Analysis of the Discriminative Generalized Hough Transform for Pedestrian Detection

Eric Gabriel[1(✉)], Hauke Schramm[1,2], and Carsten Meyer[1,2]

[1] Institute of Computer Science, Kiel University of Applied Sciences, Kiel, Germany
{eric.gabriel,hauke.schramm,carsten.meyer}@fh-kiel.de
[2] Department of Computer Science, Kiel University (CAU), Kiel, Germany

Abstract. Many approaches have been suggested for automatic pedestrian detection to cope with the large variability regarding size, occlusion, background variability etc., among them deformable part models, feature-based approaches (e.g. histograms of oriented gradients), and recently deep learning-based algorithms. Current deep learning-based frameworks rely either on a proposal generation mechanism (e.g. "Faster R-CNN") or on inspection of image quadrants/octants (e.g. "YOLO"), which are then further analyzed with deep convolutional neural networks (CNN). In this work, we analyze the Discriminative Generalized Hough Transform (DGHT), which operates on edge images, for pedestrian detection. The analysis motivates to use the DGHT as an efficient proposal generation mechanism, followed by accepting or rejecting the proposals (based on image data) using a deep CNN. Due to the low false negative rate of the DGHT and the high accuracy of the CNN we obtain competitive performance on several pedestrian detection databases.

Keywords: Pedestrian detection · Hough transform · Error analysis · Proposal generation · Patch classification · Convolutional neural network

1 Introduction

In the last decades, automatic pedestrian detection has been a very important and still challenging task [4] in computer vision exhibiting many sources of large variability, i.a. regarding the object size and pose, occlusion and background. A lot of detection approaches have been suggested, among them feature-based detectors such as Viola-Jones [33] and Two-layer histograms of oriented gradients (HOG) [36], deformable part models [12,18], Random Forest-based approaches [25] and recently deep learning algorithms. The latter mainly consist of architectures using region proposals and a subsequent patch analysis with convolutional neural networks (CNN) as e.g. in Faster R-CNN [29]. Alternatively, approaches with a constant trivial region generation scheme and a subsequent bounding box regression or those that directly operate on full images have been proposed, e.g. R-CNN minus R [24], YOLO [27,28] and SSD [35], respectively.

S. Battiato et al. (Eds.): ICIAP 2017, Part II, LNCS 10485, pp. 104–115, 2017.
https://doi.org/10.1007/978-3-319-68548-9_10

The Discriminative Generalized Hough Transform (DGHT) [30] is an efficient voting-based localization approach and has successfully been applied in single-object localization tasks with limited variability, such as joint [30] and epiphyses [15] localization in medical images or state-of-the-art iris localization [14].

In this work, we analyze in detail the performance of the DGHT in a pedestrian detection task with many sources of variability (background, object size, pose etc.). In particular, we suggest to use the DGHT as an efficient proposal generation mechanism, accepting or rejecting the generated candidates using a deep convolutional neural network. We compare our approach to state-of-the-art algorithms, obtaining competitive performance on three different databases.

2 Methods

2.1 Structured Edge Detection

We use the real-time edge detection approach of [10], which learns information on the object of interest. Here, a Random Forest [5] maps an input image patch to an output edge image patch using pixel-lookups and pairwise-difference features of 13 (3 color, 2 magnitude and 8 orientation) channels. The approach incorporates ideas of Structured Learning [22] for handling the large amount and variability of training patch combinations as well as for efficient training. While testing, densely sampled, overlapping image patches are fed into the trained detector. The edge patch outputs which refer to the same pixel are locally averaged. The resulting intensity value can be seen as a confidence measure for the current pixel belonging to an edge. Subsequently, a non-maximum suppression (NMS) can be applied in order to sharpen the edges and reduce diffusion. For an example see Fig. 1b; further details can be found in [10].

2.2 Discriminative Generalized Hough Transform

The Generalized Hough Transform (GHT) [2] is well-known as a general model-based approach for object localization. Each model point m_j of the shape model M (Fig. 1c) is represented by its coordinates with respect to the reference point. The Discriminative Generalized Hough Transform (DGHT) [30] extends the

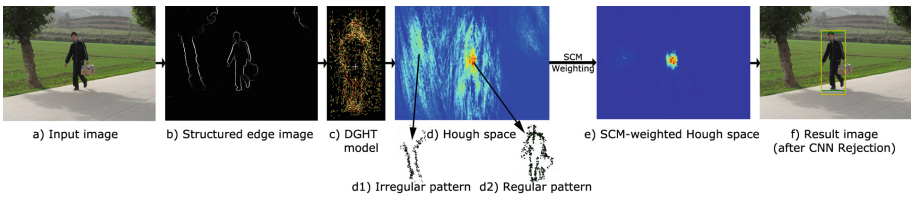

Fig. 1. Application of structured edge detection (Sect. 2.1), DGHT (Sect. 2.2) and SCM (Sect. 2.3) to an input image. Only a single image scale is shown.

GHT by individual model point weights λ_j for the J model points, which are optimized by a discriminative training algorithm that also accounts for automatic generation of M. Using this shape model M, the DGHT transforms a feature image X – in our work an edge image as outlined in Sect. 2.1 – into a parameter space H, called Hough space, by a simple voting procedure (see Fig. 1d):

$$H(c_i, X) = \sum_{\forall m_j \in M} \lambda_j f_j(c_i, X) \text{ with} \tag{1}$$

$$f_j(c_i, X) = \sum_{\forall e_l \in X} \begin{cases} 1, & \text{if } c_i = \lfloor (e_l - m_j)/\rho \rfloor \text{ and } |\varphi_{e_l} - \varphi_{m_j}| < \Delta\phi \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The Hough space H (we set the quantization parameter $\rho = 2$) consists of discrete Hough cells c_i, which accumulate the weighted number of matching pairs of all model points m_j (with corresponding weight λ_j) and feature points e_l. $f_j(c_i, X)$ determines how often model point m_j votes for Hough cell c_i given a feature image X. A vote, however, is only counted, if the absolute orientation difference of the model and feature point, φ_{m_j} and φ_{e_l}, respectively, is below $\Delta\phi$. Each Hough cell c_i represents a target hypothesis with its image space coordinates given by $\lfloor (c_i + 0.5) \cdot \rho \rfloor$. The number of weighted votes for each c_i corresponds to the degree of matching between model M and feature image X.

For ensuring good localization quality, a high correlation with the feature image at target point locations and a small correlation at confusable objects is desired. The DGHT achieves this by an iterative training procedure starting with an initial model of superimposed annotated feature images at the reference point. In each iteration, the model point weights λ_j are optimized using a Minimum Classification Error (MCE) approach and, afterwards, the model is extended by target structures from training images which still have a high localization error. To reduce model size, those points with low absolute weights are eliminated. This procedure is repeated until all training images are used or have a low localization error. Further details on this technique can be found in [30].

2.3 Rejection of Proposals

Shape Consistency Measure (SCM). [14] suggested to analyze the model point pattern voting for a particular Hough cell c_i. More specifically, a Random Forest [5] is applied to classify the model point pattern into a class "regular shape" Ω_r (representing e.g. a frontal or a side view of a person) and a class "irregular shape" Ω_i (see Fig. 1(d1) and (d2)).

To train the Random Forest Classifier, the DGHT is applied to each training image. Afterwards, the class labels Ω_r and Ω_i are assigned to the individual Hough cells of the training images: Cells with a localization error $<\varepsilon_1$ are labeled as class Ω_r while those with an error $>\varepsilon_2$ are assigned to class Ω_i.

For a test image X, a DGHT model is applied to generate a Hough space H. For each local maximum c_i in H, the Random Forest Classifier is used to calculate the probability $p = p(\Omega_r, c_i)$ that the set of model points voting for c_i

has a regular shape. The obtained probability is used as an additional weighting factor for the Hough space votes, i.e. $S(c_i, X) = H(c_i, X) \cdot p(\Omega_r, c_i)$ (see Fig. 1e). The local maxima in H are now sorted according to decreasing $S(c_i, X)$ to provide an ordered list $C = \{c_i\}$ of most probable object positions c_i.

Deep Convolutional Neural Networks. Deep convolutional neural networks (CNN) have been successfully used for image classification tasks achieving state-of-the-art classification performance e.g. on the large-scale ImageNet classification challenge [23,31,38]. In combination with a separate region proposal generation step, deep CNNs have been successfully applied in object detection taks, e.g. R-CNN [19] or Fast R-CNN [20]. Furthermore, Faster R-CNN [29] combined these two components into one network sharing convolutional features.

In this work, we use a deep CNN to individually accept or reject each proposal c_i out of the list C generated by the DGHT+SCM (see Sects. 2.2 and 2.3). Specifically, each candidate position $c_i \in C$ (seen as a proposal) is transferred from Hough space to image space. Then, a bounding box corresponding to the mean object size (60×160 px) is centered around that position (Fig. 1f), and the image patch corresponding to the bounding box is rescaled to an input size of 64×64. The patch pixel intensities of all three color channels are normalized to $[0, 1]$, and then used as input to a deep CNN. We use the standard VGG16 classifier as described in [31], pre-trained on ImageNet and fine-tuned on the IAIR training corpus (see Sect. 3.1). The output of the CNN is a softmax layer with 2 classes, pedestrian and background. We use $p = p(\text{pedestrian}, c_i)$ for candidate rejection (see Sect. 3.2). With an appropriate rejection threshold θ, any candidate c_i is rejected if $p(\text{pedestrian}, c_i) < \theta$ (Fig. 1f).

3 Experimental Setup

3.1 Databases

IAIR-CarPed. We perform most experiments on the IAIR-CarPed [36] database, because it has a reasonable amount of independent 2D images and additionally offers difficulty labels (e.g. occlusion, low contrast) for each annotation. We use this additional information for our detailed error analysis. As suggested in [36], we train on a random 50%-split of the available pedestrian images, i.e. in total 1046 images containing 2341 pedestrians with an object height range from 45 to 383 px (mean height: 160 px). The remaining 1046 images (2367 pedestrians with a similar object height range and mean height) are used for evaluation. Training and test corpus each contain all types of difficulties present in the IAIR corpus.

INRIA Person. We also evaluate our approach on the well-known INRIA Person [6] database. The test set contains 288 images which contain 561 annotated persons with a height range from 100–788 px (mean height: 299 px).

TUD Pedestrians. Moreover, we apply our framework to the TUD Pedestrians [1] dataset. The test set consists of 250 images containing 311 annotated pedestrians with a height range from 71 to 366 px (mean height: 213 px).

3.2 Experimental Setup and System Parameters

Our system setup for the single-frame detection of pedestrians in non-consecutive 2D RGB images is organized as follows:

Feature Image Generation: As input images for training and testing, we use the output of the Structured Edge Detector (see Sect. 2.1). We train this edge detector specifically for pedestrians on the PennFudan database [34], because this is the only database we were aware of providing segmentation information needed to train the edge detector. The Structured Edge Detector suppresses most of the background edges and thus significantly reduces background variability [13].

DGHT Model and SCM Training: In order to generate a DGHT pedestrian model including a certain amount of size variability, we allow a size range of 144–176 px (mean object height ±10%). All training images with pedestrians not in this size range are scaled to a person size selected randomly from the allowed range (uniform distribution), separately for each pedestrian in an image. To train our DGHT shape model (see Sect. 2.2), we only use those full training images containing "simple" pedestrians (IAIR difficulty type "S", 1406 pedestrians/775 images). Having trained the DGHT shape model, we train the SCM on the full IAIR training set comprising all difficulty types and all pedestrians scaled to the range 144–176 px as described above (see also Sect. 2.3). We set ε_1 for class Ω_r to 5 and ε_2 for class Ω_i to 15 Hough cells.

Testing: To handle the large range of object sizes, we scale each test image by the following heuristic set of 10 scaling factors such that each pedestrian should roughly fit into the expected object range (mean object height ±10%):

50%, 62.5%, 75%, 100%, 150%, 200%, 225%, 250%, 275%, 300%.

The trained DGHT model is applied independently to each scaled image, i.e. independent Hough spaces are generated for each image scale and, afterwards, weighted by the SCM (Sect. 2.3). In each weighted Hough space, local maxima $C = \{c_i\}$ are identified using a NMS with a minimum distance of 1/3 of the model width, i.e. 20 px. To reduce the amount of candidates, we discard those candidates c_i with $S(c_i, X) < \max S(X) \cdot 0.2$. To reject possible false positive candidates e.g. due to inconsistent voting schemes [14], we investigate two rejection mechanisms:

SCM Rejection: For each image scale independently, a candidate is rejected if $p(\Omega_r, c_i) < \theta$, see Sect. 2.3. This is a purely Hough-based rejection method.

Our results and error analysis (Sect. 4) motivate an additional rejection step:

CNN Rejection: Here, any candidate position c_i is transferred to (scaled) image space, and a bounding box corresponding to the mean model size is centered around that position. A deep CNN is then used to reject c_i if $p(\text{pedestrian}, c_i) < \theta$. This is an image-based rejection, as opposed to the SCM rejection.

We use the standard Keras VGG16 model, which is initialized on ImageNet. We fine-tune this model on our IAIR training corpus, using the annotated pedestrian bounding boxes scaled to $(64 \times 64 \times 3)$ as positive samples and the same candidates as for class Ω_i in the SCM training as negative samples, i.e. high

scoring peaks with a minimum error of 15 Hough cells. For fine-tuning we use the Adam optimizer [21] with categorical cross-entropy loss, a learning rate η of 0.001, which is reduced on plateaus, and an input dimension of $(64 \times 64 \times 3)$.

Combining Scales and Post-processing: Subsequent to the rejection step, the remaining candidate bounding boxes of all image scales are divided by the respective scaling factor to match the original image dimensions. Afterwards, the candidates are greedily grouped based on the mutual overlap[1] and finally a NMS is applied to each group using $S(c_i, X)$ (without CNN) or $p(\text{pedestrian}, c_i)$ as criterion, respectively, in order to avoid double detections.

Analysis: As suggested in [39], we conduct a detailed error analysis including oracle experiments: (a) localization oracle (all false positives (FP) that overlap with the ground truth are ignored) and (b) background vs. foreground oracle (all FP that do not overlap with the ground truth are ignored). In addition, we conduct another oracle experiment (c): perfect rejection oracle (DGHT as a proposal generator). For each ground truth annotation, the rejection oracle picks the best matching candidate out of the set C generated by the DGHT (including the SCM and post-processing over image scales) and rejects all other candidates. Thus, we quantify the minimal miss rate for the DGHT as proposal generator, assuming a perfect rejection mechanism.

3.3 Comparison to State-of-the-Art Approaches

We compare our approach against several state-of-the-art algorithms. For the IAIR-CarPed database, we compare to the results for the Two-layer HOG and the PASCAL deformable part model (DPM) published in [36]. We also downloaded the latest DPM release (DPMv5) [18] trained on PASCAL, the pre-trained YOLOv1 [27] full model as well as the pre-trained YOLOv2 [28] full model (both pre-trained on ImageNet and fine-tuned on PASCAL) and evaluated them on our IAIR-CarPed test corpus. Addtitionally, we used the pre-trained YOLOv1 full model and fine-tuned it on our IAIR-CarPed training set. For details on these state-of-the-art approaches see the respective references. For the other databases, we use the benchmark results from [16] and [37], respectively.

3.4 Evaluation Metrics

We evaluate our detections using the intersection over union (IoU) measure. According to [11], a detection of an object is correct if the IoU of the prediction and the ground truth exceeds 50%. As suggested in [9], for single frame evaluation we computed Detection Error Tradeoff (DET) curves plotting the miss rate against the false positives per image (FPPI) on a log-log scale by modifying the rejection threshold θ. For comparison, the miss rates at 0.5 and 1 FPPI are shown. For the TUD Pedestrians database, we use the recall at equal error rate (EER), as other groups have frequently used this measure. For measuring the candidate quality, we use the Average Best Overlap (ABO) score from [32].

[1] We set this parameter to 30%.

4 Results

4.1 SCM Rejection

Table 1 compares the DGHT + SCM (without CNN) detection performance to other approaches on the IAIR test corpus as described in Sect. 3.2.

As can be seen, the overall performance of the DGHT is comparable to the Two-Layer HOG and the published original DPM result, but worse than the current DPMv5, the fine-tuned YOLOv1 and the pre-trained YOLOv2.

Table 1. Comparison of detection results on the IAIR-CarPed test corpus[a] S: Simple, D1: Occlusion, D2: Low Contrast, D3: Infrequent Shape

Approach	Miss Rate at 0.5 FPPI					Miss Rate at 1 FPPI				
	S	D1	D2	D3	All	S	D1	D2	D3	All
DGHT + SCM	0.32	0.51	0.74	0.50	0.44	0.22	0.40	0.66	0.32	0.34
Two-Layer HOG [36]	N/A	N/A	N/A	N/A	N/A	0.25	0.47	0.44	0.50	0.35
DPM [36]	N/A	N/A	N/A	N/A	N/A	0.29	0.37	0.45	0.36	0.34
DPMv5 [18]	0.18	0.37	0.47	0.45	0.29	0.16	0.32	0.40	0.40	0.25
YOLOv1 Pre-trained [27]	0.37	0.47	0.87	0.45	0.49	0.32	0.41	0.81	0.42	0.44
YOLOv1 Fine-tuned [27]	0.06	0.18	0.23	0.18	0.13	0.06	0.17	0.22	0.17	0.13
YOLOv2 Pre-trained [28]	0.13	0.28	0.31	0.23	0.21	0.12	0.25	0.28	0.20	0.19

[a]Results for Two-Layer HOG and DPM taken from [36], i.e. potentially different training/test split of the IAIR corpus. Results for DGHT, DPMv5, YOLOv1 and YOLOv2 obtained on our test split. For details on pre-training and fine-tuning see Sect. 3.3

4.2 SCM Rejection: Error Analysis

In this section, we analyze in detail all errors of the DGHT with SCM rejection (without CNN), in comparison to the DPMv5. We specifically focus on the DPM since the code is publicly available (such that we can perform own experiments on the IAIR corpus) and since it is also a model-based approach operating on feature images. We use a similar error analysis as described in [39].

Generally, there are two basic types of errors: (I) false positives (FP), i.e. a false detection, e.g. because of confusable background structures, misaligned, larger or smaller predictions ($0\% < \text{IoU} < 50\%$) or double detections; (II) false negatives (FN), which are ground truth annotations that are not detected. FN mostly occur because of the "well known difficulty of detecting small and occluded objects" [39] as small persons are often over-/underexposed or blurry. Besides, there might be a dataset bias, e.g. side-views or cyclists are underrepresented in the training set, which hampers the detection of such instances.

We manually evaluated all detection errors (FP and FN) of the DGHT and DPMv5 [18] on the IAIR test corpus (see Sect. 3.2) at 1 FPPI:

28% (DGHT)/31% (DPM) of all FP are due to localization errors, mostly because of body part (DGHT) and double detections (DPM).

For both approaches, the vast majority of FP are background detections (69% (DGHT)/66% (DPM)) due to confusing vertical structures or trees. The DPM often detects pedestrian groups as one detection. The remaining 3% are missing annotations, which actually are true instead of false positives.

For the DGHT, 29% of all FN are due to low contrast, i.e. insufficient or no feature representation at all. Another 20% are slightly below the rejection threshold at 1 FPPI indicating that the SCM as a rejection mechanism could still work more properly. The remaining 50% mainly consist of errors at small scales (16%), localization errors (12%) and occluded pedestrians (11%).

For the DPM, the main reasons for FN are small scales (28%), occlusion (26%) and low contrast (22%; mostly also at small scales). The remaining FN mainly consist of localization errors (7%), side views (6%), and cyclists (3%).

The DPM has more FN because of small scales or occlusion. Low contrast or missing edges are a problem for both approaches. Side views/cyclists are better detected with the DGHT pipeline. Moreover, the DGHT FN often are only slightly below θ (with a lower θ, too many FP would have been generated).

4.3 SCM Rejection: Oracle Experiments

The results of the oracles (a), (b) and (c) (see Sect. 3.2 "Analysis") are shown in Table 2. Since the localization oracle only reduces the miss rate at 1 FPPI by 0.03, it shows that the DGHT detections are usually quite accurate (except for a few outliers). On the contrary, there is still much room for improvement regarding background vs. foreground errors. If we would be able to reject FP in the background (IoU = 0%), the miss rate drops from 0.34 (current DGHT + SCM result at 1 FPPI) to 0.16 at only 0.3 FPPI. This again indicates that the DGHT candidates are very accurate, but the rejection using only the model point voting pattern on structured edge images is not sufficient to properly overcome the well-known problems of small-scale detections or those of confusable background structures (see Sect. 4.2). In case of perfect rejection of DGHT proposals, we would obtain a miss rate of only 0.04 with an ABO score of 78.2% (perfect rejection oracle). This clearly indicates that the main drawbacks of the DGHT are (a) FP in the background and (b) non-optimal selection of candidates based on $S(c_i, X)$. The low miss rate of the perfect rejection oracle motivates to use the DGHT as a proposal generator, and to improve the proposal rejection. To this end, we apply the CNN proposal rejection subsequently to the SCM rejection, as outlined in Sect. 3.2.

Table 2. DGHT oracle results (at highest or max. 1 FPPI)

Experiment	S	D1	D2	D3	All
DGHT + SCM at 1 FPPI	0.22	0.40	0.66	0.32	0.34
DGHT Localization oracle at 1 FPPI	0.20	0.35	0.49	0.30	0.31
DGHT BG vs. FG oracle at 0.3 FPPI	0.09	0.24	0.24	0.23	0.16
DGHT Perfect Rejection Oracle at 0 FPPI	0.01	0.01	0.20	0.00	0.04

4.4 CNN Rejection: Detection Results

Table 3 shows the detection results using additional CNN proposal rejection on top of SCM rejection ("DGHT + SCM + VGG16") as introduced in Sect. 3.2 and motivated in Sects. 4.2 and 4.3[2]. Additionally, we show the results of this setup on TUD Pedestrians and INRIA Person in Tables 4 and 5, respectively. Note that no component of our system has been retrained on TUD or INRIA. We obtain minimal miss rates of 0.04 (78.2% ABO) at 350 candidates per image, 0.01 (75.8% ABO) at 55 candidates per image and 0.01 (76.8% ABO) at 102 candidates per image on IAIR, TUD Pedestrians and INRIA Person, respectively.

Table 3. Comparison of detection results on the IAIR-CarPed test corpus S: Simple, D1: Occlusion, D2: Low Contrast, D3: Infrequent Shape. Results of other state-of-the-art algorithms are partly repeated from Table 1

Approach	Training data	Miss Rate at 0.5 FPPI				
		S	D1	D2	D3	All
DGHT + SCM	IAIR	0.32	0.51	0.74	0.50	0.44
DGHT + SCM + VGG16	ImageNet/IAIR	0.09	0.30	0.40	0.20	0.19
DPMv5 [18]	PASCAL	0.18	0.37	0.47	0.45	0.29
YOLOv1 Pre-trained [27]	ImageNet/PASCAL	0.37	0.47	0.87	0.45	0.49
YOLOv1 Fine-tuned [27]	Im.Net/PASC./IAIR	0.06	0.18	0.23	0.18	0.13
YOLOv2 Pre-trained [28]	ImageNet/PASCAL	0.13	0.28	0.31	0.23	0.21

Table 4. Recall at EER on TUD Pedestrians without retraining

Approach	DGHT + SCM + VGG16	PartISM [1]	HoughForests [17]	Yao et al. [37]
Training data	IAIR	TUD/INRIA	TUD/INRIA	TUD/INRIA
Recall at EER	0.88	0.84	0.87	0.92

Table 5. Miss Rate at 1 FPPI on INRIA Person without retraining. Ours: DGHT + SCM + VGG16

Approach	Ours	HOG	ICF [7]	Yao [37]	FPDW [8]	VeryFast [3]	Spat.Pool. [26]
Training data	IAIR	INRIA	INRIA	INRIA	INRIA	INRIA	INRIA/Caltech
Miss rate	0.14	0.23	0.14	0.12	0.09	0.07	0.04

5 Discussion

The experiments have shown that the DGHT in general is suitable for proposal generation due to the low false negative rate and the comparably small number

[2] Evaluation at 0.5 FPPI since this is the highest FPPI rate for DGHT+SCM+VGG16.

of candidates. The trained DGHT pedestrian model for proposal generation also generalizes well to other pedestrian databases without retraining any of the components. Our pedestrian detection pipeline achieves comparable results to other state-of-the-art approaches. In comparison to Selective Search [32] (2,000–10,000 candidates) or the region proposals of Faster R-CNN [29] (300+ candidates), the DGHT outputs a smaller number of candidates (see Sect. 4.4). Example detections are shown in Fig. 2.

Our approach still has some limitations: due to the edge feature images, we intrinsically miss those objects which are of low contrast, since they do not generate well pronounced edges or no edges at all. This limitation can be seen in Tables 2 and 3 at difficulty type D2. Additionally, we currently do not perform any bounding box refinement step which might further improve detection accuracy. However, the ABO scores of >75% for all three databases indicate that the candidates are already of good quality. Currently, our implementation does not aim at real-time performance. However, due to the independent voting of model points, the DGHT exhibits a high potential for parallelization.

Fig. 2. Example DGHT + SCM + VGG16 detections on IAIR. (green): ground truth, (yellow): correct detection, (blue): FP, (red): FN; best viewed in color

6 Conclusions

In this work, we applied the DGHT as a proposal generator - in combination with proposal rejection by a deep CNN - to a real-world multi-object detection task exhibiting many sources of variability, namely pedestrian detection. We obtained comparable performance to state-of-the-art approaches on the IAIR-CarPed, the TUD Pedestrians and the INRIA Person databases, demonstrating that our framework (trained on IAIR) generalizes well also to other datasets. The main advantages of the DGHT proposal generation are (a) the relatively low amount of training images needed for training of DGHT and SCM (on the order of 100 or less per variability class), (b) the low amount of resources needed at test time, (c) the relatively low amount of proposals generated per image. Thus, our framework could be useful especially for detecting specific object categories with limited available training material.

Acknowledgement. This work was funded by the Department of Social Affairs, Health, Science and Equality of Schleswig-Holstein, Germany. Thanks to Andrew J. Richardson for providing the VGG16 framework.

References

1. Andriluka, M., et al.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR (2008)
2. Ballard, D.H.: Generalizing the Hough transform to detect arbitrary shapes. Pattern Recogn. **13**(2), 111–122 (1981)
3. Benenson, R., et al.: Pedestrian detection at 100 frames per second. In: CVPR (2012)
4. Benenson, R., et al.: Ten years of pedestrian detection, what have we learned? In: ECCV (2014)
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
7. Dollar, P., et al.: Integral channel features. In: BMVC (2009)
8. Dollar, P., et al.: The fastest pedestrian detector in the west. In: BMVC (2010)
9. Dollar, P., et al.: Pedestrian detection: an evaluation of the state of the art. In: PAMI (2012)
10. Dollar, P., et al.: Fast edge detection using structured forests. In: PAMI (2015)
11. Everingham, M., et al.: The PASCAL VOC challenge. In: IJCV (2010)
12. Felzenszwalb, P., et al.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
13. Gabriel, E., et al.: Structured edge detection for improved object localization using the discriminative generalized Hough transform. In: Proceedings of VISAPP (2016)
14. Hahmann, F., et al.: A shape consistency measure for improving the generalized Hough transform. In: Proceedings of VISAPP (2015)
15. Hahmann, F., Böer, G., Deserno, T.M., Schramm, H.: Epiphyses localization for bone age assessment using the discriminative generalized Hough transform. In: Deserno, T.M., Handels, H., Meinzer, H.-P., Tolxdorff, T. (eds.) Bildverarbeitung für die Medizin 2014. I, pp. 66–71. Springer, Heidelberg (2014). doi:10.1007/978-3-642-54111-7_17
16. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/
17. Gall, J., Lempitsky, V.: Class-specific Hough forests for object detection. In: CVPR (2009)
18. Girshick, R.B., Felzenszwalb, P.F., McAllester, D.: Discriminatively Trained Deformable Part Models, Release 5. http://people.cs.uchicago.edu/~rbg/latent-release5/
19. Girshick, R.B., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
20. Girshick, R.B.: Fast R-CNN. In: ICCV (2015)
21. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: ICLR (2015)
22. Kontschieder, P., et al.: Structured class-labels in random forests for semantic image labelling. In: ICCV (2011)
23. Krizhevsky, A., et al.: ImageNet classification with deep CNNs. In: NIPS (2012)
24. Lenc, K., Vedaldi, A.: R-CNN minus R. In: BMVC (2015)
25. Marin, J., et al.: Random forests of local experts for pedestrian detection. In: ICCV (2013)
26. Paisitkriangkrai, S., et al.: Strengthening the effectiveness of pedestrian detection. In: ECCV (2014)

27. Redmon, J., Farhadi, A.: YOLO: unified, real-time object detection. In: CVPR (2016)
28. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. arXiv:1612.08242 (2016)
29. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS (2015)
30. Ruppertshofen, H.: Automatic Modeling of Anatomical Variability for Object Localization in Medical Images. BoD-Books on Demand, Norderstedt (2013)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
32. Uijlings, J.R.R., et al.: Selective search for object recognition. In: IJCV (2012)
33. Viola, P., et al.: Detecting pedestrians using patterns of motion and appearance. In: IJCV (2005)
34. Wang, L., et al.: Object detection combining recognition and segmentation. In: ACCV (2007)
35. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). doi:10.1007/978-3-319-46448-0_2
36. Wu, Y., Liu, Y., Yuan, Z., Zheng, N.: IAIR-CarPed: a psychophysically annotated dataset with fine-grained and layered semantic labels for object recognition. Pattern Recogn. Lett. **33**(2), 218–226 (2012)
37. Yao, C., et al.: Human detection using learned part alphabet and pose dictionary. In: ECCV (2014)
38. Zeiler, M., Fergus, R.: Visualizing and understanding ConvNets. In: ECCV (2014)
39. Zhang, S., et al.: How far are we from solving pedestrian detection? In: CVPR (2016)