Journal of Computers Vol. 29 No. 5, 2018, pp. 15-28
doi:10.3966/199115992018102905002

Deep Convolution Neural Networks Cascaded Improved Boosted Forest for Pedestrian Detection

Zhi-Tong Xu[1,2], Yan-Min Luo[1,2*], Pei-Zhong Liu[3], Yong-Zhao Du[3]

[1] College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
{xzt, lym}@hqu.edu.cn

[2] Key Laboratory for Computer Vision and Pattern Recognition of Xiamen City, Huaqiao University, Xiamen 361021, China

[3] College of Engineering, Huaqiao University, Quanzhou 362000, China
pzliu@hqu.edu.cn, yongzhaodu@126.com

Received 18 May 2017; Revised 18 September 2017; Accepted 6 October 2017

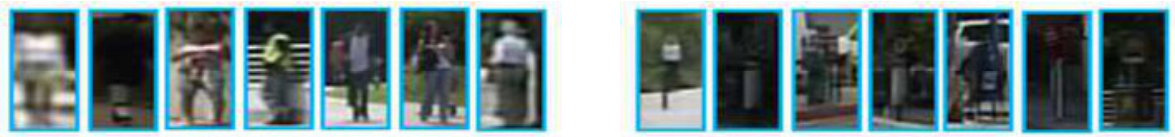
Abstract. Due to the resolution of small size pedestrian is relatively low, and the hard negative background is very similar to people, therefore, detecting small size pedestrian or detecting pedestrian from hard negative background still a challenging problem in computer vision. In order to effectively address these problem, we propose a novel deep convolution neural networks, and cascade an improved boosted forest classifier method to detect pedestrian. Firstly, by using selective search method to propose pedestrian candidate boxes with confidence scores for utmost retaining image resolution; then, based on these proposed confidence values, adopting convolution neural network model to extract candidate regions feature maps; finally, we improve the boosted forest classifier and cascade it to classify candidate boxes for achieving efficiently pedestrian detection. Extensive experiments on Caltech and KITTI benchmarks demonstrate the proposed method outperforms the state-of-the-art, achieves promising precision on KITTI and the lowest miss rate of 11.53% on Caltech, outperforming the second best method (CompACT-Deep) by 0.17%.

Keywords: deep convolution neural networks, hard negative background, improved boosted forest classifier, pedestrian detection, small size pedestrian

1 Introduction

Pedestrian detection as the hot research topic in the domain of computer vision, which has drew wide attentions for decades [1-2] and has diverse practical applications such as video surveillance, human-computer interaction, automatic safe driving and human behavior analysis. With the constantly optimize of deep learning models [3-5], pedestrian detection has reached a good detection performance on still image benchmarks (e.g. INRIA [6], which contains only upright persons with person height > 100 pixels) and has achieved significant improvement in both effectiveness [7] and efficiency [8]. Unfortunately, detecting pedestrian from dynamic videos captured in real scenes (e.g. Caltech Pedestrian Benchmark [9]) are susceptible to complex background, shooting angle, illumination variation and other factors, which leads to the miss rates of pedestrian detection are quite high, especially in detecting small size or low resolution pedestrian (as shown in Fig. 1(a)) from videos, as well as detecting pedestrian from hard negative background (as shown in Fig. 1(b)), where the patches of positive samples and hard negative samples are difficult to distinguish.

* Corresponding Author



(a) Low-resolution samples (Pedestrian) (b) Hard negative samples (Background)

Fig. 1. Distinguishing small pedestrians from the hard negatives samples with high visual similarity in Caltech [10]. In which (a) represents pedestrian object and (b) stands for people-like background samples

Most prior pedestrian detection methods require strictly constraint application scenes, such as low noise or static background, while such assumptions are almost impossible to be satisfied in real life. Moreover, these traditional pedestrian detection models need hand-crafted design complex feature extraction method, to obtain pedestrian feature representation information from the original image for further learning the pedestrian classifier. To a great extent, these models rely on the specific detection tasks, therefore, conventional pedestrian detection models generalization ability are greatly limited.

More recently years, researches on deep learning model have been in full swing. Deep learning model is proposed by Hinton and Salakhutdinov [11], which can layer-wise learn feature in bottom-up approach, and has proved that multilayer neural networks can learn more discriminatively pedestrian features. Convolution neural network is a kind of deep learning model, by combining multilayer artificial neural network and the convolution operation, which can avoid complex feature extraction stage, and has a good robustness in pedestrian scale deformation. Yet due to the high visual similarity [3] between pedestrian and complex hard-negative background, conventional convolution neural networks are easy to have confused pedestrian with hard negative samples, especially in detecting small size or low resolution pedestrian.

In order to effectively detect low-resolution pedestrian or detect pedestrian from hard negative backgrounds. We attempt to improve the boosted forest classifier and cascade it to deep convolution neural network, few ideas are borrowed from papers [5, 12], and propose an efficient pedestrian detection framework. Many experiments shown that the proposed approach outperforms all prior methods and have yielded the state-of-the-art in Caltech [9], ETH [13] and KITTI [14] pedestrian datasets. The main contributions of our work are several folds as follows:

- We use selective search strategy to generate pedestrian candidate regions, which can efficiently reserve image resolution; we also apply multiple reference boxes of different sizes to produce fixed aspect ratio candidate locations at different scales, which can capture a broader range of image scales and has a good detection effect on low-resolution pedestrians in dynamic scenes.
- We adopt deep convolution network to extract ROI fixed-length feature maps, and fully consider the ranking of confidence scores of proposal bounding boxes, to train and improve the boost forest classifier for pedestrian detection.
- Our approach outperforms the second best method (CompAct-Deep) by 0.17% and reaches the state-of-the-art.

2 Related Work

Pedestrian detection is considered as an essential stage for most computer vision tasks, researchers have been making much effort to improve pedestrian detection performance, and their research results have achieved remarkable successes in some specific scene [15]. Now, we give a brief review of related generic pedestrian detection methods.

Many pedestrian detection methods have been proposed in early period of the research. For example, Dollár et al. [16] proposed Integral Channel Features (ICF), which is also known as ChnFtrs and considered as one of the most successful pedestrian detector, many of its variants (ACF [1], LDCF [17], Checkboards [18]) have been proposed successively. In order to further boost visual content representation capability, Dollár et al. [1] have combined ACF with other low-level features, such as HOG [6], LBP [19]. Zhang et al. [20] proposed InformedHarr, while it is specifically designed for up-right human body detection. Nam et al. [17] use Linear Discriminant Analysis to decorrelate LDCF

features, make it more suitable for orthogonal decision trees. Zhang et al. [18] generalized ICF as Checkerboards for further improving pedestrian detection performance. However, according to paper [21], we know that all those methods detection speed are slow, and those methods are difficult to efficiently detect small size pedestrian from dynamic scenes or detect pedestrian from hard negative background.

In recent years, deep learning network model have gradually replaced traditional machine learning method, especially methods based on convolution neural network, which can learn rich discriminatively features from image raw pixels and have achieved impressive pedestrian detection performance [22-26]. For example, Girshick et al. [27] have combined region proposals with CNNs and propose a R-CNN method, which achieves a good performance boost by fine-tuning regions-of-interest (ROI) for pedestrian detection. More recently, Tian et al. [3] use ACF detector to produce regions, and jointly optimizes pedestrian detection with related semantic tasks by training R-CNN network. Tomé et al. [5] and Tian et al. [24] have proposed the DeepPed detector and DeepPart detector respectively, both of which have employed LDCF to proposal regions, while we found that these proposal strategies can not to maintain the original image resolution very well. Uijlings et al. [28] employed selective search method to generate proposal regions [29-30], which can keep the original image resolution through region merge, and have achieved good performance. Cai et al. [23] push features of high complexity to the later cascade stages by optimizing classification risk under a complexity constraint to learn boosted classifiers.

In this work, our goals are to effectively detect low resolution pedestrian from dynamic scenes, as well as detect pedestrian from hard negative background. Zhang et al. [12] adopt the region proposal network followed by boosted forest and propose an effective baseline for pedestrian detection, which demonstrated the classification performance of boosted forest in Caltech and KITTI. Our approach is quite similar to this architecture, but differs in that we fully consider the confidence scores of proposal bounding boxes in training classifier stage, trained and improved the boost forest classifier for pedestrian detection. Table 1 presents the miss rate comparison of pedestrian detection between our approach and the related works, extensive experiments show that the proposed method can effectively decline miss rate in detecting pedestrian.

Table 1. Comparisons of the miss rate between our method and the related works in the Caltech dataset

Methods	Miss rate (%)
ChnFtrs [16]	41.88
InformedHaar [20]	34.60
LDCF [17]	24.79
RCNN [27]	23.30
TA-CNN [3]	20.86
DeepPed [5]	19.92
Checkedboards [18]	18.47
RPN+BF [12]	12.60
DeepParts [24]	11.89
CompACT-Deep [23]	11.70
Ours	**11.53**

3 Our Methods

The pipeline our proposed for pedestrian detection is shown in Fig. 2. The whole process including three steps: pedestrian candidate region proposal, ROI feature extraction and candidate region classification. Given a small size and low resolution video clip image is shot by on-board camera, in order to reserve image original resolution, the pedestrian candidate bounding boxes with confidence scores are firstly generated by using selective search method is proposed in [28]; and then, convolution neural network is employed to extract ROI convolution feature maps; at last, based on [12], through comprehensive consider region proposal scores, we improved the boosted forest classifier and cascade it to deep convolution neural network for detecting pedestrian.

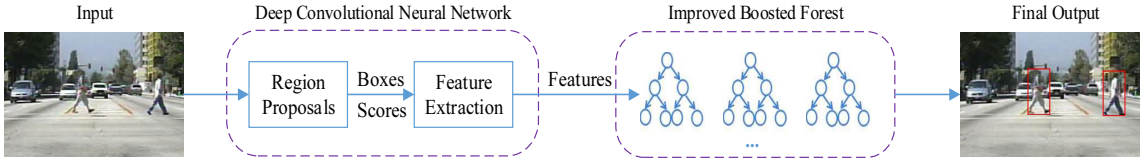


Fig. 2. Our pipeline for pedestrian detection

3.1 Selective Search for Pedestrian Candidate Region Proposal

Pedestrian candidate region proposal is a considerable crucial step of our pipeline for pedestrian detection, which need as many retain the positive regions and discard the negative regions as possible for generating possibly pedestrian locations. The selective search approach was proposed in [28] as a proposer, for multi-category object recognition have yielded surprisingly good results. As a single-category object detection task, pedestrian detection can cater to use selective search detector. Thus, we have adopted selective search (SS) strategy in this stage for pedestrian detection, the details are described as follows.

Firstly, the hierarchical grouping algorithm [28] is taken to achieve selective search for region proposal, pedestrian average aspect ratio is then specified as 0.41 the same as [9], and as the fixed aspect ratio of proposal box, because too many arbitrary scale proposal boxes are not only high computation cost, but also easy to cause noise, thereby result in region proposal accuracy decrease. Moreover, the possibly candidate locations at different scales can normally generated by using 16 reference boxes of different sizes, starting from 18 pixels width (minimum pedestrian width in Caltech [9]) with 1.2× scaling stride. This strategy satisfies to capture a broader range of image scales, and getting rid of the defect of traditional features pyramid in detecting multi-scale objects, which has a good effect on detecting small size pedestrian objects from dynamic scenes or detecting pedestrian from hard negative background.

The region proposal flowchart is shown in Fig. 3, cropping multiple patches for each scales of proposal boxes respectively, and those cropped patches that have intersection-over-union (IoU) with the ground truth box greater than or equal to threshold are considered as positive proposal, otherwise, the patches are viewed as negative proposal (patches containing background, rather than pedestrian), then normalize these patches into a fixed scale of 192×64 pixels to form a proposal region pool. The IoU is the intersection of the proposal box and ground truth, which can be defined as equation (1),

$$IoU = \frac{Pb \cap GT}{Pb \cup GT}, \tag{1}$$

where Pb and GT indicates the region proposal box and ground truth respectively. For instance, arbitrary scales an image is inputted, resize image scale to 227×227 and adopt the method [28] to propose candidate region and to further form a region pool. According to the annotated pedestrian data set, the intersection-over-union (IoU) between the proposal regions and ground truth box are calculated, the illustration of which is shown in Fig. 4. And obtaining confidence scores of the proposal boxes, decided by the overlap rate between it and ground truth to estimate object label by threshold, in other words, whether a proposal region encloses a pedestrian or not. In our experiment, the threshold of IoU is set as 0.6.

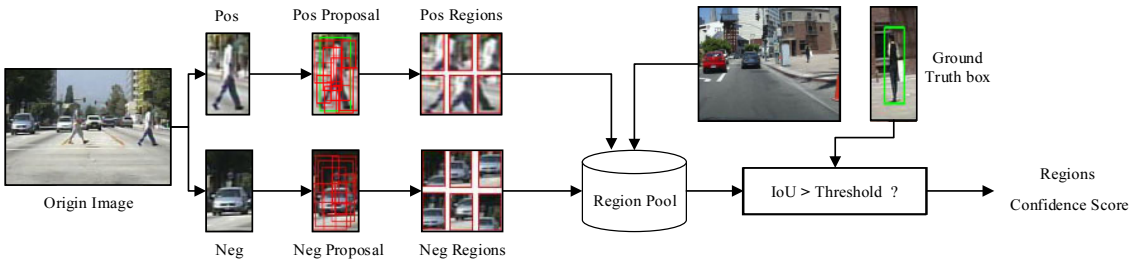


Fig. 3. The flowchart of pedestrian candidate region proposal

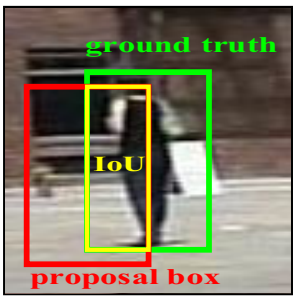


Fig. 4. The illustration of intersection-over-union between proposal box and ground truth

3.2 Region-of-Interest Feature Extraction

After obtaining the proposed windows with confidence score, all of them are chose as the region of interest, and the ROI relevant features are extracted. Firstly, the training set S is constructed by combing the cropped patches from N scales of proposal boxes, let $S=\{(X_n,s_n)\}_{n=1}^{N}$ be a set of image patches and their scores, where $X_n=(x_n,y_n,l_n,m_n)$ is a four-tuple, in which, (x_n,y_n) representing image patches top-left vertex abscissa and ordinate, and (l_n,m_n) indicating image patches width and height, respectively. Then, inspired by the successful application in image classification of AlexNet [31], and the particularity of pedestrian detection is considered, four average pooling layers, four convolution layers alternately and two fully connected layers are stacked together and replace the first fully connected layer with a max pooling layer, the rectified linear function are employed as the activation function for all layers, and adopt ROI pooling algorithm [32] to extract ROI fixed-length feature maps for training classifier that can address the problem of feature dimension restriction, our unconstrained feature dimension can effectively detect pedestrian from low resolution image, the overview of our deep model is shown in Fig. 5.

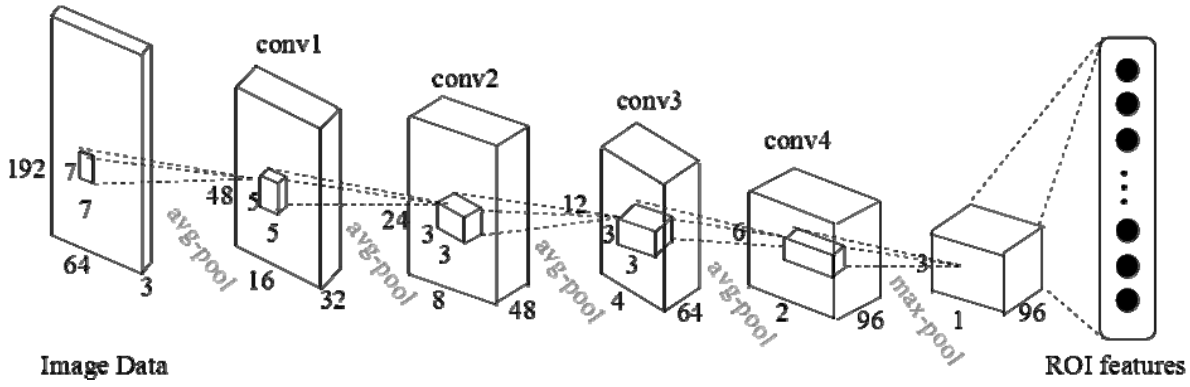


Fig. 5. The overview of our deep model

The ROI features can be extracted on Conv3 and Conv4, and concatenated these different layers features to form a feature maps, then pool the feature maps into a fixed resolution of 7×7. For a low resolution input image, the features of ROI be not discriminative, thus, we need to increase resolution for computing convolution feature maps. For example, the third pool layer stride can be set 1 and the stride of all the fourth convolution layer be reduced half by expanded filters twice, in our implement, the enlarged filter is used to perform convolution operation for extracting ROI feature.

3.3 Improved Boosted Forest Classifier

After obtaining region proposals with confidence scores as well as the ROI features, we have improved the Real AdaBoost according to the proposal bounding box confidence score based on the preliminary work of the Real AdaBoost algorithm and the configuration of the learned cascade in [23], and trained an improved boosted forest classifier for pedestrian detection.

According to [33], in the initial training phase, the training set Ψ consists of all positive regions

$\mathrm{Z}=\{\mathrm{Z}_m\}_{m=0...M}$ and the same number of negative regions $\zeta=\{\zeta_m\}_{m=0...M}$, and the forest has only 32 trees. Then, the training set was bootstrapped 6 times, and the remaining hard negative regions were added to the training set after each bootstrap, during the former 5 times bootstrap, the forest has {128, 256, 512, 1024, 1536} trees respectively. After the last bootstrapping, the forest classifier is trained by cascading 2048 decision trees with a depth of 5.

In the original boosted forest classifier training, the region proposal confidence score is used as initial reference only, which has ignored its impact on the bootstrap process. According to the rankings of bounding box confidence scores, all proposed candidate regions are unequally treated, that is, the higher confidence score of proposal region bounding box, the greater probability it belongs to pedestrian y, thus, we set different weights to different proposal boxes. First, by summing the confidence scores s of all proposed regions, and calculate the percentage p_i of each confidence score s_i account for confidence scores sum S; then, we use the initial weight $w_0=\frac{1}{2}\log\frac{s_0}{1-s_0}$ and replace the rest of all bootstrap weights $w_i=\frac{1}{M}, i=1...M$ in the original Real AdaBoost with probability $p_i=\frac{s_i}{S}, i=1...M$. Based on weight $w_i, i=1...M$, we use weak learning method to fit the proposal boxes probability $p_m(x_i)=P(x_i \mid y=1)\in[0,1], x_i\in\Psi, m=0...M$, and based on probability estimation $p_m(x_i)$, we further calculate classifiers $f_m(x)$ to boost forest, the $f_m(x)$ can be defined as equation (2),

$$f_m(x)=\begin{cases}\frac{1}{2}\log\frac{s_0}{1-s_0} & ,m=0\\ \frac{1}{2}(1-\lambda)\log\frac{p_m(x_i)}{1-p_m(x_i)}+\frac{1}{2}\lambda\log\frac{s_0}{1-s_0}, & m=1...M\end{cases}, \tag{2}$$

where $\lambda=0.3$ is a parameter, which controls the importance of the initial weight in the bootstrap process. The training procedure is summarize in Algorithm 1.

Algorithm 1: Improved Real AdaBoost Forest Classifier Training

Input: Training set: $\Psi=\{\mathrm{Z},\zeta\}$
Confidence score of each proposal region: s_i
Initial weight: w_0

Output: All improved weak forest classifiers: $f_m(x)$

Compute the confidence scores sum S of all proposed regions;
Calculate the percentage $p_i, i=1...M$ of each confidence score in sum S;
Replace all bootstrap weights w_i with the probability p_i;
for *m=0 to M* **do**
  Use weak learning method to fit the proposal boxes probability $p_m(x_i)$, according to $p_m(x_i)$, calculate weak classifiers $f_m(x)$ and update
  $w_i=w_i\exp(-y_i f_m(x_i)),\ i=1...M.$;
end
Output $f_m(x), m=0...M$.

Until now, all improved weak classifiers have already been obtained, next, these improved weak classifiers should be concatenated one by one to form a strong boosted forest classifier. At the test stage, in order to reserve image resolution, we first use selective search method to generate candidate regions, and the corresponding confidence scores of these regions are obtained by the Intersection-over-Union (IoU) between them and ground truth boxes. Then, ranking these proposal regions based on their respective confidence scores, we believe that possible pedestrian parts will occur in areas with high confidence scores. Finally, choosing ranks the top 300 proposal regions and using the improved strong

boosted forest classifier to classify candidate regions for predicting pedestrian.

4 Experiments

Following most of recent work on pedestrian detection, we evaluate the performance of our method on the Caltech [9], ETH [13] and KITTI [14] Pedestrian Benchmark. In order to augment the number of training samples, we use each frame in the video as a training sample, thus contains about 51k pedestrian bounding boxes as positive training patches. Selective search approach is utilized to produce pedestrian candidate boxes with proposal scores in both training and testing stage, and we consider pedestrian candidate boxes with proposal score rank in the top 30% as the input of the convolution neural network.

According to the experimental evaluation criteria proposed in [9], measuring the log-average miss rate over nine points False-Positive-Per-Image (FPPI) evenly spaced in log-space ranging from 10-2 to 100. In the experiments, the miss rate against False-Positive-Per-Image curves are plotted to evaluate the overall performance of our method in the reasonable subsets (pedestrians that at least 50 pixels tall, under no occlusion or partial occlusion), and compare our method with the state-of-the-art approaches, which are reported by the Caltech benchmark using the evaluation code provided in [9].

4.1 Comparison on Architectures of Our Method

In order to more accurately detect low resolution pedestrian from dynamic scenes or detect pedestrian from hard negative background, we adopt selective search method to produce pedestrian candidate boxes with confidence scores at the testing stage, and extract ROI features in fixed aspect ratio proposal boxes instead of multi-scale proposal boxes.

Different classifier components are adopted to cascade our DCNN model, for further analyzing the effectiveness of different architectures of our method, which is demonstrated in Fig. 6, where the average miss rates show a slight decline trend when cascaded classifier is more suitable to classify low resolution pedestrian from hard negative background.

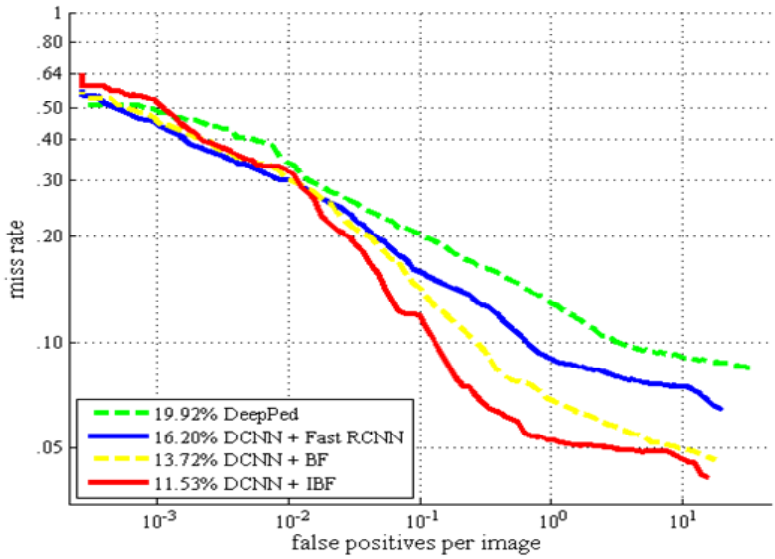


Fig. 6. The average miss rate of different architectures on Caltect-Test

For example, at first, the primordial DeepPed detector, also known as DCNN, achieves 19.92% miss rate on Caltech-Test, and then, the faster RCNN classifier cascaded on DCNN have achieved 16.20% miss rate, with 3.72% performance improvement, which is a small upgrade, partially because the low resolution features of ROI, afterwards, the faster RCNN is replaced with boosted forest classifier to cascaded DCNN model, and have reduced the miss rate by 2.48%, because the boosted forest classifier can effectively detect pedestrian from hard negative background and is more flexible to various resolution, at last, we fully consider the confidence score of the proposal bounding box and take advantage of proposed regions features to improve the boosted forest classifier for further decreasing the

average miss rate, and have finally reached miss rate of 11.53%, with 2.19% performance improvement. The details is shown in Table 2.

Table 2. Comparisons of different classifiers cascaded on DCNN model in the Caltech benchmark

Methods	Miss Rate (%)	Improvement (%)
DeepPed (DCNN)	19.92	—
DCNN+Fast RCNN	16.20	3.72
DCNN+BF	13.72	2.48
DCNN+IBF	**11.53**	2.19

4.2 Overall Evaluation on Caltech

To evaluation our model on Caltech dataset, we employ the Caltech-Trian for training and the Caltech-Test for testing. Approximately 4000 positive samples and 60000 negative samples are used for training from the Caltech-Train dataset, the performance of our DCNN cascaded IBF method is compared with the existing best performing approaches, including VJ [34], HOG [6], Crosstalk [8], MT-DPM [35], ACF-Caltech+ [17], SpatialPooling [36], LDCF [17], katamari [21], RCNN [27], TA-CNN [3], DeepPed [5], CompACT-Deep [23]. Those methods have used multiple features, includes Haar-like (VJ), HOG (HOG, MT-DPM), Motion (katamari) and Channel feature (ACF-Caltech+, LDCF); Various classifiers, such as cascade classifiers (Crosstalk, CompACT-Deep) and boosting classifiers (ACF-Caltech+, SpatialPooling) as well as different deep models (e.g. RCNN, TA-CNN, DeepPed and CompACT-Deep). Both DeepPed and CompACT-Deep have combined object region proposal with CNN model, our approach is largely similiar with that. In which, DeepPed have log-average miss rate of 19.92%, and CompACT-Deep have reached the current lowest log-average miss rate of 11.70%.

Fig. 7 shows the overall experimental results on the Caltech reasonable subset. Compared to those existing approaches, our DCNN cascaded IBF method have achieved the lowest log-average miss rate of 11.53%, which have reduced 8.39% and 0.17% the average miss rate compared with DeepPed and CompACT-Deep respectively. Although it only outperforms the second best method (CompACT-Deep [23]) by 0.17 percent, it retains image resolution and trains high level classifier in bootstrapping scheme, instead of simple optimizing classifier to reduce classification risk.

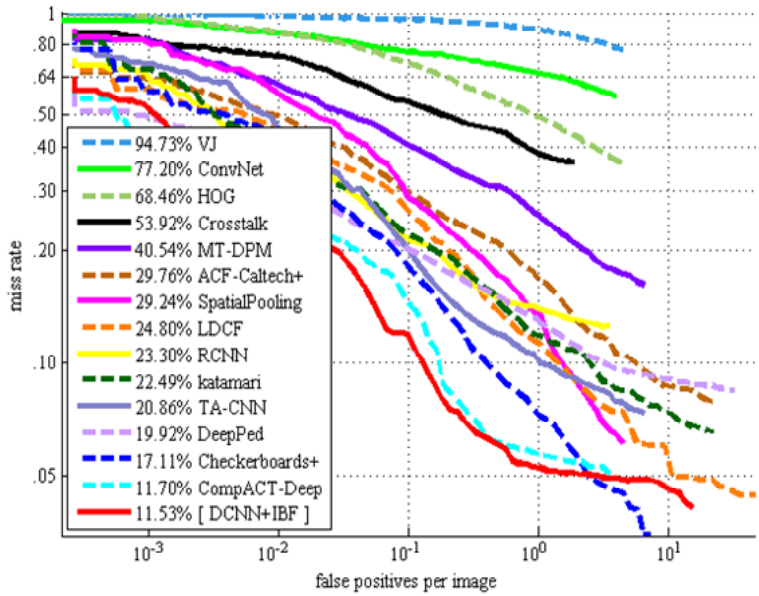
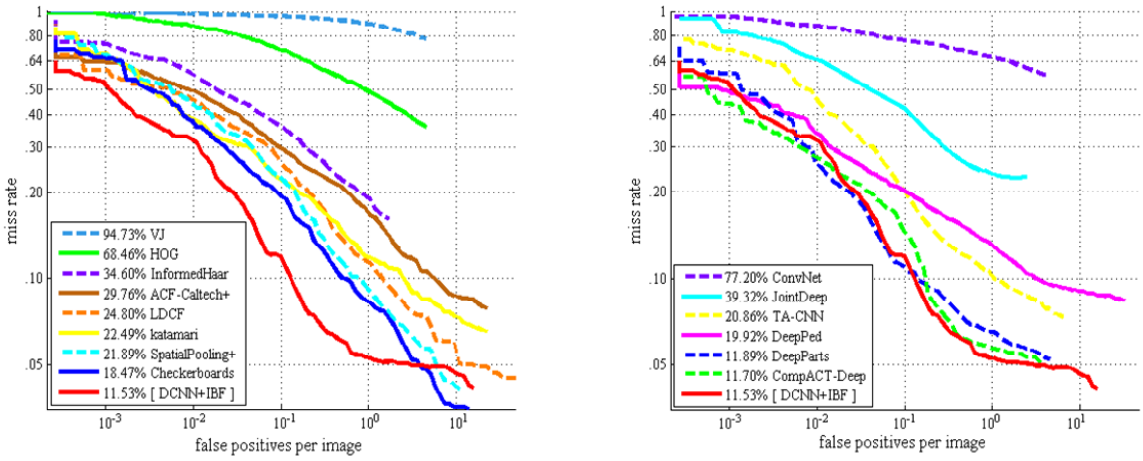


Fig. 7. Overall comparison average miss rate on Caltech reasonable subset

Our approach outperforms those traditional hand-crafted features to a large extent, such as Haar, HOG, channel, motion and context feature, due to our method have high discriminative ROI features learning, the performance comparison between DCNN cascaded IBF approach and those models based on hand-crafted features is shown in Fig. 8(a). Note that our model have achieved the log-average miss rate of

11.53%, with 10.96% and 10.36% performance improvement compared to katamari and SpatialPooling+. While both katamari method and SpatialPooling+ method have combined context information with multiple features, such as HOG, covariance, channel and motion feature.

Fig. 8(b) shows that the comparison results of our model with other deep models and indicates the superiority of our method. In order to efficiently detect pedestrian from complex background, we use DCNN model to learn high level feature representation and cascade an improved boosted forest to classify those region proposal boxes. Our method has 27.79% and 9.33% performance improvement compared to those methods that view pedestrian detection as a binary classification task, such as JointDeep, TA-CNN.



(a) Comparison with models based on hand-crafted feature (b) Comparison with other deep models

Fig. 8. Results on Caltech reasonable subset

4.3 Overall Evaluation on ETH and KITTI

In order to evaluate the generalization capacity of our DCNN cascaded IBF model, we compare the method with the state-of-the-art on ETH [13] and KITTI [14] pedestrian dataset.

Most best-performing approaches are both trained and tested on the INRIA training dataset, which is the common training set for existing methods on ETH pedestrian dataset, thus, we train our model on this dataset also, which includes approximately 2000 positive samples and 60000 negative samples.

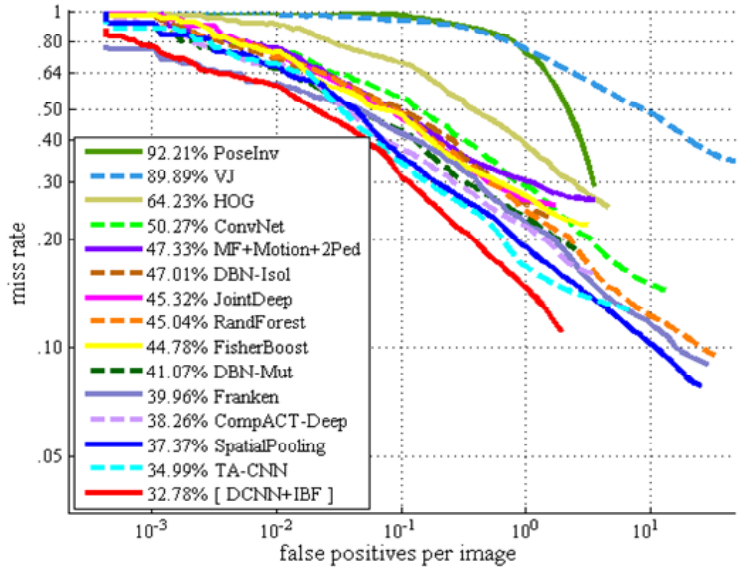
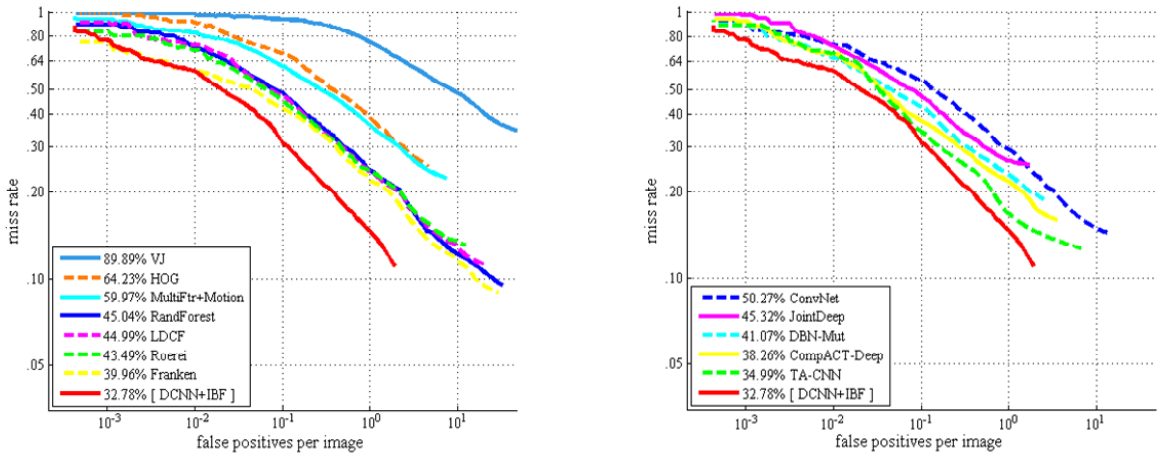


Fig. 9. Overall comparison average miss rate on ETH dataset

Comparing the performance of our method with the existing best performing approaches, Fig. 9 shows the overall experimental results on the ETH, the lowest log-average miss rate of the existing method is 34.99% by [3], while ours decreases to 32.78%, which has 2.21% improvement in detection performance, and if we consider the miss rate at 1 FPPI, the best is 17.4% by [3], while ours further decreases to 15.2%. We can see that the cascaded classifier can better detect the pedestrians and yet greatly reduces the false detection.

The performance comparison of our method and those models based on hand-crafted features is shown in Fig. 10(a). Note that our method have achieved the log-average miss rate of 32.78%, and outperforms the existing three best-performing hand-crafted models, SpatialPooling, Franken and Roerei, by 4.59%, 7.18% and 10.71%, respectively. And Fig. 10(b) shows that the comparison results of our model with other deep models, and shows us the truth that our method outperforms the two best-performing deep models, TA-CNN and CompACT-Deep, by 2.21% and 5.48%, respectively.



(a) Comparison with models based on hand-crafted feature (b) Comparison with other deep models

Fig. 10. Results on ETH dataset

In addition, our method use selective search strategy to propose candidate bounding boxes, which can maximize to reserve image original resolution, and extract the not hand-crafted ROI features. In decreasing miss rate and improving accuracy on the ETH dataset, our approach presents that effective image resolution retention and the classifier bootstrap are more important than extract hand-crafted features.

Our model is trained by using train data of Caltech-train on KITTI dataset. Fig. 11 shows the mean average precision on KITTI moderate subset and Table 3 summarized the performance of our DCNN cascaded IBF method with state-of-the-art on KITTI easy, moderate and hard subsets respectively. Since test images in KITTI are larger than in Caltech, running times are higher on this pedestrian benchmark, nevertheless except CompACT, the DCNN cascade improved boosted forest classifier method is the fastest of all the state-of-the-art detectors. Note that the proposed methods is much more accurate and faster than the conventional machine learning method, include DPM [7], FilteredICF [18] and pAUCEnstT [36-37] and the like. Compared with other deep models, we found that the proposed model superior to RCNN [29], DeepParts [24] and CompACT-Deep [23] in accuracy and speed, in addition, although our method little slower than compACT [23] about 0.05 second, the mean average precision of ours larger than its by 8.18%, 4.24% and 4.59% in easy, moderate and hard subsets respectively, which is a big gap. The only competitive detector is Regionlets [30], which should thank for additional supervision, and outperforms our model by 1.99% and 1.39% in moderate and hard subsets respectively. However, the detector has ignored the regional proposal time, which about two seconds and remains less slower than ours.

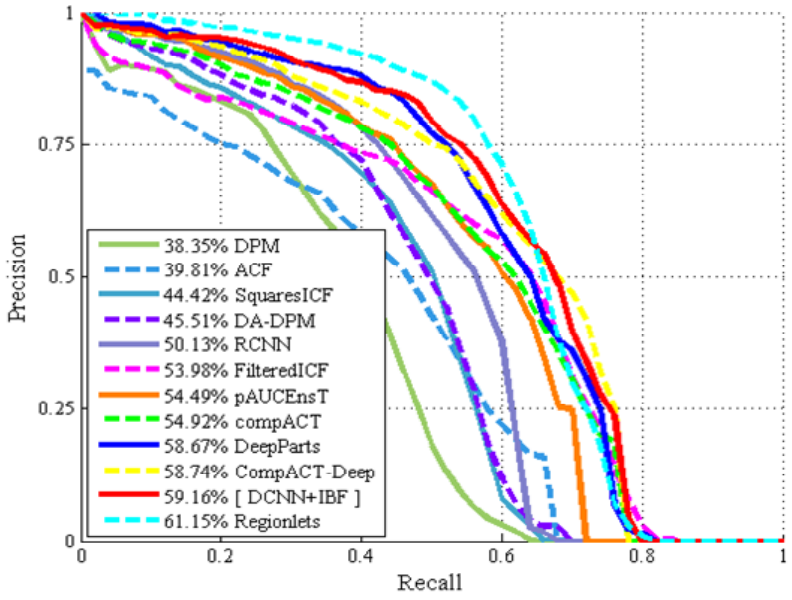


Fig. 11. Comparison mean average precision on KITTI moderate subset

Table 3. Comparisons mAP with state-of-the-art detectors on the KITTI easy, moderate and hard datasets

Methods	Easy (%)	Moderate (%)	Hard (%)	Times (s)
DPM	45.50	38.35	34.78	10
ACF	48.24	39.81	34.92	40
squaresICF	52.37	44.42	39.88	40
DA-DPM	56.36	45.51	41.08	21
RCNN	61.61	50.13	44.79	4
FilteredICF	61.14	53.98	49.29	40
pAUCEnstT	65.26	54.49	48.60	60
compACT	65.35	54.92	49.23	**0.75**
DeepParts	70.49	58.67	52.78	1
CompACT-Deep	70.69	58.74	52.71	1
Regionlets	73.14	**61.15**	**55.21**	1*
DCNN+IBF[ours]	**73.53**	59.16	53.82	0.80

Note. * denotes without take regional proposal time into account.

5 Conclusion and Future Work

In this paper, we propose a novel deep convolution neural network cascaded improved boosted forest classifier method for pedestrian detection. First of all, selective search method is employed to proposal pedestrian candidate bounding boxes with confidence scores, then, adopting convolution neural network to extract ROI features, and finally, by cascading improved boosted forest classifier to optimize the classification performance for pedestrian detection. Experiments on Caltech Pedestrian Benchmark, ETH and KITTI dataset demonstrate the proposed method outperforms the state-of-the-art.

Due to many existing works using context information [35, 38] have achieved an excellent pedestrian detection performance. Therefore, in our future works, we firstly may consider the ground plane constraint and video spatial-temporal information as context cues to further improve pedestrian detection performance from video. Secondly, high computational complexity remains a problem of our method, we tends to optimize the algorithm for speeding up detection, even achieving real-time detection.

Acknowledgements

The authors would like to thank their laboratory team member's assistance. This work was supported by the Talent project of Huaqiao University (Grant No. 14BS215), Quanzhou scientific and technological planning projects of Fujian, China (Grant No. 2015Z120).

References

[1] P. Dollar, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Transactions on Pattern Analysis & Machine Intelligence 36(8)(2014) 1532-1545.

[2] J.J. Zhu, O. Javed, J.G. Liu, Q. Yu, H. Cheng, H. Sawhney, Pedestrian detection in low-resolution imagery by learning Multi-scale Intrinsic Motion Structures (MIMS), in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[3] Y.L. Tian, P. Luo, X.G. Wang, X.O. Tang, Pedestrian detection aided by deep learning semantic tasks, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[4] W.L. Ouyang, X.Y. Zeng, X.G. Wang, Learning Mutual Visibility Relationship for Pedestrian Detection with a Deep Model, Kluwer Academic Publishers Hingham, MA, 2016.

[5] D. Tomè, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, S. Tubaro, Deep convolutional neural networks for pedestrian detection, Signal Processing Image Communication 47(C)(2016) 482-489.

[6] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2005.

[7] P.F. Felzenszwalb, R.B. Girshick, D. Mcallester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis & Machine Intelligence 32(9)(2010) 1627-1645.

[8] P. Dollár, R. Appel, W. Kienzle, Crosstalk cascades for frame-rate pedestrian detection, in: Proc. 12th European Conference on Computer Vision, Springer-Verlag Berlin, Heidelberg, 2012.

[9] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Transactions on Pattern Analysis & Machine Intelligence 34(4)(2012) 743-761.

[10] California Institute of Technology Pedestrian Detection Benchmark, Caltech pedestrian detection benchmark. <http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/>, 2016.

[11] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313(5786)(2006) 504-507.

[12] L. Zhang, L. Lin, X. Liang, K. He, Is faster R-CNN doing well for pedestrian detection?, in: Proc. European Conference on Computer Vision, Springer International Publishing, 2016.

[13] A. Ess, B. Leibe, K. Schindler, L.V. Gool, Robust multiperson tracking from a mobile platform, IEEE Transactions on Pattern Analysis & Machine Intelligence 31(10)(2009) 1831-1846.

[14] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving?, The KITTI vision benchmark suite, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2012.

[15] X.G. Wang, M. Wang, W. Li, Scene-specific pedestrian detection for static video surveillance, IEEE Transactions on Pattern Analysis & Machine Intelligence 36(2)(2014) 361-374.

[16] P. Dollár, Z.W. Tu, P. Perona, S. Belongie, Integral channel features, in: Proc. British Machine Vision Conference, 2009.

[17] W. Nam, P. Dollár, J.H. Han, Local decorrelation for improved pedestrian detection, in: Proc. 27th International Conference on Neural Information Processing Systems, 2014.

[18] S.S. Zhang, R. Benenson, B. Schiele, Filtered channel features for pedestrian detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[19] M. Heikkilä, M. Pietikäinen, C. Schmid, Description of interest regions with local binary patterns, Pattern Recognition 42(3)(2009) 425-436.

[20] S.S. Zhang, C. Bauckhage, A.B. Cremers, Informed Haar-Like features improve pedestrian detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2014.

[21] R. Benenson, M. Omran, J. Hosang, B. Schiele, Ten years of pedestrian detection, what have we learned?, in: Proc. European Conference on Computer Vision, CVRSUAD workshop, 2014.

[22] W.L. Ouyang, X.G. Wang, Joint deep learning for pedestrian detection, in: Proc. IEEE International Conference on Computer Vision, 2013.

[23] Z.W. Cai, M. Saberian, N. Vasconcelos, Learning complexity-aware cascades for deep pedestrian detection, in: Proc. IEEE International Conference on Computer Vision, 2015.

[24] Y.L. Tian, P. Luo, X.G. Wang, X.O. Tang, Deep learning strong parts for pedestrian detection, in: Proc. IEEE International Conference on Computer Vision, 2015.

[25] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image tecognition. <https://arxiv.org/abs/1409.1556>, (2014).

[26] X.G. Wang, A discriminative deep model for pedestrian detection with occlusion handling, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[27] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[28] J.R. Uijlings, K.E. Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, International Journal of Computer Vision 104(2)(2013) 154-171.

[29] J. Hosang, M. Omran, R. Benenson, B. Schiele, Taking a deeper look at pedestrians, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[30] X.Y. Wang, M. Yang, S.H. Zhu, Y.Q. Lin, Regionlets for generic object detection, in: Proc. IEEE International Conference on Computer Vision, IEEE Computer Society, 2013.

[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proc. International Conference on Neural Information Processing Systems, Curran Associates Inc, 2012.

[32] R. Girshick, Fast R-CNN. <https://arxiv.org/abs/1504.08083>, 2015.

[33] D.H. Tang, Y. Liu, T.K. Kim, Fast pedestrian detection by cascaded random forest with dominant orientation templates, in: Proc. British Machine Vision Conference, 2012.

[34] P. Viola, M.J. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance, in: Proc. 9th IEEE International Conference of Computer Vision, 2003.

[35] J.J. Yan, X.C. Zhang, Z. Lei, S.C. Liao, S.Z. Li, Robust multi-resolution pedestrian detection in traffic scenes, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2013.

[36] S. Paisitkriangkrai, C.H. Shen, A.V.D. Hengel, Strengthening the effectiveness of pedestrian detection with spatially pooled features, Lecture Notes in Computer Science 8692(2014) 546-561.

[37] S. Paisitkriangkrai, C.H. Shen, A.V.D. Hengel, Efficient pedestrian detection by directly optimizing the partial area under the ROC curve, in: Proc. IEEE International Conference on Computer Vision, 2013.

[38] Y.Y. Ding, J. Xiao, Contextual boost for pedestrian detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012.