IMPROVING CHANNEL FEATURES USING STATISTICAL ANALYSIS FOR PEDESTRIAN DETECTION

Chen Zhang and Joohee Kim

Department of Electrical and Computer Engineering
Illinois Institute of Technology, Chicago, USA

ABSTRACT

As one of the most successful features in vision-based pedestrian detection, filtered channel features have drawn considerable attention in the research community. In this paper, we improve the channel features by performing a statistical analysis of each channel and taking both the average map and variance map into account. The average map informs us the local structure of the human body parts and how it looks like in general. The variance map points out the region where high variation in human poses takes place. Based on both the average and variance information, we create two types of templates, each employs a different design strategy and generates the feature value in different ways. We also utilize the co-existence of strong responses that take place in non-neighboring region pairs to enrich the feature pool. Experimental results show that the proposed method achieves good performance on the INRIA and Caltech-USA benchmark with small feature size, short training time, and less training examples are required.

Index Terms— pedestrian detection, channel features, feature extraction, statistical model.

1. INTRODUCTION

Vision-based pedestrian detection has been one of the most active research areas in computer vision for a long time. The modern pedestrian detection system includes both image-based feature extraction and machine learning based classification. Despite the huge progress made in the last decade, pedestrian detection still remains as a difficult problem due to the well-known challenges such as cluttered background and huge variations in pedestrian's clothes and poses [1] [2]. Based on the analysis of the top-performing detectors [3], designing better features is one of the key components to improve the detection performance. To obtain a rich and discriminative feature pool, many methods generate features by convolving an image with various filters and a large number of features are obtained. However, many applications like advanced driver assistance systems (ADAS) and autonomous driving require real-time pedestrian detection. In these applications, fast processing speed and compact features are desired.

In this paper, we propose a method that focuses on improving the feature design using the statistical analysis of positive training examples. The remainder of the paper is organized as follows: In Section 2, related works regarding vision-based pedestrian detection are discussed. In Section 3, our proposed method is introduced. In Section 4, experimental results are shown. We conclude in Section 5.

2. RELATED WORK

A typical pedestrian detection system can be divided into two main components: feature generation and classification. Since our work is related to enhancing the filtered channel features, our review of related works will focus on feature generation in pedestrian detection.

Histogram of Oriented Gradients (HOG) features [4] are the most well-known features that show great success in both pedestrian and other general object detection. In integral channel features (ICF) [5], multiple channel images such as CIE-LUV color channels, gradient magnitude, and gradient orientation histograms are computed. Then, ICF randomly generates rectangular boxes for each channel and uses the sum of pixel values inside each box as the feature.

Since ICF, many methods [6] [7] [8] [9] [10] have been proposed that generate high-level features based on the 10 channels introduced in ICF. The main idea behind these methods is to extract local structures from channel maps that represent the shape of human body parts such as head and shoulder. A large template pool is designed, and each template in it is aimed to yield a high response when a certain local structure appears. Since human poses are highly variant and it is hard to estimate the exact location and shape of local structure in them, an exhaustive filtering strategy is usually used. For methods [7] [8] [9] [10] that employ the exhaustive filtering strategy, each location in the channel map is filtered by all templates in the template pool. Thus, a large number of features are produced.

In InformedHaar [11], the prior knowledge of human shape is used to avoid the exhaustive filtering with a large number of templates. The method proposed in [11] uses a statistical human model to design a specific template at each

position in the channel space. A small feature size is achieved by filtering each position in the channel space with a dedicated template. However, the up-right human shape model is derived only from the gradient magnitude channel, and this single model does not fit the rest of 9 channels very well. Besides, since the templates are designed from the logical components instead of real average channel map, it is hard to represent the nature of how the local structure appears. Finally, the human shape model derived from the average information does not handle the effect of high variation in human poses very well.

Recently, deep learning based pedestrian detectors [12] [13] [14] featuring convolutional neural networks (CNN) [15] [16] show a great improvement in terms of detection accuracy. During the training of CNN, not only the optimal rules of decision making but also the optimal way to generate features are learned. It should be mentioned that in CNN, the key to success relies on large number of training examples and the powerful GPU with the memory large enough to store features, which limits the popularity of CNN based methods in non-GPU implementation situation.

The proposed method is inspired by [11] and our contributions are as follows: First, we extend the statistical analysis to each channel so that each channel's own behavior is examined and used to design the corresponding features. During the statistical analysis, both the average pixel value and the variance are analyzed. Second, we design two types of templates to extract features that characterize the local human body structure: One is the low variance template that follows the structure on average channel map. The other is the high variance template that uses random pooling to deal with high variation in human poses. Third, we also examine the relationship between non-neighboring regions in each channel. The co-existence of strong responses from non-neighboring region pairs are used as additional features to improve pedestrian detection.

3. PROPOSED METHOD

We start introducing our method with the statistical analysis of channel images. Then the details of designing the templates and features are introduced.

3.1. Statistical analysis of channel images

We begin the analysis with the introduction of the channels in ICF [5]. Given an input image, a channel is a registered map in which each pixel on the map is computed from corresponding patches of input pixels. By doing so, the layout of the input image is kept for each channel. The 10 channels used in ICF [5] are 3 CIE-LUV color channels, 1 gradient magnitude (GM), and 6 gradient orientation histogram channels (Hist1-Hist6).

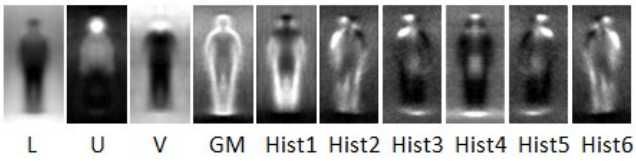

Fig. 1. Average map for each channel.

We perform a statistical analysis using the INRIA dataset [4], which is one of the most commonly used training and validation dataset. We resize all 1235 positive examples to 128×64 pixels and calculate the average map for 10 channels. The average maps are shown in Figure 1. It is observed that each channel has unique local structures that represent the shape of human body parts. Some channels like L channel and GM channel show a complete human contour, while others show a strong response on local structures like the head and the shoulder. In conclusion, the average channel map gives us the basis of how the local structures of human body parts look like and where they are. We design templates based on the shape of local structures at each position on the average channel and use these templates to extract features. Since each template only controls one specific position, exhaustive filtering is no longer needed. Thus, a smaller feature size is achieved.

In addition to the average map, we compute the variance map for each channel using the same positive examples. The variance map for each channel is shown in Figure 2. A bright pixel indicates a high variance. It is observed that many regions that contain structure information about shoulders and arms have both high average value and high variance. This observation suggests that when we try to capture local structures around these regions, a template designed directly from the average channel map cannot handle the high variation very well. A more flexible approach should be applied to deal with these regions.

Finally, we perform an analysis of the relationship between non-neighboring 8×8 pixel regions using their co-existence. For each positive example, its 10 channels are computed, and each channel's average pixel value is calculated. For each region, if its average pixel value is higher than its channel's average pixel value, it is considered to have a strong response. We define that a co-existence of strong responses between two regions takes place if both regions have a strong response. We examine all possible non-neighboring region pairs in each channel and determine if the co-existence takes place. This process is repeated for all positive examples and a co-existence probability map for each channel is obtained by accumulating the frequency of the co-existence for all region pairs from all positive examples. One of the co-existence probability maps is shown in Figure 3. Note that the axis is transformed from 2D coordinate to 1D.

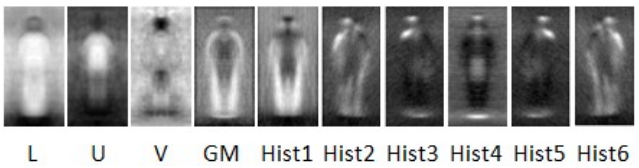

Fig. 2. Variance map for each channel.

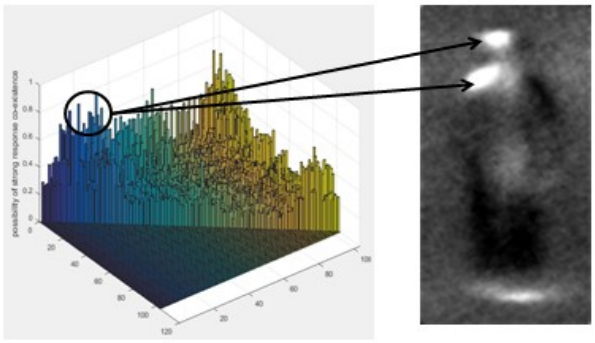

Fig. 3. Co-existence probability map for Hist3 channel. The circle indicates that the shoulder and the head regions have a high probability to have strong responses at the same time.

By analyzing the co-existence probability map, we observe that in each channel, some regions have much higher probability to show strong response together than others e.g. the head and shoulder regions in Hist3 channel. Based on this observation, we extend the feature pool by adding features that come from the two non-neighboring regions which tend to have strong responses together.

3.2. Feature design

In this section, we explain the details of designing the features. The feature design is divided into two parts: Local structure based feature design and co-existence based feature design.

3.2.1. Local structure based feature design

We use the 10 HOG + LUV channels for our feature design. The input image patch size is set to 128×64 pixels, which is suitable for detecting pedestrians with a height of 100 pixels. We compute the average map and the variance map for each of the 10 channels using all positive examples from the INRIA pedestrian training set. Then both maps for each channel are downsampled to size 32×16 pixels. The downsampling helps reduce the computational complexity.

We segment each channel's average and variance maps by thresholding and label each pixel in them. Each map's threshold is determined by its average pixel value. Specifically, a '1' is assigned to the pixel that is larger than the threshold. Otherwise, a '0' is assigned. We denote the labeled average map and variance map AM and VM, respectively. Each of them is a 32×16×10 volume.

To design the local structure based feature, we slide a 3×3 rectangular window over AM and VM. At each sliding position (x, y, c) where x and y are the 2D coordinates and

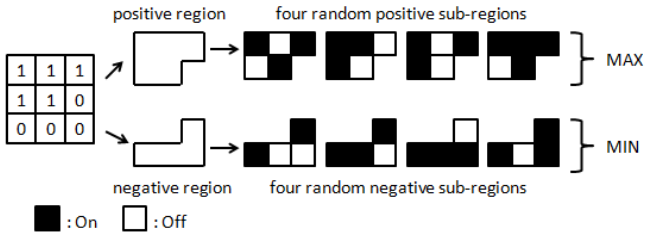

Fig. 4. Example of generating eight sub-regions for a 3x3 high variance template.

c is the channel, a unique 3×3 template is generated. There are two template types. One template is defined as a high variance template $T_H\{x, y, c\}$ if the majority of its pixels has the variance label of '1' on VM. Otherwise, the template is defined as a low variance template $T_L\{x, y, c\}$. These two types of templates are designed in different ways and generate feature values differently.

For low variance templates, the template design follows the local structure in the average map. Suppose a low variance template $T_L\{x, y, c\}$ at the location (x, y) in channel c is a 3x3 matrix and its element at (i, j) is denoted as $T_L\{x, y, c\}(i, j)$. For each element, if the corresponding label in the average map is '1', then the element is given the value of $+1$, otherwise, -1 is assigned to that element. After all 9 elements in the template are assigned either $+1$ or -1, normalization is carried out. We denote the collection of the 10 channels from the input image as I. The output feature value for a low variance template $T_L\{x, y, c\}$ is obtained by

$$F(T_L\{x, y, c\}) = \sum_{i=0}^{2}\sum_{j=0}^{2} I(x+i, y+j, c)\, T_L\{x, y, c\}(i+1, j+1) \quad (1)$$

For high variance templates, the areas that they cover contain unpredictable local structures brought by high variance in human poses. We apply random pooling to generate the feature value for this type of local structures. To do this, we first segment the template into two regions: the positive region (P) and the negative region (N). The positive region contains all pixels that are labeled '1' while the negative region contains pixels with the label '0' in the labeled average map AM. Then sub-regions for both positive and negative region are created by randomly setting each of the pixels 'on' and 'off'. Note that each sub-region should contain the amount of 'on' pixels that are equal to or larger than half the number of total pixels in the corresponding region. See Figure 4 for example. We randomly generate 4 sub-regions P_i $(i = 1,2,3,4)$ and N_i $(i = 1,2,3,4)$ for P and N, and the output feature value of any high variance template $T_H\{x, y, c\}$ can be expressed as:

$$F(T_H\{x, y, c\}) = \max_{i=1,2,3,4} \frac{S_{P_i}}{Num_{P_i}} - \min_{i=1,2,3,4} \frac{S_{N_i}}{Num_{N_i}} \quad (2)$$

where S_{P_i} and S_{N_i} are the sum of pixels that are 'on' in the sub-regions P_i and N_i, respectively. Num_{P_i} and Num_{N_i} are the total number of pixels that are 'on' in that sub-region.

In total, there are 30×14×10=4200 templates because some templates go out of the image border and are discarded. Among the 4200 templates, 2854 are low variance templates, and 1346 are high variance templates. Since each template gives a feature value, there are 4200 local structure features in the feature pool.

3.2.2. Co-existence based feature design

We calculate the co-existence probability map using the method described in Section 3.1 without downsampling the channels. The region's size is set as 8×8 pixels. From each channel's co-existence map, we select the region pairs in each channel that have the co-existence probability higher than 0.7. For each region pair R and R′, 4 random rectangular sub-regions R_i (i = 1,2,3,4) and R'_iR′i (i = 1,2,3,4) are generated inside each region, respectively. There is no length and height ratio restriction to these sub-regions, but each of them must have an area that is equal or more than half the area of R and R′. The feature value of region pair R and R′ is:

$$F(R, R') = \delta(\max_{i=1,2,3,4} A_{R_i} + \max_{i=1,2,3,4} A_{R'_i}), \quad (3)$$

where A_{R_i} and $A_{R'_i}$ are the average pixel values in sub-regions R_i and R'_i. δ is a penalty value that is inversely proportional to the absolute difference between the maximum of A_{R_i} and $A_{R'_i}$.

In total, there are 3116 pairs of regions that have the co-existence probability higher than 0.7 and we add 3116 co-existence features into the feature pool. In total, the size of the final feature is 4200+3116=7316. A strong classifier consisting of 2048 depth-2 decision trees is trained [17] as the classifier.

4. EXPERIMENTAL RESULTS

In this section, we compare our method with other state-of-the-art methods on the INRIA dataset [4] and the Caltech-USA pedestrian dataset [18]. 2470 and 3263 positive training examples are used to train the classifier for the evaluation on the INRIA and the Caltech dataset, respectively. Because the original annotation is not accurate on the Caltech dataset, the refined annotation [19] is used during the training process, which gives us a 7% improvement in the miss rate.

For each full image detection, pedestrian proposals are generated using the sliding windows strategy. For multi-scale pedestrian detection, the input image is scaled using a scale stride of 1.07. The stride of the sliding window is always set to 4 pixels no matter what the scale is. For each proposal, we generate features described in Section 3 and send them to a trained classifier. After the whole image is scanned and the classification scores of all proposals are computed, non-maximum-suppression (NMS) is performed to eliminate the overlapping detections.

Table 1 compares our method with other state-of-the-art methods. The log-average miss rate (MR) is used as the measurement. A smaller MR indicates better performance.

It can be seen from Table 1 that our method outperforms most channel feature based methods. Besides, Table 2 shows that the feature size of the proposed method is smaller than others, which results in a short training time. Compared with the CNN-based methods [12] [13] [14], our hand-crafted feature generation method is outperformed in terms of detection accuracy. However, these CNN-based methods have a large feature size and require a large number of training examples, which is around 100000, while the proposed method only requires less than 15000 examples. Most of the CNN based methods are implemented on expensive GPUs, while our method runs at a higher frame rate on a PC with Intel i7 CPU only.

Table 1. Log-average miss rate comparison with other non-CNN based state-of-the-art methods.

Method	MR INRIA	MR Caltech-USA
ChnFtr [5]	22.18%	56.34%
Roerei [7]	13.06%	43.90%
ACF [6]	17.28%	51.36%
LDCF [9]	13.79%	24.80%
InformedHaar [11]	14.43%	34.60%
Checkerboards [8]	N/A	18.47%
NNNF-L4 [10]	N/A	16.20%
CompACT [12]	N/A	11.70%
SAF-RCNN [13]	N/A	9.32%
RPN+BF [14]	N/A	9.60%
Ours	11.38%	16.27%

Table 2. Comparison of feature size, training time and running speed with other state-of-the-art methods.

method	feature size	training time	fps
Checkerboards [8]	312320	4 hours	0.50
Roerei [7]	718080	2.5 days	N/A
LDCF [9]	72000	N/A	3.62
InformedHaar [11]	12760	1 hour	0.63
NNNF-L4 [10]	20480	N/A	1.13
CompACT [12]	25088	N/A	2 (GPU)
SAF-RCNN [13]	50176	N/A	1.7 (GPU)
RPN+BF [14]	37632	>1 day	2 (GPU)
Ours	7316	30 min	3.4

5. CONCLUSION

We extend statistical analysis from [11] to each channel and use both the average pixel value and the variance as prior information to design local structure templates and features. The features that represent the co-existence of strong response in non-neighboring regions is also generated to enrich the feature pool. Evaluation on INRIA and Caltech benchmark shows that our method performs well with a small feature size and fast speed on CPU implementation.

6. REFERENCES

[1] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239-1258, May 2009.

[2] M. Enzweiler, and D. B. Gavrila, "Monocular pedestrian detection: survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179-2195, Dec. 2009.

[3] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned," in *Proc. Eur. Conf. Computer Vision*, pp. 613-627, Sep. 2014.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 886-893, Jun. 2005.

[5] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. British Machine Vision Conference*, pp. 99.1-99.11, Sep. 2009.

[6] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fastest feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532-1545, Aug. 2014.

[7] R. Benenson, M. Mathias, T. Tuytelaars, and L. V. Gool, "Seeking the strongest rigid detector," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3666-3673, Jun. 2013.

[8] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1751-1760, Jun. 2015.

[9] W. Nam and P. Dollar, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural information processing systems (NIPS)*, pp. 424-432, 2014.

[10] J. Cao, Y. Pang, and X. Li, "Pedestrian detection inspired by appearance constancy and shape symmetry," *IEEE Trans. Image Processing*, vol. 25, no. 12, pp. 5538-5551, Dec. 2016.

[11] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed Haar-like features improve pedestrian detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 947-954, Jun. 2014.

[12] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos. "Learning complexity-aware cascades for deep pedestrian detection," in *IEEE International Conference on Computer Vision*, Dec. 2015.

[13] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. "Scale-aware Fast R-CNN for Pedestrian Detection," arXiv:1510.08160, 2015.

[14] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. " Is Faster R-CNN doing well for pedestrian detection?," arXiv:1607.07032, 2016.

[15] [1] Y. Lecun, L. Bottou, and Y. Bengio, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp 2278 - 2324, Nov. 1998.

[16] [2] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural information processing systems (NIPS)*, pp 1097-1105, 2012.

[17] R. Appel, T. Fuchs, P. Dollár, and P. Perona, "Quickly boosting decision trees-pruning underachieving features early," in *Proc. Intl. Conf. Machine Learning*, pp. 594-602, 2013.

[18] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743-761, Apr. 2012.

[19] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, "How far are we from solving pedestrian detection" in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2016.