3D HAAR-LIKE FEATURES FOR PEDESTRIAN DETECTION

Xinyi Cui[1,2]*, Yazhou Liu*[1,2]*, Shiguang Shan*[2,3]*, Xilin Chen*[2,3]*, Wen Gao*[1,2]
[1]*Harbin Institute of Technology, Harbin, China*
[2] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
[3]*Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, Beijing, China*
{xycui, yzliu, sgshan, xlchen, wgao}@jdl.ac.cn

ABSTRACT

One basic observation for pedestrian detection in video sequences is that both appearance and motion information are important to model the moving people. Based on this observation, we propose a new kind of features, 3D Haar-like (3DHaar) features. Motivated by the success of Haar-like features in image based face detection and differential-frame based pedestrian detection, we naturally extend this feature by defining seven types of volume filters in 3D space, instead of using rectangle filter in 2D space. The advantage is that it can not only represent pedestrian's appearance, but also capture the motion information. To validate the effectiveness of the proposed method, we combine the 3DHaar with support vector machine (SVM) for pedestrian detection. Our experiments demonstrate the 3DHaar are more effective for video based pedestrian detection.

1. INTRODUCTION

Human detection is important for a variety of applications, such as visual surveillance, smart room, and automatic driver-assistance system. But it is a challenging task because of the wide variability in appearance due to clothing, articulation and illumination conditions.

In the last few years, many approaches to pedestrian detection have been proposed in both still images and video sequences. For pedestrian detection in still images, researchers mainly focus on the features for modeling pedestrians. In the early stage, Papageorgiou et al [1] used Haar-based representation combining with a polynomial SVM to classify pedestrians. In order to detect pedestrians with partial occlusion, Mohan et al. [2] improved the method in [1] by dividing human body into fours parts: head-shoulder, legs and left/right arms. Later, researches start describing pedestrians by local descriptors in Implicit Shape Model (ISM) [3], which is a popular method for object detection and recognition. Also some modified SIFT descriptors (i.e. Histogram of Oriented Gradient) are used in pedestrian detection with other classifiers such as SVM [4] and cascade AdaBoost [5]. Recently, Wu and Yu [6] model pedestrians in a Markov Random Field, to solve the problem of non-rigid shape and partial occlusion. Munder and Gavrila [7] carry out an extensive experimental study on various features and classifiers for pedestrian detection.

Much progress has been made in detection and tracking of pedestrians in video sequences [8-11]. However, most methods rely on segmentation of a foreground motion blob. Motion segmentation by background modeling is simple and effective when camera is stationary and changes in illumination are gradual. But for many applications the camera may move and illumination may change suddenly. In such case, direct detection of human pattern can solve the problem. Considering this requirement, we propose a method detecting pedestrian directly from the video sequences, while being independent of motion detection.

Recently, Viola et al.'s [11] and Dalal et al.'s [12] work indicate the combination of static and dynamic information can improve the detection accuracy. Viola et al. [11] presented a pedestrian detection algorithm with Haar-like features between two frame difference, considering both the appearance and motion information. Our method is a natural generalization of this algorithm. Instead of presenting features between two frames, we extract Haar-like features among multiple frames, which can capture more motion information representative of people. Since they are extracted in a space-time volume, we call them 3D Haar-like features. This kind of features is distinctive and robust to represent the motion and appearance pattern of moving people. The experimental results further verify the effectiveness of our method.

The remaining of this paper is organized as follows. Section 2 reviews the relevant algorithm of the Haar-like features proposed in [11]. Section 3 gives the detailed description of our 3D Haar-like features. Section 4 presents our experimental results, and section 5 gives the summation and the focus of our future work.

2. RELATED WORKS

The proposed method can be regarded as a natural generalization of Viola et al.'s algorithm [11]. So we start with a short description of Viola's algorithm. Given a pair of images I_t and I_{t+1} in time, five differential images are computed. Δ is the difference image between image

1-4244-1017-7/07/$25.00 ©2007 IEEE

I_t and I_{t+1} , U is the difference image between I_t and I_{t+1} with one pixel shifting up, and L, R, D between I_t and I_{t+1} with one pixel shifting down, left and right respectively.

Then four types of rectangle filters are defined on these five images. One type of filters compares sums of absolute differences between Δ and one of $\{U, L, R, D\}$. It extracts information related to the likelihood that a particular region is moving in a given direction. The second type of filters compares sums within the same motion image, with rectangle filters similar to [13]. It measures something closer to motion shear. The third type of filters measures the magnitude of motion images, which simply compute the sum within the detection window. They also use appearance filters which operate on the first input image I_t . All the filters can be evaluated rapidly using the Integral Image. Then the training process uses AdaBoost to select a subset of features and construct the cascade classifier. This method achieves good result.

3. 3D HAAR-LIKE FEATURES

The success of Viola et al.'s algorithm [11] just lies in that it uses the motion information between consecutive two images. But when person is moving slowly, the motion pattern between the two images is not obvious, thus the features from two frame differences are not so distinctive. In order to capture the long-term motion patterns among multiple frames and record the person's appearance features at the same time, we extract Haar-like features from a series of consecutive frames instead of two frames.

Although we can obtain more motion information from multiple frames, we can't use the whole frames of the video to detect a person. It consumes a lot of time and inapplicable for a detection task. In addition, a target may not always stay in one position along the whole video. Thus, we divide videos into small space-time volumes, which only contain several frames and looks like a cubic window in a video sequences. The space-time volumes in our method is similar to the 2D search window in Viola et al.'s algorithm [11]. A space-time volume can be seen as an independent and whole unit, which various 3DHaar features are extracted from. Our goal is to give a judgment on whether a space-time volume has a person or not. In the next section, we will give the detailed description of 3D Haar-like features.

3.1. Detailed Description of 3D Haar-like Features

We give the detailed description of 3DHaar features in this section. 3DHaar features are extracted in a space-time volume. They can be seen as cubic filters. Specifically, we adopt seven types of 1-order 3DHaar features. See Figure 1 for a detail. For every type of the cubic filters, the feature value is the absolute difference of the pixel intensity sum between the black and white regions. Unlike the 2D Haar-like features using the difference value [13], we only use the absolute value. It is because the intensity variations of pedestrians due to clothes, articulation and attachment are more complicated, and the structures are less definite than face. Similar cubic filters have been used by [14] for visual event detection, but only filters(a) (b) (g) are used in optical flow field in their work.

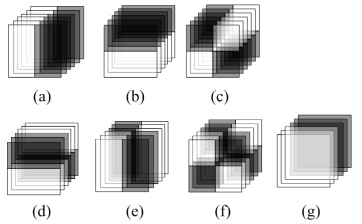

Fig.1. Seven types of 3D Haar-like feature

The cubic filters showed in Figure 1 (a) (b) (c) are the static features, which are similar to the 2D Haar-like features used in [13]. They only compare sums of the same regions in temporal coordinate. Such features are used to describe the pedestrian's appearance information.

The features showed in Figure 1 (d) (e) (f) (g) are the dynamic features. They compare sums of the different regions in temporal space. Take (d) for example; it computes the difference between diagonal pairs of cubic in temporal dimension. Since the feature value is computed among multiple frames, it can better describe the motion information in the scene and capture more motion patterns of pedestrians.

3.2. Integral Volume

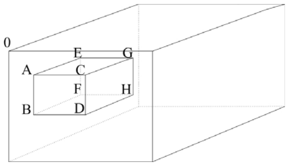

Fig. 2. The sum of the pixels within the box ABCDEFGH can be computed with seven plus/minus operations. The sum is $iv(H)-iv(D)-iv(F)-iv(G))+iv(B)+iv(C)+iv(E)-iv(A)$, where 0 is the origin.

Integral Image is a fast method to compute 2D Haar-like features. This led to a real-time face detection system [13] and human detection system [11]. To compute 3DHaar feature values efficiently, we also use the idea of Integral Image. The only difference is that we compute Integral Image in three dimensions and we refer it as to Integral Volume. Given the origin of coordinate frame, the value of Integral Volume at location (x,y,t) contains the sum of the pixels which location indices are less than the current location. More specifically:

$$iv(x,y,t) = \sum_{x'\leq x, y'\leq y, t'\leq t} i(x',y',t'),$$

where $iv(x,y,t)$ is the integral volume and $i(x,y,t)$ is the original volume. Using the Integral Volume any cubic sum can be computed in eight array references (seven plus/minus operations). See Figure 2 for example. Since the two-box cubic filters (Figure 1 (a) (b) (g)) involve adjacent boxes, they can be computed in 12 array references. Figure 1 (c) (d) (e) need 18 array references; Figure 1 (f) needs 27 array references. Since plus/minus operation is fast, the feature extraction can be processed rapidly.

3.3. Feature Representation

In order to represent the space-time volume thoroughly, we scan the volume densely using the seven types of cubic filters. Generally, each cubic filter is third or half overlapped by another one. Given a space-time volume with size (H,W,T) , representing the height, width and frame number respectively, we define the cubic filter size (h,w,t) and cubic filter scanning step $(hStep,wStep,tStep)$. Then the feature number N in a space-time volume is

$$\lceil (H-h)/hStep \rceil \times \lceil (W-w)/wStep \rceil \times \lceil (T-t)/tStep \rceil \times 7 .$$

The total N features form a N -element vector representation of a space-time volume. For detection task, we use the Gaussian Kernel SVM as classifier. In the experimental evaluation section, we will analyze the experiments on how the cubic window parameters and frame number influence the detection performance.

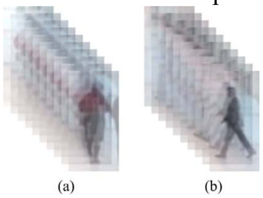

Fig. 3. Two example space-time volumes from CAVIAR dataset. The space-time volume is 60×30×10

4. EXPERIMENTAL EVALUATION

In this section, we first discuss the data preparation and outline the experiment setup. Then we give the results of our experiments.

We train and test our detector on CAVIAR data set [15], which has 52 scenes along and across the hallway in a shopping center. According to our experimental goal, we segment the regions containing persons from the origin videos, and add a margin of 2 pixels to preserve contour information. Then the segmented images are rescaled to 60×30, and consecutive 10 segmented images form a space-time volume as described in Section 3. Figure 3 shows two examples. It can be seen from the examples that such kind of volumes has obvious appearance characteristics as well as certain motion information. The negative examples are made in the same way; the only difference is that they are randomly sampled from videos known containing no pedestrian. In this way, we gather 10,000 positive and 10,000 negative space-time volumes.

We carry out three experiments to explore the effectiveness of 3DHaar features. First, we evaluate the effect of tunable window size. Then, we analyze how the frame number of cubic filter influences detection performance. Third, we evaluate 3DHaar features in real applications.

Given a cubic filter with size (h,w,t), we discuss how the three parameters influence the detection performance, and also to confirm our hypothesis of the effectiveness of 3DHaar features. First, we evaluate the window size (h,w) of the cubic filter. In this experiment, we use 5000 positives and 3000 negatives as training set, and 5000 positives and 5000 negatives as testing set. In addition, we set $t=9$ and the scanning step is third or half of the window side. We change the window size from 4×4 to 20×20 , and Figure 4 plots the error rate. It can be seen from the figure that window size of 16 pixel height and 4-8 pixel width performs best. In fact, this window size is interestingly coincident with the proportion of standing persons, and the cubic filters are just trying to capture the characteristics of pedestrians.

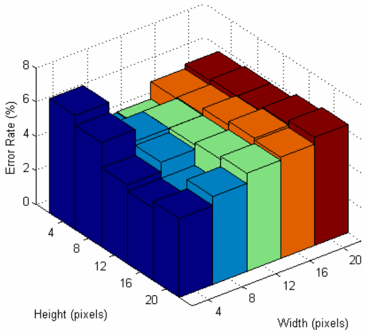

Fig. 4. Performance comparisons of 3D Haar-like features with different window sizes

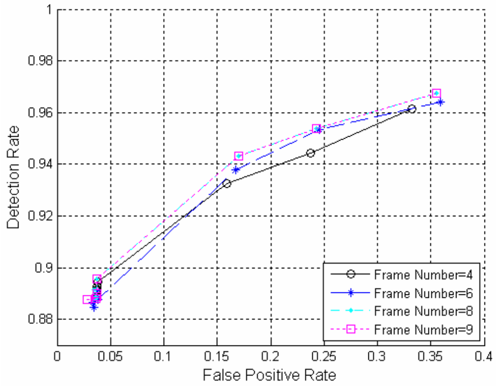

Fig. 5. Comparison of detection performance given different frame numbers

We next analyze the effect of frame number t on pedestrian detection. We use the window size 16×8 since it achieves good performance in the first experiment. This experiment uses 4000 positives and 4000 negatives as training set, and 2000 positives and 2000 negatives as testing set. Figure 5 plots the ROC curves with frame number t of 4, 6, 8 and 9. It shows that when t becomes larger, the detection performance increases. It means

3DHaar features capture more distinctive information with more frames. Thus this experiment confirms our hypothesis that 3DHaar features extracted in multiple frames are effective and distinctive.

To evaluate the performance variations in practice, the third experiment is to determine the ROC variance by varying training and test sets. This test procedure is similar to [7]. We divide the training set into three sub training sets and test set into two sub test sets. Each sub set has 2000 space-time volume samples. For each experiment, we use two out of the three sub training sets to train the SVM classifier, and then test this classifier on the two sub test set respectively. In this way, we can obtain six different ROC curves. When taken as six independent tests which follow a normal distribution, a confidence interval of the true mean detection rate can be estimated as follows:

$$\bar{y}\pm t(\alpha/2,N-1)\frac{s}{\sqrt{N}}\approx\bar{y}\pm1.05s,$$

where $\bar{y}$ and s denote the estimated mean and standard deviation respectively, $1-\alpha=0.95$ is the desired confidence interval, and $N=6$ is the number of tests.

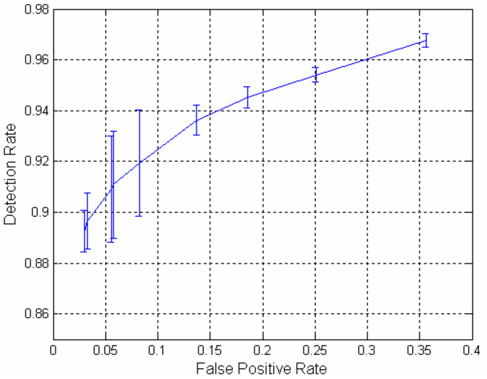

Fig. 6. Performance of 3DHaar features for pedestrian detection

Resulting ROC curve is given in Figure 6. We can see that 3DHaar features achieve a good performance of detection rate 91% given false positive rate 5%. The confidence interval is less than 5 %, so the detection performance is stable.

5. CONCLUSION

In this paper, we propose a new kind of feature, 3D Haar-like feature. This feature naturally extends 2D Haar-like feature to 3D space. We adopt seven types of volume filters here to represent pedestrian's appearance and motion information. We extract these features in a space-time volume and use Gaussian kernel SVM as classifier. The system is evaluated on CAVIAR dataset. We studied the influence of various parameters, including the window size and frame number on detection performance. Finally we have shown that 3D Haar-like features give good result for pedestrian detection task.

The Future Work. Although our current SVM detector achieves a good result, the performance is limited by the fixed window size. In the future, we will improve the detector using AdaBoost cascade to select strong features with multiple scales.

6. ACKNOWLEDGEMENT

This paper is partially supported by National Natural Science Foundation of China under contract No.60332010, and No. 60673091, Hi-Tech Research and Development Program of China (No.2006AA01Z122), and ISVISION Technology Co. Ltd.

7. REFERENCES

[1] T. P. Constantine Papageorgiou, "A trainable System for Object Detection", *IJCV*, vol. 38, pp. 15-33, 2000.

[2] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-Based Object Detection in Images by Components," *IEEE Trans. PAMI*, vol. 23, pp. 349-361, 2001.

[3] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian Detection in Crowded Scenes", CVPR, vol.1, pp.878-885, 2005

[4] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", CVPR, vol. 2, pp.886-893, 2005

[5] Q. Zhu, S. Avadan, M.-C. Yeh, and K.-T. Cheng, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients", CVPR, vol. 2, pp.1491-1498, 2006

[6] Y. Wu and T. Yu, "A Field Model for Human Detection and Tracking", *IEEE Trans. PAMI*, vol. 28, pp. 753-765, 2006.

[7] S. Munder and D. M. Gavrila, "An Experimental Study on Pedestrian Classification," *IEEE Trans. PAMI,* vol. 28, pp. 1863-1868, 2006.

[8] S. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-Time Surveillance of People and Their Activities", *IEEE Trans. PAMI*, vol. 22(8), pp. 809-830, 2000.

[9] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time Tracking of the Human Body", *IEEE Trans. PAMI*, vol. 19(7), pp. 780-785, 1997.

[10] N. T. Siebel and S. Maybank, "Fusion of Multiple Tracking Algorithm for Robust People Tracking", ECCV, pp. 373-387, 2002.

[11] P. Viola, M. Jones, and D. Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance", ICCV, vol. 2, pp. 734-741, 2003.

[12] N. Dalal, B. Triggs, and C. Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance", ECCV, vol. 2, pp. 428-441, 2006.

[13] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features", CVPR, vol. 1, pp. 511-518, 2001.

[14] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient Visual Event Detection using Volumetric Features", ICCV, vol. 1, pp. 166-173, 2005.

[15] "http://homepages.inf.ed.ac.uk/rbf/CAVIAR."