Pattern Recognition Letters 30 (2009) 148–156

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Contour-motion feature (CMF): A space–time approach for robust pedestrian detection

Yazhou Liu [a,*], Xilin Chen [b], Hongxun Yao [a], Xinyi Cui [a], Chaoran Liu [b], Wen Gao [b]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, PR China
[b] Institute of Computing Technology, Chinese Academy of Science, Beijing 100080, PR China

ARTICLE INFO

Article history:
Available online 26 March 2008

Keywords:
Contour-motion feature
Space–time analysis
Pedestrian detection
Human activity analysis
Event analysis
3D Haar filter

ABSTRACT

This paper presents a contour-motion feature for robust pedestrian detection. The space–time contours are used as the low level representation of the pedestrian. Then we apply 3D distance transform to extend the 1-dimensional contour into 3-dimensional space. By this way, the relations between the local contours can be maintained implicitly. Further, by encapsulating the static and dynamic information by 3D Haar-like filters, we can generate the middle level pedestrian representation: contour-motion features. Then we use boosting method to select the most representative features. Our experiments demonstrate that the proposed approach can outperform Viola's well-known pedestrian detector in both detection accuracy and generalization ability. In addition, even though our approach is presented in pedestrian detection scenario, it has been extended to human activity recognition application and remarkable performance has been achieved.

1. Introduction

The research of pedestrian detection has received more and more attention in recent years because of increasing demands in practical applications, such as smart surveillance, on-board driving assistance system and content based image/video retrieval. Remarkable progress has been made to improve the speed and robustness of the detection procedure (Dalal and Triggs, 2005, 2003, 2006). However, the fundamental problems for reliable pedestrian detection are still far from being completely solved. The human body is highly articulated so its shape may vary radically; the clothes of a person may have varying colors and texture, so it lacks in appearance consistency; viewing directions and lighting may also change the image of a human body. All these make pedestrian detection an especially difficult task in object recognition. Different aspects are addressed in different applications. For image retrieval and objects recognition, researchers are focusing on solving the critical issues which are caused by highly articulated human body and variations in clothing. Typical problems include pose, occlusion and lighting (Mikolajczyk et al., 2004, 2005). For surveillance and driving assistance system, more attention has been focused on improving the speed and the accuracy of detection (Gavrila, 2000, 2004, 2003). In order to improve the detection speed, many researchers follow the "segmentation to recognition" routine. The segmentation can be accomplished either by background subtraction (Zhao and Nevatia, 2003) or by some special image acquisition device such as stereo (Liu and Fujimura, 2004) or infrared cameras (Xu et al., 2005). These approaches may provide satisfiable results in some specific applications. However, their performance relies on the segmentation results of the foreground blobs. Therefore, the segmentation-free pedestrian approaches may have wider application.

Our approach is based on two basic assumptions. The first basic assumption is both static and dynamic patterns are important for identifying a moving human. So modeling pedestrian in space–time domain is a possible solution for fast and robust pedestrian detection. Many research results have demonstrated that pedestrian can be identified by either dynamic information (Cutler and Davis, 2000) or static appearance (Dalal and Triggs, 2005,, 2004). However, these approaches are either not robust enough or not efficient enough for real time implementation. The work of Viola et al. (2003) and Dalal et al. (2006) indicate that the combination of static and dynamic information can improve the detection accuracy. Recent promising works on video analysis have also demonstrated effectiveness of the space–time analysis, such as video based alignment (Ukrainitz and Irani, 2006), behavior correlation (Shechtman and Irani, 2005) and in painting (Wexler et al., 2004). The second basic assumption is the gradient-based descriptors are more robust for representing the appearance of the highly articulated pedestrian. This assumption is mainly inspired by the recent success of the gradient-based local descriptors, such as SIFT (Lowe, 1999), HOG (Dalal and Triggs, 2005), edgelet (Wu and Nevatia, 2005) and edgel (Ferrari et al., 2006). Thus, using the gradient

* Corresponding author. Fax: +86 451 86416485.
E-mail address: yzliu@vilab.hit.edu.cn (Y. Liu).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.03.007

information in space–time domain is another starting point of our research.

Based on the above two assumptions, we developed our space–time pedestrian detection approach, which represents the pedestrians in 3D distance transform volume and extracts its contour-motion features by 3D Haar-like filters. The advantages of our approach include: (1) the combination of static contour feature and long term (five frames) motion feature can provide more discriminative information and better generalization ability. (2) Our 3D Haar-like filter can handle the appearance and motion information in a consistent and efficient framework. (3) Even though our approach is presented in pedestrian detection scenario, the proposed space–time analysis technique has been easily extended to other video analysis applications, such as human activity analysis.

The remaining parts of this paper are organized as follows: Section 2 gives a brief review of the state of the art pedestrian detection approaches; followed by Section 3 which describes our methods in detail. Subsequently, Section 4 presents the evaluation results of our method against other baseline method and lastly, we shall state our conclusion and future focus in Section 5.

2. Previous works

For pedestrian detection in static images, most of the detection approaches fall into two categories according to different human body representation methods. Some researchers model the pedestrian as an integrated whole by its appearance, its shape or both. For these methods, the detection result is determined by a single whole-body detector. Earlier work includes the dense Haar+SVM detector by Papageorgiou and Poggio (2000) and contour based chamfer matching detector by Gavrila (1998, 2000). Later, with the rapid development of local descriptors, more researchers tried to use local gradient histogram to represent the appearance of human body. Dalal and Triggs (2005) used histograms of oriented gradients (HOG) for human detection and Zhu et al. (2006) extended this work by combining HOG with a cascade real time detector. Some researchers used gradient-based local descriptor to represent the appearance and contour as the global shape constrain, such as two-layer field model by Wu and Yu (2006) and the implicit shape model (ISM) by Leibe et al. (2005). Recently, Seemann et al. (2007) generalized Leibe et al.'s work to make it capable for describing both the general object class and specific object instance.

Representing human body by its parts is another popular modeling method. These methods divide human body into different parts and several part-detectors are learned separately. The final result can be inferred by fusing the outputs of part-detectors (Mohan et al., 2001, 2004). Ramanan and Forsyth (2003) searched body segments by matching the puppet models. Mikolajczyk et al. (2004) learned seven part-detectors using position-orientation histogram features. Sabzmeydani and Mori (2007) developed a two-layer Adaboost to select Shapelet features for different parts of human body. Wu and Nevatia (2005) divided the human body into four parts and use edgelet features representing the contours. Bayesian inference was used for combining the results of part-detectors. Tuzel et al. (2007) project the covariance matrices of image patches onto a Riemannian manifold and use LogitBoost to build a classifier. These methods are normally robust to partial occlusions. Recently, Munder and Gavrila (2006) presented an experimental study on pedestrian classification.

Representing the pedestrian by the combination of its appearance and motion pattern is another promising way. Viola et al. (2003) combined the appearance and motion information into his famous cascade face detection framework (Viola and Jones, 2001). Dalal et al. (2006) extended HOG features into both spatial gradient field and optical flow field. These work demonstrated that space–time analysis technique can yield better performance in comparison to the solely appearance based detectors. An intuitive but not strict classification can be seen in Fig. 1.

3. Space–time representation of pedestrian

In order to develop an effective way to represent the pedestrian, we start our research from the two assumptions mentioned above: the combination of appearance and motion information can increase the discriminative power for human detection and the gradient-based descriptors are more robust than intensity based ones for representing the appearance of pedestrians. The first assumption has been verified by Dalal et al. in their recent work (Dalal et al., 2006) and we verify the second one in our experiments. Based on the first assumption, we represent the pedestrian by a space–time volume (STV) and extract its static and dynamic features by 3D Haar-like filters. The space–time volume representation can be seen in Fig. 2, which contains several adjacent frames of the moving pedestrian. The size of STV is 30 × 15 × 5 in our approach. In addressing the second assumption, we use the 3D

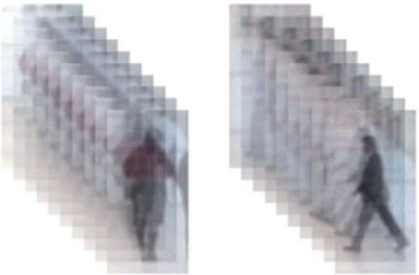


Fig. 2. The space–time volume representation of pedestrians.

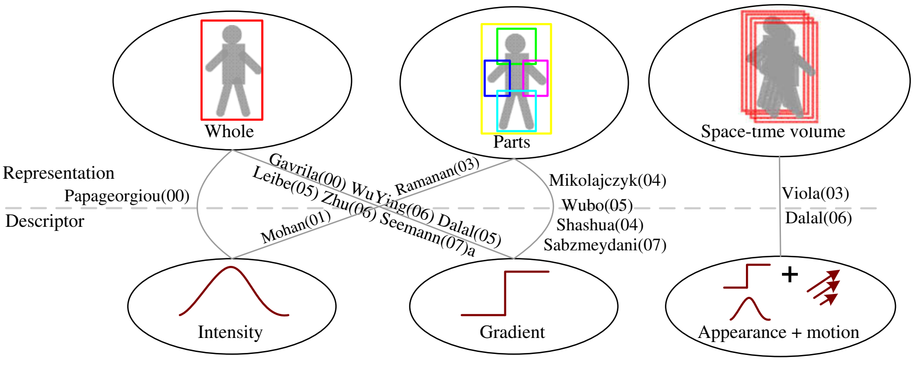


Fig. 1. Classification of the state of the art static pedestrian detectors.

distance transform to build a contour representation in space–time domain.

An overview of our feature extraction and object detection procedure can be seen in Fig. 3. Firstly, we apply the edge detection and distance transform on the space–time volume to generate the distance transform volume (DTV). Then 3D Haar-like filters are used to extract the static and the dynamic features. The candidate feature number is about 50,000, which are uniformly subsampled from the much larger set of all filters that fit in a 30 × 15 × 5 voxel volume. Then, we uses Realboost (Schapire and Singer, 1999) to select a subset of features and construct the cascade classifier.

In the following parts of this section, we use $V(p)$ to denote space–time point $p(x,y,t)$'s intensity and V_x, V_y, V_t to denote the gradient along each axis direction. The gradient and the distance transform volume are denoted by $G(p)$ and $D(p)$, respectively. We refer to the video sequence without pedestrian as the background video or just the background. More specifically, the backgrounds that captured by the static cameras are named as static backgrounds and the ones that captured by the moving cameras are named as dynamic backgrounds.

3.1. Distance transform volume

Contour is an effective way to represent the shape of the nonrigid human body, since the gradient is more robust than the intensity for varying clothes and illumination. Since we model the pedestrian in a 3D space–time volume, we need to find an effective way to formulate the contour in this 3-dimensional space. In order to address both the static and dynamic patterns, the definition of gradient contains two terms, the spatial gradient and the temporal gradient, which can be seen in Eq. (1). The first term on the right side of the equation is the spatial gradient that can capture the appearance information and the second term is the temporal gradient that can capture the variation along the temporal axis. By the parameter α we can allocate different emphasis between the spatial and temporal gradients.

$$G(p) = \alpha\sqrt{(V_x)^2 + (V_y)^2} + (1-\alpha)|V_t| \tag{1}$$

We can further get the contour/edge representation by thresholding the gradient $G(p)$ in Eq. (1). However, this representation is not very effective. There exist two critical problems. Firstly, the contour is a 1-dimensional signal (like the space curve) and theoretically it should be convolved with the 1-dimensional filters for feature extraction. But the amount of 1-dimensional filters in 3-dimensional space is too huge, making the computation inapplicable. Secondly, the non-contour regions have not been fully used, which form the majority of the space–time volume. In addressing these two problems, a sensible way is to extend the 1-dimensional contour into 3-dimensional space by filling in the non-contour regions with the values that can reflect the information of neighboring contours. An intuitive explanation is that by contours we can get the skeletons and we now need to attach the muscles to these skeletons to make the pedestrian become chubby and recognizable.

To fill in the empty regions, a sensible way is to assign every non-contour point a value reflecting its relative position within the shape. One popular example is the distance transform, which assigns to every point a value reflecting its minimal distance to the boundary contour and has been widely used for binary template matching. More sophisticated representation methods can be found in (Blank et al., 2005). Here we just use the definition of distance transform and extend this concept into 3D space–time domain naturely. The distance transform volume can be defined as:

$$D(p) = \min(\mathrm{disy}(p,p^*)), \;\; p^* \in \{p' | G(p') > \theta\} \tag{2}$$

where the threshold θ is used for edge detection and $\mathrm{dis}(\cdot,\cdot)$ is a distance metric which can be Euclidean or block distance. Thus far, we have turned the binary contour volume into a continuous distance transform volume in which each voxel's value represents the minimum distance between the current position and the contour. By this way, we can expend the 1-dimensional contour into 3-dimensional space which makes the fast integral image based feature extraction possible. In addition, the relations between the local contours can be also maintained implicitly.

3.2. 3D Haar-like filter for contour-motion feature extraction

The success of Viola and Jones' algorithm (Viola et al., 2003) lies in that it uses the motion information between two consecutive images. But when person is moving slowly, the motion pattern between the two images is not obvious, thus the features from two-frame difference are not so informative. In order to capture the long-term motion patterns among multiple frames and record the person's appearance feature at the same time, we extract 3D Haar-like features from a series of consecutive frames instead of two frames. Similar volume filters has been used by Ke et al. (2005) for visual event detection, but only filters S1, S2, D4 in Fig. 4a, b and g and a sum filter are defined in their work. Our experiments demonstrate that features D2 and D3 in Fig. 4e and f

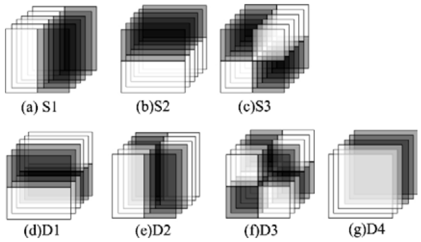


Fig. 4. Seven types of 1-order 3D Haar-like features.

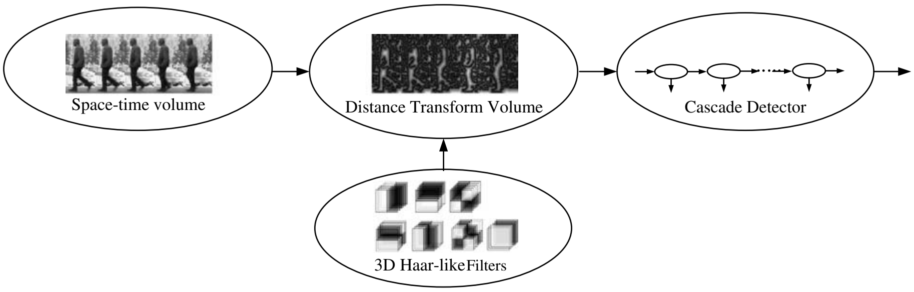


Fig. 3. The overview of our approach.

are more important for representing the motion pattern of the pedestrians.

3D Haar-like features are extracted in a 3D distance transform volume. They can be seen as the convolution results of the 3D Haar-like filters and space–times volumes. The feature value is just the difference value between the intensity sums of the dark and bright regions. We develop seven types of 1-order 3D Haar-like filters to represent both the static and dynamic information. More specifically, we use three filters to represent appearance features which referred to as S1, S2 and S3, and four filters to represent motion features which referred to as D1, D2, D3 and D4. See Fig. 4 for details. The cubic filters in Fig. 4a–c are the static features, which have similar meanings as the 2D Haar-like filters used in (Viola et al., 2003), the only difference is that these filters are calculated along several consecutive frames. We use these filters to describe the pedestrian's appearance.

The filters D1–D4 in Fig. 4d–g are the dynamic features. They are used to capture the different kinds of motion information in the space–time domain.

Translation, is modeled by feature D1 and D2 in Fig. 4d and e. More specifically, D1 is used to model the vertical movement of a horizontal edge and D2 is used to model the horizontal movement of a vertical edge. Take feature D2 for example, it computes the difference between vertical diagonal pairs of cubic in temporal dimension. So if a vertical edge moving horizontally, the response of this filter should be large. This motion pattern is very common for a moving human, such as waving legs and arms. Especially, when we observe a pedestrian from a distance, this kind of periodical motion is very distinctive and helpful to identify the pedestrian. Intuitively, we expect this filter to be a powerful one. Our experiments verify this point. Corresponding to D2, we use feature D1 to represent vertical motions. This kind of motion may occur when the pedestrian's moving plain is not parallel to the camera's optical axis, such as most of the surveillance videos where the cameras look down at the pedestrians.

Rotation, is modeled by feature D3 in Fig. 4f. By D3, we intend to model rotation of an edge. If an edge rotate by 90°, the response of this filter should be large. This kind of pattern may happen when the pedestrians swing their arms and legs.

Appearance and disappearance, is modeled by feature D4 Fig. 4g. We intend to use this feature to modeled the appearance and disappearance of an edge. When the camera is static, this feature can be used for removing the edges of the background, for which the response of D4 should be small.

Integral image is a fast method to compute 2D Haar-like features. This led to a real-time face detection system (Viola and Jones, 2001) and human detection system (Viola et al., 2003). To compute 3D Haar-like feature value efficiently, we also use the idea of the integral image. The only difference is that we compute integral image in 3-dimensional space, and we refer to it as integral volume. The value of integral volume at location (x,y,t) is the sum of the intensity of the voxels which location indices are less than the current location. Specifically:

$$\mathrm{IV}(x,y,t) = \sum_{x'<x,y'<y,t'<t} D(x',y',t') \qquad (3)$$

where $\mathrm{IV}(x,y,t)$ is the integral volume and $D(x,y,t)$ is the distance transform volume. Using the integral volume, any cubic sum can be computed in seven plus/minus operations. If we denote the vertexes of the volume as in Fig. 5, the sum of the volume can be calculated as $\mathrm{sum}(V) = \mathrm{IV}(H) - \mathrm{IV}(D) - \mathrm{IV}(F) - \mathrm{IV}(G) + \mathrm{IV}(B) + \mathrm{IV}(C) + \mathrm{IV}(E) - \mathrm{IV}(A)$, where $\mathrm{sum}(V)$ denote the sum of the voxels' intensity in the volume V. So the cubic filters in Fig. 4a, b and g need 14 operations, and the ones in Fig. 4c–e need 28 operations, and filter in Fig. 4f needs 56 operations.

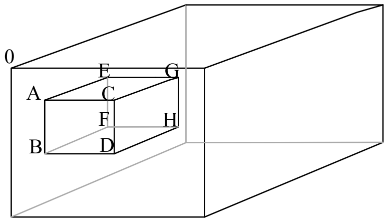


Fig. 5. Calculation of the space–time volume.

The computational complexity of calculating the 3D distance transform volume and integral volume is $O(m \cdot n)$, where m is the size of each frame and n is the number of frames. For each newly observed image, we just make it the first frame of the space–time volume and discard the oldest frame, and make the judgment on the new image by the detection results of current space–time volume. By this way, both 3D distance transform and integral volume can be calculated incrementally, the real computational complexity is $O(m)$. Only the starting stage's computational complexity is $O(m \cdot n)$. Therefore, the over all the feature extraction procedure can be very fast. Our detector can scan 3–8 352 × 288 frames per second on a PC with 3.0 GHz processing speed without any specially designed speedup routines.

3.3. Feature selection by Realboost

In our approach, the size of space–time volume of the pedestrian is $30 \times 15 \times 5$. By convolving this volume with the 3D Haar-like filters of different sizes and locations, we can obtain about 50,000 features for each scan window. We use Realboost to select the most discriminative features and build a cascade classifier as in (Viola et al., 2003). We present the first five weak classifiers (features) in Fig. 6. From Fig. 6 we can see that only the fourth feature S3 is the static feature and all the other four features are dynamic features. Take the first dynamic feature D2 in Fig. 6a, for example. Since our negative training set contains both static and dynamic backgrounds, D2 is selected by Realboost to discard static backgrounds. By this single feature, more than 86% static backgrounds can be discarded. Therefore, when detecting the pedestrians from the video sequences that captured by the static cameras, our method can be very efficient. From Fig. 6b and e, we can see that the motion patterns of human legs are very important cues for our detector to identify a pedestrian which is consistent with human perception.

Our final cascade detector contains 20 stages and about 1000 weak classifiers (features). We also calculate the probability distributions of these seven types of filters. The results can be seen in Fig. 7, by which we can get an intuition on the importance of different types of features. As we expected, the contribution of the dynamic features is greater than the static features'. Especially for feature D2, Fig. 7 demonstrates that it is very important for

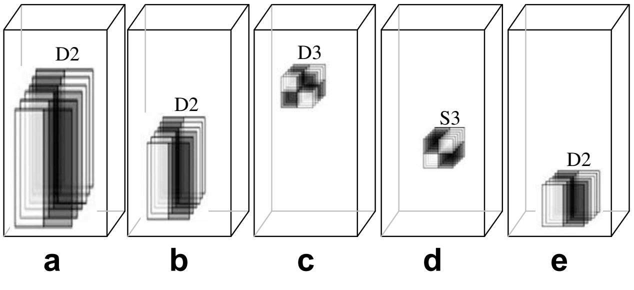


Fig. 6. The first five weak classifiers (features) selected by Realboost.

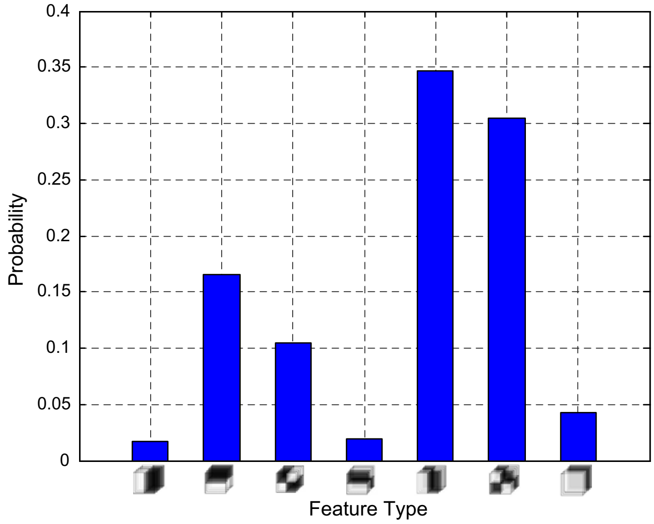


Fig. 7. The probability distribution of the selected features.

identifying the moving pedestrian. This result verifies our analysis in Section 3.2, which is the periodical motion patterns of the arms and legs are important cues for pedestrian detection. Quantitative evaluation of the different contributions of static and dynamic features can be fund in Section 4.

4. Experiments

In this section, we verify the effectiveness of our contour-motion feature by applying it to pedestrian detection and human activity recognition. In both applications, we use Realboost to select the most representative features.

4.1. Pedestrian detection

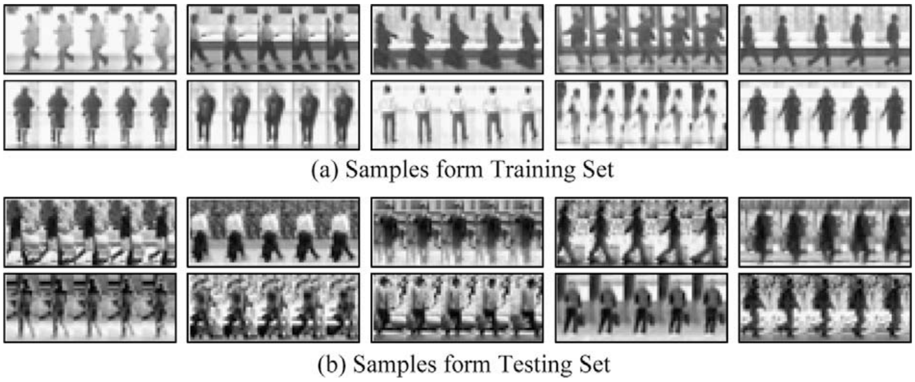
The evaluation of our method for pedestrian detection contains three parts: first, the quantitative analysis of the contributions of static and dynamic features; second, the performance of our method under different frame rate; third, the comparison with other baseline methods.

The frame rates of all the sequences in our experiments are range from 24 to 30 fps. The CAVIAR (CAVIAR, 2004) database is used as the benchmark database, all the positive training samples (pedestrians) are collected from this database. The positive samples are cropped out from the sequences and resize into $30 \times 15 \times 5$ space–time volumes (STV). Our experiments indicate that longer temporal length (up to 10 frames) can yields better performance. But for practical application, if we take more frames, the adaptivity (reaction speed) of the detector will be sacrificed. So in order to maintain a satisfiable performance and reaction speed, we choose five frames as the length of our training samples. If the frame rate of the video is quite different from 24 to 30 fps, this normalization number may need to be adjusted.

The positive training set contains 20,000 such volumes. Some selected positive samples can be seen in Fig. 8a. The negative training set contains 1500 clips selected from the surveillance video sequences, movies and our self-captured videos which contain no human bodies. For a 20-stage cascade classifier, it will take 3–4 days for training.

In order to verify the detection accuracy and generalization ability of the proposed method, we maintain two testing sets. The testing set 1 is also from the CAVIAR database and contains 118 video clips; the testing set 2 is selected from our self-capture video sequences and contains 761 video clips. The variation of scenes in the testing set 2 is larger than the testing set 1, which contain both indoor and out door scenes, varying lighting and complex backgrounds. Some pedestrian samples from testing set 2 are shown in Fig. 8b. The full frame images can be seen in the top row of Fig. 13. During detection, the size of our scan window is also $30 \times 15 \times 5$, the same as our training samples'. We resize the test window by 1.2 (horizontally and vertically) for each scale, and the slide step for both directions is 2 pixels for all the scales. In addition, the detection results in the following section are the results of detection window merging.

In the first experiment, we build four cascade detectors to evaluate the contributions of static and dynamic features. For the first one, all of the seven types of features in Fig. 4 are used, we refer to this detector as combined feature detector (CMF). For the second detector, we intend to evaluate the contribution of the dynamic features, so only dynamic features D1–D4 in Fig. 4 are used. In addition, we build two detectors to evaluate the static features, and both of these detectors are using the static features S1–S3 in Fig. 4. The difference between these two detector is that the one using the combined spatial-temporal gradient ($\alpha = 0.5$ in Eq. (1), the same as above two detectors) and the other using the spatial gradient only ($\alpha = 1.0$ in Eq. (1)). The receiver operating characteristic (ROC) curves of these four detectors on testing set 1 are presented in Fig. 9. The results reveal some observations: (1) the combined feature detector can achieve the best performance; (2) the dynamic features are more powerful than the static features in describing the pedestrian and (3) the temporal gradient is also helpful for improving performance, but the contribution is not as large as dynamic features. This result verifies our former assumption, the pedestrian is a space–time entity and both motion and appearance information are important for robust pedestrian detection.


(a) Samples form Training Set

(b) Samples form Testing Set

Fig. 8. Some training and testing samples for our experiments.

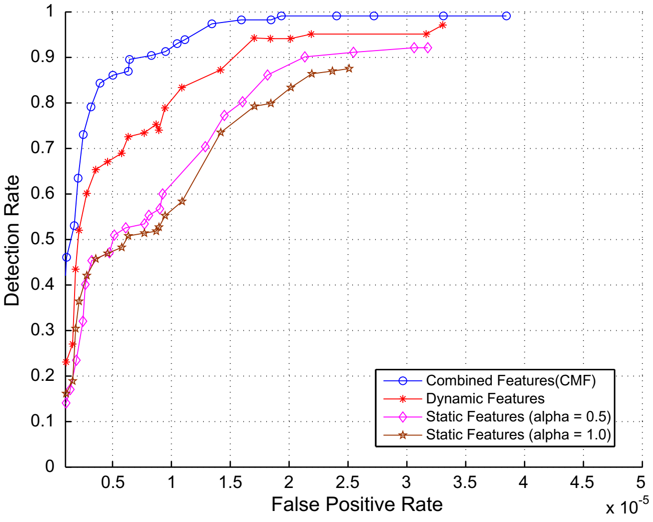


Fig. 9. The performance evaluation of different feature types.

Our second experiment is to evaluate the influence of the frame rate on the proposed method. For most of the object detection methods, in order to detect the same object of different scale, they normally search all the scales exhaustively. Take face detection for example, they just resize the image into every interested scale and detect the faces. But for video based detection, we should take the temporal dimension into consideration. A straightforward solution is just checking all the combinations of spatial scales and temporal scales. But it is not plausible for the applications where high detection speed is required. So it would be favorable if the detection method can be robust, at least partially, to the variation in the temporal dimension. In this experiment, we vary the sample steps of the images in the STVs from 1 to 5. Then we use the STVs with sample step 2 to train the detector and use STVs of other sample step for evaluation. The results are shown in Fig. 10, from which we can see that the ROC curves of sample step 1–3 are very close. The performances of step 4 and 5 drop slightly, but are still satisfiable. These results indicate that the propose method is robust to small variations of frame rate.

Based on the experimental results above, in practical application, we handle spatial scale and temporal scale separately. The spatial scales are covered by resizing the input frames during the detection procedure and the spatial scales are covered by using the training samples with different speed and frame rate.

In the third experiment, we manage to verify two critical issues of the proposed method. Are the gradient-based features really superior to the intensity based features for representing the pedestrian? Are the 3D Haar-like filters really effective for representing both the static and dynamic information? These two issues are closely related to our former two assumptions. We select two baseline approaches for comparison. In the first approach, we apply the 3D Haar-like features directly on the voxels of the space–time volume. This method is referred to as Intensity+3DHaar. Since the framework and performance evaluation criteria of our approach are very similar to Viola et al.'s well-known pedestrian detector (Viola et al., 2003), we select their method as the second baseline approach and refer to this method as Viola03. We refer to the proposed method in this paper is as CMF+3DHaar.

The testing results on dataset 1 are presented in Fig. 11. From these ROC curves, we can see all these three methods perform well. CMF+3DHaar is slightly better than Viola03 and Intensity+3DHaar when the false alarm rate is large than 0.4×10^{-5}.

We further test these three approaches on testing set 2. The ROC curves are presented in Fig. 12. The results of CMF+3DHaar are still satisfiable. The Intensity+3DHaar's performance is decreased. But the results of Viola03 are somewhat surprising. Its performance is deteriorated radically, even worse than Intensity+3DHaar. One possible explanation is that the backgrounds of these testing images are much complex than ones in the training sets, which can be seen in Fig. 8b. Another possible reason is that Viola03 relies on motion information between two consecutive images. But when person is moving slowly and the background is cluttered, the motion pattern between the two images is not obvious, thus the features from two frame difference are not so informative and discriminative. In order to verify this assumption, we decrease the number of frames used by the proposed method from five frames to two frames, and present the evaluation results in the Fig. 12 (referred to as CMF+3DHaar(n frame)). As we expected, when decrease the number of frame to 2, the method will yield comparable results as the Viola's one. These experiments reveal the following observations: (1) the long-term motion pattern is more distinctive for describing the pedestrian and can increase the generalization ability of the method and (2) the gradient-based descriptors are more robust than intensity based ones for representing the appearance of pedestrians.

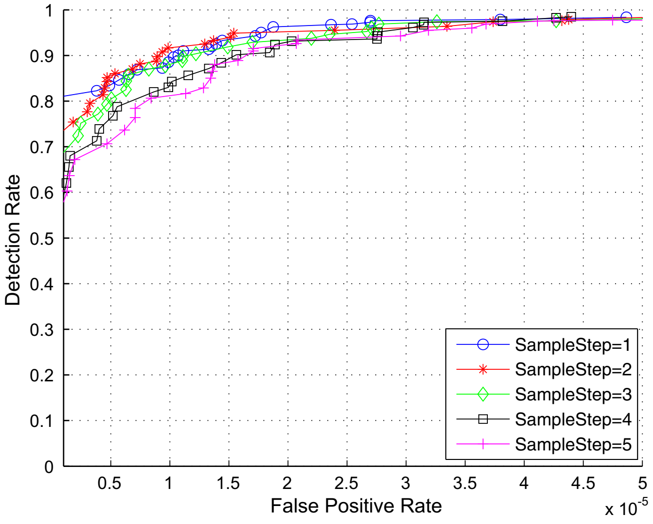


Fig. 10. The effect of different sample step.

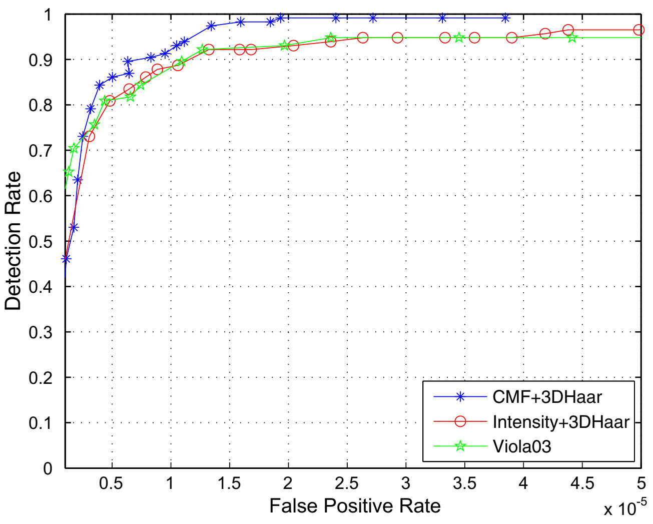


Fig. 11. Evaluation results on testing set 1.

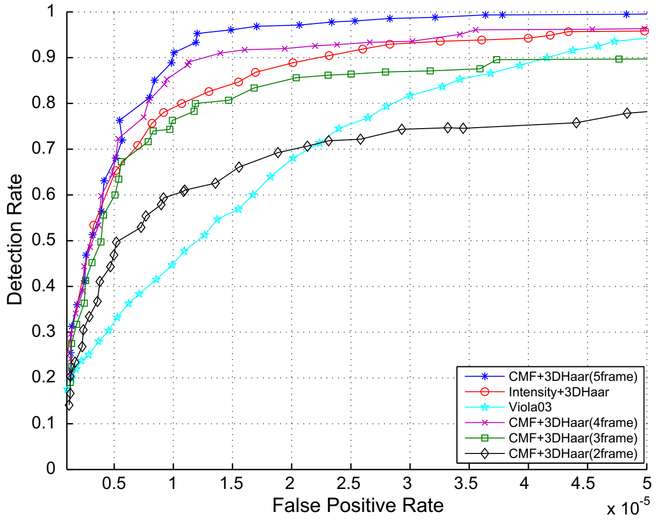


Fig. 12. Evaluation results on testing set 2.

We also present some detection results in Fig. 13. The top two images are selected from our self-captured sequence; and the bottom two images are selected from PETS06 (PETS06, 2006) dataset. In these experiments, the complex background (the top left image), the illumination conditions (the top two images) and the viewing directions (the bottom two images) are quite different from the training samples. Our approach can still achieve satisfiable detection results. The false detection window in the top-right image (the third detected window from left to right) is mainly because of our unpolished window fusion method. Here, we just combine the seriously overlapped windows into a single one; this simple strategy will not work well when two pedestrians are close to each other. Bayesian inference based window fusion strategy may provide better result.

4.2. Human activity recognition

We put forward our method into human activity recognition. The benchmark data set used in this experiment is the same as Schuldt et al. (2004). This dataset contains six types of human actions (walking, jogging, running, boxing, hand waving and hand

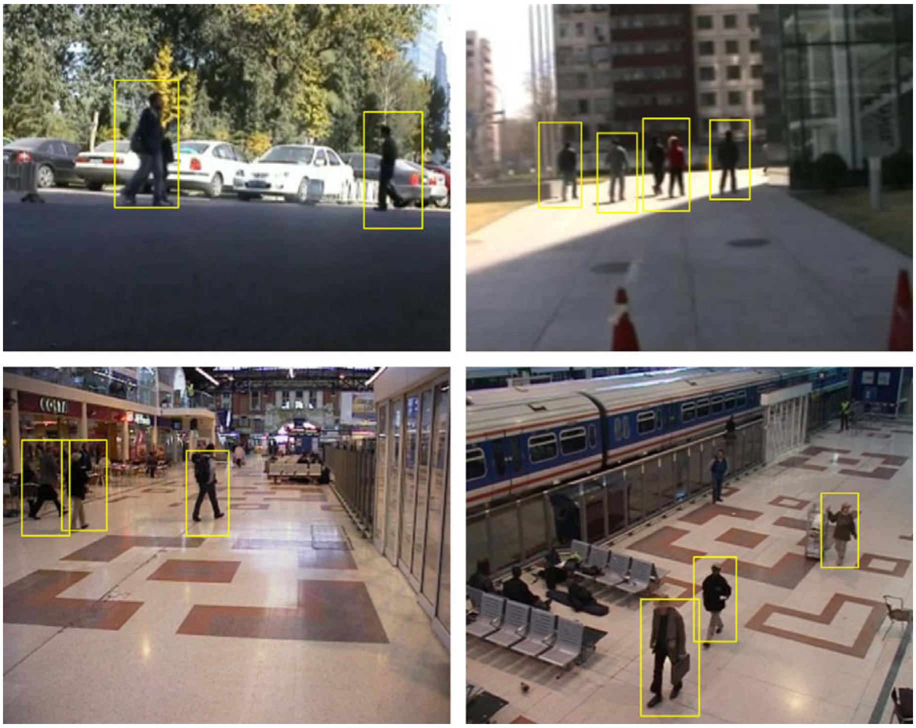


Fig. 13. Some detection results of our approach.

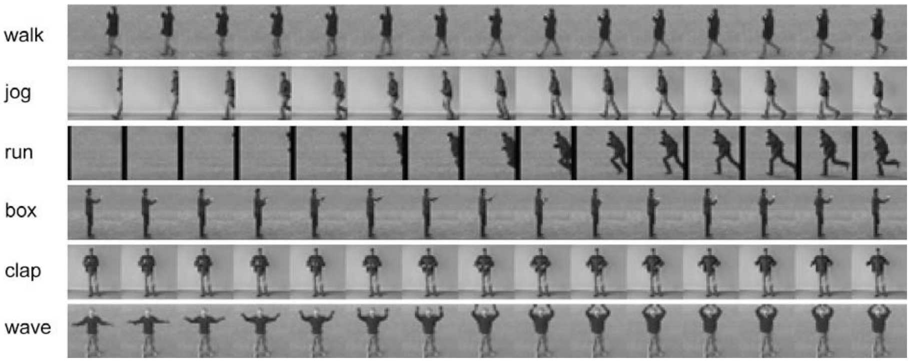


Fig. 14. Samples for human activity analysis.

Table 1
The confusion matrix of the proposed method with the trace 463.1

Ours Tr = 463.1	Walk	Jog	Run	Box	Clap	Wave
Walk	80.1	11.3	6.1	0.8	0.6	1.1
Jog	10.6	63.7	22.4	0.7	1.8	0.8
Run	4.6	14.1	77.6	2.1	1.2	0.5
Box	1.3	1.8	2.1	87.4	6.1	1.3
Clap	0.6	0.3	5.6	15.1	75.6	2.8
Wave	2.7	2.6	0.9	7.3	7.8	78.7

Table 2
The confusion matrix of the Keyan's method with the trace 377.8

YanKe Tr = 377.8	Walk	Jog	Run	Box	Clap	Wave
Walk	80.6	11.1	8.3	0	0	0
Jog	30.6	36.1	33.3	0	0	0
Run	2.8	25	44.4	0	27.8	0
Box	0	2.8	11.1	69.4	11.1	5.6
Clap	0	0	5.6	36.1	55.6	2.8
Wave	0	5.6	0	2.8	0	91.7

Table 3
The confusion matrix of the Schuldt's method with the trace 430.3

Schuldt Tr = 430.3	Walk	Jog	Run	Box	Clap	Wave
Walk	83.8	16.2	0	0	0	0
Jog	22.9	60.4	16.7	0	0	0
Run	6.3	38.9	54.9	0	0	0
Box	0.7	0	0	97.9	0.7	0.7
Clap	1.4	0	0	35.4	59.7	3.5
Wave	0.7	0	0	20.8	4.9	73.6

clapping) performed several times by 25 subjects in four different scenarios. The size of STV is 30 × 30 × 15, and some samples can be seen in Fig. 14. The average number of STVs for each action is about 16,000 (with vertical flip).

Our experimental setting is the same as Schuldt et al. (2004), and we also use eight persons' actions for training, eight persons' for validation and nine persons' for testing. We build a 1-to-rest cascade classifier for each action, then use the validation set to determined the best threshold. We compare our method against (Schuldt et al., 2004; Ke et al., 2005). The confusion matrices of these three methods are presented in Tables 1–3. The trace of the confusion matrix is the measure of classification accuracy, bigger trace indicate better classification performance. The trace of the proposed method is 463.1, which outperform the Keyan's method with the trace 377.8 and Schuldt's method with trace 430.3.

5. Conclusion and further focus

In this paper, we present a segmentation-free pedestrian detection approach in space–time domain. The camera can be either static or dynamic. Our two basic assumptions are: the combination of appearance and motion information can yield better results for pedestrian detection; the gradient-based descriptor can be more robust in representing the highly articulated human body. Based on these two assumptions, we represent the pedestrians in 3D distance transform volume and extract its contour-motion features by 3D Haar like filters. Experiments show that our contour-motion feature has remarkable generalization ability. Further, our space–time analysis method has been easily extended to human action recognition application and satisfiable performance have been achieved. This further verifies that our contour-motion feature can capture the appearance and motion feature simultaneously, which is a basic property of human activities.

There are two aspects should be addressed in our future work. First, our current detector is a full body detector and is not robust enough to the occlusion problem. One possible solution is to extend the current method into a multiple-part detector and using Bayesian inference to fuse the detection results. Second, we only use amplitude of the gradient. Many research results indicate that the orientation of the gradient can provides very useful information for complex object recognition. So we will focus on exploiting the efficient gradient orientation based pedestrian detection approach in space–time domain in our future work.

Acknowledgements

This paper is partially supported by National Natural Science Foundation of China under Contract Nos. 60332010, 60673091 and 60772071, Hi-Tech Research and Development Program of China under Contract Nos. 2006AA01Z122 and 2007AA01Z163, 100 Talents Program of CAS, Program of New Century Excellent Talents in University (NCET-04-0320), Natural Science Foundation of Beijing municipal under Contract Nos. 4061001 and 4072023, and ISVISION Technology Co. Ltd.

References

Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R., 2005. Actions as space–time shapes. In: 10th IEEE Internat. Conf. Computer Vision, vol. 2, pp. 1395–1402.

CAVIAR, 2004. <http://www.dai.ed.ac.uk/homes/rbf/caviar/>.

Cutler, R., Davis, L.S., 2000. Robust real-time periodic motion detection, analysis, and applications. IEEE Trans. Pattern Anal. Machine Intell. 22 (8), 781–796.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Schmid, C., Soatto, S., Tomasi, C. (Eds.), Internat. Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 886–893.

Dalal, N., Triggs, B., Schmid, C., 2006. Human detection using oriented histograms of flow and appearance. In: Eur. Conf. Computer Vision, vol. 2, pp. 428–441.

Ferrari, V., Tuytelaars, T., Gool, L.V., 2006. Object detection by contour segment networks. In: Eur. Conf. Computer Vision, vol. 3, pp. 14–28.

Gavrila, D.M., 1998. Multi-feature hierarchical template matching using distance transforms. In: IEEE Internat. Conf. on Pattern Recognition, vol. 1, Brisbane, Australia, pp. 439–444.

Gavrila, D.M., 2000. Pedestrian detection from a moving vehicle. In: Eur. Conf. on Computer Vision, vol. 2, Dublin, Ireland, pp. 37–49.

Ke, Y., Sukthankar, R., Hebert, M., 2005. Efficient visual event detection using volumetric features. In: Internat. Conf. Computer Vision, vol. 1, pp. 166–173.

Leibe, B., Seemann, E., Schiele, B., 2005. Pedestrian detection in crowded scenes. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 878–885.

Liu, X., Fujimura, K., 2004. Pedestrian detection using stereo night vision. IEEE Trans. Vehicular Technol. 53 (6).

Lowe, D.G., 1999. Object recognition from local scale-invariant features. In: IEEE Internat. Conf. Computer Vision, pp. 1150–1157.

Mikolajczyk, K., Schmid, C., Zisserman, A., 2004. Human detection based on a probabilistic assembly of robust part detectors. In: Eur. Conf. on Computer Vision, vol. 1, pp. 69–81.

Mohan, A., Papageorgiou, C., Poggio, T., 2001. Example-based object detection in images by components. IEEE Trans. Pattern Anal. Machine Intell. 23 (4), 349–361.

Munder, S., Gavrila, D.M., 2006. An experimental study on pedestrian classification. IEEE Trans. Pattern Anal. Machine Intell. 28 (11), 1863–1868.

Papageorgiou, C., Poggio, T., 2000. A trainable system for object detection. Internat. J. Comput. Vision 38 (1), 15–33.

PETS06, 2006. In: IEEE Internat. Workshop on Perform. Evaluation of Tracking and Surveillance.

Ramanan, D., Forsyth, D.A., 2003. Finding and tracking people from the bottom up. In: Computer Vision and Pattern Recognition, vol. 2. pp. 467–474.

Sabzmeydani, P., Mori, G., 2007. Detecting pedestrians by learning shapelet features. In: IEEE Conf. Computer Vision and Pattern Recognition.

Schapire, R.E., Singer, Y., 1999. Improved boosting algorithms using confidence-rated predictions. Machine Learn. 37 (3), 297–336.

Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: A local svm approach. In: Internat. Conf. on Pattern Recognition, vol. 3. pp. 32–36.

Seemann, E., Fritz, M., Schiele, B., 2007. Towards robust pedestrian detection in crowded image sequences. In: IEEE Conf. Computer Vision and Pattern Recognition.

Shashua, A., Gdalyahu, Y., Hayun, G., 2004. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In: IEEE Intell. Vehicles Sympos. pp. 1–6.

Shechtman, E., Irani, M., 2005. Space–time behavior based correlation. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1. pp. 20–25.

Tuzel, O., Porikli, F., Meer, P., 2007. Human detection via classification on riemannian manifolds. In: IEEE Conf. on Computer Vision and Pattern Recognition.

Ukrainitz, Y., Irani, M., 2006. Aligning sequences and actions by maximizing space–time correlations. In: Eur. Conf. Computer Vision, vol. 3, pp. 538–550.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, Hawaii, pp. 511–518.

Viola, P., Jones, M.J., Snow, D., 2003. Detecting pedestrians using patterns of motion and appearance. In: IEEE Internat. Conf. on Computer Vision, vol. 2, Nice, France, pp. 734–741.

Wexler, Y., Shechtman, E., Irani, M., 2004. Space–time video completion. In: Computer Vision and Pattern Recognition, vol. 1, pp. 120–127.

Wu, B., Nevatia, R., 2005. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: IEEE Internat. Conf. on Computer Vision, vol. 1, pp. 90–97.

Wu, Y., Yu, T., 2006. A field model for human detection and tracking. IEEE Trans. Pattern Anal. Machine Intell. 28 (5), 753–765.

Xu, F., Liu, X., Fujimura, K., 2005. Pedestrian detection and tracking with night vision. IEEE Trans. Intell. Transport. Systems 6 (1), 63–71.

Zhao, T., Nevatia, R., 2003. Bayesian human segmentation in crowded situations. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, Madison, WI, pp. 459–466.

Zhu, Q., Avidan, S., Yeh, M.-C., Cheng, K.-T., 2006. Fast human detection using a cascade of histograms of oriented gradients. In: Computer Vision and Pattern Recognition, vol. 2, pp. 1491–1498.