Group Cost-Sensitive Boosting for Multi-Resolution Pedestrian Detection

Chao Zhu and **Yuxin Peng***
Institute of Computer Science and Technology, Peking University
Beijing 100871, China
{zhuchao, pengyuxin}@pku.edu.cn

Abstract

As an important yet challenging problem in computer vision, pedestrian detection has achieved impressive progress in recent years. However, the significant performance decline with decreasing resolution is a major bottleneck of current state-of-the-art methods. For the popular boosting-based detectors, one of the main reasons is that low resolution samples, which are usually more difficult to detect than high resolution ones, are treated by equal costs in the boosting process, leading to the consequence that they are more easily being rejected in early stages and can hardly be recovered in late stages as false negatives. To address this problem, we propose in this paper a new multi-resolution detection approach based on a novel group cost-sensitive boosting algorithm, which extends the popular AdaBoost by exploring different costs for different resolution groups in the boosting process, and places more emphases on low resolution group in order to better handle detection of hard samples. The proposed approach is evaluated on the challenging Caltech pedestrian benchmark, and outperforms other state-of-the-art on different resolution-specific test sets.

Introduction

Pedestrian detection is a challenging problem in computer vision, and has attracted plenty of attention for decades for its importance in practical applications such as video surveillance, driving assistance, etc. Thanks to various effective detection techniques, pedestrian detection has achieved great progress in recent years. However, detecting multi-resolution pedestrians (as in Fig. 1) is still hard, and a major bottleneck of current state-of-the-art methods is their significant performance decline with decreasing resolution. For example, the mean miss rate of the best detector (Paisitkriangkrai, Shen, and van den Hengel 2014a) achieves 8% for pedestrians taller than 80 pixels in Caltech pedestrian benchmark (Dollár et al. 2012), while significantly increases to 63% for pedestrians of 30-80 pixels high. Nevertheless, robust detection of low resolution pedestrians is also very important in certain occasions like driving assistant systems to provide enough time for reaction.

*Corresponding author.

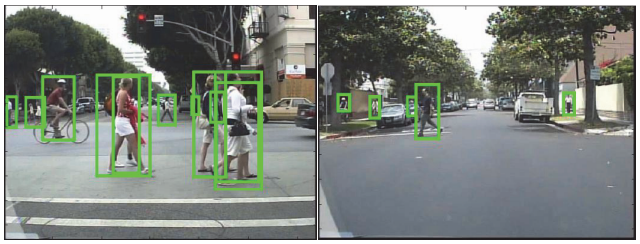


Figure 1: Example test images in Caltech pedestrian benchmark and ground truth annotations. Note that people are in a wide range of resolutions.

The boosting-based approaches are popular for training pedestrian detectors due to their both high effectiveness and efficiency. The basic idea is to linearly combine a series of weak classifiers to produce a strong classifier. Each training sample is assigned by a weight which is updated iteratively in order to emphasize on those wrongly classified samples. In pedestrian detection, the fact is that rare positive targets need to be detected from enormous negative windows. Thus more weights should be placed on positive samples during training to achieve a higher detection rate. To do so, several cost-sensitive boosting algorithms have been proposed (Viola and Jones 2001) (Sun et al. 2007) (Masnadi-Shirazi and Vasconcelos 2011) to penalize more on false negatives than on false positives. However, these methods are not optimal for multi-resolution detection, since they still treat the whole positive samples equally and ignore the intra-class variations. Features extracted from low resolution pedestrians are usually less discriminative than that from high resolution ones, so that they are more easily being rejected in early stages and can hardly be recovered in late stages as false negatives. Thus the trained detector can be biased towards high resolution pedestrians, leading to poorer performance on low resolution pedestrians. To address this problem, we propose a new multi-resolution detection approach based on a novel group cost-sensitive boosting algorithm, which can explore different costs for different resolution groups in the boosting process, and emphasize more on low resolution pedestrians in order to better handle detection of hard samples. The experimental results show its superior performance to other state-of-the-art on different resolution-specific test sets.

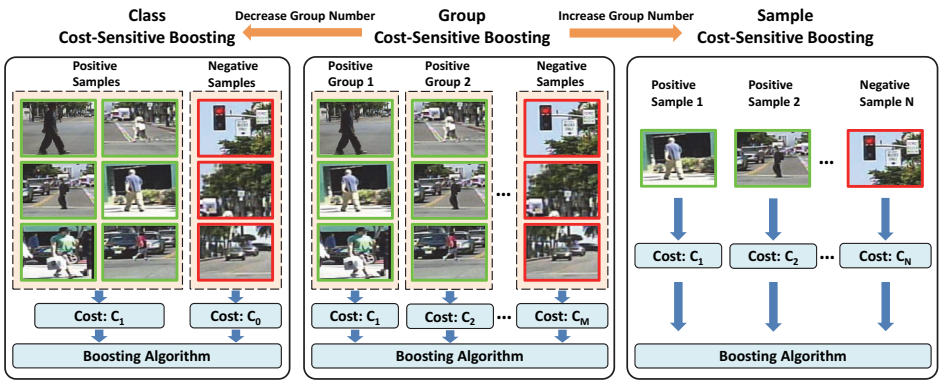


Figure 2: Comparison of different cost-sensitive boosting strategies.

Related Work

The research on pedestrian detection has lasted for decades, and remarkable progress has been made thanks to various detection approaches (Dollár et al. 2012). However, limited works focus on multi-resolution detection problem. (Park, Ramanan, and Fowlkes 2010) propose a multi-resolution pedestrian model where a rigid HOG template is used for low resolution samples, and a part-based model is used for high resolution samples. (Benenson et al. 2012) propose to reduce the number of scales for feature computation by a factor K without image resizing, and mainly focus on speedup more than quality. (Costea and Nedevschi 2014) propose scale independent features and use one single classifier for all scales. Recently, (Yan et al. 2013) take pedestrian detection in different resolutions as different but related problems, and propose a multi-task model to jointly consider their commonness and differences. Nevertheless, this method relies on deformable part-based model (DPM) and thus has a relatively high computational complexity.

To achieve efficient detection, boosting is popular in detector training. Several cost-sensitive boosting algorithms have been proposed in the literature to address data imbalance problem, and can be divided into two kinds: (1) class cost-sensitive (CCS) boosting such as AsymmetricAdaBoost (Viola and Jones 2001), AdaCost (Fan et al. 1999), CSB0-CSB2 (Ting 2000), AdaC1-AdaC3 (Sun et al. 2007) and Cost-sensitive Boosting (Masnadi-Shirazi and Vasconcelos 2011); (2) sample cost-sensitive (SCS) boosting (Abe, Zadrozny, and Langford 2004). They share the same main idea of putting more costs on the positive samples by modifying the weight update rules, so that false negatives are penalized more than false positives. Although they distinguish positive samples from negative ones in the boosting process, they still ignore the variations in positive set. Different from these methods, our proposed approach is based on a new group cost-sensitive (GCS) boosting which explores different costs for different resolution groups in positive set during boosting in order to better handle multi-resolution detection. Note that our approach is related to both CCS boosting and SCS boosting, and can be considered as a generalized form of them. As shown in Fig. 2, in the special case of decreasing group number to treat positive samples as a whole, GCS boosting is simplified to CCS boosting; in the special case of increasing group number to treat each sample as an individual group, GCS boosting scales up to SCS boosting.

Multi-Resolution Detection via Locally Decorrelated Channel Features and Group Cost-Sensitive AdaBoost

In this section, we present a new multi-resolution detection approach based on a novel group cost-sensitive boosting algorithm, which extends the popular AdaBoost by exploring different costs for different resolution groups in the boosting process, so that more emphases can be placed on low resolution group in order to better handle detection of hard samples. Particularly, we apply the idea to the popular LDCF detector (Nam, Dollár, and Han 2014) and propose a multi-resolution LDCF detector.

Locally Decorrelated Channel Features (LDCF) Detector

We apply the Locally Decorrelated Channel Features (LDCF) detector as a baseline, because of its good detection quality. Given an input image, LDCF detector first computes several feature channels, where each channel is a per-pixel feature map such that output pixels are computed from corresponding patches of input pixels. An efficient feature transform is then applied to remove correlations in local image neighborhoods, since effective but expensive oblique splits in decision trees can be replaced by orthogonal splits over locally decorrelated data. Totally 10 image channels (1 normalized gradient magnitude channel, 6 histogram of oriented gradients channels and 3 LUV color channels) are used and 4 decorrelating filters per channel are applied, resulting in a set of 40 locally decorrelated channel features. For detector training, AdaBoost is used to train and combine decision trees over these features to obtain a strong classifier. For more details of the LDCF detector, please refer to (Nam, Dollár, and Han 2014).

Detection via AdaBoost For the convenience of the following presentation, we first give a formal problem defini-

tion of detection via AdaBoost.

Given a set of samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ for pedestrian detection, where $\mathbf{x} = (x_1, \ldots, x_N)^T \in \mathbf{X} = \mathbb{R}^N$ is the feature vector of each sample, and $y \in Y = \{-1, 1\}$ is the class label, a detector (or binary classifier) is a function h which maps each feature vector $\mathbf{x}$ to its corresponding class label y, and is usually implemented as:

$$h(\mathbf{x}) = \text{sgn}[f(\mathbf{x})] \tag{1}$$

where $f(\mathbf{x})$ is a predictor, $\text{sgn}[\cdot]$ is the sign function that equals 1 if $f(\mathbf{x}) \geq 0$ and equals -1 otherwise. The detector will be optimal if it minimizes the risk $E_{\mathbf{X},Y}[Loss(\mathbf{x}, y)]$, where $Loss(\mathbf{x}, y)$ is a loss function to measure the classification error. The AdaBoost algorithm applied in LDCF uses the following loss function:

$$Loss(\mathbf{x}, y) = \begin{cases} 0, & \text{if } h(\mathbf{x}) = y \\ 1, & \text{if } h(\mathbf{x}) \neq y \end{cases} \tag{2}$$

and learns a predictor $f(\mathbf{x})$ by linear combination of weak learners:

$$f(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m g_m(\mathbf{x}) \tag{3}$$

where α_m is a set of weights and $g_m(\mathbf{x}) = \text{sgn}[\phi_m(\mathbf{x}) - t_m]$ is a set of decision stumps with $\phi_m(\mathbf{x})$ a feature response and t_m a threshold.

Specifically, the predictor is learned by gradient descent with respect to the exponential loss:

$$E_{\mathbf{X},Y}[\exp(-yf(\mathbf{x}))] \tag{4}$$

that weak learners are selected iteratively to minimize the classification error at each iteration:

$$g_m(\mathbf{x}) = \arg\min_g (err_{(m)}) \tag{5}$$

where

$$err_{(m)} = \sum_{i=1}^{n} \omega_i^{(m)} [1 - I(y_i = g_m(\mathbf{x}_i))] \tag{6}$$

is the classification error and $I(\cdot)$ is an indicator function:

$$I(y = a) = \begin{cases} 1, & \text{if } y = a \\ 0, & \text{if } y \neq a \end{cases} \tag{7}$$

The weight of weak learners is calculated by:

$$\alpha_m = \frac{1}{2} \log \left(\frac{1 - err_{(m)}}{err_{(m)}} \right) \tag{8}$$

and the weight $\omega_i^{(m)}$ assigned to the sample $\mathbf{x}_i$ is updated accordingly to increase the importance of wrongly classified samples as well as to decrease the importance of correctly classified samples at the next iteration:

$$\omega_i^{(m+1)} = \omega_i^{(m)} \exp(-y_i \alpha_m g_m(\mathbf{x}_i)) \tag{9}$$

Group Cost-Sensitive AdaBoost

The loss function defined in (2) is cost-insensitive, since the costs of false positives ($y = -1, h(\mathbf{x}) = 1$) and false negatives ($y = 1, h(\mathbf{x}) = -1$) are the same. In order to better handle the detection of multi-resolution, we propose here a novel group cost-sensitive AdaBoost algorithm by exploring different importance of the samples from different resolutions so that more emphases can be placed on hard low resolution samples.

Group Cost-Sensitive Loss To introduce different importance for the samples of different resolution, positive samples are further divided into groups with different resolution. Here we consider the case of two resolution groups: low resolution samples (30-80 pixels high, denoted as $\mathbf{x}_L$), and high resolution samples (taller than 80 pixels, denoted as $\mathbf{x}_H$), as defined in the Caltech pedestrian benchmark (Dollár et al. 2012). We propose a group cost-sensitive loss function as follows:

$$Loss(\mathbf{x}, y) = \begin{cases} 0, & \text{if } h(\mathbf{x}) = y \\ C_{fp}, & \text{if } y = -1, h(\mathbf{x}) = 1 \\ C_{fnl}, & \text{if } y = 1, h(\mathbf{x}_L) = -1 \\ C_{fnh}, & \text{if } y = 1, h(\mathbf{x}_H) = -1 \end{cases} \tag{10}$$

with $C_* > 0$. The four scenarios considered in this loss function are respectively correct detections ($h(\mathbf{x}) = y$), false positives ($y = -1, h(\mathbf{x}) = 1$), false negatives (miss detections) of low resolution samples ($y = 1, h(\mathbf{x}_L) = -1$) and false negatives (miss detections) of high resolution samples ($y = 1, h(\mathbf{x}_H) = -1$). Note that when $C_{fnl} = C_{fnh}$, this group cost-sensitive loss will be equivalent to the canonical class cost-sensitive loss.

The costs C_{fnl}, C_{fnh} and C_{fp} can be decided according to different problems. For pedestrian detection, the intuition tells us that C_{fnl} and C_{fnh} should be greater than C_{fp}, since miss detections are harder to be recovered than false positives, and C_{fnl} should be greater than C_{fnh}, since low resolution samples are harder to be detected than high resolution ones. The optimal values of these costs will be chosen experimentally by cross-validation. When C_{fp}, C_{fnl} and C_{fnh} are specified, the group cost-sensitive exponential loss is:

$$\begin{aligned} E_{\mathbf{X},Y} [&I'(y = 1, \mathbf{x} \in \mathbf{x}_L) \exp(-yC_{fnl} f(\mathbf{x})) \\ &+ I'(y = 1, \mathbf{x} \in \mathbf{x}_H) \exp(-yC_{fnh} f(\mathbf{x})) \\ &+ I'(y = -1, \mathbf{x} \in \mathbf{x}) \exp(-yC_{fp} f(\mathbf{x}))] \end{aligned} \tag{11}$$

where $I'(\cdot)$ is an extended indicator function:

$$I'(y = a, \mathbf{x} \in b) = \begin{cases} 1, & \text{if } y = a \text{ and } \mathbf{x} \in b \\ 0, & \text{others} \end{cases} \tag{12}$$

Group Cost-Sensitive Adaboost The proposed group cost-sensitive AdaBoost algorithm is derived by gradient descent on the empirical estimate of the expected loss in (11). Given a set of training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the definition of the predictor $f(\mathbf{x})$ as in (3) and three groups defined as:

$$\begin{aligned} \mathcal{G}_{+L} &= \{i | y_i = 1, \mathbf{x}_i \in \mathbf{x}_L\} \\ \mathcal{G}_{+H} &= \{i | y_i = 1, \mathbf{x}_i \in \mathbf{x}_H\} \\ \mathcal{G}_{-} &= \{i | y_i = -1\} \end{aligned} \tag{13}$$

the weak learner selected at iteration m consists of an optimal step α_m along the direction g_m of the largest descent of the expected loss in (11) as:

$$\begin{aligned}(\alpha_m, g_m) = \arg\min_{\alpha,g} &\sum_{i\in\mathcal{G}_{+L}} \omega_i^{(m)} \exp(-C_{fnl}\alpha g(\mathbf{x}_i)) \\ &+ \sum_{i\in\mathcal{G}_{+H}} \omega_i^{(m)} \exp(-C_{fnh}\alpha g(\mathbf{x}_i)) \\ &+ \sum_{i\in\mathcal{G}_{-}} \omega_i^{(m)} \exp(C_{fp}\alpha g(\mathbf{x}_i))\end{aligned} \tag{14}$$

The optimal step α along the direction g is the solution of:

$$\begin{aligned}&2C_{fnl}\cdot err_{+L}\cdot\cosh(C_{fnl}\alpha) - C_{fnl}\cdot\Omega_{+L}\cdot e^{-C_{fnl}\alpha} \\ &+2C_{fnh}\cdot err_{+H}\cdot\cosh(C_{fnh}\alpha) - C_{fnh}\cdot\Omega_{+H}\cdot e^{-C_{fnh}\alpha} \\ &+2C_{fp}\cdot err_{-}\cdot\cosh(C_{fp}\alpha) - C_{fp}\cdot\Omega_{-}\cdot e^{-C_{fp}\alpha} = 0\end{aligned} \tag{15}$$

with

$$\Omega_{+L} = \sum_{i\in\mathcal{G}_{+L}} \omega_i^{(m)}, \Omega_{+H} = \sum_{i\in\mathcal{G}_{+H}} \omega_i^{(m)}, \Omega_{-} = \sum_{i\in\mathcal{G}_{-}} \omega_i^{(m)} \tag{16}$$

$$err_{+L} = \sum_{i\in\mathcal{G}_{+L}} \omega_i^{(m)}[1 - I(y_i = g(\mathbf{x}_i))] \tag{17}$$

$$err_{+H} = \sum_{i\in\mathcal{G}_{+H}} \omega_i^{(m)}[1 - I(y_i = g(\mathbf{x}_i))] \tag{18}$$

$$err_{-} = \sum_{i\in\mathcal{G}_{-}} \omega_i^{(m)}[1 - I(y_i = g(\mathbf{x}_i))] \tag{19}$$

Given the step α and the direction g, the total loss of the weak learner (α, g) is calculated as:

$$\begin{aligned}err_T = &(e^{C_{fnl}\alpha(g)} - e^{-C_{fnl}\alpha(g)})err_{+L} + e^{-C_{fnl}\alpha(g)}\Omega_{+L} \\ &+ (e^{C_{fnh}\alpha(g)} - e^{-C_{fnh}\alpha(g)})err_{+H} + e^{-C_{fnh}\alpha(g)}\Omega_{+H} \\ &+ (e^{C_{fp}\alpha(g)} - e^{-C_{fp}\alpha(g)})err_{-} + e^{-C_{fp}\alpha(g)}\Omega_{-}\end{aligned} \tag{20}$$

and the direction of the largest descent is selected to have the minimum loss:

$$g_m = \arg\min_{g}(err_T) \tag{21}$$

The weight $\omega_i^{(m)}$ of each sample $\mathbf{x}_i$ at the next iteration is updated as follows:

$$\omega_i^{(m+1)} = \begin{cases} \omega_i^{(m)} e^{-C_{fnl}\alpha_m g_m(\mathbf{x}_i)}, & \text{if } i\in\mathcal{G}_{+L} \\ \omega_i^{(m)} e^{-C_{fnh}\alpha_m g_m(\mathbf{x}_i)}, & \text{if } i\in\mathcal{G}_{+H} \\ \omega_i^{(m)} e^{C_{fp}\alpha_m g_m(\mathbf{x}_i)}, & \text{if } i\in\mathcal{G}_{-} \end{cases} \tag{22}$$

The possible descent directions are defined by a set of weak learners $\{g_k(\mathbf{x})\}_{k=1}^K$. The optimal step α along each direction is obtained by solving (15), and can be done efficiently with standard scalar search procedures. Given step α and direction g, the loss associated with the weak learner is calculated as in (20), and the best weak learner with the minimum loss is selected as in (21). A summary of the proposed group cost-sensitive AdaBoost algorithm is presented in Algorithm 1.

Algorithm 1 Group Cost-Sensitive AdaBoost

Input: Training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i$ is the feature vector of the sample and $y_i \in \{1, -1\}$ is the class label, costs $\{C_{fnl}, C_{fnh}, C_{fp}\}$ for different groups, set of weak learners $\{g_k(\mathbf{x})\}_{k=1}^K$, and the number M of weak learners in the final classifier.

Output: Strong classifier $h(\mathbf{x})$ for GCS-LDCF detector.

Initialization: Set uniformly distributed weights for each group:
$\omega_i^{(0)} = \frac{1}{2|\mathcal{G}_{+L}|}, \forall i \in \mathcal{G}_{+L}$; $\omega_i^{(0)} = \frac{1}{2|\mathcal{G}_{+H}|}, \forall i \in \mathcal{G}_{+H}$; $\omega_i^{(0)} = \frac{1}{2|\mathcal{G}_{-}|}, \forall i \in \mathcal{G}_{-}$.
for $m = \{1, \ldots, M\}$ **do**
  for $k = \{1, \ldots, K\}$ **do**
    Compute parameter values as in (16)-(19) with $g(\mathbf{x}) = g_k(\mathbf{x})$;
    Obtain the value of α_k by solving (15);
    Calculate the loss of the weak learner $(\alpha_k, g_k(\mathbf{x}))$ as in (20).
  end for
  Select the best weak learner $(\alpha_m, g_m(\mathbf{x}))$ with the minimum loss as in (21);
  Update the weights ω_i according to (22).
end for
return $h(\mathbf{x}) = \text{sgn}\left[\sum_{m=1}^M \alpha_m g_m(\mathbf{x})\right]$.

Multi-Resolution LDCF Detector

Finally, by incorporating the proposed group cost-sensitive AdaBoost algorithm into the LDCF detector, we obtain a new group cost-sensitive LDCF detector (denoted as "GCS-LDCF" in the experiments) which has a better capability of handling multi-resolution detection. For pedestrian detection, the learned detector is applied on each test image using a multi-scale sliding window strategy, and non-maximal suppression is used to obtain the final detection results.

Experimental Evaluation

In order to evaluate the proposed approach, the experiments are conducted on the Caltech pedestrian detection benchmark (Dollár et al. 2012), which is by far the largest, most realistic and challenging pedestrian dataset. It consists of approximately 10 hours of 640×480 30Hz video taken from a vehicle driving through regular traffic in an urban environment. The data were captured over 11 sessions, and are roughly divided in half for training and testing. It contains a vast number of pedestrians – about 250,000 frames in 137 approximately minute long segments with a total of 350,000 bounding boxes and 2300 unique pedestrians were annotated. This benchmark is challenging mainly because it contains many low resolution pedestrians and has realistic occlusion frequency. Also the image quality is somewhat lacking, including blur as well as visible JPEG artifacts (blocks, ringing, quantization).

Table 1: Log-average miss rate (%) of popular multi-resolution detection methods on different subsets of Caltech.

	MultiResC [ECCV 2010]	**Roerei** [CVPR 2013]	**MT-DPM** [CVPR 2013]	**WordChannels** [CVPR 2014]	**GCS-LDCF** [Proposed]
Reasonable	48.45	48.35	40.54	42.30	**20.20**
Large Scale	17.84	16.07	16.26	21.97	**3.62**
Near Scale	21.19	21.79	19.23	26.54	**5.84**
Medium Scale	73.16	74.16	66.63	73.16	**59.53**
Mean	40.16	40.09	35.67	40.99	**22.30**

Experimental Setup

We follow a common training-testing protocol as in the literature: the pedestrian detector is trained on the training set (set00-set05), and the detection results are reported on the test set (set06-set10). To train the detector, we choose the image regions labeled as "persons" that are non-occluded with different resolutions as positive samples, and negative samples are chosen at random locations and sizes from the training images without pedestrians.

The training parameters in the proposed approach are set as follows: The optimal value of the costs for different groups are selected from $C_{fp} = 1$, $C_{fnh} \in [1 : 0.1 : 10]$ and $C_{fnl} \in [C_{fnh} : 0.1 : C_{fnh} + 10]$ by cross-validation. 4096 weak classifiers are trained and combined to a strong classifier, and the nodes of the decision trees are constructed using a pool of random candidate regions from image samples. The multi-scale models are used to increase scale invariance. Three bootstrapping stages are applied with 25,000 additional hard negatives each time.

For evaluation of the results, we use the bounding boxes labels and the evaluation software (version 3.2.1) provided by Dollár *et al.* on his website[1]. The per-image evaluation methodology is adopted, i.e. all the detection results are compared using miss rate vs. False-Positive-Per-Image (FPPI) curves. The *log-average miss rate* is also used to summarize the detection performance, and is computed by averaging the miss rate at 9 FPPI points[2] that are evenly spaced in the log-space in the range from 10^{-2} to 10^{0}. There exist various experimental settings on Caltech to evaluate detectors in different conditions. In order to validate the effectiveness of the proposed approach for multi-resolution detection, our experiments are conducted on the most popular "Reasonable" subset (pedestrians of ≥ 50 pixels high and less than 35% occluded) and three resolution-specific subsets: "Large Scale" (pedestrians of ≥ 100 pixels high and fully visible); "Near Scale" (pedestrians of ≥ 80 pixels high and fully visible) and "Medium Scale" (pedestrians of 30-80 pixels high and fully visible).

Comparison with Other Popular Multi-Resolution Detection Methods

We first compare the proposed approach with other popular multi-resolution detection methods in the literature, including MultiResC (Park, Ramanan, and Fowlkes 2010), Roerei (Benenson et al. 2013), MT-DPM (Yan et al. 2013) and WordChannels (Costea and Nedevschi 2014). Table 1 reports the *log-average miss rate* of these methods on the "Reasonable" and three resolution-specific subsets of the Caltech benchmark. It can be observed that the proposed approach significantly outperforms the other multi-resolution methods on all the test sets. By averaging the performances on four test sets, the proposed approach outperforms the other methods by at least 13.4%, validating the effectiveness of the proposed approach for multi-resolution detection.

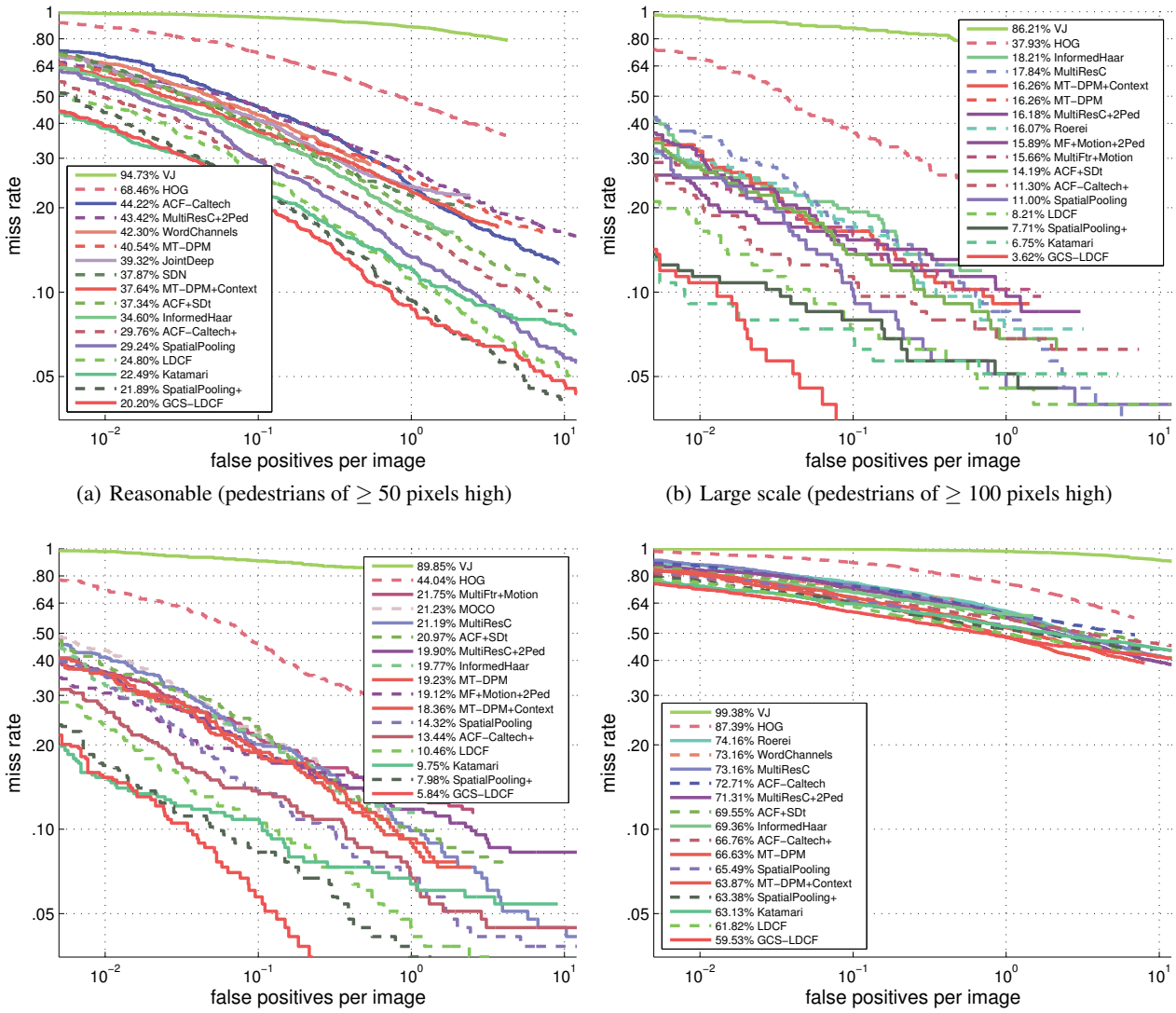
Comparison with Other State-of-the-art Pedestrian Detection Methods

We also compare the proposed approach with many other state-of-the-art pedestrian detection methods in the literature, including VJ (Viola, Jones, and Snow 2005), HOG (Dalal and Triggs 2005), AFS (Levi, Silberstein, and Bar-Hillel 2013), ChnFtrs (Dollár et al. 2009), HOG-LBP (Wang, Han, and Yan 2009), ConvNet (Sermanet et al. 2013), CrossTalk (Dollár, Appel, and Kienzle 2012), FeatSynth (Bar-Hillel et al. 2010), FPDW (Dollár, Belongie, and Perona 2010), HikSVM (Maji, Berg, and Malik 2008), LatSVM (Felzenszwalb et al. 2010), MultiFtr (Wojek and Schiele 2008), MOCO (Chen et al. 2013), pAUCBoost (Paisitkriangkrai, Shen, and van den Hengel 2013), Pls (Schwartz et al. 2009), PoseInv (Lin and Davis 2008), Shapelet (Sabzmeydani and Mori 2007), RandForest (Marín et al. 2013), MultiSDP (Zeng, Ouyang, and Wang 2013), ACF (Dollár et al. 2014), SDN (Luo et al. 2014), DBN-Mut (Ouyang, Zeng, and Wang 2013), Franken (Mathias et al. 2013), InformedHaar (Zhang, Bauckhage, and Cremers 2014), LDCF (Nam, Dollár, and Han 2014), ACF-Caltech+ (Nam, Dollár, and Han 2014), SpatialPooling (Paisitkriangkrai, Shen, and van den Hengel 2014b), SpatialPooling+ (Paisitkriangkrai, Shen, and van den Hengel 2014a) and Katamari (Benenson et al. 2014). These methods adopt various types of features and different modeling strategies. We obtain the results of these methods directly from the same website as the evaluation software.

Fig. 3 presents the ROC curves (miss rate vs. FPPI) and the corresponding *log-average miss rate* (reported in the legend of the figure) of different methods on four test sets of the Caltech benchmark. Note that only the results of top 15 methods plus the classic VJ and HOG are displayed in the figure due to the space limitation. We can clearly observe that: (1) The proposed approach performs significantly bet-

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

[2] The mean miss rate at 0.0100, 0.0178, 0.0316, 0.0562, 0.1000, 0.1778, 0.3162, 0.5623 and 1.0000 FPPI.



(a) Reasonable (pedestrians of $\geq$ 50 pixels high)

(b) Large scale (pedestrians of $\geq$ 100 pixels high)

(c) Near scale (pedestrians of $\geq$ 80 pixels high)

(d) Medium scale (pedestrians of 30-80 pixels high)

Figure 3: Comparison with state-of-the-art methods on the Caltech benchmark.

ter than the baseline detector (LDCF) on all of the four test sets (4.60% better on "Reasonable", 4.59% better on "Large Scale", 4.62% better on "Near Scale" and 2.29% better on "Medium Scale" respectively), demonstrating that the proposed detector truly benefits from exploring different costs for different resolutions by the group cost-sensitive boosting in the training phase; (2) The proposed approach outperforms all the other state-of-the-art methods both in terms of the ROC curves and the *log-average miss rate* on all of the four test sets, indicating that our approach is an effective way for pedestrian detection, especially in multi-resolution cases; (3) Note that some methods in the literature also utilize the additional motion or context information to help detection, while our approach focuses on pedestrian detection in static images and does not take such information into consideration. Nevertheless, a possible future work is to consider motion and context information for further improvement.

Conclusions

In this paper, we have proposed a new group cost-sensitive boosting based approach for handling multi-resolution pedestrian detection. Different from the canonical boosting-based methods in which low resolution samples are treated by equal costs as high resolution ones, so that they are more easily being rejected in the early stage, the proposed approach extends the popular AdaBoost by exploring different costs for different resolution groups in the boosting process, and places more emphases on the hard low resolution samples to better handle multi-resolution detection. The effectiveness of the proposed approach has been validated by its superior performance to other state-of-the-art on different

resolution-specific test sets of the Caltech benchmark.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grants 61371128 and 61532005, National Hi-Tech Research and Development Program of China (863 Program) under Grants 2014AA015102 and 2012AA012503, and China Postdoctoral Science Foundation under Grant 2014M550560.

References

Abe, N.; Zadrozny, B.; and Langford, J. 2004. An iterative method for multi-class cost-sensitive learning. In *SIGKDD*, 3–11.

Bar-Hillel, A.; Levi, D.; Krupka, E.; and Goldberg, C. 2010. Part-based feature synthesis for human detection. In *ECCV*, 127–142.

Benenson, R.; Mathias, M.; Timofte, R.; and Gool, L. J. V. 2012. Pedestrian detection at 100 frames per second. In *CVPR*, 2903–2910.

Benenson, R.; Mathias, M.; Tuytelaars, T.; and Gool, L. J. V. 2013. Seeking the strongest rigid detector. In *CVPR*, 3666–3673.

Benenson, R.; Omran, M.; Hosang, J. H.; and Schiele, B. 2014. Ten years of pedestrian detection, what have we learned? In *Computer Vision - ECCV 2014 Workshops*, 613–627.

Chen, G.; Ding, Y.; Xiao, J.; and Han, T. X. 2013. Detection evolution with multi-order contextual co-occurrence. In *CVPR*, 1798–1805.

Costea, A. D., and Nedevschi, S. 2014. Word channel based multiscale pedestrian detection without image resizing and using only one classifier. In *CVPR*, 2393–2400.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*, 886–893.

Dollár, P.; Appel, R.; and Kienzle, W. 2012. Crosstalk cascades for frame-rate pedestrian detection. In *ECCV*, 645–659.

Dollár, P.; Belongie, S.; and Perona, P. 2010. The fastest pedestrian detector in the west. In *BMVC*, 7–17.

Dollár, P.; Tu, Z.; Perona, P.; and Belongie, S. 2009. Integral channel features. In *BMVC*, 5–15.

Dollár, P.; Wojek, C.; Schiele, B.; and Perona, P. 2012. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* 34(4):743–761.

Dollár, P.; Appel, R.; Belongie, S.; and Perona, P. 2014. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(8):1532–1545.

Fan, W.; Stolfo, S. J.; Zhang, J.; and Chan, P. K. 1999. Adacost: Misclassification cost-sensitive boosting. In *ICML*, 97–105.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D. A.; and Ramanan, D. 2010. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(9):1627–1645.

Levi, D.; Silberstein, S.; and Bar-Hillel, A. 2013. Fast multiple-part based object detection using kd-ferns. In *CVPR*, 947–954.

Lin, Z., and Davis, L. S. 2008. A pose-invariant descriptor for human detection and segmentation. In *ECCV*, 423–436.

Luo, P.; Tian, Y.; Wang, X.; and Tang, X. 2014. Switchable deep network for pedestrian detection. In *CVPR*, 899–906.

Maji, S.; Berg, A. C.; and Malik, J. 2008. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 1–8.

Marín, J.; Vázquez, D.; López, A. M.; Amores, J.; and Leibe, B. 2013. Random forests of local experts for pedestrian detection. In *ICCV*, 2592–2599.

Masnadi-Shirazi, H., and Vasconcelos, N. 2011. Cost-sensitive boosting. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(2):294–309.

Mathias, M.; Benenson, R.; Timofte, R.; and Gool, L. J. V. 2013. Handling occlusions with franken-classifiers. In *ICCV*, 1505–1512.

Nam, W.; Dollár, P.; and Han, J. H. 2014. Local decorrelation for improved pedestrian detection. In *NIPS*, 424–432.

Ouyang, W.; Zeng, X.; and Wang, X. 2013. Modeling mutual visibility relationship in pedestrian detection. In *CVPR*, 3222–3229.

Paisitkriangkrai, S.; Shen, C.; and van den Hengel, A. 2013. Efficient pedestrian detection by directly optimizing the partial area under the roc curve. In *ICCV*, 1057–1064.

Paisitkriangkrai, S.; Shen, C.; and van den Hengel, A. 2014a. Pedestrian detection with spatially pooled features and structured ensemble learning. *CoRR* abs/1409.5209:1–19.

Paisitkriangkrai, S.; Shen, C.; and van den Hengel, A. 2014b. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *ECCV*, 546–561.

Park, D.; Ramanan, D.; and Fowlkes, C. 2010. Multiresolution models for object detection. In *ECCV*, 241–254.

Sabzmeydani, P., and Mori, G. 2007. Detecting pedestrians by learning shapelet features. In *CVPR*, 1–8.

Schwartz, W. R.; Kembhavi, A.; Harwood, D.; and Davis, L. S. 2009. Human detection using partial least squares analysis. In *ICCV*, 24–31.

Sermanet, P.; Kavukcuoglu, K.; Chintala, S.; and LeCun, Y. 2013. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*, 3626–3633.

Sun, Y.; Kamel, M. S.; Wong, A. K. C.; and Wang, Y. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* 40(12):3358–3378.

Ting, K. M. 2000. A comparative study of cost-sensitive boosting algorithms. In *ICML*, 983–990.

Viola, P. A., and Jones, M. J. 2001. Fast and robust classification using asymmetric adaboost and a detector cascade. In *NIPS*, 1311–1318.

Viola, P. A.; Jones, M. J.; and Snow, D. 2005. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision* 63(2):153–161.

Wang, X.; Han, T. X.; and Yan, S. 2009. An hog-lbp human detector with partial occlusion handling. In *ICCV*, 32–39.

Wojek, C., and Schiele, B. 2008. A performance evaluation of single and multi-feature people detection. In *DAGM-Symposium*, 82–91.

Yan, J.; Zhang, X.; Lei, Z.; Liao, S.; and Li, S. Z. 2013. Robust multi-resolution pedestrian detection in traffic scenes. In *CVPR*, 3033–3040.

Zeng, X.; Ouyang, W.; and Wang, X. 2013. Multi-stage contextual deep learning for pedestrian detection. In *ICCV*, 121–128.

Zhang, S.; Bauckhage, C.; and Cremers, A. B. 2014. Informed haar-like features improve pedestrian detection. In *CVPR*, 947–954.