ISBN 2354-0575

A FLEXIBLE APPROACH FOR REAL-TIME PEDESTRIAN DETECTION WITH FOREGROUND-BASED CASCADE CLASSIFIER

Hong-Son Vu[1], **Kuan-Hung Chen**[2]

[1]*Hung Yen University of Technology and Education*

[2]*Feng Chia University*

Received: 01/10/2016
Revised: 31/10/2016
Accepted for publication: 15/11/2016

Abstract

Almost all existing state-of-the-art pedestrian detection methods require heavy computing cost from their feature descriptors, which cannot detect pedestrians reliably in real-time. In this paper, we take advantage of Background Subtraction (BS) technique to extract moving objects region on whole natural scene images in complicated environments. Then, Haar-like or Histograms of Oriented Gradients (HOG) features are used to classify the detected moving objects to the categories they belong to. The proposed fusion method achieves a speedup of at least 4.5x compared to conventional approaches based on Haar-Like and HOG descriptors only for high resolution images (768 x 576), with detection rate of 97.76% and a minor false detection rate of 2.66%.

Keywords: *moving object detection, pedestrian detection, fusion method.*

1. Introduction

Pedestrian detection is one of the most important tasks in computer vision, with several applications that may be potential to positively influence quality of human life [1], such as video surveillance, advanced driver assistance systems, and intelligent robotics. Therefore, detecting and tracking pedestrian is an important domain of research. Nevertheless, pedestrian detection is still challenging due to their variety in pose, clothing, illumination variations, articulation, partial occlusion, shadow, and complicated background in the real-world environments.

In general, the objective of pedestrian detection is to determine the presence of human in natural scene images and videos, and then return information about their locations and sizes. To obtain a reliable pedestrian detection, a robust feature set describing visual human recognition is required. These feature sets have been proposed by researchers, such as Haar-like features [2], HOG [3], and combination of Haar-like features along with HOG descriptor [4]. These descriptors along with AdaBoost and Support Vector Machine (SVM) classifiers can be reliably classified the detected objects into human or non-human.

In HOG-based pedestrian detection method, the processing unit is a 64x128-pixel detection window that divided into 7 blocks horizontally and 15 blocks vertically, for a total of 105 blocks. Each block contains 4 cells with a 9-bin histogram for each cell. Thus, a detection window comprises 7 x 15 x 4 x 9 = 3780 values. The HOG algorithm applies the sliding window technique in order to slide the detection window from left-to-right and top-to-bottom across the whole image. Although HOG-based pedestrian detection method achieves excellent detection results, its heavy computing cost requirements makes the system cannot detect objects in real-time. Viola and Jones have proposed a boosted cascade of simple features for rapid object detection [2]. Nevertheless, the proposed techniques in [2] would generate many false alarms on the whole scene images. Recently, for moving objects detection, Background Subtraction (BS) techniques are well known for its rapid processing time, precisely and robustly performance in a fixed camera scene [5].

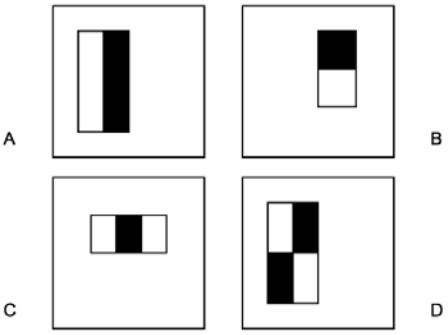
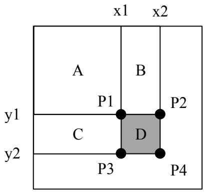


Fig.1. *Haar-like feature masks [2]. Two-rectangle features are illustrated in (A) and (B). Three-rectangle and four-rectangle features are respectively shown in (C) and (D).*



Rectangle D = P4 – P2 – P3 + P1, P2, P3, P4 are the values of th image at coordinates (x1,y1), (x1,y2), and (x2,y2), respectively.

Fig.2. *Calculation of the sum of the pixels within rectangle D. P1 is the sum of the pixels in rectangle A. P2 is the sum of the pixels in rectangle A and rectangle B. P3 is the sum of the pixels in rectangle A and rectangle C. P4 is the sum of the pixels in rectangle A, rectangle B, rectangle C, and rectangle D.*

Although these methods are very robust and can achieve a high detection rate because of their exhaustive search strategy at all potential candidate regions, they are not able to meet for real-time applications. Therefore, in this paper, we aim to deal with real-time pedestrian detection by fusing the advantages of these types of methods. The proposed fusion method consists of detection and classification modules, i.e., BS technique for detection task, AdaBoost or SVM classifiers for recognition task. To reduce the search space and detection time across the whole scene image, we first identify possible moving objects proposals based on motion information. For appearance-based pedestrian detection, we use Adaboost or SVM classifiers. As a result, the proposed method can detect pedestrian in real-time (24 frames per second) with excellent detection rate. Experimental results show that the proposed fusion method achieves a speedup of at least 4.5x compared to conventional approaches based on Haar-Like and HOG descriptors only for high resolution images (768 x 576), with detection rate of 97.76% and a minor false detection rate of 2.66%.

The rest of this paper is organized as follows. Previous pedestrian detection algorithms are reviewed in Section II. The fusion techniques of the proposed work are described in Section III. Section IV presents experimental results and performance comparison. Finally, the conclusion is drawn in Section V.

2. Related Work

Viola and Jones proposed a rapid and robust object detector by using the AdaBoost algorithm [2]. They proposed a feature extraction method, i.e., Haar-like feature for weak classifiers, and a cascade structure of a classifier to obtain rapid object detection. In their method, a strong classifier is represented by combining many weak-classifiers. In other words, this strong classifier is formed as a linear combination of weighted results of weak classifiers, and the weights of weak classifiers are trained by a large number of positive and negative sample images. The combination of the strong classifiers in a cascade leads to high precision rate and computational efficiency. Since object detection extracts a lot of candidate regions that need to be calculated and classified, the computation cost for each region should be kept in small level. For such requirements, the AdaBoost-based algorithm can achieve accurate classification with small computational cost. This technique can accelerate computation time by determining if the sample is a successful candidate to move on to the next stage or rejecting the negative sub-windows that do not include objects of interest, so that the detector only concentrates on successful candidates.

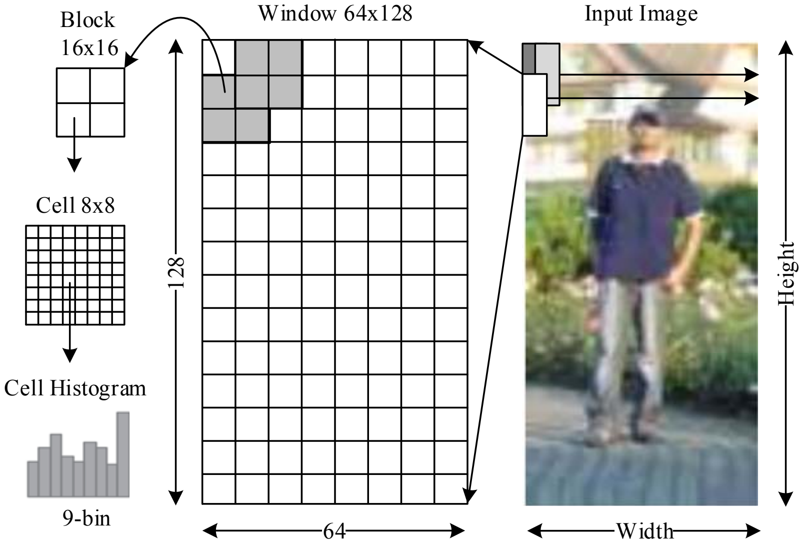


Fig.3. *An illustration of HOG descriptor for sliding detection window. An input image is divided into overlapping 64x128 pixels detection windows. The detection window is then partitioned into overlapping blocks that consist of 2x2 cells. Each cell is presented by 9 bins of gradient orientation histogram.*

Haar-like features are extracted by calculating the difference of the sums of the pixel values at the corresponding location of the black and white rectangles. The features are extracted by sliding four Haar-like masks on the whole input images. These four kinds of Haar-like feature masks are shown in Fig. 1. Viola and Jones also proposed a new image representation called an "integral image" to accelerate calculation of the sums of the pixel values in the black and white rectangles. The integral image *ii(x,y)* at location *(x,y)* contains the sum of the pixels to the left and above of that point, as shown in Fig. 2.

$$ii(x,y) = \sum_{x' \le x, y' \le y} i(x',y') \quad (1)$$

Where *i(x,y)* is a pixel value in the original image. In the integral image, the sum of pixels in the region from (x_1,y_1) to (x_2,y_2) is presented as follows:

$$r(x_1,x_2,y_1,y_2) = \sum_{x_1 \le x' \le x_2, y_1 \le y' \le y_2} i(x',y')$$

$$= \sum_{x' \le x_2, y' \le y_2} i(x',y') - \sum_{x' \le x_2, y' \le y_1} i(x',y') - \sum_{x' \le x_1, y' \le y_2} i(x',y') + \sum_{x' \le x_1, y' \le y_1} i(x',y')$$

$$= ii(x_2,y_2) - ii(x_2,y_1) - ii(x_1,y_2) + ii(x_1,y_1) \quad (2)$$

HOG feature along with SVM classifier that have been introduced by Dalal and Triggs [3] is the most widely used approach for pedestrian and object detection currently. In their approach, HOG feature is extracted by the following steps. First, an input image is divided into overlapping 64x128 pixels detection windows. Then, the detection window is segmented into 7x15 blocks that are further divided into 2x2 cells. Next, the direction and magnitude of the gradient in each cell is calculated and then the histogram of each block can be achieved through accumulating the direction and magnitude of the gradient in all cells of the block. Finally, the histograms of all blocks are concatenated into final feature vector of 3780 values. An illustration of HOG descriptor is further depicted in Fig. 3.

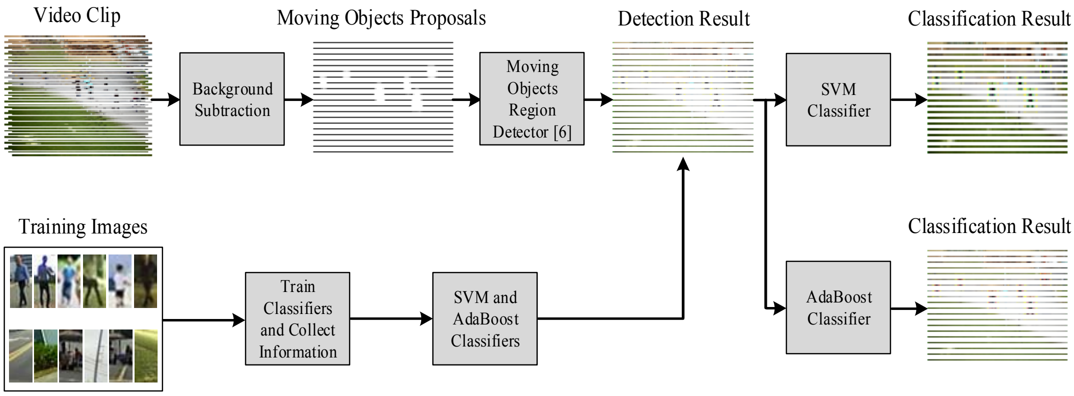


Fig.4. *The framework of the proposed fusion method.*

Although these methods are very robust and can achieve a high detection rate because of their exhaustive search strategy at all potential candidate regions, they are not able to meet for real-time applications. For moving objects detection, BS techniques are well known for its rapid processing time, precisely and robustly performance in a fixed camera scene [5]. Therefore, in this paper, we aim to deal with real-time pedestrian detection by fusing the advantages of these types of methods. To reduce the search space and detection time across the whole scene image, we first identify possible moving objects proposals based on motion information. For appearance-based pedestrian detection, we use Adaboost or SVM classifiers.

3. Proposed Method

The framework of the proposed fusion method is illustrated in Fig. 4, where yellow rectangles denote moving objects proposals, red rectangles and green rectangles respectively show pedestrian detection results using AdaBoost and SVM classifiers. The goal of detection module is to propose the positions of moving objects in natural scene images. Conventional approaches often use the sliding detection window based on either HOG features along with SVM classifier or Haar-like features along with AdaBoost classifier to classify the detected objects into their categories, where the HOG and Haar-like descriptors slide the detection window from left-to-right and top-to-bottom across the whole scene image. Both these approaches lead to a high computation cost, this results from their large search space strategy on whole scene images, where the probability to find desired objects is not always existed. This paper proposes an approach using BS technique to reduce search space for detection module. These techniques are really helpful not only to reduce the number of candidate regions, but also to avoid extracting regions such as sky or regions of interests (ROIs) inconsistent with perspective, which generate the potential number of false alarms.

The steps of the proposed method are as follows: First, BS technique is used to determine moving objects proposals, as shown in Fig. 4. Second, under some critical situations such as complicated background, shadows, and illumination variations, the detected moving objects around foreground are partitioned into separate parts. This leads to failure in determining moving objects proposals for classification module, since moving objects proposals in such cases may only contain separated parts such as head-shoulder, torso, or legs, etc. To conquer this problem, we adopt the proposed technique in [6] to merge the bounding boxes around foreground objects. Finally, moving objects proposals are classified into their categories using AdaBoost and SVM classifiers, as illustrated in Fig. 4.

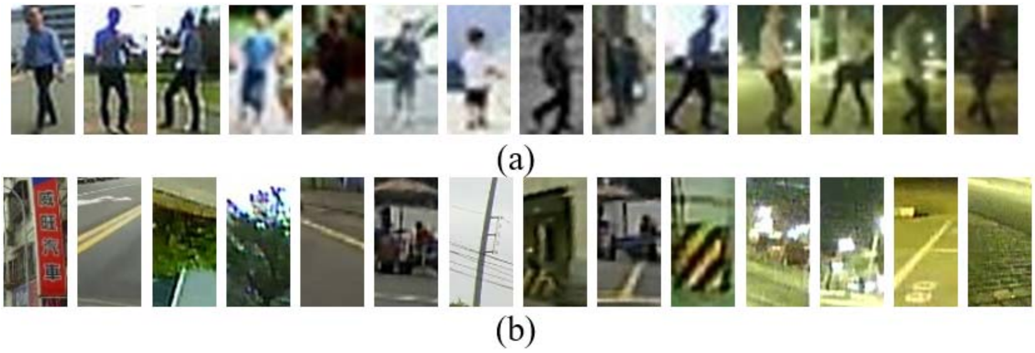


Fig. 5. Some training samples from the dataset. (a) Positive samples, (b) Negative samples.

Table 1. *Performance Evaluation on our Dataset, With the Resolution of 768x576. Scaling Factors of Adaboost and Svm Classifiers are respectively 1.1 and 1.03.*

Input videos	Total number of frames	Detection rate (%)	Miss rate (%)	False detection rate (%)	Processing speed (FPS)	Methods
In the daytime (multi-object moving)	795	80.23	19.77	93.51	1.7	Haar-like
		100	0	0	40	BS
		97.76	2.24	2.66	24	Fusion
In the daytime (multi-object moving)	795	56.06	43.94	0	2	HOG
		100	0	0	40	BS
		63.59	36.41	0	9	Fusion

4. Experimental Results

In our experiments, the training dataset contains two sets, i.e., 1) the first one consists of 27,596 positive pedestrian samples and 12,960 negative samples in the daytime, and 1,008 positive pedestrian samples and 1,853 negative samples at night; 2) the second one comes from the INRIA training dataset. At first, we use all training images in the first set to train the AdaBoost classifier, while the SVM classifier is trained by the INRIA person dataset, as described in [3]. Some training samples from the dataset are depicted in Fig. 5. To evaluate the proposed method in practical scenarios, we collect a dataset from natural scene images in complicated environments with a high-resolution for surveillance. The dataset comprises a total number of frames of 795 with the corresponding resolution of 768 x 576, where a large number of pedestrians with variety in pose, clothing, articulation, partial occlusion, and complicated background. Our experiments are conducted in an Intel Core i7-3770 CPU at 3.40 GHz and 16G DDR2 memory. The code has parts in C++ (i.e. background subtraction method and moving objects region detector) and others (i.e. AdaBoost and SVM classifiers) in OpenCV library. No parallel implementation or specific algorithm optimization are used in experiments. In addition, we define the Detection Rate (DR), Miss Rate (MR), and False Detection Rate (FDR) as three performance indexes for evaluating the proposed fusion method in our datasets. The equations are expressed as follows:

$$DR = \frac{\#\,of\ True\ Positives\ detected\,(TP)}{Total\ \#\,of\ \ Pedestrian\ Collections\,(TPC)} * 100\% \quad (3)$$

$$MR = 100\% - DR \quad (4)$$

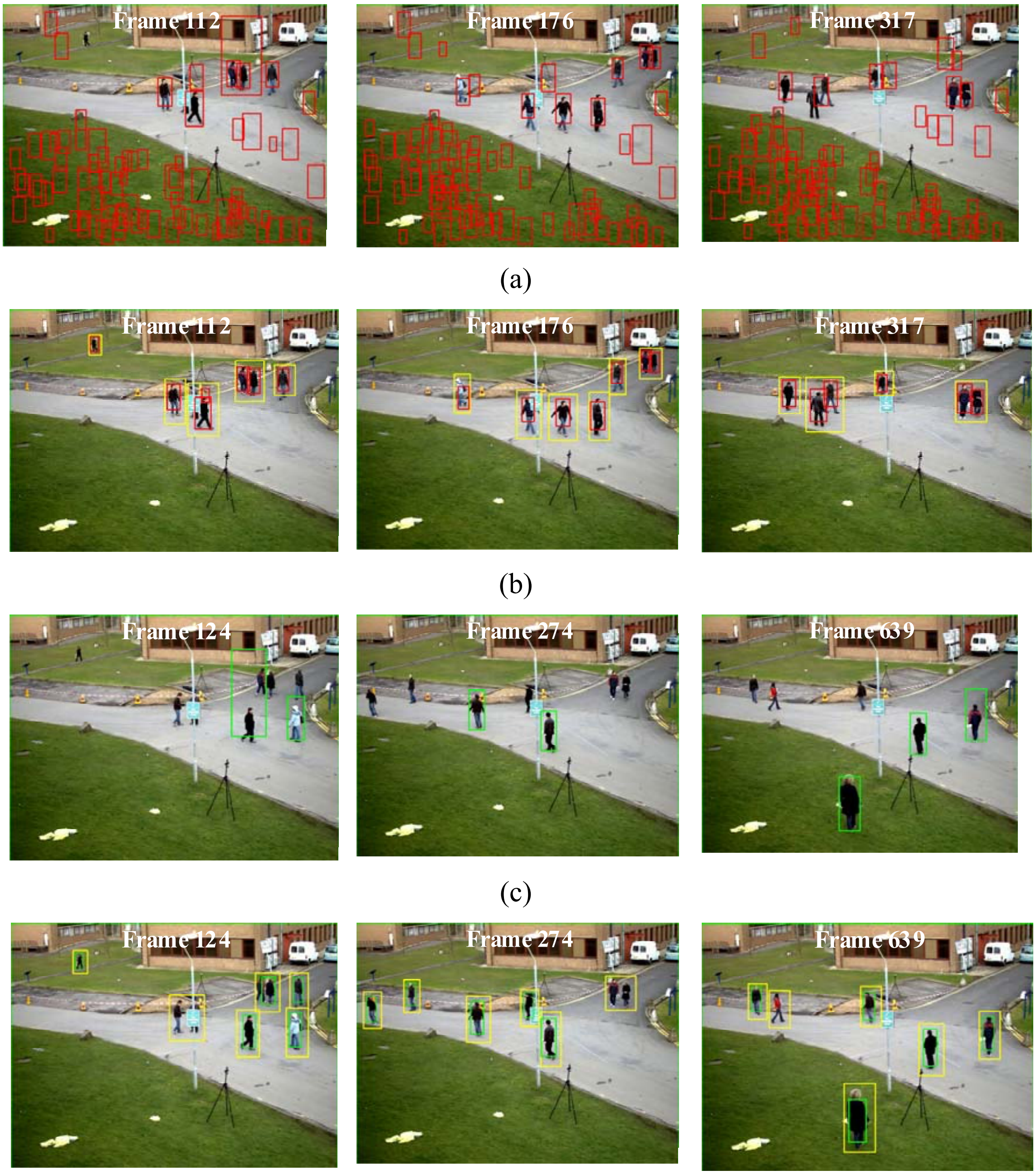
$$FDR = \frac{\#\,of\ \ False\ Positives\ detected\,(FP)}{TP + FP} * 100\% \quad (5)$$

where *TPC* presents total number of pedestrian collections, *TP* is true positive illustrating the number of the pedestrian samples that are detected as pedestrians, *FP* is false positive presenting the number of the non-pedestrian samples that are detected as pedestrians.

Experimental results show that our approach can speed up at least 4.5 times as compared to conventional methods, with significantly improved detection rate, i.e., 17.53% detection rate increment and 90.85% false detection rate decrement. Table I shows the detection rate, miss rate, false detection rate, and processing speed of the proposed fusion method when compared to those of the classical HOG/SVM and Haar-like/AdaBoost pedestrian detectors. It is valuable to mention that pedestrian detection is still challenging in pattern recognition and computer vision due to its heavy computation requirements for accurate and robust recognition against complicated environments, and especially real-time implementation. Despite these challenges, our method can detect and classify pedestrian at 24 Frames Per Second (FPS) on 768x576 images. This makes the proposed method possible to be applied to real-time automated surveillance systems. The detected results using the classical methods and the proposed fusion methods are further illustrated in Fig. 6.

5. CONCLUSION

In this paper, we aim to address the problem of real-time pedestrian detection. Through fusing the advantages of BS technique and the classical pedestrian detectors, the proposed fusion method is really robust not only to improve the processing time and detection rate, but also to significantly reduce the false detection rate. Experimental results show that the proposed method can speed up at least 4.5x as compared to conventional methods for high resolution images (768 x 576), with detection rate of 97.76% and a minor false detection rate of 2.66%. This method has highly potential to be applied on real conditions that include moving objectssuch as automated surveillance systems. A possible extension of this work is real-time implementation on embedded systems.



(a)

(b)

(c)

(d)

Fig.6. *Pedestrian detection results by different detectors. (a) Haar-like/AdaBoost,(b) Fusion result of BS/AdaBoost, (c) HOG/SVM, and (d) Fusion result of BS/SVM.*

References

[1] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, 2012.

[2] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. Comput. Vis. Patt. Recognit.* (CVPR), 2001, pp. 511–518.

[3] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Comput. Vis. Patt. Recognit.* (CVPR), 2005, pp. 886–893.

[4] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. Comput. Vis. Patt. Recognit.* (CVPR), 2006, pp. 1491–1498.

[5] S. J. Noh, M. Jeon, "A new framework for background subtraction using multiple cues," in *Proc. 11th Asian Conf. on Comput. Vis.*, 2013, pp. 493–506.

[6] H. S. Vu, J. X. Gou, K. H. Chen, S. J. Hsieh, and D. S. Chen, "A real-time moving objects detection and classification approach for static cameras," in *Proc.IEEE Int. Conf. on Consumer Electronics-Taiwan (ICCE-TW)*, 2016, pp. 1–2.

MỘT CÁCH THỨC TIẾP CẬN MỀM DẺO CHO PHÁT HIỆN NGƯỜI ĐI BỘ THỜI GIAN THỰC VỚI BỘ PHÂN LOẠI TẦNG CĂN CỨ VÀO TIỀN CẢNH

Tóm tắt:

Phần lớn các phương pháp phát hiện người đi bộ hiện đại hiện nay yêu cầu chi phí tính toán cao từ các bộ mô tả đặc trưng của chúng, cái không thể phát hiện người đi bộ đáng tin cậy ở thời gian thực. Ở bài báo này, chúng tôi lấy kỹ thuật thuận lợi của phương pháp trừ nền để rút trích vùng các đối tượng đang di chuyển ở toàn bộ các ảnh nền tự nhiên trong môi trường phức tạp. Sau đó, các đặc trưng Haar-like và HOG là được sử dụng để phân loại các đối tượng di chuyển đã được phát hiện vào loại chúng thuộc về. Phương pháp hợp nhất đã đề xuất đạt được tốc độ nhanh hơn ít nhất 4.5 lần khi so sánh với các cách tiếp cận truyền thống căn cứ chỉ vào các bộ mô tả Haar-like và HOG cho các ảnh độ phân giải cao (768x576), với tỉ lệ phát hiện 97.76% và một tỉ lệ phát hiện lỗi thấp 2.66%.

Từ khóa: *Phát hiện đối tượng đang di chuyển, phát hiện người đi bộ, phương pháp hợp nhất.*