Fast Pedestrian Detection by Cascaded Random Forest with Dominant Orientation Templates

Danhang Tang
http://www.iis.ee.ic.ac.uk/~dtang
Yang Liu
http://www.iis.ee.ic.ac.uk/~yliu
Tae-Kyun Kim
http://www.iis.ee.ic.ac.uk/~tkkim

Department of Electrical Engineering,
Imperial College,
London, UK

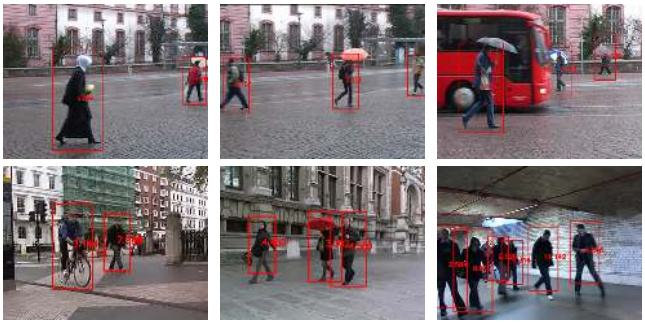

Figure 1: 1^{st} row: results of TUD pedestrian dataset; 2^{nd} row: results of our own video sequences. Numbers within bounding boxes indicate the voting scores.

In this paper, we present a new pedestrian detection method combining Random Forest and Dominant Orientation Templates(DOT) to achieve state-of-the-art accuracy and, more importantly, to accelerate run-time speed. Our method consists of a 2-level cascade: At first the scale space is divided into G overlapped groups, and a holistic detector(1^{st} level), designed with RF and our novel split function based on DOT, is applied on the first layer of each group and thus areas of interests are identified. After that a patch-based detector(2^{nd}), improved from HF [1] using the novel split function and clustering votes, is applied within these areas to achieve final bounding boxes.

Main contribution The prerequisite of this method is to adopt DOT as a descriptor, for its binary form is apt for bitwise operation. Besides orientations, we also encode the hue channel of colour information into binary format as another feature. Utilising DOT allows down-sampling images, which is the key reason for speeding up. However, since a significant amount of information(magnitude) is discarded, it loses some discriminative information. To compensate, we propose a novel similarity measurement to incorporate more dimensions. This measurement develops into a non-linear split function which better split the feature space whist maintaining the complexity of an axis-align split. This novel split function drastically improves both the detection *accuracy* of RF with 2-pixel tests on *DOT*, and the detection *speed* of RF with 2-pixel tests on *HOG* .

We define a template $\mathcal{T}$ as a n-dimension DOT sample selected from a positive training set $\mathbf{S}^P$. The similarity between $\mathcal{T}$ and any sample $\mathcal{S}$ can be measured with:

$$F(\mathcal{S},\mathcal{T}) = \sum_{\substack{P_d^{\mathcal{S}} \in \mathcal{S} \\ P_d^{\mathcal{T}} \in \mathcal{T}}} \delta(P_d^{\mathcal{S}} \otimes P_d^{\mathcal{T}} \neq 0), d = 1,...,n \tag{1}$$

where $P_d^{\mathcal{S}}$ and $P_d^{\mathcal{T}}$ refer to the d^{th} dimension of $\mathcal{S}$ and $\mathcal{T}$ respectively. $\otimes$ is the bitwise AND operation. δ is an impulse function that is zero except when any bit in $P_{\mathcal{S}} \otimes P_{\mathcal{T}}$ is 1. To accelerate this matching process, SSE optimization is employed similar to [2]. Therefore although this function measures the distance between samples in a feature space as a non-linear split, it is efficient since only binary bitwise operations and addition are involved.

With the measurement above, we can then define the split function h_i as:

$$h_i(\mathcal{S}) = \begin{cases} 0, & F(\mathcal{S},\mathcal{T}_i) \leq \tau_i \\ 1, & F(\mathcal{S},\mathcal{T}_i) > \tau_i \end{cases}, \tag{2}$$

where $\mathcal{T}_i$ means a chosen template and τ_i is a threshold of the i^{th} node.

During training, a set of positive $\mathbf{S}^P$ and negative $\mathbf{S}^N$ samples are used to construct a set of randomised decision trees. At the i-th non-leaf node,

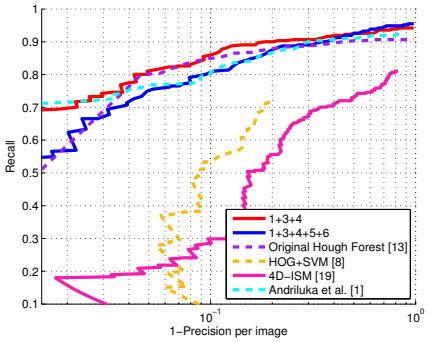

Figure 2: Comparison of accuracy with different components: 1. Dominant Orientations; 2. 2-Pixel Test; 3. Template Matching; 4. Dominant Colours; 5. Cascade; 6. Clustered Votes. (b) Comparison between our detector and State-of-the-art.

Method	Time(s)
Original Hough Forest	4.15
Our method(patch-based+DOT)	0.62
Our method(cascade)	0.33
Our method(cascade+clustering)	0.20

Table 1: Comparison of efficiency with 24 scales(0.17∼0.87) on 640x480 images.

we have these parameters:

$$\theta_i = \{\mathcal{T}_i, \tau_i\}, \quad \mathcal{T}_i \in \mathbf{S}_i^P, \tag{3}$$

where $\mathcal{T}_i$ is a template chosen from positive samples, and τ_i is a threshold. We randomly generate a set of θ_i, and select the optimal one in terms of information gain.

Although adopting DOT allows down-sampling in scanning-windows, we still need to perform dense classification and voting to obtain satisfying results. Thus it is a natural option to construct a cascade to filter out unlikely regions before performing the patch-based detector. The design of this cascade and training tricks are described in detail in the full version of our paper.

Result In the experiment section, we first revisit different methods of training RF and come up with an optimal combination. The rest tests are performed accordingly. We compare our method against the original HF and other state-of-the-art methods. Figure 2 shows that adopting DOT with orientation and colour has better accuracy than the original HF and achieve 85% correct rate at 10^{-1} 1-precision. Cascaded version achieves more than 20 times of speed improvement whilst keeping a comparable accuracy. (Table 1) Note that our speed optimisation is mainly done about features rather than scales, therefore it can be combined with those works optimising multi-scale detection and obtain further speed-up. Also, we emphasise the inherent benefits of our RF framework for scalability, quick training, multi-class or multi-part detection.

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

[2] S. Hinterstoisser, V. Lepetit, S. Ilic, P. Fua, and N. Navab. Dominant orientation templates for real-time detection of texture-less objects. In *CVPR*, 2010.