P\left(\mathbf{x}|\mathbf{D}\right) - log\ Z = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c), \tag{1}$$

where Z is a normalization constant, $\mathbf{D}$ is the observed data, $\mathcal{C}$ is the set of considered cliques defined by the neighborhood relationship $\mathcal{N}$ and $\psi_c(\mathbf{x}_c)$ is a potential function of the defined clique. The Maximum a Posteriori (MAP) labeling $\mathbf{x}^*$ is then found by minimizing this energy term as

$$x^* = \underset{x \in \mathbf{L}}{\operatorname{argmin}}\, E\left(\mathbf{x}\right). \tag{2}$$

To obtain reasonable label assignments it is important to select powerful potentials, which range from simple unary terms (e.g. evaluating class likelihoods), pairwise terms (e.g. the widespread Potts model) or even higher order terms. While unary potentials ensure correct label assignments, pairwise potentials aim at providing a smooth label map, e.g. by avoiding the assignment of identical

labels over high image gradients. The overall goal is to provide a smooth labeling consistent with the underlying data.

Our energy is modeled as the sum of unary and pairwise potentials. In the following, we assume that we are given a set of seed variables $\mathcal{S}$ which consist of L subsets $\mathbf{S_l} \in \mathcal{X}$ of random variables for which we know the assignment to a specific label in the image. We again assume that the set $\mathbf{S_0}$ represents the background. These assignments are an essential step of our method and how to obtain them is discussed in Section 3.2.

Based on the given assignments $\mathbf{S_l}$, we define our instance-labeling problem as minimizing the following energy term

$$E(x) = \sum_{X_i \in \mathcal{S}_l} \Theta\left(X_i\right) + \sum_{X_i \in \mathcal{S}_l} \Upsilon\left(X_i\right) + \sum_{X_i \in \mathcal{S}_l} \Omega\left(X_i\right) + \sum_{X_i, X_j \in \mathcal{N}} \Psi\left(X_i, X_j\right) , \quad (3)$$

which contains a unary, instance-specific class potential Θ, a color based unary potential Υ, a distance based unary potential Ω and a pairwise potential Ψ. The first unary potential Θ contains the estimated likelihoods for a pixel taking a certain category label as provided by the generalized Hough voting method as

$$\Theta = -\log\ p\left(C|X_i, D_i\right), \quad (4)$$

i.e. the unary potentials Θ describe the likelihood of getting assigned to one of the identified seed regions or the background. Since we do not have a background likelihood, the corresponding values for l_0 are set to a constant value p_{back}, which defines the minimum class likelihood we want to have. The term Θ drives our label assignments to correctly distinguish background from actual object hypotheses.

The unary potential Υ analyzes the likelihood for assigning a pixel to one of the defined seed regions, e.g. analyzing local appearance information in comparison to the seed region. In general any kind of modeling scheme is applicable in this step. We model each subset $\mathbf{S_l} \in \mathcal{X}$ by Gaussian Mixture Models (GMM) $\mathcal{G}_l$ and define color potentials for assigning a pixel to each instance by

$$\Upsilon = -\log\ p\left(X_i|\mathcal{G}_l\right). \quad (5)$$

The corresponding likelihoods for the background class l_0 are set to $\log(1 - \max_{l \in \mathcal{L}} p\left(X_i|\mathcal{G}_l\right))$ since the background is mostly too complex to model. The term Υ ensures that pixels are assigned to the right instances by considering the appearance of each instance.

The third unary potential Ω defines a spatial distance function analyzing how close each pixel is to each seed region $\mathbf{S_l}$ by

$$\Omega = \Delta\left(X_i, \mathbf{S_l}\right), \quad (6)$$

where Δ is a distance function. The term Ω ensures that correct assignments to instances are made considering constraints like far away pixels with diverging center votes are not assigned to the same instance.

Finally, our pairwise potentials encourage neighboring pixels in the image to take the same label considering a contrast-sensitive Potts model Ψ analyzing pixels included in the same cliques c defined as

$$\Psi = \begin{cases} 0 & \text{if} \quad x_i = x_j \\ \nabla & \text{if} \quad x_i \neq x_j \end{cases}, \tag{7}$$

where ∇ is an image specific gradient measure or local color difference. The term Ψ mainly ensures that smooth label assignments are achieved.

This way we can effectively incorporate not only the probability of a single pixel belonging to an object, but also consider the spatial extent and the local appearance of connected pixels in our inference. Please note that this formulation can e.g. be extended to superpixels to include higher-order potentials [21].

Finding an optimal solution without knowledge of the seeds $\mathbf{S_l}$ is infeasible since inference in such a dense graph is NP-hard. Therefore, in the next section we propose a greedy algorithm for solving the above presented energy minimization problem, which alternately finds optimal seed assignments and then segments instances analyzing the hough vector support. The final result of our method is a segmentation of the entire image into background and individual category instances.

3.2 Instance Labeling Inference

We propose a novel, greedy inference concept to solve the energy minimization problem defined in Eq. 3. The core idea is to alternately find a single, optimal seed region analyzing the provided Hough space (Eq. 8), and afterwards to use this seed region to find an optimal segmentation of the corresponding instance in the image space (Eq. 3). The obtained segmentation is then used to update the Hough space information (Eq. 10), and we greedily obtain our final image labeling.

We assume that we are given the votes $\mathcal{H}_i$ of each pixel into the Hough vote map, providing a connection between pixelwise feature responses and projected Hough centers. Hence, we can build a two-layer graph for any image, where the nodes in the first layer $\mathbb{I}$ (image graph) are the underlying random variables X_i for all pixels in the image, and the second layer $\mathbb{H}$ (Hough graph) contains their transformed counterparts $\mathcal{H}(X_i)$ in the Hough vote map. Figure 2 illustrates this two-layer graph setup.

The first step is to optimally extract a seed-region from the corresponding Hough graph $\mathbb{H}$. Therefore, we propose a novel paradigm denoted as *Hough regions*, which formulates the seed pixel extraction step itself as a segmentation problem. A *Hough region* is defined as connected, arbitrarily shaped subset of graph nodes $\mathbf{H_l} \in \mathbb{H}$, which are the hypothesized center votes for the object instance into the Hough space. In the ideal case, all pixels of the instance would vote to the same center pixel, unfortunately in the real world these Hough center votes are quite imperfect. Even despite recent research to decrease the Hough vector impurity [17, 19], the Hough center is never a single pixel. This arises from

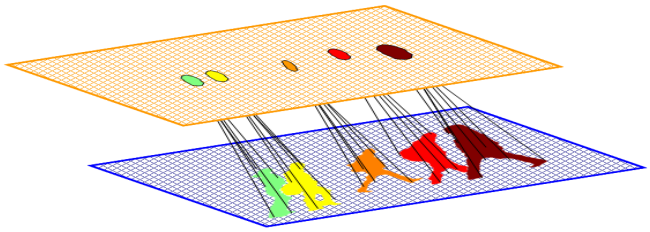

Fig. 2. Two layer graph design (top: Hough and bottom: image) to solve proposed image labeling problem. We identify *Hough regions* as subgraphs of the Hough graph $\mathbb{H}$. The backprojection into the image space, in combination with color information, distance transform values and contrast sensitive Potts potentials defines instance segmentations in the image graph $\mathbb{I}$.

various sources of error such as changes in global and local scales, aspect ratios, articulation, etc. Perfect Hough maxima are unlikely and there will always exist inconsistent centroid votes. The idea and benefit of our *Hough regions* paradigm is now to exactly capture this uncertainty by considering regions, and not single Hough maxima.

Thus, our goal is to identify *Hough regions* $\mathbf{H_l}$ in the Hough space $\mathbb{H}$, which then allows to directly define the corresponding seed region $\mathbf{S_l}$ by the back-projections into the image space, i.e. the random variables $X_i \in \mathbf{S_l}$ are the nodes in the image graph $\mathbb{I}$ which project into the *Hough region* $\mathbf{H_l}$. For this reason, we define a binary image labeling problem in the Hough graph by

$$E(x) = \sum_{X_i \in \mathcal{X}} \Theta(\mathcal{H}(X_i)) + \sum_{\mathcal{N}} \Phi(\mathcal{H}(X_i), \mathcal{H}(X_j)), \tag{8}$$

which contains the projected class-specific unary potentials Θ and a pairwise potential Φ. The potential Θ for each variable X_i is projected to the Hough graph using $\mathcal{H}(X_i)$. The unary potential at any node $\mathcal{H}(X_i)$ is then the sum of the classifier responses voting for this node, as it is common in general Hough voting methods. The pairwise potential Φ is defined on this very same graph $\mathbb{H}$ as a gradient-sensitive Potts model, where the gradient is calculated as the difference between the unary potentials of two nodes $\mathcal{H}(X_i)$ and $\mathcal{H}(X_j)$ in the Hough graph $\mathbb{H}$. This binary labeling problem can be solved in a global optimal manner using any available inference algorithm as e.g. graph cuts [22], and the obtained Hough region $\mathbf{H_l}$ is then back-projected to the seed region $\mathbf{S_l}$.

The second step, after finding the optimal seed region $\mathbf{S_l}$, is to identify all supporting pixels of the category instance in the image graph $\mathbb{I}$. Since we have now given the required seed region $\mathbf{S_l}$, we can apply our energy minimization problem defined in Eq. 3 and again solve a binary labeling problem, for assigning each pixel to the background or the currently analyzed category instance. The last part is the distance function $\Delta(X_i, \mathbf{S_l})$, which we define as

$$\Delta(X_i, \mathbf{S_l}) = \begin{cases} 0 & \text{if} \quad \mathcal{H}(X_i) \in \mathbf{H_l} \\ DT(X_i) & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathcal{H}(X_i)$ is the Hough transformation of an image location to its associated Hough nodes and $DT(X_i)$ is the distance transform over the elements of the current seed pixels, which are given by the *Hough region* $\mathbf{H_l}$. Again this binary labeling problem can be solved in a global optimal manner using e.g. graph cut methods [22], which returns a binary segmentation mask $\mathcal{M}_l$ for the current category instance.

After finding the optimal segmentation $\mathcal{M}_l$ for the currently analyzed category instance in the image space, we update the Hough vote map, considering the already assigned Hough votes. In such a way our framework is not error-prone to spurious incorrect updates in the Hough voting space as it is common in related methods. This directly improves the detection performance, as occluded object instances are not removed. On the contrary, occluded instances are now more easily detectable by their visible segmentation, as they require less visibility with competing object instances or other occlusions.

In detail, the update considers each image location and its (independent) votes, which are accumulated in the nodes of the Hough graph. An efficient update is possibly by subtracting the previously segmented object instance from the classwise, unary potentials, which are initially $\Theta_0(X_i) = \Theta(X_i)$, by

$$\Theta_{t+1}(X_i) = \Theta_t(X_i) - \Theta(X_{\mathcal{M}_l}), \tag{10}$$

where each random variable $X_{\mathcal{M}_l}$ within the segmentation mask $\mathcal{M}_l$ of the category instance is used in the update. Using our obtained segmentations, we focus the update solely on the areas of the graph where the current object instance plays a role. This leads to a much finer dissection of the image and Hough graphs, as shown in Figure 2.

After updating the Hough graph $\mathbb{H}$ with the same update step, we repeat finding the next optimal seed region and afterwards segmenting the corresponding category instance in the image graph $\mathbb{I}$. This guarantees a monotonic decrease in $max(\Theta_{t+1})$ and our iteration stops when $max(\Theta_{t+1}) < p_{back}$, i.e. we have identified all Hough regions (object instances $\mathcal{L}$) above a threshold.

3.3 Comparison to Related Approaches

Our implicit step of updating the Hough vote maps by considering optimal segmentations in the image space, is related to other approaches in the field of non-maximum suppression. In general, one can distinguish two different approaches in this field.

Bounding boxes are frequently selected as underlying representation to perform non-maximum suppression and are the de-facto standard for defining the number of detections from a Hough image. In these methods the Hough space is analyzed using a Parzen-window estimate based on the Hough center votes. Local maxima in the Parzen window estimate determine instance hypotheses by placing a bounding box considering the current scale onto the local maxima. Conflicting bounding boxes, e.g. within a certain distance and quality with respect to each

other are removed (NMS). Integrating such a setup in our method would mean that each bounding box represents one seed region $\mathbf{S_l}$ and additionally defines a binary distance function $\Delta(X_i, \mathbf{S_l})$, where $\Delta = 1$ for all pixels within the bounding box and $\Delta = 0$ for all other pixels.

Non-maximum suppression comes in many different forms like a) finding all local maxima and performing non-maximum suppression in terms of mutual bounding box overlap in image space to discard low scoring detections; b) extracting global maxima iteratively and eliminating these by deleting the interior of the bounding box hypothesis in Hough space [17]; c) iteratively finding global maxima for the seed pixels [3] and estimating bounding boxes and subtracting the corresponding Hough vectors. All pixels inside the bounding box are used to reduce the Hough vote map. This leads to a coarse dissection of the Hough map which leads to problems since it includes much background information and overlapping object instances.

Clustering methods allow more elegant formulation since they attempt to group Hough vote vectors to identify instance hypotheses. The Implicit Shape Model (ISM) [5], for example, employs a mean-shift clustering on sparse Hough center locations to find coherent votes. The work by Ommer and Malik [23] extends this to clustering Hough voting lines, which are infinite lines as opposed to finite Hough vectors. The benefit lies in extrapolating the scale from coherent Hough lines, as the line intersection in the 3D x–y–scale space determines the optimal scale. Such an approach relates to ours, since the clustering step can be interpreted as elegantly identifying the seed pixels we analyze. However such approaches assume well-distinguishable cluster distributions and are not well-suited for scenarios including many overlaps and occlusions. Furthermore, similar to the bounding box methods, which require a Parzen-window estimate to bind connected Hough votes, clustering methods require an intra- to inter cluster balance. In terms of a mean-shift approach this is defined by the bandwidth parameter considered. An additional drawback of clustering methods is also the limited scalability in terms of number of input Hough vectors that can be efficiently handled. An interesting alternative is the grouping of dependent object parts [24], which yields significantly less uncertain group votes compared to the individual votes.

Our proposed approach has several advantages compared to the discussed related methods. A key benefit is that we do not have to fix a range for local neighborhood suppression, as it is e.g. required in non-maximum suppression approaches, due to our integration of segmentation cues. As it is also demonstrated in the experimental section, our method is robust to an increased range of variations in object articulations, aspect ratios and scales, which allows to reduce the number of analyzed scales during testing while maintaining the detection performance. We implicitly also provide segmentations of all detected category instances, without requiring segmentation masks during training. In overall, the benefits of our *Hough Regions* approach lead to higher precision and recall compared to related approaches.

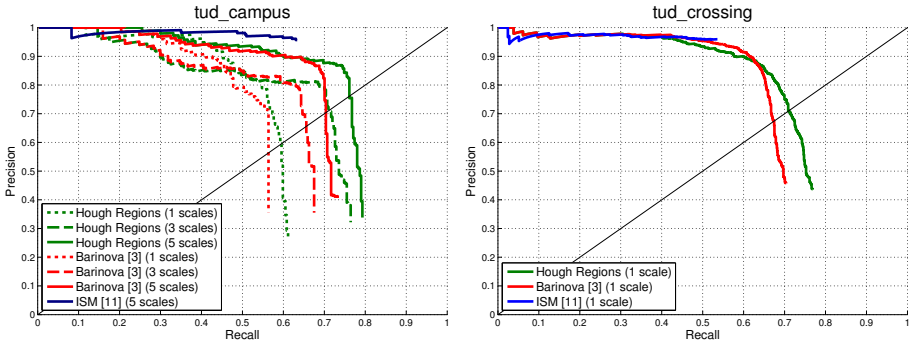

Fig. 3. Recall/Precision curve for the TUD campus and TUD crossing sequences showing analysis at multiple scales for highly overlapping instances. See text for details.

4 Experimental Evaluation

We use the publicly available Hough Forest code [17] to provide the required class likelihoods and the centroid votes. For non-maximum suppression (NMS), we apply the method of [3] (also based on [17]), which outperforms standard NMS in scenes with overlapping instances. It is important to note that for datasets without overlapping instances like PASCAL VOC, the performance is similar to [17]. For this reason, we mainly evaluate on datasets including severely overlapping detections, namely TUD Crossing and TUD Campus. Additionally, we evaluate on a novel window detection dataset GT240, designed for testing sensitivity to aspect ratio and scale variations. We use two GMM components, equal weighting between the energy terms and fix $p_{back} = 0.125$.

4.1 TUD Campus

To demonstrate the ability of our system to deal with occluded and overlapping object instances (which results in an increased recall using the PASCAL 50% criterion), we evaluated on the TUD Campus sequence, which contains 71 images and 303 highly overlapping pedestrians with large scale changes. We evaluated the methods on multiple scale ranges (one, three and five scales) and Figure 3 shows the Recall-Precision curve (RPC) for the TUD Campus sequence. Aside the fact that multiple scales benefit the detection performance for all methods, one can also see that our *Hough regions* method surpasses the performance at each scale range (+3%, +8%, +8% over [3]) and over fewer scales. For example, using *Hough regions* we only require three scales to achieve the recall and precision, which is otherwise only achieved using five scales. This demonstrates that our method handles scale variations in an improved manner, which allows to reduce the number of scales and thus runtime during testing. The ISM approach of Leibe et al. [5] reaches good precision, but at the cost of decreased recall, since the MDL criterion limits the ability to find all partially occluded objects.

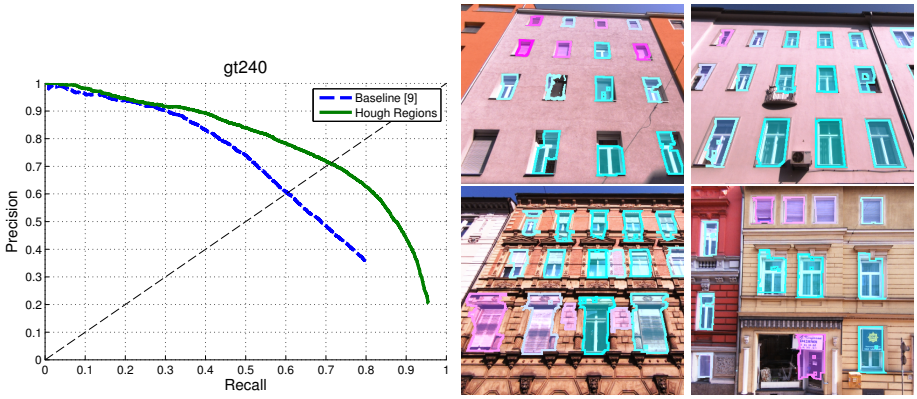

Fig. 4. Recall/Precision curve for the street-level window detection dataset GrazTheresa240 with strong distortion and aspect ratio changes in 5400 images.

4.2 TUD Crossing

The TUD Crossing dataset [25] contains 201 images showing side views of pedestrians in a relatively crowded scenario. The annotation by Andriluka et al. [25] contains 1008 tight bounding boxes designed for pedestrians with at least 50% visibility, ignoring highly overlapping pedestrians. For this reason we created a pixelwise annotation. The new annotation is based on the original bounding box annotation and now contains 1212 bounding box annotations with corresponding segmentations. In addition to bounding boxes we also annotated the visibility of pedestrians in fully visible body parts: head, front part of upper body, back part of upper body, left leg or right leg.

Typically, in this sequence three scales are evaluated, however to show ability to handle scale, we evaluate only on a single scale for all methods. Figure 3 shows the Recall-Precision curve (RPC) for the TUD crossing sequence in comparison to an Implicit Shape Model (ISM) [5] and the probabilistic framework of [3]. Our method achieves a better recall compared to the other approaches. We increase the recall as well as the precision indicating that our method can better handle the overlaps, scale and articulation changes.

4.3 GrazTheresa240

We also evaluated our method on a novel street-level window detection dataset (denoted as GrazTheresa240), which consists of 240 buildings with 5400 redundant images with a total of 5542 window instances. Window detection itself is difficult due to immense intra-class appearance variations. Additionally, the dataset includes a large range of diverging aspect ratios and strong perspective distortions, which usually limit object detectors. In our experiment we tested on three different scales and a single aspect ratio. As shown in Figure 4 we can substantially improve the detection performance in both recall and precision

Fig. 5. Segmentations for TUD Crossing and TUD Campus datasets.

(12% at EER) compared to the baseline [17]. Our *Hough regions* based detector consistently delivers improved localization performance, mainly by reducing the number of false positives and by being less sensitive to diverging aspect ratios and perspective distortions.

4.4 Segmentation

As final experiment we analyze achievable segmentation accuracy on the TUD Crossing sequence [25], where we created binary segmentations masks for each object instance for the entire sequence. Segmentation performance is measured by the ratio between the number of correctly assigned pixels to the the total number of pixels (segmentation accuracy). We compared our method to the Implicit Shape Model (ISM) [5], which also provides segmentations for each detected instance. Please note that the ISM requires segmentation masks for all training samples to provide reasonable segmentations whereas our method learns from training images only. Nevertheless, we achieve competitive segmentation accuracy of 98.59% compared to 97.99% for the ISM. Illustrative results are shown in Figure 5.

5 Conclusion

In this work we proposed *Hough regions* for object detection to jointly solve instance localization and segmentation. We use any generalized hough voting method, providing pixel-wise object centroid votes for a test image, as starting point. Our novel *Hough regions* then determine the locations and segmentations of all category instances and implicitly handles large location uncertainty due to articulation, aspect ratio and scale. As shown in the experiments, our method jointly and accurately delineates the location and outline of object instances, leading to increased recall and precision for overlapping and occluded instances. These results confirm related research where combining the tasks of detection and segmentation improves performance, because of the joint optimization benefit over separate individual solutions. In future work, we plan to consider our method also during training for providing more accurate foreground/background estimations without increasing manual supervision.

References

1. Ramanan, D.: Using segmentation to verify object hypotheses. In: CVPR. (2007)
2. Ladicky, L., Sturgess, P., Alahari, K., Russell, C., Torr, P.: What, Where & How Many? Combining Object Detectors and CRFs. In: ECCV. (2010)
3. Barinova, O., Lempitsky, V., Kohli, P.: On the detection of multiple object instances using hough transforms. PAMI **34** (2012) 1773–1784
4. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. IJCV **95** (2011) 1–12
5. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. IJCV **77** (2008) 259–289
6. Borenstein, E., Ullman, S.: Class-specific, top-down segmentation. In: ECCV. (2002)
7. Yu, S., Shi, J.: Object-specific figure-ground segregation. In: CVPR. (2003)
8. Amit, Y., Geman, D., Fan, X.: A coarse-to-fine strategy for multiclass shape detection. PAMI **26** (2004) 1606–1621
9. Larlus, D., Jurie, F.: Combining appearance models and markov random fields for category level object segmentation. In: CVPR. (2008)
10. Gu, C., Lim, J., Arbelaez, P., Malik, J.: Recognition using regions. In: CVPR. (2009)
11. Tu, Z., Chen, X., Yuille, A., Zhu, S.: Image parsing: Unifying segmentation, detection, and recognition. IJCV **62** (2005) 113–140
12. Gould, S., Gao, T., Koller, D.: Region-based segmentation and object detection. In: NIPS. (2009)
13. Wojek, C., Schiele, B.: A dynamic conditional random field model for joint labeling of object and scene classes. In: ECCV. (2008)
14. Winn, J., Shotton, J.: The layout consistent random field for recognizing and segmenting partially occluded objects. In: CVPR. (2006)
15. Yang, Y., Hallman, S., Ramanan, D., Fowlkes, C.: Layered object models for image segmentation. PAMI **34** (2011) 1731–1743
16. Floros, G., Rematas, K., Leibe, B.: Multi-Class Image Labeling with Top-Down Segmentation and Generalized Robust P^N Potentials. In: BMVC. (2011)
17. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: CVPR. (2009)
18. Maji, S., Malik, J.: Object detection using a max-margin hough transform. In: CVPR. (2009)
19. Okada, R.: Discriminative generalized hough transform for object dectection. In: ICCV. (2009)
20. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: CVPR. (2008)
21. Kohli, P., Ladicky, L., Torr, P.: Robust higher order potentials for enforcing label consistency. IJCV **82** (2009) 302–324
22. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI **23** (2001) 1222–1239
23. Ommer, B., Malik, J.: Multi-scale object detection by clustering lines. In: ICCV. (2009)
24. Yarlagadda, P., Monroy, A., Ommer, B.: Voting by grouping dependent parts. In: ECCV. (2010)
25. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR. (2008)