Information Systems and Signal Processing Journal (2016) 1: 7-11
Clausius Scientific Press, Canada

Pedestrian Detection Based on Informed Haar-like Features and Switchable Deep Network

Gu Lingkang
College of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China
glk_81@163.com

Keywords: Haar-like features, feature extraction, pedestrian detection, Switchable Deep Network

Abstract: As pedestrians usually appear up-right in image or video data, we therefore employ a statistical model of the up-right human body where the head, the upper body, and the lower body are treated as three distinct components. As we incorporate different kinds of low-level measurements, the resulting multi-modal & multi-channel Haar-like features represent characteristic differences between parts of the human body yet are robust against variations in clothing or environmental settings. Then we use a Switchable Deep Network(SDN) for pedestrian detection. The SDN automatically learns features of different body parts. Experimental results on many pedestrian datasets show that the proposed algorithm significantly improves the detection rates at 0.1FPPI compared with the state-of-the-art domain adaptation methods and that it is robust and accurate against cluttered dynamical background, occlusion and the object deformation.

1. Introduction

Pedestrian detection is a challenging task of great interest in computer vision. Pedestrian detection is an important topic in computer vision [1]. Significant progress has been achieved in recent years [2]. But this problem is particularly challenging because pedestrian images undergo large variations of visual appearance due to the changes of poses, viewpoints, clothing, lighting, and resolutions. Background clutters in a detection window also confuse the detectors.

Many pedestrian detectors have been developed to address these challenges. They extract manually designed features, such as HOG[3] and Haar-like descriptors [4] or their combinations [5], from images, and then employ classifiers such as boosting [6], SVM [3], and structure SVM [7] to decide whether a detection window should be classified as a pedestrian. In order to handle more complex and larger variations, a mixture of templates is learned for each body part [8]. Such templates (e.g., poselets [8]) are learned through clustering pose annotations and region appearance.

Over the last decade, the question of how to detect pedestrians in images has been thoroughly investigated [9]. A noticeable trend in this domain is that researchers increasingly rely on huge feature pools and high dimensional feature vectors since it is commonly believed that more features integrate more information and thus lead to better performances. As a consequence, many recent approaches rely on the availability of powerful computers and GPU computation in order to be capable of real-time detection. Also, aspects due to the peculiar geometry of high dimensional spaces, *e.g*. concentration of measure and neighborliness, appear to be disregarded[10].

2. The algorithm flow

In this paper, we propose more compact features which simultaneously ensure effectiveness and efficiency. In particular, we argue that by incorporating prior information as to the appearance of the up-right human body, one can design reasonable features for pedestrian detection. In fact, from the point of view of visual perception, pedestrians form a class of high intra-class similarity. This is because strong regularities of up-right body shapes limit how pedestrians may appear in image data. In particular the head-shoulder area of the human body shows a geometry seldom found among other natural objects. Based on a careful exploration of these characteristics, we design new features that enable efficient, state-of-the-art pedestrian detection.

Our approach is motivated by prior work on detecting objects of rather low intra-class variability. In particular, HOG and cascaded Haar-like features have become the de-facto methods of choice in this area. Yet, we note that corresponding features are either determined by means of exhaustive searches over all possible variations or by means of less exhaustive random sampling. In this paper, we propose a method that marks a middle ground; we design compact, discriminative Haar-like features selected from a particular template pool that reflects prior information about the pedestrian up-right body shape. Extensive experiments indicate that these features are highly characteristic and therefore enable very robust detection.

In recent years, deep learning has been applied to pedestrian detection and achieved promising results.Instead of using handcrafted features, it can automatically learn features in an unsupervised or supervised fashion, such as restricted Boltzmann machine(RBM), and discriminative RBM. They are often stacked into multiple layers so as to map the raw data into gradually higher-level representations. Then, the entire network is fine-tuned with label information and the top layer output is often adopted as features to train classifiers. However, the hierarchical representations learned by deep models do not have semantic meanings as in previous hierarchical deformable part-based models. Ouyang and Wang extend DPM to a deep model by learning feature representations and jointly optimizing the key components of DPM. However, they did not explicitly model mixture of templates for each body parts as in and did not depress the influence of background clutters.

We propose a novel Switchable Deep Network (SDN) for pedestrian detection. The SDN automatically learns hierarchical feature representations that correspond to body parts and the whole body. The key contribution of the model is that it introduces a new Switchable Restricted Boltzmann Machine (SRBM) to explicitly model the complex mixture of visual appearance at multiple levels. SRBM is used to build switchable layers added into the hierarchy of the SDN. At each feature level, SRBM estimates saliency maps (indicating a pixel is on the background or a pedestrian) for each test sample. For instance, in the root layer, the saliency map separates background clutter from discriminative regions for pedestrian detection. In a part layer, the saliency map also helps to localize each part in the same way. In addition, our deep model learns a mixture of templates for each part to represent it in different views and poses. SRBM can infer the most appropriate template for each part or the whole body. A new generative algorithm is devised to effectively pre-train the SDN and then fine-tune it with back-propagation.

3. Informed Haar-like features

In the following, traditional Haar-like features will be referred to as binary modalities as they only carry two possible weights (+1 and −1) for different rectangles. However, this binary modality is ill suited to represent cusps or corner-like structures of the human silhouette. That is to say, that it hardly adapts to the description of the content of bounding boxes that contain three different logical components such as, head, upper body, and background. Yet, for efficient subsequent classification

we are interested in computing the difference between parts w.r.t. two of them at a time.

To integrate color and gradient information, we build a multi-channel descriptor for each cell. We consider a total of 10 different channels as it is done in the detector: 3 channels for LUV colors, 1 channel for gradient magnitude information, and 6 channels for histograms of oriented gradients. Assume we are given a template $t = (x, y, (w, h), W)$. We first count how often the weights +1 and −1 appear and denote these counts as n_{add} and n_{sub}. There are thus n_{add} additive cells and n_{sub} subtractive cells and we normalize each cells weight by the total number of corresponding cells covered by a rectangle. This results in an average weight matrix:

$$Wavg = sgn(W)/n_{add} + sgn(-W)/n_{sub}. \tag{1}$$

The feature value of any template t for any channel k,*e.g.* color or gradient information, can then be computed as a weighted sum:

$$f(t,k) = \sum_{i=1}^{h} \sum_{j=1}^{w} \sigma(x+i, y+j, k) W_{avg}(i, j) \tag{2}$$

where, $\sigma(i, j, k)$ denotes the sum of values in $cell(i, j)$along channel k.

4. Switchable Deep Network (SDN)

4.1 Switchable Restricted Boltzmann Machine(SRBM)

We employ both the input data and the labels as observed variables other than only using the data as in RBM, because supervised information can improve classification performance. The energy function is formulated as:

$$\begin{aligned} E(x, y, h, s, m; \Theta) = -\sum_{k=1}^{K} s_k h_k^T (W_k (x \circ m_k) + b_k) - \\ \sum_{k=1}^{K} s_k c_k^T (x \circ m_k) - y^T U \sum_{k=1}^{K} s_k h_k - d^T y \end{aligned} \tag{3}$$

In which K indicates the number of components in the mixture and $\Theta = \{W, b, c, U, d\}$, where U is a fully-connect weight matrix to transform the features to labels and d is the bias vector of the label. The switch variable $s_k \in [0,1], \sum_{k=1}^{K} s_k = 1$ indicates which component is activated.

4.2 SND

We stack a convolutional layer, four switchable layers(that is, modeled with SRBM), and one logistic regression layer into the SDN for pedestrian detection. The convolutional layer learns to extract low- and mid-level features, the switchable layers model high-level mixture representations and salience maps of the entire body and different body parts (head-shoulder, upper-body, and lower-body), and the logistic regression layer predicts labels. This architecture is designed for pedestrian detection. More layers can be added to handle more complex object hierarchies.

The input image data x^0 have six channels, each of which is in the size of 108×36. The first three channels are obtained by resizing the bounding box centered on the pedestrian with three different scales and then extract the Y-channels of these three images in the YUV color spaces. The last three channels are the edge maps of the first three channels by using Sober edge detector. This is to encourage the SDN to learn features with multi-scales and boundary cues.

The convolutional layer outputs 64 channels by learning 64 filters, each with a size of 9×9×6. This layer can be formulated as below:

$$x_j^1 = \tanh^{abs}\left(\sum_{i=1}^{I} W_i^1 * x_i^0 + b_j^1\right) \tag{4}$$

where $\tanh^{abs}(\cdot) = |\tanh(\cdot)|$ is the absolute values of the hyperbolic tangent function, * indicates convolution, and i = 1…6 and j = 1…64 are the indicates of the input and output channels, respectively. W^1 and b^1 are the filter matrixes and bias vector. The output x^1 are then sub-sampled by a max pooling layer to obtain more compact representation.

5. Experiments and results analysis

In this paper the algorithm program runs in Pentium (R) CPU 987 dual core 1.5 GHz, 64 Windows 7 operating system, 4G memory on the computer. Fig.1 is part pedestrian sample figure, and Fig.2 is the result of pedestrian detection.

Fig.1 part pedestrian sample figure

Fig. 2 Schematic diagram of pedestrian detection results

The average detection rate of the method of Haar+AdaBoost in the reference[4] is 96.72%, the method of HOG+Haar+AdaBoost is 97.80%, the method of HOG+IKSVM is 96.19%, the method of FBP-CNN is 97.52%, and the method of DLSSC is 98.34%. The results indicated that the average detection rate of the proposed algorithm in this paper is 99.03%.In these algorithms, the detection rate of this algorithm is better. Furthermore, we analysis the missing rate and false detection rate of these algorithms, and take 0.1 FPPI as the reference point. The result is shown in Fig. 3.

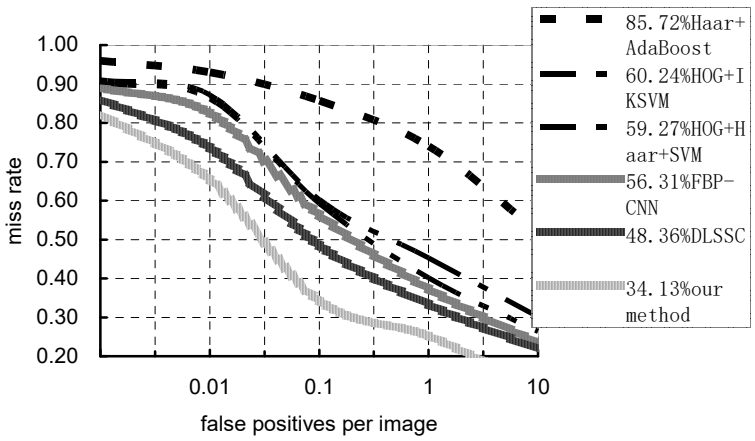

Fig. 3 Overall performance of different detectors

6. Conclusions

The experimental results show that, this method's detection precision compared to the common pedestrian detection algorithm, is obviously improved, at the same time, also improves the real-time performance, so that the overall performance of pedestrian detection system is greatly improved. In general, out method provides an excellent balance between high detection accuracy and time efficiency both at training and test time.

Acknowledgements

This work was supported by the project "Anhui province higher education to enhance the general project plan of Provincial Natural Science Research (NO.TSKJ2014B11)". The authors wish to thank the Education Department of Anhui Province for their help.

References

[1] Gu Lingkang. Fast pedestrian detection based on feature of local model. Journal of Computational Methods in Sciences and Engineering, v 15, n 3, p 387-393, August 3, 2015.

[2] K. Sohn, G. Zhou, C. Lee, and H. Lee. Learning and selecting features jointly with point-wise gated Boltzmann machines.ICML, 2013.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2015.

[4] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. IJCV, 63(2):153–161, 2005.

[5] P. Luo, L. Lin, and H. Chao. Learning shape detector by quantizing curve segments with multiple distance metrics.ECCV, 2010.

[6] P. Doll´ar, Z. Tu, P. Perona, and S. Belongie. Integral channel features. BMVC, 2009.

[7] X. Wang, X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. CVPR, 2009.

[8] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In ICCV, 2009.

[9] P. Doll′ar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: an evaluation of the state of the art. IEEE Trans.PAMI, 34(4):743–761, 2011. 1, 5

[10] Shanshan Zhang, Christian Bauckhage, Armin B. Cremers. Informed Haar-like Features Improve Pedestrian Detection. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, 126, 947-954.

[11] J. Gall, A. Yao, N. Razavi, L. V. Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. IEEE Trans. PAMI, 33(11):2188–2202, 2011. 2