Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 820203, 8 pages
http://dx.doi.org/10.1155/2014/820203

Research Article

Video Pedestrian Detection Based on Orthogonal Scene Motion Pattern

Jianming Qu, Zhijing Liu, and Wenhua He

School of Computer Science and Technology, Xidian University, Xi'an 710071, China

Correspondence should be addressed to Jianming Qu; sancoder.q@gmail.com

Received 29 April 2014; Accepted 12 June 2014; Published 8 July 2014

Academic Editor: Yuping Wang

In fixed video scenes, scene motion patterns can be a very useful prior knowledge for pedestrian detection which is still a challenge at present. A new approach of cascade pedestrian detection using an orthogonal scene motion pattern model in a general density video is developed in this paper. To statistically model the pedestrian motion pattern, a probability grid overlaying the whole scene is set up to partition the scene into paths and holding areas. Features extracted from different pattern areas are classified by a group of specific strategies. Instead of using a unitary classifier, the employed classifier is composed of two directional subclassifiers trained, respectively, with different samples which are selected by two orthogonal directions. Considering that the negative images from the detection window scanning are much more than the positive ones, the cascade AdaBoost technique is adopted by the subclassifiers to reduce the negative image computations. The proposed approach is proved effectively by static classification experiments and surveillance video experiments.

1. Introduction

Pedestrian video detection as one of the difficulties and hotspots in the field of computer vision has attracted increasing attention in recent years. The purpose of video pedestrian detection is to find out whether there are pedestrians in the scene and then where they exactly are. The results of detection can be regarded as the input for higher level analysis including pedestrian tracking, individual behavior recognition, and crowd motion learning or as the update knowledge and further validation information for more precise detections. Used for pedestrian detection or combined with higher level visual computations, these algorithms play important roles in a wide range of application domains like security surveillance and autonomous driving.

Although significant advancement of the research has been made in the past decade, detecting individuals in fixed scenes is still a challenge for several reasons such as the selectivity performed by pedestrians which are forced by the layout of scenes (e.g., zebra crossings or barriers), the directivity of pedestrian caused by the attraction of scenes in nature (e.g., store entrances or bus station), multistate of pedestrian caused by the interobject interactions (e.g., occlusion or pushing), and the uncertainty of pedestrian behavior (e.g., suddenly turning back).

To overcome some of these challenges, a feasible idea is to take different detection strategies for the observed object in different states. The idea is predicated on a statistical regularity reflected by the fact that pedestrians in specific region of scene regularly perform similar behaviors. For instance, people standing at a bus station or an entrance usually face the same direction. In order to increase the accuracy of the pedestrian detection by reasonably making use of the scene pattern, we have developed a novel algorithm in this paper based on this statistical regularity by learning the influences of the space layout and environment on pedestrian in general density scenes.

According to our algorithm, different regions of scene are simply divided into two classes as path and holding area based on two phenomena. (1) Objects in path area usually perform stronger directivity than those in holding area. (2) Objects in path area move faster than those in holding area statistically since most of them have specific destination. To utilize the phenomena mathematically, the scene is overlaid with a grid composed of cells. A 2D histogram is defined for each cell as scene motion pattern to describe the speed

and the direction of the object. To calculate the histograms, a crowd Kanade-Lucas-Tomasi (KLT) [1] tracking algorithm is employed before the detection. And the cells are labeled by categories based on the statistic tracking results.

Furthermore, a set of overcomplete features introduced in [2] including both upright and 45° rotated Haar-like features is applied in our approach. Since most of detection algorithms are using a slidable window with all possible scales scanning over the whole scene image, the feature calculation usually costs considerable time and system resources. However, this efficiency problem of the algorithm is overcome significantly by using the integral channel features according to Dollar et al. [3].

Instead of using a unitary classifier, the employed classifier is composed of two directional subclassifiers trained, respectively. The subclassifiers share the same negative training sample sets, while the two kinds of positive sample sets of them are built according to the different pedestrian motion directions of the samples, that is, vertical and horizontal directions (resp., parallel to the image columns and rows). And then a group of pattern-classify mapping strategies are adopted to input the feature vectors of sliding windows to the subclassifiers. Because of the common characteristics of the feature vectors, it is shown experimentally that the separate training algorithm can give a better accuracy rate than that of the traditional ones.

The subclassifiers are based on boosting technique and, more specifically, the AdaBoost algorithm which is numerically robust, rapidly converging, and scale controllable. The main idea of boosting is that a set of weak classifiers are combined to find a highly accurate hypothesis. In general density scenes, the sliding images containing pedestrians only occupy a quiet small portion of the whole input images. Thus, large numbers of negative images need to be classified and rejected rapidly. And the procedures can be further accelerated by using the cascade method [4] in which a negative image sample needs only one of the layers to be rejected, but not all of them.

The rest of the paper is structured as follows. Firstly, the related work is introduced in Section 2. Then our algorithm based on a model named orthogonal scene motion pattern (OSMP) and a detector named orthogonal cascades classifier is introduced and the idea of the algorithm is discussed theoretically and mathematically in Section 3. Subsequently, two steps of experiments are designed to verify the algorithm in Section 4. The final conclusions are given in Section 5.

2. Related Work

In resent year, a significant amount of research has been put in the pedestrian detection field. The overwhelming majority of the pedestrian detection methods use classification algorithms of machine learning to detect people in a sliding window with single or multiple scales scanning over the entire image [4–7]. The features extracted from sliding window, like HOG [5] or Haar-like wavelet [8], are sent to a classifier trained previously on the set of labeled samples. However, in previous studies, very few pedestrian detection algorithms have taken the prior knowledge of scenes environment and the historical information of crowd motion into consideration. Since there are lots of achievements in the study of motion patterns of environment and scene [9–11], putting this knowledge in pedestrian detection is an extremely rewarding work. In [9], the concept of the floor field has been defined as the influence of the environment and the crowd on an individual, and a scene structure based force model which consists of three types of floor fields has been introduced for tracking in crowd scenes. Rodriguez et al. have assumed that pedestrian motion in any location of the scene is generated by a set of behavior proportions and employed the Correlated Topic Model to solve the pedestrian tracking problem in crowd scenes [10].

Meanwhile, in the image classification field, a lot of significant work has been done either on classification techniques [4, 12–17] or on feature analysis area [4–6, 8]. Since Freund and Schapire proposed AdaBoost algorithm in [14], much attention has been given. In [15], the real-valued confidence-rated prediction $h(x)$ has been used, and the predicted label assigned to instance x has been described by the sign of $h(x)$. They have also proposed, proved, and applied the theorem that minimizing the normalization factor Z_t on each round of boosting is the key to minimize training error as Z_t is related to the upbound of training error. Friedman et al. have noticed the boosting classification problem as an additive logistic regression model and introduced Gentle AdaBoost [13] classifier which has been proved to outperform the Discrete AdaBoost and Real AdaBoost by Lienhart et al. [2]. On the other hand, in the subject of feature extraction, lots of good algorithms, including Histograms of Oriented Gradients, Haar-like wavelet, and HOG-LBP, have been created and improved.

Nevertheless, as a matter of fact, the usage of motion patterns in combination with classifiers to the video pedestrian detecting and tracking fields is far from the completeness and the final success. There has been still a long way to go in pursuing the high accuracy and efficiency of the detection. In order to emphasize the direction characteristic of the scene motion patterns in local regions and to strengthen the correlations between classifier and the direction characteristic, an OSMP model and a specific designed classifier are introduced in this paper. The OSMP containing the information of the environment and the history of pedestrian motion provides a reliable prior knowledge at the key stage of the classification and is indisputably beneficial to improve the accuracy of the detection.

3. Pedestrian Detection Based on Scene Motion Pattern

3.1. Scene Constraint and Motion Pattern. In a common scene, performances of people moving are always influenced by the layout and moving trajectory of other scene elements. These influences, called floor field defined in [9], come from several aspects including social constraints (e.g., traffic lights), other moving objects (e.g., pedestrians and cars), and environmental layout. Even though it is difficult to obtain

the high-level information of the floor field by learning the scene itself, it can be obtained by treating each pedestrian as an independent sample and statistically analyzing its motion since the influences of the floor field apply to people's movement all the time. In general density scenes, without considering the interaction of individuals (e.g., road congestion and mutual jostling among the crowd), the influences can be seen as an independent and identically distributed function. To estimate the function, a method based on object observation and motion statistic is developed. As long as the number of pedestrians is big enough, this Monte-Carlo-statistic-like learning method acquires an approximate distribution of the real floor field.

The construction of the OSMP model begins with computing the sparse optical flows which are defined as the apparent motion vectors of the brightness patterns at the KLT corner pixels. The whole scene is divided into squared cells, and each cell has the same proper size, not to be oversized to ignore the details or undersized to waste computing resources. For a given video, the KLT corners of sparse optical flow as the interest points of each frame are extracted. Thus, the floor field of the cells of every optical flow passing by is modeled as a 2D histogram $h_{i,j}(v, d)$, where v is the magnitude of optical flows and d is the direction over the cells. According to the magnitude, the feature vectors of the optical flows are classified as two histogram components: high-speed one and low-speed one. Usually, people moving by path are having a higher speed than hanging around since the movement with purpose is faster. Thus, the high-speed components of the optical flow describe the main directions of all the paths over the cell. And the low-speed components describe the complexity of the pedestrian moving directions because the waiting or the wandering pedestrians (e.g., the subconscious moving of people at a bus stop) always present a characteristic of low-speed movements. The probability of one region being a holding area can be estimated by calculating the reciprocal of the variance of the low-speed optical flow components. The bigger the variance, the smaller the probability. Thus, the motion pattern which describes the probability can be estimated and represented.

For better scale invariance of different distances from camera, a threshold function $T(j)$ which decreases with increasing row-coordinate of different cells is used to correct the optical flows since the cameras usually used are all set as downward ones. By the geometric projection principle of downward camera, as in Figure 1, we have

$$\tan^{-1}\frac{x}{h} - \tan^{-1}\frac{x_0}{h} = \tan^{-1}\frac{y}{f}, \tag{1}$$

where h is the height of the position of the camera and f is the focal length. By differentiating both sides of the above equation, we have the relation:

$$\frac{h}{h^2 + x^2}dx = \frac{f}{f^2 + y^2}dy. \tag{2}$$

Thus, we obtain

$$\frac{dy}{dx}dx = \frac{1}{hf}\frac{(fh - yx_0)^2}{h^2 + x_0^2}. \tag{3}$$

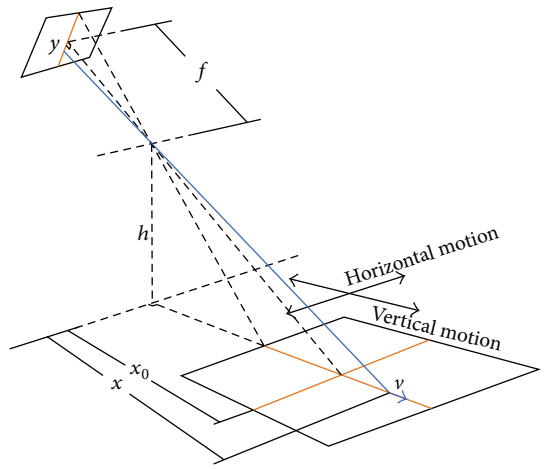

Figure 1: The geometric projection principle of a downward camera.

And so,

$$\frac{y'}{y'|_{x=x_0}} = \left(1 - \frac{x_0}{fh}y\right)^2 = \frac{v}{v_0}. \tag{4}$$

Let $\alpha = \tan^{-1}x_0h^{-1}$ be the depression angle of camera; then,

$$\frac{v}{v_0} = \left(1 - f^{-1}y\tan\alpha\right)^2. \tag{5}$$

Since the quadric expression in (5) reaches the minimum at $y = f/\tan\alpha$, that means, the object is at infinity, we employ a linear threshold function $T_\alpha(y)$ which related to the specific surveillance condition to approximate (5), correcting the magnitudes of optical flows at different row-coordinates. Based on (5), the linear approximation error will decrease with the focal length of camera increasing. In other words, the error will be small when the surveillance system uses a middle or a long focal lens. Before using the threshold function, it is needed to filter the noises by separately setting up an upper and a lower bound of the optical flow magnitude. Thus, every corrected vector is voted into the high-speed or the low-speed bin of 2D histogram of each cell according to its direction angle, and the voting weight is the magnitudes of the vector. In consideration of balancing the calculation, we space four orientation bins (d_1, d_2, d_3, and d_4 in Figure 2) evenly over $[0, \pi)$. The values of d_1 and d_4 reflect the horizontal optical flows in every cell, and the d_2 and d_3 reflect the vertical ones. In addition, the linear interpolation, which is proved to effectively reduce the error generated by lots of optical flows in the vicinity of cell boundaries, is used in the voting procedure.

Although the movement of individual changes all the time, the changing of crowd motion pattern which depends on the observation scene, observing time, and possible happenstances is relatively slow. Hence, the histogram calculation is executed before the detection and is updated without excessive frequency to decrease the computation of the whole procedure.

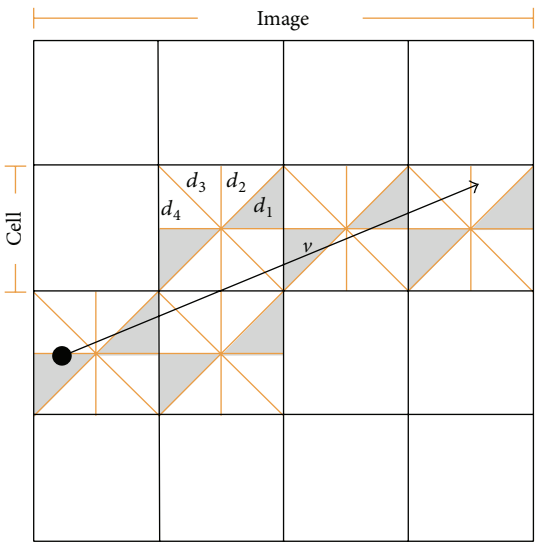

FIGURE 2: The direction vote of motion pattern histogram for cells. Four orientation bins are spaced evenly over $[0, \pi]$.

3.2. Classifier Training and Detection. Corresponding to the OSMP, the orthogonal cascades classifier is composed of two directional subclassifiers trained, respectively. The subclassifiers employed Gentle AdaBoost (GAB) technique due to the fact that it not only has better performance and faster convergence ability than those of other boosting algorithms, for example, Descreat AdaBoost and Real AdaBoost proved and practiced in [2, 13], but also can choose the most effective features from a huge number of upright and 45° rotated over-complete Haar-like features. It is shown that the Pareto principle in economics works here as well; that is, only a small set of features as the important ones are the determining factors in the classification. The weak learners of the subclassifiers are designed as one for each feature called decision stamps (similar to single-level decision trees). During the whole classification process, the GAB is in charge of selecting the good ones of the weak learners which are corresponding to the important features.

The GAB algorithm is based on a set of training samples $(x_1, y_1), \ldots, (x_n, y_n)$, where $y_i \in \{-1, +1\}$ is the class label and x_i is the sample feature. An iterative procedure is provided to fit an additive regression model $F(x) = \sum^{t=1} Tf_t(x)$, where $f(x) = p_D(y = 1 \mid x) - p_D(y = -1 \mid x)$ is the difference of weighted probability of two classes based on one weak learner. By taking adaptive Newton steps, the Gentle AdaBoost can minimize the weighted conditional expectation of squared error $E[e^{-yF(x)}(y - f(x))^2 \mid x]/E[e^{-yF(x)} \mid x]$ which is a Lagrange quadratic approximation of $J[F(x) + f(x)] = E[(e^{-yF(x)e^{-yf(x)}})^2 \mid x]/E[e^{-yF(x)} \mid x]$ in each iteration. The details of GAB are shown in Algorithm 1.

In practice, for choosing one optimal weak learner in each loop, a look-up-table (LUT) method proposed in [18, 19] is used to get the weighted probability for fitting the regression function in the loop of Algorithm 1. The detail of LUT method is shown in Algorithm 2.

To reject the negative images efficiently, the two subclassifiers are trained, respectively, with horizontal and vertical motion training sets corresponding to the two orthogonal directions of region motion pattern. Because the cascade algorithm has the ability of giving a negative result with few classified stages as shown in Figure 3, our approach can run very efficiently.

The final result of the classification is derived from the outputs of the subclassifiers and the regional OSMP. A group of pattern-classify mapping strategies are applied to determine the different classification procedures for different patterns. Since the subclassifiers include both a horizontal one F_h and a vertical one F_v, there are three types of classification procedures: using each classifier and using both. Defining the 2D histogram h_C of the cell C at the center of detection window, the variance of the low-speed optical flow components is

$$D_{\text{low}}(h_C) = \text{Var}[h_C(\text{low}, d_i)]. \tag{6}$$

Define the function

$$S(x, y) = \begin{cases} 1, & \text{if } x = 1 \text{ or } y = 1 \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

and we listed the mapping strategies in Table 1.

Based on the motion pattern of the scene and the mapping strategies, the orthogonal pedestrian detection chain is shown in Figure 4.

4. Experiment Designs and Results

Two steps of experiments have been designed specifically to test our approach from different perspectives. In the first step, a series of static detection experiments are designed to test the difference between orthogonal cascades classifier based on OSMP model and classic GAB cascade classifier. To compare the overall dynamic detection property of the two algorithms, several videos labeled manually are tested with the orthogonal cascades classifier trained in static experiments in the second step.

4.1. Static Experiments. To test our algorithm proposed above, two static pedestrian classification systems have been built: one with a classic GAB cascade classifier and the other with the orthogonal cascades classifier based on OSMP model. Both of the detectors are sharing one structure with 38 cascade layers but are trained with different sample sets. We set the first layer of cascade with 5 features and 5 features are added to each layer after.

Positive images are collected from two influential person datasets with two principles: pedestrian centered and no significant occlusion. Specifically, our positive sample sets are constituted by 2276 mirrored (defined as both the original and the horizontally flipped) positive training images (including 1472 vertical motion samples and 804 horizontal motion ones) selected manually from INRIA person dataset [5] and 1780 mirrored pedestrian images (including 1128 vertical motion samples and 652 horizontal motion ones)

Given: $(x_1, y_1), \ldots, (x_n, y_n), x_i \in X, y_i \in \{-1, +1\}$;
Initialize: $D_i = 1/n, i = 1, 2, \ldots, n, F(x) = 0$;
  for $t = 1, 2, \ldots, T$ **do**
    Fit the regression function $f_t(x)$ by weighted least-squares of y_i to x_i with weights D_i;
    Set $F(x) \leftarrow F(x) + f_t(x)$;
    Set $D_i \leftarrow D_i \exp[-y_i f_t(x_i)], i = 1, 2, \ldots, n$, and normalize D_i so that $\sum D_i = 1$;
  end for
Output: the classifier $\operatorname{sign}[F(x)] = \operatorname{sign}[\sum f_t(x) + b]$ where b is threshold.

Algorithm 1: Gentle AdaBoost.

for each weak learner $h \in H$ **do**
  Partition feature space into J disjoint blocks $x_1, \ldots, x_J$ that at least one sample in each block;
  Under the distribution D_t calculate $\widetilde{W}_l^j = p\left(y_i = l \mid x_i \in X_j\right)$ where $l = \pm 1$;
  Calculate $Z_h = \sum_i \{D_i \exp[-(1/2) y_i \sum_j (\widetilde{W}_{+1}^j - \widetilde{W}_{-1}^j) B_j(x_i)]\}$ where $B_j(x) = \begin{cases} 1, & x \in X_j; \\ 0, & x \notin X_j; \end{cases}$
end for
Output: $f_t(x) \stackrel{\text{def}}{=} \arg_h \min Z_h$.

Algorithm 2: Look-up-table.

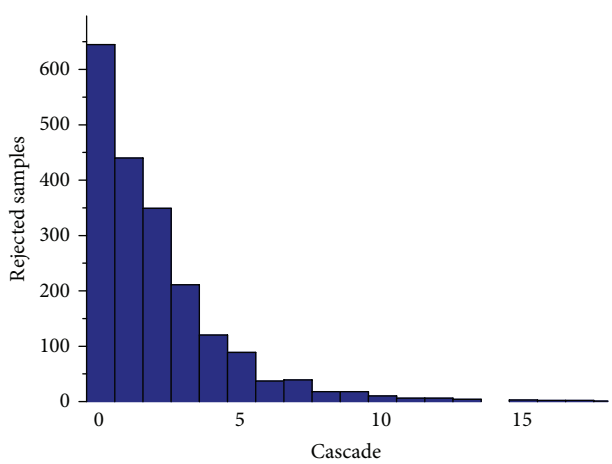

Figure 3: The statistical histogram of rejected samples in each stage of cascade from static experiments.

selected from Caltech [20] pedestrian dataset. The key point to differentiate the two kinds of samples is that both hands and feet can be seen from vertical images if there exists no interindividual occlusion from other objects. The nonperson samples are collected from thousands of pictures of the Internet with different sizes and resolutions. And 60,000 negative samples are extracted from these pictures by using a scan-window. All the input images of the sample sets are scaled and cut to a unified size with 60 pixels in width and 120 pixels in height. 1,030,530 upright and rotated Haar-like features are extracted from one input image, and only 5,880 features are selected with GAB to participate in classification to avoid overfitting and improve efficiency.

A 5-fold cross validation technique is taken for our static experiments. The positive sample sets are split into the training and the test sets. The vertical and horizontal subclassifiers are trained with 2080 vertical and 1164 horizontal positive samples, respectively, and the remaining 812 images are set as the positive test data. Meanwhile, all the 3244 positive training samples are used to train the GAB cascade classifier. The classifiers are all trained with 10,000 negative samples including the false positive samples from the prior layer and the rest of the samples selected randomly.

We compare the GAB cascade average results with the results of logic or operation on two orthogonal cascade subclassifiers. As shown in Figure 5(a), by the performance comparisons of the receiver operating characteristic (ROC) curves, it is indicated that the logic or operation results are generally better than those of the GAB cascade classifier. The average detection rate of 5-fold cross test results achieves 95.47% with a 92.74% positive predictive value. Because of the test sets including both vertical and horizontal motion of pedestrians, the performances of two subclassifiers are not optimal compared with the GAB cascade classifier. The logic or operation obviously reduces the number of false negative instances and thus the orthogonal cascades classifier has the best performance in our test.

We also try increasing the subclassifiers of the detector by training them in left, right, front, and back directions, respectively. Using the same training procedure above, the training sets are separated manually to four directions. The same test sets are employed and the results are shown in Figure 5(b). Compared with orthogonal cascades classifier, the four-direction version is shown without obvious advantage. One possible reason is that there is only slight difference between the left and the right direction versions of a pedestrian, as well as that between the front and the back ones. Besides,

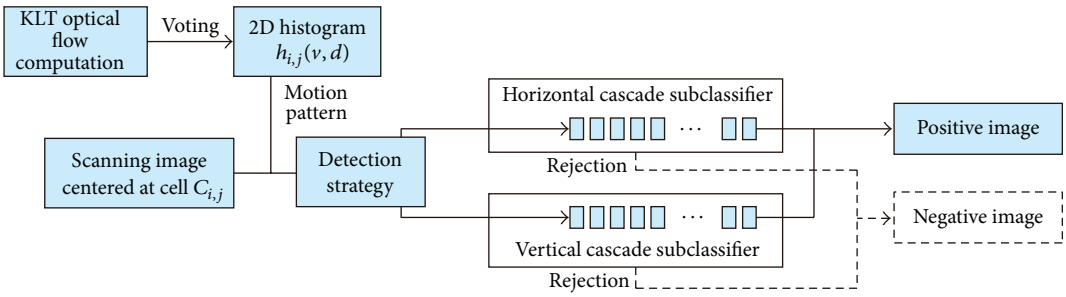

FIGURE 4: An overview of our pedestrian detection chain.

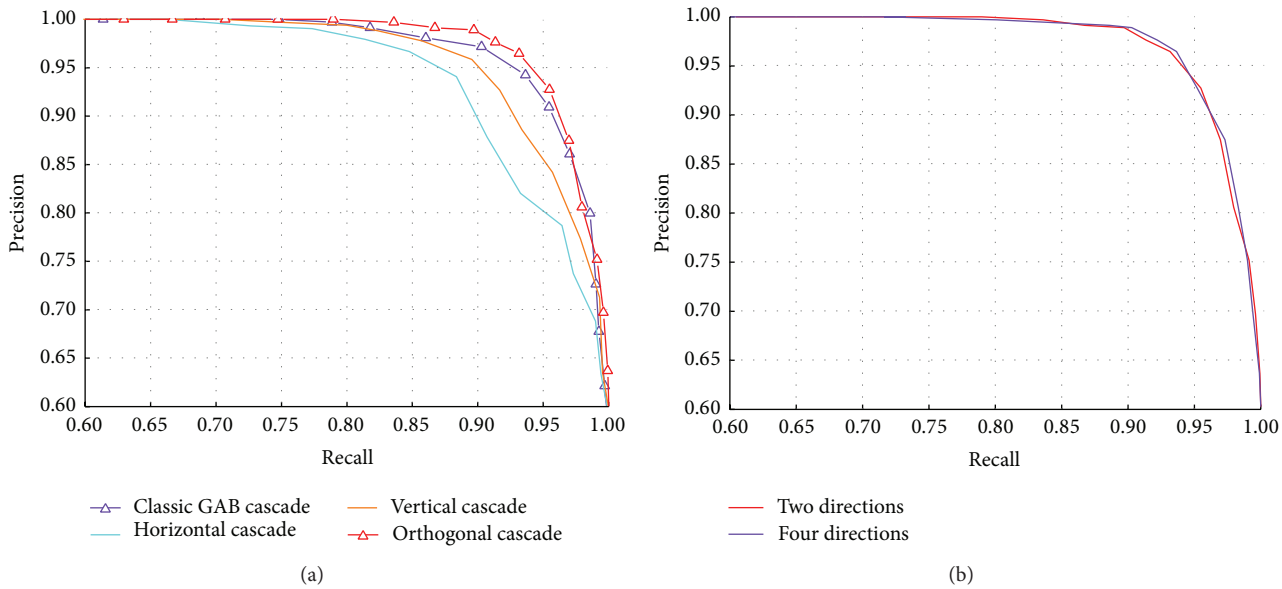

FIGURE 5: ROC curves of static experiments. (a) is the average precision-recall curves of orthogonal cascades classifier (including horizontal subclassifier, vertical subclassifier, and the result of logic or operation on both) and the GAB cascade classifier. (b) is the comparison of orthogonal cascades classifiers with different orientation bins.

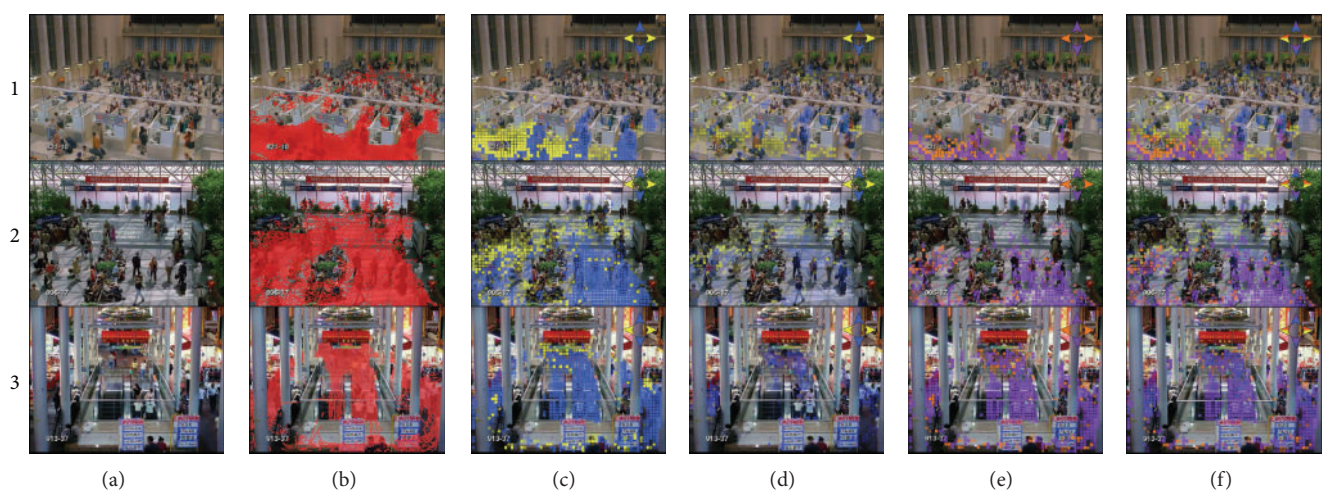

FIGURE 6: OSMP Images. According to different speeds and directions, motion pattern images of three scenes were arranged in columns. The images in columns (a) and (b) show the scenes and crowd flow images, respectively. Column (c) shows the vertical and the horizontal motion patterns. Columns (d) and (e), respectively, show the low-speed and high-speed velocity components of the motion patterns. The final OSMP images are shown in column (f). The opacity of the cells represents the energy of the histograms.

Table 1: The pattern-classify mapping strategies of orthogonal cascade classifier.

Pattern-classify mapping strategy	
Condition	Output
(1) If $D_{\text{low}}^{-1}(h_C) > v$, where v is the threshold of holding area,	$S\{\text{sign}[F_h(x)], \text{sign}[F_v(x)]\}$.
(2) Else if $\gamma \in (u, u^{-1})$, where $\gamma \stackrel{\text{def}}{=} (h_C(\text{high}, d_1) + h_C(\text{high}, d_4))/(h_C(\text{high}, d_2) + h_C(\text{high}, d_3)) > 0$ and $u \in (0, 1)$ is the threshold to distinguish a cross as the intersection of two orthogonal paths,	$S\{\text{sign}[F_h(x)], \text{sign}[F_v(x)]\}$.
(3) Else if $\gamma \in (0, u)$,	$S\{\text{sign}[F_v(x)], 0\}$.
(4) Else $\gamma \in (0, u)$,	$S\{\text{sign}[F_h(x)], 0\}$.

Table 2: The comparison of video experiment recalls classifier.

Video	Recall (8 pixels)	Recall (16 pixels)	Recall (24 pixels)
(1)	82.23%	79.8%	76.97%
(2)	82.52%	84.74%	84.19%
(3)	80.77%	81.51%	79.98%

four-direction classifier experiment takes twice as much time as two-direction one does. Hence, according to our tests two orthogonal cascades with logic or operation are ideal for pedestrian detection.

4.2. Video Experiments. The trained classifiers described above are then loaded into the second step of experiments with several public surveillance videos as inputs from detection datasets of University of Central Florida [21] and CAVIAR [22] project. To test the relationships of different image scales and different cell sizes, the resolutions of these selected videos range from 384×288 to 720×576. There are about 400 frames in the shortest video and about 3000 frames in the longest one. The motion states of different objects with different speeds and directions are extracted from the first 200 frames of each sequence. The motion patterns are learned by voting the states to the corresponding cells. The cells are tested in three different sizes: 8 pixels, 16 pixels, and 24 pixels. The detection rate of each video is counted, respectively.

The motion patterns of the scenes of a railway station hall, airport security checkpoints, and a shopping mall entrance by using 8 pixels as cell size are shown in Figure 6. According to different speeds and directions, the motion pattern images are arranged in columns. It is seen from column (c) that there are always some entrances, exits, and several primary paths in each scene. There are also holding areas with clutter crowd flows on account of the existence of some functional scene elements like service desk, shop, information board, waiting benches, and so forth. As can be seen in column (d), these areas are always overlaid by the low-speed cells, while the high-speed cells in column (e) overlay the primary paths. Using the strategies described in Section 3.2, the set of all cells is utilized as motion pattern to obtain the results. It is generated from our experiments that the most accurate detection rate our approach achieved is 86.77% with fewer than 2% false positive per frame while the rate of traditional GAB without motion pattern is 76.44%. The experiments confirmed that our approach can offer a relatively satisfied detection result.

Setting up different cell sizes, the comparable results are filled into Table 2. As can be seen, the best rates are achieved by scene (1) with the 8 pixels of cell size and by scene (2) and scene (3) with 16 pixels of cell size. Based on the analysis, the main reason is that the average size of human body images in scene (1) is smaller than that in scenes (2) and (3). For this reason, in order to obtain a better rate, using the same size of cells as the width of human body is recommended. Furthermore, observing angle and occlusion could have affected the detection results. It can be seen in the recall comparison of three scenes that the second scene achieves the best rate.

5. Conclusion

A novel approach to detect pedestrian in video using crowd motion pattern is introduced in the paper. Compared with traditional methods, this approach utilizes the motion pattern of scene as the prior information to allocate the detection strategies. By modeling the crowd motion pattern, the primary paths of crowd and holding areas in scene can be obtained and represented by a probability grid model called OSMP. Meanwhile, an orthogonal cascades classifier is introduced to replace traditional ways. Different pattern-classify mapping strategies of two subclassifiers are employed for different scene areas. The static experiments show that without raising false positive rate the orthogonal cascades classifier based on OSMP model achieves detection rate of 86.8% which is higher 8.3% than that of the GAB cascade classifier. In the meantime, the video experiments confirmed the improvement of video pedestrian detection algorithm. Future work will focus on two aspects: one is exploring richer motion pattern information, and the other is improving the efficiency of classifier, especially in feature extraction.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (no. 61173091).

References

[1] J. Shi and C. Tomasi, "Good features to track, computer vision and pattern recognition," in *Proceedings of the IEEE Computer Society Conference*, vol. 1, pp. 593–600, 1994.

[2] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," *Pattern Recognition*, vol. 2781, pp. 297–304, 2003.

[3] P. Dollar, Z. Tu, P. Perona et al., "Integral channel features," in *Proceedings of the British Machine Vision Conference*, vol. 2, p. 5, BMVC, 2009.

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. I511–I518, December 2001.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 886–893, San Diego, Calif, USA, 2005.

[6] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proceedings of the IEEE 12th International Conference on Computer vision*, pp. 32–39, IEEE, Kyoto, Japan, September 2009.

[7] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: survey and experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[8] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 193–199, San Juan, Puerto Rico, USA, June 1997.

[9] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *Computer Vision—ECCV 2008*, vol. 5303 of *Lecture Notes in Computer Science*, pp. 1–14, Springer, Berlin, Germany, 2008.

[10] M. Rodriguez, S. Ali, and T. Kanade, "Tracking in unstructured crowded scenes," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 1389–1396, October 2009.

[11] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1003–1016, 2012.

[12] M. Hu, S. Ali, and M. Shah, "Learning motion patterns in crowded scenes using motion flow field," in *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, pp. 1–5, December 2008.

[13] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[14] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the 13th International Conference on Machine Learning*, vol. 1, pp. 148–156, 1996.

[15] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, pp. 297–336, 1999.

[16] Y. Wang and C. Dang, "An evolutionary algorithm for global optimization based on level-set evolution and latin squares," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 5, pp. 579–595, 2007.

[17] Y. Wang, Y. Jiao, and H. Li, "An evolutionary algorithm for solving nonlinear bilevel programming based on a new constraint-handling scheme," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 35, no. 2, pp. 221–232, 2005.

[18] B. Wu, H. Ai, and C. Huang, "LUT-based Adaboost for gender classification," in *Audio- and Video-Based Biometric Person Authentication*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 104–110, Springer, Berlin, Germany, 2003.

[19] B. Wu, H. Ai, C. Huang, and S. Lao, "Fast rotation invariant multi-view face detection based on real adaboost," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '04)*, pp. 79–84, May 2004.

[20] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: a benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 304–311, IEEE, 2009.

[21] S. Ali and M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–6, June 2007.

[22] CAVIAR dataset, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.