A Two Phase Approach for Pedestrian Detection

Soonmin Hwang, Tae-Hyun Oh, In So Kweon

Robotics and Computer Vision Lab., KAIST, Korea

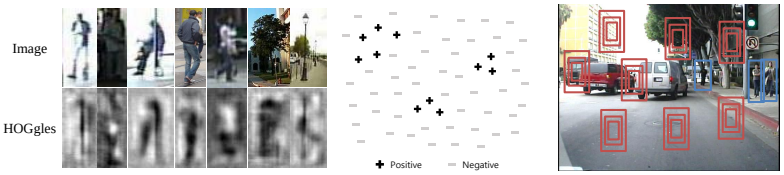
Abstract. Most of current pedestrian detectors have pursued high detection rate without carefully considering sample distributions. In this paper, we argue that the following characteristics must be considered; 1) large intra-class variation of pedestrians (multi-modality), and 2) data imbalance between positives and negatives. Pedestrian detection can be regarded as one of *finding needles in a haystack* problems (rare class detection). Inspired by a rare class detection technique, we propose a two-phase classifier integrating an existing baseline detector and a hard negative expert by separately conquering recall and precision. Main idea behind the hard negative expert is to reduce sample space to be learned, so that informative decision boundaries can be effectively learned. The multi-modality problem is dealt with a simple variant of a LDA based random forests as the hard negative expert. We optimally integrate two models by learned integration rules. By virtue of the two-phase structure, our method achieve competitive performance with only little additional computation. Our approach achieves 38.44% mean miss-rate for the reasonable setting of *Caltech Pedestrian Benchmark*.

1 Introduction

Pedestrian (or Human) detection has been an open research problem in computer vision community for more than decades due to the complexities of human variations and environment. The state-of-the-art approaches still show very high mean miss rate which limits the practical usage [1]. In recent years, pedestrian detection has impressively progressed in terms of feature representations [2–5], learning model [6–15], efficiency [10, 12, 13].

A challenge mainly comes from the large intra-class variations of human like pose and illumination changes. In addition, a lack of positive (human) samples comparing to negative (non-human) causes high asymmetricity in classification problem. These factors are on data characteristics. We are aware that there are very limited works comprehensibly considering the characteristics.

We argue that by considering the characteristics, one can develop a new effective model from a existing method. Based on our analysis of the data characteristics for pedestrian detection, pedestrian detection can be regarded as a *finding needles in a haystack* problem (rare class detection) [16,17], which is one of generic concepts of data mining. Inspired by one of the rare class detection approach [16], we propose a two-phase classifier for pedestrian detection. The proposed two-phase classifier consists of a baseline detector and hard negative expert. We exploit modern successful methods as the first-phase baseline method to reduce sample space to be learned for the second-phase. By virtue of the two-phase approach, we can improve the overall performance with little additional



(a) Multi-modality (b) Data imbalance (c) Sparsity of pedestrian

Fig. 1. Characteristics of pedestrian detection problem. For HOGgles representation, refer to Vondrick *et al*. [20].

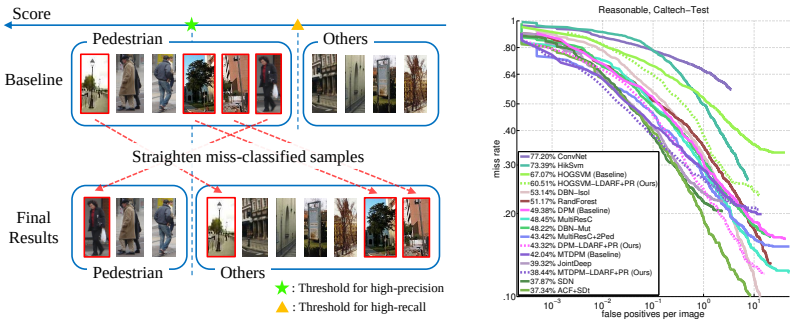
computation without re-computing features. Particularly for the expert model, we extend Random Forest (RF) [18] model to more discriminative one based on the criterion of Fisher's Linear Discriminant Analysis (LDA) [19]. Its purpose is to deal with multi-modality of data automatically and discriminatively, which is not covered by the first-phase. We propose a conjunction rule to effectively fuse the responses of the baseline and expert. As addendum, we present three learning schemes for the expert model to improve discriminative power.

We validate our two-phase model on the challenging *Caltech Pedestrian Benchmark*, and our method achieves the competitive performance against the state-of-the-art methods, although we only use a single feature instead of other rich representations. For reasonable subset, our method achieves at most 38.44% mean miss rate over the baseline. This achievement is based on the following analyses of pedestrian data.

Analysis of Pedestrian Detection Data We concentrate on two aspects which make the pedestrian detection problem challenging: 1) multi-modality among intra-class samples (*i.e.* intra-class variations), and 2) data imbalance of positive/negative samples. To achieve more accurate detection, this kinds of characteristics should be seriously considered and reflected to the designed detector. The following analyses go for other single object detection problems such as face detection.

The multi-modality of pedestrians is formed by high intra-class variations due to pose deformation, view points, appearance, resolution, camera hardware, illumination change, background clutters, skin color, and so forth (Some examples of modalities on HOG domain are shown in Fig. 1-(a)). Based on this fact, we believe that positive samples would conform multi-modality rather than uni-modality (see Fig. 1-(b)). It requires a complex learning model.

The data imbalance of pedestrian detection comes from natural statistics. Only pedestrians are considered as positive class and all the others are regarded as negative. In a image pyramid for multi-scale detection, there are only few pedestrians among millions of sliding windows even in crowd scene as illustrated in Fig. 1-(c). We naturally get a tremendous number of negative samples incomparable to the positives. The imbalance can affect the overall performance of the designed detector through both learning and detection steps. The imbalance on the learning step could cause bias of decision boundary to be negative-oriented, or could induce less informative decision boundaries because some random deci-



(a) Concept of the two-phase model (b) The state-of-the-arts [1]

Fig. 2. Illustrations of motivation. (a) We propose a two-phase model which starts from the detection results of existing one and straighten miss-classified samples to achieve low miss-rate and low false-alarms. (b) Existing methods already achieve low miss-rate at high false positives per image, but they did not achieve low false-alarms still.

sion boundary might be misread as work well; *e.g.* For the ensemble model, the imbalance can induce high sub-optimality on the selection of simple weak classifiers of AdaBoost or randomized forests, because they misread their capacity due to easily achieved high recall. Also, most of detection algorithms have a trade-off between false positive and false negative rates. The imbalance on detection step disturbs finding a good trade-off. For this class imbalance problem, a special treatment may be necessary as rare class detection problems did in [16].

2 Related Works

As pedestrian detection is one of attention-getting topics in computer vision, it has long history and many related works. In this section, we focus on the relevant works to our method. One can refer the thorough review on pedestrian detection approaches to [1, 21].

Many works notice that a main challenge of pedestrian detection comes from multi-modality (intra-class variations) of data. Some researches develop robust and distinctive feature representation such as HOG [2], CSS [4], LBP [3], integral channels [5], and temporal feature [22] which invariant to some modalities like illumination changes or color variances.

On the top of rich feature representation, many learning models are also applied to improve accuracy. Most of works try to deal with some specific variations of pedestrians by advanced learning models. Popular Deformable Part Model (DPM) [7] combines a static root detector and part detectors by latent SVM approach.

It allows flexibility to handle deformation and partial occlusions by latent variables. Also other works [3, 14, 15] have been proposed to handle deformation and occlusions. Ouyang *et al.* [23] learn a background diversity for a limited case, two pedestrians being together. Recently, Park *et al.* [24] and Yan *et al.* [6] argue

[1] In case of MTDPM [6], we use executable code provided by the authors without context model and get 42.04% log-average miss rate.

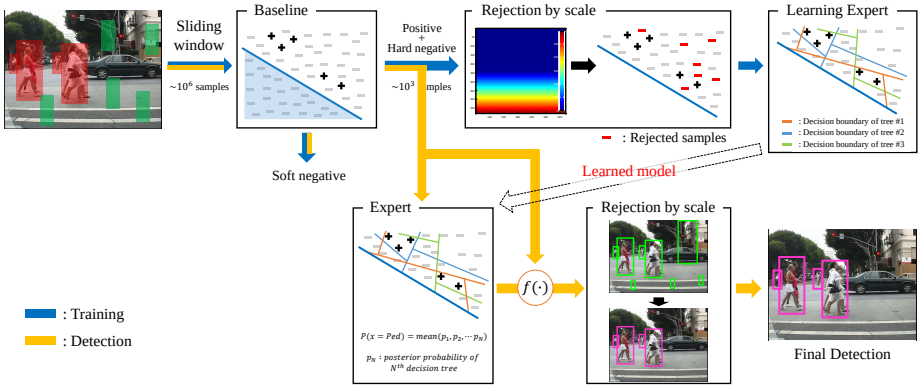


Fig. 3. The proposed two-phase framework. [**Blue line**] In training step, the baseline detector and rejection scheme help gathering positive and hard negative samples for learning the expert discriminatively. [**Yellow line**] In detection step, the baseline discards many data as soft negative. This rejection in the baseline makes the expert check only a small subset of samples. So, a limited amount of additional computation is required to improve performance.

that low and high resolution pedestrians share commonness, but different characteristics should be considered. All these methods provide improved accuracy and higher robustness, but each method focuses on one or two specific intra-class variations. Particularly, the existing linear models are not enough to deal with many kinds of intra-class variations due to its limited parameterizations; *e.g.* The latent variable in DPM is the only parameter to deal with pose variations.

To allow flexibility, ensemble based classifiers are presented for pedestrian detection. Many of them show fast and efficient approach with satisfactory performance. Among the ensemble models, Boosting [8–13] and RF [14, 15] based classifiers are popularly applied for pedestrian detection. They learn non-linear decision boundaries with many weak-classifiers, which share a single nature. Our expert model is developed as an extension of RF. For more relationships with other RFs, we will further discuss in Sec. 3.2. Although these approaches allow to learn multi-modality of data, it would not be enough to handle many different kinds of multi-modality (*e.g.* In DPM, latent variables are only parameters to handle deformable parts. The ensemble models dump the flexibility on uni-nature weak classifiers). Our method utilizes two heterogeneous models, and encourages to capture complementary information during learning time.

We are aware that many methods pass over some traditional data mining rules. Among the contexts of data mining, we found that *finding needles in a haystack* problem [16,17] is very relevant to pedestrian detection problem, which detects rarely occurring phenomena in the data. They define a class that has very rare occurrence due to its nature as 'rare class'. The rare class detection problem is especially challenging. High recall can be easily achieved due to class imbalance in the rare class case. Conventional learning models try to achieve high recall and high precision simultaneously. They are prone to find low precision decision boundary, because it easily achieves high recall. This induces performance degra-

dation. They point out that conventional sequential techniques are inadequate for the rare class problem. Joshi *et al.* [16] propose the two-phase rule induction. In the two-phase induction, recall and precision are separately optimized, rather than simultaneously optimizing two measures as most of learning models did. Their experimental result indicate that the two-phase model outperforms AdaBoost for rare class, and consistently produces competitive performance for generic cases. More general concept of two-phase induction model can be found in [16, 17]. Our method is built on the top of these philosophy.

3 Two-phase Classifier Model under *PNrule*

One of effective approaches for the rare class problem is to separately conquer high recall first and high precision next, which is called as *PNrule* [16]. Inspired by *PNrule*, we propose a two-phase classifier model which minimizes miss rate first, then optimizes our detector to minimize false positives. In the first phase, a detector classifies pedestrians allowing many false positives. Then, the second phase classifier (called as expert) straightens the miss-classifications to achieve high precision. This procedure is illustrated in Fig. 2-(a). By this way, we can achieve low miss-rate at low false positives per image (FPPI) by *PNrule*.

The proposed two-phase classifier model can be viewed as a variant of cascade classifier structures. The conventional approaches learn weak learners which have same properties. Rather than cascading uniform weak learners, exploiting heterogeneous classifiers is more helpful for achieving different objectives (in our case, recall and precision). Even when the same training set is given, heterogeneous classifier models bring out different characteristics in decision boundary or classification results, as well as commonness (intersection regions on feature space among different classifiers). We expect that combining heterogeneous classifiers learns complementary information even from same data, when we carefully choose the classifiers by their properties and data's characteristics.

Our proposed two-phase detector consists of a baseline and a hard negative expert detector as illustrated in Fig. 3. The baseline initially rejects soft negatives which are easily classified with high confidence, and measures how likely positive. Then, the remaining negatives (hard negatives) and positives are passed to the next phase expert. The expert classifies into hard negative or positive on the reduced sample space. By combining results from the baseline and expert, the mis-classifications by the baseline are straighten. If the baseline classifies non-pedestrian as high score, the final results are corrected to have low value. This procedure works like Re-ranking approach [25]. Even though the baseline allows many false positives, the number of samples that have to be checked at the expert detector are surprisingly reduced. Thus, our method only require a limited amount of additional computation, while enhancing overall accuracy.

3.1 First Phase: Baseline Detector

Baseline detector filters out soft negatives, and leaves hard negatives and positives. The main objective of the baseline detector in the two-phase classifier is to minimize miss rate of pedestrians (high recall), while minimizing the number of hard negatives is a subject class designated by the baseline as positive.

The minimized number of hard negatives by the baseline can help to reduce the feature space to be learned by the expert. Since the hard negatives guide the learning decision boundaries of the expert, a well designed baseline detector is preferred to grasp informative hard negatives.

We found that many existing detectors already achieve low miss-rate at high FPPI as shown in Fig. 2-(b). At the first phase, one of the existing algorithms can be used for the baseline. Among publicly available methods, we test linear detectors, HOGSVM [2], DPM [7], and MT-DPM [6], as our baseline detector. Our expert adopts an ensemble model, so that linear detector which has different characteristics would be a reasonable choice. These could be a guideline for selecting baseline, but not strict restriction. Users can select the baseline by their own criteria. We observe that improvements are achieved for all of our experiments among the several baselines. We set the rejection threshold of the baseline lower than usual settings suggested by the authors to allow high recall.

3.2 Second Phase: Hard-negative Expert Detector

We build a expert classifier which classifies hard negatives and positives at the second phase. We can improve the performance by applying the second phase detector once again to the detection results of the first phase. Rather than using two models independently with training by the same sample set, we train the expert with the baseline's results, where hard negatives are designated. It learns complementary information from the reduced learning space, where the union of positives and hard negatives. Also, since the baseline eliminates soft negatives, the data imbalance on both detection and learning steps are partially relaxed.

We try to handle multi-modality of features in the expert model, so we exploit a RF model [18][2] which can better handle multi-modality than linear models. Nevertheless, discriminating the non-pedestrians from the hard negatives is still challenging, because it was failed in the baseline once. There are recent variants of discriminative RF [15, 27–29]. A fundamental difference is how to learn a linear decision boundaries of each node to obtain more discriminative power ([29]: Ridge regression, [15,28]: SVM, [27]: LDA). According to [18], the standard weak learners in [15,27–29] select few dimensions of data vectors randomly before finding split function to construct forests with less-correlation between trees. The feature selection function $\phi(\cdot)$ maps a high dimension vector to a low dimension vector, and can be represented in a matrix form as: $\mathbf{Y} = \phi(\mathbf{X}) = \mathbf{W} \cdot \mathbf{X}$, where $\mathbf{Y} \in \mathbf{R}^{M\times 1}, \mathbf{X} \in \mathbf{R}^{N\times 1}$, $\mathbf{W} \in \{W|\{0,1\}^{M\times N}, \mathbf{W}\cdot\mathbf{1} = \mathbf{1} \in 1^{N\times 1}\}$ is a binary selection matrix, and $M \ll N$. Marín *et al.* [15] extend it to the random subgroup feature selection for encouraging part configuration. Our hypothesis is that feature selection may reduce the probability to find good discriminative boundaries, because high dimension feature is more preferable to find a linear separable space than low dimension one in general. Thus, we instead generalize the feature selection in a soft manner, which can automatically select informative features and $\mathbf{W}$ to be $\mathbf{R}^{M\times N}$.

[2] Since RF model is a general concept of AdaBoost, it has more flexibility for complex decision boundary than AdaBoost [26].

We build our expert model with LDA criterion, because it has an analytical closed-form solution, while SVM based methods cannot avoid expensive numerical optimization. However, the LDA based RF [27] deterministically decides splitting criteria for each node under the strict assumption that each class has exactly same covariance, but the assumption is not satisfied the estimated split function is no longer optimal. By proposing an alternative threshold estimation, we relax the problem. We will explain it later.

Given the hard negative and positive samples, our method learns important and discriminative regions of the pedestrian template based on LDA for each node. By the general definition of Criminisi *et al.* [18], the split function of the j-th node is defined as $h(\mathbf{s};\boldsymbol{\psi}_j,\tau_j) = \mathbf{1}[\boldsymbol{\psi}_j^\top \cdot \mathbf{s} < \tau_j]$, where $\mathbf{s}$ denotes a data sample, $\mathbf{1}[\cdot] \in \{0,1\}$ denotes the indicator function, $\boldsymbol{\psi}$ denotes a transformation that maps the data to a separable space, and τ is a threshold for classification. For a sample $\mathbf{s}$, if $h(\mathbf{s};\boldsymbol{\psi}_j,\tau_j) = 0$, $\mathbf{s}$ is passed to left (or right), and otherwise vise versa. We obtain the parameter $\boldsymbol{\psi}_j$ from LDA. For each node, we use a maximally separable axis $\boldsymbol{\psi}_j$ computed by the following equation.

$$\boldsymbol{\psi}_j = \boldsymbol{\Sigma}_{W,j}^{-1}(\mu_{j,y=\mathcal{P}} - \mu_{j,y=\mathcal{N}}), \tag{1}$$

where $\mu_{j,y=\{\mathcal{P},\mathcal{N}\}}$ represents the mean vector of the positive and negative data of the node j respectively, and $\boldsymbol{\Sigma}_{W,j}$ represents the within-class scatter matrices of the node j for 2-class case [30].

Also we can easily obtain an optimal decision threshold by $\tau_j = \frac{1}{2}\boldsymbol{\psi}_j^\top(\mu_{y=\mathcal{P}} + \mu_{y=\mathcal{N}})$ with the assumption that two groups have the same covariance matrices. However, the same covariance assumption would be too strict. We propose an alternative threshold computation as the following equation:

$$\tau_j(\alpha) = \boldsymbol{\psi}_j^\top(\alpha \cdot \mu_{y=\mathcal{P}} + (1-\alpha)\cdot \mu_{y=\mathcal{N}}), \quad \alpha \in [0,1]. \tag{2}$$

Instead of finding τ, we apply brute-force search on the sampled $\alpha \in [0,1]$ maximizing the following information gain:

$$I(\mathcal{S},\Theta) = H(\mathcal{S}) - \sum\nolimits_{i\in\{L,R\}} \frac{|\mathcal{S}_i|}{|\mathcal{S}|} H(\mathcal{S}_i), \tag{3}$$

where Θ is the set of parameters defining the split function $\boldsymbol{\psi}_j$ and τ_j, $H(S)$ is Shannon's entropy defined as $-\sum_{c\in\mathcal{C}} p(c)\log(p(c))$, and $\mathcal{C}$ is possible class sets (in our case, positive $\mathcal{P}$ and negative $\mathcal{N}$). When each covariance of positive and negative at a node is different, the optimal decision threshold are likely to be between the two means $\mu_{y=\{\mathcal{P},\mathcal{N}\}}$ by Bayes decision rule [30]. This α parameterization gives the bounded sample range, while the range of τ should be estimated from data.

In implementation, we reuse already extracted features by the baseline for all the forests to avoid re-computation. We now discuss the remaining issues of the proposed LDARF.

Discussion of the proposed LDARF LDA is optimal in the sense of Bayes error under the assumptions that multivariate normality and equal covariances

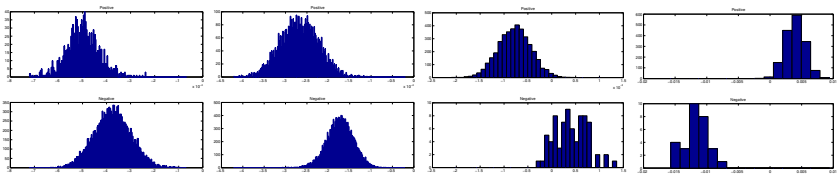


Fig. 4. Distributions of *Caltech* data in four sampled nodes of our proposed LDA based RF. Data samples of each node are projected on LDA axis of the node. (**Top**) Positive distributions. (**Bottom**) Negative distributions. Each column indicates the distributions of the same node. This shows that the data approximately conforms Gaussian distribution in each local partition.

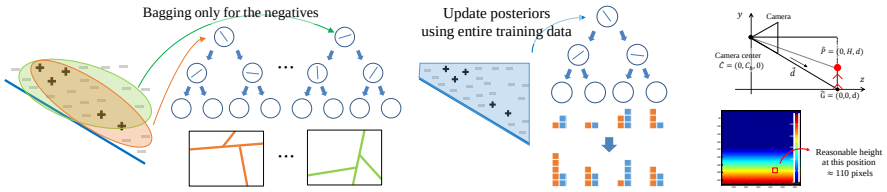
are satisfied by class data [31]. While we relax the equal covariance assumption by introducing α parameter, we still assume Gaussian distribution of samples.

It is true that the assumed model may not be supported by the given data. The general PDF can be approximated by a non-parametric PDF estimation, but Devijver and Kittler *et al.* [32] claim that the errors on nonparametric PDF estimate may significantly exceed those of simple parametric models, such as Gaussian, when the sample size is limited. Also, parametric models could be better in terms of both accuracy and simplicity under the situation. When learning RF, the data is partitioned as growing trees, so that the sample size of a node is exponentially decreased. Thus, Gaussian approximation would valid by the statement of [32]. Fig. 4 show the distribution of the dimension reduced features of *Caltech* data at intermediate nodes. It shows validity of our local Gaussian assumption of each node.

Moreover, the Gaussian assumption is beneficial when determining split functions. In the standard RF [18], Eq. (3) is utilized to measure the goodness of the split function, but Shannon's entropy is defined on discrete distribution which is constructed resultantly when a split function is given. Rather than, suppose that samples follows a parametric model like Gaussian distribution in a small partition. A parametric model has generative property on continuous space and locally spans its feature space, so it is helpful for learning a generalized boundary.

3.3 Integrating the Baseline and Expert

As depicted in Fig. 3, a single final score should be resulted from two scores of the baseline and expert. A well-designed combining rule could straighten misclassified data from the baseline due to complementary characteristics of two detectors. However, hand-crafted combination rules may not be desirable, so we learn a score integration function with the two scores and labels. For simplicity, we model the score integration function with a linear model as $r(\mathbf{x}) = \mathbf{f}^\top \mathbf{x}$, where $\mathbf{f}$ is a weight vector to be learned, and $\mathbf{x}$ is a elementary score vector of which entries come from the baseline and expert scores. To encode several basis rules, we construct the vector $\mathbf{x}$ of each sample as $\mathbf{x} = [s_b | s_e | s_b \cdot s_e | s_b^2 | s_e^2]$, where s_b and s_e are the scores obtained by the baseline and expert respectively. This construction can be regarded as kernelization that maps low-dimensional features into a higher dimensional non-linear space.



(a) (b) (c)

Fig. 5. Illustrations of learning schemes. (a) Soft bagging scheme keeping all the positives to make less-correlated decision trees. (b) Updating the distribution of leaf nodes with whole training set. (c) The estimated reasonable scale map for *Caltech* dataset.

Our goal is to learn the score function that satisfy the relative score order $r(\mathbf{x}_p) > r(\mathbf{x}_n)$ for all the positive samples $\mathbf{x}_p$ and the negative samples $\mathbf{x}_n$. More explicitly, the pairwise order constraint can be represented as:

$$\mathbf{f}^\top \mathbf{x}_p > \mathbf{f}^\top \mathbf{x}_n \quad \Rightarrow \quad \mathbf{f}^\top (\mathbf{x}_p - \mathbf{x}_n) > 0, \quad \forall p, n. \tag{4}$$

Finding $\mathbf{f}$ satisfying all the constraints could not be possible due to the presence of outliers or insufficient sorts of the basic rules. Instead of the hard constraint, we encourage the constraints in a soft manner with maximizing margin similar to SVM classification. Then, with adding a regularization term for f, the learning problem leads the following optimization problem as:

$$\arg\min_{\mathbf{f}} \|\mathbf{f}\|_p + \frac{C}{|P| \cdot |N|} \sum_{i \in P} \sum_{j \in N} \max\left(0, 1 - \mathbf{f}^\top (\mathbf{x}_i - \mathbf{x}_j)\right)^2. \tag{5}$$

where P, N are the positive and negative sample sets of the training set. This formulation shares the similar spirit of the learning to rank technique [33]. Eq. (5) encourage the order constraints by the squared hinge loss, and can be regarded as a l_p regularized SVM. The optimization is effectively solved by the off-the-shelf l_p regularized SVM solvers[3] in the l_1 and l_2 cases. For $p \geq 1$, Eq. (5) is convex formulation, so we can obtain a global optimum solution.

We found the optimal parameter $\mathbf{f}$ for both l_1 and l_2 cases, and empirically observed that the l_1 formulation produces slightly better results. Since l_1 is known to have a sparse selection property, it can be possible that only few informative entries of $\mathbf{f}$ have non-zeros, while discarding unhelpful basic rules in $\mathbf{x}$. Thus, we use l_1-norm for all our experiments. Again, we would like to notice that, although we simply model $r(\cdot)$ with a linear function, $\mathbf{x}$ is constructed by kernelizing the baseline and expert scores, so nonlinear integration rules can be considered directly.

4 Learning Schemes for the Expert Model

Additionally, we describe three more simple learning schemes for expert under our analysis in Sec. 1. The described approaches are simple, but improve the performance of our detector.

[3] We use the G-SVM package used in [34].

Bagging with Preserving Positive Distribution We apply the conventional bagging scheme only for the negative samples with keeping all the positives. It reduces the gap between the number of positive and negative, while positive distribution is preserved. We expect that diverse decision boundaries are around positive distributions. It helps the expert model to well learn decision boundaries and to be generalized by making trees uncorrelated despite the rare positives.

We learn each tree of the hard negative expert ensemble from differently sampled sets by the above bagging scheme. Each tree can share some commonness by sharing the same positives set, while takes different characteristics from different negative subsets.

We simply apply random sampling of negative samples, and it shows plausible results.

Updating Leaf Distribution of LDARF In RF, each leaf node stores the positive and negative posterior distributions of the training samples arrived at the node. The posterior of a tree $p_t(c|\mathbf{v})$ is defined by the posterior of the leaf node that the sample $\mathbf{v}$ reaches. Given this, the final decision for $\mathbf{v}$ is determined by $c^* = \arg\max_{c\in\{\mathcal{P},\mathcal{N}\}} p(c|\mathbf{v})$, where $p(c|\mathbf{v}) = \frac{1}{T}\sum_{t=1}^{T} p_t(c|\mathbf{v})$ is the average for every tree by aggregation rule [18].

Constructing accurate posterior distributions is as important as finding good split functions. As we use bagging scheme, each tree is built only with sampled training data, and initially constructed posteriors do not reflect entire training data in our framework. We update the positive and negative posteriors in leaf nodes by expectation with the remaining training data after bagging. Thus, the estimate $p(c|l)$ at a leaf l is calculate as

$$n_l = \frac{1}{n(\mathcal{P})}\cdot n_{p,l} + \frac{1}{n(\mathcal{N})}\cdot n_{n,l}, \qquad n_{c,l} = \sum_{\mathbf{v}\in l} \mathbf{1}\left[\mathbf{v}\in c\right], \tag{6}$$

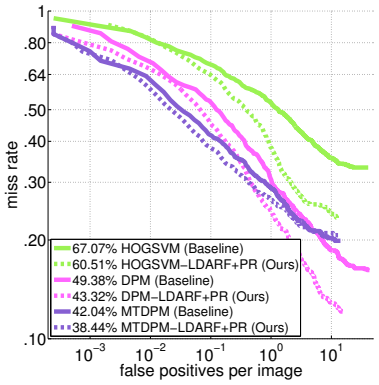
$$p(c|l) = \frac{1}{n(c)}\cdot\frac{n_{c,l}}{n_l}, \qquad c\in\{\mathcal{P},\mathcal{N}\}, \tag{7}$$

where $n(x)$ is the cardinality of a set x, $\mathbf{1}[\cdot]$ is indicator function. $p(c|l)$ is weighted posterior for relaxing data imbalance between positive and negative.

Learning by Perspective Aware Rejection Ground plane information is effectively utilized in Hoeim *et al.* [35] and Park *et al.* [24], and is shown to be beneficial for validating the detected locations and scales. Although the perspective information in detection step has been commonly utilized in some cases like the driving scenario, we extent its usage in the learning step.

It is possible that the training sets given by the baseline include unreasonable samples with respect to scale and location in the perspective world. Since samples from different resolution have different natures as argued in [6, 24], the hard negative samples with unreasonable scale could introduce unnatural artificial features in the sample space to be learned.

To reject these unreasonable scale samples, we estimate reasonable height map using intrinsic camera parameters given in [1] and geometric relation between a camera and pedestrians. As you can see in Fig. 5-(c), we estimate rough



(a)

Miss rate(%)	Methods
49.38	DPM
47.50	DPM+RF
47.06	DPM+PR
43.32	DPM+RF+PR
42.04	MTDPM
38.52	MTDPM+RF
42.15	MTDPM+PR
38.44	MTDPM+RF+PR

Step	Time
Feat. pyramid	2.17 sec
Baseline detector (HOGSVM+GPU)	0.80 sec
Baseline detector (DPM)	1.31 sec
Baseline detector (MTDPM)	1.63 sec
Expert detector (LDARF)	**0.17 sec**

(b)

Fig. 6. Quantitative evaluations of the proposed two-phase classifier on the reasonable subset of *Caltech*. (a) Miss-rate comparisons between the baseline and the proposed two-phase approach. (b)-[**Top**] Performance comparisons according to combinations of the proposed modules. (b)-[**Bottom**] Computation times.

heights of pedestrians in pixels utilizing following relation; $f : d = h : H \Rightarrow h = \frac{fH}{d}$ (f : focal length, d : depth, h : height in image, and H : height in real world)

We allow margins for rejecting samples to alleviate error of the estimated height map. Thus, we reject the candidate boxes which is taller than 1.5 times or shorter than 0.5 times of reasonable height. It could reduce the variance of negative samples in feature space. This seems to be very simple, but it improves the performance of the final detector (see Sec. 5.3).

5 Experimental Results

To focus on the effects of each module, we fixed the used feature with HOG [2]. In this section, the comparisons between our method and other approaches based on HOG feature are only shown to easily compare the effects of the proposed method. Comparisons with more than other recent approaches can be found in the supplementary material.

We evaluated on *Caltech* benchmark [1] which is the challenging and latest pedestrian benchmark. For training both the baselines and expert in our framework, we used both set00-set05 in *Caltech* and training set in *INRIA* dataset [2] due to the lack of information from high-resolution pedestrians in *Caltech*.

To compare performances, we followed full image evaluation and miss rate against FPPI (False Positives Per Image) plot by varying the threshold on the detection confidence as in [1]. For testing, we used set06-set10 in *Caltech*. We repeated the whole training and evaluation process 5 times and report averaged values.

Since the existing detectors find more than 80% of pedestrians at 10 FPPI as shown in Fig. 2-(b), we set the rejection threshold of the baseline detectors to a value corresponding to 10 FPPI. Total 72 different scale images are used for all the experiments, as in the default setting of [7].

(a) Examples of the improved results

Miss rate	Methods
49.38	DPM (Baseline)
45.72	DPM+oRF-LDA [29]+PR
44.71	DPM+StdRF [26]+PR
44.39	DPM+SVMRF [15]+PR
43.32	DPM+LDARF(Ours)+PR
47.50	DPM+RF
47.06	DPM+PRD
47.77	DPM+RF+PRL
43.50	DPM+RF+PRD
43.32	DPM+RF+PRL+PRD
43.85	DPM+RF+PR (+)
43.98	DPM+RF+PR (x)
43.32	DPM+RF+PR (Opt.)

(b) Effect of each module

Fig. 7. Effects of the expert model. (a) Sampled results of Straightened detection at 1 FPPI ([**Top**] DPM [7], [**Bottom**] Ours (DPM+LDARF+PR)). We denote the false positives and true positives as *Yellow* and *Magenta* respectively. (b) Comparisons according to the expert types (**Top**, Sec. 3.2), the perspective-aware rejection scheme in the learning and detection step (**Middle**, Sec. 4), and integration rules (**Bottom**, Sec. 3.3)

The computation times shown in the bottom of Fig. 6-(b) are measured on a PC with 3.40GHz i7-4770 CPU and 32Gb RAM. The computation time of the proposed framework depends on the choice of baseline detector. For **DPM** [7] and **MT-DPM** [6], it takes 3.48 and 3.8 sec respectively (feature pyramid construction + applying DPM or MTDPM). **HOGSVM+GPU** [36] takes 0.80 sec including the feature pyramid construction and detection times. Our proposed two-phase classifier model only takes additional 0.17 sec per an image on MATLAB+MEX code to improve the performance. It can easily speed up by parallelization with modern GPU techniques due to the independent structure of trees. Also, the score integration takes 2.3 ms in average.

5.1 Evaluations of Two-phase Classifier Model

We compare the performances of existing methods with the boosted performances by our two-phase model. We apply our approach to **HOGSVM** [2], **DPM** [7] and **MT-DPM** [6] as a baseline, of which the cores are the linear classifier model based on HOG feature.

As depicted in Fig. 3, our expert detector is trained from the detection results of the baseline detector. Thus, a training set for expert depends on the baseline detectors, so each expert is learned differently according to the baseline. As shown in Fig. 6-(a), our model improves the accuracy of the baseline detector from 3.60% to 6.56%.

Notice that, for **MT-DPM**, we use executable code provided by the authors without context model, so that the baseline detector trained by only *Caltech benchmark*. Also, the expert is trained with fewer sampled data, because the executable only provides sampled results by non-maximum suppression (NMS) [1], while other experts for **HOGSVM** and **DPM** are learned from samples without NMS. Despite this handicap, the proposed two-phase model still improves the performance of **MT-DPM** as 3.60%. On *Caltech* benchmark, most of methods

exploiting only HOG feature did not achieve below 40 % miss rate except MT-DPM+Context [6][4] which utilizes another vehicle detector to utilize a high level contextual relationship between vehicles and pedestrians.

The proposed two-phase model mines HOG feature for more information to improve performance without additional advanced features and show the consistent improvement which allow better performance than the previous HOG-based algorithms.

5.2 Evaluations of Discriminative Random Forests

We compare expert detectors with the axis-aligned model (**StdRF**) [18], Marín *et al*. (**SVMRF**) [15], the ridge regression based LDA model (**oRF-LDA**) in [29] and the proposed LDA based discriminative RF (**LDARF**). Contrary to Sec. 5.1, in order to maintain training data to be same for all the expert detectors, we fixed the baseline with DPM [7]. For fair comparison, parameters such as the depth of tree and the number of trees were set to 6 and 100 respectively for all the experts.

As shown in the top of Fig. 7-(b), our **LDARF** shows better performance compared to other RF as a expert. This implies that LDARF was learned more discriminatively. Fig. 7-(a) shows that mis-classified instances by the baseline are well straightened by our expert.

5.3 Influences according to Each Components

Fig. 7-(b) shows the influence according to each component of our framework, such as the choice of the expert types, the perspective-aware rejection (PR), and the integration rule.

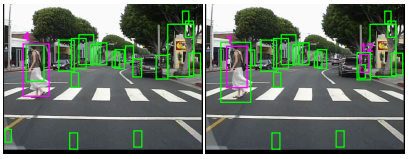
As mentioned in Sec. 4, we use PR at both detection and learning step. It means that we intend to completely ignore particular instances that are not matched with reasonable height according to their position during both learning and detection time. As shown in the middle of Fig. 7-(b), applying both **PRL** (PR in Learning step) and **PRD** (PR in Detection step) (denoted by **DPM+RF+PRL+PRD**) improves the performance compared to the single usage of PRL or PRD. We notice that **DPM+RF+PRL** would produce more false positives than DPM+RF, because in learning step, **DPM+RF+PRL** ignores the samples with unreasonable heights and supposes that the given candidates for the expert are already filtered by PR. In this case, **DPM+RF+PRL** has not been learned for other resolution candidates, so that could not distinguish un-reasonable height instances from positive. When both PRL and PRD are applied, we can expect that the sample space to be learned get reduced and focused to distinguish the positive and hard negatives with excluding effects of un-reasonable resolutions of instances.

In the bottom of Fig. 7-(b), we try to find a good integration rule of two scores from the baseline and the expert. We compare a simple addition rule ($a \cdot s_B + b \cdot s_E$), multiplication rule ($(a \cdot s_B + b) \cdot s_E$), and the proposed optimal integration rule in Sec. 3.3. For the addition and multiplication rule, parameter sweeping for a and b is applied to empirically find the best combination. Our

[4] Codes for MT-DPM+Context was not provided by author when this paper is submitted.

Rank in an image	**Average rank**	1	2	3	4	5	6	7	8
DPM	**2.7391**	429	190	75	39	34	24	6	5
Proposed	**2.3072**	457	200	80	37	24	14	20	11

(a) Statistics for rank of true positives



(b) DPM (c) DPM+RF

Fig. 8. (a) Summary of true positive ranks. Each entry of the table denotes the number of true positives with i-rank. (b,c) Sample results by DPM and the proposed two-phase model without PR (DPM+RF). True positives are denoted by *Margenta* color with their rank value.

optimized integration rule shows better performance. Although the differences of the performances are marginal, the proposed method can suggest more plausible rules than the heuristically found best rules with high probability due to stable performance from the convex formulation. If one adds more complex rules by aggregating to the rule vector, better integration could be automatically found.

5.4 Analysis for performance improvement

In the performance measure suggested by [1], miss rates at sampled FPPIs are calculated by varying threshold. This implies that only relative scores (*i.e.* ranks) between true positives and false positives are important. As many true positives get higher ranks, less miss rate can be achieved. Therefore, to analyze why and how our approach improve the performance, we count the number of instances for each rank. We count rank of true positives in each image and summarize them for test images. The proposed two-phase model and the integration of two scores operate as a re-ranking process. The average rank of true positives is reduced from 2.7391 to 2.3072 by applying the proposed method (Fig. 8-(a)). As shown in Fig. 8-(b,c), the re-ranking caused by the proposed method allows higher threshold (less false positives) while true positives are kept.

6 Conclusion

We present a two-phase framework for pedestrian detection inspired by data mining philosophy, especially from the rare class detection. The baseline and expert detectors, which have different characteristics, optimally integrated by max-margin criteria without any heuristics. We validate our method on the systematical experiments and analyze re-ranking effects. We believe that the consistent improvements by our method are mainly comes from the samples space reduction to be learned. The beauty of our approach is that it can be easily adopted to an existing detector as an add-on module, and can improve the performance with little additional computation.

Acknowledgement This work was supported by the Development of Autonomous Emergency Braking System for Pedestrian Protection project funded by the Ministry of Trade, Industry and Energy of Korea. (MOTIE)(No.10044775)

References

1. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE Trans. on PAMI (2012)
2. Dalal, N., Triggs, B.: Histogram of oriented gradient for human detection. CVPR (2005)
3. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. ICCV (2009)
4. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. CVPR (2010)
5. Dollár, P., Tu, Z., Perona, P., Belonggie, S.: Integral channel features. BMVC (2009)
6. Yan, J., Zhang, X., Lei, Z., Liao, S., Li, S.Z.: Robust multi-resolution pedestrian detection in traffic scenes. CVPR (2013)
7. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. on PAMI (2010)
8. Bourdev, L., Brandt, J.: Robust object detection via soft cascade. CVPR (2005)
9. Zhang, C., Viola, P.A.: Multiple-instance pruning for learning efficient cascade detectors. NIPS (2007)
10. Dollár, P., Appel, R., Kienzle, W.: Crosstalk cascades for frame-rate pedestrian detection. ECCV (2012)
11. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV (2004)
12. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. CVPR (2012)
13. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Trans. on PAMI (2014)
14. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. CVPR (2009)
15. Marín, J., Vazquez, D., Lopez, A.M., Amores, J., Leibe, B.: Random forests of local experts for pedestrian detection. ICCV (2013)
16. Joshi, M.V., Agarwal, R.C., Kumar, V.: Mining needles in a haystack: classifying rare classes via two-phase rule induction. ACM SIGMOD (2001) 91–102
17. Weiss, G.M.: Mining with rarity: a unifying framework. ACM SIGKDD (2004)
18. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations and Trends in Computer Graphics and Vision (2011)
19. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of eugenics (1936)
20. Vondrick, C., Khosla, A., Malisiewicz, T., Torralba, A.: Hoggles: Visualizing object detection features. In: ICCV, IEEE (2013)
21. Geronimo, D., Lopez, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. IEEE Trans. on PAMI (2010)
22. Park, D., Zitnick, C.L., Ramanan, D., Dollar, P.: Exploring weak stabilization for motion feature extraction. CVPR (2013)
23. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. CVPR (2013)
24. Park, D., Ramanan, D., Fowlkes, C.: Multiresolution models for object detection. ECCV (2010)
25. Hsu, W.H., Kennedy, L.S., Chang, S.F.: Reranking methods for visual search. IEEE MultiMedia (2007)

26. Breiman, L.: Random forests. Machine Learning (2001)
27. Lemmond, T.D., Chen, B.Y., Hatch, A.O., Hanley, W.G.: An extended study of the discriminant random forest. Data Mining (2010)
28. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. CVPR (2011)
29. Menze, B.H., Kelm, B.M., Splitthoff, D.N., Koethe, U., Hamprecht, F.A.: On oblique random forests. In: Machine Learning and Knowledge Discovery in Databases. Springer (2011)
30. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley-Interscience (2001)
31. Hamsici, O.C., Martinez, A.M.: Bayes optimality in linear discriminant analysis. IEEE Trans. on PAMI (2008)
32. Devijver, P.A., Kittler, J.: Pattern recognition: A statistical approach. Prentice-Hall London (1982)
33. Joachims, T.: Optimizing search engines using clickthrough data. In: ACM SIGKDD. (2002)
34. Flamary, R., Jrad, N., Phlypo, R., Congedo, M., Rakotomamonjy, A.: Mixed-norm regularization for brain decoding. Computational and Mathematical Methods in Medicine (2014)
35. D. Hoiem, A. Efros, M.H.: Putting objects in perspective. IJCV (2008)
36. : Opencv 3.0. (In: http://opencv.org/)