Multimed Tools Appl (2018) 77:30891–30910
https://doi.org/10.1007/s11042-018-6173-4

Real-time moving pedestrian detection using contour features

Kai Zhao[1] · **Jingjing Deng**[1] · **Deqiang Cheng**[1]

Received: 13 November 2017 / Revised: 12 May 2018 / Accepted: 21 May 2018 /
Published online: 2 June 2018

Abstract Pedestrian detection is one of the most fundamental research in computer vision. However, many high performance detectors run slowly. In this paper, we propose a real-time moving pedestrian detector by using efficient contour features. Firstly, the moving targets are detected by background subtraction. By combining the elliptic Fourier descriptors and the normalized central moments, we propose the Elliptic Fourier and Moments Descriptors (EFMD) to describe the moving target contours. Secondly, the moving targets are classified by the trained Support Vector Machine (SVM). In addition, we introduce a novel overlap handling algorithm based on linear fitting and normalized central moments, which improves the detection performance by reducing both false positives and miss rate. The experimental results on PETS 2009 and CAVIAR datasets show that our approach achieves a miss rate of 14% (PETS 2009) and 13% (CAVIAR) at 10^{-1} False Positives Per Image (FPPI) and an average runtime per frame of 30 ms (PETS 2009) and 25 ms (CAVIAR), which significantly outperforms several state-of-the-art detectors in both detection performance and runtime.

Keywords Pedestrian detection · Elliptic Fourier descriptors · Normalized central moments · Support vector machine · Linear fit

1 Introduction

As the hot topic in the field of computer vision, pedestrian detection, which is widely used in the video surveillance, automatic driving, event detection [4, 5], robot technology, and human-computer interaction [3], has attracted considerable attention in computer vision recently. The task of pedestrian detection is to recognize the pedestrian and output the location and the size of that in the image automatically [20]. With the increasing number of video surveillance devices installed in public places, like airports, subways, railway stations and road junctions, it

✉ Deqiang Cheng
dqcheng@cumt.edu.cn

[1] School of Information and Control Engineering, China University of Mining and Technology, No.1, Daxue Road, Xuzhou, Jiangsu 221116, People's Republic of China

is difficult for people to keep watching and analyzing these huge video data. The pedestrian detection system is more efficient and more intelligent than human. However, pedestrian detection is confronted with many challenges due to frequent occlusion, complex background and the diversity of pedestrian postures, clothes, camera angles and illuminations [17]. Although many high-performance methods for pedestrian detection have been proposed, the issue of the fast and reliable pedestrian detection is still far from being solved.

This paper aims to propose a fast and accurate pedestrian detector, which detects pedestrians in real-time and solves the overlap problem in video surveillance. The main contributions of this paper are summarized as follows:

(1) An efficient method for describing moving object contours is proposed in this paper. As shown in Fig. 1, the elliptic Fourier descriptors [12] and the normalized central moments are combined into the proposed EFMD features.
(2) We design a novel overlap handling algorithm based on linear fitting and normalized central moments, which improves the detection performance by reducing both false positives and miss rate.

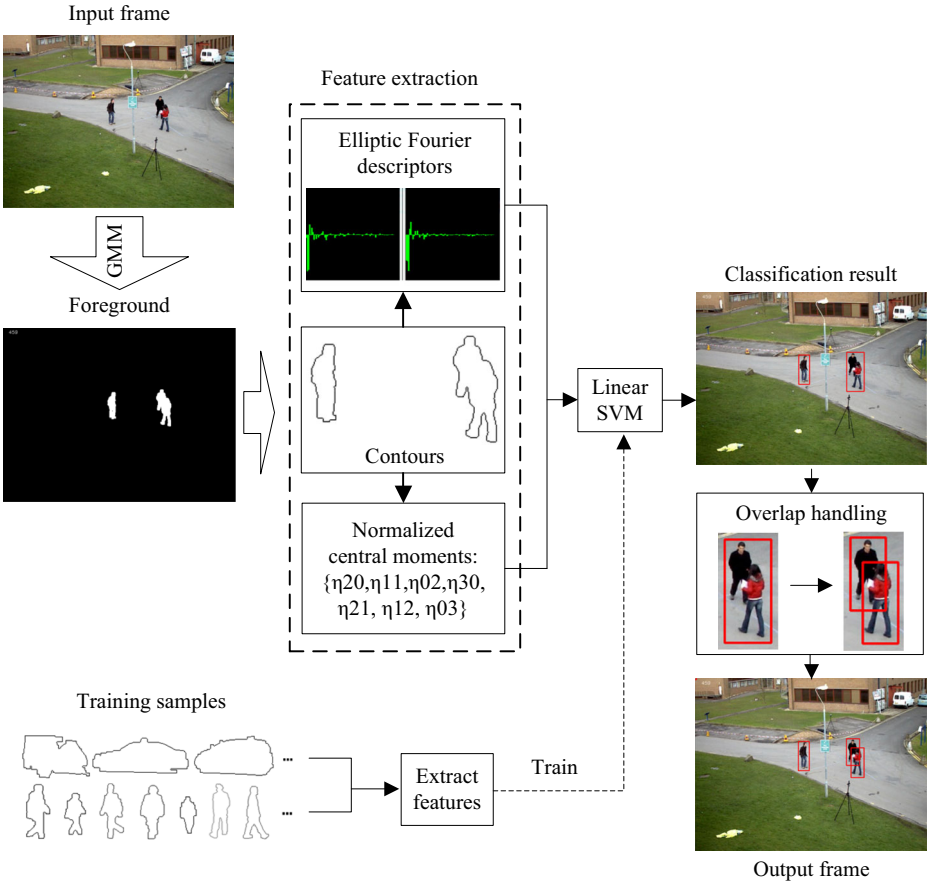


Fig. 1 The flow chart of our approach

The remainder of this paper is organized as follows: in Section 2, we will review the related work. In Section 3, we introduce the proposed method. The experimental results are illustrated in Section 4. Finally, we summarize all the conclusions in Section 5.

2 Related work

At present, the features used in pedestrian detection can be divided into two categories: the traditional handcrafted features and the data-driven features [25]. The former, which is designed by people, becomes more and more complicated. While the latter, as the new emerging feature, is learnt from a large amount of training data via artificial neural network.

Numerous researchers are devoted into the handcrafted features. Papageorgiou and Poggio [21] employed the SVM to learn a pedestrian detector using Haar wavelets as descriptors. Viola and Jones [27] proposed the integral images to increase the speed of computation of Haar-like features, and the cascade classifier was trained utilizing the adaboost, which achieved excellent performance in face recognition. Inspired by SIFT [18], Dalal and Triggs [6] proposed the histogram of oriented gradients (HOG) feature and combined it with the linear SVM classifier to get the detector, which obtained almost perfect detection performance in the early pedestrian detection dataset. Gradient-based features built the milestone and significantly improved the performance of pedestrian detectors. Afterwards the gradient information was widely used in the following pedestrian detectors.

In addition, the shape feature is another significant cue in pedestrian detection. Gavrila and Philomin [11] took advantage of the Hausdorff distance transform and the template hierarchy to quickly match the edge of the image with the shape templates tree. The edgelet feature consisting of a large number of short lines and curve segments was proposed by Wu and Nevatia [31] to represent the local shape and the Boosting algorithm was employed to learn the various parts of the body. Sabzmeydani and Mori [23] proposed the Shapelet feature and the global detector was generated from the combination of the local descriptors by means of the Boosting algorithm. Lin and Davis [16] used a shape-based template tree to model various parts of the human body.

As feature fusion leads to better detection performance, the handcrafted features evolved from single channel feature into multi-channel features that contains multiple information. Some multi-channel features even contained 10 channels, in which the gradient [6, 27, 32], the color and the texture [15, 34] are usually included. Wang et al. [29] combined the local binary model (LBP) [15] and the HOG as HOG-LBP characteristic set that obtained better detection function. Wojek and Schiele [30] fused the Haar-like feature, the HOG feature, the Shapelet feature and the shape context into the MultiFtr feature, which performs better than any single feature of those. MultiFtr+CSS and MultiFtr+Motion were proposed by Walk et al. [28], which combined the MultiFtr with local color self-similar features and motion features respectively. Dollár et al. [7] proposed the ChnFtrs by calculating the Haar-like feature on a number of channels such as gradient amplitude channel, gradient direction quantization channel, LUV color channel and gray-scale channel, etc. After that, they proposed the FPDW [8] and ACF [10] by approximating features from nearby scales. However, the multi-feature fusion improves the detection capability at the cost of increasing time for computation, which is not suitable for real-time pedestrian detection. Some researchers reduced the feature space dimension, which made the subsequent computation more efficient [2, 14].

Recently, the deep learning model has shown very powerful detection performance, such as convolution neural network (CNN) [24] and deep belief network (DBN) [13], which overcomes the previous manual designed detectors.

The performance of the detector using global description features is not excellent, when the pedestrian is occluded. The more serious the occlusion is, the lower the detection rate goes. Therefore, many researchers proposed detection methods based on several main parts of human body. For example, Lin and Davis [16] proposed a template tree model by matching the upper body, waist and legs, respectively. Mohan et al. [19] extracted Haar wavelet features for four parts of the human body (head, left arm, right arm and leg), and utilized SVM as classifier to train detector. In spite of this, Dollár et al. [9] demonstrated that the performance of part based detectors in the case of partial occlusion is still not satisfactory by experiments on Caltech Pedestrian Dataset.

Many pedestrian detection methods use background subtraction to eliminate the interference caused by clutter background. The pedestrians in video surveillance are mostly in motion, so the moving pedestrians can be segmented from the background by moving target detection. Zhang et al. [33] proposed pedestrian detection method based on motion analysis, which improved the performance and efficiency of the detector. They utilized motion segmentation to detect pedestrian candidate area. Gaussian Mixture Model (GMM) [26] is a widely used background modeling method, and the model is continually updating according to background changes. Optical flow method [1] is a motion target detection method that adapts to the movement of the camera, but the computation cost is high.

3 Proposed approach

The flow chart of our approach is shown in Fig. 1. The system consists of two parts: training and detecting. In the training phase, a large number of moving targets are separated from the surveillance video as training samples, including pedestrians and vehicles. And then the contour features of these moving targets are extracted. Finally, the features and labels of these training samples are input into the SVM to train the classifier. In the detecting phase, firstly, the system reads a frame from the video sequence, and the GMM is used to extract the foreground. Secondly, the EFMD features of the moving target contours are extracted. Finally, the SVM classifier gives the detection results. If two pedestrians overlap, the overlap handling is added in the end.

3.1 Background modeling and foreground extraction

Using a statistical model to represent the scene is the most typical solution for background foreground segmentation, such as GMM proposed by Stauffer and Grimson [26]. The GMM is highly adaptable to light change, reciprocating motion interference and scene changes. In this paper, we use GMM for background subtraction to extract moving targets.

In GMM, the value of each pixel in video frame is represented by a mixture of K Gaussian distributions. Usually, the value of K is between 3 and 5. The Gaussian distributions are ranked by the value of ω/σ, where ω and σ are the weight and

variance of the i-th Gaussian distribution in the mixture model respectively. Select the first B Gaussian distributions as background, where

$$B = \text{argmin}_b \left(\sum_{i=1}^{b} \omega_i > T \right) \tag{1}$$

and T is the weight threshold of the Gaussian distribution that should be chosen as background. The range of T is between 0.5 and 1, and the empirical value is 0.7.

In the surveillance scene, the trees rippled in the breeze lead to small noise particles in the binary image, and cavities are usually observed in the foreground. Because the follow-up processing is affected by these noise particles and cavities, the morphological operation on the binary image is used to eliminate the isolated particles and fill the cavities in the foreground. The erosion operation is used to eliminate small objects in the image. The erosion operation of structure A by an element B is defined as:

$$A \Theta B = \left\{ x \middle| \left[(B)_x \subseteq A \right] \right\} \tag{2}$$

The dilating operation is the sister of erosion. We apply the dilation to the foreground after the erosion. In order to fill the cavities effectively, the dilation operation should be performed at least twice. The dilating operation of structure A by an element B is defined as:

$$A \oplus B = \left\{ x \middle| \left[(B)_x \cap A \neq \varnothing \right] \right\} \tag{3}$$

As shown in Fig. 2, there are two defects in the extraction of moving targets by using GMM only: the noise particles caused by the branches and leaves flickering in the wind, and the cavities in foreground. Obviously, these two problems have been properly solved by the morphological operation. In our experiment, the erosion is performed once firstly, and the dilation is performed three times next. The result shows that the morphological operation is effective.

3.2 Elliptic Fourier and moments descriptors

Contour is a significant cue to identify targets. We focus on analyzing the contours of the moving targets after getting the foreground. Small targets, such as flags and birds, bring about interference, therefore the threshold are set to filter out the small contours. In order to describe the target contour more accurately, the elliptic Fourier descriptors and the normalized central

(a) (b) (c)

Fig. 2 **a** Original image, **b** Moving objects detection by GMM, **c** Morphological operation on the foreground

moments are used to describe the contours. These descriptors are connected together to form the EFMD features. Algorithm 1 introduces the extraction process of the EFMD features.

Algorithm 1 Extracting EFMD features

Input: video frames;
Output: EFMD features;
1. **for** each frame **do**
2. detect moving objects by GMM;
3. do morphological operation;
4. extract moving object contours: $\{C_1, C_2, \ldots Cn\}$;
5. **for** $i = 1, \ldots, n$ **do**
6. compute contour C_i elliptic Fourier descriptors: $\left\{\frac{a_1}{A_1}, \frac{c_1}{A_1}, \frac{b_1}{B_1}, \frac{d_1}{B_1}, \ldots, \frac{a_{20}}{A_1}, \frac{c_{20}}{A_1}, \frac{b_{20}}{B_1}, \frac{d_{20}}{B_1}\right\}$;
7. compute contour C_i normalized central moments: $\{\eta 20, \eta 11, \eta 02, \eta 30, \eta 21, \eta 12, \eta 03\}$;
8. merge descriptors: $\mathrm{EFMD} = \left\{\frac{a_1}{A_1}, \frac{c_1}{A_1}, \frac{b_1}{B_1}, \frac{d_1}{B_1}, \ldots, \frac{a_{20}}{A_1}, \frac{c_{20}}{A_1}, \frac{b_{20}}{B_1}, \frac{d_{20}}{B_1}, \eta 20, \eta 11, \eta 02, \eta 30, \eta 21, \eta 12, \eta 03\right\}$;
9. **end for**
10. **return** EFMD;
11. **end for**

The elliptic Fourier descriptors describe the closed curve in a two-dimensional space by considering that the image space defines the complex plane. Based on the Fourier series, elliptic Fourier descriptors approximate the boundary by multiple harmonics. A two-dimensional closed curve is defined as:

$$c(t) = x(t) + jy(t) \tag{4}$$

The curve $c(t)$ can be expressed in Fourier expansion form as:

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} a_0 \\ c_0 \end{bmatrix} + \sum_{k=1}^{\infty} \begin{bmatrix} a_k & b_k \\ c_k & d_k \end{bmatrix} \begin{bmatrix} \cos(k\omega t) \\ \sin(k\omega t) \end{bmatrix} \tag{5}$$

where

$$\begin{aligned} a_k &= \frac{2}{N}\sum_{i=1}^{N} x_i \cos(k\omega i\tau), \quad b_k = \frac{2}{N}\sum_{i=1}^{N} x_i \sin(k\omega i\tau) \\ c_k &= \frac{2}{N}\sum_{i=1}^{N} y_i \cos(k\omega i\tau), \quad d_k = \frac{2}{N}\sum_{i=1}^{N} y_i \sin(k\omega i\tau) \end{aligned} \tag{6}$$

Here, ω defines the fundamental frequency which is equal to $T/2\pi$, T is the period of the curve, n is the number of sampling points, and τ is the sampling period that is equal to T/N. The k-th harmonic coefficients a_k, b_k, c_k, d_k define the k-th order ellipse, and the curve can be approximated by the summation of all ellipses.

Since pedestrians are upright, we do not need to construct descriptors with rotation invariance. In addition, the car contour rotated 90 degrees is similar to the pedestrian contour, which causes interference to pedestrian recognition. Thus, we only consider the invariance of translation and scale. The translation invariance can be achieved by discarding the direct-

current component. To achieve invariance of scale, the semi-major axis A_1 and semi-minor axis B_1 of the ellipse for $k = 1$ are used.

$$A_1 = \sqrt{{a_1}^2 + {c_1}^2}, B_1 = \sqrt{{b_1}^2 + {d_1}^2} \tag{7}$$

In general, the elliptic Fourier descriptors with scale and translation invariance can be simply defined as:

$$\mathrm{EFD} = \left\{ \frac{a_1}{A_1}, \frac{c_1}{A_1}, \frac{b_1}{B_1}, \frac{d_1}{B_1}, \cdots, \frac{a_k}{A_1}, \frac{c_k}{A_1}, \frac{b_k}{B_1}, \frac{d_k}{B_1}, \cdots \right\} \tag{8}$$

In the frequency domain, the low frequency ellipses describe the general shape of the contour, while the high frequency ellipses describe the details. Generally, the energy of the Fourier transform for the object in the nature world tends to be concentrated in the low frequency. Therefore, we do not need too high order harmonic coefficients to describe the object contour. Moreover, higher order harmonic coefficients mean greater computation cost. In order to determine the appropriate number of coefficients, we use different number of harmonic coefficients to approximate the original contour. Figure 3 shows the reconstruction of two sample contours. The Fourier approximation of vehicle and pedestrian contours are illustrated respectively. We can see that the reconstructed contour is refined until the curve represents an accurate approximation of the original contour when adding higher order harmonic coefficients.

We compute the absolute error between the reconstructed contours and the original contours by L2-norm:

$$S = \sqrt{\sum_{i=1}^{n} (y_i - f(x_i))^2} \tag{9}$$

where y_i is the target value and $f(x_i)$ is the estimated value. As shown in Fig. 4, we calculated the reconstruction errors of different order elliptic Fourier descriptors for vehicle and

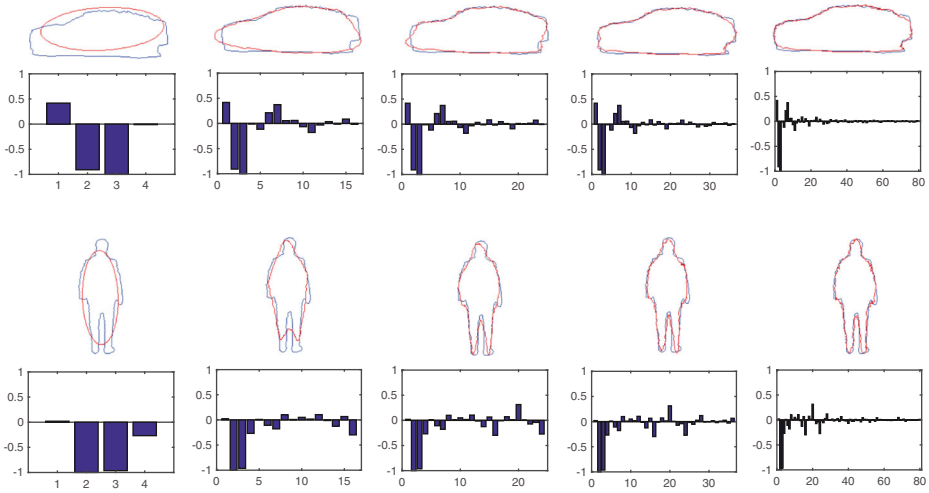


Fig. 3 Examples of elliptic Fourier descriptors reconstruction

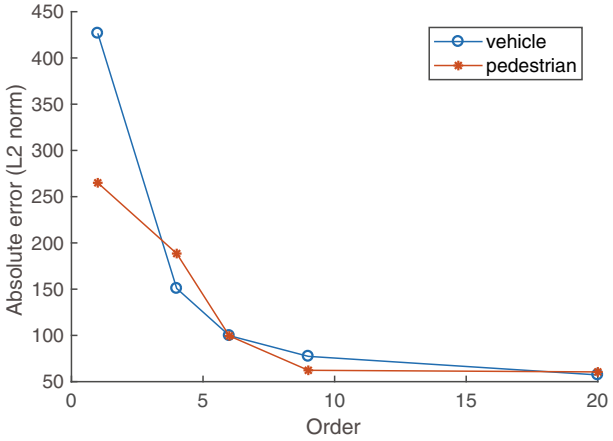


Fig. 4 Reconstruction errors

pedestrian respectively. From the error curves, we can see that higher order harmonic coefficients contribute to the decline of reconstruction error. The error decreases dramatically before reaching the 10th order, and tends to flatten after that. In this paper, we choose up to 20th order harmonic coefficients to describe the target contour.

In addition to elliptic Fourier descriptors, the moments are another widely used shape descriptors, which integrates all the points of the contour. In the discrete case, the $p+q$-th order geometric moments of the image $f(x,y)$ are defined as:

$$m_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} x^p y^q f(x,y) \tag{10}$$

The $p+q$-th order central moments are defined as:

$$\mu_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} \left(x-\overline{x}\right)^p \left(y-\overline{y}\right)^q f(x,y) \tag{11}$$

where $p,q = 0,1,2\ldots$ and $(\overline{x},\overline{y})$ is the mass center:

$$\overline{x} = \frac{m_{10}}{m_{00}}, \overline{y} = \frac{m_{01}}{m_{00}} \tag{12}$$

The normalized central moments are computed as:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(p+q)/2+1}} \tag{13}$$

where $p+q = 2,3\ldots$ The normalized center moments are invariant to translation and scale. We combine the second and third order normalized center moments with the elliptic Fourier descriptors mentioned above to form the EFMD features. That is:

$$\text{EFMD} = \left\{ \frac{a_1}{A_1}, \frac{c_1}{A_1}, \frac{b_1}{B_1}, \frac{d_1}{B_1}, \cdots, \frac{a_{20}}{A_1}, \frac{c_{20}}{A_1}, \frac{b_{20}}{B_1}, \frac{d_{20}}{B_1}, \eta 20, \eta 11, \eta 02, \eta 30, \eta 21, \eta 12, \eta 03 \right\} \tag{14}$$

3.3 Overlap handling

When a pedestrian is partially occluded by another pedestrian, the EFMD detector only identifies them as one pedestrian, since these two pedestrians are overlapped and the foreground targets of them merge into one. In order to solve this problems, we propose an overlap handling method to accurately detect these overlapped pedestrians, which significantly reduces the false alarm rate and the miss rate.

As shown in Fig. 5, the height of one pedestrian in the video is positively correlated with the y-axis value of the center point. Moreover, we find that there is an approximately linear relation between them by establishing the camera imaging model. The relation between height and the y-axis value of center point can be obtained by linear fitting, so that the pedestrian height can be estimated by the y-axis value. The target whose height is significantly higher than the estimated value is usually made up of overlapped pedestrians. In this paper, the proposed overlap handling algorithm identifies the overlapped pedestrians according to the difference between the estimated height and the target height, and then obtains the bounding box of each pedestrian accurately by further segmentation.

Figure 6 is the schematic diagram of imaging model in typical pedestrian surveillance system. Table 1 explains the notation used in this paper. Generally, the surveillance camera is fixed on a pedestal above the pedestrians as illustrated in Fig. 6. Therefore, the parameters a, L, f, and θ that are related to the intrinsic attributes of the camera are fixed values. While the variables S, P, h and y_c that are associated with the pedestrians are variable when they move in the surveillance region. In this paper, we assume that the actual height of the pedestrian obeys the Gaussian distribution $N(\mu_P,\sigma_P^2)$.

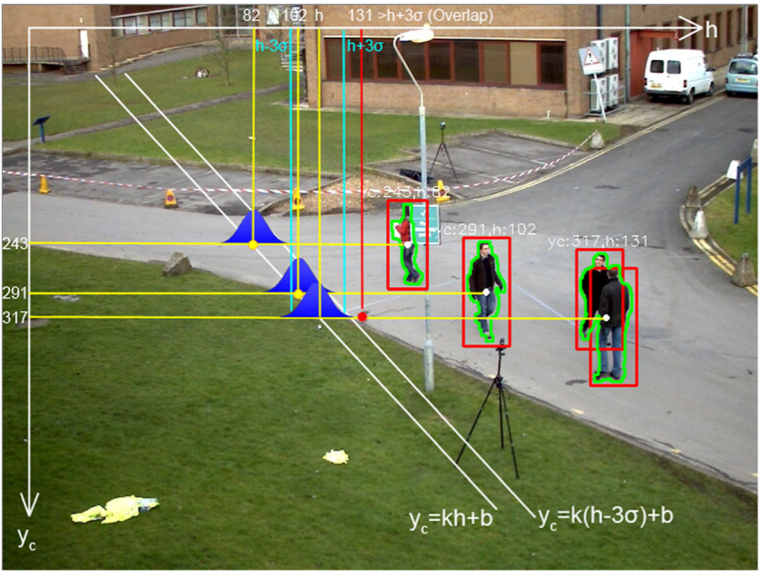


Fig. 5 The pedestrian height and the y-axis value of center point

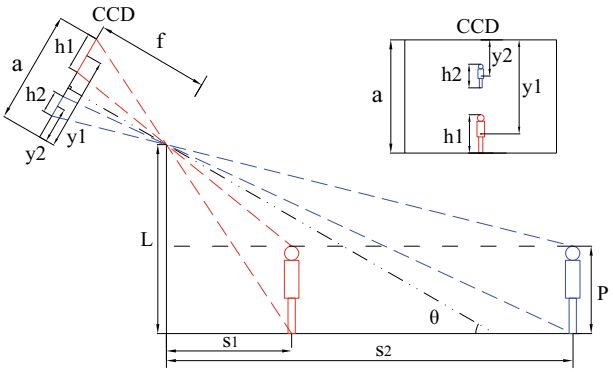


Fig. 6 The schematic diagram of imaging model

According to the geometry of the camera imaging model, the h and y_c can be expressed as:

$$h = f\left[\tan\left(\arctan\frac{L}{S}-\theta\right)-\tan\left(\arctan\frac{L-P}{S}-\theta\right)\right] \tag{15}$$

$$y_c = \frac{a}{2}+\frac{h}{2}+f\tan\left(\arctan\frac{L-P}{S}-\theta\right) \tag{16}$$

The relation between h and y_c can be obtained by Eqs. (15) and (16):

$$y_c = \frac{a}{2}+\frac{h}{2} + f\tan\left(\arctan\frac{2h(L-P)}{fP(1+\tan^2\theta)-h(2L-P)\tan\theta+\sqrt{[h(2L-P)\tan\theta-fP(1+\tan^2\theta)]^2-4h^2L(L-P)\tan^2\theta}}-\theta\right) \tag{17}$$

Figure 7 illustrates the relation between h and y_c. The value of model parameters are set according to a typical case. Here, $f = 1$, $a = 1$, $L = 6$, $P = 1.8$ and the range of θ is 0.1~0.5. Note that, all the angles in this paper are measured in radians. As shown in Fig. 7a, the relation between h and y_c remains linear in general. In addition, we also evaluated the relation between h and y_c at the case of different camera mounting heights. Here, $L = 2$~8, $P = 1.8$, $f = 1$, $a = 1$, $\theta = 0.3$. Similarly, the Fig. 7b shows that the relation between h and y_c remains linear generally as well.

Table 1 The notation used in this paper

Symbol	
a	The height of CCD plane
f	The focal length of camera
L	The mounting height of camera
θ	The dip angle of camera axis
S	The horizontal distance of a pedestrian to camera
P	The height of pedestrian in real word
h	The height of pedestrian on CCD plane
y_c	The y-axis value of the pedestrian center point on CCD plane

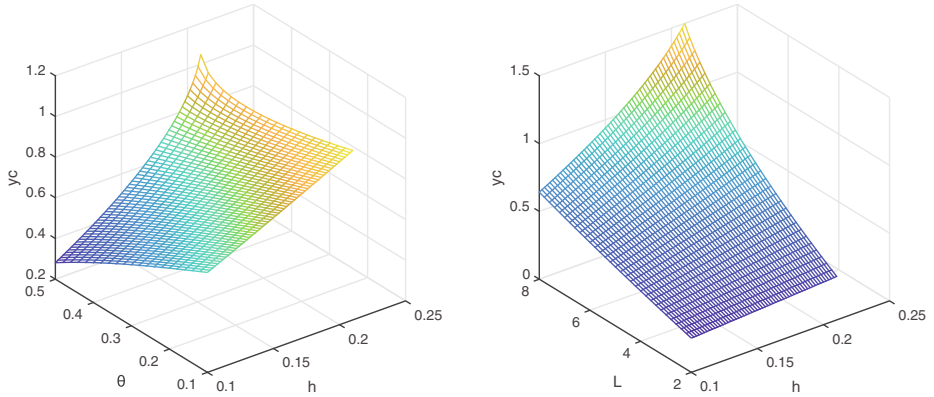


Fig. 7 The linearity of the height estimation equation

In conclusion, the relation between h and y_c is approximately linear in most video surveillance scenarios. We use the liner fitting method to obtain the liner Eq. (18) as the approximation of Eq. (17) by learning a large number of pedestrians in the video sequence.

$$y_c = kh + b \tag{18}$$

Algorithm 2 introduces the fitting process, and the fitting result is shown in Fig. 8. Assuming that the variable h obeys the Gaussian distribution $N(\mu,\sigma^2)$ when the y-axis value of these pedestrians center points are the same. In order to confirm the parameters of the distribution, we firstly choose a horizontal line where the pedestrians most frequently occurred in the video, and then compute the mean value and the variance of these pedestrians height h as

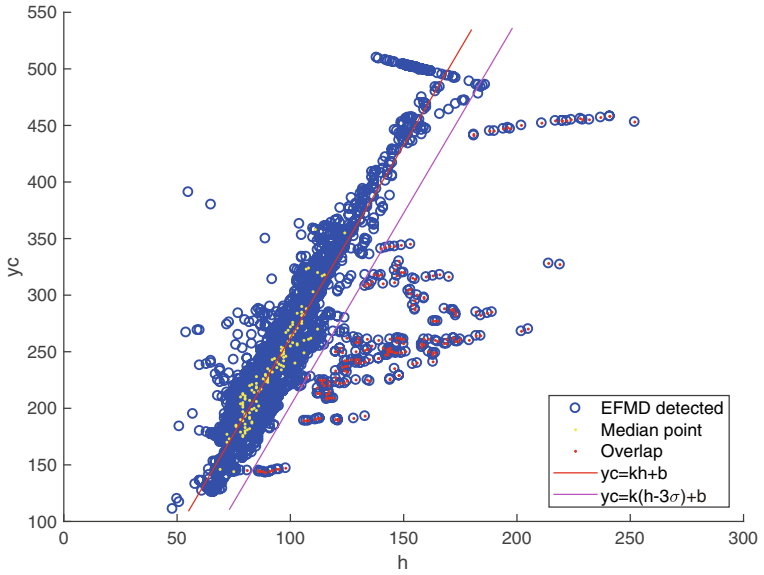


Fig. 8 The fitting line and the detected targets

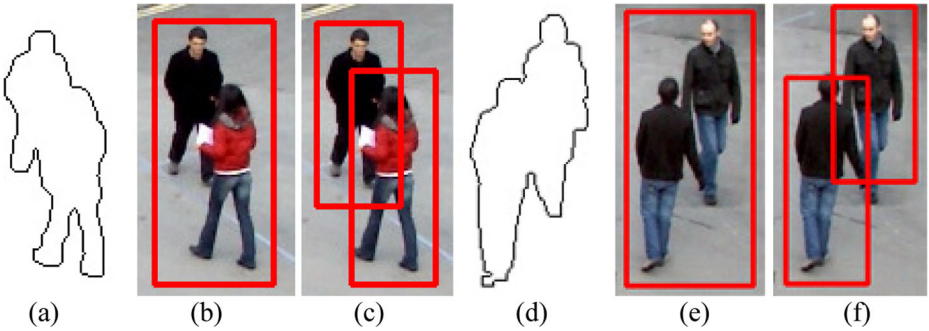


Fig. 9 Two examples of overlap handling. **a** the first contour and the moment $\eta 11 = 0.047$, **b** detection result of EFMD, **c** final result after overlap handling, **d** the second contour and the moment $\eta 11 = -0.060$, **e** detection result of EFMD, **f** final result after overlap handling

the initialization of $N(\mu,\sigma^2)$. By translating Eq. (18) to the right by 3σ, the overlap evaluation line is defined as:

$$y_c = k(h-3\sigma) + b \tag{19}$$

Algorithm 2 Linear fitting for the EFMD detected targets

Input: training frames;
Output: linear equation parameters k and b;
1. **for** each frame **do**
2. extracting EFMD features;
3. classify the feature vector by linear SVM;
4. compute height h and yc of the pedestrian bounding box BB;
5. **end for**
6. get $H=\{h_1,h_2,...h_n\}$ and $YC=\{yc_1, yc_2,...yc_n\}$ from all training frames;
7. **if** the number of the same value elements in YC is less than 10 **then**
8. abandon these elements in YC and the corresponding BBs' h in H;
9. **end if**
10. construct $YCV=\{ycv_1, ycv_2,\ldots ycv_k\}$ containing all YC elements' values;
11. **for** $i = 1$ to k **do**
12. append the same y_c value (ycv_i) BBs' h to the collection HSY_i;
13. compute the median hm_i of all elements in HSY_i;
14. **end for**
15. use the least square linear fitting to process the data $HM=\{hm_1, hm_2,\ldots hm_k\}$ and $YCV=\{ycv_1, ycv_2,\ldots ycv_k\}$;
16. **return** linear equation parameters k and b;

All the positive samples detected by the EFMD are plotted in Fig. 8, where the samples labeled with red dots represent the overlapped pedestrians. Obviously, most of the overlapped pedestrians are distributed on the right side of the line of Eq. (19), which can be utilized to determine whether the detected target is a single pedestrian. If the target is a single pedestrian, then the bounding box is output directly. If the target consists of two pedestrians that overlap each other, then the normalized central moments that analyze the pedestrian space distribution of its contour are used to calculate the bounding box of each pedestrian. The normalized

central moments contain seven moments, where the second moment $\eta 11$ reflects the spatial distribution of the overlapped target. As shown in Fig. 9, according to $\eta 11$, the pedestrians are on the lower left and the upper right, or on the upper left and the lower right conversely. Algorithm 3 explains the segmentation and the position of the two overlapped pedestrians.

Algorithm 3 Overlap handling

Input: bounding box height h, width w, x-axis value x, y-axis value y, y_c, $\eta 11$; linear equation parameter k and b;

Output: bounding boxes height h', width w', x-axis value x', y-axis value y';

1. initialize average aspect ratio r ;
2. initialize standard deviation σ of the same y_c bounding boxes heights;
3. initialize non-single classification line : $y_c = k(h\text{-}3\sigma)+b$;
4. compute the estimated height $H = (y_c - b)/k+3\sigma$;
5. **if** $h < H$ **then**
6. break;
7. **else if** $\eta 11 < 0$ **then**
8. $h'_1 = 0.9 * H$, $w'_1 = r * H$, $x'_1 = x + (w - w'_1)$, $y'_1 = y$;
9. $h'_2 = 1.1 * H$, $w'_2 = r * H$, $x'_2 = x$, $y'_2 = y + (h - H)$;
10. **else**
11. $h'_1 = 1.1 * H$, $w'_1 = r * H$, $x'_1 = x + (w - w'_1)$, $y'_1 = y + (h - H)$;
12. $h'_2 = 0.9 * H$, $w'_2 = r * H$, $x'_2 = x$, $y'_2 = y$;
13. **end if**
14. **end if**
15. **return** bounding boxes parameters $h'_1, w'_1, x'_1, y'_1, h'_2, w'_2, x'_2, y'_2$;

3.4 Classifier training

In this paper, we use linear SVM as classifier, which has been widely used in pedestrian detection system. Unlike conventional training samples, we make samples by extracting moving object contours from surveillance videos. We make a total of 7146 samples consisting of 3222 positive samples and 3924 negative samples. Some positive and negative samples are shown in Fig. 10. The negative samples are the contours of varieties of vehicles while the positive samples are the contours of moving pedestrians in various poses. In the training phase, we compute EFMD feature vectors from all training samples, and input the vectors and labels to SVM classifier for training. After the training process, the system outputs a XML file as the result, which is used to compute decision score for the classification procedure.

4 Experimental results

Since our moving target detection method is based on GMM, we use the camera fixed video dataset to test the performance of our approach. The PETS 2009 benchmark data[1] and the CAVIAR test case scenarios[2] are widely used benchmark datasets for pedestrian detection. In this paper, we choose the scenario S2_L1 with timestamp 12–34 using view 001 in PETS 2009

[1] http://www.cvg.reading.ac.uk/PETS2009/a.html
[2] http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1

Fig. 10 Some training samples

for the first experimental dataset, which consists of 795 JPEG images with 768 × 576 resolution. The second experimental dataset is a surveillance video in CAVIAR, and we choose 810 JPEG images with 384 × 288 resolution from one of the corridor view clips (OneStopEnter2cor). The validation is performed by using manually labeled ground truth. In addition, we also test our approach on a surveillance video recorded on campus to verify the robustness of our method in the scenarios which contain non-pedestrian moving objects (e.g., moving cars).

In this section, we introduce the evaluation protocol, and then examine the effectiveness of our overlap handling and compare our approach with the state-of-the-art. Furthermore, we analyze the runtimes of all detectors mentioned in Section 4.2.

4.1 Evaluation protocol

We employ the PASCAL criterion to determine whether the pedestrian is accurately detected. That is, the intersection area of the detection bounding box *BB*dt and the pedestrian ground truth bounding box *BB*gt must account for more than 50% of their union area:

$$a = \frac{area\left(BB_{dt} \cap BB_{gt}\right)}{area\left(BB_{dt} \cup BB_{gt}\right)} > 0.5 \tag{20}$$

DET curve is the most popular detection performance evaluation method in pedestrian detection, which plots miss rate versus false positives per image in logarithmic scale by

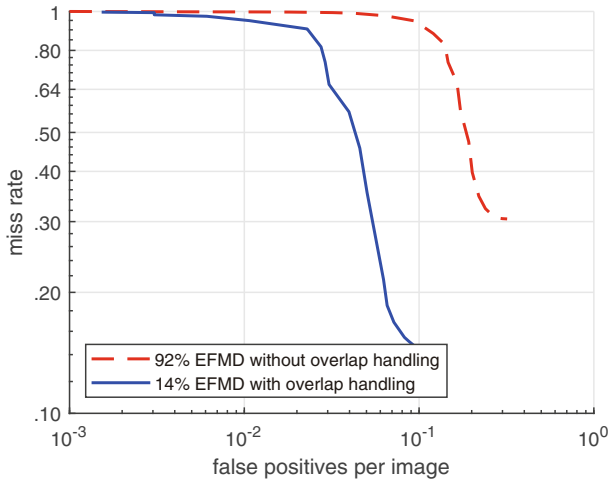


Fig. 11 Effectiveness of our overlap handling algorithm

changing the threshold of classifier in the pedestrian detector. The miss rate and the FPPI can be defined as:

$$\text{Miss Rate} = \frac{\text{the number of false negatives}}{\text{the number of positives}} \tag{21}$$

$$\text{FPPI} = \frac{\text{the number of false positives}}{\text{the number of frames}} \tag{22}$$

Lower DET curve means better performance. In addition to the DET curve, the miss rate at 10^{-1} FPPI is also regarded as a general evaluation index of detector performance, which is marked in the legend of the DET graph.

4.2 Performance evaluation

In order to examine the improvement of our overlap handling algorithm, we define two detectors: EFMD without overlap handling and EFMD with overlap handling. The latter is based on the former by adding overlap handling algorithm. In this section, we test the two detectors on PETS 2009 dataset and plot their DET curves in Fig. 11. We can see that the overlap handling reduces the miss rate significantly from 92 to 14% at 10^{-1} FPPI. Moreover, the FPPI declines as well. Generally, the EFMD with overlap handling detector shows high performance as the result of adding overlap handling.

We compare our detector with several state-of-the-art detectors: HOG [6] detector trained on the INRIA dataset; VJ [27] as a detector of Haar wavelets feature; LBP [15] as a detector of texture feature; ACF [10] as a detector of multi-channel features; Faster R-CNN [22] as a detector of deep learning. We perform an evaluation of the above detectors on PETS 2009 and CAVIAR datasets using the same evaluation protocol illustrated in Section 4.1. The DET curves of these detectors are shown in Fig. 12, and the miss rate at 10^{-1} FPPI of each detector is indicated in the legend.

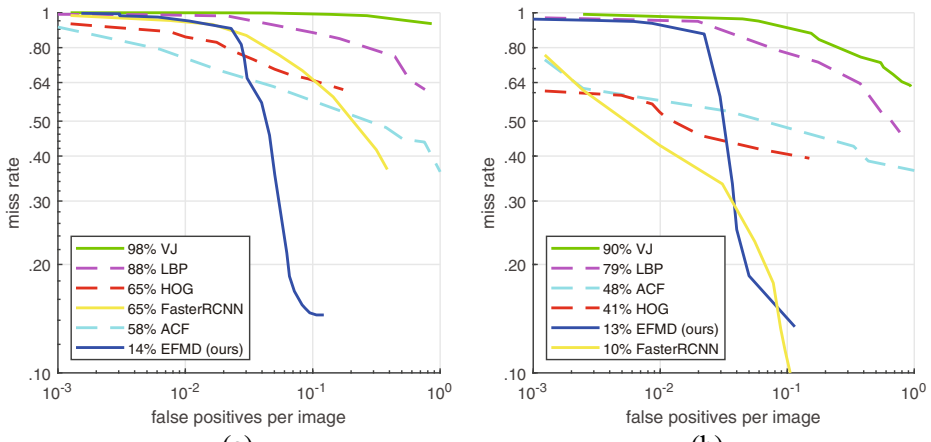


Fig. 12 Comparison with the state-of-the-art on PETS 2009 (a) and CAVIAR (b)

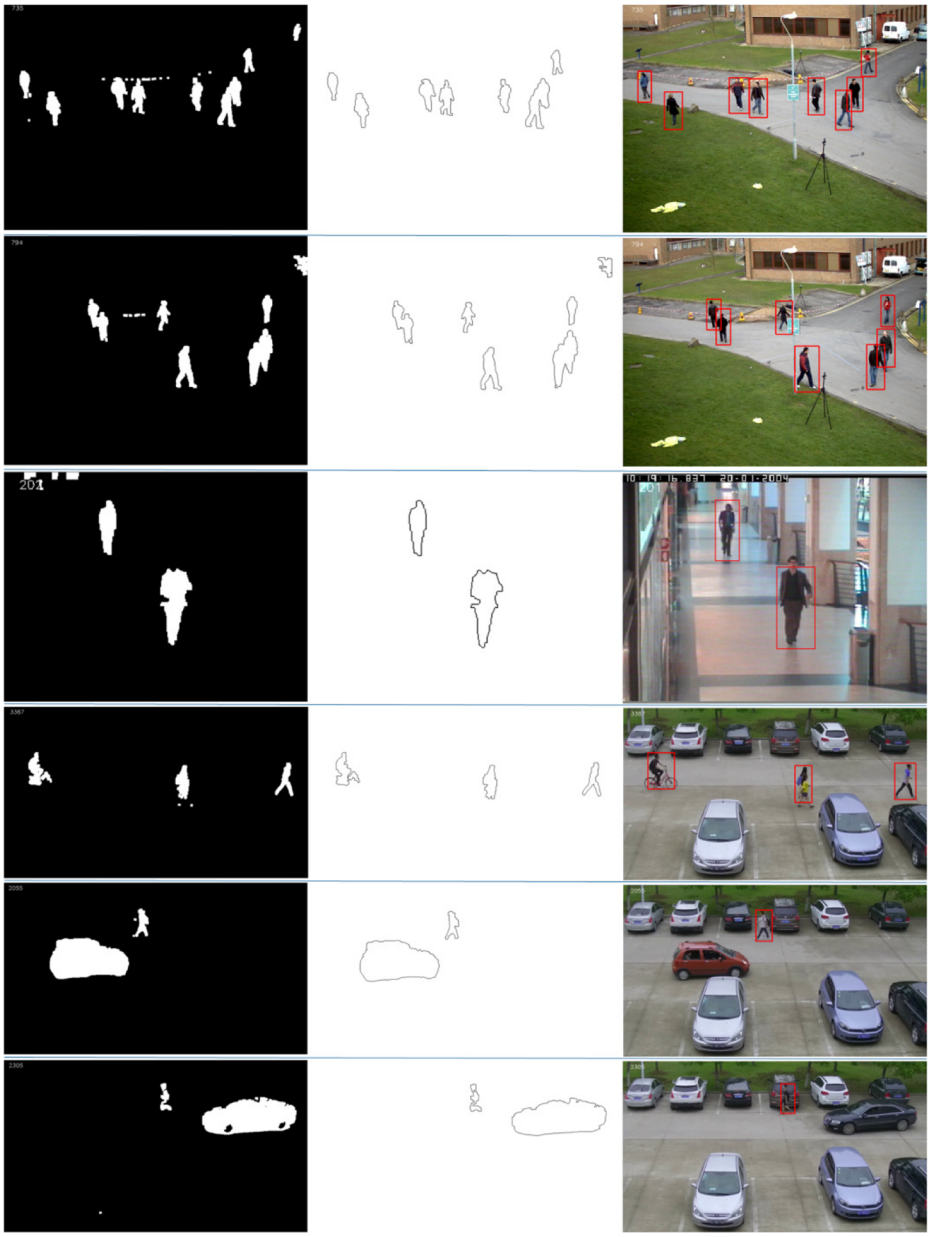


Fig. 13 Some detection results of our approach in three videos. The top two rows are extracted from the test on PETS 2009 dataset, the third row is extracted from the test on CAVIAR dataset, and the last three rows are extracted from the test on a video recorded on campus. In each row, the moving foreground, contours and final detection results are shown from left to right

From the test on PETS 2009, our detector significantly outperforms the others obviously. The DET curves show that the miss rate at 10^{-1} FPPI of our detector is 14%, while others range from 58 to 98%. Due to the low resolution and less pedestrian overlap, all the detectors

Table 2 The runtimes of all detectors

Detector	Feature	Classifier	Time(ms)/ Frame (PETS 2009)	Time(ms)/ Frame (CAVIAR)
VJ [27]	Haar	cascade	535	103
LBP [15]	LBP	cascade	70	34
HOG [6]	HOG	SVM	192	154
ACF [10]	ACF	decision trees	278	128
FasterRCNN [22]	CNN	neural network	214	211
ours	EFMD	SVM	30	25

perform better on CAVIAR dataset. From the DET curves, the miss rate at 10^{-1} FPPI of the Faster R-CNN is 10%, which is the lowest one. Our detector ranks second with the miss rate of 13% at 10^{-1} FPPI.

Generally, the evaluation results indicate that our detector shows very good performance.

As shown in Fig. 13, our approach detects pedestrians accurately and robustly despite the overlap and the changes of their postures. The pedestrian on bicycle is detected as well. In addition, our detector successfully distinguishes the pedestrian from the moving vehicles.

4.3 Runtime analysis

For fair comparison, we run programs on the same personal computer with Intel® Core™ i5–7500 CPU @ 3.4GHz and NVIDIA® GeForce™ GTX1060. The runtime test session also uses the PETS 2009 and CAVIAR datasets that mentioned above. We run 10 times for each detector and use the average time as the final result. The runtimes of all detectors are listed in Table 2.

From Table 2, we can see that our approach is the fastest detector with an average time of 30 ms per frame (768 × 576) on PETS 2009 and 25 ms per frame (384 × 288) on CAVIAR, containing only 15 ms for EFMD features computation and classification, which is owing to the low feature vector dimension. By contrast, the runtimes of other detectors range from 70 ms to 535 ms on PETS 2009 and from 34 ms to 211 ms on CAVIAR, respectively. It is worth mentioning that the running speed of our approach is 33 frames per second, which is able to achieve real-time processing.

5 Conclusion

In this paper, we proposed a real-time moving pedestrian detector and an overlap handling method. The main idea is to utilize GMM to obtain foreground for extracting the contours of moving objects, and then the elliptic Fourier descriptors and the normalized central moments are combined in series to form the EFMD features. Moreover, the overlap evaluation line and the moment $\eta 11$ are used for overlap handling. Linear fitting is introduced to fit the overlap evaluation line by processing a large number of detected targets.

Our approach significantly reduced the detection computation by focusing on moving foreground and using efficient contour descriptors, which achieved real-time processing. Furthermore, an overlap handling method based on linear fitting and normalized central moments improved the detection performance by reducing false positives and miss rate. The experimental results on PETS 2009 and CAVIAR dataset showed that our approach significantly outperformed the conventional pedestrian detectors in both detection performance and runtime. However, our approach cannot handle the target that consists of more than three overlapped pedestrians, which is the next problem to be solved.

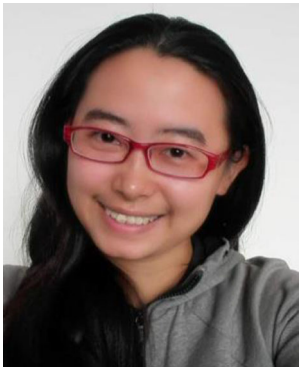
Acknowledgments This work was supported by the Fundamental Research Funds for the Central Universities (No.2014ZDPY32).

References

1. Barron JL, Fleet DJ, Beauchemin SS, Burkitt TA (1994) Performance of Optical Flow Techniques. Int J Comput Vis 12(1):43–77
2. Chang X, Yang Y (2017) Semisupervised feature analysis by mining correlations among multiple tasks. IEEE Trans Neural Netw Learn Syst 28(10):2294–2305
3. Chang X, Ma Z, Lin M, Yang Y, Hauptmann A (2017) Feature Interaction Augmented Sparse Learning for Fast Kinect Motion Detection. IEEE Trans Image Process 26(8):3911–3920
4. Chang X, Ma Z, Yang Y, Zeng Z, Hauptmann A (2017) Bi-level semantic representation analysis for multimedia event detection. IEEE Trans Cybern 47(5):1180–1197
5. Chang X, Yu YL, Yang Y, Xing EP (2017) Semantic pooling for complex event analysis in untrimmed videos. IEEE Transactions on Pattern Analysis & Machine Intelligence 39(8):1617–1632
6. Dalal N, Triggs B (2005) Histograms of Oriented Gradients for Human Detection. IEEE Computer Society Conference on Computer Vision & Pattern Recognition 1(12):886–893
7. Dollár P, Tu Z, Perona P, Belongie S (2009) Integral channel Features. British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings DBLP
8. Dollár P, Belongie S, Perona P (2010) The fastest pedestrian detector in the West. British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings DBLP, 2010:1–11
9. Dollár P, Wojek C, Schiele B, Perona P (2011) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34(4):743–761
10. Dollár P, Appel R, Belongie S, Perona P (2014) Fast feature pyramids for object detection. IEEE Trans Pattern Anal Mach Intell 36(8):1532–1545
11. Gavrila DM (2007) A Bayesian Exemplar-based approach to hierarchical shape matching. IEEE Trans Pattern Anal Mach Intell 29(8):1408–1421
12. Granlund GH (1972) Fourier preprocessing for hand print character recognition. IEEE Trans Comput 21: 195–201
13. Hinton GE (2009) Deep belief networks. Scholarpedia 4(6):5947
14. Li Z, Nie F, Chang X, Yang Y (2017) Beyond trace ratio: weighted harmonic mean of trace ratios for multiclass discriminant analysis. IEEE Trans Knowl Data Eng 99:1–1
15. Liao S, Zhu X, Lei Z, Zhang L, Li SZ (2008) Learning Multi-scale Block Local Binary Patterns for Face Recognition. Advances in Biometrics 4642:828–837
16. Lin Z, Davis LS (2008) A pose-invariant descriptor for human detection and segmentation. European Conference on Computer Vision Springer-Verlag, 2008:423–436
17. Liu Y, Chen X, Yao H, Cui X, Liu C, Gao W (2009) Contour-motion feature (CMF): A space–time approach for robust pedestrian detection. Pattern Recogn Lett 30(2):148–156
18. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
19. Mohan A, Papageorgiou C, Poggio T (2001) Example-Based Object Detection in Images by Components. IEEE Trans Pattern Anal Mach Intell 23(4):349–361
20. Paisitkriangkrai S, Shen C, Hengel AVD (2016) Pedestrian Detection with Spatially Pooled Features and Structured Ensemble Learning. IEEE Trans Pattern Anal Mach Intell 38(6):1243–1257

21. Papageorgiou C, Poggio T (2000) A trainable system for object detection. Int J Comput Vis 38(1):15–33
22. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. International Conference on Neural Information Processing Systems 39: 91–99
23. Sabzmeydani P, Mori G (2007) Detecting pedestrians by learning shapelet features. Computer Vision and Pattern Recognition, 2007. IEEE Conference on IEEE, 2007:1–8
24. Sermanet P, Kavukcuoglu K, Chintala S, LeCun Y (2013) Pedestrian detection with unsupervised multi-stage feature learning. IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 2013:3626–3633
25. Shen J, Zuo X, Li J, Yang W, Ling H (2017) A novel pixel neighborhood differential statistic feature for pedestrian and face detection. Pattern Recogn 63:127–138
26. Stauffer C, Grimson WE (1999) Adaptive background mixture models for realtime tracking. Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on IEEE Xplore, 1999(2):252
27. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
28. Walk S, Majer N, Schindler K, Schiele B (2010) New features and insights for pedestrian detection. IEEE Conference on Computer Vision and Pattern Recognition 2010, 119(5):1030–1037
29. Wang X, Han TX (2009) An HOG-LBP human detector with partial occlusion handling. Proc.IEEE Int.conf.on Computer Vision Kyoto Japan Sept, 30(2):32–39
30. Wojek C, Schiele B (2008) A performance evaluation of single and multi-feature people detection. In: Proceedings of the symposium of the german association for pattern recognition (DAGM)
31. Wu B, Nevatia R (2008) Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. Computer Vision and Pattern Recognition, 2008. IEEE Conference on IEEE, 2008:1–8
32. You X, Du L, Cheung YM, Chen Q (2010) A blind watermarking scheme using new nontensor product wavelet filter banks. IEEE Trans Image Process 19(12):3271–3284
33. Zhang S, Bauckhage C, Klein D, Cremers A (2016) Fast moving pedestrian detection based on motion segmentation and new motion features. Multimed Tools Appl 75(11):6263–6282
34. Zhu Z, You X, Chen CL, Tao D, Ou W, Jiang X, Zou J (2015) An adaptive hybrid pattern for noise-robust texture analysis. Pattern Recogn 48:2592–2608

Kai Zhao received the master's degree in control engineering from China University of Mining and Technology, Xuzhou, China, in 2016. He joined the Video Lab at China University of Mining and Technology as a PhD candidate in 2016. His general interests lie in machine learning and pattern recognition and their application to computer vision.



Jingjing Deng received the Master Degree in Control Engineering from China University of Mining and Technology, Xuzhou, China, in 2016. She is currently a doctor candidate in China University of Mining and Technology. Her research interests include machine learning, artificial intelligence, Terahertz Time-Domain Spectroscopy data analysis and linearization of sensor characteristics.

Deqiang Cheng received the PhD degree from China University of Mining and Technology, China, in 2007. He is currently a professor at China University of Mining and Technology. He was a visiting scholar of "Key Laboratory of Intelligent Perception and Image Understanding" of Xidian University in 2012 and a visiting professor of University of Alberta in 2014. His research interests include image processing and pattern recognition.