Neural Process Lett (2017) 45:1025–1037
DOI 10.1007/s11063-016-9561-7

Supervised Local High-Order Differential Channel Feature Learning for Pedestrian Detection

Jifeng Shen[1] · Xin Zuo[2] · Hui Liu[1] · Haoran Wang[3] · Wankou Yang[4] · Chengshan Qian[5]

Published online: 31 October 2016

Abstract In this paper, a novel supervised local high-order differential channel feature is proposed for fast pedestrian detection. This method is motivated by the recent successful use of filtering on the multiple channel maps, which can improve the performance. This method firstly compute the multiple channel maps for the input RGB image, and average pooling is acted on the channel maps in order to reduce the effect of noise and sample misalignment. Then, each of the pooled channel maps is convolved with our proposed local high-order filter bank, which can enhance the discriminative information in the feature space. Finally, due to the increasing memory consumption incurred by the higher dimension of resulting feature, we have proposed a local structure preserved supervised dimension reduction method which aims to keep the manifold structure of samples in the feature space. This method is formulated as a classical spectral graph embedding problem which can be solved by the LPP algorithms. Thorough experiments and comparative studies show that our method can achieve very competitive result compared with many state-of-art methods on the INRIA and Caltech datasets. Besides, our detector can run about 20 fps in 480 × 640 resolution images.

Keywords High-order differential feature · Pedestrian detection · Manifold learning

✉ Jifeng Shen
shenjifeng1980@hotmail.com

[1] School of Electronic and Informatics Engineering, Jiangsu University, Zhenjiang 212013, Jiangsu, China

[2] School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, Jiangsu, China

[3] College of Information Science and Engineering, Northeastern University, Shenyang 110819, Liaoning, China

[4] School of Automation, Southeast University, Nanjing 210096, Jiangsu, China

[5] School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing 210044, Jiangsu, China

1 Introduction

In recent years, object detection [1–3] has attracted much more attention from the worldwide computer vision researchers. Although object detection have gained much progress in the recent years, detecting pedestrians in arbitrary image is still an open problem to be addressed, due to many complicated factors, such as view variation, occlusion and illumination changes, etc. As a hot topic in the domain of computer vision, pedestrian detection can be applied to many utilitarian areas, such as autonomous-driven vehicles, video surveillance and human activity classification [4], etc.

The past decades witness a large amount of work [5,6] devoted to human detection, but in this paper, we will cover just a few closely related to our method. Most of the classical methods for pedestrian detection are generally treated as a rigid or half-rigid object detection problem with less deformation. So it can be formulated as a traditional pattern recognition problem which can be handled with carefully designed feature and the off-the-shelf classifiers. Massive efforts have been made to the feature extraction step from the raw RGB or grey images, which is a crucial step to build a successful object detector. Without loss of generality, the feature extraction methods can be categorized into two main groups: hand-crafted features [7–10] and learning based features [11,12]. The representative works of the former one are HOG [7], LBP [13], InformedHaar [14] and ICF [9], etc, they are proved to be very effective in many applications. The most popular learning based features are obtained by the recent typical deep learning methods [15–17], such as CNN [18] and DBN [19], etc. The hand-craft features are advantageous in desirable computational efficiency, and thus achieves state-of-the-art results in many applications. However, they're task-specific and suffering from subjective human experimence, which leads to their sub-optimal performance in various visual applications. The emerging features of deep features have achieved best performance in many computer vision applications, such as image classification [20], object localization [21] and speech recognition [22], etc. Although deep learning based applications grow fast in recent few years, the intrinsic mechanisms are insufficiently stated. Furthermore, a high-end computer workstation with high capacity GPUs is indispensable for training a robust model.

In this paper, we will focus on the light-weighted discriminative feature which can be used to build efficient pedestrian detection system. a novel high-order differential channel feature is proposed for fast pedestrian detection task. Motivated by the filtered channel features [14,23], the proposed method also utilize an intermediate layer filtering on the multiple channel maps. It can not only improve the accuracy of the detector, but also have a real-time running speed.

Our contributions of this work are summarized as follows:

(1) A novel light-weighted filter bank is proposed to enhance the discriminative information from the pooled channel maps of the training data;
(2) Label information is integrated into the procedure of feature dimension reduction which preserves the structure of the manifold space.
(3) Our detector is very efficient to train and also can run about 20 fps in 480 × 640 resolution images.

2 Review of Related Works

2.1 Multiple Channel Map

The multiple channel map [9] models the feature C of image I as a channel generation function Ω, so the feature of image I can be represented as $C_i = \Omega_i(I)$, where $\Omega =$

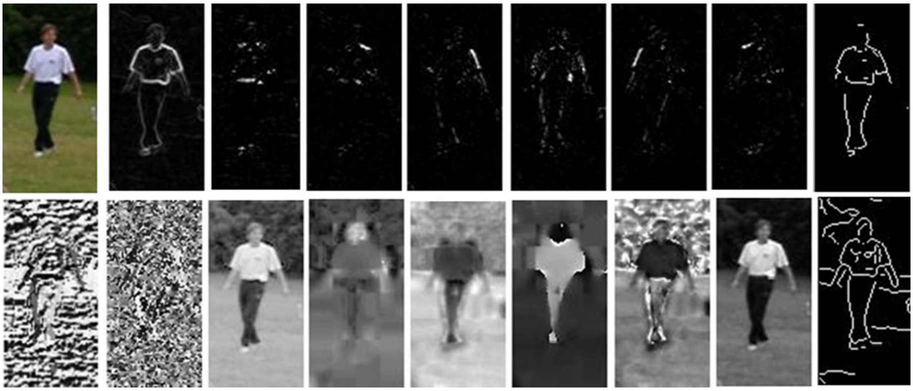


Fig. 1 Different channels of the input image (the *first row* gives the original, magnitude, gradient angle discrete from 1 to 6, sobel image, while the *second row* illustrates the lbp, orientation, luv, hsv and canny image)

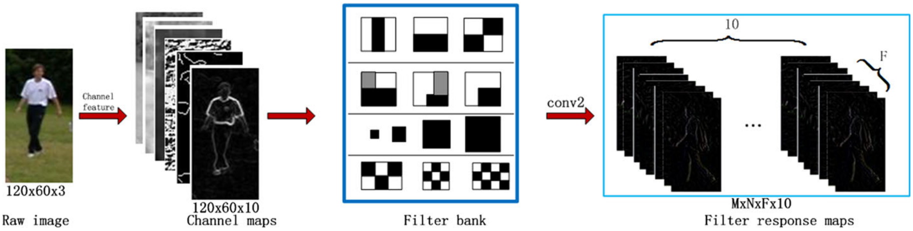


Fig. 2 Filter banks for channel features

$\{\Omega_i\}, i = 1, 2, \ldots, n$. C_i is the ith channel map for image I, Ω_i is the ith channel generation function. The channel generation function can be a linear function, such as gray-level image of original image I or nonlinear function, such as gradient image. Each channel represents a different feature space which is derived from the original image. The different channel of the original RGB image is shown in Fig. 1. It can be observed that different channels can reflect different aspect of input image, such as magnitude, gradient orientation, edge map and color channels, etc. In this paper, we only use ten channels (LUV color space, gradient magnitude and six gradient orientation maps) for our detection system due to its fast computation and high informative feature.

2.2 Filtered Channel Feature

Recent studies show that filtering on the channel maps can greatly improve the performance of the detector [23]. Several popular effective filters have been developed for object detection, such as Haarlike filter [24], informedHaar feature [14], checkboard [23], square filters [25] and SSD filters [26], etc. These filter banks are shown in Fig. 2. The Haarlike filter, informed-Haar, square filter and checkboard filters are shown from top to bottom at the third step in Fig. 2. It is shown that all the filters require calculating the mean pixel value in a local area and then computing the difference between these local areas. Each filter only returns a single feature value which represents the response in a local area. Then all the channel maps are convolved with these filters to produce the final response maps which are fed to

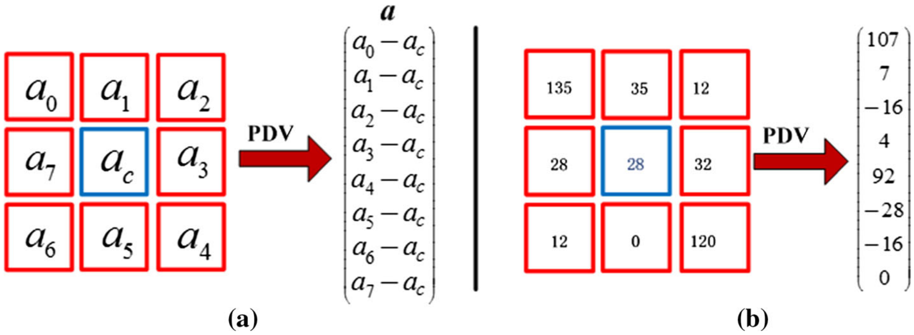


Fig. 3 Local differential filter bank

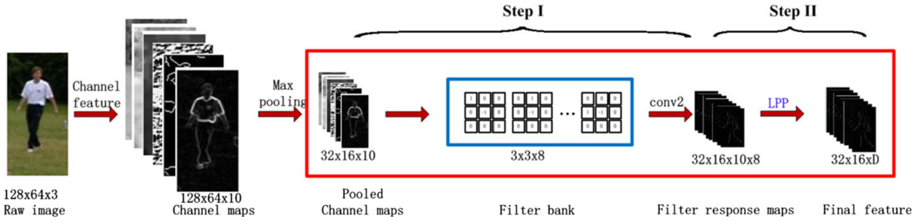


Fig. 4 Filtering with local high-order differential filter bank

the AdaBoost classifiers. Albeit simple, this straightforward step significantly benefits the performance improvement [23] (Fig. 3).

3 Our Proposed Method

3.1 Local Differential Filter Bank

The filtered channel features [14,23] can be used as an intermediate layer filtering imposed on the low-level features, such as InformedHaar [14], CheckBoard [23], SSD [26], etc. Inspired by this, we model this type of feature as a two layers network in this paper, which is shown in Fig. 4.

The computation procedure of our filter bank is shown in Fig. 4 (step I). The calculation of the local differential filter bank is similar to the calculation of Pixel Differential Vector (PDV), which is shown in Fig. 3. The computation of PDV a for each reference pixel and the dimension of the PDV equals to 8 (Fig. 3a). The local filtering is used for enhancing the shape information of the channel map, which is more discriminative than the original channel map. In order to illustrate the procedure of calculating PDV, we give a toy sample as shown in Fig. 3b.

We have also demonstrated four channels of the filtering results with our filter banks corresponding to the LU colorspace, magnitude and gradient angle channel map, which are shown in Fig. 5. It can be seen from Fig. 5, the average contour of the pedestrians is highlighted after filtering with our local differential filter bank.

It is worth mentioning that there exist two main differences between our method and the methods referred in Sect. 2.2. Firstly, our method depicts the details of statistic information for

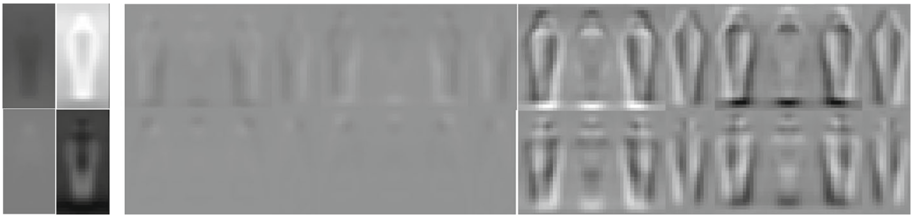


Fig. 5 Four average channel maps and filtered high-order differential maps

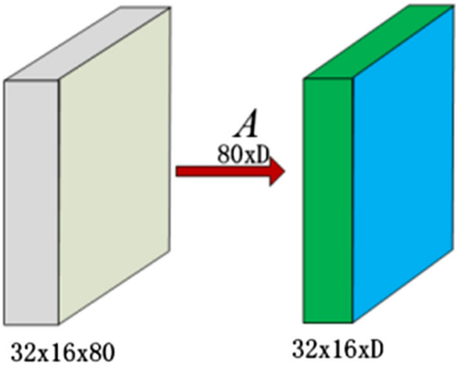


Fig. 6 Our proposed feature extraction framework

each position of channel maps instead of each block region. Our fine-grained filter bank allows capturing more discriminative information of the pedestrians. Secondly, in order to preserve the pedestrian manifold in the feature space, we make use of the class label information to reduce the feature dimension. It can be considered as a feature transformation from high dimensional feature space to low discriminative manifold space which will be discussed in Sect. 3.2.

3.2 Supervised Filtered Channel Feature Dimension Reduction

Although our method can distill discriminative information, it lead to a much higher feature dimension (40,960 dim) than the original channel map (5120 dim) for each training sample. Despite a number of unsupervised dimension reduction methods proposed in recent years, class-specific discriminative information is often ignored. In this paper, we attempt to utilize a supervised dimension reduction algorithm to distill discriminative information which is effective and computationally efficient. As shown in Fig. 4 (step II), our method makes use of a much simpler filter bank which enables capturing statistic information in each local position. Besides, our method is deeper than other methods (e.g. haarlike, square and informedhaar, etc.), which can integrate the class label information into the feature extraction procedure.

The step II of Fig. 4 is enlarged in Fig. 6. We formulate the supervised dimension reduction of the channel map as a classical spectral graph embedding problem which can be solved by the LPP algorithms [27]. This LPP method not only preserves the manifold structure of the filter response maps, but also reduces the side effect of misalignment of training data. The eigenmap (projection matrix) of A can be formulated as a generalized eigenvector problem, which is shown in Eq. (1).

$$XLX^T a = \lambda XDX^T a \tag{1}$$

where $L = D - W$, $D_{ii} = \sum_j W_{ji}$. The eigenmap A is constructed by the eigenvectors of solution of Eq. (1), correspond to the top d eigenvalues. W is calculated by the heat kernel based on the pooled channel maps in the positive training data set, which is shown in Eq. (2).

$$W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}} \tag{2}$$

where t is the window size, x_i is the feature vector corresponding to ith sample, $x_i \in R^{80}$. In this paper, we calculate the feature vector x for each sample with average PDV for all channels, which is efficient and effective. It can be formulated in Eqs. (3) and (4).

$$\bar{x} = \frac{1}{512N} \sum_{i=1}^{N} \sum_{j}^{512} m_j^i \tag{3}$$

$$x = \frac{1}{512} \sum_{j}^{512} (m_j - \bar{x}) \tag{4}$$

where $\bar{x}$ is the mean PDV for the entire positive samples at all positions, N is the number of positive samples, $m_j \in R^{80}$ is the PDV at the jth position of the channel map.

It is worth to mention that the filter response map is reshaped from 4-dimensional tensor (32 × 16 × 10 × 8) to a three dimensional tensor (32 × 16 × 80), which can be readily to implemented. The dimension D of the final feature (last step in Fig. 6) equals to 10 × d.

3.3 Computational Complexity Analysis

Suppose the input image size is $H \times W \times 3$, the pooling size is s, then the resulted channel map is $\frac{H}{s} \times \frac{W}{s} \times 10$. The dimension of the local differential filter is $3 \times 3 \times 8$ and the dimension of the filtered result is $\left(\frac{H}{s} - 2\right) \times \left(\frac{W}{s} - 2\right) \times 10 \times 8$. Besides, we will learn a feature projection matrix with size $80 \times d$, then the feature dimension is reduced to $\left(\frac{H}{s} - 2\right) \times \left(\frac{W}{s} - 2\right) \times 10 \times d$. In this paper, we use H = 128, W = 64 for the training samples, pooling size s = 4, each feature is a float type and there're about 50,000 samples in each training round. So on the training session, the features of original and reduced feature will consume 6.26 and 3.14 Gigabytes memory respectively in total. In the test section, the input image size is 640 × 480 × 3, the extract feature is 120 × 160 × 40, so each image will consume about 2.8 megabytes memory. It can be inferred that our method is memory efficient in both training and testing section, and our detector also can run on the common desktop computers.

The computation complexity of multiple channel maps is about $O(10 \times W \times H)$, the complexity of calculating filter banks is $O\left(\frac{H}{s} \times \frac{W}{s} \times 40\right)$, and the complexity of feature reduction is $O\left(\frac{H}{s} \times \frac{W}{s} \times 80 \times d\right)$. In this paper, the parameters are set as W = 480, H = 640, s = 4, d = 4, which is computationally efficient. Furthermore, we have implemented the convolution function with multi-thread operation which is very fast to calculate.

4 Experiments

4.1 Experimental Setting

The detailed experiment setting is described as follows. The size of the pedestrian window is 128 × 64, and every positive sample is cropped from the annotated images. Each annotation of pedestrian is jittered to deal with misalignments problem, which yields approximately

24,740 positive samples for the INRIA dataset and 24,498 positive samples for the Caltech dataset. The pooling template size is 4×4 pixels, which shrink the original channel maps (size $128 \times 64 \times 10$) into pooled maps (size $32 \times 16 \times 10$). We make use of GentleBoost algorithm for feature selection, with depth-2 decision tree as weak classifiers. The soft cascade structure is employed in our model training, and the final detector comprised of 2048 weak classifiers. We use the Piotr's toolbox [17] to calculate the channel features and also utilize the evaluation code [17] publicly available.

4.2 Comparisons of ACF, PDV and Different Dimensions of LPP

In this section, we have compared our methods with the baseline (ACF) which is shown in Fig. 7. It can be seen that our method PDV is slightly better than ACF, which indicates that statistic information of each pixel in the channel maps do helps to promote the performance of the detector. It is also shown that the ROC curves of the baseline and the PDV feature have an intersection when FPPI = 0.2 and miss rate equals 0.13. When the FPPI is less than 0.2 the PDV is better than ACF, and when the FPPI is greater than 0.2 the ACF is slightly better than PDV. It is can be explained with less positives produced in our detection framework, which indicates that PDV is more discriminative than ACF. The PDV have a miss rate of 7 %@FPPI = 1, whereas the ACF is 9 %@FPPI = 1.

We also make a comparison between different hyper parameters of the LPP method. The results are shown in Fig. 7. We can see that our PDV + LPP (d = 4) achieves the best result (average miss rate equals to 14.07 %) when the reduced dimension equals to 4 that is the half of the original dimension. The deteriorating performance is also revealed when the dimensions exceeds 4. Based on this experimental observation, we fixed our hyper parameter d = 4 in the following experiments. Besides, it is can be concluded that distill discriminative information can improve the performance of the detector.

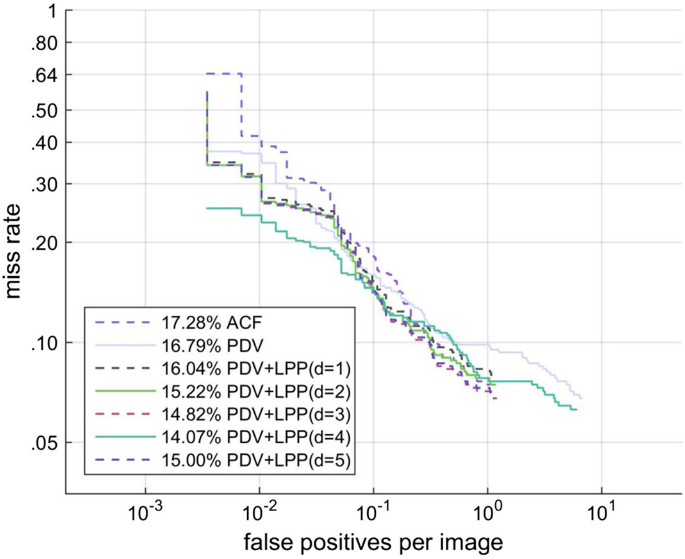


Fig. 7 Parameter selection for LPP

4.3 Comparisons with State-of-the-Arts Methods on INRIA Dataset

In order to validate the effectiveness of our proposed method, we have compared our methods PDV and PDV + LPP with other state-of-the-art methods on the INRIA dataset. The INRIA dataset containing high resolution pedestrians in complex background, such as deformation, occlusion.

The experimental result is shown in Fig. 8. In Fig. 8a, we evaluate our detector with ROC curve with miss rate@FPPI. We can see that the PDV is slightly better than ACF [28] and DPM [29] based method, but still slightly worse than InformedHaar feature [14] which is most close to our method. But our PDV + LPP method can surpass the InformedHaar feature [14] which is due to its supervision of labels when getting discriminative information by manifold structure preserving dimension reduction. Our PDV + LPP reports a miss rate of 6 %@FPPI = 1 and miss rate of 12 %@FPPI = 0.1. It is shown that PDV + LPP achieves state-of-the-arts result when compare with most of the hand-crafted features.

The evaluation result of our detector with the recall-precision curve is shown in Fig. 8b. The results indicate that our method works marginally better than others and reports the same performance as wordchannel, randforest and NAMC. It is suggested that evaluation with average precision is less significant of difference than average miss rate@FPPI evaluation on this dataset.

4.4 Comparisons with State-of-the-Arts Methods on the Caltech Dataset

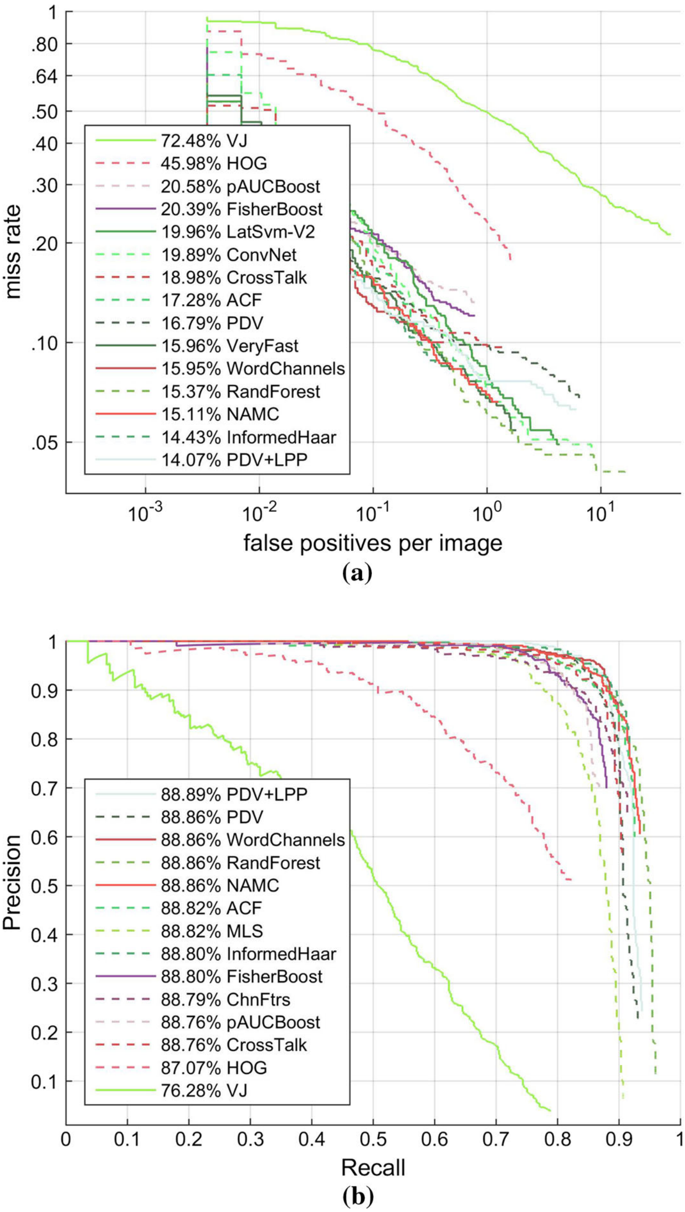
We have also conducted experiments on the Caltech dataset, which is the large-scale dataset for pedestrian detection. This dataset features low resolution of pedestrians in the city road, which is captured by a camera fixed in the car. It's one of the most challenging datasets, due to low resolution and occlusion in the video.

We evaluated our detector with ROC curve with miss rate@FPPI in Fig. 9a. The comparison results indicate that both of the PDV and PDV + LPP methods can achieve state-of-the arts results on this low resolution dataset. The PDV + LPP achieves an average miss rate of 24.42 % which is best among all the compared methods, such as ACF [28], InformedHaar [14] and LDCF [30]. The phenomenon also coincides with the results on the INRIA dataset. The results also indicate that PDV performs much better for the low resolution pedestrians and the PDV + LPP performs best for both high resolution and low resolution dataset. Furthermore, our method has achieved best performance for almost all the hand-crafted feature on this dataset.

We also evaluated our detector with the recall-precision curve, which is shown in Fig. 9b. Our method (PDV + LPP) is also better than other methods, but the difference of average precision between PDV and PDV + LPP is small (less than 1 %). Surprisingly, our PDV is slightly worse than LDCF in miss rate@FPPI curve, but the results reversed in average precision of the recall-precision curve. We think the reason behind that these two detectors have different performance when the recall approximately equal to 0.78. Our detector is better than inforHaar when recall rate is less than 0.78 and slightly worse when recall is greater than 0.78.

4.5 Sample Detections on the INRIA and Caltech Datasets

In this section, some sample detections on the INRIA and Caltech dataset are provided, which are shown in Figs. 10 and 11 respectively. From Fig. 10, it can be seen that the most challenge scenario in this dataset is that many pedestrians overlap with others and some pedestrians carry

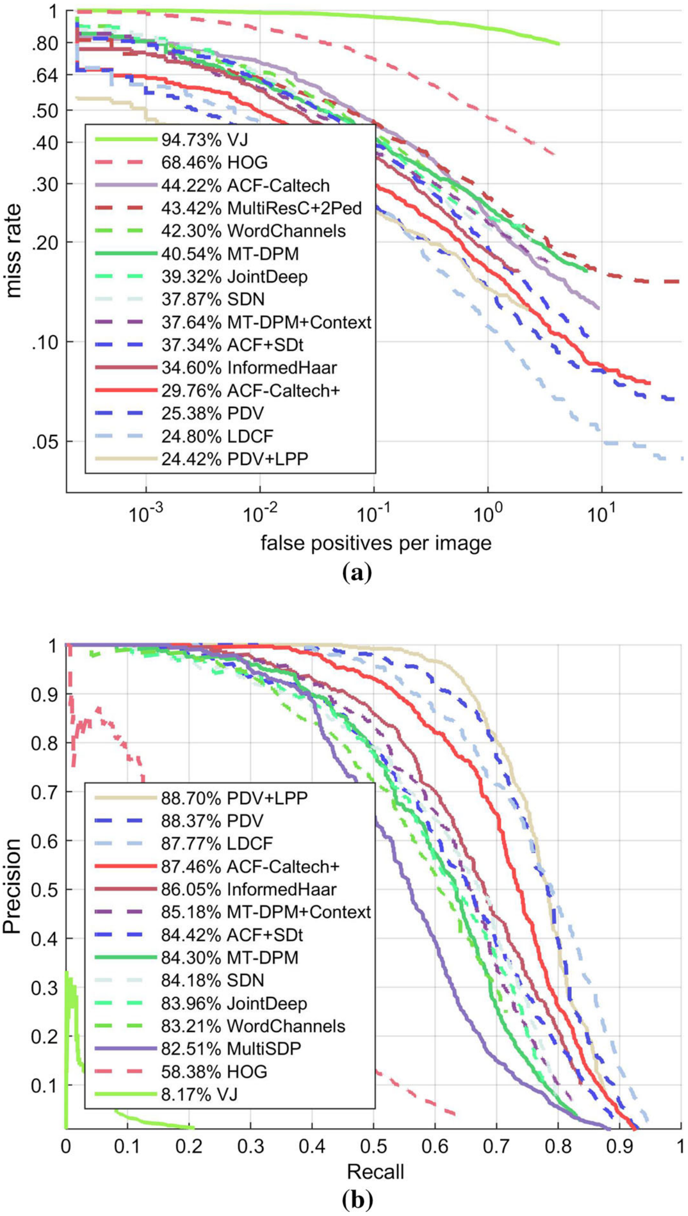


(a)

(b)

Fig. 8 Comparisons with state-of-arts in INIRA dataset

package can also lead to false negative. Another situation is the deformation of pedestrians which is quite difficult to detect for a single rigid pedestrian detector. From Fig. 11, it can be concluded that the current best detector still have a long way to go. There're lots of false positives appear in the image which is due to the small size of the pedestrians. It also can be found that, in some scenarios, the quality of the image is quite difficult even for human being to observe due to image blurring and low resolution.



(a)

(b)

Fig. 9 Comparison with state-of-the-arts methods on Caltech dataset

4.6 Runtime Comparison

Our detector is implemented with Matlab R2014b and visual studio 2013 on DELL workstation of Precision T7610 (8 core CPU E5-2650 ×2, 2.6 GHz, 64G). It takes about 6 h to train a four-stage detector and the detection speed for a 640 × 480 image can get approximately 20 fps. The speed of our detectors with different features using average number of weak

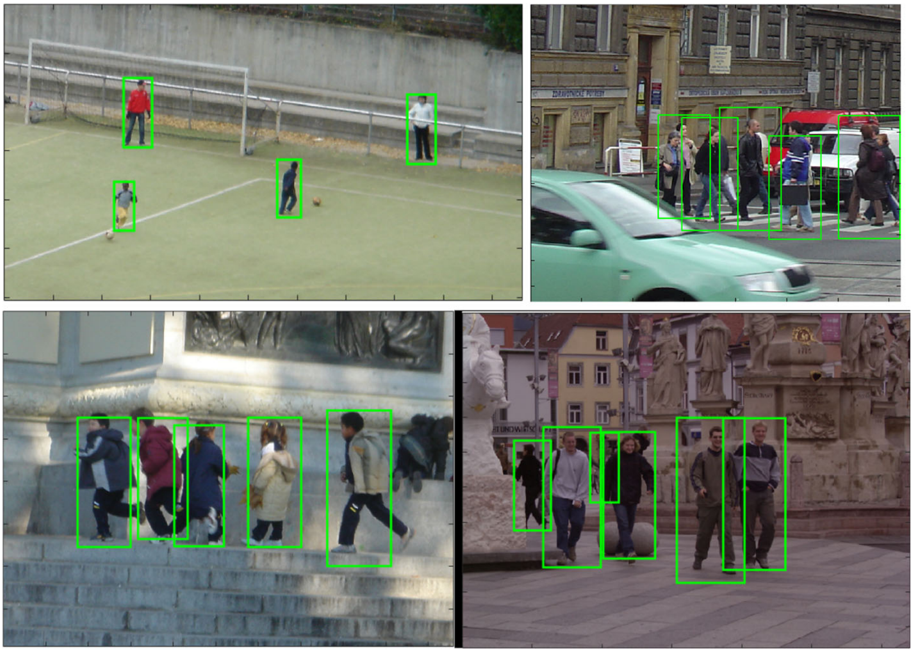


Fig. 10 Sample detections on INRIA dataset

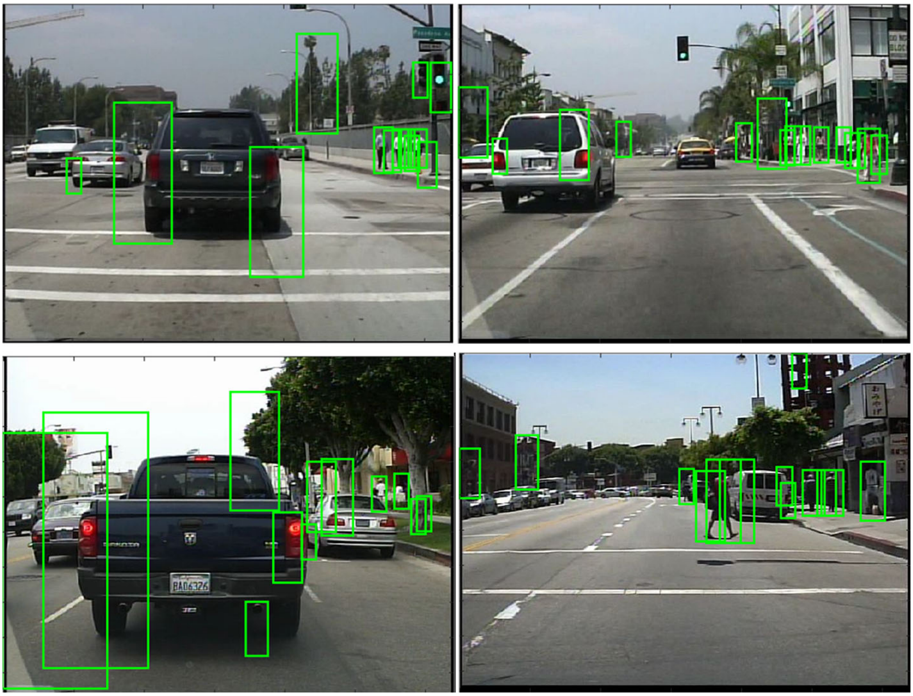


Fig. 11 Sample detections on Caltech dataset

Table 1 Runtime comparison

Name	Running speed (fps)	Average reject number
ACF	30	3.86
PDV	20	3.09
PDV + LPP	19	2.94

classifiers to reject a background patch is called Average Reject Number in Table 1. We can see that both of the PDV and PDV + LPP method are slower than the original ACF features.

5 Conclusion

We have proposed a supervised local high-order differential channel feature for pedestrian detection, which make use of the labels of the positive samples in the feature design process. This feature can achieve state-of-the-art results compared with other hand-crafted features on both the INRIA and Caltech datasets, which indicate that the supervision of label is helpful in designing discriminative features. Furthermore, we also point out that the most challenge problems of these two dataset, which still need to be focus on. Our future work will study on the occlusion reasoning for overlapping pedestrians which have a large space to improve the performance.

Acknowledgements This project is supported by the NSF of China (61305058, 61473086), the Fundamental Research Funds for the Jiangsu University (13JDG093), the NSF of the Jiangsu Higher Education Institutes of China (15KJB520008), the NSF of Jiangsu Province (Grants Nos. BK20130471, BK20140566, BK20150470, BK20130501), China Postdoctoral science Foundation (2014M561586) and the Fundamental Research Fund for the Central Universities of China (N150403006).

References

1. Htike KK, Hogg D (2016) Adapting pedestrian detectors to new domains: a comprehensive review. Eng Appl Artif Intell 50:142–158
2. Girshick R, Donahue J, Darrell T, Malik J (2016) Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans Pattern Anal Mach Intell 38(1):142–158
3. Hosang J, Omran M, Benenson R, Schiele B (2015) Taking a deeper look at pedestrians. In: IEEE conference on computer vision and pattern recognition, pp 4073–4082
4. Wang H, Yuan C, Hu W et al (2014) Action recognition using nonnegative action component representation and sparse basis selection. IEEE Trans Image Process 23(2):570–581
5. Geronimo D, Lopez AM, Sappa AD (2010) Survey of pedestrian detection for advanced driver assistance systems. IEEE Trans Pattern Anal Mach Intell 32(7):1239–1258
6. Dollár P, Wojek C, Schiele B, Perona P (2012) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34(4):743–761
7. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceeding of IEEE conference on computer vision and pattern recognition, vol 1, pp 886–893
8. Zhe L, Davis LS, Doermann D, DeMenthon D (2007) Hierarchical part-template matching for human detection and segmentation. In: IEEE 11th international conference on computer vision, pp 1–8
9. Dollar P, Tu Z, Perona P, Belongie S (2009) Integral channel features. In: British machine vision conference
10. Shen J, Yang W, Sun C (2013) Real-time human detection based on gentle MILBoost with variable granularity HOG-CSLBP. Neural Comput Appl 23(7):1937–1948
11. ouyang W, Wang X (2015) A discriminative deep model for pedestrian detection with occlusion handling. In: IEEE conference on computer vision and pattern recognition, pp 3258–3265

12. Yang B, Yan J, Lei Z, Li SZ (2015) Convolutional channel features. In: Proceedings of the IEEE international conference on computer vision, pp 82–90
13. Wang X, Han T, Yan S (2009) An HOG-LBP human detector with partial occlusion handling. In: IEEE international conference on computer vision
14. Zhang S, Bauckhage C, Cremers AB (2014) Informed haar-like features improve pedestrian detection. In: IEEE conference on computer vision and pattern recognition
15. Girshick R (2015) Fast R-CNN. In: IEEE international conference on computer vision (ICCV), pp 1440–1448
16. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: IEEE conference on computer vision and pattern recognition
17. Angelova A, Krizhevsky A, Vanhoucke V (2015) Pedestrian detection with a large-field-of-view deep network. In: IEEE international conference on robotics and automation (ICRA), pp 704–711
18. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
19. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507
20. Krizhevsky A, Ilya S, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
21. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2014) OverFeat: integrated recognition. In: Localization and detection using convolutional networks, proceedings of international conference on learning representations
22. Mohamed AR, Dahl GE, Hinton G (2012) Acoustic modeling using deep belief networks. IEEE Trans Audio Speech Lang Process 20(1):14–22
23. Zhang S, Benenson R, Schiele B (2015) Filtered channel features for pedestrian detection. In: IEEE conference on computer vision and pattern recognition
24. Zuo X, Shen J et al (2015) Haarlike Feature revisited: fast human detection based on multiple channel maps. In: 12th International symposium on neural networks, pp 240–247
25. Benenson R, Mathias M, Tuytelaars T, Van Gool L (2013) Seeking the strongest rigid detector. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 3666–3673
26. Shen J, Zuo X, Yang W, Yu H, Liu G (2016) Learning discriminative shape statistics distribution features for pedestrian detection. Neurocomputing 184:66–77
27. He X, Niyogi P (2004) Locality preserving projections. Adv Neural Inf Process Syst 16:153–160
28. Dollar P, Appel R, Belongie S, Perona P (2014) Fast feature pyramids for object detection. IEEE Trans Pattern Anal Mach Intell 36(8):1532–1545
29. Felzenszwalb P, McAllester D, Ramanan D (2008) A discriminatively trained, multiscale, deformable part model. In: IEEE conference on computer vision and pattern recognition
30. Nam W, Dollar P, Han JH (2014) Local decorrelation for improved pedestrian detection. Adv Neural Inf Process Syst 27:424–432