PEDESTRIAN DETECTION VIA LOGISTIC MULTIPLE INSTANCE BOOSTING

Junbiao Pang[*,†,‡], *Qingming Huang*[*,†,‡], *Shuqiang Jiang*[†,‡], *Wen Gao*[*,§]

[*]Graduate School of Chinese Academy of Sciences, Beijing,100190, China
[†]Key Lab. of Intelligent Information Processing, Chinese Academy of Sciences(CAS)
[‡]Institute of Computing Technology, CAS, Beijing 100190, China
[§]Institute of Digital Media, Peking University, Beijing, 100190, China
{jbpang | qmhuang | sqjiang | wgao}@jdl.ac.cn

ABSTRACT

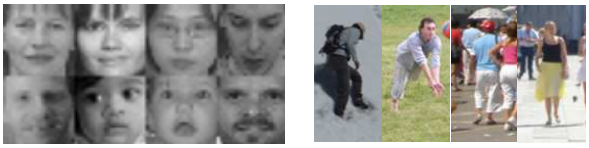
Pedestrian detection in still image should handle the large appearance and pose variations arising from the articulated structure and various clothing of human bodies as well as view points. So it is difficult to design effective classifier for this problem. In this paper, we address these variations in detection via multiple instance learning, specifically logistic multiple instance boosting (LMIB). In LMIB, a example is represented as a set of instances, which implicitly encode the variations. Giving different confidence to the instances in a bag, the LMIB will automatically reduce the influence of the variations at training stage. To obtain rapid detection speed, the LMIBs are grouped into the cascaded structure. The proposed detection algorithm is tested on MIT and INRIA human datasets where promising detection results are comparable with the baseline algorithms.

Index Terms— pedestrian detection, multiple instance learning, boosting, object detection, machine learning

1. INTRODUCTION

When machine learning is used for object detection, the positive examples should be well normalized for training. For instance, the face examples illustrated in Fig. 1(a) are approximately aligned according to eyes in face detection. However, in pedestrian detection, the effective normalization is ill-posed. Compared with the face examples, pedestrian examples have the large appearance and pose variations which are caused by the articulated structure and variable clothing of human bodies. Some examples shown in Fig. 1(b) describe the problem. Moreover, the test sets have also large discrepancy with training examples (Compared with Fig. 1(a), some detection results in Fig. 7 partly illustrate the discrepancy). Pedestrian detection is considered among the hardest examples of object detection problems [1].

This work was supported in part by National Science Foundation of China under Grant 60773136 and 60702035, and in part by National Hi-Tech Development Program (863 Program) of China under Grant 2006AA01Z117 and 2006AA01Z320.



(a) (b)

Fig. 1. Training samples for different object detection. (a) the aligned frontal faces have little appearance and pose variations; and (b) the pedestrians have large variations, although the samples are already aligned according to shoulders.

Several existing publications have been aware of the appearance and pose variations, and handle it by"divide and conquer" [7, 10, 11]. In [10], body parts are represented by co-occurrences of local orientation features, and detectors are trained separately for each part using Adaboost. Human location is determined by maximizing the joint likelihood of part occurrences according to the geometric relation. In [11], local appearance feature and their geometric relation are combined with global cues by segmentation based on per pixel likelihoods. However, "divide and conquer" approaches have two drawbacks. First, different detectors have to be applied to the same image patch. This will reduce the detection speed. Second, labeling and aligning the local parts are tedious and time-costing work.

Inspired by face detection [2], we handle the variations by explicitly acknowledging that the object detection is a Multiple Instance Learning (MIL) [8, 3] problem. In MIL, training examples are not singletons; instead, they come in "bag", where all of the instances in a bag share a single label. A positive bag means that at least one instance in the bag is positive, while a negative bag means that all instances in the bag are negative. Intuitively objects are located in some region of the image, but the exact position of object is not known. Therefore, the training object can be represented as a bag having a set of instances, and the "well" aligned instances are expected to be located by MIL to train a bag-level classifier.

In pedestrian detection setting, A pedestrian example is

represented as a bag of instances, which do not need to be well normalized, but to be a set of instances with large variations between each other. In training stage, MIL learner can automatically give high confidence to "well" aligned instances, and train a bag-level classifier. During detection, if one of the instances occurs, the pedestrian is located. The seemingly most similar work to ours is [2]. In [2], multiple instance learning, noise-OR boosting is used for face detection in teleconfencing setting. Here, we introduce the MIL into a more challengeable problem: pedestrian detection. The main contributions of this paper are two folds:

- Multiple instance learning is first introduced into pedestrian detection for handling the large variations between training examples (alignment problem). Specifically, logistic multiple instance boosting (LMIB) [3] is exploited to learn the non-aligned pattern of pedestrian.
- Considering the pedestrian detection problem, a method to generate individual instances is proposed.

Experimental results show that by learning with MIL framework, the detector significantly outperforms Adaboost [5] with strong discriminative histogram of oriented gradients (HOG) feature [4], and slightly outperforms the kernel Support Vector Machine(SVM) [4] with HOG feature.

2. LOGISTIC MULTIPLE INSTANCE BOOSTING

In detection, the appearance and pose variations are represented within each instances and within the uncertain instance labels in multiple instance learning. For pedestrian detection, most person are standing stance. Therefore, we propose one method to generate the instances: the instance window is shifted around the body. The created instances can take advantage of all information of the "omega shape" of heads and the rectangle shape of bodies in Fig. 2(a). During training, "well" aligned instances is automatically given higher bag-level class confidence.

Compared with supervised learning, an instance x_{ij} is indexed with two indices: i which indexes the bag, and j which indexes the instance within the bag. Here, we assume that all instances contribute equally and independently to a bag's class label. Given a bag $\mathbf{x}_i$, the probability of the bag-level class label $\mathbf{y}_i$ is given by

$$p(\mathbf{y}_i|\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} p(y_{ij}|x_{ij}) \tag{1}$$

where n_i is the number of instances in the i-th bag. y_{ij} is the instance-level class label. The instance-level class probability is given as $p(y|x) = 1/(1+e^{\beta x})$, where β is the parameter to be estimated. Controlling β value gives different instance-level probability, which contributes different confidence to bag-level probability. Ideally, "well" aligned instances should be given higher probability than the non-aligned. Based on Eq. (1), the parameter β can be estimated by maximizing the bag-level binomial log-likelihood function

$$L = \sum_{i}^{N} [\mathbf{y}\log p(\mathbf{y}=1|\mathbf{x}) + (1-\mathbf{y})\log p(\mathbf{y}=0|\mathbf{x})] \tag{2}$$

Eq. (2) can not be solved analytically. Xu el al [3] propose an boosting method to maximize the log-likelihood function. Given a collection of N $i.i.d$ bags $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and every bag $\mathbf{x}_i$ having $x_{i1}, \ldots, x_{ij}$ instances, we need to learn a bag-level function $\mathbf{F}(\mathbf{x}) = \sum_m c_m \mathbf{f}_m(\mathbf{x}), m = 1, \ldots, M$ and the corresponding *strong* classifier $\mathbf{H} = sign(\mathbf{F}(\mathbf{x}))$. The parameters $c_1, \ldots, c_M \in \mathbb{R}$, and the $\mathbf{f}$ is the bag-level weak hypothesis. The empirical loss

$$E[I(\mathbf{F}(\mathbf{x}) \neq \mathbf{y})] = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i \mathbf{F}(\mathbf{x}_i) \tag{3}$$

where $I(\cdot)$ is the indicator function. We interest in wrapper the bag-level weak hypothesis $\mathbf{f}$ using the instance-level weak hypothesis f. Combining Eq. (1), Eq. (3) is converted into the instance-level's exponential loss $E_{\mathbf{x}}E_{\mathbf{y}|\mathbf{x}}[e^{-yf}]$ as $e^{-\mathbf{yf}} \geq I(\mathbf{H}(\mathbf{x}) \neq \mathbf{y}), \forall M$. One searches for the optimal updating $c_m f_m$ to minimize

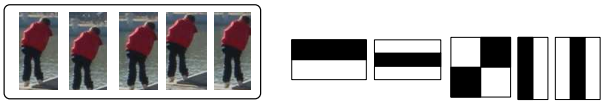
$$E_{\mathbf{x}}E_{\mathbf{y}|\mathbf{x}}\left[e^{-y_j F_{m-1}(x_{ij}) - c_m y_j f_m(x_{ij})}\right] = \sum_i w_i e^{[(2\epsilon_i - 1)c_m]} \tag{4}$$

where error $\epsilon_i = \sum_j 1_{f_m(x_{ij}) \neq y_i}/n_i$. The ϵ_i indicates the discrepancy between the bag label and instance label.

The result of the LMIB is not only a bag-level classifier, but also a set of the instance-level classifier. The instances in positive bags with higher scores $f(x_{ij})$ give higher confidence to the bag's label, even there are some negative instances (to detection, non-aligned examples) occurring in positive bags. Therefore, the final classifier decides these bags as positive. The discrepancy in training bags will be automatically reduced. The LMIB is summarized in Algorithm. 1.

Algorithnm. 1 Logistic Multiple Instance Boosting

1: Initialize weight of each bag $W_i = 1/N, i = 1, 2, \ldots, N$.
2: **for** $m = 1$ to M **do**
3: Set instance weights $w_{ij} \leftarrow W_i/n_i$, and find an instance-level weak classifier f_m which minimizes Eq. (4).
4: Calculate the i-th bag's ϵ_i.
5: **If** $\epsilon_i < 0.5$ **for** all i, **go to** 8.
6: Compute $c_m = argmin \sum_j W_j e^{((2\epsilon_i - 1)c_m)}$.
If $c_m \leq 0$, **go to** 8.
7: Update weight $W_j \leftarrow W_j e^{((2\epsilon_i - 1)c_m)}$
8: **end for**
9: return $\mathbf{H} = sign\left(\sum_j \sum_m c_m f_m(x_j)\right)$



(a) A positive bag (b) 5 type rectangle feature

Fig. 2. Some instances in one positive bag and 5 type feature.

3. PEDESTRIAN DETECTION

To achieve the fast detection speed, we adopt the cascade structure of detector [9] shown in Fig. 3. Each level is designed to achieve high detection rate and modest false positive rate. Let N_{pi} and N_{ni} be the number of positive and negative training examples at i stage. Considering the influence of asymmetric training data on the classifier and computer RAM limitations, we constrain N_{pi} and N_{ni} to be approximately equal.

Assuming that we are training the k-th stage, we classify all the possible detection windows on the negative training images with the cascade of the previous $(k-1)$ LMIB classifier. The examples which are misclassified form the possible negative training set. The positive training examples do not change during bootstrap.

According to "There is no free lunch" theorem, it is very important to choose suitable number of the instances for training and detection. The more instances in a bag will reduce more variations and improve the detection accuracy, but decrease the training and detection speed. We experimentally set 5 instances for training and 3 instances for detection.

To test the power of the LMIB, efficient descriptor HOG is not taken into the detection framework. On the contrary, five type weak discriminative Haar rectangle feature [9] is used as descriptor in Fig. 2(b). During training, 3617 weak feature are designed for each instance normalized as 128×64 pixel. We build the 30 stage cascade detector. Each level of cascade classifier is optimized to correctly detect at least 95% of the positive bags, while reject at least 50% of the negative bags. Here 99.9% hit rate is not adopted for two reasons: (1) the rectangle feature are too weak discriminative to obtain better classification ability; (2) some bags is hard to be classified correctly at the cost of higher false positive rate.

4. EXPERIMENTAL RESULTS

We evaluate our algorithm by experimenting on two different datasets. One is the MIT pedestrian datasets [7, 6], a commonly used database for evaluation of pedestrian systems. Another is INRIA dataset [4]. Note that MIT dataset only has the upright, frontal and back viewpoint pedestrian images. This set only contains 923 pedestrian examples and does not contain a negative set. The positive samples are also not separated into training and testing set. We use 600 of them as positive training set and the left for testing, and negative examples supplied in INRIA dataset are used as negative training set. In

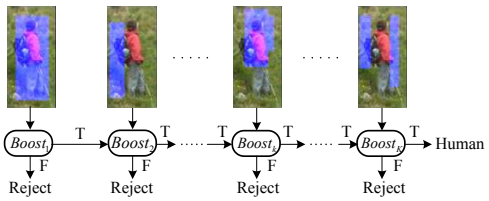


Fig. 3. The cascade of LMIB detector. The k-th classifier selects the Haar rectangle feature (blue region represents it).

Fig. 4, we plot the detection error tradeoff curves on a log-log scale [4] as evaluation criterion. The y-axis corresponds to the miss rate, and the x-axis corresponds to false positives per window (FPPW). The curve for our method is generated by adding one cascade level at a time. As illustrated in Fig. 4, our approach achieves near zero false negative results.

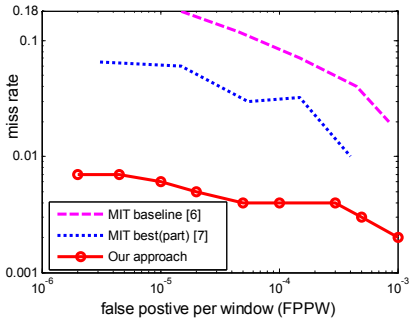


Fig. 4. Comparative results on MIT dataset.

We perform the comparative experiments on another more challenging dataset, the INRIA dataset [4]. The database contains 1239 pedestrian images (2478 with their left-right reflections) and 1218 person-free images for training. In the test set, there are 566 images containing pedestrian. In Fig. 5, we perform the same separation of training-testing sets to directly compare the results [4, 5, 9]. Note that motion patterns in [9] are ignored for still image, and we consider both the kernel and linear SVM method of [4]. L2-norm in HOG feature, the best performance, is only considered. Our method significantly improves the classification ability of rectangle feature [9]. Although the weak discriminative rectangle feature is utilized, our method achieves better results than Zhu et al [5]. Multiple instance boosting can handle more variations and brings more discriminative ability for detector than Adaboost. The detection results can be compared with the kernel SVM. If we consider the 10^{-4} as an acceptable FPPW, our miss rate is 9.4%, while the kernel SVM is 9.3%. However, kernel SVM is significantly computationally expensive. Compared with our work, [4, 1] all focus on developing stronger feature, while our work focuses on the classifier. These are two complementary directions.

We scan the image at 0.8 scale and 2 pixel step. In Fig. 6,

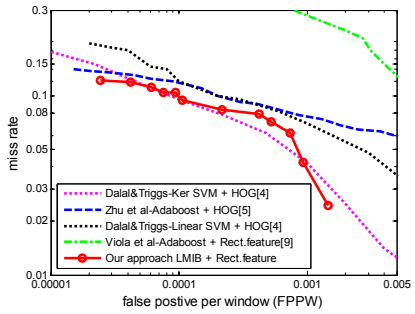


Fig. 5. Comparative results on INRIA dataset. The curves for approach [4, 5] are generated from the respective papers.

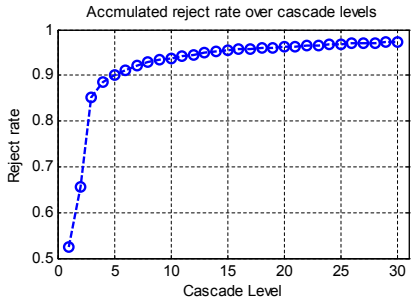


Fig. 6. the accumulated rejection rate over the cascade level.

we plot the accumulated rejection rate over the cascade levels. The first five levels reject 90% of the negative windows (without pedestrian). A total of 620 weak classifiers are used in 30 cascade levels. On average our method requires to evaluate 55.07 rectangle features per negative detection windows. Scanning a 320×240 image needs average 100 ms under PC with 2.8 GHz CPU and 512 RAM, while 250 ms for 320×240 image is reported in Zhu et al [5].

In Fig. 7, several detection results are shown for different scenes with human having variable illumination, appearance and stance. Although there are still few false positive, strong discriminative descriptors will reduce more false positive.

5. CONCLUSION AND DISCUSSIONS

We present a new approach to solve the large appearance and pose variations in pedestrian detection utilizing multiple instance learning, specifically LMIB. The training examples do not need to be well aligned, but to be represented as a set of instances. During training, the variations are automatically reduced with multiple instance learning. For detection, the LMIBs are builded into cascade detector to achieve rapid detection speed. The promising performance is shown on INRIA and MIT datasets.

Although we show that multiple instance learning has achieved good performance, weighted average instance label

Fig. 7. Some detection results on INRIA dataset. Note that significantly overlapping detection windows are averaged into a single window.

may be unsuitable for detection, when only one instance is used for detection. Better method to model the bag label can enhance the power of multiple instance learning. In future, strong discriminative feature, such as, HOG will be exploited to reduce the number of average features per negative detection window and to reduce false positive.

ACKNOWLEDGEMENT

This work was supported by "Science100 plan" of China Academy of Sciences under Grant 99T3002T03. We would also thank the anonymous reviewers for their valuable comments.

6. REFERENCES

[1] O.Tuzel, F.Porikli and P.Meer. Human detection via classification on riemannian manifolds. In CVPR 07.

[2] P. Viola, J. C. Platt, and C. Zhang. Multiple instance boosting for object detection. In NIPS, 2006.

[3] X. Xu and E. Frank, Logistic regression and boosting for labeled bags of instances, In PAKDD, pages 272-281, 2004.

[4] N. Dalal and B.Triggs, Histograms of oriented gradients for human detection, In CVPR,pages 886-893, 2005

[5] Q. Zhu, S. Avidan, M.C. Yeh, and K.T. Cheng. Fast human detection using a cascade of histogram of oriented gradients, In CVPR,pages 1491-1498, 2006.

[6] P. Papageorgiou and T. Poggio, A trainable system for object detection, IJCV, pages 15-33, 2000

[7] A. Monhan and C. Papageorgiou and T. Poggio, Example-based object detection in images by components, IEEE Trans. PAMI, vol. 23, pages 349-360, 2001

[8] O. Mahon and T. Lozanno-Perez, A framework for multiple-instance learning, In NIPS, pages 570-576, 1998.

[9] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance, In ICCV, 2003.

[10] K. Mikolajczyk, C. Schmid, and A.Zisserman, Human detection based on a probabilistic assembly of robust part detectors, In ECCV, 2004.

[11] B.leibe, E. Seemann, and B. Schiele, Pedestrian detection in crowed scenes. In CVPR, page 878-885, 2005.