3D-Guided Multiscale Sliding Window for Pedestrian Detection

Alejandro Gonzalez, Gabriel Villalonga, German Ros, David Vázquez, and Antonio M. López
{agalzate,gvillalonga,gros,dvazquez,antonio}@cvc.uab.es

Computer Vision Center and Universitat Autònoma de Barcelona
Bellaterra, Barcelona, Spain.

Abstract. The most relevant modules of a pedestrian detector are the candidate generation and the candidate classification. The former aims at presenting image windows to the latter so that they are classified as containing a pedestrian or not. Much attention has being paid to the classification module, while candidate generation has mainly relied on (multiscale) sliding window pyramid. However, candidate generation is critical for achieving real-time. In this paper we assume a context of autonomous driving based on stereo vision. Accordingly, we evaluate the effect of taking into account the 3D information (derived from the stereo) in order to prune the hundred of thousands windows per image generated by classical pyramidal sliding window. For our study we use a multi-modal (RGB, disparity) and multi-descriptor (HOG, LBP, HOG+LBP) holistic ensemble based on linear SVM. Evaluation on data from the challenging KITTI benchmark suite shows the effectiveness of using 3D information to dramatically reduce the number of candidate windows, even improving the overall pedestrian detection accuracy.

1 Introduction

Pedestrian detection is a key technology for many applications related to safety (e.g., autonomous driving) and security (e.g., video-surveillance). In this paper we focus on the context of autonomous driving, relying on stereo vision. Thus, we want to detect pedestrians using 2D image information as well as the 3D information provided by processing the stereo images, here assuming still images.

The most relevant modules of a *pedestrian detector* are *candidate generation* and *candidate classification*. The former presents image windows to the latter so that they are classified as containing a *pedestrian* or *background*. In the pedestrian detection literature much attention has being paid to the classification module in terms of image descriptors, classifiers, and models. As a result different concepts are nowadays of common use. For instance, descriptors such as Haar, EOH, HOG, and LBP; classifiers such as SVM, AdaBoost, Random Forest, CNNs; and models such as holistic, deformable part-based, and patch-based ensembles. All these concepts can be applied to different image modalities such as RGB, far infrared, and 3D-based stereo. For an in deep review the reader can check [6].

In comparison, candidate generation has mainly relied on a dominant approach, namely, the popular (multiscale) sliding window pyramid, which normally generates hundred of thousands candidates (image windows) even for images with VGA resolution. Therefore, candidate generation is critical for achieving real-time, a mandatory requirement to fulfil in driver assistance and autonomous driving. Of course, a faster computation of the particular descriptors is also a proper direction towards real-time (e.g., see [2] for HOG or [3] for Integral Channels). However, each new descriptor will require its own optimization. Therefore, an orthogonal approach (yet complementary) is to design a candidate generation scheme able to generate a relatively small number of candidates, without harming the detection accuracy and even improving such accuracy by directly discarding tricky candidates that may confuse the classifier.

In this line, here we evaluate the effect of taking into account 3D information (derived from stereo) in order to dramatically prune the hundred of thousands windows per image generated by classical sliding window pyramid. We remark that, in our application context, such information is not used in exclusive for pedestrian detection but also for other tasks such as the navigation of an autonomous vehicle [11]. Thus, it does not involves an extra cost. In addition, it allows to have an accurate distance estimation for the detected pedestrians.

For our study we use a multi-modal (RGB, disparity map) and multi-descriptor (HOG, LBP, HOG+LBP) holistic ensemble based on linear SVM. Evaluation on data from the challenging KITTI benchmark suite [5] shows how, in deed, our proposal (based on 3D information) dramatically reduces the number of candidate windows, and even significantly improves the overall detection accuracy.

2 Related Work

Beyond the sliding window pyramid there are other methods for 2D candidate selection [9]. They can provide accurate object detectors depending on the number of classes under consideration and the typical sizes covered. However, these methods tend to be very time consuming since they normally rely on some sort of segmentation or classification procedure (e.g., see the selective search [13] and edge boxes [15], two of the best methods according to [9]). In a way, candidate generation and classification rely on the same kind of information, i.e. visual appearance.

When 3D information is available geometric constraints can be applied [1], [7]. In short, pedestrians must be standing at the ground plane. In [7] candidate generation is based on detecting the road plane from the 3D information, and then uniformly distribute candidate windows sitting on the road and projecting them to the image plane (e.g. to the left image of the stereo pair). We call this approach *Linear-To-Road* (LtR) strategy. Moreover, after classifying the candidate windows, for those passing the pedestrian-test, a further post-processing is performer to remove *incoherent* detections. Here, incoherent refers to the fact that the 3D data corresponding to the image pixels contained in a 2D candidate window (considered a *detection*) should be consistent with the 3D window posi-

tion and that the window content fulfils the pedestrian size constraints. In [1], candidates are generated according to a clustering based on 3D point density. Then, starting from the 2D window corresponding to each 3D cluster, a set of neighbour windows are generated to have 15 windows for each cluster.

In this paper, we investigate LtR strategy and also an alternative called *Linear-to-Image* (LtI). LtI is based on the combination of sliding window pyramid and the effective use of depth. In short, each candidate window is back-projected to 3D and accepted as a candidate if it *hits* the road surface and agrees with pedestrian size. Therefore, we work densely in 2D and more sparsely in 3D, but with respect to [1] we are more robust to 3D clustering errors, and with respect to [7] we avoid the post-processing step for rejecting false positives. Our results, show better performance for LtI with respect to LtR regarding processing time without losing detection accuracy. In addition, both LtR and LtI show better detection accuracy than the sliding window pyramid.

3 Candidate Proposal via Structural Constraints

One of the biggest bottle necks on detection strategies is the generation of a set of candidates to be classified. A classical way is to evaluate all possibilities in an exhaustive fashion (sliding window). However, due to the inefficiency of sliding window, more sophisticated techniques for object proposal have gained popularity. These techniques propose candidates showing certain degree of "objectness", and can be seen as a more basic detection system, using low-level features. Some example of this trend are SelectiveSearch [12] and the work by Gu et. al [8] (see [9] for a complete survey on the topic). However, using candidate proposal based on "objectness" involves an extra level of computation, since a new (soft) classifier needs to be run over the entire image.

As an alternative, we propose to improve candidate generation by using structural information —that anyhow will be available— for the detector, i.e., the image disparity. Since 3D information has to be computed for other tasks involved in autonomous navigation, we can assume that image disparities are available at no extra cost. Then, we can exploit this information to establish constraints on the location of the candidate windows, i.e., pedestrians have to be touching the ground and meet clear size constraints. This process is explained in 3.1 and 3.2.

3.1 Ground Plane Estimation

The estimation of the ground plane is a fundamental step that allows for fast candidate rejection. The ground plane $\Pi_G = (\mathbf{n}^T, h)^T$ is here defined by its normal vector $\mathbf{n}$ and its height h with respect to the camera center. This magnitudes can be easily computed through the V-disparity map. Given a disparity map $D_{i,j} \in \{0, 255\}^{m \times n}$, its V-disparity V_D is an $m \times n$ matrix representing the normalized distribution of disparities per row. This can be formally expressed as follows:

$$V_D = \begin{bmatrix} H(D_{m,*})/\max(H(D_{m,*})) \\ \vdots \\ H(D_{1,*})/\max(H(D_{1,*})) \end{bmatrix} \in [0,1]^{m \times 256}, \quad (1)$$

where $H(\cdot) : \{0, 255\}^n \rightarrow \mathbb{R}^{256}$, is a function computing a histogram and $D_{i,*}$ is a n-vector with the information of the i-th row. From V_D one can estimate $(\mathbf{n}^T, h)^T$ by following Labayrade's proposal [10]. To achieve this, a Hough transform is applied to V_D to estimate the dominant line, $\mathfrak{l}$. Then, the estimation of $(\mathbf{n}^T, h)^T$ is given by

$$\theta = \arctan\left(c_v - v_{r_0}/f\right) \tag{2}$$

$$h = B\cos\left(\theta\right)/\bar{\theta} \tag{3}$$

$$\mathbf{n} = [X, Y, Y\cos(\theta)/h\sin(\theta)]^T, \quad \text{for arbitrary X, Y,} \tag{4}$$

assuming that the intrinsic parameters of the camera, c_v (v-coordinate of the principal point), f (camera focal length) and the baseline B of the stereo rig are known; and that $\bar{\theta}$ is the slope of $\mathfrak{l}$ and v_{r_0} is the row for which $\mathfrak{l}$ has zero disparity in V_D.

3.2 Fast Candidate Rejection

As mentioned before, it is critical to reduce the number of candidates for classification, in order to achieve real-time capabilities. Here we propose two alternative strategies that make use of the plane Π_G to reject unlikely candidates.

Linear-to-Road strategy The *Linear-to-Road* (LtR) policy, consists of positioning candidate windows directly in the 3D scene. To this end, we establish a practical operation range for $X = \pm 20$ meters, $Y = -h$ meters (i.e., on the ground) and $Z = 1, \ldots, 50$ meters with an increment of 3.5 meters. At each specific location $(X, -h, Z)$, we set three candidate windows with respective heights of 1.50, 1.75 and 2.00 meters (see Fig. 1-Top). These parameters give us an effective and practical coverage of the scene, while still producing good results as show in section 5. However, it is possible to increase the efficiency of the LtR candidate search at the cost of a minimal accuracy loss, by using a second strategy called Linear-to-Image (LtI).

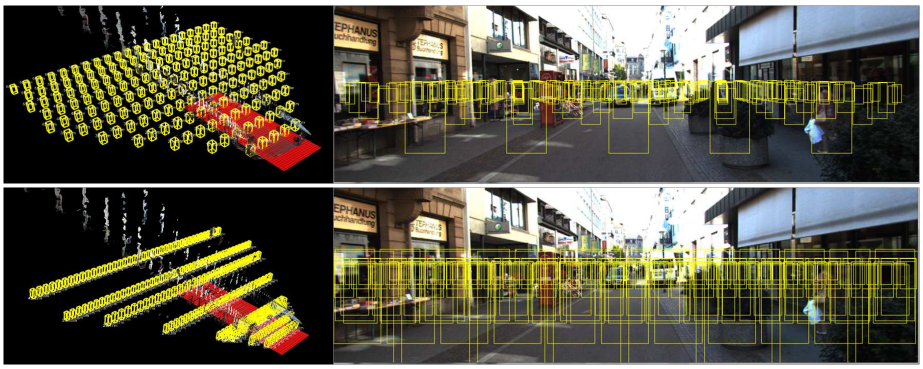

Fig. 1. Example of candidate generation for *Linear-to-Road* (top) and *Linear-to-Image* (bottom). Notice how in the case of the Linear-to-Road strategy, windows are uniformly distributed in 3D, while for Linear-to-Image candidates are not uniformly distributed due to perspective distortion.

Linear-to-Image strategy This second strategy is based on the combination of a classical sliding window pyramid (SWP) and the effective use of depth to generate candidates at multiple scales. First, SWP generates a large amount of candidates across all the scales. Each candidate is back-projected to 3D by using the available disparity information and the known calibration parameters. However, this step requires special care due to the presence of noise on the stereo data. In order to be robust to noise we select N_p random points $\{(u,v)^{(i)}\}_{i=1}^{N_p}$ within the given bounding box. The depth of each point $(u,v)^{(i)}$ is independently computed as $Z_i = Bd_i/f$, where $d_i = D_{v,u}$ is the disparity of that point. Then, each of the depths $\{Z_i\}_{i=1}^{N_p}$ is used to back-project the four corners defining the bounding box $\{(u',v')^{(t,l)}, (u',v')^{(t,r)}, (u',v')^{(b,l)}, (u',v')^{(b,r)}\}$, as follows

$$\begin{aligned} Y_i^{(j)} &= Z_i(v'^{(j)} + c_v)/f \\ X_i^{(j)} &= Z_i(u'^{(j)} + c_u)/f, \quad \text{for } j = (t,l),\ (t,r),\ (b,l),\ (b,r). \end{aligned} \tag{5}$$

Then, for each of the N_p hypotheses we check if the bottom of the back-projected bounding box is touching the ground, i.e., $\frac{1}{2}(|Y_i^{(b,l)}| + |Y_i^{(b,r)}|) < \epsilon$; otherwise, the candidate is rejected. If the size of the bounding box exceeds a given maximum size, i.e., $||Y_i^{(t,l)} - Y_i^{(b,r)}||_{\ell_2} > \max_S$ the candidate is also reject. These two criteria are able to reduce the amount of candidates dramatically, while maintaining a good recall and being highly efficient (see Fig. 1-Bottom).

4 Ensemble of Multi-modal Features

The proposed classification stage is inspired on one of the most accepted pipelines of the literature [14], i.e., a holistic detector based on the combination of HOG and LBP as input features and linear SVM as the learning method. Such an approach has shown to be both efficient and accurate, and it is the starting point for more modern and sophisticated detectors [4].

Here, we propose to boost the accuracy of this basic classifier by performing an ensemble over multiple image modalities, which in our case are the standard RGB space and the space of disparities. The idea is as follows. We consider the candidates proposed by one of the strategies introduced in section 3.2. These candidates guide the process of features extraction for HOG and LBP on both modalities, originating four different set of features. Such features are just extracted there where candidates are present, avoiding to perform dense feature extraction. This process can be run in parallel —since the image modalities and the features are assumed to be independent— speeding up the process four times.

Then, for each type of feature (i.e., HOG, LBP and HOG-LBP) and a given modality (i.e., RGB and Disparity), a linear SVM, $\mathcal{W}_{\text{feat}}^{\text{mod}}$, is trained. During testing time, the ensemble is performed by applying all the classifiers to a given sample S_i and then combining their outputs via the direct application of the max function,

$$\max(\mathcal{W}_{\text{HOG}}^{\text{RGB}} S_i, \mathcal{W}_{\text{LBP}}^{\text{RGB}} S_i, \mathcal{W}_{\text{HOG-LBP}}^{\text{RGB}} S_i, \mathcal{W}_{HOG}^{\text{Disp}} S_i, \mathcal{W}_{\text{LBP}}^{\text{Disp}} S_i, \mathcal{W}_{\text{HOG-LBP}}^{\text{Disp}} S_i). \tag{6}$$

We would like to highlight that, this simple way of performing the ensemble on the raw output of linear SVMs, although rough at first sight fulfil all our goals. It is very fast and turns out to increase the final accuracy as we will show in next section.

5 Experimental Evaluation

In this section we assess the benefits of the proposed approach, analysing the impact that each of the ingredients has in both, the accuracy and the computational performance of the final detector. For comparison, we have chosen a consolidated baseline consisting in HOG-LBP linear-SVM [14] in combination with a spatial window pyramid, which is one of the most popular techniques for holistic pedestrian detection. To this purpose we make use of the object detection dataset of the challenging KITTI benchmark suite [5], using 3738 samples for training and 3740 for testing through all our experiments.

5.1 Candidate Proposal Assessment

For evaluating the different candidate proposal techniques we have considered a metric based on the number of false positives per image (FPPI) and the miss rate of the final detector. For ease of comparison we run all the techniques, i.e., *Linear-to-Road* (LtR), *Linear-to-Image* (LtI) and the classic sliding sindow pyramid, for training detectors on just the RGB data modality and HOG features, using a linear-SVM. We also evaluate the number of candidates generated by each method in order to have a clear indicator of the computational efficiency of the resulting detector. This comparison is shown in Fig. 2.

Notice that both proposed strategies, LtR and LtI, reduce the miss rate of the final detector in more than 7 points when compared against the standard SWP. For this experiment SWP has been set up to use 6 scales with factors $f_s = \{1, 1.14, 2, 2.8, 4, 5.6\}$, generating 673×10^3 candidates (on average). LtR can drastically reduce this number to an average of 246×10^3 candidates (36% of candidates w.r.t. SWP) to produce the most accurate results. However, it seems very convenient to select the LtI strategy instead, since for the price of a small sacrifice in accuracy (less than one point) we just need to examine 75×10^3 windows on average (11% of candidates w.r.t. SWP).

5.2 Features Assessment

Our second experiment measures how the use of different types of features and image modalities affects detection accuracy. In this case all the candidates are generated by following the *Linear-to-Image* strategy. Fig. 3 shows detection results with respect to FPPI and miss rate for the different modalities (RGB, Disparity and RGB+Disparity). As one can observe, the ensemble of HOG-LBP features over multiple modalities (**12.37**% of miss rate) leads to better results than using single modalities. This represents an improvement of almost 2 points with respect to the baseline (HOG-LBP linear-SVM on RGB), which obtained a miss rate of 14.10%.

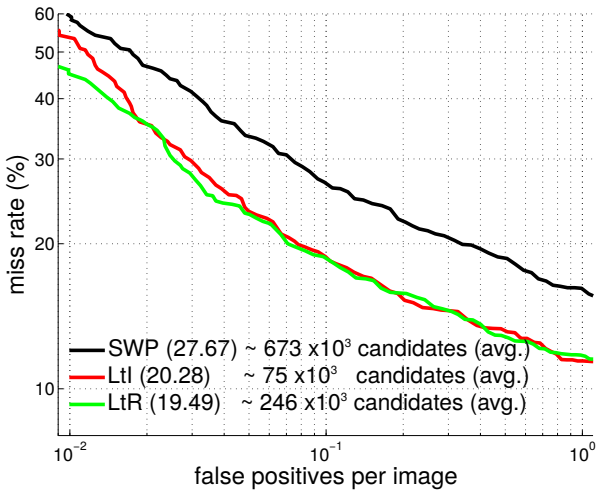

Fig. 2. Comparison of candidate proposal strategies: *Linear-to-Road* (LtR), *Linear-to-Image* (LtI) and the classic sliding window pyramid (SWP), with respect to the number of false positives per image (FPPI) and the detection miss rate, for a HOG linear-SVM detector applied on the RGB modality of the KITTI benchmark suite.

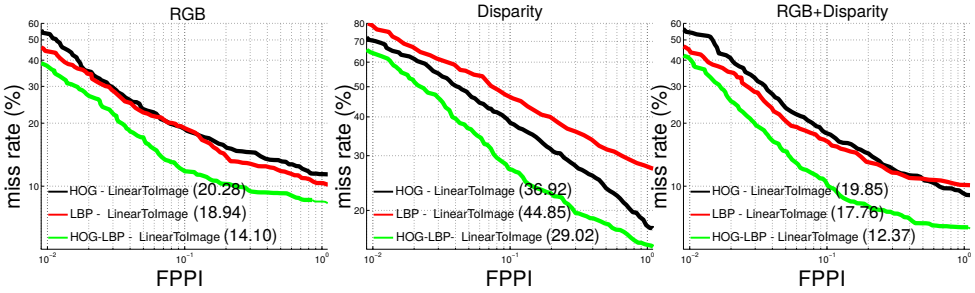

Fig. 3. Feature comparison (HOG, LBP and HOG-LBP) at the level of FPPI and miss rate for three image modalities: RGB (left), Disparity (middle) and RGB-Disparity (right).

6 Conclusion

We have presented an approach that exploits 3D structure to dramatically reduce the number of candidate windows for pedestrian detection, making use of the *Linear-to-Image* strategy. Furthermore, 3D structure also served to improve the overall pedestrian detection accuracy. This is due to the LtI strategy itself, as it discards tricky samples, but also due to a proposed multi-modal (RGB, disparity) and multi-descriptor (HOG, LBP, HOG+LBP) holistic ensemble based on linear SVM. We showed the superiority of our technique by comparing it against one of the most accepted baselines, i.e., the HOG-LBP SVM detector, in the challenging KITTI benchmark suite.

Acknowledgments

This work is supported by the Spanish MICINN projects TRA2011-29454-C03-01 and TIN2011-29494-C03-02.

References

1. Alonso, I.P., Llorca, D.F., Sotelo, M.A., Bergasa, L.M., Toro, P.R.d., Nuevo, J., Ocana, M., Garrido, M.A.: Combination of feature extraction methods for svm pedestrian detection. Trans. Intell. Transport. Sys. (2007)
2. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA (2012)
3. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: Proc. of the British Machine Vision Conference, London, UK (2009)
4. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Mahine Intelligence **32**(9) (2010) 1627–1645
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR'12. (2012)
6. Geronimo, D., Lopez, A.: Vision-based pedestrian protection systems for intelligent vehicles. Springer Briefs in Computer Science (2013)
7. Gerónimo, D., Sappa, A., Ponsa, D., López, A.: 2D-3D based on-board pedestrian detection system. Journal of Computer Vision and Image Understanding **114**(5) (2010) 583–595
8. Gu, C., Lim, J.J., Arbelez, P., Malik, J.: Recognition using regions. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2009)
9. Hosang, J., Benenson, R., Schiele, B.: How good are detection proposals, really? In: Proc. of the British Machine Vision Conference. (2014)
10. Labayrade, R., Aubert, D., Tarel, J.P.: Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation. In: IEEE, Intelligent Vehicle Symposium. (2002)
11. Ros, G., Ramos, S., Granados, M., Bakhtiary, A., Vazquez, D., Lopez, A.: Vision-based offline-online paradigm for autonomous driving. In: Winter Conference on Applications of Computer Vision (WACV). (2015)
12. Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. International Journal of Computer Vision (2013)
13. van de Sande, K.E.A., Uijlings, J.R.R., Gevers, T., Smeulders, A.W.M.: Segmentation as selective search for object recognition. In: Proc. of the IEEE International Conference on Computer Vision. (2011)
14. Wang, X., T.X. Han, Yan, S.: An HOG-LBP human detector with partial occlusion handling. In: Proc. of the IEEE International Conference on Computer Vision, Kyoto, Japan (2009)
15. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Proc. of the IEEE European Conference on Computer Vision. (September 2014)