International Journal of
Science and Engineering Investigations vol. 1, issue 3, April 2012
ISSN: 2251-8843

Spatial-Temporal Words Learning for Crowd Behavior Recognition

Chuanxu Wang[1], Chenchen Dong[2]
[1,2] Institute of Informatics, Qingdao University of Science and Technology, Qingdao China
([1]wangchuanxu_qd@163.com, [2]mable.101@163.com)

Abstract-In this paper we introduce a method to detect abnormal behavior in crowd scenes using spatial-temporal words, in which Spatial Temporal Interest Points (STIPs) are extracted as crowd behavior movement features. For this purpose, three methods for STIPs extraction are compared and analyzed, which are Harris corner, Gabor wavelet and Hessian Matrix. Then we select Hessian matrix which can get scale-invariant STIPs. Gradient histogram, optical flow histogram and spatio-temporal Haar feature are used to build descriptors for STIPs. In normal behavior modeling Bag-of-words strategy is used, the keywords of which are produced by GMM based on EM estimation. Then training samples are divided into several clips which are described in probability vectors using keywords, we combine all vectors as a normal behavior codebook. In recognition phase, we calculate the similarity distance between the coding vector of the test samples and the codebook, the abnormal behavior can be detected when the minimal distance exceeds a threshold. We verify the algorithm in UMN and UCF datasets, which shows that the proposed algorithm has effective identification for crowd abnormal behavior, and it has good robustness against scale variant and illumination changing.

Keywords- *Crowd abnormal behavior; Spatial-temporal interest points; Bag-of-words; GMM*

I. INTRODUCTION

Crowd abnormal behavior detection gradually attracts widespread attentions of researchers in video surveillance. Ramin Mehran et al.[1] propose the establishment of particles in the image, they use SFM to describe the interaction between particles and the surrounding space, the strength of which can represent the pedestrians behavior in video. Shandong Wu[2] adopts a particle flow mode, and describes the local trajectory through the particle track, which realizes the abnormal behavior detection and localization. With respect to abnormal behavior detection in extreme crowded scene, Vijay Mahadevan et al. [3] use MDT [4] modeling video sequence, and detect the abnormity of the model in space and time respectively, then combine them to determine whether the abnormal behavior appears; Ramin Mehran et al.[5] propose using streakline to characterize the crowded scene, and clearly describe the behavior in complicated crowd scenes.

This paper presents a crowd abnormal behavior detection method based on STIPs. We compare the performance of three main methods of STIPs extraction; then introduce optical flow histogram, gradient histogram and Haar feature to build STIPs descriptors. Finally, we accomplish the normal behavior modeling and abnormal behavior detection by using Bag-of-words, and verify its effectiveness with UCF and UMN datasets.

II. EXTRACTION OF STIPS

STIPs methods do not require background modeling, which can effectively overcome the problems existing in global spatial-temporal information representation ways [6]. They provide a simple means to understand and analyze interest events in videos. The typical methods are as followings: the spatio-temporal Harris corner proposed by Ivan Laptev[7], the Gaussian and Gabor wavelet by Dollár[8], the Hessian matrix by Geert Willems[9].

A. *Spatio-temporal Harris corner detection*

Firstly the video stream is modeled in scale space:

$$L(x,y,t;\sigma_l^2,\tau_l^2) = g(x,y,t;\sigma_l^2,\tau_l^2) * I \qquad (1)$$

with I is the input image, σ_l^2 and τ_l^2 are the spatial and temporal scales respectively. $g(x,y,t;\sigma_l^2,\tau_l^2)$ is the separable spatial-temporal Gaussian smoothing filter.

Then we build a 3×3 spatio-temporal second-moment matrix:

$$\mu = g\left(;\sigma_i^2;\tau_i^2\right) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \qquad (2)$$

where L_x, L_y, L_t are the first-order derivatives of L in the x, y, t direction, which are defined as $L_\xi\left(;\sigma_l^2;\tau_l^2\right) = \partial_\xi\left(g * f\right)$.

So response function of STIPs can be defined as:

$$\begin{aligned} H &= \det(\mu) - k * trace^3(\mu) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3 \end{aligned} \qquad (3)$$

with $\lambda_1, \lambda_2, \lambda_3$ are the eigenvalues, and k should be large enough. If H gets the positive local maxima at the point (x,y,t), it can be regarded as a STIP.

For suitable scale selected, Ivan uses the normalized Laplacian. Figure 1 shows the STIPs. The positions of blue circles correspond to the locations where the STIPs exist, while the radiuses stand for the scales.

This method gets lesser feature points and relies on the scale selections, which is quite sensitive with different scales.

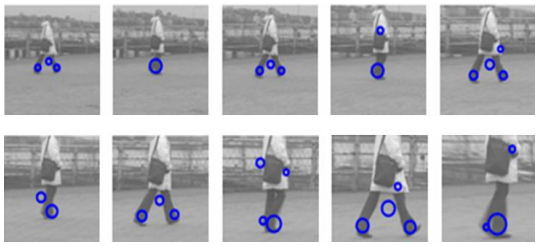


Figure 1. Extraction results of STIPs based on Ivan Laptev

B. *Gaussian and Gabor wavelet detection*

This approach determines the STIPs through calculating 2-D spatial Gaussian and 1-D temporal Gabor wavelet response function. The extreme point will be STIP, if it reaches local maxima and larger than a certain threshold.

The response function is defined as:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \qquad (4)$$

where I is the input image, $g(x, y, \sigma)$ is the 2-D spatial Gaussian smoothing filter, h_{ev} and h_{od} are the 1-D Gabor temporal filters.

（a） （b） （c）

Figure 2. STIPs results based on Dollár

Figure 2 shows STIPs which are extracted from walking behavior in simple scenes. This method produces lavish STIPs, but still subject to the scale variations.

C. *Hessian-based detection*

In the pre-processing stage of the algorithm, the integral video and box-filter are used. The Hessian matrix with second partial derivatives is expressed as:

$$H(\cdot;\sigma_i^2;\tau_i^2) = \begin{pmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{yx} & L_{yy} & L_{yt} \\ L_{tx} & L_{ty} & L_{tt} \end{pmatrix} \qquad (5)$$

where $L_{xx} = \frac{\partial}{\partial x^2} g(\sigma) * I(x,y,t)$, $g(\sigma) = \frac{1}{\sqrt{(2\pi)^3 \sigma^2}} e^{-(x^2+y^2+t^2)/2\sigma^2}$, the same as the others.

In order to speed up computing, a box filter is used to approximate the calculation of the Gaussian partial derivative, and the number of STIPs can be controlled according to the threshold, the scale selections may be obtained through salience testing, which avoids the iterative or the combination of scales. This method can give substantial and scale-invariant STIPs.

Table 1 shows the comparisons of these three approaches in amount of STIPs, scale selection and computation time. Willems method has the advantages of computation speeding, quantity controllable and scale invariant features. We select this method in crowd abnormal behavior detection, for crowed scenes are usually of great variability, diverse number of human beings and complex movements.

TABLE I. COMPARISON OF THREE METHODS IN EXTRACTING STIPS

	Extraction	**Types of STIPs**	**Scale selection**	**Computation time**
Ivan method	Harris corner	sparse	iterative	Not considered
Dollár method	Gaussian(2D)+ Gabor(1D)	dense	Combination	Not considered
Willems method	Hessian matrix	adjustable	significant test	integral video and box-filter

Figure 3. Normal behavior STIPs in UMN database.

In time and space domain we use a scale octave respectively, in which there are 5 small scales. Figure 3 and 4 show the distribution of STIPs which detected by Willems method in both normal and abnormal crowd behavior in UMN dataset, where the center of a box represents the position of STIP, the size of the box stands for the scales, and different colors of boxes refer to different STIPs.

Figure 4. Abnormal behavior STIPs in UMN database.

III. THE DESCRIPTORS OF STIPS

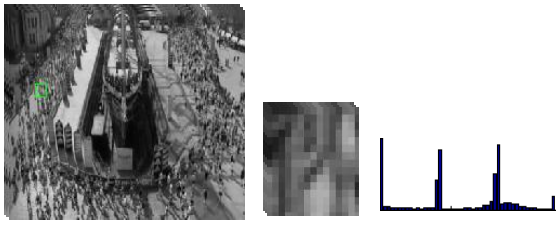
The combination of spatio-temporal information around a STIP is required to describe it. In this paper, we construct descriptor for each STIP by histogram of gradient, histogram of optical flow and spatial-temporal Haar feature. In these three methods we select a spatio-temporal cube C centered at a STIP, its size is defined as 6 times of spatial and temporal scales.

A. Histogram of gradient

Gradient describes grayscale change in the image, the gradient in the video space includes three directions of *x*, *y*, *t*. Then calculate the gradient (*Lx*,*Ly*,*Lt*) of each pixel in C with the following formula:

$$L_x = \partial_x(g * f), L_y = \partial_y(g * f), L_t = \partial_t(g * f). \quad (6)$$

In order to maintain scale-invariant features, we use normalized Gaussian model, that is $L_{x^m y^n t^k} = \sigma^{m+n}\tau^k(\partial_{x^m y^n t^k} g) * f$, where σ and τ are the spatial and temporal scales of 3-D Gaussian smoothing filter *g*, and *f* is input image, while *m*,*n*,*k* are the orders of differential. We calculate the sub-histogram *Bx*, *By*, *Bt* respectively for each gradient component, then combine them as a whole vector *D* = (*Bx*, *By*, *Bt*) to be the descriptor for a STIP. Figure 5 shows the location of one STIP, its neighborhood cube and the descriptor vector D.



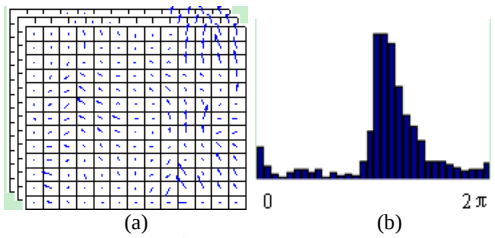
(a) (b) (c)

Figure 5. one STIP and its gradient histogram descriptor

B. Histogram of optical flow

Lucas-Kanade method is a sparse optical flow calculation algorithm which can reduce computation complexity through setting the computation window of optical flow according to the image size.

In order to build the descriptor of optical flow, we calculate the optical flow in C. We divide the optical flow orientation range of $0-2\pi$ into 32 sub-spaces to build a 32-dimensional histogram *B*. We vote each corresponding sub-space in *B* according to optical flow direction and amplitude information.

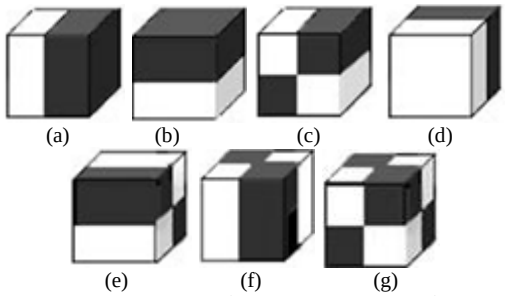


(a) (b)

Figure 6. Optical flow and its orientation histogram

Figure 6(a) is the optical flow which is calculated from a STIP cube shown in Figure 5(b), the length of the arrow represents the speed amplitude. Figure 6(b) shows the optical flow orientation histogram *B*. In order to adapt the changes of scene and human movements, we normalize the amplitude of each component in histogram B.

C. Spatial-temporal Haar feature



(a) (b) (c) (d)

(e) (f) (g)

Figure 7. Seven types of spatial-temporal Haar feature

Seven spatial-temporal Haar characteristics [10] are shown in Figure 7, where (a), (b) and (c) describe the behavior static features, while (d), (e), (f) and (g) describe the movement traits. Every point in C is described by a 7-D feature vector $(L_x, L_y, L_t, L_{yt}, L_{xt}, L_{xy}, L_{xyt})$. We sum the all of the feature vectors within this cube as the Harr descriptor of the STIP: $D = [\sum L_x, \sum L_y, \sum L_t, \sum L_{yt}, \sum L_{xt}, \sum L_{xy}, \sum L_{xyt}]$.

In order to adapt different scales and background light variations, this descriptor is then normalized with its maximum and minimum.

IV. NORMAL CROWD BEHAVIOR MODELING AND ABNORMAL BEHAVIOR DETECTION

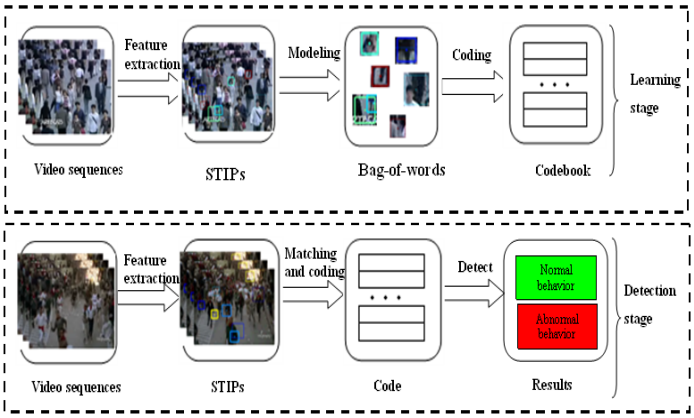


Figure 8 Integral structure of algorithm

The global algorithm is shown in Figure 8. We use the Bag-of-words strategy to describe the generality of normal behavior, and then determine the abnormal behavior by calculating the distance of descriptor vectors between the test video and the normal behavior.

A. *Normal behavior learning based on Bag-of-words*

Bag-of-words comes from the text retrieval field primarily, in which each document would be expressed as an N-dimensional vector, and then the massive documents can be classified through computer intelligent clustering. Human behaviors appear regular and calm in normal crowd fluids, although the distributions of STIPs are chaotic and disordered. So the constant attributes of normal crowd behaviors can be modeled with Bag-of-words strategy.

Each keyword in Bag-of-words is described by a Gaussian model, which represents the distribution of a certain amount STIPs feature. The keywords in the bag can be modeled by GMMs, we adopt EM algorithm to estimate its centers, weights, variances.

We build K keywords for all spatio-temporal features in training samples, namely the GMMs with *K*-cores, and then get a probability distribution vector *H*, its $n-th$ component is:

$$P_n(\Gamma \mid \Phi) = w_n G_n(\Gamma; \mu_n, \sigma_n) \qquad (7)$$

Where w_n is the weight of *n-th* keyword, Γ is the feature vector, G_n is the Gaussian probability dense function, *n = 1, 2,... , K.*

In order to describe the spatio-temporal motion characteristics in the training samples in detail, all the training samples are grouped into *M* sub-samples (the length is approximately 50 frames, about 2s long). We calculate probability distribution vectors of all STIPs in each sub-sample, and sum them up as one K-dimensional vector for this sub-sample. So all *M* training sub-samples compose a K-dimensional codebook as the normal behavior code.

B. *The procedures of abnormal behavior detection*

Firstly the test video is cut into multiple clips with a certain length (50 frames or so), in order to verify whether there is abnormal behavior happened in such period precisely. Then STIPs are extracted and their descriptors are established for each clip. We test the abnormal behavior occurrences from one clip to another continuously.

Secondly we use each interest point in a clip to match with the above training keywords that is to calculate its K matching probabilities then form a K-dimensional vector, we accumulate all these vectors to be a encoding vector, which represents the behavior movement distributions for this clip.

Finally, we measure the Euclidean distance between the encoding vector and *M* vectors in training samples codebook, if the minimal distance is still greater than a certain threshold, it will be judged as abnormal behavior.

V. EXPERIMENT RESULTS AND ANALYSIS

A. *Experiment conditions*

Experiment is simulated with Matlab7.0, 2.80GHz computer CPU and 4G bytes memory. The test databases are UMN and UCF dataset.

UMN is a low population density dataset, which contains normal behaviors, such as walking, standing, talking etc., while abnormal behaviors, for instance escaping and fearing, and its resolution is 483 × 361.

UCF is high population density dataset, which contains different urban environmental surveillance videos. Its normal behaviors include pedestrians walking, jogging etc.; abnormal behaviors are crowd escaping, protesting, assaulting etc.. And the resolution is 320×240. Both of datasets contain different scenes, illumination changes, scale variations and occlusion disturbances.

B. *Experiment results*

- Feature Extraction Tests

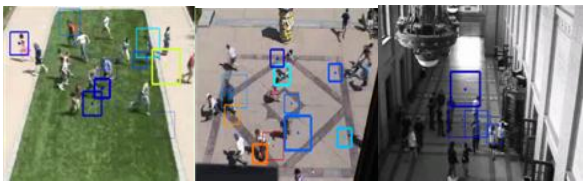


Figure 9 STIPs locations of UMN dataset

Figure10 STIPs Locations of UCF dataset

Figure 9 and Figure 10 show the STIPs extracted from two datasets, including the normal and abnormal behaviors.

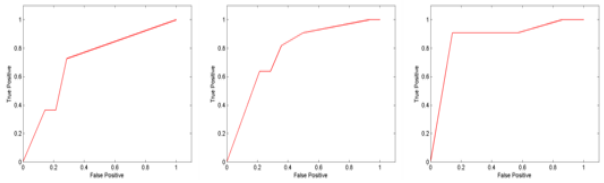
It can be observed that in the same scene the number of the STIPs extracted from normal behaviors is less than that from abnormal, for strenuous movements in normal behaviors are less than those in abnormal behaviors.

Additionally STIPs extracted from UCF dataset is much more than those of UMN dataset, because video scene is simpler in UMN dataset with lower population density.

- Abnormal behavior test

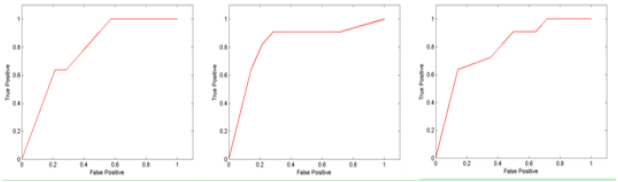
Crowd abnormal behavior detection is a two-typed classification problem, so we use the ROC curve to reflect the recognition accuracy, the positive alarm and false alarm rate are two important specifications.

We test algorithm performances by setting the different numbers of keywords in the two datasets. When K= 30 and K=50 respectively, STIPs are extracted with Willems method, and their three kinds of descriptors are built as gradient histograms, optical flow direction histograms, spatio-temporal Haar features; and then abnormal behavior recognition results are compared with ROC curves, as shown in Figure 11-14.



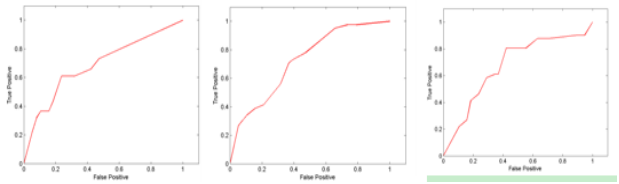
(a) (b) (c)

Figure 11 The ROCs for detection of three descriptors in UMN dataset when K=30



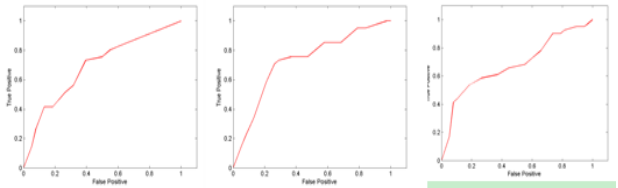
(a) (b) (c)

Figure 12 The ROCs for detection of three descriptors in UMN dataset when K=50



(a) (b) (c)

Figure 13 The ROCs for detection of three descriptors in UCF dataset when K=30



(a) (b) (c)

Figure 14 The ROCs for detection of three descriptors in UCF dataset when K=50

The area under the ROC curve reflects the algorithm quality, the larger the area, the better the classifier performance. Table 2 shows areas occupied by the ROC curve of the three descriptors when the keywords number *k* is 30 or 50 respectively.

TABLE 2 AREAS UNDER ROC OF DESCRIPTORS ON UMN AND UCF

	UMN (K=30)	**UMN** (K=50)	**UCF** (K=30)	**UCF** (K=50)
gradient histogram	0.707792	0.775974	0.686457	0.688383
optical flow histogram	0.769481	0.821429	0.721759	0.725610
Haar feature	0.870130	0.792208	0.672013	0.680359

In UMN dataset, the best performance is the spatio-temporal Haar features of K=30 according the ROC curve and its area under the curve reaches 0.870130. It shows that low-dimensional Bag-of-words can get better results in simple scene.

The global effectiveness of the algorithm in UMN is greatly different between K=50 and K=30, which shows that the number of keywords causes a greater impact on the algorithm in this scene.

In UCF dataset, the effects of three descriptors vary greatly compared to those in UMN dataset, and optical flow histogram performances best when K=50.

The descriptors have slight changes at K=30 and K=50, which shows that K=30 is proper for high population density scenes, and enlarging the volume of Bag-of-words to 50 cannot improve algorithm quality clearly.

VI. SUMMARY AND FUTURE WORK

STIPs approach is proposed for crowd abnormal behavior detection in this paper, we use three methods to construct their descriptors, which are gradient histogram, optical flow histogram, and spatio-temporal Haar features. Then GMM based on EM estimation is used to build Bag-of-words for behavior modeling. Algorithm achieves good results in the two datasets.

The limitation of this method is that it cannot recognize abnormal behavior types, just generally differs them from the normal. Therefore our future work is focus on semantic understanding of abnormal behaviors.

REFERENCES

[1] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pages 935–942.

[2] Shandong Wu, Brian E. Moore, Mubarak Shah. Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp.2054-2060.

[3] Vijay Mahadevan, Weixin Li, Viral Bhalodia, Nuno Vasconcelos. Anomaly Detectionin Crowded Scenes, IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, 2010.

[4] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. PAMI, 30(5):909–926, May 2008.

[5] Ramin Mehran1, Brian E. Moore2, and Mubarak Shah. A Streakline Representation of Flow in Crowded Scenes, ECCV 2010, Part III, LNCS 6313, pp. 439–452.

[6] Ronald Poppe. A survey on vision-based human action recognition[J]. Image and Vision Computing, 2010, 28:976–990.

[7] Ivan Laptev, Tony Lindeberg. Space-Time Interest Points. International Conference on Computer Vision(ICCV), Nice, France, 2003, pp.I:432-439.

[8] Dollar P,Rabaud V,Cotterll G,Belongie S. Behavior recognition via sparse spatio-temporal features. Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveil-lance and Performance Evaluation of Tracking and Surveillance, Beijing,China,2005:65-72.

[9] G. WIllems, T. Tuytelaars, and L. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. Proceedings of the European Conference on Computer Vision (ECCV), 2008, pages 650-663.

[10] Xinyi Cui, Yazhou Liu, Shiguang Shan, Xilin Chen, Wen Gao. 3D Haar-Like Features for Pedestrian Detection. IEEE International Conference on Multimedia and Expo, July 2007, pp.1263-1266.

Chuanxu Wang. Born in 1968, he received the Bachelor of Applied Electronic Technology and Master degree of Industrial Automation in China University of Petroleum; received Ph.D. in Ocean University of China. He has a academic access in Australia wollongong university.

He is an associate professor of Computer Vision in Qingdao University of Science & Technology. Since 2000, he has involved in the completion of seven projects with national, provincial or departmental level. And now he has three projects for researching. He has published over 30 academic papers including 18 for EI or ISTP retrieval.

Chenchen Dong. Born in 1988, she received the Bachelor of Information Technology in Qingdao University of Science & Technology; and now studying in Qingdao University of Science & Technology as a graduate of Computer Vision.