Group Cost-sensitive Boosting with Multi-scale Decorrelated Filters for Pedestrian Detection

Chengju Zhou
zhou0271@e.ntu.edu.sg
Meiqing Wu
meiqingwu@ntu.edu.sg
Siew-Kei Lam
assklam@ntu.edu.sg

School of Computer Science and Engineering
Nanyang Technological University (NTU), Singapore

Abstract

We propose a novel two-stage pedestrian detection framework that combines multi-scale decorrelated filters to extract more discriminative features and a novel group cost-sensitive boosting algorithm. The proposed boosting algorithm is based on mixture loss to alleviate the influence of annotation errors in training data and explores varying cost for different types of misclassification. Experiments on Caltech and INRIA datasets show that the proposed framework achieves the best detection performance among all state-of-the-art non-deep learning methods. In addition, the proposed approach runs 88X faster than the best performing method from the widely-known Filtered Channel Feature framework.

1 Introduction

Over the past decade, vision-based pedestrian detection has attracted plenty of attention in the research community [1]. However, [30] suggests that there is still a ten fold improvement needed before existing methods can match human perception. Even though the recently reported deep learning methods (e.g. [3, 27]) have demonstrated impressive performance, the high computational complexity of Convolutional Neural Network (CNN) based methods prohibit their applicability for real time processing on embedded systems with tight computational and energy constraints [4]. Hence, state-of-the-art pedestrian detection methods are still unable to meet the joint performance, runtime, and energy requirements imposed by practical systems (e.g. autonomous driving).

In this paper, we focus on improving the detection performance of non-CNN based methods, while at the same time, lowering the computational complexity. In particular, we propose a novel two-stage detection framework to boost the performance of decorrelated channel feature methods. Unlike LDCF (Local Decorrelated Channel Feature) [20], which learns single scale local decorrelated filters for each feature channel, the proposed method explores multi-scale local decorrelated filters to extract more discriminative features. Our proposed multi-scale filters not only improves the detection performance of existing decorrelated channel feature methods, but also contributes to lowering computational complexity due to the use of smaller filter sizes (proposed filter sizes are 2×2 and 3×3, while LCDF requires 5×5 filters).

To further improve the detection performance, we proposed a new boosting algorithm that explores different costs for varying misclassification types and exploits a mixture loss to alleviate the sensitivity to annotation errors. Concretely, the training samples are divided into groups based on the resolution, and each group is assigned different cost so that more emphasis is given to the harder samples in the training process. Mixture loss that combines the exponential and logistic loss is employed to deal with outliers, which have significant effects on the boosting algorithm. The proposed mixture loss not only maintains the efficiency of exponential loss, but also reduces its sensitivity to outliers caused by the annotation errors.

The main contributions of this paper are summarized as follows:

1) Our work is the first to learn multi-scale decorrelated filters from training samples. Unlike existing decorrelated filters method [20] that learns single scale filters, our proposed method can integrate richer local information and extract more discriminative features.

2) We propose a novel group cost-sensitive RealBoost algorithm based on a new mixture loss for training pedestrian detector. The training samples are divided into groups based on the resolution and different costs are assigned to each group. The different costs enforce the learning algorithm to pay more attention to the harder samples. A new mixture loss is proposed to alleviate the influence of annotation errors.

3) We perform extensive evaluations on the proposed framework using the widely known Caltech and INRIA datasets. Our proposed framework achieves the best detection performance (i.e. 14.63% log-average miss rate (MR) on Caltech dataset) among all the state-of-the-art non-CNN methods. In addition, the proposed method runs 88 times faster than the best performing method in the widely-known FCF framework (i.e. Checkerboards [29]).

2 Related Work

Existing pedestrian detection methods can be categorized into three families based on the classification strategy adopted: DPM (Deformable Part Models) variants [13, 14, 16], Convolutional Neural Network (CNN) [3, 11, 27], and Decision Forest (DF) [6, 10, 28, 29]. The current ranking on Caltech dataset [9] shows that the top-performing methods belong to the CNN and DF based approaches. The former adopts deep CNN to learn discriminative features and they typically require high-end GPUs to meet real-time constraints. Experiments undertaken in [4, 17] show that energy and thermal constraints will limit the maximum achievable accuracy and run-time of deep learning algorithms on embedded GPUs for applications, e.g. autonomous driving. As such, CNN-based methods are currently not well suited for realization on affordable and mass volume deployable embedded platforms that have tight computational constraints. In the remaining section, we will present a review of the DF based methods, which have achieved notable successes recently.

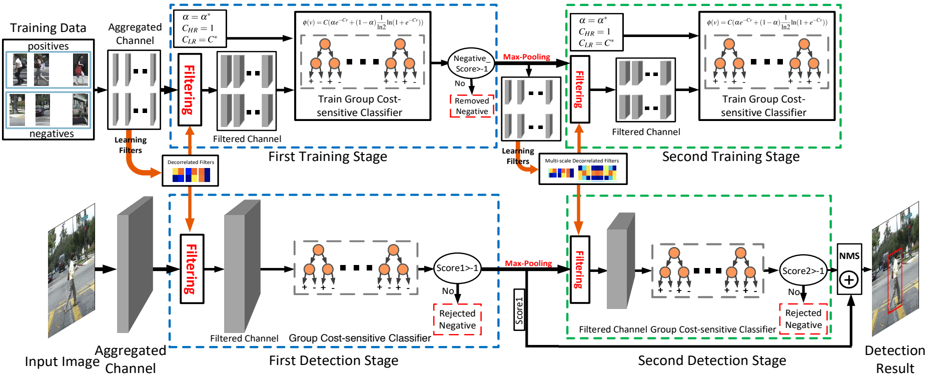


Figure 1: Proposed framework. The upper and lower parts are the training and testing procedures respectively. For each stage, the decorrelated filters are learned from feature channels and used to extract filtered channel features. Then the classifier is trained by exploiting proposed group cost-sensitive RealBoost algorithm. During testing, the aggregated channels are extracted from the images and fed into our multi-stage detector. The final score comprises of scores from the first and second stage and the detection result is obtained after Non-Maximum Suppression (NMS).

Current top-performing DF-based approaches belong to the Filtered Channel Feature (FCF) framework [29], which consists of three steps: aggregated feature channel extraction, filtering over feature channel and object classification. The aggregated feature channel includes 10 channels: 3 for LUV color, 1 for gradient magnitude, and 6 for Histogram of Oriented Gradients (HOG). These channels are used for extracting mid-level or high-level features through filtering. The Adaboost algorithm is then employed to train a cascade strong classifier. The existing FCF methods vary according to the filters adopted in the channel filtering step. For example, Checkerboards [29] uses a set of checkerboard pattern filters, while RotatedFilters [30] employs rotated filters based on the orientation of HOG channel. All of the above mentioned approaches adopt hand-crafted filters based on domain knowledge and are independent from the training data used in the learning process. To integrate prior information from training samples, LDCF [20] employs local decorrelation strategy to learn the filters. While LDCF has been shown to achieve notable gains over ACF (Aggregated Channel Feature) which use single pixel lookups in the aggregated channels as feature, its performance is still far lower than current top-performing DF-based approaches. Besides, LDCF exploits large decorrelated filter size (5×5) that leads to high computational complexity at the channel filtering step.

All of the above mentioned methods [10, 20, 28, 29, 30] focus on feature representation and employ cost-insensitive Adaboost learning algorithm [15] when training the classifier. The cost-insensitive Adaboost assumes that different types of misclassification have equal importance and hence are assigned identical cost. However, this assumption does not usually hold in real-world applications [18]. For instance, pedestrians that are far from the camera are more difficult to detect and localize due to lower resolution and image blurring. To achieve a better detection performance, more emphasis should be given to these harder samples. Various cost-sensitive boosting methods have been proposed to assign different costs to different types of misclassification [12, 19, 24, 25]. However, these algorithms are designed to deal with class-level cost-sensitivity by only assigning different costs to inter-class misclassification. In order to better handle multi-resolution pedestrian detection, [31] proposed a group cost-sensitive Adaboost which explores different costs for different resolution subsets from positive samples during training. However, the groups are formulated based on the resolution of positive training samples, and hence they are affected by the quality of the positive samples that are often subjected to annotation errors [30] as shown in Fig. 3(b). The negative samples with complex background have much larger intra-variants than positive samples but are not explored. Consequently, the detection performance of [31] is condiderably lower than existing FCF methods (e.g. Checkerboards [29]).

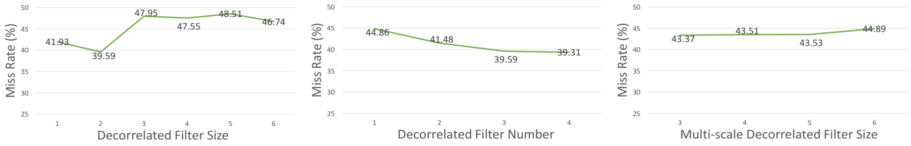
3 Proposed Method

In this section, we present our multi-stage pedestrian detection framework (see Fig. 1), which integrates the proposed multi-scale local decorrelated filters and group cost-sensitive RealBoost algorithm to simultaneously improve detection performance and runtime.

3.1 Multi-scale Local Decorrelated Filters

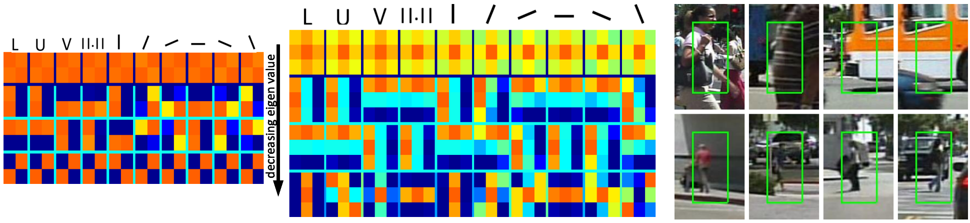
LDCF [20] applies feature transform to remove correlation in local neighbourhoods so that the filtered channel features are more suitable for orthogonal split instead of oblique split in decision trees. Specifically, LDCF uses the top 4 learned PCA eigenvectors extracted from 5×5 patch as filters and implements filtering over each aggregated channel. The filtered features are fed into a strong classifier learned by Adaboost. Compared to ACF [10], LDCF achieved higher performance on Caltech dataset. However, the decorrelated filters learned by LDCF is from single scale (size of 5×5) that can only extract discriminative features from a specific local neighbourhood, which limits their detection performance potential. To extract more discriminative features, we propose to learn multi-scale decorrelated filters from training samples and investigate the impact of different scales of filters on the detection performance. We setup a validation environment by splitting the Caltech 10x training set into two for training and one for testing. The parameters used in the cross-validation experiments are identical to those used in the experiments described in Section 4. The log-average miss rate (MR), which is calculated as False Positive Per Image (FPPI) in $[10^{-1}, 10^{0}]$, is used to evaluate the detection performance.

The detection performance with varying decorrelated filter size on Caltech 10x validation set is shown in Fig. 2(a). Note that decorrelated filters corresponding to top three PCA eigenvectors of each scale are used in the experiments. The aggregated channel feature is used when the decorrelated filter size is one. It can be observed that average MR increase when the filter size is larger than two. This phenomenon is mainly caused by overfitting as larger filters capture pixel differences at larger distances and hence, are less correlated. Fig. 2(b) shows the detection performance with varying number of filters with the filter size fixed as 2×2. It can be observed that the average MR decreases with more filters which implies more decorrelated filters can extract more discriminative feature for pedestrian detection. The detection performance gain is very small when the filter number increases from three to four, but the time for filtering over aggregated channels increases about 33%. Therefore, in the first stage of the proposed framework, we employ top three decorrelated filters with size 2×2. Our pedestrian detector using these three decorrelated filters achieves MR of 15.56%, which is much better than LDCF (MR of 24.80% [20]) and outperforms existing top-performing FCF methods (e.g. Checkerboards with MR of 18.47%, and RotatedFilter with MR of 19.20% [29, 30]). This demonstrates the effectiveness of using small scale decorrelated filters compared to larger ones.



(a) (b) (c)

Figure 2: Average MR on Caltech 10x validation set with varying (a) decorrelated filter size, (b) number of decorrelated filters, and (c) multi-scale decorrelated filter size.



(a) (b)

Figure 3: a) Learned multi-scale decorrelated filters with size of 2×2 (left) and 3×3 (right) from Caltech10x training set, the top bar indicates the corresponding aggregated channel. b) examples of error in Caltech annotations, upper and lower rows are due to false annotations and poor alignments respectively.

To further increase the detection performance, multi-scale decorrelated filters are explored in the second detection stage of our framework. Based on the analysis in Fig. 2(a), we propose to employ filter size of 2×2 and another larger filter in a multi-scale setting. The average MR of varying multi-scale decorrelated filter sizes in Caltech 10x validation set is shown in Fig. 2(c). It can be observed that the average MR increase with larger filters. This is consistent with the results of using single scale filters shown in Fig. 2(a). As such, we propose to exploit 2×2 and 3×3 decorrelated filters when training the second stage detector. The learned multi-scale decorrelated filters are illustrated in Fig. 3(a). By using multi-scale decorrelated filters in the two-stage framework, the proposed method achieves MR of 14.63% which is the best detection performance on Caltech dataset among all non-deep learning methods in the literature.

3.2 Group Cost-sensitive RealBoost Algorithm

We first briefly describe detection via cost-insensitive boosting and then derive our proposed group cost-sensitive RealBoost algorithm based on a new mixture loss.

Detection via Boosting: Given a set of training samples for pedestrian detection $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ is the feature vector of each training sample, and $y \in \{-1, 1\}$ is the label of the training sample. The detector aims to learn a function $G(\mathbf{x})$ that maps feature space to label space and can be expressed as:

$$G(\mathbf{x}) = sgn[F(\mathbf{x})] \tag{1}$$

$F(\mathbf{x})$ is a predictor, $sgn[.]$ is the sign function that returns 1 if $F(\mathbf{x}) > 0$ and -1 otherwise. In RealBoost, the predictor $F(\mathbf{x})$ is learned from a linear combination of weak learners in a greedy forward stagewise fashion [15]:

$$F(\mathbf{x}) = \sum_{m=1}^{M} f_m(\mathbf{x}) \tag{2}$$

The predictor is updated at each iteration according to:

$$F^{(t)}(\mathbf{x}) = F^{(t-1)}(\mathbf{x}) + f^{(t)}(\mathbf{x}) \tag{3}$$

where $f^{(t)}(\mathbf{x})$ is the learned weak learner in iteration t. The detector $G(\mathbf{x})$ is optimal if it minimizes the risk $E_{\mathbf{X},Y}[Loss(\mathbf{x},y)]$, where $Loss(\mathbf{x},y)$ is a loss function that measures the misclassification. The zero-one loss is often used to evaluate the misclassification:

$$Loss(\mathbf{x},y) = \begin{cases} 0, & \text{if } G(\mathbf{x}) = y \\ 1, & \text{if } G(\mathbf{x}) \neq y \end{cases} \tag{4}$$

Group Cost-sensitive Boosting: The zero-one loss assigns identical cost to all kinds of misclassification and is not well suited for pedestrian detection. For instance, samples with low resolution often leads to misclassification. As such, we propose a group cost-sensitive RealBoost algorithm that explores different costs for misclassification of pedestrian samples with different resolutions. Specifically, the training samples are divided into two groups based on their resolution. The group cost-sensitive loss can be expressed as:

$$Loss(\mathbf{x},y) = \begin{cases} 0, & \text{if } G(\mathbf{x}) = y \\ C_{HR}, & \text{if } G(\mathbf{x}_{HR}) \neq y \\ C_{LR} & \text{if } G(\mathbf{x}_{LR}) \neq y \end{cases} \tag{5}$$

where $\mathbf{x}_{HR}$ and $\mathbf{x}_{LR}$ represents set of high and low resolution samples respectively. In the proposed method, the misclassification of low resolution samples need to be given higher emphasis, i.e. $C_{HR} < C_{LR}$. Since the minimization of risk with respect to zero-one loss is usually difficult, the boosting family chooses to minimize alternative risk, based on convex upper-bounds of zero-one loss. These risk are of the form:

$$R_\phi = E_{\mathbf{X},Y}[\phi(yF(\mathbf{x}))] \tag{6}$$

where $\phi(v)$ is a convex upper bound of loss in Eq. 5. In particular, exponential loss and logistic loss are commonly used in boosting algorithms e.g. Adaboost, RealBoost and LogitBoost [15]. The former two employs exponential loss while the last one exploits logistic loss. The exponential loss is known to be overly sensitive to outliers [8, 26]. As can be seen from Fig. 4(a), the sensitivity stemming from the weight of misclassifed samples increase exponentially, especially when $yF(\mathbf{x}) \to -\infty$. In object detection tasks, errors from annotations are common, as shown in Fig. 3(b) from Caltech pedestrian dataset [9]. Compared to exponential loss, the logistic loss is more moderate and the speed of weight increasing is much slower. To alleviate the sensitivity of exponential loss and explore samples with different resolutions, we propose a novel group cost-sensitive mixture loss which exploits a combination of exponential and logistic loss as:

$$\phi(v) = C(\alpha e^{-Cv} + (1-\alpha)\frac{1}{\ln 2}\ln(1+e^{-Cv})) \tag{7}$$

where α is a trade-off between the exponential and logistic loss. C is the cost for samples with different resolution as defined in Eq. 5. The proposed loss is illustrated in Fig. 4(a). It can be observed that the proposed loss is the convex upper-bound of loss defined in Eq. 5.

Note that the hypotheses that pass the first stage should have similar response in decorrelated channels. In order to distinguish the hypothesis that are wrongly classified in the first stage, we need to expand the receptive field and integrate more local information. In the proposed framework, two methods are employed: using larger filter and feature channel shrinkage. Larger filters allow us to learn decorrelated filters from larger patch and integrate more local information. Feature channel shrinkage aims to shrink the feature channels prior to filtering which allows filters of the same size to cover larger region. Concretely, in the second stage, we learn decorrelated filters with size of 2×2 and 3×3 since these two scale filters show lower average MR as illustrated in Fig. 2(a). We adopt 2×2 max-pooling operation with stride of 2 in vertical and horizontal orientations. The feature channel shrinkage operation leads to 2×2 receptive field expansion from the local regions if we use the same filters as in the first stage while keeping some degrees of invariance with respect to translations and distortions.

Given C_{HR}, C_{LR} and α, the optimal detector can be learnt by minimizing empirical risk as:

$$\begin{aligned} R_\phi(F) = \frac{1}{N}[&\sum_{\mathbf{x}_i\in\mathbf{x}_{LR}} C_{LR}(\alpha e^{-y_iC_{LR}F(\mathbf{x}_i)} + (1-\alpha)\frac{1}{\ln 2}\ln(1+e^{-y_iC_{LR}F(\mathbf{x}_i)})) \\ &+ \sum_{\mathbf{x}_i\in\mathbf{x}_{HR}} C_{HR}(\alpha e^{-y_iC_{HR}F(\mathbf{x}_i)} + (1-\alpha)\frac{1}{\ln 2}\ln(1+e^{-y_iC_{HR}F(\mathbf{x}_i)}))] \end{aligned} \tag{8}$$

The above problem can be optimized using greedy forward stagewise fashion [15]. After we have estimated $F^{(t-1)}(\mathbf{x})$, the weak learner $f^{(t)}(\mathbf{x})$ can be learned by solving the following:

$$\begin{aligned} f^{(t)}(\mathbf{x}) = \arg\max_f -[&\sum_{\mathbf{x}_i\in\mathbf{x}_{LR}} C_{LR}(-\alpha y_iC_{LR}e^{-y_iC_{LR}F^{(t-1)}(\mathbf{x}_i)} - \frac{(1-\alpha)y_iC_{LR}}{\ln 2(1+e^{y_iC_{LR}F^{(t-1)}(\mathbf{x}_i)})}) \\ &+ \sum_{\mathbf{x}_i\in\mathbf{x}_{HR}} C_{HR}(-\alpha y_iC_{HR}e^{-y_iC_{HR}F^{(t-1)}(\mathbf{x}_i)} - \frac{(1-\alpha)y_iC_{HR}}{\ln 2(1+e^{y_iC_{HR}F^{(t-1)}(\mathbf{x}_i)})})]f^{(t)}(\mathbf{x}_i) \\ = \arg\max_f &\sum_{\mathbf{x}_i\in\mathbf{x}_{LR}} w^{(t)}_{i_{LR}} y_i f^{(t)}(\mathbf{x}_i) + \sum_{\mathbf{x}_i\in\mathbf{x}_{HR}} w^{(t)}_{i_{HR}} y_i f^{(t)}(\mathbf{x}_i) \end{aligned} \tag{9}$$

where

$$w_i^{(t)} = \begin{cases} C_{LR}(\alpha C_{LR}e^{-y_iC_{LR}F^{(t-1)}(\mathbf{x}_i)} + \frac{(1-\alpha)C_{LR}}{\ln 2(1+e^{y_iC_{LR}F^{(t-1)}(\mathbf{x}_i)})}), & \text{if } \mathbf{x}\in\mathbf{x}_{LR} \\ C_{HR}(\alpha C_{HR}e^{-y_iC_{HR}F^{(t-1)}(\mathbf{x}_i)} + \frac{(1-\alpha)C_{HR}}{\ln 2(1+e^{y_iC_{HR}F^{(t-1)}(\mathbf{x}_i)})}), & \text{if } \mathbf{x}\in\mathbf{x}_{HR} \end{cases} \tag{10}$$

is the weight of training samples for different groups at iteration t. Note that the weight rule becomes cost-insensitive if $C_{LR} = C_{HR}$. The costs defined in Eq. 5 is decided by the ratio C_{LR}/C_{HR}, C_{HR} can be set to one and the search for optimal cost becomes one-dimensional. The optimal C_{LR} and α can be selected from cross-validation experiments. The detector of each stage can be learned by solving corresponding problem in Eq. 8.

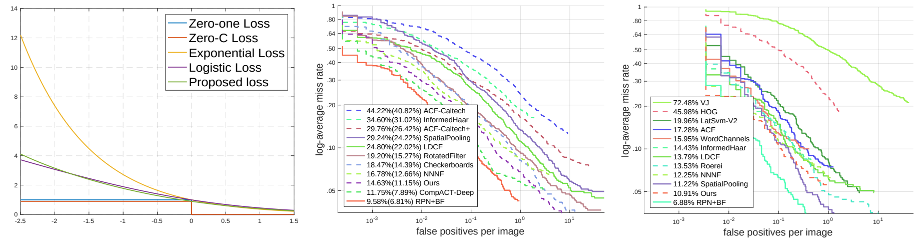
4 Results and Discussion

In this section, we evaluate the detection performance and execution time of the proposed method on two widely used datasets. To ensure a fair comparison for the execution time, we implemented all the methods on the common platform, i.e. 3.5GHz Intel Xeon E5-1650 CPU with single thread execution. We have not relied on GPUs in our experiments.

Datasets: Our experiments are based on two public datasets: Caltech [9] [1] and INRIA [6] [2]. For the Caltech dataset [9], the training data is augmented by extracting one of every

[1] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

[2] http://pascal.inrialpes.fr/data/human/



(a) (b) (c)

Figure 4: a) Loss function $\phi(v)$. Detection performance comparison with state-of-the-art methods on b) Caltech and c) INRIA datasets

Table 1: Average execution time per frame (seconds) and MR of FCF methods (Caltech)

	Aggregated Channel	Filtering	Classification	Total Time (s)	MR(MR$_{\_N}$) (%)
ACF	0.045	-	0.061	0.106	29.76(26.42)
LDCF	0.046	0.220	0.035	0.301	24.80(22.02)
RotatedFilters	0.384	9.309	6.931	16.625	19.20(15.27)
Checkerboards	0.496	22.117	20.709	43.319	18.47(14.39)
Ours	0.193	0.135	0.163	0.491	14.63(11.15)

3 frames (similar to [29]). 42782 images are used to train our model. The Caltech test set consists of 4024 images which includes 1014 positive images. The evaluation metric is MR on False Positive Per Image (FPPI) in $[10^{-2}, 10^{-0}]$ under *reasonable* setup (pedestrians that are at least 50 pixels tall and at least 65% visible [9]). In addition, we also tested our model on the new annotations of Caltech test set provided by [30], which has corrected some errors in the original annotations. We denote the results of the original and new annotations as MR and MR$_{\_N}$ respectively. For the INRIA dataset [6], there are 614 positive images and 1218 negative images in the training set. The trained model is evaluated on 288 testing images using MR on FPPI ranges of $[10^{-2}, 10^{-0}]$.

Model Parameters: We use a model with size 64×128 when training the detector. For each stage, three rounds of hard negative mining (32, 512, 1024, 5120 trees respectively) are used and 100000 negatives are added to the training set in each round. During decision tree learning, we randomly selected 1/16 features and the depth of the decision tree is limited to 5. The strides of both sliding window and aggregated channel shrinkage factor are 4, and each image is upsampled by one octave. The height threshold of samples that splits training data into low and high resolution group is 55 pixels. The optimal value of costs for different groups and coefficient α are selected from $C_{HR} = 1$, $C_{LR} \in [1.0 : 0.05 : 1.3]$ and $\alpha \in [0.5 : 0.05 : 1]$ in the cross-validation experiments as described in Section 3.1.

4.1 Comparison with State-of-the-art Methods on Caltech dataset

The detection performance of proposed method and state-of-the-art methods are shown in Fig. 4(b). It can be observed that the proposed method outperforms all non-CNN methods in detection performance. Compared with FCF methods, the proposed method achieves a much lower MR i.e 14.63% while the MR of LDCF [20], RotatedFilters [30] and Checkerboards [29] are 24.80%, 19.20% and 18.47% respectively. The MR of proposed method is about 2% lower than NNNF [5]. The proposed method still achieves lowest MR among non-CNN methods when evaluating on the new annotations provided by [30]. Among the

Table 2: Average execution time per image (seconds) and MR of FCF methods (INRIA)

	Aggregated Channel	Filtering	Classification	Total Time (s)	MR (%)
ACF	0.024	-	0.012	0.036	17.28
LDCF	0.051	0.243	0.105	0.399	13.79
Ours	0.079	0.198	0.022	0.299	10.91

CNN methods, RPN+BF that combines a region proposal network and boosted forest to achieve the lowest MR. However, when using RPN as stand-alone pedestrian detector [27], the MR is 14.90% which is a little higher than proposed method. Though RPN+BF [27] and compACT-Deep [2] have a better performance than our proposed method, their performance are achieved using pre-trained deep models (e.g., VGG [23]) on ImageNet [22] that requires a large amount of convolution operations. The RPN+BF runs at about 0.5 second per frame (similar to our runtime on a single thread CPU) on the Tesla K40 GPU (which is reported to have 10x computation power of parallel processing on a 16-core 3.1GHz CPU [3]).

The detection time of proposed method and state-of-the-art FCF methods are illustrated in Table 1. Note that we only perform comparisons with methods that have released their models as we can run them on a common platform. It can be observed that the ACF and LDCF run faster than other methods. This is partly due to their adoption of smaller model size (64 × 32) compared to others that exploit larger model size (128 × 64). In addition, ACF does not incorporate the filtering process which is the most time-consuming step in FCF framework. The RotatedFilters and Checkerboards achieve much better detection performance compared to ACF and LDCF, but the detection time is very high due to the high resolution channels and large amount of filters employed. As shown in Table 1, the proposed method achieves a much lower MR and still runs about 33 and 88 times faster than RotatedFilters and Checkerboards respectively. The proposed method also runs significantly faster than the current top-performing non-CNN method in the literature, NNNF [5] (i.e. 0.877 seconds per frame on a similar platform) while achieving a lower MR. These results clearly demonstrate that our proposed method achieves the best trade-off between detection performance and speed among all the state-of-the-art pedestrian detection methods.

4.2 Comparison with State-of-the-art Methods on INRIA dataset

The INRIA dataset has much lesser training images compared to Caltech. Therefore, we adjust some settings to adapt to the INRIA dataset. We train a single-stage detector via three rounds of hard negative mining (32, 128, 512, 2688 trees respectively) and 11000 negatives are added to each round. The depth of decision tree is constrained to 2 with other settings identical to the description in Section 4.1. The multi-scale decorrelated filters are learned with 2 × 2 and 4 × 4 filter sizes. Note that we only compare the detection time with FCF methods that have released their models for INRIA dataset.

The detection performance of the proposed method and the state-of-the-art methods are shown in Fig. 4(c). We can observe that the proposed method still achieves the best detection performance among the non-CNN methods. The MR of LDCF [20] is about 2.88% higher than proposed method which demonstrates the effectiveness of the proposed multi-scale decorrelated filter learning strategy. The NNNF [5] and SpatialPooling [21] employ more complex Haar-like features but their detection performance are still lower than ours. The detection time of ACF [10], LDCF [20] and proposed method are shown in Table 2. It can be observed that ACF runs at 0.036 seconds per image but with a very high MR which

[3] https://www.nvidia.com/content/tesla/pdf/nvidia-tesla-k40-2014mar-lr.pdf

prohibits its application in real-time systems. LDCF achieves a much lower MR than ACF but its detection time is about 1.3 times higher than the proposed method and its MR is also about 2.88% higher than the proposed method. These results further demonstrate the effectiveness and efficiency of the proposed method.

5 Conclusion

In this work, we presented a novel method for improving the performance and runtime of pedestrian detection. To integrate more local information, we proposed to learn multi-scale decorrelated filters for each channel and exploit max-pooling operation prior to the filtering step in a multi-stage detection framework. We proposed a new group cost-sensitive Real-Boost algorithm based on a new mixture loss which gives higher emphasis to harder samples and alleviate the sensitivity to outliers. By combining multi-scale decorrelated filters and cost-sensitive learning in a multi-stage detection framework, the proposed method achieves best detection performance among all non-CNN methods. In addition, the proposed method can run an order of magnitude faster than top-performing FCF methods.

References

[1] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *European Conference on Computer Vision*, pages 613–627. Springer, 2014.

[2] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3361–3369, 2015.

[3] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision*, pages 354–370. Springer, 2016.

[4] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

[5] Jiale Cao, Yanwei Pang, and Xuelong Li. Pedestrian detection inspired by appearance constancy and shape symmetry. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[7] Floris De Smedt, Dries Hulens, and Toon Goedemé. On-board real-time tracking of pedestrians on a uav. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2015.

[8] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.

[9] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012.

[10] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (8):1532–1545, 2014.

[11] Xianzhi Du, Mostafa El-Khamy, Jungwon Lee, and Larry S Davis. Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. *arXiv preprint arXiv:1610.03466*, 2016.

[12] Wei Fan, Salvatore J Stolfo, Junxin Zhang, and Philip K Chan. Adacost: misclassification cost-sensitive boosting. In *Icml*, pages 97–105, 1999.

[13] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 2241–2248. IEEE, 2010.

[14] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[15] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[16] Ross Brook Girshick. *From rigid templates to grammars: Object detection with structured models*. Citeseer, 2012.

[17] Shaoshan Liu, Jie Tang, Zhe Zhang, and Jean-Luc Gaudiot. Caad: Computer architecture for autonomous driving. *arXiv preprint arXiv:1702.01894*, 2017.

[18] Aurlie C Lozano and Naoki Abe. Cost-sensitive boosting with p-norm loss functionsand its applications. *MI lecture note series*, 12:65–74, 2008.

[19] Hamed Masnadi-Shirazi and Nuno Vasconcelos. Asymmetric boosting. In *Proceedings of the 24th international conference on Machine learning*, pages 609–619. ACM, 2007.

[20] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2014.

[21] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *European Conference on Computer Vision*, pages 546–561. Springer, 2014.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[24] Yanmin Sun, Andrew KC Wong, and Yang Wang. Parameter inference of cost-sensitive boosting algorithms. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 21–30. Springer, 2005.

[25] Kai Ming Ting. A comparative study of cost-sensitive boosting algorithms. In *In Proceedings of the 17th International Conference on Machine Learning*. Citeseer, 2000.

[26] Yichao Wu and Yufeng Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

[27] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European Conference on Computer Vision*, pages 443–457. Springer, 2016.

[28] Shanshan Zhang, Christian Bauckhage, and Armin B Cremers. Informed haar-like features improve pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 947–954, 2014.

[29] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Filtered channel features for pedestrian detection. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1751–1760. IEEE, 2015.

[30] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2016.

[31] Chao Zhu and Yuxin Peng. Group cost-sensitive boosting for multi-resolution pedestrian detection. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3676–3682. AAAI Press, 2016.