Multiview Random Forest of Local Experts Combining RGB and LIDAR data for Pedestrian Detection

Alejandro González[1,2], Gabriel Villalonga[1,2], Jiaolong Xu[2], David Vázquez[1,2], Jaume Amores[2], Antonio M. López[1,2]
{agalzate,gvillalonga,jiaolong,dvazquez,jaume,antonio} @cvc.uab.es
[1]Autonomous University of Barcelona, [2]Computer Vision Center

Abstract— **Despite recent significant advances, pedestrian detection continues to be an extremely challenging problem in real scenarios. In order to develop a detector that successfully operates under these conditions, it becomes critical to leverage upon multiple cues, multiple imaging modalities and a strong multi-view classifier that accounts for different pedestrian views and poses. In this paper we provide an extensive evaluation that gives insight into how each of these aspects (multi-cue, multimodality and strong multi-view classifier) affect performance both individually and when integrated together. In the multimodality component we explore the fusion of RGB and depth maps obtained by high-definition LIDAR, a type of modality that is only recently starting to receive attention. As our analysis reveals, although all the aforementioned aspects significantly help in improving the performance, the fusion of visible spectrum and depth information allows to boost the accuracy by a much larger margin. The resulting detector not only ranks among the top best performers in the challenging KITTI benchmark, but it is built upon very simple blocks that are easy to implement and computationally efficient. These simple blocks can be easily replaced with more sophisticated ones recently proposed, such as the use of convolutional neural networks for feature representation, to further improve the accuracy.**

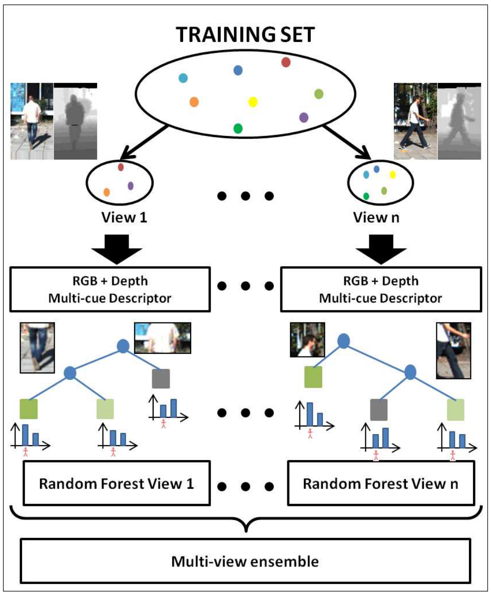

Fig. 1. Scheme Overview.

I. Introduction

Developing a reliable pedestrian detector enables a vast range of applications such as video surveillance and the practical deployment of autonomous and semi-autonomous vehicles. With more than a decade of history by now [1], pedestrian detection is still a very challenging task when applied under realistic conditions [2], [3], [4]. In order to obtain a detector that successfully operates under these conditions, it becomes critical to exploit sources of information along three orthogonal axis: i) the integration of multiple feature cues (contours, texture, etc.), ii) the fusion of multiple image modalities, and iii) the use of multiple views of the pedestrian by learning a strong classifier that accommodates for both different 3D points of view and multiple flexible articulations. In this paper we perform an extensive evaluation providing insight about how each of these three aspects affect accuracy, both individually and when integrated together.

In order to integrate multiple image modalities, we considered the fusion of dense depth maps with visible spectrum images. The use of depth information has gained attention thanks to the appearance of cheap sensors such as the one in Kinect, which provides a dense depth map registered with an RGB image (RGB-D). However, the sensor of Kinect has a maximum range of approximately 4 meters and is not very reliable in outdoor scenes, thus having limited applicability for pedestrian detection. On the other hand, Light Detection and Ranging (LIDAR) sensors such as the Velodyne HDL-64E have a maximum range of up to 50 meters and are appropriate for outdoor scenarios. Although they produce a sparse cloud of points and they are only recently starting to receive attention for application to pedestrian detection. In this work we analyze the use of registered pairs of RGB and Velodyne 3D point clouds.

Our analysis reveals that, although all the aforementioned components (the use of multiple feature cues, multiple modalities and a strong multi-view classifier) are important, the fusion of visible spectrum and depth information allows to boost the accuracy significantly by a large margin. The resulting detector not only ranks among the top best performers in the challenging KITTI benchmark, but it is built upon very simple blocks that are easy to implement and computationally efficient. Before explaining all the components of our system in detail, we provide a brief summary of each component and how they relate/differ to more recent work.

A. Multi-cue feature representation

In this work we integrate gradient-based features such as HOG [5], that provides a good description of the object contours, with texture-based features such as LBP [6]. These two types of features provide complementary information and the fusion of both types of features has been seen to

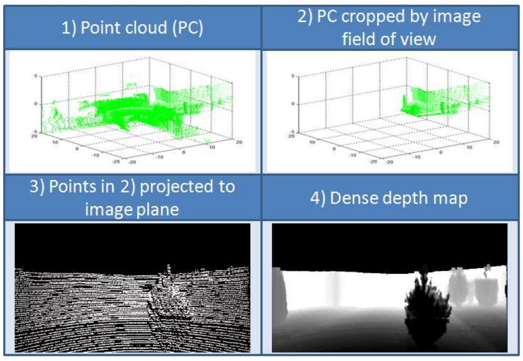
1) Point cloud (PC)
2) PC cropped by image field of view
3) Points in 2) projected to image plane
4) Dense depth map

Fig. 2. Dense Depth map generation scheme.

boost the performance of either feature separately [7], [8]. Both types of features are extracted for each image modality. We show that by appropriately choosing the parameters used in the computation of these features for each modality we can obtain an important gain in accuracy.

From the seminal work of Dalal and Triggs [5] it has been seen that using different types of gradient-based features and their spatial distribution, such as in the HOG descriptor [5] provides a distinctive representation of both humans and other objects classes. The integral channel features proposed by Dollar et al. [9] allows to integrate multiple kinds of low-level features such as the gradient orientation over the intensity and LUV images, extracted from a large number of local windows of different sizes and at multiple positions, allowing for a flexible representation of the spatial distribution. In this work and related ones [10], [11] it has been seen that including color boosts the performance significantly, being this type of feature complementary to the ones we used in this study. Context features have also been seen to aid [12], [13] and could be easily integrated as well. Exploring alternative types of spatial pooling of the local features in HOG or LBP is also beneficial as shown in [14] and is also complementary to the approach used in this paper.

B. Fusion of multiple image modalities

In this work we explore the fusion of dense depth maps obtained with RGB images. Dense depth maps are obtained by first registering the 3D cloud of points captured by a Velodyne sensor with the RGB image captured with the camera, and then interpolating the resulting sparse set of pixels to obtain a dense map where each pixel has an associated depth value [15]. Descriptors such as HOG and LBP are extracted from each modality independently. Following [16], the information provided by each modality can be fused using either an early-fusion scheme, at the feature level, or a late-fusion scheme, at the decision level. In our study, using a simple early fusion scheme, where descriptors from each modality are concatenated, provided the best results.

Pedestrian detection based on data coming from multiple modalities has been a relatively active topic of study [1], and in particular the use of 2D laser scanners and visible spectrum images has been studied in several works, for instance [17], [15]. Only recently authors are starting to study the impact of high-definition 3D LIDAR [15], [17], [18], [19], [20], [21], [22]. Most of these works propose specific descriptors for extracting information directly from the 3D cloud of points [17], [18], [19], [20], [21], [22]. A common approach is to detect pedestrians independently in the 3D cloud of points and in the visible spectrum images, and then combining the detections using an appropriate strategy [18], [19], [22]. Following the steps of [15], we opt for a simple yet effective strategy where the 3D cloud of points is used to obtain a dense depth map. Given this map, well-known description techniques such as HOG and LBP can be applied in order to obtain a highly distinctive object representation. Our work differ from [15] in that we use multiple descriptors and adapt them to have a good performance in dense depth images. While [15] employs a late fusion scheme, our experimental analysis revealed that using an early fusion scheme provided best results in the given multi-cue, multi-modality framework.

C. Multi-view classifier

In this work we make use of Random Forests (RF) of local experts [23], which has a similar expressive power than the popular Deformable Part Models (DPM) [24] and less computational complexity. In this method, each tree of the forest provides a different configuration of local experts, where each local expert takes the role of a part model. At learning time, each tree learns one of the characteristic configurations of local patches, thus accommodating for different flexible articulations occurring in the training set. In [23] the RF approach consistently outperformed DPM. An advantage of the RF method is that only a single HOG or HOG-LBP descriptor needs to be extracted for the whole window, and each local expert re-uses the part of the descriptor that corresponds to the spatial region assigned to it. Its computational cost is further significantly reduced by applying a soft cascade, operating in close to real time. Contrary to the DPM, the original RF method learns a single model, thus not considering different viewpoints separately. In this work, we extend this method to learn multiple models, one for each 3D pose, and evaluate both the original single model approach and the multi model approach.

Several authors have proposed methods for combining local detectors [24], [25] and multiple local patches [26], [27], [28]. The method in [29] also makes use of RF with local classifiers at the node level, although it requires to extract many complex region-based descriptors, making it computationally more demanding than [23].

D. Multi-cue/modality/view representation

Most relevant to this paper is the approach presented in [30] where the authors combine multiple views (front, left, back, right), modalities (luminance, depth based on stereo, and optical flow), and features (HOG and LBP). The main differences between [30] and our work are as follows: i)in order to complement RGB information, we make use of a sensor modality, high-definition 3D LIDAR, which has received relatively little attention in pedestrian detection until now, and ii) while [30] makes use of an holistic classifier, we make use of a more expressive patch-based model, and iii)

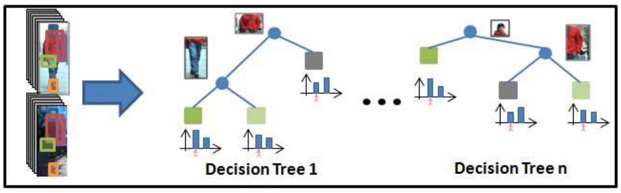

Fig. 3. Random Forest scheme.

in [30] multiples cues are combined following late-fusion style, while we consider also early-fusion, which, in fact, give better results in our framework.

The rest of the paper is organized as follows. In Sect. II we develop our proposal. Section III presents the experiments carried out to assess our proposal step by step, and discuss the obtained results. Finally, section IV draws our main conclusions.

II. Multiview RGBD-RF for pedestrian detection

We propose a complete framework in which our final model incorporates the multi-cue characteristic by extracting HOG and LBP descriptors. Also the multi-modal characteristic by extracting information from RGB and depth modalities, which will be combined at feature level (early fusion) or at decision level (late fusion). Finally we will use a multi-view model in which we will separate the problem into $n-$ views for combining them with late fusion (See Figure 1). In addition we will model pedestrians both holistically and as a set of relevant patches. In the former case the model will be learnt with linear SVM; and in the latter with a Random Forest of Local Experts.

A. Multi-cue feature representation

In order to improve the pedestrian detection accuracy it is widely used the incorporation of different cues or features. This incorporation looks for complementarity by using different cues for describing the same object. In order to incorporate different cues in our framework we use the HOG [5] descriptor (shape) and the LBP [31] descriptor (texture). Both descriptors are combined using an early fusion technique (concatenating HOG and LBP descriptor), obtaining a robust descriptor with complementary information (HOGLBP). HOG descriptor is composed by a histogram of gradient orientations. Given a window candidate the histograms are calculated on overlapped blocks inside it. LBP descriptor calculate histograms of texture patterns over the same overlapped blocks than HOG. This texture patterns are based on value differences between the central pixel and the surrounding ones in a 3×3 neighborhood. We use our own implementation that includes some modifications that improve the final detection rate. The first modification is included in the image pyramid construction. The image re-size process is done by bilinear interpolation with antialiasing,which helps the gradient calculation and thereby the HOG descriptor classification accuracy. The second modification is included in the LBP descriptor. When the value differences are calculated we accept as equal values the ones included in a defined range, this range (defined as $ClipTh$) allows that small noises (small value changes) do not affect the texture pattern (More details in [32]).

B. Fusion of multiple image modalities

Keeping in mind that more complementarity is better for pedestrian detection, we want to explore the integration of different modalities. Usually information is extracted from a single-modal sensor (RGB camera), but we combine this visual information with 3D information extracted from a LIDAR sensor. In order to transform the point cloud obtained using the LIDAR into a dense depth map, we follow the approach presented by Premediba *et al.* [15]. In this method, the 360° 3D point cloud from the LIDAR sensor is filtered in order to take only those points included in the viewfield of the RGB camera. The filtering process start by applying calibration matrices to project the 3D coordinates in image plane coordinates. Then the points that fall inside the image are selected, while the others are rejected. The filtered points form a sparse depth image, time and space synchronized with the visual image. At this step by defining a neighborhood for each valid pixel of the depth map we interpolate the information for filling the missing values. After this process, the pixels without depth information will be filled, ending up in a dense depth map (see Fig. 2).

At this point, for each candidate window we extract HOG and LBP features over each modality (visual and depth). Then, we combine these features into a single detector. There are two approaches for performing this combination. The first one is to use a late ensemble of detectors; in this case we train two separate detectors, one per modality. The second one is to combine at feature level the two modalities; in this case we train a single model using as descriptor the concatenation of the features computed at each modality.

C. Multi-view detectors

For facing the problem of intra-class variability, if we reduce the variability inside the target object, we will discriminate better this class from the background. One of the biggest causes of the large variability in object detection is the pose and orientation of the object. In order to solve this problem we propose to use a multi-view approach. Given a set of annotated pedestrians for training a detector, we propose to separate them into n different views depending on its orientation. We analyze the orientations of the training samples in order to define some thresholds that split the samples into different views. By spliting in this way the samples we can adjust the canonical size for each subset, allowing the final detector to deal with pedestrians in different orientation having each orientation its own aspect ratio (*e.g.* it is not the same bounding box for a frontal-viewed pedestrian than for a side-viewed pedestrian).

D. Pedestrian Model

In our study we focus on two different models: one holistic, where the whole pedestrian view is considered as the model; and a patch-based one where only a subset of pedestrian patches are considered as model. As holistic model we use the $descriptor/SVM$ with a linear kernel

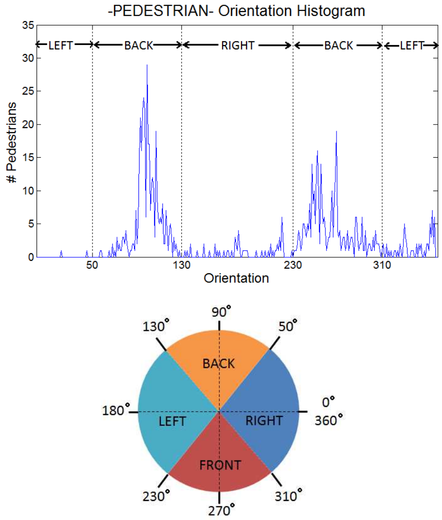

Fig. 4. Orientation Histogram and Distribution.

(*linSVM*) which has a good compromise between computation time and accuracy. As patch based model we use *descriptor*/*RandomForest*(*RF*). Each tree inside the *RF* has different configuration of patches (see Fig. 3), and at the end the decision is made by taking into account the different configuration learned in each tree. We will use the RF formed by 100 trees, 7 levels as max depth and each node in a tree will be a *linSVM* local expert (see [23] for details).

III. Experimental results

In this section we will evaluate each step of the proposed approach: multi-cue, multi-modal and multi-view. In order to fulfill this evaluation we will use our own implementation of HOG and LBP features, which provides significantly better results than the one proposed in [7], *i.e.*, removing the occlusion handling reasoning. Also we will use as classifier the SVM with linear kernel, and the Random Forest. Letting us with a bunch of possibles detectors: HOG/linSVM, LBP/linSVM, HOGLBP/linSVM, HOGLBP/RF. We will use as baseline for comparing the different steps the HOG/linSVM detector which is standard for pedestrian detection.

a) **KITTI Dataset:** in this paper we use the KITTI dataset that provides synchronized data obtained from two different sensors, a stereo-pair camera and a LIDAR. KITTI dataset for object detection includes 7481 training images and 7518 test images, comprising a total of 80.256 labeled objects. Annotations are provided only for the training set. For this reason we split the training set into a training set (the first 3740 images) and a validation set (the last 3741 images) as done in [15], these subsets are used for the evaluation of each step of our approach. The original training and testing set will be used for the detector configuration finally selected, *i.e.*, in order to compare with the state-of-the-art methods using the KITTI web page for submitting results on test set.

b) **Evaluation protocol:** As evaluation methodology we follow the de-facto Caltech standard for pedestrian detection [4], *i.e.*, we plot curves of false positives per image (FPPI) *vs* miss rate. The average miss rate (*AMR*) in the range of 10^{-2} to 10^{0} FPPI is taken as indicative of each detector accuracy, *i.e.*. the lower the better. Also we will evaluate using the KITTI evaluation framework in which the precision-recall curve is calculated for ranking the methods by the average precision (*AP*), *i.e.*, the higher the better. During training we consider pedestrian higher than 25 pixels not occluded (Reasonable subset). For testing we use the reasonable subset in the caltech evaluation and the KITTI evaluation is performed over 3 different subsets depending on height and occlusion level: *easy subset* (Min height: 40 px; max occlusion level: fully visible; max truncation: 15%), *moderate subset* (Min height: 25 px; max occlusion level: partly occluded; max truncation: 30%), *hard subset* (Min height: 25 px; max occlusion level: difficult to see; max truncation: 50%). This KITTI evaluation will be performed in the validation set and in the final testing set.

c) **Multi-cue:** We start by evaluating the gain obtained by using multiples cues, for that reason we start the evaluation by comparing the single-view (SV) detectors. However, first of all we have tuned the LBP parameters, getting $ClipTh_{RGB} = 4$ and $ClipTh_{Depth} = 0.2$. These parameters mean that, for calculating the texture pattern, we will treat as the same value those in the range on 4 luminance units for the RGB modality and 0.2 meters in depth. In Table I, comparing the SV experiments HOG/linSVM against HOGLBP/linSVM, we can see that the gain in *AMR* is around 12% with RGB modalities, around 4% with depth and around 2% when combining the two modalities. The same behavior can be seen also if we compare the SV-LBP/linSVM against the SV-HOGLBP/linSVM where we obtain improvements of around 10% , 3% and 4% respectively.

d) **Multi-modal:** Regarding the evaluation of the multi-modal approach, we compare the SV-HOG/linSVM detector over RGB and depth against its combination (RGB + depth). In order to select the best type of modality fusion we evaluate both the late and early fusion techniques. In Fig. 5 we can see that the early fusion method improves the performance (lower *AMR*) with respect to the late fusion for all the proposed models. Taking into account this fact we include in Table I only the modality fusion experiments based on early fusion technique. In Table I, comparing the SV experiments HOG/linSVM in RGB, Depth and RGB+Depth we can see how the multi-modal experiments outperform the single-modal ones obtaining an *AMR* gain of $\sim 18\%$ against RGB and $\sim 10\%$ against Depth. This behavior is repeated if we look at the different SV proposed models: LBP/linSVM ($\sim 14\%$ / $\sim 7\%$), HOGLBP/linSVM ($\sim 8\%$ / $\sim 8\%$) and HOGLBP/RF ($\sim 4\%$ / $\sim 4\%$).

e) **Multi-view:** In order to show the gain obtained by the introduction of a multi-view (MV) model we will compare the SV-HOG/SVM against the MV version. For pedestrian detection we propose a two view approach, one grouping the front (270°) and back (90°) orientations and the other one grouping the left (180°) and right (0°) orientations. Analyzing the orientation histogram (see Figure 4) for pedestrians included in the KITTI training set, we find out

TABLE I

Results using different subsets for training (Single-view (SV), Frontal (F), Lateral (L), Multi-view (MV)), modalities, and detectors, tested over the validation set. *AMR* is indicated for Caltech evaluation and *AP* for KITTI evaluation. Best accuracy for each detector in each modality is indicated in bold, while the best detector for each modality in red

Evaluation	Detector	RGB				Depth				RGB + Depth (Early Fusion)			
		SV	F	L	MV	SV	F	L	MV	SV	F	L	MV
AMR	HOG/SVM	50.29	55.58	62.67	**46.77**	42.17	46.71	45.23	**40.25**	32.49	38.28	33.12	**31.44**
	LBP/SVM	48.91	53.23	57.63	**45.37**	41.42	43.24	43.24	**38.51**	34.66	38.33	30.85	**29.85**
	HOGLBP/SVM	38.66	44.25	45.91	**35.66**	38.05	41.67	39.42	**36.14**	30.78	34.08	30.85	**28.33**
	HOGLBP/RF	30.05	39.95	41.76	**27.94**	30.47	36.06	34.10	**30.13**	25.99	32.49	30.44	**25.59**
AP (Easy)	HOG/SVM	0.50	0.44	0.36	**0.54**	0.59	0.54	0.57	**0.62**	0.71	0.65	0.70	**0.73**
	LBP/SVM	0.52	0.47	0.42	**0.56**	0.62	0.60	0.59	**0.65**	0.69	0.66	0.74	**0.75**
	HOGLBP/SVM	0.64	0.58	0.54	**0.68**	0.65	0.61	0.64	**0.67**	0.74	0.71	0.74	**0.76**
	HOGLBP/RF	0.73	0.62	0.59	**0.75**	0.74	0.67	0.70	**0.75**	0.79	0.73	0.74	**0.79**
AP (Moderate)	HOG/SVM	0.38	0.33	0.28	**0.41**	0.46	0.42	0.42	**0.47**	0.57	0.53	0.54	**0.58**
	LBP/SVM	0.41	0.37	0.34	**0.44**	0.48	0.46	0.46	**0.50**	0.57	0.54	0.59	**0.61**
	HOGLBP/SVM	0.50	0.46	0.43	**0.54**	0.51	0.47	0.49	**0.52**	0.61	0.58	0.60	**0.62**
	HOGLBP/RF	0.59	0.50	0.46	**0.60**	0.58	0.52	0.53	**0.58**	0.65	0.59	0.59	**0.66**
AP (Hard)	HOG/SVM	0.33	0.29	0.24	**0.35**	0.40	0.37	0.36	**0.41**	0.50	0.46	0.47	**0.51**
	LBP/SVM	0.36	0.32	0.29	**0.38**	0.42	0.40	0.40	**0.43**	0.50	0.47	0.52	**0.53**
	HOGLBP/SVM	0.43	0.40	0.37	**0.47**	0.45	0.42	0.43	**0.46**	0.53	0.51	0.52	**0.55**
	HOGLBP-RF	0.51	0.43	0.40	**0.52**	0.50	0.45	0.46	**0.50**	0.56	0.52	0.52	**0.57**

a separation between the four views: $310^\circ - 50^\circ$ for right, $50^\circ - 130^\circ$ for back, $130^\circ - 230^\circ$ for left and $230^\circ - 310^\circ$ for front. Once defined the orientation clusters we calculate the canonical aspect ratios (*AR*) for each view. This *AR* is important for defining the sliding window size, and in order to allow detection with different AR in the final ensemble. For the defined frontal (F) view, *i.e.*, (Front + Back) we obtain an $AR = 2.6$ and for the lateral (L) view, *i.e.*, (Left + Right) an $AR = 2.4$. In Table I are presented the results obtained by training the model using the different subsets based on each view (F/L) and the ensemble of them (MV), comparing the SV-HOG/SVM against its MV counterpart we obtain an *AMR* gain of $\sim 4\%$ (RGB), $\sim 2\%$ (Depth) and $\sim 1\%$ (RGB+Depth). The same behavior is obtained by comparing the other SV models against its MV counterpart: LBP/linSVM ($\sim 3\%$ / $\sim 3\%$ / $\sim 5\%$), HOGLBP/linSVM ($\sim 3\%$ / $\sim 2\%$ / $\sim 2\%$) and HOGLBP/RF ($\sim 2\%$ / $\sim 1\%$ / $\sim 1\%$).

f) **Discussion:** Each of the mentioned detectors is developed using RGB, Depth and Early Fusion (RGB + Depth) information sources in order to compare the accuracy under the different conditions. Also for evaluating the multi-view performance the experiments are carried out using a single-view (all samples) and a multi-view (samples divided in two views). In Table I there are the accuracy measurements over the validation set. The measurements include the Caltech evaluation methodology for reasonable pedestrians subset and the KITTI evaluation methodology for *easy*, *moderate* and *hard* pedestrian subset. Regarding the obtained results it is easy to see the performance improvements at each step of the proposed method. First we can see the improvement introduced by the RF over the other detectors. Comparing the results obtained in each column (training subset and information source) we obtain always the best accuracy in the HOGLBP/RF detector. The second improvement is introduced by the multi-view proposed method, comparing each row (detector) we obtain the best performance for each of the information sources (RGB, Depth, RGB+Depth) when we perform the multi-view ensemble classifier. The third improvement is introduced by the early fusion of information sources, in this case for each detector and given a training subset we obtain the best performance in the RBG+Depth experiment.

TABLE II

Evaluation and comparison of Multi-view RGBD RF detector using the final test set

Rank	Method	Moderate	Easy	Hard
1	Regionlets	61.15 %	73.14 %	55.21 %
2	**MV-RGBD-RF**	**54.56 %**	**70.21 %**	**51.25 %**
3	pAUCEnsT	54.49 %	65.26 %	48.60 %

Finally if we compare the baseline method SV-HOG/linSVM against our proposed multi-cue, multi-modal and multi-view Random Forest of Local Experts we obtain an *AMR* gain of $\sim 25\%$ in the validation set. Regarding the final approach MV-HOGLBP/RF early fusion of RGB and depth in Table II, we obtain an *AP* of 70.37%, 54.67%, 50.35% for the *easy*, *moderate* and *hard* subset respectively, ranking the second best pedestrian detector in the challenge. Fig. 5 shows the precision-recall curve obtained over each subset using the final approach.

IV. Conclusions

In this paper we develop a complete multi-cue, multi-modal and multi-view framework for pedestrian detection. We have shown the applicability to different models (holistic, patch-based), obtaining significant performance improvements. In this paper we focus on pedestrian detection using HOG/linSVM as baseline applying the different proposed method: different cues (HOG and LBP), diferent modalities (RGB and depth) and different views (Frontal and lateral), thus, our immediate future work will focus on detection of different classes like cyclist and car in which the intra-class variability is higher depending on the orientation. Also the candidate generation and re-localization based on segmentation as in [33] could be integrate in this pipeline improving the obtained results.

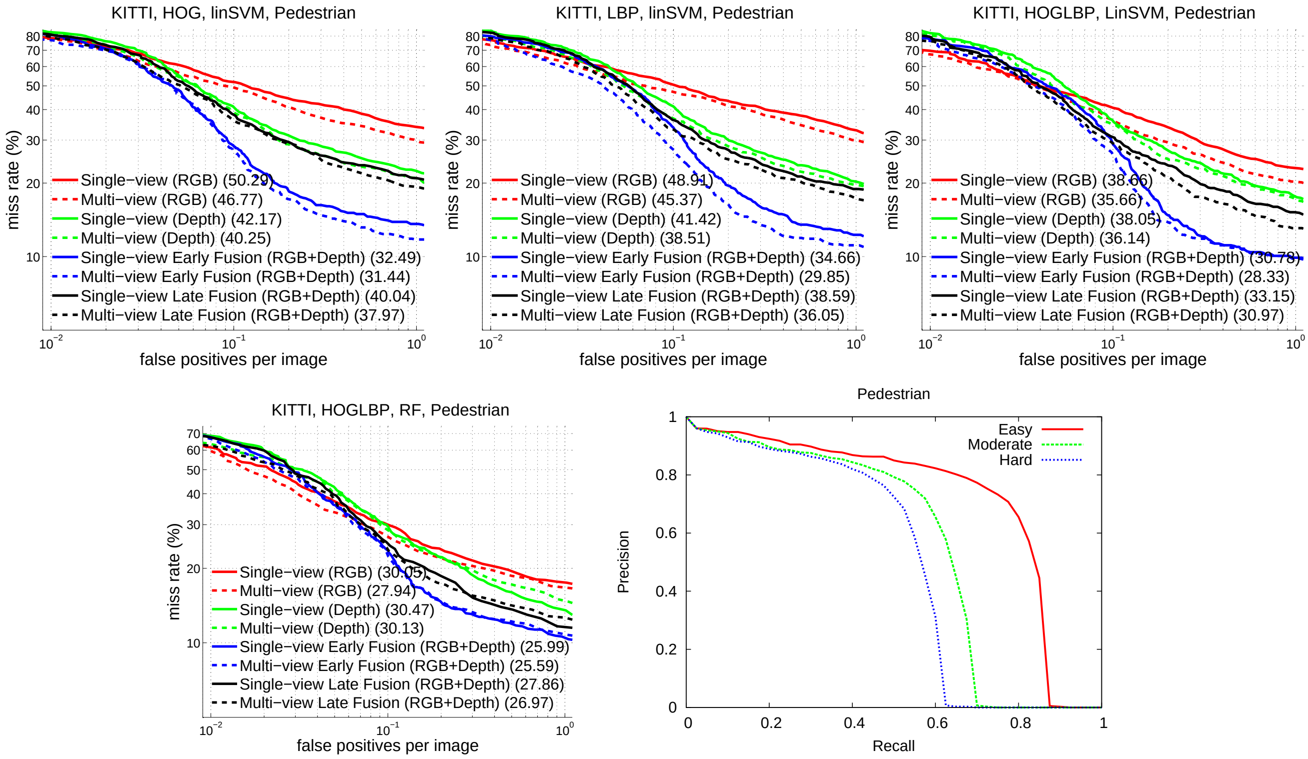

Fig. 5. First 4 plots represent the results over validation set using HOG/linSVM, LBP/linSVM, HOGLBP/linSVM, HOGLBP/RF under the different sources: RGB, Depth and RGB+Depth. The last plot shows the final approach precision-recall curve of the testing set for each subset: *easy*, *moderate* and *hard*.

V. Acknowledgments

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

References

[1] D. Gerónimo and A. López, *Vision-based Pedestrian Protection Systems for Intelligent Vehicles*. Springer, 2013.

[2] M. Enzweiler and D.M. Gavrila, "Monocular pedestrian detection: survey and experiments," *T-PAMI*, vol. 31, no. 12, 2009.

[3] D. Gerónimo, A.M. López, A.D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *T-PAMI*, vol. 32, no. 7, pp. 1239–1258, 2010.

[4] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *T-PAMI*, vol. 34, no. 4, pp. 743–761, 2012.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, San Diego, CA, USA, 2005.

[6] T. Ahonen, A. Hadid, and P. M., "Face recognition with local binary patterns." in *ECCV*, 2004.

[7] X. Wang, T.X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *ICCV*, Kyoto, Japan, 2009.

[8] J. Marin, D. Vázquez, A. López, J. Amores, and L. Kuncheva, "Occlusion handling via random subspace classifiers for human detection," *Cyber*, 2013.

[9] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *BMVC*, London, UK, 2009.

[10] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," in *CVPR*, 2013.

[11] M. Rao, D. Vázquez, and A. López, "Color contribution to part-based person detection in different types of scenarios," in *CAIP*, 2011.

[12] M. Wang and X. Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *CVPR*, 2011.

[13] G. Chen, Y. Ding, J. Xiao, and T. Han, "Detection evolution with multi-order contextual co-occurrence," in *CVPR*, Portland, 2013.

[14] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *CVPR*, 2013.

[15] C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining rgb and dense lidar data," in *IROS*, 2014.

[16] D. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[17] L. Spinello, K. O. Arras, R. Triebel, and R. Siegwart, "A layered approach to people detection in 3d range data." in *AAAI*, 2010.

[18] K. Kidono, T. Miyasaka, A. Watanabe, T. Naito, and J. Miura, "Pedestrian recognition using high-definition lidar," in *IV*, 2011.

[19] K. Kidono, T. Naito, and J. Miura, "Reliable pedestrian recognition combining high-definition lidar and vision data," in *ITSC*, 2012.

[20] L. E. Navarro-Serment, C. Mertz, and M. Hebert, "Pedestrian detection and tracking using three-dimensional ladar data." *IJRR*, vol. 29, no. 12.

[21] J. Behley, V. Steinhage, and A. B. Cremers, "Laser-based segment classification using a mixture of bag-of-words," in *IROS*, 2013.

[22] J. Xu, K. Kim, Z. Zhang, H.-w. Chen, and Y. Owechko, "2d/3d sensor exploitation and fusion for enhanced object detection," in *CVPR*, 2014.

[23] J. Marin, D. Vázquez, A. López, J. Amores, and B. Leibe, "Random forests of local experts for pedestrian detection," in *ICCV*, 2013.

[24] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *T-PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[25] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," *IJCV*, 2009.

[26] J. Gall and V. Lempitsky, "Class-specific hough forests for object detection," in *CVPR*, 2009.

[27] D. Tang, Y. Liu, and T.-K. Kim, "Fast pedestrian detection by cascaded random forest with dominant orientation templates," in *BMVC*, 2012.

[28] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *IJCV*, 2008.

[29] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in *CVPR*, 2011.

[30] M. Enzweiler and D.M. Gavrila, "A multi-level mixture-of-experts framework for pedestrian classification," *T-IP*, vol. 20, no. 10, 2011.

[31] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *T-PAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[32] D. Vázquez, A. López, J. Marín, D. Ponsa, and D. Gerónimo, "Virtual and real world adaptation for pedestrian detection," *T-PAMI*, 2013.

[33] C. Long, X. Wang, M. Yang, and Y. Lin, "Accurate object detection with location relaxation and regionlets relocalization," in *ACCV*, 2014.