Pedestrian Detection with Deep Convolutional Neural Network

Xiaogang Chen, Pengxu Wei, Wei Ke, Qixiang Ye, Jianbin Jiao

School of Electronic,Electrical and Communication Engineering, University of Chinese Academy of Science, Beijing, China

Abstract. The problem of pedestrian detection in image and video frames has been extensively investigated in the past decade. However, the low performance in complex scenes shows that it remains an open problem. In this paper, we propose to cascade simple Aggregated Channel Features (ACF) and rich Deep Convolutional Neural Network (DCNN) features for efficient and effective pedestrian detection in complex scenes. The ACF based detector is used to generate candidate pedestrian windows and the rich DCNN features are used for fine classification. Experiments show that the proposed approach achieved leading performance in the INRIA dataset and comparable performance to the state-of-the-art in the Caltech and ETH datasets.

1 Introduction

Pedestrian detection has been one of the most extensively studied problems in the past decade. One reason is that pedestrians are the most important objects in natural scenes, and detecting pedestrians could benefit numerous applications including video surveillance and advanced driving assistant systems. The other reason is that pedestrian detection has been the touchstone of various computer vision and pattern recognition methods. The improvement of pedestrian detection performance in complex scenes often indicates the advance of relevant methods.

Two representative works in pedestrian detection are the VJ [1] detector and HOG [2] detector. The VJ detector employed the framework of using simple Haar-like features and cascade of boosted classifiers, achieved a very fast detection speed. This framework is further developed by Dollár *et al.* who proposed the Integral Channel Features [3], including multiple types of features(grayscale, LUV color channels, gradient magnitude, etc.) that can be quickly computed using integral images. The Integral Channel Features is simple but effective, and widely used in many state-of-the-art pedestrian detectors [4–8].On the other hand, the success of HOG detector encouraged the usage of complex features, like Local Binary Pattern [9], Dense SIFT [10] and Covariance Descriptor [11], etc.. Also, based on HOG, Felzenszwalb *et al.* proposed the Deformable Part Based Model (DPM) [12] which made a breakthrough in pedestrian detection.

Since the feature extraction pipelines in above methods are designed manually, they can be categorized as hand-craft features. In recent researches, with

the steady advance of deep learning [13] and unsupervised feature learning [14], learnable features gain significant attentions. Specially, the Deep Convolutional Neural Network (DCNN) proposed by Krizhevsky *et al.* [15] achieved record-breaking results in ImageNet Large Scale Visual Recognition Challenge 2012. Afterwards, its specific network structure has been widely used in image classification and object detection [16–19]. In [16], Donahue *et al.* showed that features generated from a classifying CNN perform excellently in related vision tasks, implying that DCNN can be used as a generic feature extractor.

In the field of pedestrian detection, many feature learning and deep learning methods have been introduced recently. In [20], Sermanet *et al.* proposed a two layers convolutional model and layers were pre-trained by convolutional sparse coding. In [21], Ouyang *et al.* conducted Restricted Boltzmann Machine (RBM) in modeling mutual visibility relationship for occlusion handling. And in [22] authors further cooperated with Convolutional Neural Network, and proposed a joint deep learning framework that jointly consider four key components in pedestrian detection: feature extraction, deformation model, occlusion model and classifier.

In [20], Convolutional Neural Network has been successfully applied in pedestrian detection, where the used network structure have only 2 layers. In contrast, Krizhevsky's CNN [15] that has 7 layers is much deeper. In this paper, we try to make it clear that whether the usage of a larger and deeper Convolutional Neural Network for feature extraction can further improve the performance of pedestrian detection or not. When using large CNN for feature extraction, the commonly used "sliding-window" detection paradigm is hard to work for the computational efficiency problem [17]. To improve detection efficiency, pre-localization approaches such as the selective search method [23] has been used to generate proposal regions, which is a "recognition-using-regions" paradigm [24]. In pedestrian detection task, however, it is observed in experiments that the "recognition-using-regions" paradigm based on "Selective Search" is infeasible. The reason is that pedestrian detection requires precise localization before it can obtain a good detection performance, but the selective search method cannot provide precise localization of pedestrians.

In this paper, we propose to cascade simple Aggregated Channel Features (ACF) and rich Deep Convolutional Neural Network (DCNN) features for efficient and effective pedestrian detection in complex scenes. To generate precisely localized candidate pedestrian windows, we employ a cascade of Adaboost classifiers on Aggregated Channel Features (ACF detector) [5]. We reduce the stage number of the original ACF detector to two so that most of pedestrian windows could be kept for fine classification. We propose to use the DCNN pre-trained on a large image set so that the network parameters are well learned. On the candidate pedestrian windows, DCNN features and a linear SVM classifier are used to perform pedestrian classification and fine detection. Our proposed approach is also an attempt to combine hand-craft features with learned features, which is seldom investigated in existing works. The flowchart of the proposed pedestrian detection approach is as Fig1.

The rest of this paper is organized as follows: in section 2 we introduce our pedestrian detection approach. In section 3, we report the experiment results on three benchmark datasets, INRIA, Caltech, ETH. In section 4, we conclude our works and have a discussion.

2 Pedestrian Detection Approach

The proposed pedestrian detection approach can be regarded as a two-stages system. In the first stage, we conduct simple channel features and boosted classifiers, in order to rapidly filter out as many negative windows as possible, while keeping all the positive windows. Then in next, the fine detection stage, we use DCNN to extract features from these windows, and SVM for classification. The flowchart of proposed approach is draw in Fig.1

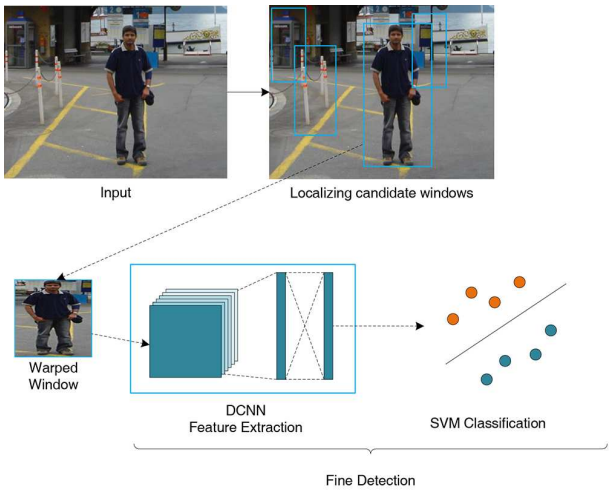


Fig. 1. Flowchart of the proposed pedestrian approach. A fast candidate window localization method is first applied to input image, then all candidate windows are warped to require size of the fine detection stage, where they will be classified. For details, see the following sections.

2.1 Localizing Candidate Windows

In this stage, we employ the Aggregate Channel Features detector [5] for localizing candidate windows. Following the "channel features + boosted classifier" schema, three types of channel features are used: normalized gradient magnitude, histogram of oriented gradients and LUV color channels.

These channels are first generated from input images, then summed up in 4×4 pixel grid and smoothed, yields a 5120 dimensions feature pool. Next, a

Table 1. Comparison of region proposal methods.

	Selective Search	Objectness	ACF	Selective Search on Caltech Dataset
Cover rate	97.62%	93.55%	**98.13%**	1.9%
Runtime	∼4s	∼4s	**<0.5s**	∼4s

bootstrapping iteration is conducted over the feature pool to construct a cascade of classifiers. Here we built a cascade of 2 stages, that combined 32 and 128 classifiers in each.

Since our purpose is to quickly filter out as many negative windows as possible, while keeping all positive windows, it is important to measure performance with ground truth cover rate and runtime. The ground truth cover rate measures how many positive windows can be detected by proposal windows. A ground truth window is consider detected by a proposal given that their area of overlap exceed 50% [25]. In Table 1, we compare the proposed method with Selective Search [23] and Objectness [26] on INRIA pedestiran dataset, using the measures indicated above.

It can be seen that although Selective Search achieves comparable cover rate on INRIA dataset, the processing time is much slower than ACF. Notice that Selective Search is a segmentation based method that performance is strongly affected by image quality. So we further conduct Selective Search on Caltech dataset, where the cover rate dramatically down to 1.9 percents because the image quality of Caltech dataset is much worse than INRIA. The Objectness method did not show advantages in either cover rate or runtime.

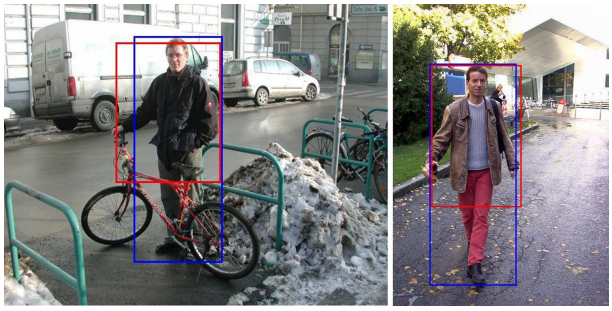


Fig. 2. Fail cases of using Selective Search. The candidate windows (red rectangles) that mostly overlapping with ground truth (blue rectangles) only cover half of the pedestrian.

On the other hand, pedestrian detection requires the bounding box tightly surround pedestrians, where the general region proposal methods might fail

because most of them are designed to capture object in any aspect ratio, ignoring the fact that pedestrians are more like rigid object. This inaccuracy will affect applications built on pedestrian detection results. As in Fig.2, the blue box indicates ground truth annotation, and the red box is the candidate window overlapping mostly with ground truth. Due to the region grouping setting, Selective Search methods generate candidate windows that cover only half of the pedestrian body. Since ACF detector is designed for pedestrian detection, it won't suffer from this problem.

2.2 Fine Detection

In the fine detection stage, to classify the candidate windows passed the previous stage, DCNN is employed. Following the network architecture proposed by Krizhevsky *et al.* [15], we used the RCNN package [17] which utilize the Caffe [27] to implement DCNN.

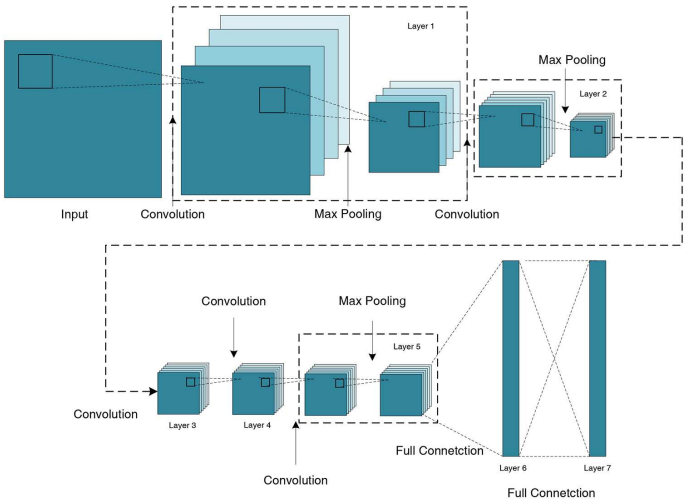


Fig. 3. Architecture of DCNN

The architecture of used DCNN is presented in Fig.3, which has 7 layers. Notice that the DCNN requires input images of 227×227 pixels size, so first we simply warp candidate windows to the required size. Notice that the warping causes a distortion of images which will affect the information carried within, however it is observed from experiments that warping works well. In the first layer, the warped images are filtered with 96 kernels of size 11×11×3 pixels with a stride of 4 pixel, then max-pooling is applied in 3×3 grid. The second layer has the same pipeline as first layer, with 256 kernels of size 5×5×48, and max-pooling in 3×3 grid. Afterwards, there are two convolution layers without pooling, which

both contains 384 kernels. In the fifth layer, again, the output of previous layer is first convoluted with 256 kernels then applied spatial max-pooling in 3×3 pixel grid. The last two layers of the network are fully connected layer, which both contains 4096 nodes respectively. The DCNN eventually output features of 4096 dimensions from the last layer. The activation function used in the convolution and full connected layer is Rectified Linear function $f(x) = max(0, x)$. For more details about network parameters and training protocol, we refer reader to [15].

After obtaining the training features, we train a linear SVM for classification. As common practice [12], a bootstrapping process is conducted to improve classification. We mine hard negative samples from training dataset and retrain SVM with it. It is worth nothing that, the bootstrapping converges quickly in single iteration, compare with DPM [12] that runs in multiple iterations, indicating the capacity of DCNN in modeling complex images.

3 Experiments

In this section, we evaluate the proposed pedestrian detection approach on three well-known benchmarks: INRIA, Caltech and ETH datasets. Before getting into specific experiments, there are some issues in using DCNN, which are the usage of pre-trained model and the feature layers. In first subsection, we discuss these problems and show the comparison experiment results. The results on INRIA, Caltech and ETH are presented in second and third subsections. All evaluations follow the protocols proposed by Dollár *et al.* [28].

3.1 Model Setting

As common practice, we train the detector with INRIA dataset. However, it is obviously insufficient to train the DCNN model with INRIA dataset, since the model contains millions of parameters, that will easily leads to over-fitting. In [16], Donahue *et al.* generated features from a network pre-trained with the ImageNet dataset, and successfully apply them to other vision tasks, shows generalization capacity of DCNN. Here we follow the same strategy that use a pre-trained CNN as a blackbox feature extractor. We use the pre-trained models provided in the RCNN package, which were trained on the PASCAL VOC 2007, 2012 and ImageNet dataset, respectively.

On the other hand, another issue concerned is the usage of feature layers, recent DCNN based research [17] suggested that the FC6 (Short for Fully Connected layer 6) features usually outperforms the FC7 and POOL5 (Short for Pooling layer 5) layer features. To achieve comprehensive understanding, in our experiments, we compare the performance of different model-layer combinations. The result are presented in Fig.4.

It can be seen that the DCNN pre-trained from PASCAL VOC 2007 dataset generally outperforms models trained from VOC 2012 and ImageNet datasets. The model pre-trained in ImageNet dataset perform unexpectedly poor, considering that both PASCAL VOC and ImageNet datasets contain the category of

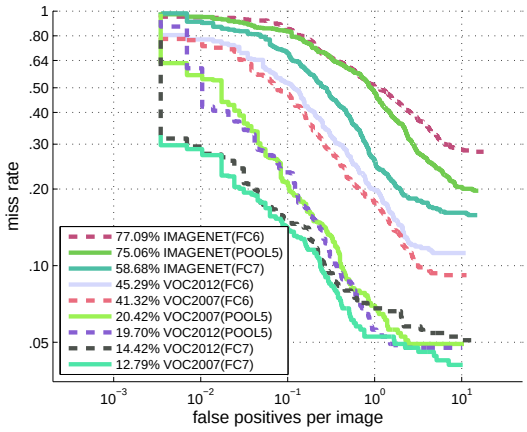


Fig. 4. Comparison of different model settings on INRIA dataset

person, which is no exactly pedestrian but share most characteristics with it, so all models are expected to have similar performances. We consider the performance gap might be that the PASCAL VOC dataset contain less categories of objects than ImageNet dataset, which is 20 to 200, then the DCNN model trained from VOC can have more parameters to characterize persons that leads to better performance.

For feature layer selection, with different pre-trained model, the FC7 features performed best. These experiments gave the hints for constructing better DCNN model for pedestrian detection, so in this work we choose to use the pre-trained model from VOC 2007 and FC7 layer feature for further study.

3.2 INRIA Dataset

We first evaluate the proposed approach in INRIA dataset. In this experiment, we mirror the positive samples for augmentation, then generate features from the DCNN for SVM training, and run single iteration of bootstrapping.

Evaluation is based on the fixed INRIA annotations provided by [20], which include additional "ignore" labels for pedestrians miss labelled in original annotations [2]. To compare with major state-of-the-art pedestrian detectors, the log-average miss rate is used and is computed by averaging the area under curve (AUC) from 9 discrete false positive per image (FPPI) rates [25]. We plot all comparing DET curves (miss rate versus FPPI) in Figure 5. The abbreviation DCNN short for Deep Convolutional Neural Network indicates our results.

The proposed system outperforms most state-of-the art pedestrian detectors with an average 12.79% miss rate. We can see that about 30% performance improvement is obtained compare with ConvNet [20], the substantial gain proved that the larger and deeper Convolutional Neural Network indeed improve pedestrian detection. In the other hand, 21% improvements is gained from ACF detector [5], which is used for region proposal in our approach. Notice that original

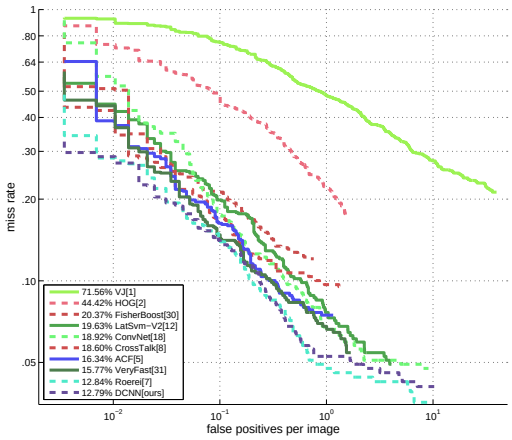
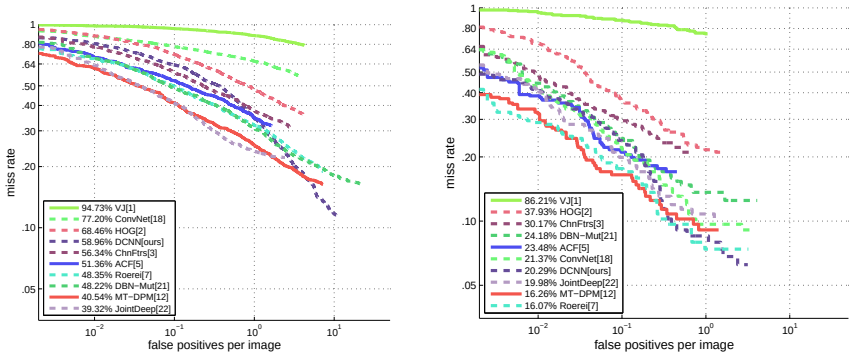


Fig. 5. Comparison of different methods on INRIA dataset. DCNN refers to our proposed detection approach.

ACF detector, 4 stages of boost classifiers are used while 2 stages in our approach, indicating that the DCNN outperforms the higher stages boost classifiers.

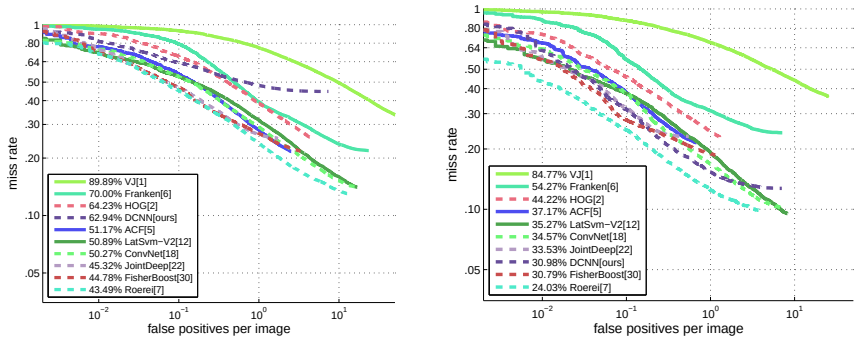
3.3 Caltech and ETH Datasets

In this section, we train the proposed system with INRIA dataset, then apply it to Caltech and ETH dataset. We do not bootstrap our system on these datasets, in order to discover if the features extracted from DCNN can generate to other datasets. Comparison results are plot in Fig.6 and Fig.7 respectively.



(a) (b)

Fig. 6. Experimental results on Caltech Dataset. (a) Results on '*Reasonable*' pedestrians. (b) Results on '*Large*' pedestrians



(a) (b)

Fig. 7. Experimental results on ETH Dataset. (a) Result on '*Reasonable*' pedestrians. (b) Result on '*Large*' pedestrians

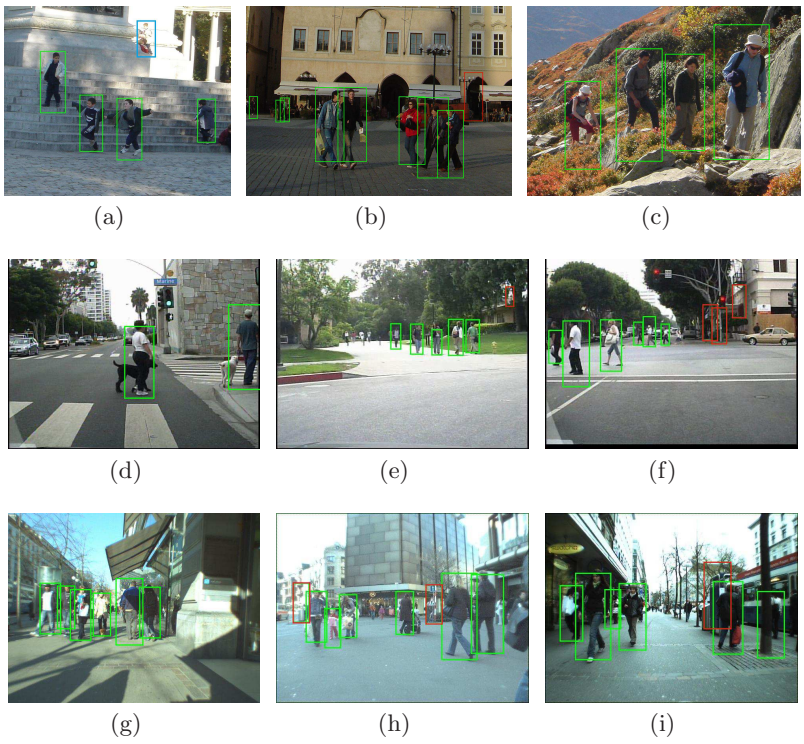
We can observe that, although the proposed system achieve excellent performance on the INRIA dataset, results on Caltech and ETH datasets are less impressive. The proposed approach performs poorly in the "Reasonable" subset, however, in the "Large" subset it is better than ConvNet as in the INRIA experiments. Notice that the performance of DCNN is comparable with another CNN based method JointDeep [22] which employed deformation and occlusion handling pipelines in their method.

We conclude the decrease of performance might have several reasons. First is that we do not conduct any specific fine tuning on both dataset which might affect the performance. Another reason might relate to image quality, the INRIA dataset has much better image quality than the other datasets, the resolution is higher and images are more distinct. Due to the large multi-layer network structure, DCNN is good at capture detail characteristics. The performance gap between small and large subsets is shared with [20], which points out a very interest problem for future study, how to obtain good detection performance on low resolution imagery with DCNN features.

3.4 Detection Examples

We show some detection examples in Fig.8. In INRIA dataset,Fig.8(a)-Fig.8(c), most pedestrians are correctly located with few false positives. In Fig.8(a), there is a missed positive (marked with blue rectangle), because of the strong sunlight, the little boy is hard to distinguish from background.

In Caltech and ETH datasets,Fig.8(d)-Fig.8(i), more false positives are observed. Most false positives appeared in clustered backgrounds, where trees, trash cans and billboards are prone to be recognized as pedestrians. As we analyzed in previous section, a specific hard samples mining procedure will help to reduce these types of false alarms. Also, conducting deformation and occlusion handling pipeline in pedestrian detection will boost the performance in crowd scenes.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 8. Detection Examples

4 Conclusions

We proposed a state-of-the-art pedestrian detection method, which combines the successful Aggregated Channel Features detector and Deep Convolutional Neural Network. An ACF detector is used to generate candidate pedestrian windows, and a DCNN based detector is used to extract features for classification. Benefitting from the large network structure, the proposed method gains substantial improvement over previous CNN based methods, and achieves leading performance in INRIA dataset and comparable performance in Caltech and ETH datasets.

The proposed method does not conduct fine tuning on the experiment datasets and does not include a specified pipeline for occlusion handling, leaving room for further improvements. In addition, improving the performance of DCNN in low resolution images is worth working.

Acknowledgement. This work was supported in Part by National Basic Research Program of China (973 Program) with Nos. 2011CB706900, 2010CB731800, and National Science Foundation of China with Nos. 61039003, 61271433 and 61202323.

References

1. Viola, P.,Jones, M.J.: Robust real-time face detection. International journal of computer vision **57** (2004) 137–154
2. Dalal, N., Triggs, B. : Histograms of oriented gradients for human detection. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2005) 886–893
3. Dollár, P., T.Zhuowen, Perona, P., Belongie, S.: Integral Channel Features. British Machine Vision Conference **2** (2009)
4. Dollár, P., Belongie, S., Perona, P.: The Fastest Pedestrian Detector in the West. British Machine Vision Conference **2** (2010)
5. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Trans. on Pattern Analysis and Machine Intelligence **36** (2014) 1532 - 1545
6. Mathias, M., Benenson, R., Timofte, R., Gool, L.V.: Handling occlusions with franken-classifiers In: Proc. of IEEE International Conference on Computer Vision (2013) 1505–1512
7. Benenson, R., Mathias, M., Tuytelaars, T., Gool, L.V.: Seeking the strongest rigid detector. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2013) 3666–3673
8. Dollár, P., Appel, R., Kienzle, W.: Crosstalk cascades for frame-rate pedestrian detection. IEEE European Conference on Computer Vision (2012) 645–659
9. W.Xiaoyu, Han, T., Y.Shuicheng: An HOG-LBP human detector with partial occlusion handling. In: Proc. of IEEE International Conference on Computer Vision (2009) 32–39
10. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: Proc. of IEEE International Conference on Computer Vision (2009) 606–613
11. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008) 1713–1727
12. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010) 1627–1645
13. Bengio, Y.: Learning deep architectures for AI. Foundations and trends in Machine Learning **2** (2009) 1–127
14. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013) 1798–1828
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems (2012)
16. Donahue, J., J.Yangqing, Vinyals, O., Hoffman, J., Zh.Ning, Tzeng, E. Darrell, T.: ImageNet Classification with Deep Convolutional Neural Networks. International Conference on Machine Learning (2014)
17. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE conference on Computer Vision and Pattern Recognition (2014)
18. Sermanet, P., Eigen, D., Zh.Xiang, Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. International Conference on Learning Representations (2014)

19. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional neural networks. IEEE European Conference on Computer Vision (2014) 818-833
20. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2013) 3626–3633
21. OY.Wanli, Z.Xingyu, W.Xiaogang: Modeling mutual visibility relationship in pedestrian detection. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2013) 3222–3229
22. OY.Wanli, Xiaogang: Joint deep learning for pedestrian detection. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2013) 2056–2063
23. Van de Sande, K., Uijlings, J., Gevers, T., Smeulders, A.: Segmentation as selective search for object recognition. In: Proc. of IEEE International Conference on Computer Vision (2011) 1879–1886
24. G.Chunhui,Lim, J., Arbeláez, P., Malik, J.: Recognition using regions. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2009) 1030–1037
25. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state-of-the-art. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 743–761
26. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 2189–2202
27. J.Yangqing: Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding. http://caffe.berkeleyvision.org/ (2012)
28. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2009) 304–311
29. L.Ping, T.Yonglong, W.Xiaogang, T.Xiaoou: Switchable deep network for pedestrian detection. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2013) 899–906
30. S.Chunhua, W.Peng, Sakrapee P., Anton, v.d.H.: Training Effective Node Classifiers for Cascade Classification. International journal of computer vision **103** (2013) 326-347
31. Gool, L.V., Mathias, M., Timofte, R., Benenson, R.: Pedestrian detection at 100 Frames Per Second. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2012) 2903-2910