On-Board Object Detection: Multi-cue, Multi-modal and Multi-view Random Forest of Local Experts.

Alejandro González, IEEE, David Vázquez, IEEE, Antonio M. López, IEEE and Jaume Amores

***Abstract*—Despite recent significant advances, object 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 object views and poses. In this paper we provide an extensive evaluation that gives insight into how each of these aspects (multi-cue, multi-modality and strong multi-view classifier) affect accuracy both individually and when integrated together. In the multi-modality component we explore the fusion of RGB and depth maps obtained by high-definition LIDAR, a type of modality that is starting to receive increasing attention. As our analysis reveals, although all the aforementioned aspects significantly help in improving the accuracy, 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.**

***Index Terms*—Multi-cue, Multi-modal, Multi-view, Object detection.**

I. Introduction

DEVELOPING a reliable object detector enables a vast range of applications such as video surveillance and the practical deployment of autonomous and semi-autonomous vehicles. In order to obtain a detector that successfully operates under realistic 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 (color, depth, etc.), and iii) the use of multiple views (frontal, lateral, etc.) of the object 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 insights about how each of these three aspects affect accuracy, both individually and when integrated together. The proposed method (General scheme in Fig. 1) will be evaluated in key objects for autonomous and semi-autonomous vehicles such as pedestrians, cyclists and cars. With more than a decade of history, by now pedestrian detection [1] is still a very challenging task [2], [3], [4]. Car detection is very relevant for driver assistance (*e.g.*ACC) and autonomous vehicles [5], [6], [7], [8].

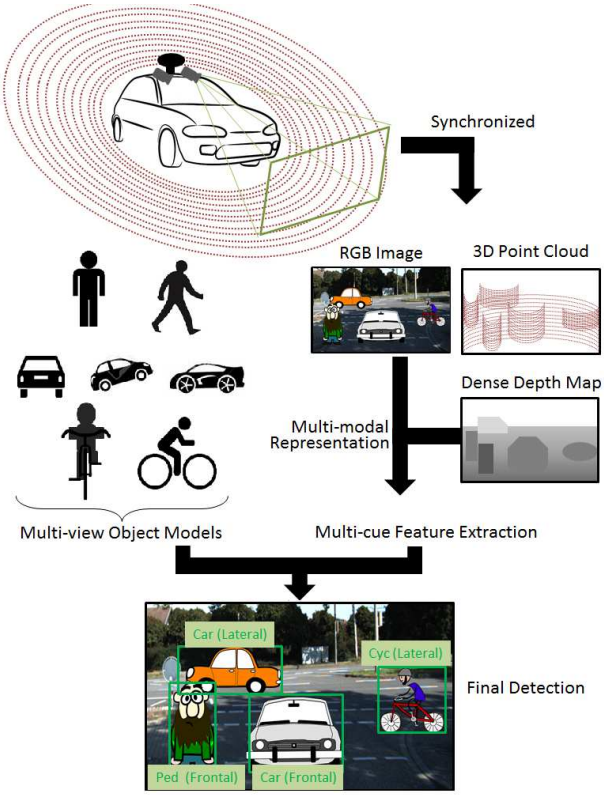

Fig. 1. General scheme: From RGB images and LIDAR data to object detection. RGB images and LIDAR data synchronized for multi-modal representation. Multi-modal representation based on RGB images and dense depth maps (obtained from LIDAR sparse data). Multi-cue feature extraction over the multi-modal representation. Multi-view detection of different objects.

In order to integrate different cues we use HOG [9], that provides a good description of the object contours, and LBP [10] as texture-based feature. These two types of features provide complementary information and the fusion of both types of features has been seen to boost the performance of either feature separately [11], [12], [13]. From the seminal work of Dalal and Triggs [9] it has been seen that using different types of gradient-based features and their spatial distribution, such as in the HOG descriptor [9] provides a distinctive representation of both humans and other objects classes. However, there exist in the literature other approaches such the integral channel features proposed by Dollar et

A. González, D. Vázquez and A.M. López are with the Departament de Ciències de la Computació and the Computer Vision Center (CVC), Universitat Autònoma de Barcelona, Spain. E-mail: agalzate@cvc.uab.es

J. Amores is with United Technologies Research Center

This work is supported by the Spanish MICINN projects TRA2014-57088-C2-1-R, by the Secretaria dUniversitats i Recerca del Departament dEconomia i Coneixement de la Generalitat de Catalunya (2014-SGR-1506), by TECNIOspring with the FP7 of the EU and ACCI, and also by DGT (SPIP2014-01352). Our research is also kindly supported by NVIDIA Corporation in the form of different GPU hardware.

al. [14] that 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 [15], [16] 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 [17], [18] and could be easily integrated as well. Exploring alternative types of spatial pooling of the local features is also beneficial as shown in [6] and is also complementary to the approach used in this paper.

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 objects detection in on-board sequences. 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. In this work we explore the fusion of dense depth maps (obtained based on the sparse cloud of points) with RGB images. Following [19], the information provided by each modality can be fused using either an early-fusion scheme, *i.e.* at the feature level, or a late-fusion scheme, *i.e.* at the decision level. In our study, using an early fusion scheme, where descriptors from each modality are concatenated, provided the best results.

Object 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 [20], [21]. Only recently authors are starting to study the impact of high-definition 3D LIDAR [21], [20], [22], [23], [24], [25], [26]. Most of these works propose specific descriptors for extracting information directly from the 3D cloud of points [20], [22], [23], [24], [25], [26]. A common approach is to detect objects independently in the 3D cloud of points and in the visible spectrum images, and then combining the detections using an appropriate strategy [22], [23], [26]. Following the steps of [21], 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. Given this map, 2D descriptors in the literature can be extracted in order to obtain a highly distinctive object representation. Our work differ from [21] in that we use multiple descriptors and adapt them to have a good performance in dense depth images. While [21] employs a late fusion scheme, in our experimental analysis we evaluate both early and late fusion approaches in the given multi-cue, multi-modality framework.

Learning a model flexible enough for dealing with multiple views and multiple positions of an articulated object is a hard task for a holistic classifier. In order to fulfill this aspect we

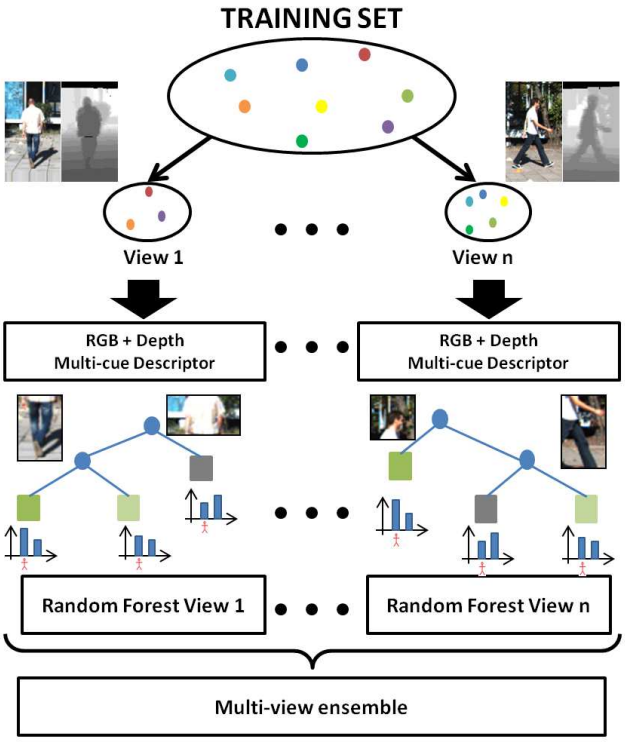

Fig. 2. Multi-view, Multi-cue, Multi-modal detector scheme. 1) Split training set samples in different views. 2) Generate a multi-modal representation using RGB and depth. 3) Extract multi-cue features. 4) Train a random forest of local experts for each view. 5) Ensemble different views detection.

make use of Random Forests (RF) of local experts [27], which has a similar expressive power than the popular Deformable Part Models (DPM) [28] 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 [27] the RF approach consistently outperformed DPM. An advantage of the RF method is that only a single 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 [28], [29] and multiple local patches [30], [31], [32]. The method in [33] 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 [27].

Most relevant to this paper is the approach presented in [13] where the authors combine multiple views (front, left, back, right), modalities (luminance, depth based on stereo,

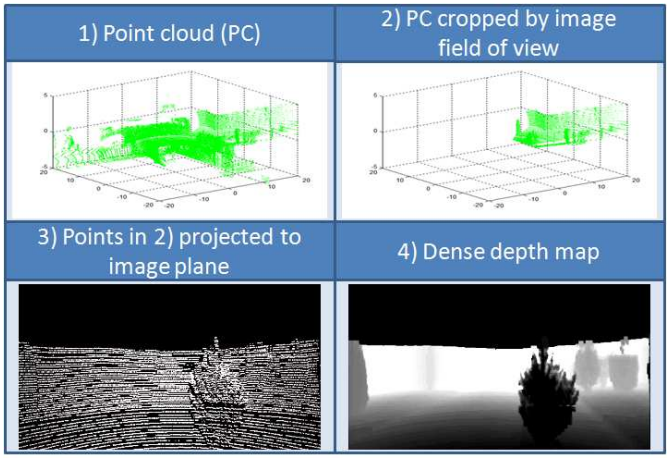

Fig. 3. Dense depth map generation scheme. From a cloud of points to a dense depth map: Filter cloud of points for synchronize with view field of image. Projection of 3D points into 2D image coordinates. Interpolate depths for getting a dense depth map.

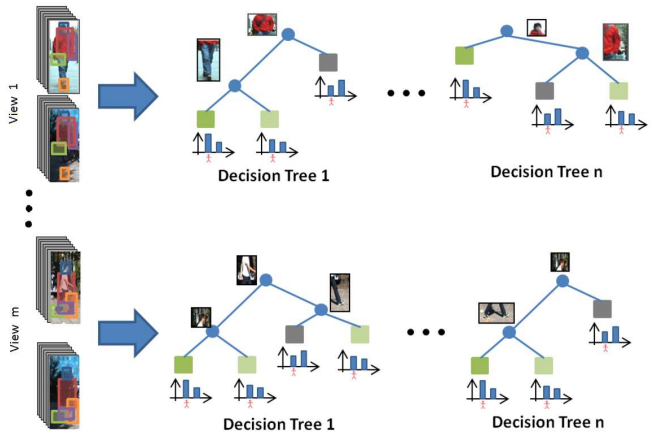

Fig. 4. Multi- View Random Forest scheme. For each view is learnt a different random forest, and each tree has different configuration of random patches.

and optical flow), and features (HOG and LBP). The main differences between [13] 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, but it is being used for autonomous driving, and ii) while [13] makes use of an holistic classifier, we make use of a more expressive patch-based model, and iii) in [13] multiples cues are combined following late-fusion style, while we consider also early-fusion, which, in fact, gives better results in our framework.

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.

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 Object 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 in a final ensemble. In Fig. 2 is shown a scheme of the proposed method applied for pedestrian detection. In addition we will model objects 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 object 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 [9] descriptor (shape) and the LBP [34] descriptor (texture). Both descriptors are combined using an early fusion technique, concatenating them, obtaining a robust descriptor with complementary information (HOGLBP). HOG descriptor is composed by a histogram of gradient orientations. Given a candidate window the histograms are calculated on overlapped blocks inside it. LBP descriptor calculates 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 resize 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 [35]).

B. Multi-modal image fusion

Keeping in mind that more complementarity is better for object 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.* [21]. 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. In order to do this each point P_i is projected into the image plane using the calibration and projection matrices provided in the dataset, using $TM = P2 \times R0 \times VtC$, where, $P2$ is the projection matrix from camera coordinate system to left color image coordinate system, $R0$ is the rectification matrix, and VtC is the projection matrix from velodyne coordinate system to camera coordinate system. Once we have the transformation matrix (TM) we can project any 3D point (defined by its 3D coordinates $[x_{3D}, y_{3D}, z_{3D}]$) to its correspondent point in the image plane (defined by its 2D coordinates $[x_{2D}, y_{2D}]$) by applying $[x_{2D}, y_{2D}, 1] = TM * [x_{3D}, y_{3D}, z_{3D}, 1]$. Then the points that fall inside the image borders are selected, while the others are rejected, ending up with points that form a sparse depth image, time and space synchronized with the visual image. At this step by defining a neighborhood (N) for each valid pixel of the depth map we interpolate the information for filling the missing values. In order to calculate the missing values we use the bilateral filtering formalism [36]: $D_p = \frac{1}{W_p} * \sum_{q \varepsilon N} G_d(\|p-q\|) * G_i(|I_q|) * I_q$ where I_q is the depth value of the point q, G_d weights points q inversely to their distance to position p, G_i penalizes as function of their range values, and W_p is a normalization factor. After this process, the pixels without depth information will be filled, ending up in a dense depth map (see Fig. 3).

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 an ensemble of detectors (late fusion); in this case we train two separate detectors, one per modality. The second one is to combine at feature level the two modalities (early fusion); in this case we train a single model using as descriptor the concatenation of the features computed at each modality.

C. Multi-view classifier

In general, reducing intra-class variability is a good way to better discriminate a class from potential false positives (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 and aspect ratio; obtaining in this way a partition of n subsets where the variability among the samples in each subset is lower than in the original set. The partition is obtained by using a cluster method with regular-spaced seeds in orientation space, in particular using K-Means [37]. After splitting the training samples using this automatic approach, we can set the canonical size of the detection window for each subset (aspect ratio) by selecting the mean size among the samples in each partition set. Thus,

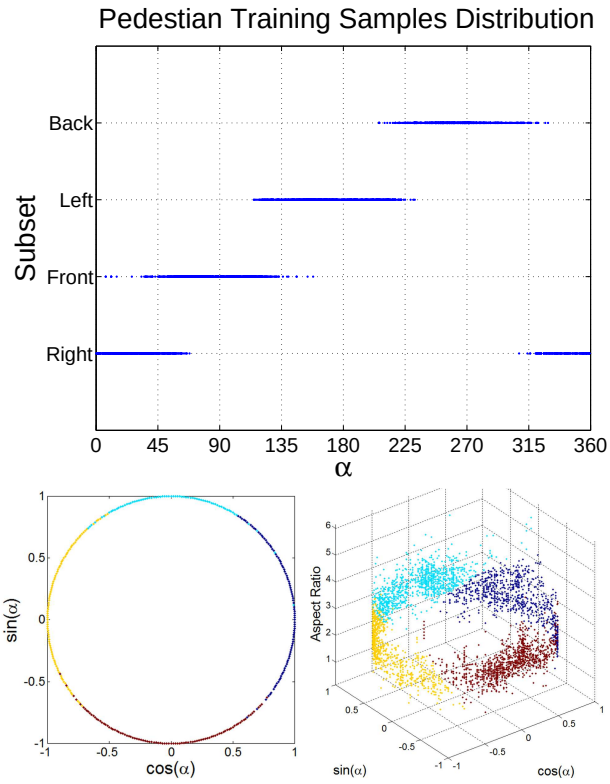

Fig. 5. Pedestrian Orientation Histogram and Distribution. Upper image shows the assigned views against the angle for each training sample. Bottom images show the training samples distribution in the clustering space (angle and aspect ratio).

this process allows the final detector to deal with objects in different orientations, having each orientation its own aspect ratio (*e.g.* we use a different bounding box aspect ratio for modeling frontal-viewed pedestrians than for modeling side-viewed ones). In figure 5, 6, and 7 it is shown the training samples and their views definition based on the clustering process for the pedestrians, cyclists, and cars classes respectively. In order to cluster we use the orientation angle (α) and the aspect ratio (AR) of the sample.

D. Object model

In our study we focus on two different models: one holistic, where the object descriptor takes into account the candidate detection window as a whole; and a patch-based one where random subsets of patches are used for generating different object configurations which are further assembled to form the overall object model. As holistic model we use the linear SVM ($linSVM$) classifier which has a good compromise between computation time and accuracy. This model learns the max-margin hyperplane that better splits the positive and negative samples in the descriptor space (either HOG, LBP or HOGLBP in our case). As patch-based model we use our Random Forest (RF) of Local Experts. The forest is composed of trees. Each tree is able to return a classification probability (of being object) given a candidate window that must be classified. The forest classification probability is obtained by averaging the classifications probabilities of the

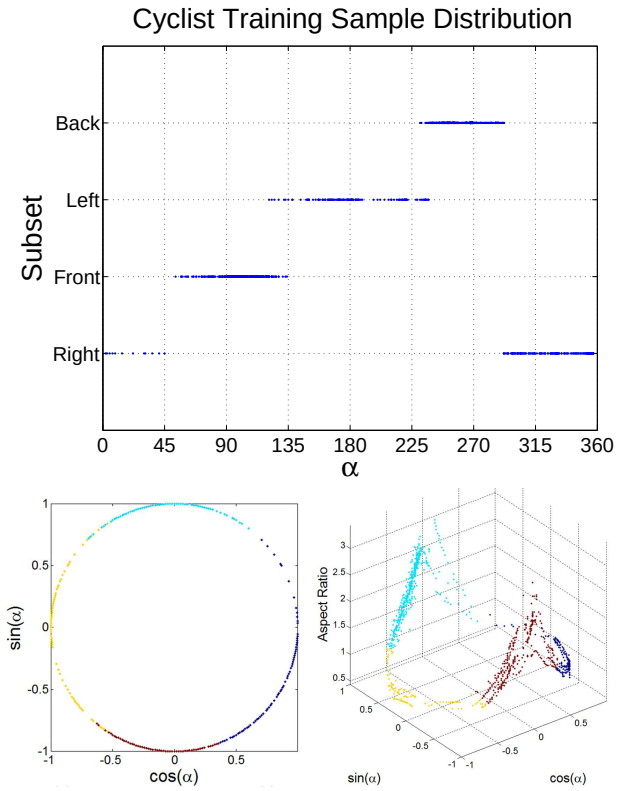

Fig. 6. Cyclist Orientation Histogram and Distribution. Upper image shows the assigned views against the angle for each training sample. Bottom images show the training samples distribution in the clustering space (angle and aspect ratio).

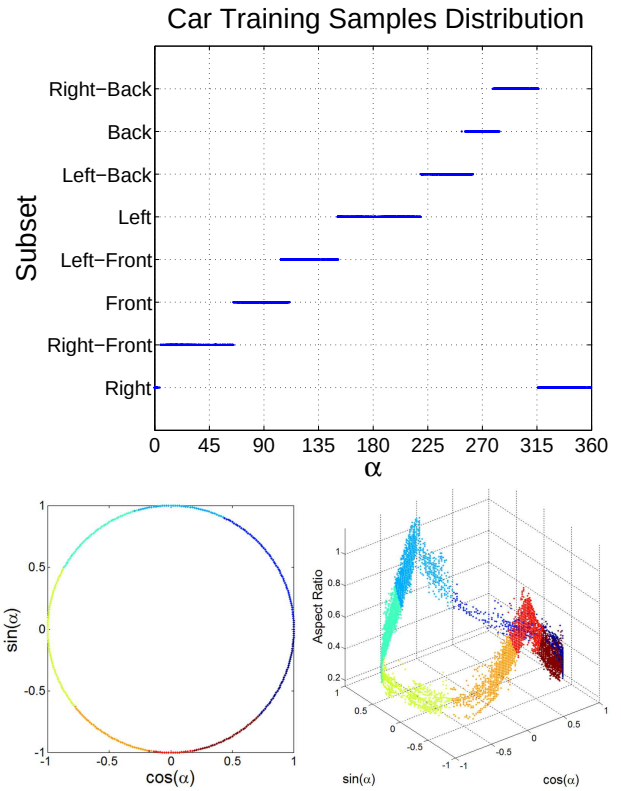

Fig. 7. Car Orientation Histogram and Distribution. Upper image shows the assigned views against the angle for each training sample. Bottom images show the training samples distribution in the clustering space (angle and aspect ratio).

trees. Each tree of the forest is a binary tree, where at each node there is either one local expert (decision node) or a leaf node (with the corresponding probability distribution of object *vs* background). The local experts decide whether to route the tree search towards the left or the right; a search that starts in the root of the tree and ends in a leaf node. The probability distribution at the reached leaf provides the classification probability of the tree. In this case, each expert consists in a *linSVM* classifier that takes into account the descriptor (HOG/LBP/HOGLBP) of only a pre-assigned (during the learning stage) patch within the candidate window under classification. Thus, each tree corresponds to a different configuration of patches (see Fig. 4), and thus the forest is an ensemble of patch-based configurations which where automatic learned. In this paper we use the RF formed by 100 trees, 7 levels as maximum depth, we refer to [27] for details about the training of such RF of local experts.

III. Experimental results

In this section we will evaluate each step of the proposed approach: multi-cue, multi-modal and multi-view as we have described in previous sections, in order to fulfill this evaluation we will use HOG and LBP features and 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 was the first milestone in pedestrian detection.

a) **KITTI Dataset:** in this paper we use the KITTI dataset since it provides synchronized camera and LIDAR data. KITTI dataset for object detection includes 7,481 training images and 7,518 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 3,740 images) and a validation set (the last 3,741 images) as in [21], these subsets are used for the evaluation of each step of our approach. The original training and testing set will be used for training and testing the optimal configuration of the detector, *i.e.*, in order to compare with the state-of-the-art methods using the KITTI web page for submitting results. During training we consider pedestrians, cyclists and cars higher than 25 pixels and not occluded (Reasonable subset).

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. For testing we use the reasonable subset in the caltech evaluation and the KITTI evaluation is performed over three different subsets depending

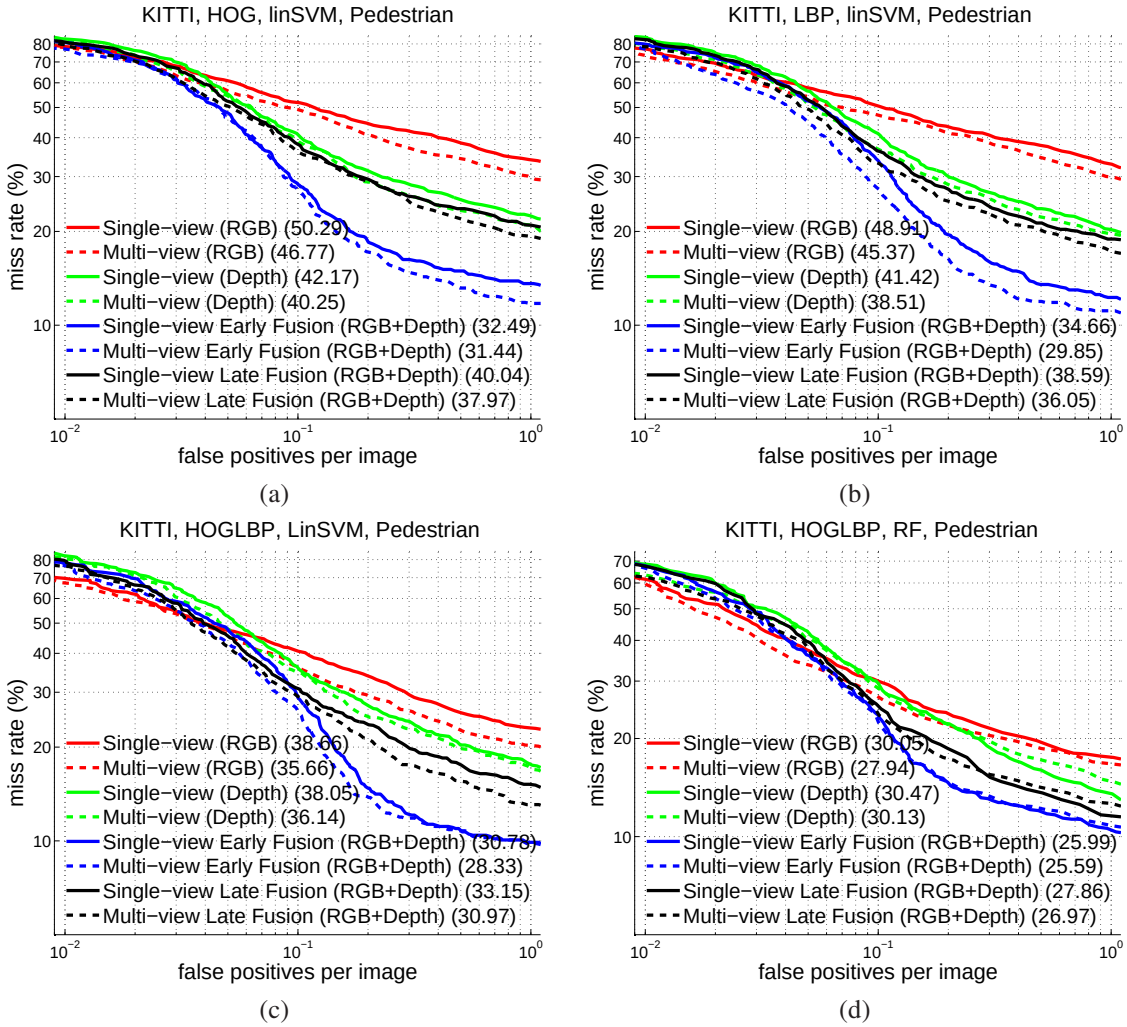

Fig. 8. Results over validation set using a) HOG/linSVM, b) LBP/linSVM, c) HOGLBP/linSVM, d) HOGLBP/RF. All detectors under the different sources: RGB (Red), Depth (Green) and RGB+Depth (Early fusion (Blue) and late fusion(Black)); and using a single-view (Continuous line) and multi-view (Dot line) approaches. The number in parenthesis for each curve represents the average miss rate (*AMR*).

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 multiple cues, for that reason we start by comparing each single-view (SV) detector. 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. As it is usual for pedestrian detection to use the Caltech standard evaluation method, we plot the FPPI curves in Fig. 8. Comparing the *AMR* in SV experiments HOG/linSVM (Fig. 8a) against HOGLBP/linSVM (Fig. 8c), we can see that the gain in *AMR* is around 12 points with RGB modalities, around 4 with depth, around 2 when combining the two modalities using an early fusion approach, and 7 when a late fusion approach is used. The same behavior can be seen also if we compare the LBP/linSVM (Fig. 8c) against the HOGLBP/linSVM where we obtain improvements of around 10 , 3, 4 and 5 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. 8 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 for now ahead when we compare against RGB+Depth we are making reference to the early fusion approach. In the rest of the paper it will be used the KITTI evaluation method in order to compare the results using all the different classes. In Tables I, II, and III (Pedestrian, Cyclists and Cars results respectively) 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. In pedestrian class (Table I) we obtain an *AP* gain of ~ 21 against RGB and ~ 12

TABLE I

RESULTS FOR PEDESTRIAN DETECTION USING DIFFERENT SUBSETS FOR TRAINING (SINGLE-VIEW (SV), MULTI-VIEW (MV)), MODALITIES, AND DETECTORS, TESTED OVER THE VALIDATION SET. FOR EACH DETECTOR *AP* FOR KITTI EVALUATION IS SHOWN . BEST *AP* FOR EACH DETECTOR IN EACH MODALITY IS INDICATED IN BOLD, WHILE THE BEST DETECTOR ACROSS THE DIFFERENT MODALITIES IN RED

Evaluation	Detector	RGB		Depth		Early Fusion		Late Fusion	
		SV	MV	SV	MV	SV	MV	SV	MV
AP (Easy)	HOG/linSVM	0.50	**0.54**	0.59	**0.62**	0.71	**0.73**	0.63	**0.65**
	LBP/linSVM	0.52	**0.56**	0.62	**0.65**	0.69	**0.75**	0.66	**0.68**
	HOGLBP/linSVM	0.64	**0.68**	0.65	**0.67**	0.74	**0.76**	0.70	**0.73**
	HOGLBP/RF	0.73	**0.75**	0.74	**0.75**	**0.79**	**0.79**	0.77	**0.79**
AP (Moderate)	HOG/linSVM	0.38	**0.41**	0.46	**0.47**	0.57	**0.58**	0.49	**0.51**
	LBP/linSVM	0.41	**0.44**	0.48	**0.50**	0.57	**0.61**	0.52	**0.54**
	HOGLBP/linSVM	0.50	**0.54**	0.51	**0.52**	0.61	**0.62**	0.56	**0.58**
	HOGLBP/RF	0.59	**0.60**	0.58	**0.58**	0.65	**0.66**	0.62	**0.63**
AP (Hard)	HOG/linSVM	0.33	**0.35**	0.40	**0.41**	0.50	**0.51**	0.43	**0.44**
	LBP/linSVM	0.36	**0.38**	0.42	**0.43**	0.50	**0.53**	0.45	**0.47**
	HOGLBP/linSVM	0.43	**0.47**	0.45	**0.46**	0.53	**0.55**	0.49	**0.51**
	HOGLBP-RF	0.51	**0.52**	0.50	**0.50**	0.56	**0.57**	0.54	**0.55**

TABLE II

RESULTS FOR CYCLIST DETECTION USING DIFFERENT SUBSETS FOR TRAINING (SINGLE-VIEW (SV), MULTI-VIEW (MV)), MODALITIES, AND DETECTORS, TESTED OVER THE VALIDATION SET. FOR EACH DETECTOR *AP* FOR KITTI EVALUATION IS SHOWN . BEST *AP* FOR EACH DETECTOR IN EACH MODALITY IS INDICATED IN BOLD, WHILE THE BEST DETECTOR ACROSS THE DIFFERENT MODALITIES IN RED

Evaluation	Detector	RGB		Depth		Early Fusion		Late Fusion	
		SV	MV	SV	MV	SV	MV	SV	MV
AP (Easy)	HOG/linSVM	0.43	**0.52**	**0.44**	0.42	0.62	**0.66**	0.48	**0.51**
	LBP/linSVM	0.34	**0.48**	**0.48**	0.46	0.62	**0.62**	0.50	**0.53**
	HOGLBP/linSVM	0.49	**0.60**	0.48	**0.49**	0.69	**0.69**	0.55	**0.59**
	HOGLBP/RF	0.64	**0.70**	0.49	**0.49**	0.72	**0.73**	0.54	**0.57**
AP (Moderate)	HOG/linSVM	0.31	**0.41**	**0.30**	0.29	0.44	**0.49**	0.34	**0.39**
	LBP/linSVM	0.29	**0.41**	**0.34**	0.33	0.48	**0.50**	0.38	**0.43**
	HOGLBP/linSVM	0.39	**0.50**	0.34	**0.35**	0.52	**0.54**	0.42	**0.48**
	HOGLBP/RF	0.50	**0.57**	0.33	**0.35**	0.52	**0.55**	0.41	**0.45**
AP (Hard)	HOG/linSVM	0.28	**0.38**	**0.28**	0.27	0.41	**0.45**	0.32	**0.36**
	LBP/linSVM	0.26	**0.38**	**0.31**	0.30	0.45	**0.46**	0.35	**0.39**
	HOGLBP/linSVM	0.35	**0.46**	0.32	**0.33**	0.48	**0.50**	0.38	**0.44**
	HOGLBP/RF	0.45	**0.52**	0.31	**0.32**	0.47	**0.50**	0.38	**0.41**

TABLE III

RESULTS FOR CAR DETECTION USING DIFFERENT SUBSETS FOR TRAINING (SINGLE-VIEW (SV), MULTI-VIEW (MV)), MODALITIES, AND DETECTORS, TESTED OVER THE VALIDATION SET. FOR EACH DETECTOR *AP* FOR KITTI EVALUATION IS SHOWN . BEST *AP* FOR EACH DETECTOR IN EACH MODALITY IS INDICATED IN BOLD, WHILE THE BEST DETECTOR ACROSS THE DIFFERENT MODALITIES IN RED

Evaluation	Detector	RGB		Depth		Early Fusion		Late Fusion	
		SV	MV	SV	MV	SV	MV	SV	MV
AP (Easy)	HOG/linSVM	0.26	**0.72**	0.22	**0.78**	0.29	**0.77**	0.17	**0.78**
	LBP/linSVM	0.11	**0.62**	0.04	**0.70**	0.11	**0.71**	0.10	**0.71**
	HOGLBP/linSVM	0.16	**0.66**	0.18	**0.70**	0.21	**0.72**	0.06	**0.72**
	HOGLBP/RF	0.29	**0.81**	0.38	**0.81**	0.37	**0.82**	0.24	**0.82**
AP (Moderate)	HOG/linSVM	0.21	**0.67**	0.17	**0.56**	0.24	**0.69**	0.18	**0.71**
	LBP/linSVM	0.11	**0.60**	0.03	**0.61**	0.11	**0.65**	0.12	**0.67**
	HOGLBP/linSVM	0.14	**0.65**	0.16	**0.63**	0.19	**0.67**	0.11	**0.68**
	HOGLBP/RF	0.26	**0.75**	0.28	**0.61**	0.29	**0.76**	0.24	**0.75**
AP (Hard)	HOG/linSVM	0.17	**0.52**	0.14	**0.44**	0.19	**0.54**	0.15	**0.57**
	LBP/linSVM	0.10	**0.48**	0.03	**0.49**	0.09	**0.52**	0.09	**0.54**
	HOGLBP/linSVM	0.11	**0.52**	0.13	**0.50**	0.14	**0.54**	0.09	**0.55**
	HOGLBP/RF	0.21	**0.61**	0.22	**0.48**	0.23	**0.62**	0.20	**0.62**

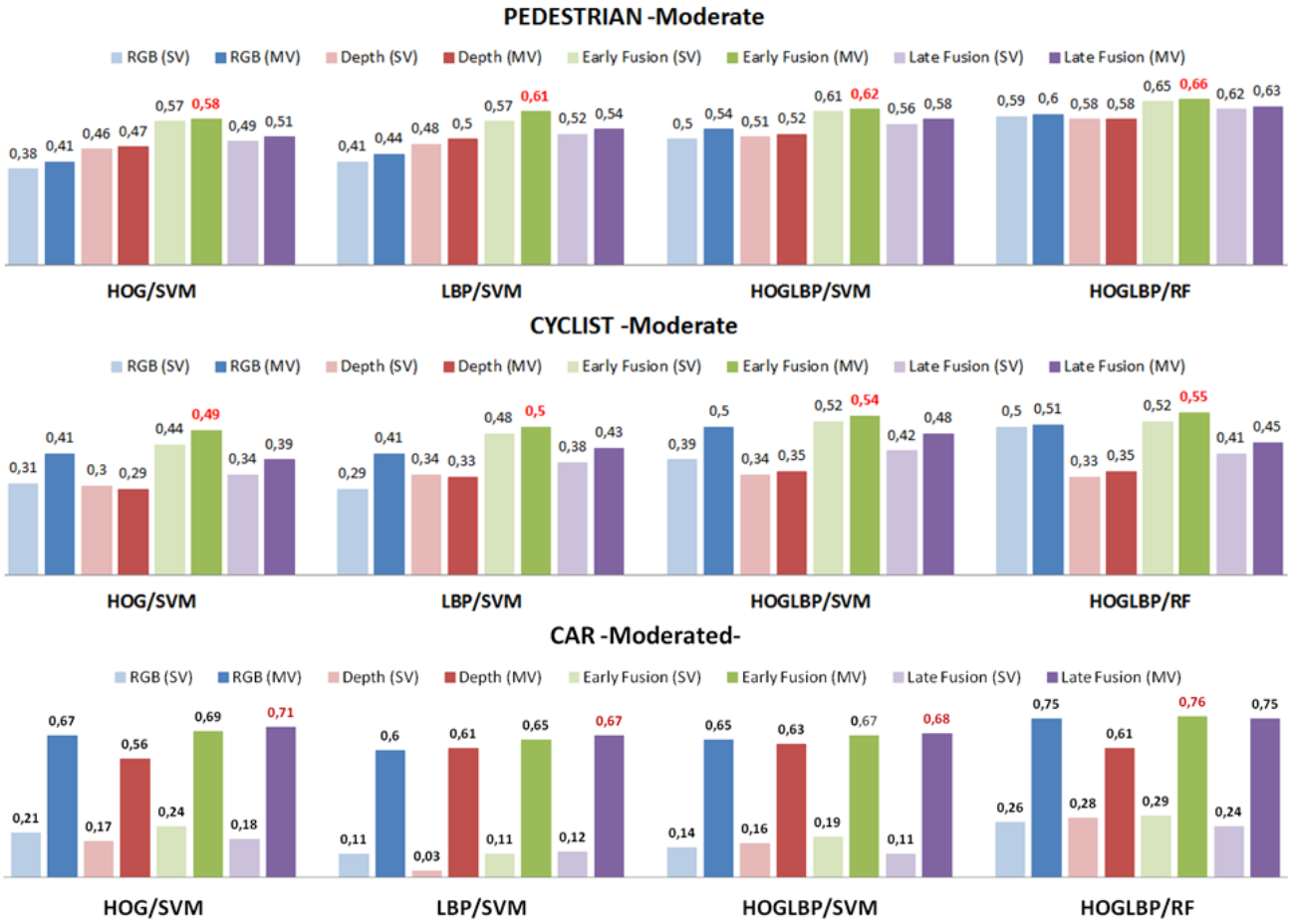

Fig. 9. Detection results for pedestrians, cyclists, and cars, in the moderate case (Tables I, II, and III for details). In red the best result of each configuration.

against Depth. This behavior is repeated if we look at the different SV proposed models: LBP/linSVM (~ 17 / ~ 7), HOGLBP/linSVM (~ 10 / ~ 9) and HOGLBP/RF (~ 6 / ~ 5). Repeating this analysis in cyclists (Table II)and cars classes (Table III) we observe improvements in all different detectors if we compare the multi-modal version against its single-modal counterpart.

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/linSVM against the MV version. In Table IV shows the different view's relevant values, min and max angle, aspect ratio and number of samples. The number of views for each class is defined according to the appearance symmetries of the class. For instance, pedestrians present a very similar pose appearance when imaged from frontal and rear views; thus, we use a single frontal-rear view. Analogously, for left and right views; thus, we use a single left-right view. Note that frontal-rear and left-right views look different, mainly due to the pose of the legs and arms. Vehicles are very different than pedestrians in this sense. For instance, for some vehicles even frontal and rear views are really different. Therefore, we decided to incorporate more views, in particular eight: frontal, rear, right, left, right-frontal, left-frontal, right-rear, and left-rear. Cyclists correspond to an intermediate situation between pedestrians and vehicles. However, we think they are more close to pedestrians and, therefore, for this study we have assumed frontal-rear and left-right differences, *i.e.* two views too. Looking again at Tables I, II, and III, where the results for the different classes are tabulated, and comparing the SV-HOG/SVM against its MV counterpart, for pedestrian detection (Table I) we obtain an *AP* improvement of ~ 4 (RGB), ~ 3 (Depth) and ~ 2 (RGB+Depth). The same behavior is obtained by comparing the other SV pedestrian models against its MV counterpart: LBP/linSVM (~ 4 / ~ 3 / ~ 6), HOGLBP/linSVM (~ 4 / ~ 2 / ~ 2) and HOGLBP/RF (~ 2 / ~ 1 / ~ 0). Following the same analysis in cyclists (Table II) and cars detection (Table III), we observe a similar behavior getting improvements in each one of the proposed detectors.

f) **Discussion:** Each of the mentioned detectors in Section III is developed using RGB, Depth and Early Fusion and Late Fusion 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 different views). In Tables I, II, and III we show the accuracy measurements over the validation set. The measurements include the KITTI evaluation methodology for *easy*, *moderate* and *hard* pedestrian subset. Figure 9 shows

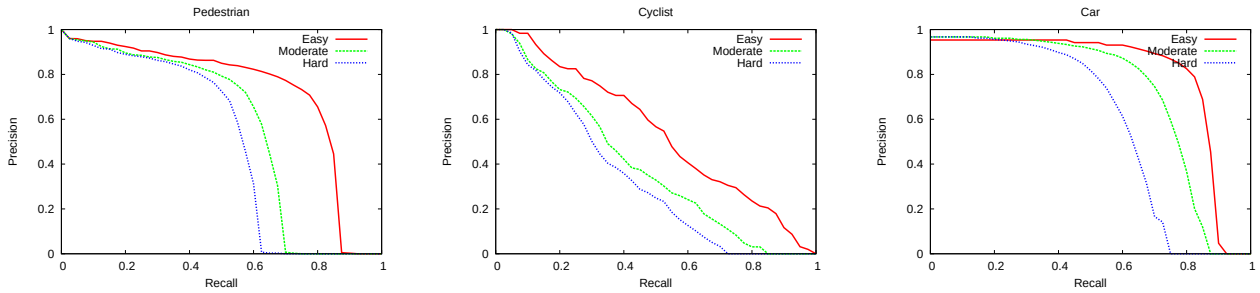

Fig. 10. Precision-recall curve of the testing set for each subset: *easy*, *moderate* and *hard*, for pedestrian, cyclist and car classes.

TABLE IV
MULTIVIEW PARTITION

Class	View	Angle min	Angle max	Aspect Ratio	Num. Samples
Pedestrian	Left	136	219	2.50	940
	Right	-42	37		
	Front	37	136	2.69	1415
	Back	219	318		
Cyclist	Left	127	234	1.85	328
	Right	-68	49		
	Front	49	127	0.97	760
	Back	234	292		
Car	Right	-44	4	0.37	902
	Right-Front	4	65	0.36	274
	Front	65	107	0.74	1713
	Front-Left	107	151	0.50	4170
	Left	151	219	0.37	1194
	Left-Back	219	257	0.57	1850
	Back	257	284	0.84	4061
	Back-Right	284	316	0.55	1542

graphically the results for the *moderate* case. Regarding the obtained results it is easy to see the accuracy 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, Early Fusion and Late Fusion) 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 Early Fusion experiment.

An interesting result can be observed in Table II (cyclist). In particular, SV and MV models report very similar accuracy when relying on Depth. In fact, only for HOG+LBP the MV models tend to perform better than the SV. The reason is that many times the bike itself is not visible for the LIDAR, so cyclists resemble pedestrians. Thus, while the SV models tend to be more blurry, in this case when using a MV model, some of the views may be especially sensitive to pedestrians; which turns out to be worse for the evaluation protocol we follow, since when a pedestrian is classified as cyclist, this is considered an error (false positive). However, this may depend a lot in the particular data we have for training. Overall, still combining multiple features, views and modalities compensate

TABLE V
EVALUATION AND COMPARISON OF MULTI-VIEW RGBD RF DETECTOR USING THE FINAL TEST SET FOR PEDESTRIAN DETECTION

Rank	Method	Moderate	Easy	Hard
1	Regionlets	61.15 %	73.14 %	55.21 %
2	**MV-RGBD-RF**	**56.59 %**	**73.30 %**	**49.63 %**
3	pAUCEnsT	54.49 %	65.26 %	48.60 %

TABLE VI
EVALUATION AND COMPARISON OF MULTI-VIEW RGBD RF DETECTOR USING THE FINAL TEST SET FOR CYCLIST DETECTION

Rank	Method	Moderate	Easy	Hard
1	Regionlets	58.72 %	70.41 %	51.83 %
2	**MV-RGBD-RF**	**42.61 %**	**52.97 %**	**37.42 %**
3	pAUCEnsT	38.03 %	51.62 %	33.38 %

this circumstance. On the other hand, it is more difficult to confuse pedestrians with cyclists, because the pose of the cyclists may be inclined on the bike, thus not common for pedestrians. This fact, together with the circumstance of having less training examples of cyclists that pedestrians and cars, lead to a poorer detection accuracy for cyclist than for the other classes, which can be seen by comparing Tables I, II, and III, as well as in Fig. 10.

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 *AP* gain of ~ 29 in pedestrians detection, ~ 30 in cyclists detection, and ~ 50 in cars detection, in the validation set (Tables I, II, and III respectively).

Regarding the final approach MV-HOGLBP/RF early fusion of RGB and Depth in Table V and comparing against the methods with an associated paper in the competition, we obtain an *AP* of 73.30%, 56.59%, 49.63% for the *easy*, *moderate* and *hard* subset respectively, ranking the second best pedestrian detector in the challenge. In Table VI, we obtain an *AP* of

TABLE VII
EVALUATION AND COMPARISON OF MULTI-VIEW RGBD RF DETECTOR USING THE FINAL TEST SET FOR CAR DETECTION

Rank	Method	Moderate	Easy	Hard
1	spLBP	77.39 %	87.18 %	60.59 %
2	Regionlets	76.45 %	84.75 %	59.70 %
3	3DVP	75.77 %	87.46 %	65.38 %
4	SubCat	75.46 %	84.14 %	59.71 %
5	AOG	71.88 %	84.36 %	59.27 %
6	**MV-RGBD-RF**	**69.92 %**	**76.40 %**	**57.47 %**

53.97%, 42.61%, 37.42% for the *easy*, *moderate* and *hard* subset respectively, ranking the second best cyclist detector in the challenge. Finally in Table VII, we obtain an *AP* of 70.40%, 69.92%, 57.47% for the *easy*, *moderate* and *hard* subset respectively, ranking the sixth best car detector in the challenge. Fig. 10, shows the precision-recall curves obtained over each subset using the final approach.

It is worth to mention that the first ranked method in pedestrian/cyclist detection, *i.e.* Regionlets [38], appeared posterior to our random forest of local experts but has common key ideas such as using HOG and LBP as features, and being patch-based. Thus, we think our conclusions also apply for them. Analogously, spLBP [39] is a better ranked method for car detection which appeared after our proposal, sharing ideas such as the subcategorization (multi-view approach) as well as using patch-based features (ACF [14]) too. Therefore, in line with our conclusions. In fact, new approaches are constantly appearing in this ranking, especially since the irruption of Convolutional Neural Networks (CNNs) in the computer vision community. In fact, as future work we plan to evaluate the use of not too deep CNNs as multi-cue, multi-view, and multi-modal local experts.

Finally, it is also worth to mention the computational complexity, in comparative terms, of applying multi-view multi-modal models. First of all, note that many methods in the literature use HOG and LBP as features, as well as pyramidal sliding window for providing candidate windows for their classification. Thus, our method is not different in this. Second, regarding the MV setting with respect to the SV setting, we remark that the same pyramids of HOG and LBP features have been used to apply SV and MV models; therefore, the main cost of the algorithm, *i.e.* feature extraction, is shared, it does not depend on the number of views. The only difference is in the number of scalar products of the object models (holistic linSVM, small linSVMs of RF nodes) times the feature vectors (HOG, LBP, HOGLBP). If the computational complexity of such scalar products for evaluating a single view is $\mathcal{O}(c)$, to evaluate n_v views, will be $\mathcal{O}(n_v c)$. Third, a similar reasoning applies when we have to detect different objects. We have used the same pyramid of features for all cases, the differences among objects is, again, the number of scalar products (models $\times$ features) required in each view of each object. Overall, lets $\mathcal{O}(c_f^m)$ be the computational complexity of generating the pyramid of features for the modality m (depth, RGB), lets $\mathcal{O}(c_v^o)$ be the average complexity of the mentioned scalar products for a single view of the object o, lets n_v^o be the number of views considered for the object o, lets N be the number of objects we want to detect, and lets M be the number of considered modalities; then, the computational complexity of the multi-view multi-modal approach is $\mathcal{O}((\sum_{o=1}^{N} n_v^o c_v^o) + (\sum_{m=1}^{M} c_f^m))$. As a reference, for an image of size of 1242×375pixels, our RGB-SV-HOGLBP/linSVM pedestrian detector runs at 2.5fps in a CPU i7-5930K, at 119fps in a NVIDIA GTX960, and at 20fps in a NVIDIA Tegra X1.

IV. Conclusions

In this paper we develop a complete multi-cue, multi-modal and multi-view framework for object detection. We have shown the applicability to different models (holistic, patch-based), obtaining significant accuracy improvements. In this paper we focus on object detection using HOG/linSVM as baseline applying the different proposed method: different cues (HOG and LBP), different modalities (RGB and Depth) and different views (Frontal, Lateral, etc.), thus, our immediate future work will focus on detection using more complex features (motion, context), classification algorithms (CNN), and modalities (disparity, far infrared). Also the candidate generation and re-localization based on segmentation as in [38] could be integrate in this pipeline improving the obtained results.

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] B. Li, T. Wu, and S.-C. Zhu, "Integrating context and occlusion for car detection by hierarchical and-or model," 2014.

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

[7] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3d voxel patterns for object category recognition," June 2015.

[8] E. Ohn-Bar and M. M. Trivedi, "Learning to detect vehicles by clustering appearance patterns," *T-ITS*, 2015.

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

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

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

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

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

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

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

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

[17] X. Wang, X. Bai, W. Liu, and L. J. Latecki, "Feature context for image classification and object detection." in *CVPR*, 2011.

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

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

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

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

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

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

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

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

[26] 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.
[27] J. Marin, D. Vázquez, A. López, J. Amores, and B. Leibe, "Random forests of local experts for pedestrian detection," in *ICCV*, 2013.
[28] 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.
[29] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," *IJCV*, 2009.
[30] J. Gall and V. Lempitsky, "Class-specific hough forests for object detection," in *CVPR*, 2009.
[31] D. Tang, Y. Liu, and T.-K. Kim, "Fast pedestrian detection by cascaded random forest with dominant orientation templates," in *BMVC*, 2012.
[32] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *IJCV*, 2008.
[33] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in *CVPR*, 2011.
[34] 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.
[35] D. Vázquez, A. López, J. Marín, D. Ponsa, and D. Gerónimo, "Virtual and real world adaptation for pedestrian detection," *T-PAMI*, 2013.
[36] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, "Bilateral filtering: Theory and applications," *Foundations and Trends in Computer Graphics and Vision*, vol. 4, no. 1, pp. 1–73, 2009.
[37] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 1967, pp. 281–297.
[38] C. Long, X. Wang, M. Yang, and Y. Lin, "Accurate object detection with location relaxation and regionlets relocalization," in *ACCV*, 2014.
[39] Q. Hu, S. Paisitkriangkrai, C. Shen, A. van den Hengel, and F. Porikli, "Fast detection of multiple objects in traffic scenes with a common detection framework," *T-ITS*, vol. 17, no. 4, pp. 1002–1014, 2016.

Alejandro González received the B.Sc. degree in Electronic Engineering from the National University of Colombia in 2010. He received his M.Sc. in Computer Vision and Artificial Intelligence in 2011 and his Ph.D. degree in Computer Science in 2015 at the Computer Vision Center (CVC/UAB). His research interests include pedestrian detection, spatiotemporal information, multi-modal detection. He is a student member of the IEEE.

David Vázquez received the B.Sc. degree in Computer Science from the Universitat Autònoma de Barcelona (UAB) in 2008. He received his M.Sc. in Computer Vision and Artificial Intelligence in 2009 and his Ph.D. degree in 2013 at the Computer Vision Center (CVC/UAB). He is currently a research scientist at CVC. His research interests include pedestrian detection, virtual worlds, domain adaptation and active learning. He is a member of the IEEE.

Jaume Amores received the Ph.D. degree from the Universitat Autònoma de Barcelona (UAB), in 2006. He has held positions in the Computer Vision Center (CVC) and in the Institut National de Recherche en Informatique et en Automatique (INRIA) and a Ramon y Cajal fellowship as part of the Advanced Driver Assistance Systems (ADAS) group. His research interests include machine learning and pattern recognition, object recognition and detection, and medical imaging.

Antonio M. López received the B.Sc. degree in Computer Science from the Universitat Politècnica de Catalunya (UPC) in 1992 and the Ph.D. degree in Computer Vision from the UAB in 2000. Since 1992, he has been giving lectures in the UAB, where he is now Associate Professor. In 1996, he participated in the foundation of the CVC, where he has held different institutional responsibilities. In 2003 he started the CVC's ADAS group, presently being its head. He is a member of the IEEE.