Reduce False Positives for Human Detection by a Priori Probability in Videos

Lei Wang, Xu Zhao, Yuncai Liu
Key Laboratory of System Control and Information Processing
Department of Automation, Shanghai Jiao Tong University
800 Dongchuan RD, Shanghai 200240, China
{wltongxing,zhaoxu,whomliu}@sjtu.edu.cn

Abstract

In this work, we address the problem of reducing the false positives for human detection in videos. We employ the motion cue to build a foreground probability model. Then the mean expectation of the pixel-level foreground probability is computed to assign a priori probability to the sliding window in detection. We combine the response of Deformable Part Models and the mean probability expectation to form the features and train a linear classifier. The proposed approach is threshold-free, and reduces the false positives in human detection by the foreground cues. As well, we describe an integral probability image for fast computation of the mean probability expectation. Experimental results show that the proposed method achieve superior performance over the baseline of Deformable Part Models.

1. Introduction

Human detection is an important research in computer vision area. The target is to locate all the people in still images or videos. It's applications include surveillance, human-computer interface, robotics, monitoring of the elderly and disabled, entertainment and content-based retrieval. The problem is challenging due to the variation of human appearance and pose, image quality and the occlusion. Thus human detection still remains active in recent years. Numerous approaches have been proposed to work on features, cope with human parts and pose, and improve the classification and learning framework.

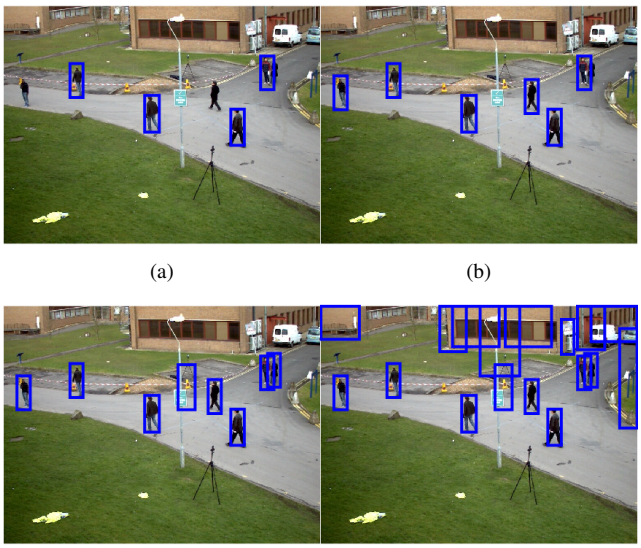
Many features have been proposed for human detection. Based on Haar wavelets [15], Viola and Jones [18] utilized integral image for fast feature computation and presented automatic feature selection. Dalal and Triggs [4] proposed the Histogram of Oriented Gradient (HOG) features for detection. Gavrila [11] presented a hierarchical shape-based object representation and a coarse-to-fine approach for fast matching. Dalal *et al*. [5] built a motion histogram based on differential optical flow, which is combined with appearance HOG for human detection in video. Wang *et al*. [19] employed a texture descriptor by combining Local Binary Patterns (LBP) and HOG. Dollár *et al*. [7] proposed the multiple-channel features where Haar-like features are computed over multiple channels of image.

To deal with the articulation in modeling human, the representations based on parts and pose have been investigated. Mohan *et al*. [14] proposed a method to train detectors for head, arm and leg in a supervised manner, and then combined the results of the component detectors. Several approaches are proposed to jointly detect humans and estimate the poses. Yang and Ramanan [20] augmented the standard pictorial structure models and described a flexible mixture model to capture the spacial relations between parts. Sun and Savarese [17] proposed an Articulated Part-based Model which represents human as a set of parts with multi-level details. In [12] Multiple Instance Learning was employed to mitigate the feature misalignment caused by pose variation and part deformation. Felzenszwalb *et al*. [9] proposed the Deformable Part Models (DPM) with the part positions as latent variables, and date-mined hard negative examples.

Another major component for human detection is the classifier. Boosting method is the basis of the detection framework of Viola and Jones [18]. A weak classier is trained in each round and in the meantime the feature is selected. The final decision is determined by the strong classifier formed by the weighted sum of all weak classifiers. Linear Support Vector Machine (SVM) was used as the classifier for HOG-LBP features in [19]. Maji *et al*. [13] approximated the histogram intersection kernel for SVMs, allowing for faster computation in the sliding window framework of detection. In [9] latent SVM was used to mine the hard examples with latent part positions, and also an iterative training algorithm was proposed to deal with the semi-convex problem of latent SVM.

Dollár *et al*. [8] presented an extensive evaluation of the state of the art for human detection. In sliding window detection, many approaches calculate a response for each window candidate, and a threshold is set to determine the final

978-1-4799-6100-9/15/$31.00 ©2015 IEEE



(a) (b) (c) (d)

Figure 1: Results of human detection on Pets2009 dataset by DPM under different thresholds. The thresholds are 0, -0.5, -1, -1.5 for (a), (b), (c) and (d), respectively.

results. Figure 1 shows the detection results by the approach of DPM under different thresholds. Lower threshold allows for lower miss rate, but results in more false positives. In real applications, threshold plays an important role, but few methods have presented instructions to select the threshold to achieve the best performance.

In this work, we propose a threshold-free approach for human detection. Our model is built in the framework of DPM, but it can also incorporated with other methods by slight modification. In the sliding window detection, each window candidate is classified as "person" or "not person" without a priori hypothesis. In fact, some windows are more likely to have a person. For example, in surveillance videos, the background is usually static, and the foreground (*e.g.* person, car) is dynamic or moving. Inspired by this, we want to reduce the false positives by a priori probability of foreground.

The contributions are as follows. i) We build a foreground probability model, and employ it to reduce the false positives in human detection. ii) Our approach is threshold-free, and achieves better performance than the baseline.

2. Method

This approach is formulated in the framework of DPM [9] which also serves as the baseline method for comparison. We train a linear classifier based on the DPM response and the average expectation of foreground probability. To cope with different sliding windows, we normalized the average expectation of foreground probability with respect to the window size. As well, we presented an "integral probability image" for fast computation, similar to [18].

2.1. Model response

For completeness, the brief description of DPM is as follows. Let a model have m components, and each has n part filters and one root filter of width w_c and height h_c, $c = 1, ..., m$. H is a pyramid of HOG features, and $p = (x, y, l)$ defines a position (x, y) in the l-th level of H. An object hypothesis is $z = (p_0, ..., p_n)$ which specifies the locations of all the filters of a model component, where p_0 is for the root filter. The score of object hypothesis z based on model component c is given by scores of all the filters minus the deformation cost:

$$\text{score}^c(p_0, ..., p_n) = \sum_{i=0}^{n} F_i' \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot \phi_d(dx_i, dy_i) + b, \quad (1)$$

where $\phi(H, p_i)$ is the feature vector at p of H, F_i' is the filter vector, $\phi_d(dx_i, dy_i)$ is the deformation feature vector, d_i is the coefficient of deformation cost, and b is the bias.

The overall score for each component is obtained according to the best possible placement of the parts. For a mixture model, the object location is defined by the highest score across all the components,

$$\text{score}(p_0) = \max_c \text{score}^c(p_0) \quad (2)$$

$$= \max_c \{ \max_{p_1,...,p_n} \text{score}^c(p_0, ..., p_n) \}. \quad (3)$$

2.2. Foreground probability model

We employ Bernoulli distribution [6] to model the pixel-level probability for background and foreground, since only two choices are involved. A pixel variable of the distribution takes value 1 for foreground with probability p_b and value 0 for background with probability $1 - p_b$.

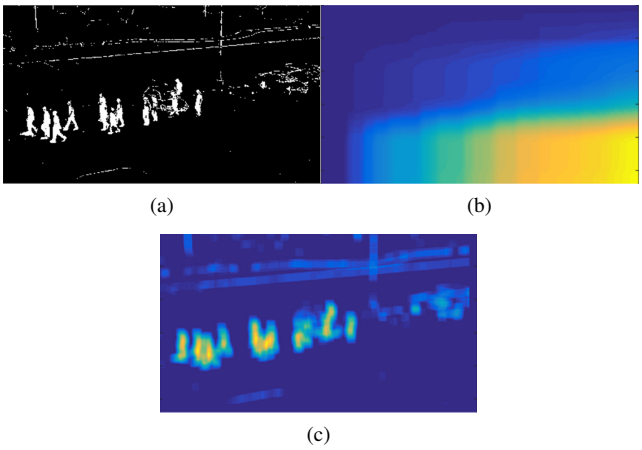
Given a sliding window W_d of width w and height h, the probability of the pixel variable $X^{i,j}$ is $p_b^{i,j}$, $i = 1, ..., h$, $j = 1, ..., w$. The expected value (expectation) of the Bernoulli distribution is:

$$E[X^{i,j}] = 1 \cdot p_b^{i,j} + 0 \cdot (1 - p_b^{i,j}) = p_b^{i,j}. \quad (4)$$

To represent the whole window, we use the mean expectation E_m of all pixels, and the mean expectation is normalized with respect to the window size. It is formulated:

$$E_m = \frac{1}{w \cdot h} \sum_{i=1}^{h} \sum_{j=1}^{w} E[X^{i,j}]. \quad (5)$$

In human detection, we need to compute E_m at all the locations in pyramid H for all the model components, and



(a) (b) (c)

Figure 2: Integral probability image for mean expectation. (a) is the pixel-level probability of foreground. (b) is the integral probability image, based on which the mean expectation is the calculated with window size of 50×50 in (c).

the computational cost is high. To alleviate it, we employ an integral probability image I which is the integral image [18] of the pixel-level foreground probability. The mean expectation at the location (x, y) of image for sliding widow of size (w, h) is written as:

$$E_m^{x,y}(w,h) = \frac{I_{x,y} + I_{x+w,y+h} - I_{x+w,y} - I_{x,y+h}}{w \cdot h}, \quad (6)$$

where $I_{x,y}$ is the pixel value at (x, y) in the integral probability image.

An example for integral probability image is shown in Figure 2. We can see that the mean expectation is helpful in human detection. Higher mean expectation gives rise to higher probability of human hypothesis.

2.3. Detection method

In the detection framework of DPM, the detections are decided by all the locations with the model response score(p_0) in (2) over threshold T. In different applications, we may need to set different T manually to obtain good results. Also, in the sliding window method all windows are treated equally as human candidates. In fact, some locations are more likely to be a candidate with the help of other cues (*e.g.* motion, a priori foreground probability).

We present a linear classified to combine the model response and the mean expectation of foreground probability. Some of the following notations are defined in Section 2.1. The size of the root filter of each component is (w_c, h_c). p_0 defines a root filter position (x, y) in the l-th level of pyramid H. The mean expectation for p_0 should be computed in an integral probability pyramid. Since the proposed mean expectation is invariant to scaling, we map the root filter location to the original image by scaling from l-th level. Thus we only need to compute the integral probability image for the original image. To obtain the location and size in the original image, the root filter located at p_0 is mapped by scaling:

$$x^0 = x/s_l, y^0 = y/s_l, w_c^0 = w_c/s_l, h_c^0 = h_c/s_l, \quad (7)$$

where s_l is the scaling parameter for l-th level of H.

We find that non-maximum suppression [9] plays an important role in the approach. Non-maximum suppression is applied based on the model response in (2) to eliminate repeated detections and the set S of detection candidates is obtained. Given S, the classifier is defined as:

$$D_s^{p_0} = \text{sign}\{\omega_1 \cdot \text{score}^{\tilde{c}}(p_0) + \omega_2 \cdot E_m^{x^0,y^0}(w_{\tilde{c}}^0, h_{\tilde{c}}^0) + \omega_3\}, \quad p_0 \in S, \quad (8)$$

where $\tilde{c} = \text{argmax}_c \, \text{score}^c(p_0)$. The detection results are $D_s^{p_0} \in \{1, 0, -1\}$, 1 for positives, -1 for negatives, and 0 for the undetermined. $\text{score}^c(p_0, ..., p_n)$ is defined in (1) and E_m is defined in (6). $(\omega_1, \omega_2, \omega_3)$ is the parameter of the linear classifier.

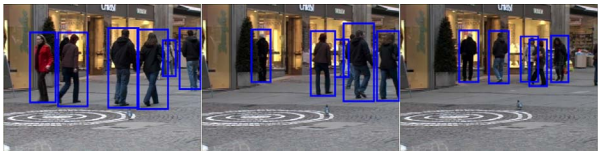
The learning procedure is divided into two parts. First, the latent SVM [9] is used to train the DPM model. Then the integral probability image is computed and the mean expectation is obtained for each model component. The parameters of the presented linear classifier is learned in the framework of the linear SVM [3]. Then the classifier is ready for human detection by the value of D_s in (8) for new images and videos.

2.4. Implementation details

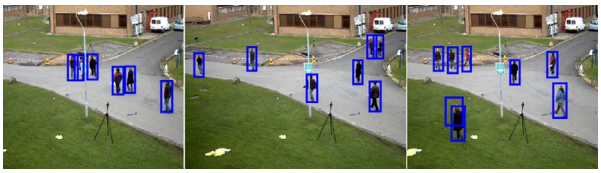
Many approaches have addressed the problem of foreground segmentation. In the probability framework, the foreground probability of each pixel is denoted as $p_b \in [0, 1]$. By considering the computation speed and detection rate, we employ the algorithm of ViBe [2] for motion detection and further foreground probability estimation. The pixel is classified as either background (static or slow moving) or foreground (moving object), so the foreground probability is formulated in a discrete form $p_b \in \{0, 1\}$. An example of pixel-level probability is shown in Figure 2a.

3. Experimental results

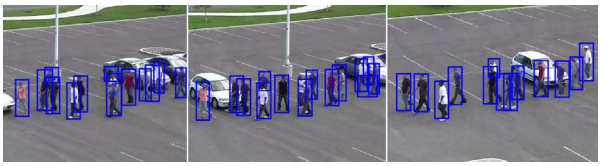
We evaluate the proposed approach on three datasets: TUD-stadtmitte [1], PETS2009 [10] S2.L1 sequence and the ParkingLot1 dataset [16]. We employ 60% of all the detection candidates from DPM as the training samples to learn the linear classifier for each dataset. As well, we evaluate the cross-dataset performance. The model trained on ParkingLot1 dataset is used to test on TUD and PETS2009 datasets.



(a)



(b)



(c)

Figure 3: Example images of human detection on datasets: (a) TUD-stadtmitte, (b) PETS2009 and (c) ParkingLot1.

We report the miss rate (MR) against false positive per image (FPPI) and the average precision (AP) by varying the detection threshold (confidence) [8]. The threshold of DMP is for the model response score(p_0) in (2), and our threshold is obtained based on the output value of the linear classifier in (8) without the "sign" operation. Note that the proposed approach does not need a threshold to get the final detection. The threshold is only used for comparison. Since we focus on the real application, we also compare the miss rate when the false positive per image is 1.

The example images of human detection by our method on the three datasets are shown in Figure 3. We can see that our method achieves high accuracy and low false positive rate. The persons in TUD-stadtmitte dataset have clear appearance, and the count of the persons is small. The PETS2009 and ParkingLot1 datasets are more challenging, especially the ParkingLot1 dataset. They have more persons and the occlusions also affect the detection.

Figure 4 shows the detailed evaluation and comparison of our method and the original DPM approach. Figure 4a, 4b and 4c show the miss rate against FPPI on TUD-stadtmitte, PETS2009 and ParkingLot1 datasets respectively. When FPPI=1, our method reduces the miss rate from 0.060 to 0.026 on TUD-stadtmitte, from 0.166 to 0.120 on PETS2009, and from 0.283 to 0.234 on ParkingLot1. The miss rate level also indicates the difficulty degree of the datasets, among which the ParkingLot1 is the most challengeable. Figure 4d, 4e and 4f show the precision-recall curve on the three datasets respectively. From Table 1, we can see that our approach improve the AP from 0.944 to 0.948 on TUD-stadtmitte, from 0.848 to 0.862 on PETS2009, and from 0.769 to 0.810 on ParkingLot1.

Method	TUD-stadtmitte	PETS2009	ParkingLot1
DPM	0.944	0.848	0.769
Ours	0.948	0.862	0.810

Table 1: Average Precision (AP) on three datasets.

For cross-dataset comparison, we detect human on TUD-stadtmitte and PETS2009 datasets using the model trained on ParkingLot1 dataset. The AP on PETS2009 is 0.862, which is same to that by the model trained on the dataset itself (Figure 4e). However, the AP on TUD-stadtmitte is 0.929, which is lower than that of DPM (0.944, Figure 4d). One reason is that PETS and ParkingLot1 datasets share a typical surveillance setting. The appearance and sizes of people in the videos are similar. Also, the foreground probability plays a key role. The humans in TUD dataset are closer to the camera, the pixel-level probability by ViBe algorithm is not as accurate as that of PETS and ParkingLot1.

4. Conclusion

In this paper, we propose a threshold-free approach to reduce the false positives in human detection by a priori foreground probability, which is obtained by the motion cue in videos. We present the mean probability expectation to model the foreground cues, and also provide an integral probability image for fast computation. The experimental results demonstrate that our approach can improve the performance of human detection.

Acknowledgement

This research has been partially supported by China 973 Program (2011CB302203) and NSFC grants (61273285, 61375019).

References

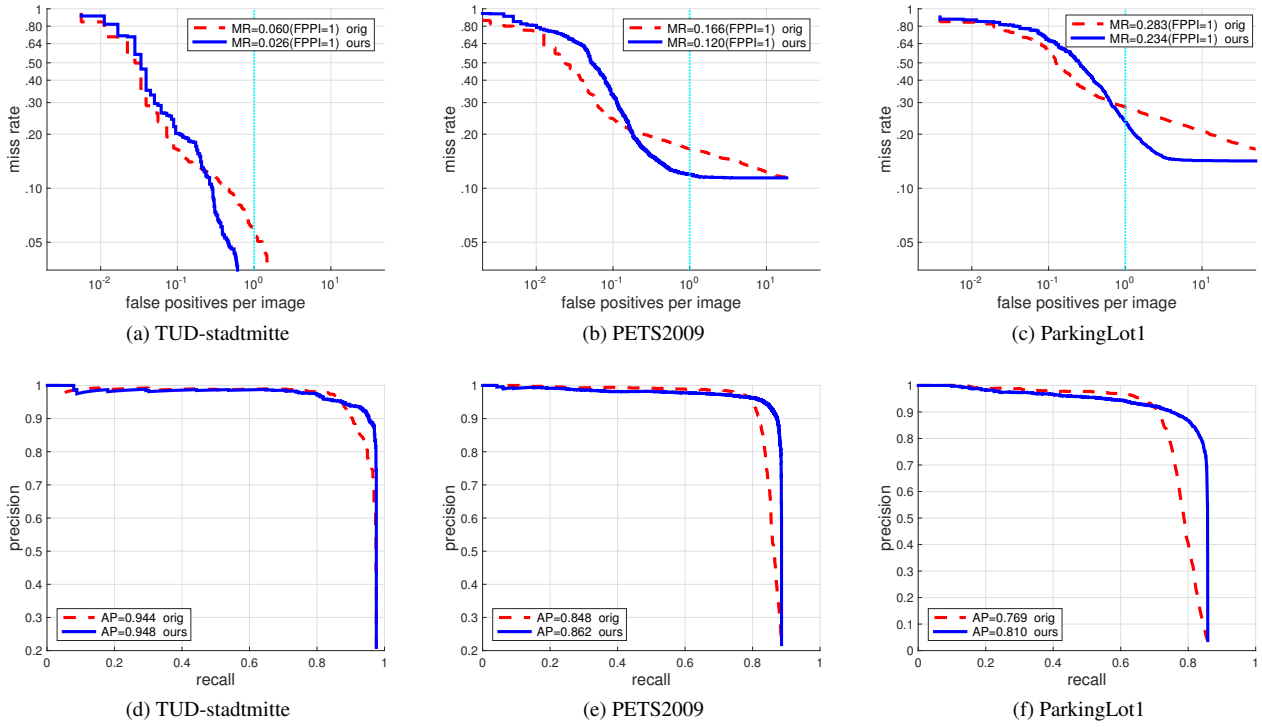
[1] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, pages 623–630. IEEE, 2010. 3

[2] O. Barnich and M. Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Trans. on Image Processing*, 20(6):1709–1724, 2011. 3

[3] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2(3):27, 2011. 3

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005. 1

[5] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, pages 428–441. Springer, 2006. 1



(a) TUD-stadtmitte (b) PETS2009 (c) ParkingLot1

(d) TUD-stadtmitte (e) PETS2009 (f) ParkingLot1

Figure 4: Comparison of recall-precision curve, average precision (AP) and miss rate (MR) against false positive per image (FPPI) on TUD-stadtmitte, PETS2009 and ParkingLot1 datasets. The upper row (a,b,c) shows the results of miss rate, and the lower row (d,e,f) shows the recall-precision curve and the average precision.

[6] Y. Dodge. Bernoulli distribution. In *The Concise Encyclopedia of Statistics*, pages 36–37. Springer New York, 2008. 2

[7] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009. 1

[8] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012. 1, 4

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 1, 2, 3

[10] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. *Winter-PETS*, 2009. 3

[11] D. M. Gavrila. A bayesian, exemplar-based approach to hierarchical shape matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(8):1408–1421, 2007. 1

[12] Z. Lin, G. Hua, and L. S. Davis. Multiple instance ffeature for robust part-based object detection. In *CVPR*, pages 405–412. IEEE, 2009. 1

[13] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, pages 1–8. IEEE, 2008. 1

[14] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001. 1

[15] C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000. 1

[16] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, pages 1815–1821. IEEE, 2012. 3

[17] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, pages 723–730. IEEE, 2011. 1

[18] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 4:51–52, 2001. 1, 2, 3

[19] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV*, pages 32–39. IEEE, 2009. 1

[20] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013. 1