Because better detections are still possible: Multi-aspect Object Detection with Boosted Hough Forest

Carolina Redondo-Cabrera
carolina.redondoc@edu.uah.es
Roberto López-Sastre
robertoj.lopez@uah.es

University of Alcalá
GRAM
Alcalá de Henares, ES

Abstract

In this work, we proceed to deconstruct the HF learning model to investigate whether a considerable better performance can be obtained detecting multi-aspect object categories. We introduce the novel Boosted Hough Forest (BHF): a HF where all the decision trees of the forest are trained in a stage-wise fashion, by optimizing a global differentiable loss function with Gradient Boosting, and using the concept of *intermediate* Hough voting spaces. This is in contrast to the local optimization performed in each tree node during the training of a standard HF. We also show how the multiple aspects of the object categories can be incorporated into the learning model by simply augmenting the dimensionality of the Hough voting spaces of the BHF. This allows our approach to naturally infer the pose of an object, simultaneously with the detection, for example. The experimental validation, considering four different datasets, confirms that the performance of the HF is improved by the new BHF.

1 Introduction

In the last few years, Random Forests [3] (RFs) have attracted a lot of attention by the computer vision community. A RF is an ensemble classifier consisting of a set of randomized decision trees. During training, a binary weak classifier is learned for each non-leaf node. At runtime, test samples are passed through the nodes of the trees, and the output is computed by averaging the distributions learned at the reached leaf-nodes. Recently, RF-based approaches have been extensively used to try to solve a fundamental regression task: object localization. Gall *et al*. [9] propose the Hough Forest (HF) approach for object detection. This framework provides a way for the combination of the RF with the Hough transform (HT), in order to learn an ensemble regressor, which is able to efficiently detect object classes. And Schulter *et al*. [22] present a new approach, the Alternating Regression Forests (ARFs), which learns a RF by optimizing a global loss function over all trees.

Inspired by these works [9, 22], we introduce the Boosted Hough Forest (BHF). Essentially, the BHF is a HF where the decision trees are trained in a stage-wise fashion, by optimizing a global loss function. As in [22], the trees grow until a maximum depth is reached, but under the minimization of the global loss, which controls the performance of

the whole forest. This is in contrast to the local optimization done during the learning of every tree node in the traditional HF [9].

We formulate the training of the BHF in accordance with the empirical risk minimization principle, and model each depth of the forest as a stage-wise *complete* Hough voting based regressor. This is also in contrast to the ARFs [22]. During the learning of an ARF, the loss for each training sample is calculated using a residual, which is obtained by evaluating a RF at the current depth of the forest. Even when the ARF is applied to an object detection problem, the actual detection performance is not considered to update these residuals.

In the BHF framework, we provide an alternative way for the computation of the residuals. In a BHF, each iteration, *i.e*. each depth of the forest, is interpreted as a stage-wise HF weak object detector. This implies that for each iteration we are able to realize a complete detection process, using intermediate Hough spaces, and employing a Hough voting mechanism for the regression of the object center in the images. It is now the difference between the maximum in these intermediate Hough spaces and the ground truth information of the training samples, what defines the residuals for our Gradient Boosting based learning. Our iterative training procedure alternates between splitting data in the tree nodes and growing the forest by one depth level and evaluating the global loss for all training samples.

In addition, in order to further improve the detection performance for multi-aspect object categories, we show how the BHF can be extended to deal with this problem. The solution is easy: augment the dimensionality of the Hough voting spaces. In this extra dimension, the aspect can be encoded. This allows us to enforce consistency of the votes for each aspect separately. For instance, a BHF for detecting cars can be trained to deal with two views (frontal/rear vs. left/right) simultaneously, having a separate Hough voting space per aspect. This way, the BHF is able to recover the idea of the Deformable Part Model [8] (DPM) of having a root filter per aspect of the class. Nothing changes during the training of the BHF when multiple aspects are integrated: the residual of each training sample is computed considering only its corresponding aspect in the augmented intermediate Hough voting space.

Our results in four different datasets confirm that the performance of the BHF is superior to the one reported by the HF and the ARF for the problem of object detection and pose estimation.

The rest of the paper is organized as follows. We present an overview of the related works in Section 2. In Sections 3 and 4, we introduce the BHF and its extension to multi-aspect object detection, respectively. Section 5 includes the experimental validation. We conclude in Section 6.

2 Related Work

There have been many successful ideas over the last years in the field of object detection [5, 8, 12, 26]. Indisputably, one of them is the Implicit Shape Model (ISM) [15, 16], which constitutes the basis for several extensions in the following years (*e.g*. [9, 24]). The ISM [15, 16] combines the ideas of appearance codebooks and the Generalized Hough Transform. During training, it augments each visual word with the spatial distribution of the displacements between the object center and the respective visual word location. At detection time, these spatial distributions are converted into Hough votes, within the Hough transform, in order to identify the object center hypothesis .

The HF model have been presented in [9] as a variant of the ISM, inspiring numerous

applications and extensions (*e.g.* [7, 20]). In contrast to an ISM, at test time the HF extracts local features densely, instead of just on interest points, resulting in the aggregation of more evidence and thus increased robustness. The fast identification of local object parts is made possible by the use of Random Forests (RF) [3] for the visual vocabulary construction. Contrary to a standard HF-based approach, our forests grow in a stage-wise fashion, where a global loss is minimized for each depth of the forest, using the residuals obtained by a Hough voting based object detection at each iteration.

Our learning strategy for the BHF differs markedly from other popular methods which use RF [4, 11, 19, 23, 25]. Note that all these RF-based approaches are based on a local minimization process in the nodes. Only recently, Schulter *et al.* [22] have introduced the ARF model, which incorporates to the RF the idea of using a global loss minimization process to govern how the trees grow. Our BHF has been directly inspired by [22]. However, there are some clear differences between an ARF and our BHF. Specifically, the BHF constitutes a regression model based on an ensemble of trees trained in a stage-wise fashion by optimizing a global differentiable loss function with Gradient Boosting [14]. Unlike [22], we introduce a complete detection process at each depth level of the forest. Our approach employs intermediate Hough voting spaces to make the forest predictions in order to compute the residuals used by the Gradient Boosting optimization. That is, we define the residuals as the difference between the maxima detected in the intermediate Hough spaces and the ground truth information of the training samples. In an ARF, neither intermediate Hough spaces, nor whole forest object detections are used. Their pseudo-targets are calculated per training sample via the given loss and the current prediction of the forest for the mean of the offset of the patches. Additionally, we enrich the BHF with the concept of augmented Hough voting spaces in order to be able to deal with the problem of multi-aspect object detection and pose estimation, simultaneously.

3 Boosted Hough Forest

A BHF is essentially a HF with a different training algorithm. Following the original HF formulation of [9], in a BHF $\mathcal{F}$, we define a set of N binary decision trees $\mathcal{T}_n(\mathcal{P}) : \mathcal{P} \rightarrow \mathcal{H}$, where $\mathcal{P} \subset \mathbb{R}^d$ is the d-dimensional feature space and $\mathcal{H} \subset \mathbb{R}^H$ represents the Hough space where the hypotheses are encoded. Within the context of object detection, this Hough space lets us recover the hypotheses for the object location at multiple scales in an image. So, each object hypothesis $\mathbf{h} \in H$ can be defined as $\mathbf{h} = (x_h, y_h, s_h)$, where x_h and y_h encode the position of the object and s_h identifies the scale. One can learn a BHF $\mathcal{F}$ for object detection, from a set of sampled image patches $\mathcal{P}_t = \{(\mathcal{A}_t, c_t, d_t)\}$. $\mathcal{A}_t = \{A_t^1, A_t^2, ..., A_t^C\}$ represents the appearance of the training patch $\mathcal{P}_t$, where A_t^j is the appearance of the j^{th} channel. $c_t \in \{0, 1\}$ is the class label: 1 for a foreground patch, and 0 for the one extracted from the background. d_t encodes the relative 2D location of the object center with respect to the sampled patch.

While a forest is learned, any patch $\mathcal{P}_t$ can be propagated through it, following the path from the root node to the leaves according to the tests that take place at each node. Leveraging this fact, in a BHF we consider each depth d of the forest as a weak object detector, and like in the ARF model [22], the gradient of a loss, for each training sample, can be calculated and exploited to optimize a global loss function over the whole forest in the next stage $d+1$.

Here we follow a Gradient Boosting [14] formulation, as proposed for the ARFs [22]. A BHF, like any other boosting based approach, combines weak learners into a single strong learner, in an iterative process. Therefore, the principal idea behind the BHF learning con-

sists in constructing the new base-learners to be maximally correlated with the negative gradient of the loss function associated to the whole forest. That is, in a BHF each depth level is built to minimize the residuals of the preceding level of the forest.

So, our goal is to learn a model F that predicts values $\hat{\mathbf{d}}_\mathbf{t} = F(\mathcal{P}_t)$ for each patch, minimizing a global differentiable loss function. In our BHF, the forest prediction corresponds to an object center, which lets us compute a relative offset for each patch, *i.e.* $\hat{\mathbf{d}}_\mathbf{t}$. At each depth d of the forest, for $1 \leq d \leq D_{max}$, being D_{max} the maximum tree depth, the BHF improves $F_{d-1}(\mathcal{P}_t)$ by constructing a new model that adds an estimator h_d to compensate the shortcomings, *i.e.* the gradients, of the existing weak learner, and to provide a better model $F_d(\mathcal{P}_t) = F_{d-1}(\mathcal{P}_t) + h_d(\mathcal{P}_t)$. During learning, for the depth d, we formulate the following greedy stage-wise optimization,

$$\arg\min_{\varphi_d} \sum_{\{\mathcal{P}_t, d_t\}} \mathcal{L}(d_t; F_{d-1}(\mathcal{P}_t, \phi) + h_d(\mathcal{P}_t, \varphi_d)), \tag{1}$$

where $F_{d-1}(\mathcal{P}_t, \phi)$ is the BHF trained up to depth $d-1$, ϕ represents all the parameters optimized up to level $d-1$, and φ_d contains the parameters to be optimized at depth d. In our case, φ_d represents all the test-based appearance function parameters for all the nodes of the forest. As in the original HF (see [9] for more details), each test-based appearance function of a BHF is characterized by the following parameters: an appearance channel, two pixel coordinates, and a threshold for the data splitting.

We start with an initial regressor $F_0 = h_0(\mathcal{P}_t, \varphi_0)$, which corresponds to the N root nodes of the trees. Each iteration d adds a new level of depth. Therefore, the regressor $F_{d-1}(\mathcal{P}_t, \phi)$, trained up to depth $d-1$, gives a prediction for each training patch $\mathcal{P}_t$. With these predictions, we proceed to compute the residuals for each training patch $\mathcal{P}_t$ as follows,

$$r_{td} = -\left[\frac{\partial \mathcal{L}(d_t; F(\mathcal{P}_t))}{\partial F(\mathcal{P}_t)}\right]_{F(\mathcal{P})=F_{d-1}(\mathcal{P})}. \tag{2}$$

Equation 2 corresponds with the negative gradient of the loss w.r.t. the output of the current regressor, *i.e.* $F(\mathcal{P}_t)$. In our model, we choose the squared loss defined by,

$$\mathcal{L}(d_t; F(\mathcal{P}_t)) = \frac{1}{2}(d_t - F(\mathcal{P}_t))^2. \tag{3}$$

So, the residuals defined in Eq. 2 can be calculated as

$$r_{td} = d_t - F_{d-1}(\mathcal{P}_t). \tag{4}$$

The question is now: How does $F_{d-1}(\mathcal{P}_t, \phi)$ obtain the *weak* prediction $\hat{\mathbf{d}}_\mathbf{t}$ for the relative offset of each patch $\mathcal{P}_t$? For doing so, we introduce here the concept of *intermediate* Hough voting spaces. In our BHF model, we consider each depth of the forest as a weak object detector, so we can estimate the object position in an intermediate Hough space incrementally, while the trees are growing. This is in contrast to the incremental process of the ARFs [22], where no complete object detection is performed.

Then, given a training patch $\mathcal{P}_t$, centered at the position y, *i.e.* $\mathcal{P}_t(y)$, we proceed to pass it through the trees, trained up to depth $d-1$, to determine the set of leaf nodes $\{L_n\}_{n=1}^{N}$ reached. The rest of training patches distributed in the selected leaves cast their corresponding votes into the intermediate Hough space. The votes accumulated by a patch $\mathcal{P}_t(y)$ into the intermediate Hough space $\mathcal{H} \in \mathbb{R}^2$, are computed by adding $\frac{C_{L_n}}{|D_{L_n}| \cdot N}$ to the locations

Algorithm 1 Training a *Boosted Hough Forest*

Require: Labeled training set $\{\mathcal{P}_t, c_t, d_t\}_{t=1}^{T}$
Require: Number of trees N, maximum tree depth D_{max}
INIT F_0 using the N root nodes
for d from 1 to D_{max} **do**
  Check stopping criteria for all nodes in depth d
  for $\mathcal{P}_t(y)$ from $t = 1$ to T **do**
    Cast votes in the intermediate Hough space $\mathcal{H} \in \mathbb{R}^2$
    Find the object center prediction $\hat{\mathbf{h}}_\mathbf{t}$, *i.e.* the local maximum in $\mathcal{H}$
    Calculate the estimated offset using Eq. 5
    Update the residual r_{td} following Eq. 4
  end for
  Learn $h_d(\mathcal{P}_t, \varphi_d)$ using the set $\{\mathcal{P}_t, c_t, r_{td}\}_{t=1}^{T}$, and build the level d of the forest
  for $\mathcal{P}_t(y)$ from $t = 1$ to T **do**
    Propagate $\mathcal{H} \in \mathbb{R}^2$ from parent node to child node in each tree
  end for
end for

$\{y - d | d \in D_{L_n}\}$, where D_{L_n} represents the set of offsets for the patches that end up to the leaf node L_n. C_{L_n} is defined as the proportion of the foreground patches (*i.e.* patches with label $c_t = 1$) in the leaf L_n.

The current object center prediction $\hat{\mathbf{h}}_\mathbf{t}$ of the training patch $\mathcal{P}_t(y)$ can be obtained by finding the local maximum on its corresponding intermediate Hough space $\mathcal{H}$. Using $\hat{\mathbf{h}}_\mathbf{t}$, we can calculate the estimated offset $\hat{\mathbf{d}}_\mathbf{t}$ for the patch $\mathcal{P}_t(y)$ as

$$\hat{\mathbf{d}}_\mathbf{t} = y - \hat{\mathbf{h}}_\mathbf{t}. \tag{5}$$

Finally, this offset estimation is used to update the residual per patch r_{td}, using Equation 4. Once these residuals have been updated, the set $\{\mathcal{P}_t, c_t, r_{td}\}_{t=1}^{T}$ is used to train the base learner $h_d(\mathcal{P}_t, \varphi_d)$ which defines the depth d of the forest. In a BHF, $h_d(\mathcal{P}_t, \varphi_d)$ is learned like in a standard HF [9]: several random tests at each node are performed, and the patches are passed to the child nodes based on the minimization of a randomly chosen uncertainty. We use both the class and regression uncertainties introduced in [9].

Once $h_d(\mathcal{P}_t, \varphi_d)$ is optimized, we can proceed to build the next depth of the forest. The process is simple. The training patches are split into left and right child nodes, and their intermediate Hough spaces are propagated (they are scaled and summed) from the parent node to the new child nodes. At this point, the BHF uses $F_d(\mathcal{P}_t, \phi)$ to cast the new predictions for the depth $d + 1$. This procedure described continues until any of the stopping criteria is reached. In Algorithm 1, we summarize the complete training procedure of a BHF.

4 Multi-aspect Object Detection with BHF

Hough voting based methods for object detection work by allowing image features to vote for the location of the object center. While this representation allows for parts observed in different training instances to support a single object hypothesis, it also produces false positives by accumulating votes that are consistent in location *but* inconsistent in the aspect,

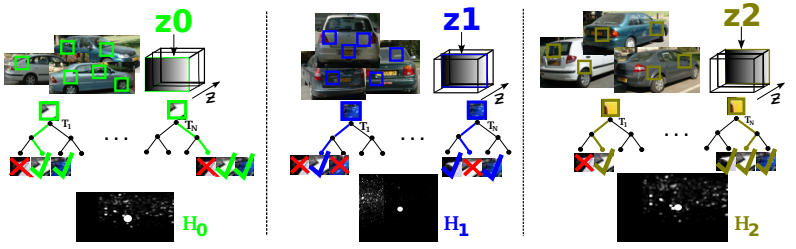


Figure 1: Toy example considering three aspects of the category car. Our model augments the Hough spaces by adding a dimension $\mathcal{Z}$, which encodes the object aspect. The training patches are passed through the trees to determine a leaf node. Only the patches with the same aspect z in the reached leaf cast probabilistic votes in the corresponding $\mathcal{H}_z$.

like for example the viewpoint. Here we propose an extension for the BHF, with the aim of improving its detection performance for multi-aspect objects: the concept of *augmented* Hough voting spaces.

A BHF employs intermediate Hough voting spaces to cast the forest predictions in order to compute the residuals used by the Gradient Boosting optimization. It is possible to enforce the consistency of the votes in these intermediate Hough spaces for each object category aspect. See the idea in Figure 1. We propose to augment the dimensionality of the intermediate Hough spaces, $\mathcal{H} \times \mathcal{Z}$. Each element z in the extra dimension $\mathcal{Z}$ summarizes global appearance changes caused for example by viewpoint changes or deformations of the object shape. This way, we only allow the patches to vote in the component z to which they belong to. This guarantees an alignment of the training data during the voting, which results beneficial for the object localization task.

During training, we associate each training patch $\mathcal{P}_t$ to a single aspect z_t. This association groups the training data into $|\mathcal{Z}|$ disjoint groups. In each depth d, we build an intermediate and *augmented* Hough voting space, defined by $\{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{|\mathcal{Z}|}\}$. Given a patch $\mathcal{P}_t$, extracted from the position y, *i.e.* $\mathcal{P}_t(y)$, we pass it through the trees to determine the set of leaf nodes $\{L_n\}_{n=1}^{N}$ where it arrives. Only the patches in these leaves with the same aspect z as $\mathcal{P}_t(y)$ are allowed to vote in the space $\mathcal{H}_z$. The votes accumulated by a patch $\mathcal{P}_t(y)$ in its corresponding Hough space $\mathcal{H}_z$, are computed by adding $\frac{C_{L_n}}{|D_{L_n}| \cdot N}$ to the locations $\{y - d | d \in D_{L_n}\}$. Note that now D_{L_n} represents the set of offsets for the patches in the leaf node L_n *but* with the same aspect z as $\mathcal{P}_t(y)$. The current object center prediction $\hat{\mathbf{h}}_\mathbf{t}$ of the training patch $\mathcal{P}_t(y)$ is obtained by finding the local maximum in its corresponding intermediate Hough space $\mathcal{H}_z$. Once again, using $\hat{\mathbf{h}}_\mathbf{t}$, we calculate the estimated offset $\hat{\mathbf{d}}_\mathbf{t}$ for the patch $\mathcal{P}_t(y)$ using Eq. 5, while the residual is updated via Eq. 4.

During inference, we also augment the hypothesis space $\mathcal{H}$ by $\mathcal{Z}$, so as to enforce consistency of the votes by the different aspects of the objects. This is done by only allowing votes that agree on the values of z to support a single detection hypothesis. We define now a hypothesis as $\mathbf{h}(x_h, y_h, s_h, z_h)$, where z_h identifies the aspect. The forest estimation is then computed by aggregating votes into the augmented $\mathcal{H} \times \mathcal{Z}$ space, $\{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{|\mathcal{Z}|}\}$, where $\mathcal{H}_j \in \mathbb{R}^2$. Moreover, to deal with the different scales of objects, we resize the test image by a set of scale factors $\{s_1, s_2, \ldots, s_S\}$, and compute their corresponding voting spaces $\{\{\mathcal{H}_1^1, \mathcal{H}_2^1, \ldots, \mathcal{H}_{|\mathcal{Z}|}^1\}, \ldots, \{\mathcal{H}_1^S, \mathcal{H}_2^S, \ldots, \mathcal{H}_{|\mathcal{Z}|}^S\}\}$.

For a particular scale s and aspect z, we first project all votes on the (x, y) positions of the

subspace $\mathcal{H}_z^s$ and compute the total score for an object hypothesis $\hat{\mathbf{h}}$ as

$$S(\hat{\mathbf{h}}, z, s) = \sum_{\mathcal{I}_i \in I} \sum_{t \in T} V(\hat{\mathbf{h}}, z, s | \mathcal{P}_t), \tag{6}$$

where $V(\hat{\mathbf{h}}, z, s|\mathcal{P}_t)$ represents the votes cast by all patches $\{\mathcal{P}_t\}_{t=1}^T$ from the training image $\mathcal{I}_i$ which aspect $z_t = z$. To obtain the final detections, we identify the maxima in the Hough spaces $\{\mathcal{H}_z^s\}$ and use non-maximum suppression to consolidate the object localizations. Each object hypothesis is then obtained as,

$$\mathbf{h}(x_h, y_h, s_h, z_h) = \arg\max_{\hat{\mathbf{h}}, z, s} S(\hat{\mathbf{h}}, z, s), \tag{7}$$

thereby the BHF also recovers the aspect z_h of the object. If we assume that the aspect encodes the discrete viewpoint of a category, the BHF is able to naturally perform a simultaneous object detection and pose estimation.

Moreover, we can enrich our approach to perform a coarse-to-fine pose estimation, *i.e.* to infer the continuous pose of the objects. Given an object hypothesis $\mathbf{h}(x_h, y_h, s_h, z_h)$, we proceed to scale and translate the aspect-specific bounding box (BB) to the identified object location. Then, within this BB, image patches are densely collected and passed again through the trees. The training patches of the leaves with $z = z_h$ are identified and used to vote for an object center hypothesis. We index the training images from which these patches come from. We finally propose to estimate the fine pose of the object $\hat{\theta}$, recovering the pose of the *most similar* training image, $\mathcal{I}_i^*$, which is defined as the training image with the largest contribution of patches to the object center hypothesis $\hat{\mathbf{h}}$.

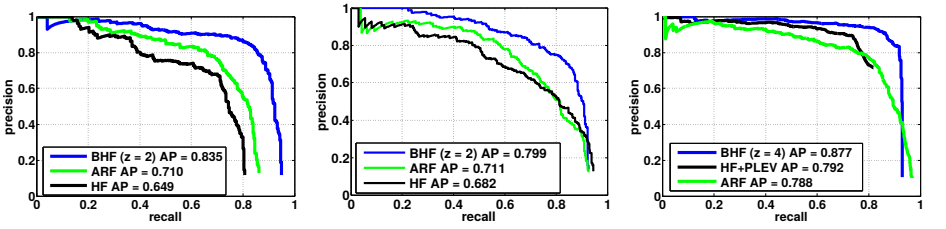
5 Experiments

5.1 TUD and TUD Multiview Pedestrian databases

We start the experimental validation using the *TUD-Pedestrian* dataset (TUD) [1]. We here compare the performance of a HF [9], an ARF [22] and the novel BHF in the same taks: pedestrian detection. For a fair comparison, we endow all methods with 10 trees, each having a maximum depth of 30, and 20000 random tests per node. Note that these parameters reported the best results for the ARF and HF in [22]. We use a training data set consisting of 16000 foreground and 16000 background patches randomly extracted. This database has no pose annotation, so we initialize the aspect z for each training example applying a clustering with K-means and using HOG features.

We present the detection results in Figure 2(a). For this experimental setup, the BHF gets a gain of 18% and 12% in terms of average precision (AP) over the HF and the ARF, respectively. In Figure 3(a) we show how the detection performance of a BHF evolves for different $|\mathcal{Z}|$ values. One can see that even a BHF where no multiple aspects are considered, *i.e.* a BHF with $|\mathcal{Z}| = 1$, achieves a gain of 11.3% in terms of AP over the ARF. This confirms that our main contribution, *i.e.* the integration of the intermediate Hough voting based regressors with the global loss minimization for the forest, results beneficial.

We now proceed to evaluate these models using the *TUD Multiview Pedestrian dataset* (TUD-Multiview) [2]. This dataset is particularly interesting, because it has been designed to deal with the problem of multi-view (multi-aspect) pedestrian detection. We use a setup



(a) (b) (c)

Figure 2: Precision-Recall curves for HFs, ARFs and BHFs on the (a) TUD-Pedestrian, (b) TUD Multiview Pedestrian and (c) WCV datasets.

Table 1: Results on WCV dataset.

					ARF		BHF $\|\mathcal{Z}\| = 4$	
	azimuth	zenith	azimuth	zenith	azimuth	zenith	azimuth	zenith
MAE(°)	36.4	8.7	**25.8**	**3.2**	81.2	5.1	30.7	4
AOS	–	–	0.763	0.791	0.5164	0.7863	**0.805**	**0.875**

Table 2: Results on PASCAL3D+.

methods	AP	AOS azimuth	AOS zenith	AVP azimuth	AVP zenith
HF	0.163	0.113	0.163	0.090	0.158
ARF	0.182	0.118	0.181	0.088	0.177
BHF ($\|\mathcal{Z}\| = 4$)	0.187	0.154	0.187	**0.112**	0.183
DPM [8]	**0.266**	–	–	–	–
BHF ($\|\mathcal{Z}\| = 4$) + verif.	0.265	**0.196**	**0.264**	0.110	**0.258**

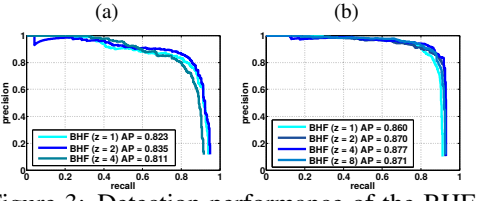


Figure 3: Detection performance of the BHF for different values of $|\mathcal{Z}|$ on (a) TUD and (b) WCV datasets.

consisting of 1600 images (200 examples per viewpoint) for training, 248 images for validation, and 248 images for testing. We use again the same parameters to train the three approaches. We initialize the aspect z for each training example applying K-means over HOG features. We report the results in Fig. 2(b). As before, our BHF outperforms the HF and ARF approaches. The BHF achieves a gain of 11.7% and 8.8% over the HF and the ARF, respectively.

Note that we report these results with $|\mathcal{Z}| = 2$. Although there are eight discrete viewpoints annotated in this dataset (front, left-front, left, etc.), we let the clustering discover the two aspects (see Figure 4(b) for the aspects discovered). With this strategy we obtain the best performance. In our experiments, we have noticed that if $|\mathcal{Z}|$ is fixed to 8, according to the ground truth information, the AP drops to 0.681. We believe that the type of viewpoint annotation provided is not convenient for an approach based on a patch voting strategy. In terms of detection performance, it is more adequate to align the pedestrian training examples into two sets with different aspect ratios, like the ones shown in Figure 4(b). This reveals that for the particular case of pedestrians, it is difficult to define the *best* aspects due to the large variability of the human body articulation changes.

5.2 Weizmann Cars Viewpoint dataset

We now evaluate the detection and pose estimation performance of our model using the *Weizmann Cars Viewpoint dataset* (WCV) [13]. This benchmark contains 1539 images of cars, divided in 22 car models partitioned into three sets s1, s2, and s3. As in [13], we use one set for testing (s3) and the other two for training (s1 and s2). For our BHF, we initialize the z aspect using the ground truth annotation provided in this dataset: we discretize the azimuth angle using $|\mathcal{Z}|$ bins, such that the bin centers have an equidistant spacing of $\frac{360}{|\mathcal{Z}|}$.

We show in Fig. 2(c) that the BHF significantly improves the detection performance achieved by the state-of-the-art reported in this dataset: the HF+PLEV [21]. As well, the BHF improves the AP of an ARF model for the same set of parameters. We use 4 aspects ($|\mathcal{Z}| = 4$) in order to incorporate the viewpoint information (frontal, rear, left and right) into our model. Note that our BHFs outperform the ARF and HF+PLEV performances for other viewpoint discretizations, see Figure 3(b). This confirms that for rigid categories, like the cars, the integration of the multi-aspect extension into the BHF model results beneficial. Moreover, this reveals that for the problem of car detection, the viewpoint information really matters.

Table 1 includes quantitative pose estimation results. Our model achieves a Mean Angular Error (MAE) equal to 30.7° for the azimuth angle. This is 4.9° higher than the state-of-the-art presented in [21]. However, the MAE metric only considers the viewpoint accuracy for the correct detections, and this aspect makes this metric not adequate to compare two detectors casting different detections [10, 17]. Note that our BHF has an AP of 0.877, compared to the AP of 0.792 of the HF+PLEV. In order to establish a fair comparison, we use the Average Orientation Similarity (AOS) evaluation metric introduced in [10], which simultaneously evaluates the detection and pose estimation performance. Now, our model outperforms the performance achieved by the ARF and the HF+PLEV for both angles. We get a gain, w.r.t [21], of 4.2% and 8.4% for azimuth and zenith, respectively. Note that we have extended the ARF with our approach for the BHF for recovering the continuous pose. In summary, the BHF establishes the new state-of-the-art performance for both car detection and pose estimation in the WCV dataset.

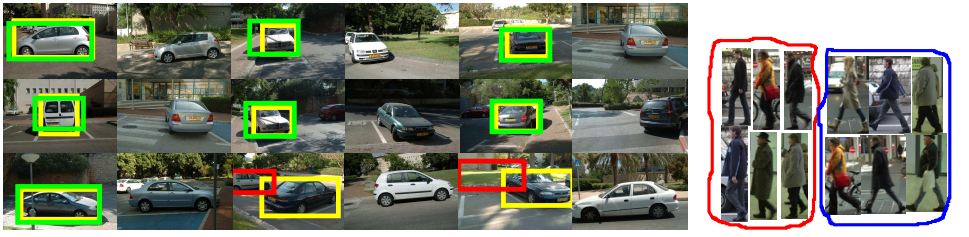
It is noteworthy that the integration of the augmented Hough spaces based on the aspects of the objects allows us to naturally infer the continuous pose by a straightforward coarse-to-fine strategy. In contrast to the HF+PLEV [21], which is also a HF based approach for simultaneous detection and pose estimation, we do not: a) need an additional uncertainty measure for the pose; b) use complex test functions, but simple pixel comparisons; c) need to incorporate any refinement for the regression of the pose using a kernel density estimation strategy. In summary, a BHF is a more compact and precise model. Fig. 4(a) shows some qualitative results for this dataset.

5.3 PASCAL3D+ dataset

We finally proceed to evaluate the models using a more challenging dataset: the novel *PASCAL3D+ dataset* [27]. Here we report detailed results for the particular class car, although in the supplementary material the rest of classes are included.

For the all the experiments, we fix the forest size to 35 trees, with a maximum depth of 20. In each node 20000 binary tests are considered. 10 positive and 10 negative patches are randomly extracted from each training image. The number of aspects considered is 4. To improve the detection results, we re-train the forest introducing hard negative samples.

In the PASCAL VOC datasets [6], as it is described by Gall *et al*. in [9], the HF based approaches achieve considerably lower performance than the state-of-the-art. These HF based methods struggle with the variation of the data that contains many truncated examples. Although these models have a very good recall, this usually comes at the cost of a low precision. However, techniques like the additional verification step proposed in [18, 20], where an extra detector is adapted to re-score the hypotheses of the forests, can be combined with the HF based approaches to improve their detection results. Therefore, we incorporate a verification step using a car DPM [8] trained on the PASCAL VOC 2007. We first run the DPM *only*



(a) (b)

Figure 4: (a) Qualitative results on WCV dataset. Columns 2,4 and 6 show the training images selected to estimate the azimuth and zenith. Ground truth in yellow, estimations in green and wrong detections in red. (b) Clusters for $|\mathcal{Z}| = 2$ with TUD-Multiview dataset.

over our BHF hypotheses. If the DPM also detects a car, we re-score (multiplying by 10) our corresponding BHF hypothesis.

Table 2 reports the results achieved by a BHF with and without the verification step. Again, the BHF without verification obtains an AP slightly higher than the reported for the HF and ARF approaches. When we incorporate the verification step to our model (see Table 2 row 5), we do not lose accuracy compared to the DPM [8], obtaining our BHF a gain, in terms of AP, of 7.8%, with respect to the BHF without verification.

For the pose estimation quantitative results, we use the AOS and the Average Viewpoint Precision (AVP) [27] metrics. For the AVP, we fixed a threshold of 15°[1]. Again, our BHF outperforms the pose estimation results obtained by the ARF and HF models, for both azimuth and zenith angles, and considering both evaluation metrics. Note that our verification step also improves the pose estimation results.

6 Conclusions

We have introduced the BHF. This novel learning model controls the performance of the forest as a whole, where the decision trees are trained in a stage-wise fashion, while a global loss is minimized. Essentially, the learning strategy follows a Gradient Boosting approach, where the residuals per sample are updated via a regression performed in each level of the forest, with intermediate Hough voting spaces for object detection. The experiments show that the BHF exhibits a higher performance than the HF and ARF. We have also shown how our BHF is able deal with the problems of multi-aspect object detection and pose estimation. These task are accomplished by an augmentation of the dimensionality of the Hough spaces, where each category aspect is represented in the novel dimension.

Acknowledgements. This work is supported by projects CCG2014/EXP-054, TEC2013-45183-R and SPIP2014-1468.

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, 2008.

[1] This value was suggested by the authors of the PASCAL3D+ dataset

[2] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. In *CVPR*, 2010.

[3] L. Breiman. Random forests. In *Machine Learning*, 2001.

[4] Antonio Criminisi, Jamie Shotton, Duncan Robertson, and Ender Konukoglu. Regression forests for efficient anatomy detection and localization in ct studies. In *International MICCAI Conference on Medical Computer Vision: Recognition Techniques and Applications in Medical Imaging*, 2011.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2010.

[7] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3D face analysis. *IJCV*, 2013.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[9] J. Gall, A. Yao, N. Razavi, L. van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. In *PAMI*, 2011.

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

[11] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *ICCV*, 2011. ISBN 978-1-4577-1101-5.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[13] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and continuous pose estimation. *IVC*, 2012.

[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York Inc., 2001.

[15] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, 2003.

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

[17] R. J. Lopez-Sastre, Tuytelaars. T., and S. Savarese. Deformable part models revisited: A performance evaluation for object category pose estimation. In *ICCV 2011, 1st IEEE Workshop on Challenges and Opportunities in Robot Perception*, 2011.

[18] S. Maji and J. Malik. Object detection using a max-margin hough transform. In *CVPR*, 2009.

[19] A. Montillo, J. Shotton, J. Winn, J.E. Iglesias, D. Metaxas, and A. Criminisi. Entangled decision forests and their application for semantic segmentation of ct images. In *IPMI*, 2011.

[20] N. Razavi, J. Gall, and L. Van Gool. Scalable multi-class object detection. In *CVPR*, 2011.

[21] C. Redondo-Cabrera, R. Lopez-Sastre, and T. Tuytelaars. All together now: Simultaneous object detection and continuous pose estimation using a hough forest with probabilistic locally enhanced voting. In *BMVC*, 2014.

[22] S. Schulter, C. Leistner, P. Wohlhart, P. M. Roth, and H. Bischof. Alternating regression forests for object detection and pose estimation. In *ICCV*, 2013.

[23] S. Schulter, C. Leistner, P. Wohlhart, P. M. Roth, and H. Bischof. Accurate object detection with joint classification-regression random forests. In *CVPR*, 2014.

[24] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. In *CVPR*, 2006.

[25] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[26] P. Viola and M. Jones. Robust real-time object detection. In *IJCV*, 2001.

[27] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *WACV*, 2014.