# Moving Pedestrian Detection Using Normed Proposals and Key Points Matching

Chen Wei[1, a], Wang Taihong[2, b] and Cai yong[2, c]

[1] College of Information Science and Engineering, Hunan University, Changsha 410082, China;

[2] Key Laboratory for Micro-Nano Optoelectronic Devices of Ministry of Education, Hunan University, Changsha 410082, China.

[a]chenwei8650@foxmail.com, [b]512195698@qq.com, [c]15616125491@163.com

**Keywords:** pedestrian detection, region of interest, key point, proposal.

**Abstract.** Occlusion detection and automatic adaption of a generic pedestrian detector to a specific scene are difficult problems in intelligent monitoring. When a detector trained in a specific scene is applied on a new scene, its accuracy will decrease greatly. To solve this problem, we propose a new detection algorithm in which motion regions of interest based on motion information are obtained quickly by a flash-bit computing method. Also we focus on the case in which a single target converts to be a difficult one due to multiple overlapping between pedestrians. Key points with BRISK feature which computed and saved before are used to match difficult targets in occlusions. Normed proposals which proved to have higher confidence are used to correct the location and shape of detection windows, results in a five percent increasing of detection accuracy. Results of comparative experiments of five different detectors on three motion pedestrian datasets show that proposed algorithm achieves not only a real time speed, but also the best accuracy that more than half of difficult targets are detected successfully.

## Introduction

In monocular monitoring system, pedestrians are the most important, complex and volatile objects. Currently, focuses of pedestrian detection algorithm include foreground segmentation, pedestrian overlapping and detector adaptability. Over the past decade, remarkable achievements have been made based on sliding window algorithm [1, 2], such as HOG SVM [3] proposed by Dalal, multi-feature detector [4] produced by Wejek, and integral channel features used in [5]. However, the results from [1] show that efficient classifier requires a lot of positive and negative samples for SVM training, so that improvements in detection accuracy are accompanied with the increased computational costs, and then detector can't adapt to the change of application scene.

In order to improve the efficiency and the adaptability of detector, YANG at al. in [6] propose a fast predicting detection algorithm based on motion information. YANG believes that moving is a commonality of pedestrians in changeable scenes. So they obtain motion regions from scenes firstly, and then predict the positions of pedestrians by a comparison of motion regions in two neighbor images, finally use key points with BRISK feature to verify the correction of predictions. Although the predicting algorithm achieves a real-time detection of speed, but its detection capability is limited by occlusions when applied in crowded scenes. In order to detect pedestrians in occlusions, DPM [7, 8, 9] has been widely used and achieved excellent results. However, the essence of DPM is the training and matching of feature samples in multiple angles, limiting the detection speed just as showed in [2].

CMT in [10] uses key points to match objects, to a certain extent, solving the detection problems in which objects may be with posture changed and partial occlusion. Detection algorithms based on motion information [6] and key points matching [10] enables faster operation efficiency but lacks credibility slightly when compared to detection algorithms using features [3, 4, 5]. BING [11] in CVPR 2014 obtains the best performance in generic object detection. It realizes an amazing real-time

speed: 300 images could be detected per second on a popular PC, and recalls 96.2% objects with only 1000 proposals. So BING is ideally suited as a final classifier to enhance the credibility of detection.

In this paper we proposed a new algorithm to detect moving pedestrians based on motion information, normed proposals and key points matching. According to foreground segmentation algorithms [6, 12, 13], the pose, position and background of a same pedestrian in two neighbor video frames may not change significantly. So we proposed a new method to obtain motion regions of interest by a flash-bit computing method. A motion region of interest which initialized in the first image is considered to contain a single target only. If a region only contains a single target, key points with BRISK feature [14] in this region are computed and saved. When a single target converts to be a difficult one due to obstructions and overlaps, the difficulty of the detection will increase significantly. Inspired by the CMT [10], the key points saved before are used to match difficult target, resulting in more than half of pedestrians in occlusions are detected successfully. We prove that normed proposals with width $2^m$ and height $2^n$ have more confidence than motion regions, so that the normed proposals outputted by BING [11] are selected partly to correct the location and shape of detection windows. Experiments on three different datasets show that the algorithm proposed in this paper not only improves the detection accuracy, but also achieves a real-time speed in comparison with other four algorithms including HOG SVM [3].

## Methodology

**Moving Regions of Interest.** Moving pedestrian can be divided into two categories: the single targets and difficult targets. A single target may convert to be a difficult one due to obstructions and overlaps in the process of its movement. Our algorithm is divided into three parts as: obtaining moving regions of interest, matching difficult targets by key points, correting detection by normed proposals. Given a sequence of images $I^0, I^1, \ldots, I^t$, two neighbor images $I^{t-1}, I^t$ are considered as a computable node. Areas which contain pedestrians are considered as moving regions of interest. In each $I^t$ with $t > 0$ we simply obtain moving regions of interest by Eq. (1).

$$MRoi^t = \sum_{i=1}^{Nt} rect(((I^{t-1} \oplus I^t)\ \&\ I^t) > Th_{mr}) \tag{1}$$

Where $Nt$ refers to the amount of regions in $MRoi^t$ and $rect$ is an operation of drawing a rectangle. Firstly, a *XOR* operating removes the same static objects in background but obtains moving objects in two neighbor images $I^{t-1}, I^t$. Then a *AND* operating removes the pixels belong to $I^{t-1}$ but retains those in $I^t$. In realistic scenarios, due to the interference of noise, only pixels whose gray value is greater than the threshold $Th_{mr}$ can be judged belonged to $MRoi^t$. Finally, drawing a rectangle to surround neighbor pixels whose binarized values is equal to 1 as much as possible.

We focus on the case in which a single target converts into a difficult one, so a variable $Ch$ is used to identify whether a region in $MRoi^t$ contains a single target or not. In order to identify each region in $MRoi^t$, five variables $center$, $w_{mr}$, $h_{mr}$, $Ch$, $i$ are used, respectively, as the center, width, height, occlusion identifier and index of the regions.

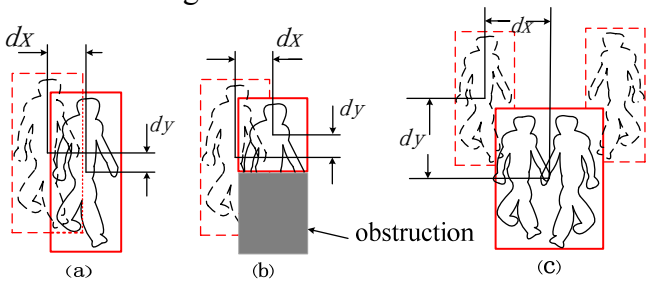



Figure.1. The potential movements of a single target in two neighbor images. (a) Target still is a single one. (b)Target converts to be a difficult one due to obstructions in the background. (c)Target converts to be a difficult one due to overlaps between pedestrians.

As shown in Fig.1, a single target converts to be difficult one due to obstructions (Fig. 1(b)) and overlaps (Fig. 1(c)). When the case occurs, the shape and center point of the region in $MRoi^t$ will change significantly in two neighbor images. We calculate the displacement of a same moving region in two images by Eq. (2).

$$(dx, dy) = center^t - center^{t-1} \tag{2}$$

$dx, dy$ identify the displacement of the region together. We find that if a single target converts to be difficult one due to obstructions, the displacement of the target may not be large but its shape changes significantly (Fig. 1(b)) when compared to Fig. 1(a). However, when overlapping, both the displacement and shape of the single target change significantly, just as showed in Fig. 1(c). We update the value of Ch by Eq. (3).

$$Ch = (\eta_c \mid dx \times dy \mid + \mid (1 - \eta_c)(w_{mr}^t \times h_{mr}^t - w_{mr}^{t-1} \times h_{mr}^{t-1}) \mid) > Th_{ch} \tag{3}$$

$\eta_C$ is an adjustment coefficient which related with the amount of obstructions and the density of pedestrians. Although we discuss above two different occurrences, but we use a same algorithm named KPM to match difficult targets. So when the sum result of displacement and shape is bigger than threshold $Th_{Ch}$, we believe that the case occurs and then update the value of identifier $Ch$.

As shown in Fig.2, we get moving regions of interest by above flash-bit computing method on three different datasets. In dataset PETS2010 V1 and V6, the background and pedestrians' shape are simple, the effect of motion regions detection is obvious. However, in database Video Seq, the lights and shadows have significantly bad effects on the detection of $MRoi$, resulting in poor motion detection.

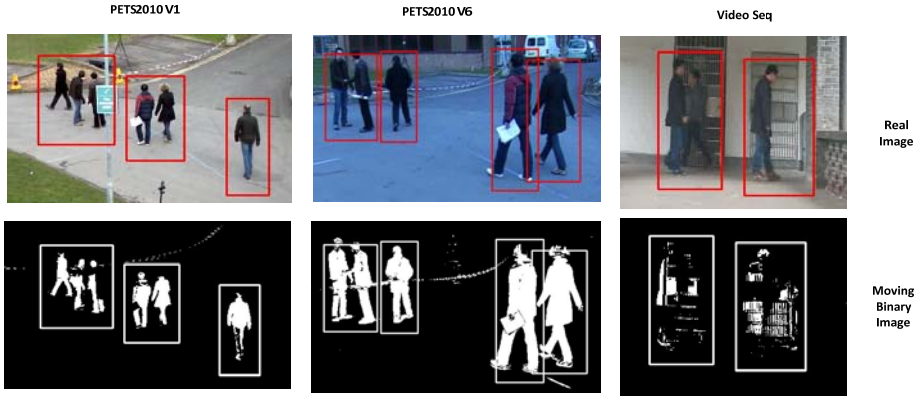



Figure.2 Single targets and difficult targets are surrounded by a red rectangle in three different datasets PETS2010 V1, V6 and Video Seq. Moving binarized images are obtained by the flash-bit computing method.

**Normed Proposals.** We can roughly estimate the location of pedestrians by motion information. However, due to changes of lights and shadows in background, the regions which contain targets may drift and expend significantly, resulting in wrong labeling of pedestrian. In view of those uncertain changes with drifting and expending, normed proposals outputted by BING algorithm are used to correct the shape and location of detection windows.

The normed proposal refers to a detection window which in a fixed value of width and height. In this paper, we size the proposal' width in $2^m$ and height in $2^n$, and then BING algorithm [11] is used to output those normed proposals.

Do those normed proposals have high confidence on matching targets?

The PASCAL criteria in [1] are used to determine the correctness of a matching between a detection window $\alpha_0$ and a ground truth bounding box $\alpha_1$. The matching is considered correct only if the overlapping ratio of $\alpha_0$ and $\alpha_1$ are above a threshold, as shown in Eq. (4).

$$f(\alpha_0, \alpha_1) = \frac{Area(\alpha_0 \cap \alpha_1)}{Area(\alpha_0 \cup \alpha_1)} > 0.5 \tag{4}$$

If the bottom left point of $\alpha_0$ and $\alpha_1$ are overlapped totally on a coordinate system, then we find four positional relationships between $\alpha_0$ and $\alpha_1$, just as shown in Fig. 3(a).

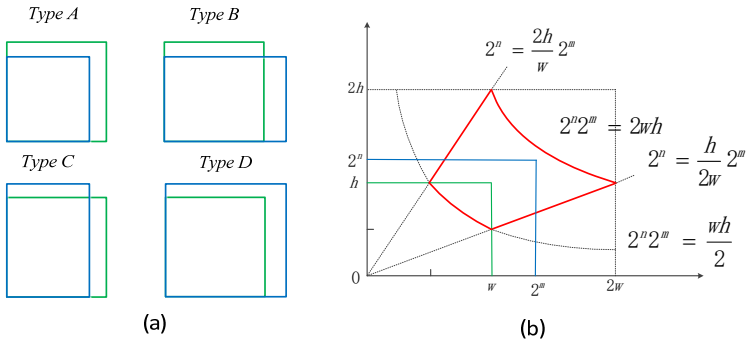



Figure.3 (a) Four positional relationships between a normed proposal $\alpha_0$ (in blue color) and a ground truth bounding box $\alpha_0$ (in green color). (b) The region surrounded by two lines and two curves (in red color) are the possible area in which top right points of the normed proposals may locate.

In order to meet the PASCAL criteria in Eq. (4), some constraints are deduced as follows

Type A: $$2^m 2^n > \frac{wh}{2} \tag{5}$$

Type B: $$\frac{w2^n}{2^m 2^n + wh - w2^n} = \left(\frac{w}{h}\right) \frac{2^n}{2^m \frac{2^n}{h} + w\left(1 - \frac{2^n}{h}\right)} > \frac{1}{2} \tag{6}$$

Since $2^n < h$, a sufficient condition of Eq. (6) is deduced as follows:

$$2^n > \frac{h}{2w} 2^m \tag{7}$$

Type C: $$\frac{2^m h}{2^m 2^n + wh - 2^m h} = \left(\frac{h}{w}\right) \frac{2^m}{2^n \frac{2^m}{w} + h\left(1 - \frac{2^m}{w}\right)} > \frac{1}{2} \tag{8}$$

Since $2^m < w$, a sufficient condition of Eq. (8) is deduced as follows:

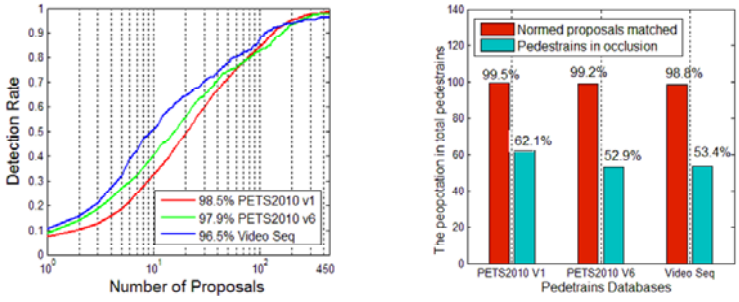
$$2^n < \frac{2h}{w} 2^m \tag{9}$$

Type D: $$2^m 2^n < 2wh \tag{10}$$

According to above inferences, the region $\alpha$ surrounded by two line (Eq. (7), Eq. (9)) and two curves (Eq. (5), Eq. (10)) are the possible area in which top right point ($2^m$, $2^n$) of matchable normed proposal may locate. Although there must be a $\alpha$ on the mathematical deduction, but in fact, $m, n \in Z$, there may not exists two integers ($m$, $n$) makes that the point ($2^m$, $2^n$) in the area of $\alpha$. We get normed proposals $NP$ outputted by BING algorithm on three datasets, and get detection rates between $NP$ and ground truth bounding boxes $R_g$ by calculating whether they meet the PASCAL criteria or not. As shown in Fig.4 (a), the detection rate of first $NP$ is only 10%, but with the amount of $NP$ increasing from 1 to 450, the detection rate increases quickly and eventually reaches at more than 96.5%.

In addition, we assume that the width and height of a pedestrian are both between 16 and 512 pixels, so the value of m and n are both between 4 and 9. For each ground truth bounding box with a width $w$ and a height $h$, we use a sliding window method in which a normed detection window starting at the point (0.5$w$, 0.5$h$) and then sliding in a step of 8 pixels. We get the proportions of pedestrians matched by above normed sliding window method and pedestrians in occlusions, just as shown in Fig.4 (b). The proportions of pedestrians which can be matched by normed sliding windows are similar to 99%, and pedestrians in occlusions are above 52%. After comparing Fig.4 (a) to Fig.

4(b), we draw a conclusion that the normed proposals outputted by BING have high confidence on the matching of pedestrians, even all the pedestrians in occlusions can be matched successfully.




(a) (b)

Figure.4 (a) Tradeoff between the number of normed proposals and detection rate for BING algorithm on three different datasets. The detection rate of first $NP$ is only 10%, but eventually reaches at more than 96.5% when the number of $NP$ increases to 450. (b) The proportions of pedestrians matched by normed proposals which outputted by a sliding window method and the pedestrians in occlusions.

For each region $MRoi_i^t$ in moving regions of interest and 450 normed proposals, if $MRoi_i^t$ and $NP$ meet the PASCAL criteria, then those $NP$ are selected to correct the location and shape of $MRoi_i^t$ in an average way, as shown in Eq. (11).

$$(center, w_{mr}, h_{mr}) = \sum_{j \in SEL}^{} avg(center_j, w_j, h_j) \tag{11}$$

Where $SEL$ are the collection of those selected $NP$, $center_j, w_j, h_j$ are the center, width and height of each selected $NP$. As shown in Fig.5, the moving regions which corrected by selected normed proposals still contain pedestrians well.

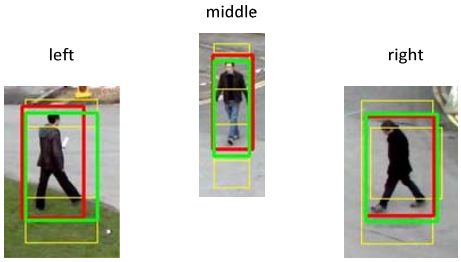



Figure.5: The moving regions which corrected by selected normed proposals still contain pedestrians well. The moving regions which before corrected are surrounded by red rectangle; the normed proposals are in yellow color; the moving regions which after corrected are surrounded by a green rectangle.

**Key Points Matching.** In order to match difficult targets, when a moving region of interest $MRoi_i^t$ contains a single target only, the key points with BRISK feature in this region are detected and saved as a unit

$$UKP_i^t = \left\{ ( p_k, n_k ) \right\}_{k=1}^{N_{kp}} \tag{12}$$

Where $p_k$ refers to the key point position in absolute image coordinates of $I^t$, $n_k$ is an index descriptor and $N_{kp}$ are the amount of key points in $MRoi_i^t$. The pose and angle of a same pedestrian in two neighbor images will not change significantly, but after a long time movement, there may exists obvious pose rotation and change in background. So we update the unit of key points $UKP$ in each single region continuously until the moment in which the region converts to be a difficult one. In that converting moment, we stop updating key points and record the unit of key points as $UKP_{Ch}^t$, the amount of key points as $N_{Ch}$ and the single target as $MRoi_{Ch}^t$.

In next $s$ frames, in order to find out the target which have converted to be a difficult one in $I^{t+s}$, we calculate corresponding key points in $MRoi_{LK}^{t+s}$ by an optical flow method in Eq. (13).

$$\begin{aligned} UKP_{LK}^{t+s} &= \left\{ CLK\left(UKP_{Ch}^{t}\right)\right\}_{j}^{N_{LK}} \\ &= \left\{\left(k_j, n_j\right)\right\}_{j}^{N_{LK}} \end{aligned} \tag{13}$$

Where $CLK$ denotes the optical flow method, $UKP_{LK}^{t+s}$ is a unit of corresponding key points, $k_j$ refers to the key point position in absolute image coordinates of $I^{t+s}$, and $n_j$ is the index of the corresponding key points in $UKP_{LK}^{t+s}$. If $N_{LK} / N_{Ch} > 0.3$, indicating that the target can be matched; otherwise, the difficult target is in a serious occlusion and could not be matched.

For each key point in $MRoi_{Ch}^{t}$, we compute the hamming distances as Eq. (14).

$$D_{k,j}(n_k^t, n_j^{t+s}) = \sum_{k=1}^{N_{Ch}} XOR(n_k^t, n_j^{t+s}) \tag{14}$$

Then we choose the median value of humming distances as the displacement of $MRoi_{Ch}^{t}$ in $I^{t+s}$, and then we find out a location and shape of the difficult target named as $MRoi_{Ch}^{t+s}$ by Eq. (15).

$$MRio_{Ch}^{t+s} = MRio_{Ch}^{t} + median(D_{k,j}(n_k^t, n_j^{t+s})) \tag{15}$$

Finally, normed proposals $NP_j^{t+s}$ in $I^{t+s}$ are used to correct the location and shape of the difficult target $MRoi_{Ch}^{t+s}$.

We name the algorithm described above as KPM. When compared to CMT [10] in detail, KPM has three differences listed as follows:

(1) CMT is a single target matching algorithm. However, KPM are aimed at matching difficult targets, and several difficult targets may exist at the same time.

(2) The unit of key points in CMT is initialized at frame 1 and would not be updated in the next image sequences. However, in KPM, the unit of key points for each single target has been updated until the single target converting to be a difficult one. KPM with an updating strategy can effectively reduce the possibility in which detection window may drift significantly due to the change of pose and obvious angle rotation.

(3) KMP processes key points simply, so that it meets a faster speed than CMT and is more suitable to be applied in pedestrian detection.

## Experiments

In order to evaluate the performance of the proposed methodology in section 2, we evaluate and compare five different detection algorithms on three different datasets whose statistic are shown in Table 1.

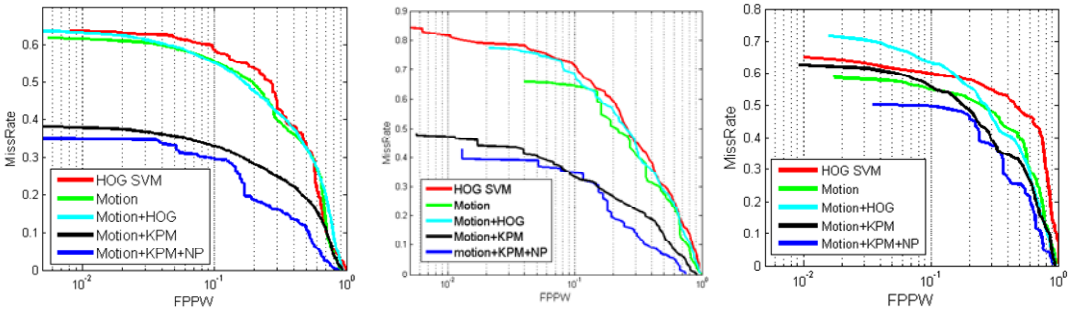
Table 1 Pedestrian datasets

| properties | PETS2010 V1 | PETS2010 V6 | VideoSeq |
|---|---|---|---|
| frames | 796 | 796 | 732 |
| pedestrians | 4880 | 3977 | 1847 |
| $Th_{mr}$ | 90 | 30 | 50 |
| $Th_{Ch}$ | 1300 | 1800 | 1800 |
| $Nc$ | 0.25 | 0.1 | 0.1 |
| resolutions | 720 * 576 | 720 * 576 | 720 * 480 |

The first algorithm is HOG SVM proposed in [3], it has been proved to be one of the most optimal detectors on some challenge datasets [1]. The second algorithm is named as Motion, in which each region of interest obtained by the flash-bit computing method (section 2.1) is thought to contain a single target only. The third algorithm is a mixed version of HOG and Motion, in which each moving region of interest is cut out alone and then as an input image of HOG detector.

The fourth algorithm is named as Motion+KPM, in which we use the flash-bit computing method to detect single targets and then use KPM to match difficult targets. The Last algorithm is based on the fourth one, and then normed proposals which proved to have higher confidence are used to correct the shape and location of detector' output windows.

**Accuracy Analysis.** We plot Detection Error Tradeoff (DET) curves to evaluate the performance of the five different algorithms described above. MissRate (FN/(FN+TP)) versus FPPW (false positives per window, FP/(FP+TN)) shown on logarithmic scale. The PASCAL criteria in Eq. (4) are used to determine whether the detection windows outputted by each algorithm are matchable with the ground truth bounding boxes of each datasets or not.




(a) PETS2010 V1 (b) PETS2016 V6 (c) Video Seq

Figure.6 Accuracy comparisons on three datasets

Fig.6 illustrates that our algorithm Motion+KPM+NP provides the best accuracy on three datasets. Especially, in datasets PETS2010 V1 and V6, KPM matches nearly more than 30% pedestrians compared with Motion. Since the proportion of pedestrians in occlusions is 52% as shown in Fig. 4(b), and then we conclude that KPM has matched nearly half of the difficult targets successfully. In addition, the detection windows corrected by normed proposals improve nearly 5% accuracy of the detector when Motion+KPM+NP is compared to Motion+KPM. In dataset Video seq, although moving regions are easily affected by the lights and shadows in complex background, our algorithm still achieves a 9% improvement of accuracy when compared to HOG SVM. Motion+HOG, Motion and HOG SVM achieve a similar effect for the reason that classes of moving objects in three datasets are pedestrians basically.

**Detection Speed.** All the experiments are conducted on a same laptop, equipped with Intel i3 CPU and 4GB memory. Resource code for five algorithms are mixed programmed in Python and C, where HOG and BING are implemented with C language; Motion, KPM and NP are produced by Python language.

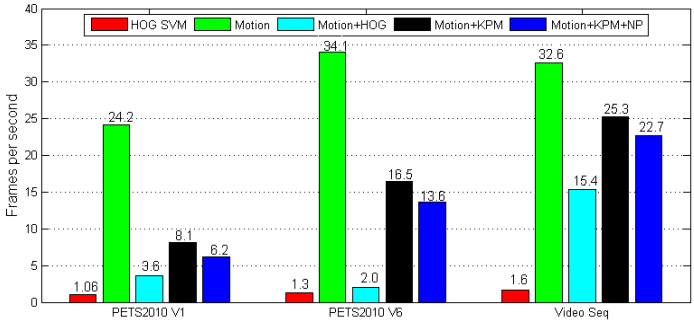
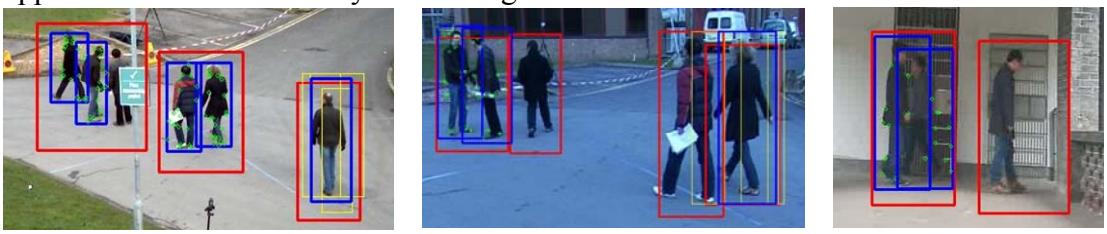



Figure. 7 Detection speed comparison

As shown in Fig.7, the fastest one is Motion detector, in which all moving regions of interest are obtained by the flash-bit computing method in Eq. (1). HOG SVM is with least efficient, for the reason that the detection speed of a sliding window detector will decrease significantly when the input image is with a higher resolution and a heavy pedestrian density. Motion+HOG achieves a 1.5 times speed up than HOG SVM, because of the input images of the detector are those moving regions cut from the image, to some extent, decreasing the size of the target image. The speed of Motion+KPM is little slower than motion, for the reason that the processing time of KPM will increase significantly when scene is crowed as in dataset PETS2010 V1. NP using those normed proposals outputted by BING, thanks to the amazing speed of BING (300fps), it has little effect on the

total speed of Motion+KPM+NP. In a medium density surveillance video, Motion+KPM+NP achieves a acceptable real-time speed with the value of 13.6fps and with the best accuracy, so that it is suitably applied in a medium density monitoring environment.



(a)PETS2010 V1 (b)PETS2010 V6 (C) Video Seq

Figure.8: Screenshot of qualitative results on three datasets (red bounding box denote the moving regions of interest obtained by the flash-bit computing method in Eq.(1); the light green circles are the corresponding key points in KPM used to find out difficult targets; light yellow bounding boxes are those normed proposals which are selected to correct the shape and location of detection windows; blue bounding boxes are the finally output detection windows of the Motion+KPM+NP.For a better view, only part of those selected normed proposals (in light yellow color) and final output window (in blue color) are shown in images.

## Conclusion

In this work, we have presented an algorithm for the detection of single targets and difficult targets in occlusions. Firstly the flash-bit computing method is used to obtain moving regions of interest quickly, and then key points are updated and saved for matching difficult targets, finally normed proposals which proved to have high confidence are used to correct the location and shape of the detection windows. Compared to other algorithm, our algorithm achieves greater accuracy in which half of difficult targets are detected successfully and meets a real-time processing speed.

In further work we will strengthen the adaptive capacity of our detector, for better improvement accuracy in complex application environments, in which background may changes frequently and density may be heavier.

## References


[1] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: A benchmark[J]. Proc Cvpr, 2009:304-311.

[2] Dollár P, Wojek C, Schiele B, et al. Pedestrian Detection: An Evaluation of the State of the Art[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 2012, 34(4):743-761.

[3] Dalal P, Triggs B. Histograms of Oriented Gradients for Human Detection[C]. IEEE Conference on Computer Vision & Pattern Recognition. 2005:886-893.

[4] Wojek C, Schiele B. A Performance Evaluation of Single and Multi-feature People Detection[M]. Pattern Recognition. Springer Berlin Heidelberg, 2008:82-91.

[5] Dollár P, Tu Z, Perona P, et al. Integral Channel Features[J]. Bmvc, 2009.

[6] Yang Z, Wang T, Deng J, et al. Fast pedestrian detection based on motion informaion[J]. Journal of Computational Information Systems, 2015, 11(6):2303-2313.

[7] Felzenszwalb P, Mcallester D, Ramanan D. A discriminatevely trained, multiscale, deformable part model[C]. In IEEE Conference on Computer Vision and Pattern Recognition CVPR-2008. 2008:1-8.

[8] Felzenszwalb P F, Girshick R B, David M A, et al. Object Detection with Discriminatively Trained Part-Based Models[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 2010, 32(9):1627-1645.

[9] Tang S, Andriluka M, Schiele B. Detection and Tracking of Occluded People[J]. International Journal of Computer Vision, 2014, 110(1):58-69.

[10] Nebehay G, Pflugfelder R. Consensus-based matching and tracking of keypoints for object tracking[C]. Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on. IEEE, 2014:862-869.

[11] Ming-ming Cheng, Ziming Zhang, Wen-yan Lin, et al. Binarized normed gradients for objectness estimation at 300fps[C]. Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014:3288-3293.

[12] Piccardi M. Background subtraction techniques: a review[C]. Systems, Man and Cybernetics, 2004 IEEE International Conference on. IEEE, 2004:3099 - 3104.

[13] Kalal Z, Mikolajczyk K, Matas J. Forward-Backward Error: Automatic Detection of Tracking Failures[C]. Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010:2756-2759.

[14] Leutenegger S, Chli M, Siegwart R Y. BRISK: Binary Robust invariant scalable keypoints[C]. Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011:2548-2555.