Object Proposals for Pedestrian Detection in Stereo Images

UYEN D. X. NGUYEN[1], FRANZ ROTTENSTEINER[1] & CHRISTIAN HEIPKE[1]

Abstract: Pedestrian detection is an active research field in computer vision and photogrammetry today due to its importance for applications related to autonomous driving, machine human interaction, and surveillance. Object proposals play a significant role in guiding a classifier where to examine image regions that may contain pedestrians. In this paper, we present a proposal framework employing both 3D information derived from stereo images and RGB cues to generate pedestrian bounding box proposals with high recall and relatively small number of candidates. Our proposal generator has two stages: (1) generating a large number of proposals to achieve a good recall value; and (2) eliminating unreasonable initial candidates based on 3D cues and pedestrian geometric constraints. Preliminary experiments on the Kitti benchmark dataset show that our proposal framework is comparable to state-of-the-art methods.

1 Introduction

Today, pedestrian detection, a specific case of object detection, is one of the most vigorous research areas in computer vision and photogrammetry, especially in applications related to autonomous driving, robotics, safety, surveillance, etc. With the support of these systems, human effort in processing huge amounts of images, which is time-consuming, expensive and subjective, can be reduced or completely avoided.

In recent years, Convolution Neural Networks (CNNs) have undergone a dramatic development, which enabled object detection to become more and more accurate and reliable (REDMON et al. 2016; GIRSHICK 2015; REN et al. 2015; GIRSHICK et al. 2014). However, every CNN object detector depends on a region proposal method to determine possible locations in an image where desired objects can appear. A typical paradigm for this is exhaustive search in the entire image with sliding windows of multiple scales (HARZALLAH et al. 2009; VEDALDI et al. 2009). This approach is easy and simple, but results in a very large number of candidates, which makes it computationally expensive to adopt a complex classifier for reliable recognition. Recently, algorithms were suggested which exploit low level features to generate a set of proposals to reduce the number of regions which have to be examined more closely (ALEXE et al. 2012; UIJLINGS et al. 2013; CHENG et al. 2014; ZITNICK & DOLLÁR 2014). These algorithms proved their effectiveness on different datasets like PASCAL VOC (EVERINGHAM et al. 2010) or ImageNet (DENG et al. 2009). However, for a dataset containing objects with large scale variation, occlusion, and truncation like the Kitti benchmark for autonomous driving (GEIGER et al. 2012), most of these methods do not achieve a satisfactory recall with a relatively small number of proposal bounding boxes (HOSANG et al. 2016).

In this work, with the focus on generating object proposals for pedestrians using a stereo camera, we tackle the problem of region proposals in two steps, and we aim to obtain a high recall with a

[1] Leibniz Universität Hannover, Institut für Photogrammetrie und GeoInformation, Nienburger Str. 1,
D-30167 Hannover, E-Mail: [nguyen, rottensteiner, heipke]@ipi.uni-hannover.de

limited number of proposals. First, a large number of proposals is produced in an initial step to recover a high recall. Then, a filtering step eliminates proposed regions that most certainly do not contain pedestrians. For that purpose, we employ geometric constraints relating to pedestrians and scene context information, and we re-rank all bounding boxes produced from the first step with a scoring function. The candidates with the highest scores are then selected for further analysis and classification. Two main assumptions are made about pedestrians: they all stand on the ground and have restricted size. Further assumptions concern the scene, which we take as being composed of a horizontal ground plane and a number of objects with primarily vertical extent. To demonstrate the efficiency of our approach, we conduct experiments on the Kitti dataset. The performance of our region proposal is analysed and compared to other state-of-the-art methods.

The remainder of this paper is arranged as follows: Section 2 reviews related works. Section 3 presents the methodology and algorithms used in our work. Section 4 illustrates our experimental results, followed by the conclusion in Section 5.

2 Related Works

In this chapter we discuss previous work concerning region proposal methods. It is the aim of a proposal generator to determine all instances of interesting objects in an image with as few false alarms as possible, delineating them, e.g., by bounding boxes. The result allows more sophisticated and accurate classifiers like CNNs to be adopted to detect the objects. Currently, several paradigms are widely applied for the region proposal task.

Objectness scoring: this approach uses a sliding window in multiple scales to examine every area in an image. Each region is then assigned an "objectness" score based on combining a number of cues such as normalised gradients, structured edges, texture, etc. Then the regions with the highest score are selected as proposal candidates. Several well-known algorithms in this group are objectness (ALEXE et al. 2012), edge boxes (EB) (ZITNICK & DOLLÁR 2014), BING (CHENG et al. 2014). EB is one of the most successful methods in this group, which efficiently evaluates a proposal candidate based on the contours located inside and on the edge of the bounding box.

Super pixel grouping: representative proposal generators of this group are selective search (SS) (UIJLINGS et al. 2013) and multiscale combinatorial grouping (MCG) (ARBELAEZ et al. 2014). This technique exploits a hierarchical segmentation to obtain a set of super pixels, which are subsequently merged and scored by a ranking function. SS can deliver high quality proposals and is widely used by many detectors (WANG et al. 2013; GIRSHICK et al. 2014) due to its comparably short execution time and high recall. However, the results are sensitive to large size variation, occlusion, and truncation of objects. In addition, small objects lead to the requirement to select small super pixels, which can lead to high computation cost.

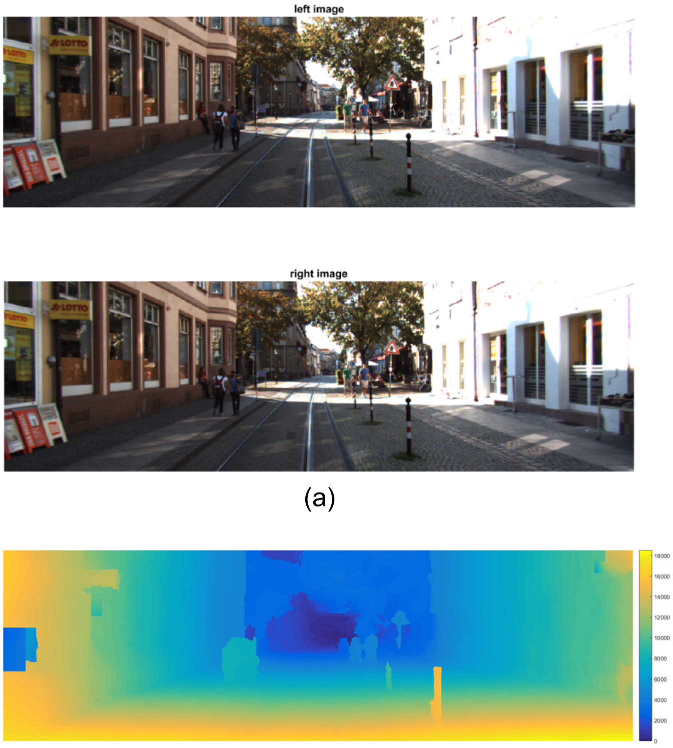
3-D object proposal: the methods mentioned above mostly focus on solving the object proposal task based on RGB images. However, depth information can also provide a valuable cue for evaluating candidate regions. A few methods like MCG-D (GUPTA et al. 2014) consider depth information to improve performance. But their complicated scoring functions make the method rather slow. Using depth cues and energy minimization to infer potential objects with 3D boxes, CHEN et al. (2015) can achieve a good recall value for the Kitti benchmark. Nevertheless, considering only 3D position and dismissing RBG information is not necessarily a good idea because many object features can only be distinguished in RGB images.

3 Methodology

In this study, we aim to develop a pedestrian proposal framework. After appropriate 3D scene modelling we generate initial proposals based on RGB image information using the EB method (ZITNICK & DOLLÁR, 2014). Then, we use 3D information obtained from the stereo image pair and geometric constraints corresponding to pedestrians to filter out unreasonable bounding boxes.

3.1 Modelling the Scene

Given a stereo image pair with known orientation parameters, the disparity map D of all pixels with respect to the left image is first estimated using a state-of-the-art dense matching approach (YAMAGUCHI et al., 2014). Then, a point cloud P is computed using the disparity values (Fig. 1).



(a)

(b)

Fig. 1: The stereo image pair (a) is used to compute the disparity map (b).

We assume our scene to be mainly composed of a more or less horizontal ground plane (e.g. a road), vertical planes (e.g. building facades) and the sky (which we don't consider further). In addition, other objects such as pedestrians are present, which are our objects of interest. Many objects in an urban scene can be considered as vertical planar surfaces supported by the ground plane. Reconstructing these vertical objects and the ground plane in object space would provide additional evidence for pedestrian detection.

3.1.1 Potential obstacles

We define potential obstacles, i.e. non-pedestrian objects, as regions in an image which are perpendicular to the ground plane. Under this definition, a number of pixels in each image column that have the same disparity value belong to an obstacle if we assume images to be taken with horizontal and parallel optical axes. A binary obstacle mask for each input image is estimated using the following steps proposed by HU & UCHIMURA (2005):

Step 1 From the disparity map D, the vertical or v-disparity image V_{dis} is computed such that each column in V_{dis} is a disparity histogram of the corresponding column in D.

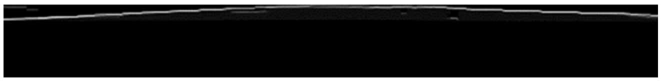


Fig. 2: v-disparity image estimated from the disparity map shown in Fig. 1b.

Step 2 The intensity of a pixel $v(x, y)$ in V_{dis} represents the number of pixels in column y of D that have approximately disparity x. Hence, the binary obstacle mask O_{mask} can be built by finding pixels in the disparity map D, which correspond to pixels with the entry in V_{dis} larger than a threshold value. Those pixels in D are considered as obstacle regions in O_{mask}.

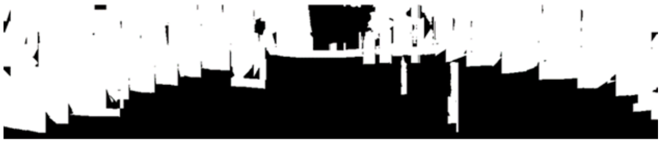


Fig. 3: The binary obstacle mask is built from the v-disparity image; obstacle pixels are shown in white.

Step 3 Morphological closing is used to join small obstacle regions together. Then, all none-obstacle regions smaller than a threshold are considered to be caused by errors and, thus, are included as obstacle areas.

In fact, we produce two obstacle masks O_{mask1} and O_{mask2} by applying two different thresholds (Fig. 4). Based on a small threshold, O_{mask1} contains nearly all vertical objects in the scene. In contrast, O_{mask2} is computed with a higher threshold, so small objects (e.g. pedestrians) should not be part of this mask.

3.1.2 Ground plane extraction

Unlike vertical object pixels, ground plane pixels should have similar depths per row. Using this assumption, the ground plane is estimated as follows:

Step 1 We use the obstacle mask O_{mask1} to eliminate most pixels related to vertical objects in the disparity map, so that the remaining pixels in a new disparity map:

$$D_{nonobs} = O_{mask1} .* D \, , \quad (1)$$

mostly belong to the ground plane. In (1), $.*$ denotes a pixel-wise multiplication of the grey values.

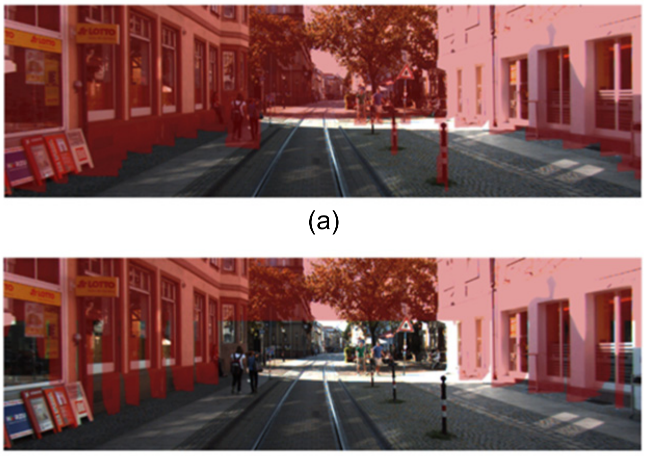
Step 2 Ground pixels in D_{nonobs} are determined in a similar way as the obstacle mask. However, instead of using the v-disparity image, we compute the disparity histogram of each row in D_{nonobs} to generate a horizontal or h-disparity image. As a result, ground pixels in D_{nonobs} with their 3D positions are collected in a set $P_{ground} = \{p_1(\mathrm{x,y,z}), \dots, p_n(\mathrm{x,y,z})\}$.

Step 3 Any 3-D point $(\mathrm{x,y,z})$ lying on the ground plane (Ω) must satisfy the planar equation

$$(\Omega): a\mathrm{x} + b\mathrm{y} + c\mathrm{z} + d = 0, \tag{2}$$

where (a, b, c) is the normal vector with length 1 and d is the distance from the origin to (Ω).

The ground plane (Ω) is determined using the 3-D ground points P_{ground}, together with RANSAC to remove outliers. Then, the ground pixels in the image which form a ground mask G_{mask} are determined as those which have an absolute distance in object space to (Ω) smaller than a value ε.



(a)

(b)

Fig. 4: Binary obstacle mask O_{mask1} which covers almost all vertical objects (red part) (a). Binary obstacle mask O_{mask2} which covers large vertical objects only (b).

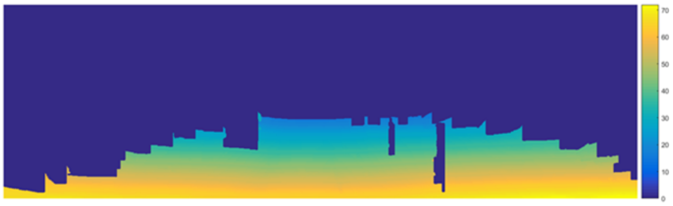


Fig. 5: The disparity map without vertical objects in the obstacle mask O_{mask1} (Fig. 2.a).

3.1.3 Areas of interest

We define areas of interest as those areas where pedestrian can appear in the image. Assuming that pedestrians are not taller than a threshold $high_{pedes}$, pixels corresponding to 3-D points having a distance to (Ω) in object space smaller than $high_{pedes}$ and larger than ε belong to the area of interest, which are represented by a binary mask PD_{mask}. Assuming a pixel $pd(c, r)$ in PD_{mask} , where *c* and *r* are the pixel's column and row indices, respectively, corresponds to the 3-D position $pd(\mathrm{x,y,z})$ in object space, PD_{mask} is derived by (3):

$$PD_{mask}(c, \mathrm{r}) = \begin{cases} 1 \text{ if } & \varepsilon < dist(pd(\mathrm{x},\mathrm{y},\mathrm{z}),\ \Omega) \leq high_{pedes} \\ 0 & otherwise \end{cases} . \quad (3)$$

In (3), $dist$ is the Euclidean distance from a point to a plane in 3-D space.

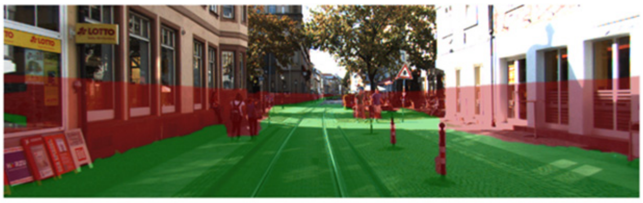


Fig. 6: Extracted ground plane (green) and areas of interest (red) in the reference image of a stereo pair.

3.2 Object Proposal and Re-ranking

In order to search for pedestrians in the image, we apply an object proposal framework consisting of two stages to the areas of interest with the purpose of achieving a high recall while only producing a small number of candidate bounding boxes. A huge number of initial candidates are produced in the first stage; we then employ additional cues related to geometrical knowledge of pedestrians and scene information to eliminate unreasonable proposals.

3.2.1 Initial proposals

Instead of combining different features, the EB algorithm (ZITNICK & DOLLÁR 2014) directly generates object proposal from edge maps generated by the structured edge detector (DOLLÁR & ZITNICK 2013). EB computes objectness as the likelihood of a bounding box (BB) to contain an object based on the number of contours lying within and on that BB, where a contour is a set of edges forming a coherent boundary, curve or line. The reader is referred to (ZITNICK & DOLLÁR 2014) for more details on how the objectness score is computed. The desired number of BB can be influenced by setting an objectness score threshold. There are three variants of EB based on different parameter values: EB 50, EB 70, and EB 90, in which the step size of the sliding window decreases and the density of sampling increases, respectively. Therefore, depending on the trade-off between accuracy, efficiency, and the number of desired proposals, one wants to achieve for a specific purpose, a suitable variant can be chosen.

In this study, the variant EB 70, which has nearly similar running time as EB 50 but can produce more precise BB, is employed to generate regions that potentially contain any type of object. In order not to eliminate pedestrians in this early stage, we choose a small threshold for the objectness score, which will result in a relatively large number of regions. However, the grey values of image pixels that are outside the areas of interest according to PD_{mask} are set to 0 before passing the image on to the EB 70 algorithm, so that no proposal regions that are completely outside the area of interest will be generated.

3.2.2 Proposal ranking

The initial proposal step produces candidate regions for unspecific types of objects, so that we want to filter out BB that are very unlikely to contain pedestrians. For that purpose, we exploit scene information and prior knowledge about pedestrians, which is not considered in the first stage. We define features that encode this additional information and use them to define a combined scoring function, which is used to rank all proposals. Only BB with a high score are maintained because they are very likely to contain a pedestrian, whereas the others are discarded. The high-scoring BB can be fed to a sophisticated classifier for the final decision whether they contain a person or not. Details about that procedure are given in the following paragraphs.

Overlap with the area of interest: this criterion expresses the fact that a good candidate region should lie inside or at least cover a part of the area of interest. We determine an overlap rate f_{int_area} for each bounding box BB:

$$f_{int\_area} = \frac{\sum_{p(c,r) \in BB}(PD_{mask}(c,r))}{|BB|}, \tag{4}$$

where $p(c,r)$ is a pixel of the proposal bounding box BB and $|BB|$ is the number of pixels covered by BB. BB having a high value of f_{int_area} are more likely to contain a pedestrian than BB with a low value of that ratio.

Overlap with the road plane and obstacles: these criteria encode the fact that a bounding box containing a pedestrian should not cover too many pixels belonging to the road or to obstacles such as buildings, fences or tree trunks. We derive two ratios f_g and f_o for the overlap of a BB with the road plane and the obstacles, respectively, to evaluate this constraint, using O_{mask2} to obtain f_o:

$$f_g = \frac{\sum_{p(c,r) \in BB}\ (G_{mask}(c,r))}{|BB|},\ f_o = \frac{\sum_{p(c,r) \in BB}\ (O_{mask2}(c,r))}{|BB|}. \tag{5}$$

Bounding box ratio: for a specific dataset, a pedestrian appearing in an image should have a size and a ratio between height and width within specific ranges; we assume this to be true even in the case of occlusions. This constraint is related to a two features f_{BBr} and f_{BBw}:

$$f_{BBr} = \frac{BB_{height}}{BB_{width}}; \qquad f_{BBw} = BB_{width}, \tag{6}$$

where BB_{width} and BB_{height} are width and height of the bounding box BB in image space.

Standing on the ground: this criterion is related to the knowledge that a pedestrian should stand on the ground. Even if just a pedestrian's upper part is visible (due to occlusion), the distance of the lowest points of the related BB part from the ground should not be too large. We define the average distance f_{feet} of the lower part of a BB from the ground to evaluate this criterion:

$$f_{feet} = \frac{1}{|BB_{feet}|} \sum_{p \in BB_{feet}} dist(p, \Omega), \tag{7}$$

where BB_{feet} is a square area in the centre of the lower boundary of the bounding box BB (Fig. 7).

Pedestrian height: this criterion is related to the knowledge about pedestrians' heights. The average distance f_{head} of the upper part of a BB from the ground is used as a feature:

$$f_{head} = \frac{1}{|BB_{head}|} \sum_{p \in BB_{head}} dist(p, \Omega)\ , \tag{8}$$

where BB_{head} is a square area in the centre of the upper boundary of the bounding box BB (Fig. 7).

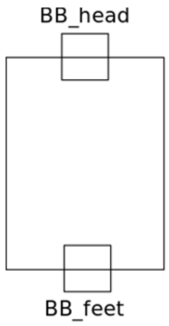


Fig. 7: Head (BB_head) and feet (BB_feet) area of a bounding box *BB*.

Depth consistency: a bounding box BB covering a person should contain a large number of pixels which correspond to that person and thus have similar depth values. This criterion is represented as the highest percentage of pixels having similar depth in BB. We use the depth histogram of all pixels in BB to derive the corresponding value f_{depth} :

$$h_{depth}(\mathrm{i}) = \frac{1}{|BB|}\sum_{i}^{n} 1(p \in BB, i * bin_{size} < depth(p) \leq (i+1) * bin_{size}) , \qquad f_{depth} = \max(h_{depth}) , \tag{9}$$

where $h_{depth}(\mathrm{i})$ is the histogram entry the i^{th} bin, bin_{size} is the size of each histogram bin, n is the number of bins, and $depth(p)$ is the depth value of pixel p.

Scoring function: our scoring function uses the features f_i described in Eqs. (3) to (8) to compute a score for each bounding box BB delivered by EB 70. We assume these features f_i to be independent and to follow normal distributions. Thus, our scoring function $p(C_{ped}|F)$ delivering a likelihood for a *BB* to contain a pedestrian given the feature vector *F* which contains all the features f_{i} defined earlier in this section is based on the following probabilistic model:

$$\begin{aligned} p(C_{ped}|F) &\propto p(C_{ped})p(F|C_{ped}), \\ &\propto p(C_{ped}) \prod_{i=1}^{n} p(f_i|C_{ped}) , \\ &\propto \prod_{i=1}^{n} p(f_i|C_{ped}) . \end{aligned} \tag{10}$$

$$p(f_i|C_{ped}) = \frac{1}{\sqrt{2\pi\sigma^2_{f_i(C_{ped})}}} e^{-\frac{(f_i - \mu_{f_i(C_{ped})})^2}{2\sigma^2_{f_i(C_{ped})}}} . \tag{11}$$

In Eq. 10, $p(f_i|C_{ped})$ is a likelihood function for f_i assuming that the BB contains a pedestrian. The factorisation can be made due to our assumption of the features to be independent. The prior $p(C_{ped})$ for a BB to contain a pedestrian is supposed to be uniform and, thus, is neglected. The index i indicates one of the features defined in Eqs. (3) to (8); n is the number of features evaluated. The individual likelihood terms are modelled as Gaussians (Eq. 11), where $\mu_{f_i(C_{ped})}$

and $\sigma^2_{f_i(C_{ped})}$ are mean and variance of the features f_i for a BB supposed to contain a pedestrian. They are determined in a training step. We rank the BB according to the scoring function $p(C_{ped}|F)$ and select the N_{opt} best BB as the final proposed regions.

4 Experimental results

Dataset: We evaluate our approach on the Kitti object detection benchmark (GEIGER ET AL., 2012), which has 7481 images in the training set. For these images, BB containing pedestrians are available. Just as the benchmark organisers, the BB are split into three sets (*hard*, *moderate*, *easy*) according to the level of difficulty to be expected for detecting the respective person based on the BB size, occlusion, and truncation levels in the input image. As the ground truth labels are not provided for the test set of the benchmark, we use only the training set for our evaluation. In order to be able rank our proposal, we first need to determine the means and variances of the normal distributions used to compute the ranking score according to Eqs. (10) and (11). Consequently, we apply cross-validation for the evaluation: we split the available images into five independent sets. In each test run, we use four sets (80% of the data) for learning the parameters of the score function and 20% for testing. We repeat this procedure five times, each time using a different image set for testing, so that in the end, each image contributes to the test set once. We report combined evaluation metrics over all test runs.

We compare the performance of our new techniques to other state-of-the-art proposal methods, namely BING, SS, EB, MCG-D, and 3DOP from (CHEN et al., 2015). Among these methods, BING, SS, and EB take RGB images as input, whereas MCG-D and 3DOP use depth as an additional cue.

Evaluation metrics: The quality of a BB proposal is estimated using its intersection over union (IoU) with the ground truth:

$$\mathrm{IoU}(A,B) = \frac{A \cap B}{A \cup B}$$

Following the criteria of the Kitti benchmark (GEIGER ET AL., 2012), a BB proposal is counted as a true positive if its IoU is equal to or larger than 0.5. We are mostly interested in the *recall*, which is the percentage of BB containing pedestrians in the reference that were detected by our proposal method and, thus, correspond to true positives. We present recall as a function of the IoU threshold and assess the impact of the number N_{opt} of selected BB proposals on the results.

Parameters settings: To estimate the obstacle masks O_{mask1} and O_{mask2}, we set the values of the two thresholds corresponding to the intensity of vertical disparity image to 40 and 200, respectively. We set the value ε in equation (2) to 0.2 (m). Further, a pedestrian is supposed not to exceed 2 meters in height. The side lengths of the squares containing the head and feet areas in Eqs. (7) and (8) (Fig. 7) are set to $1/3$ of the width of the bounding box. For estimating the depth constancy feature, we use a bin size of 1 meter for the depth histogram (see Eq. 9).

Results: We use the EB 70 method (ZITNICK & DOLLÁR 2014) to generate initial BB and select 20000 best candidates per input image based on the objectness score. The average recall of pedestrians with IoU of 0.5 or better are 99.6%, 96.3%, and 88.6% for the *easy*, *moderate*, and *hard* sets, respectively; these numbers are upper bounds for the recall that our method can achieve because we do not define new BB proposals. Fig. 8 shows the recall of EB 70 and our method as a function of the IoU threshold. The figure shows that our ranking function works very well on the *easy* and *moderate* sets. When ranking the BBs according to our scoring function and selecting the best N_{opt} = 2000 proposals, we discard hardly any correct BB for the *easy* and *moderate* sets, so that we achieve nearly the same recall as EB 70. Thus, we can reduce the number of region proposals delivered by EB 70 by a factor of 10 while losing only a very small percentage of regions corresponding to pedestrians. For the *hard* set, our method is not as successful; only when using the N_{opt} = 5.000 best proposals we can achieve a recall close to EB 70 (Fig. 8); however, this still would reduce the number of BBs to be inspected by more complex processes by a factor of 4. Fig. 9 shows recall as a function of N_{opt}. Again it shows that using N_{opt}= 5.000 results in a reasonably good performance independently from the level of difficulty. In contrast, the state-of-the-art methods need 10.000 proposals to achieve approximately 90% recall for *easy* and *moderate* sets. To obtain our recall at 1.000 proposals, those methods need at least 5.000 bounding boxes (see Figs. 9 and 10). Only 3DOP outperforms our method when N_{opt} is lager than 1.000. This is mainly because they focus on 3D object proposals using deep learning techniques together with additional data. However, when N_{opt} is smaller than 1.000, our method produces even better result than the 3DOP.

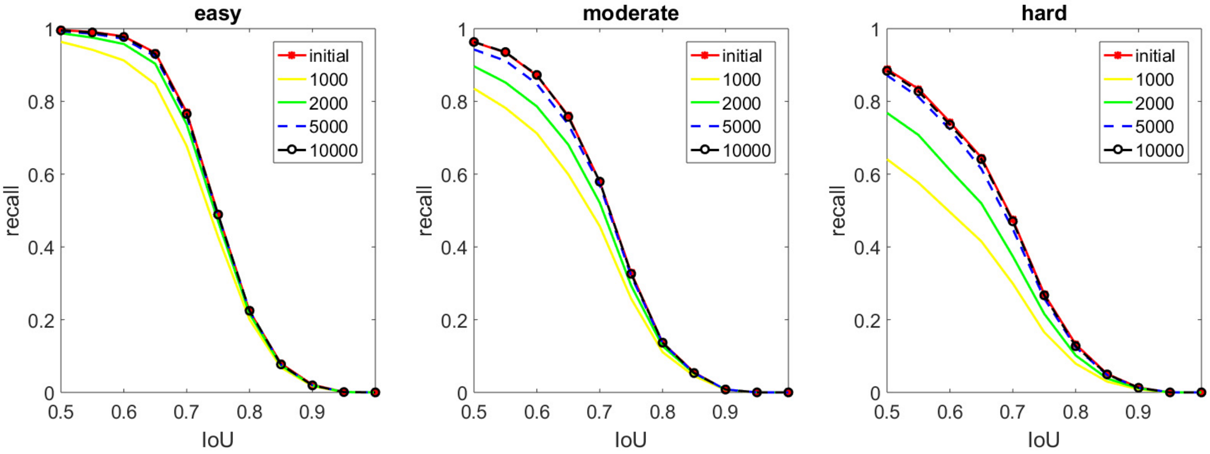


Fig. 8: Recall as a function of the IoU threshold. Init: Evaluation of the 20000 proposals generated by EB 70. The remaining curves were derived for different values for N_{opt} (1000, 2000, 5000, 10000) for selecting the final candidates after ranking.

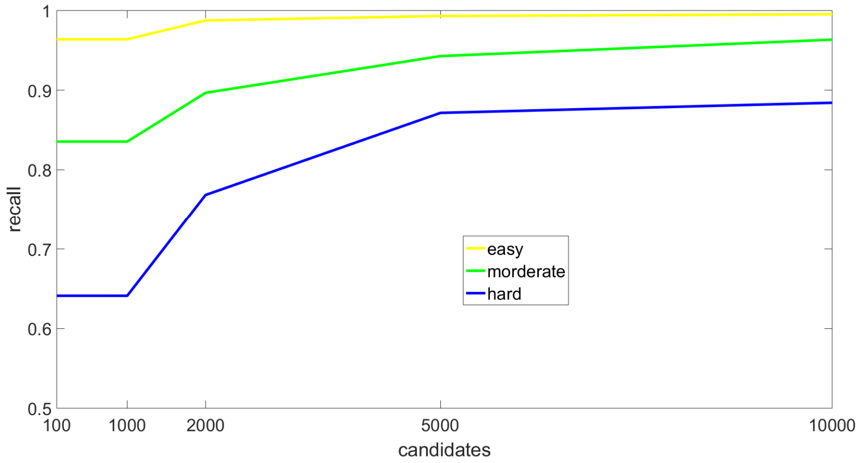


Fig. 9: Recall vs. number of candidates for pedestrian proposals on Kitti dataset of our approach.

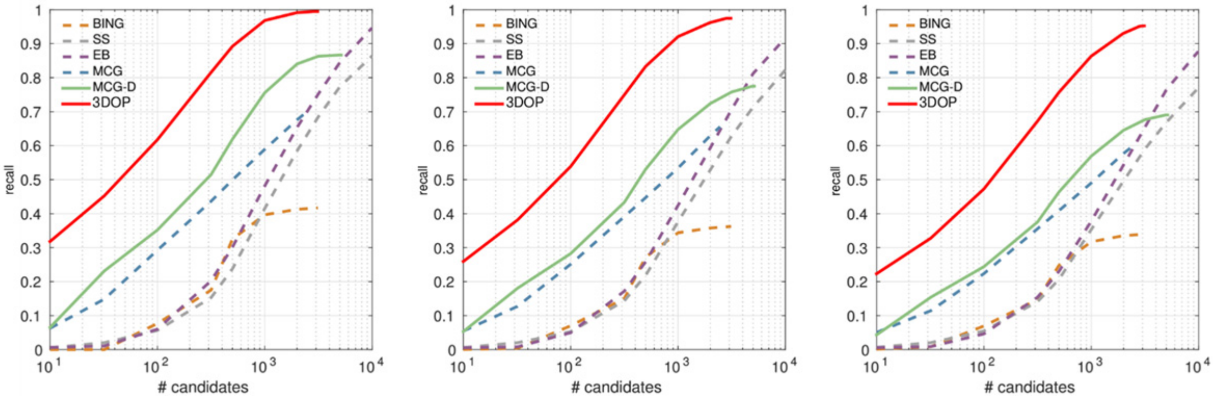


Fig. 10: Recall as a function of the number of candidates for pedestrian proposals on Kitti dataset of state of the art methods: BING, SS, EB, MCG, MCG-D, 3DOP.The graphs are adapted from CHEN et al. (2015).

5 Conclusion

In this paper, we have presented our simple but efficient framework for proposing regions that may contain pedestrians. It combines both, RGB and depth cues in two stages: initial proposal generating and filtering with a new score function. The experimental results show that by exploiting additional 3D cues, our approach can produce promising results and outperform most of the state-of-the-art methods. Specifically, we can achieve approximately 90% recall for the *easy* and *moderate* sets of pedestrians of the Kitti benchmark with 2.000 region proposals, and we can achieve reasonably good results when using 5000 proposals, thus reducing the computational costs of subsequent methods by a factor of 10 and 4, respectively. Moreover, with a very small number of proposals (less than 1.000), our proposal framework can achieve even better recall than the state-of-the art method 3DOP which is based on a sophisticated deep learning method. In the future, we will take advantage of CNN and our proposal method to develop a complete pedestrian detector, which will form the basis of pedestrian tracking.

Reference

ALEXE, B., DESELAERS, T. & FERRARI, V., 2012: Measuring the Objectness of Image Windows. IEEE Transactions on Pattern Analysis and Machine Intelligence, **34**(11), 2189-2202.

ARBELAEZ, P., PONT-TUSET, J., BARRON, J.T., MARQUES, F. & MALIK, J., 2014: Multiscale Combinatorial Grouping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 328-335.

CHEN, X., KUNDU, K., ZHU, Y., BERNESHAWI, A.G., MA, H., FIDLER, S. & URTASUN, R., 2015: 3D Object Proposals for Accurate Object Class Detection. Advances in Neural Information Processing Systems, 424-432.

CHENG, M. M., ZHANG, Z., LIN, W.Y. & TORR, P., 2014: BING: Binarized Normed Gradients for Objectness Estimation at 300fps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3286-3293.

DENG, J., DONG, W., SOCHER, R., LI, L.J., LI, K. & FEI-FEI, L., 2009: Imagenet: A Large-scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 248-255.

DOLLÁR, P., & ZITNICK, C.L., 2013: Structured forests for fast edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, 1841-1848.

EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C.K., WINN, J. & ZISSERMAN, A., 2010: The Pascal Visual Object Classes (voc) Challenge. International Journal of Computer Vision, **88**(2), 303-338.

GEIGER, A., LENZ, P. & URTASUN, R., 2012: Are We Ready for Autonomous Driving? the Kitti Vision Benchmark Suite. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 3354-3361.

GIRSHICK, R., 2015: Fast R-CNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1440-1448.

GIRSHICK, R., DONAHUE, J., DARRELL, T. & MALIK, J., 2014: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580-587.

GUPTA, S., GIRSHICK, R., ARBELÁEZ, P. & MALIK, J., 2014: Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In European Conference on Computer Vision, 345-360.

HARZALLAH, H., JURIE, F. & SCHMID, C., 2009: Combining Efficient Object Localization and Image Classification. The 12th International Conference on Computer Vision, 237-244.

HOSANG, J., BENENSON, R., DOLLÁR, P. & SCHIELE, B., 2016: What Makes for Effective Detection Proposals? IEEE Transactions on Pattern Analysis and Machine Intelligence, **38**(4), 814-830.

HU, Z. & UCHIMURA, K., 2005: U-V-disparity: An Efficient Algorithm for Stereovision Based Scene Analysis. The IEEE Intelligent Vehicles Symposium, 48-54.

REDMON, J., DIVVALA, S., GIRSHICK, R. & FARHADI, A., 2016: You Only Look Once: Unified, Real- time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779-788.

REN, S., HE, K., GIRSHICK, R. & SUN, J., 2015: Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. In Advances in Neural Information Processing Systems, 91-99.

UIJLINGS, J.R., VAN DE SANDE, K.E., GEVERS, T. & SMEULDERS, A. W., 2013: Selective Search for Object Recognition. International Journal of Computer Vision, **104**(2), 154-171.

VEDALDI, A., GULSHAN, V., VARMA, M. & ZISSERMAN, A., 2009: Multiple Kernels for Object Detection. The 12th International Conference on Computer Vision, 606-613.

WANG, X., YANG, M., ZHU, S. & LIN, Y., 2013: Regionlets for Generic Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, 17-24.

YAMAGUCHI, K., MCALLESTER, D. & URTASUN, R., 2014: Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation. In European Conference on Computer Vision, Springer, 756-771.

ZITNICK, C.L. & DOLLÁR, P., 2014: Edge Boxes: Locating Object Proposals from Edges. In European Conference on Computer Vision, Springer, 391-405.