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

In the last few years, many adaptations of the Hough Forest (HF) [2] approach for object detection have been proposed. In this work, we proceed to deconstruct the HF learning model to investigate whether a considerable better performance can be obtained detecting multi-aspect objects.

Inspired by [4], 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.

While the tress are learned, any image patch $\mathcal{P}_t$ can be propagated through them, following the path from the root to the leaves. 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 [4], 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 $d+1$.

Following a Gradient Boosting formulation, at each depth d of the forest, the BHF improves the regressor of the previous iteration $F_{d-1}(\mathcal{P}_t)$, by learning a new model $F_d(\mathcal{P}_t) = F_{d-1}(\mathcal{P}_t) + h_d(\mathcal{P}_t)$, in which the regressor $h_d(\mathcal{P}_t)$ is added.

In our BHF, the forest prediction corresponds to an object center, which lets us compute a relative offset for each patch, *i.e.* $\hat{\mathbf{d}}_\mathbf{t}$. So, following an iterative training procedure, we start with an initial regressor F_0, which corresponds to the N root nodes of the trees. Each iteration d adds a new level of depth to the forest. A regressor $F_{d-1}(\mathcal{P}_t, \phi)$ trained up to $d-1$ gives a prediction for each training patch $\mathcal{P}_t$, and these predictions are use to compute the residuals,

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

which will be used to train the base learner $h_d(\mathcal{P}_t)$, *i.e.* the depth d of the forest, according to the original HF formulation [2].

The question is now, how does $F_{d-1}(\mathcal{P}_t, \phi)$ obtain the *weak* prediction $\hat{\mathbf{d}}_\mathbf{t}$ for each patch $\mathcal{P}_t$? Here, in contrast to the ARFs [4], we introduce the concept of *intermediate Hough space*. The BHF interprets each depth of the forest as a stage-wise HF weak object detector.

During training, each training patch $\mathcal{P}_t$ traverses the trees and casts votes to the intermediate Hough space $\mathcal{H} \in \mathbb{R}^2$ based on the location stored in the "leaves" at depth $d-1$. The current object center prediction $\hat{\mathbf{h}}_\mathbf{t}$ of the training patch centered at position y, *i.e.* $\mathcal{P}_t(y)$, 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 $\hat{\mathbf{d}}_\mathbf{t}$ as $\hat{\mathbf{d}}_\mathbf{t} = y - \hat{\mathbf{h}}_\mathbf{t}$. Finally, this estimation is used to compute the residual using Eq. 1. In Algorithm 1, we summarize the complete training procedure of the BHF.

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}
1: INIT F_0 using the N root nodes
2: **for** d from 1 to D_{max} **do**
3: Check stopping criteria for all nodes in depth d
4: **for** $\mathcal{P}_t(y)$ from $t=1$ to T **do**
5: Cast votes in the intermediate Hough space $\mathcal{H} \in \mathbb{R}^2$
6: Find the object center prediction $\hat{\mathbf{h}}_\mathbf{t}$
7: Calculate the estimated offset: $\hat{\mathbf{d}}_\mathbf{t} = y - \hat{\mathbf{h}}_\mathbf{t}$
8: Update the residual r_{td} following Eq. 1
9: **end for**
10: 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
11: **for** $\mathcal{P}_t(y)$ from $t=1$ to T **do**
12: Propagate $\mathcal{H} \in \mathbb{R}^2$ from parent node to child node in each tree
13: **end for**
14: **end for**

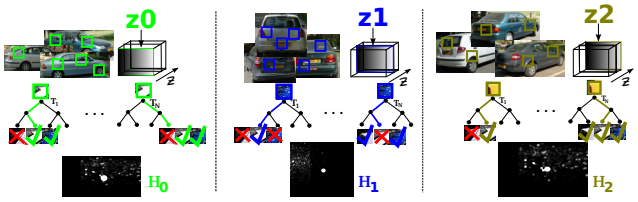
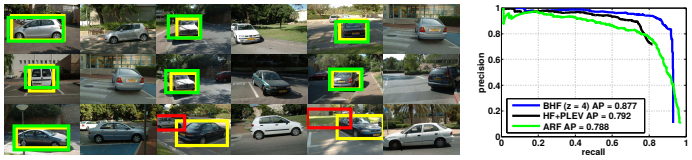


Figure 1: Toy example for 3 aspects of the car category. Our BHF model augments the Hough spaces by adding a dimension $\mathcal{Z}$, which encodes the object aspect. The training patches are passed through the trees to determine a leaf node. Only the patches with the same aspect z in the reached leaf cast probabilistic votes in the corresponding voting space $\mathcal{H}_z$.

In order to further improve the detection performance for multi-aspect objects, we show how the BHF can be naturally extended to deal with this problem. The solution is simple: augment the dimensionality of the Hough voting spaces, with one dimension to encode the different aspects. This allows us to enforce consistency of the votes for each object category 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. 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 dimension in the augmented intermediate Hough voting space (see Fig. 1).

We also extend our BHF model for object detection to simultaneously predict the continuous pose of the object. The pose is recovered from the training sample which contributes the most to the final hypothesis.

In our experimental validation, we have observed that the detection performance of our BHF increases with respect to both the traditional HF and the ARF models. We also show how augmenting the dimensionality of the Hough voting spaces, the BHF is able to deal with the problem of multi-aspect object detection and pose estimation. Figure 2 shows qualitative and quantitative results obtained by the BHF on the Weizmann Cars Viewpoint dataset [3]. Additional results using the PASCAL3D+ [5] and some TUD Pedestrian [1] datasets can be found in the paper.



(a) (b)

Figure 2: Results on the Weizmann Cars Viewpoint dataset [3]. (a) Qualitative results. Columns 2,4 and 6 show the training images selected to estimate the pose. Ground truth in yellow, estimations in green and wrong detections in red. (b) Precision-Recall curves for HF, ARF and BHF.

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

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

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

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

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