Spatio-Temporal Descriptor for Abnormal Human Activity Detection

Fam Boon Lung, and Mohamed Hisham Jaward
Electrical and Computer System Engineering, School of Engineering
Monash University
Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya Selangor, Malaysia
{fam.boon.lung, mohamed.hisham}@monash.edu

Jussi Parkkinen
University of Eastern Finland
P.O. Box 111, Fl-80101 Joensuu, Finland
jussi.parkkinen@uef.fi

Abstract

There has been an increased interest in the field of abnormal human activity detection to find a good descriptor with a lower computational cost. In this paper, we propose such a Spatio-Temporal Descriptor (STD) based on spatio-temporal features of an image sequence. Proposed descriptor is based on a texture map, known as Spatio-Temporal Texture Map (STTM) and is based on 3-dimensional Harris function. It is able to capture subtle variations in the spatio-temporal domain. Performance of the STD was illustrated with a mixture of Gaussian Hidden Markov Model (HMM) to show its potential for more complex modeling. Proposed algorithm was evaluated with UCSD dataset that has abnormal events that are not staged such as biker, skater, cart activities etc. Compared to other state of the art descriptors that are used with the same dataset, our proposed descriptor shows competitive performance with a lower computational cost.

1 Introduction

Recently, there has been an increased interest on research on human activity recognition in video surveillance due to need for better security and intelligent health monitoring [1, 2]. Abnormal human activity is defined as the human action which stands out and requires more attention and this depends on the context of scene considered. Abnormal human activity often has the properties of irregular pattern [1] and being an event that occur at a low frequency relative to a normal event [3]. Abnormal human activity detection is thus accomplished by finding the outliers of the normal event due to low occurrence of such anomaly.

Modeling of abnormal human activity detection highly depends on the performance of the descriptor. In order to clearly classify between normal and abnormal events, a good descriptor that is distinguishable among each activity and easy to model is highly sought after [1, 2]. Traditional optical flow [4] that finds the pixel changes between 2 consecutive frames, was widely used in early papers [1, 2]. The approach proposed in optical flow constrains the problem of activity detection to be between two frames with consistent illumination constraint. This result in a lot of deficiencies in the descriptor especially when the anomaly can only be observed in a longer temporal space. The problem of illuminance change especially in outdoor scene has also been an another problem with such an approach. In order to overcome these problems, different descriptors such as mixture dynamic texture [3, 5], sparse reconstruction cost [6] and frequency transform [7] have been proposed.

Among different sophisticated descriptors used in these approaches, recently proposed mixture of dynamic texture (MDT) [3, 5] has shown promising results. MDT is an image compression technique that model the probability of each pixel point using a linear dynamic system [8]. This work is further improved to include spatial constrains of the scene in the application of abnormal human activity recognition [3, 5]. The extensive modeling of each pixel points in spatial and temporal domain has lead to a good performance of the descriptor but at a high computational cost.

The application of extending detection of spatial features to spatio-temporal domain has been a new trend in activity recognition [9]. Among many spatio-temporal detector used in activity recognition, spatial temporal interest point (STIP) detector has shown promising results in many cases [10]. STIP concept is based on Harris function [11, 12] to consider the variation along the spatial and temporal domain to find geometric points, also known as landmark points, which are interest points with relatively high variances. By tracking landmark points in STIP detectors, shape and motion characteristics of the video can be recorded. STIP has been tested on action class of varying difficulty [13] and proven to be successful in tracking for many cases. STIP has an strong advantages in computational cost as the calculation of the singular value decomposition is replaced by finding the determinant and trace. The relatively faster computation speed of STIP makes the algorithm suitable to be explored for abnormal human activity detection to achieve better performance with a lower computational cost.

Motivated by the work on STIP, we propose the use of Spatial Temporal Texture Map (STTM) and its descriptor, Spatio-Temporal Descriptor (STD) for abnormal human activity detection. Instead of detecting multiple interest points in STIP, we consider spatio-temporal variations of texture in videos sequences for the detection of abnormal human activity. In this paper, the Harris corner detector as used in STIP is used to form the Spatial Temporal Texture Map (STTM) representation of the videos. By taking the histogram of the STTM, our novel STD is generated to represent the scene. The STD is generated for the all the video including the training video that contain only normal human activity and the testing video where anomaly event occurs. Both cases are respectively modeled and tested using a mixture of Gaussian hidden Markov model.

The organization of the paper is as follow: In section 2, we present the details of STD formulation. Multiple experiments are conducted to validate the performance of STTM and the discussion on the performance of STD are included in section 3. Finally, future work are concluded in section 4.

2 Motion Representation and Descriptor Generation

2.1 STTM Representation

In spatio-temporal domain, a given video $V(x,y,t)$ can be modeled by $L : \Re^2 \times \Re \times \Re^2_+ \mapsto \Re$ using linear scale-space representation [14],

$$L\left(x,y,t|\sigma_i^2,\tau_i^2\right) = g\left(x,y,t|\sigma_i^2,\tau_i^2\right) * V\left(x,y,t\right), \quad (1)$$

where $*$ is the convolution and spatio-temporal Gaussian kernel, $g\left(x,y,t|\sigma_i^2,\tau_i^2\right)$ is defined as follows with σ_i and τ_i which represent spatial variance and temporal variance respectively,

$$g\left(x,y,t|\sigma_i^2,\tau_i^2\right) = \frac{1}{\sqrt{\left(2\pi\right)^3 \sigma_i^4 \tau_i^2}} \exp\left(-\frac{x^2+y^2}{2\sigma_i^2} - \frac{t^2}{2\tau_i^2}\right) \quad (2)$$

The spatio-temporal variation of the video can be obtained from Harris function [11, 12]. Given any spatial variance, σ_i^2 and temporal variance, τ_i^2, the variation of the video along spatial domain (x,y) and temporal domain (t) can be obtained as

$$\begin{aligned} \mu\left(x,y,t|\sigma_i^2,\tau_i^2\right) &= g\left(x,y,t|\sigma_i^2,\tau_i^2\right) * \\ &\quad \left(\triangledown L\left(x,y,t|\sigma_i^2,\tau_i^2\right)\left(\triangledown L(x,y,t|\sigma_i^2,\tau_i^2\right)^T\right) \\ &= g(x,y,t|\sigma_i^2,\tau_i^2) * \\ &\quad \begin{bmatrix} (L_x)^2 & L_x\ L_y & L_x\ L_t \\ L_x\ L_y & (L_y)^2 & L_y\ L_t \\ L_x\ L_t & L_y\ L_t & (L_t)^2 \end{bmatrix} \end{aligned} \quad (3)$$

where L_x, L_y, and L_t are the partial derivatives of scale-space representation. This represent the smoothed version of spatio-temporal second moment matrix.

The main properties of μ matrix can be easily found by inspecting the significant eigenvalues λ_1, λ_2 and λ_3 of μ. The product and summation of the significant eigenvalues are used to construct extended Harris function for spatio-temporal domain. The Harris function could be expressed using the determinant and trace of μ as follows

$$\begin{aligned} H &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3 \\ &= \det(\mu) - k\ \text{trace}^3(\mu) \end{aligned} \quad (4)$$

where k is a constant. The Harris function is applied to pixel points to generate a 3 dimensional Harris-matrix. Every temporal (t) layer of the Harris-matrix forms a STTM representation of the image. This STTM as shown in figure 1 highlights the texture variation of the image pixels in the spatio-temporal domain.

STTM is normalized with a global maxima value of the training video to limit the dynamic range of data values. In STTM not only considers spatio-temporal interest points of high variance but all points are taken into account. STTM is capable of detecting subtle changes in the frame and this could be easily observed in the figure 1. High intensity values in the texture map denote the anomaly in the scene and they are shown in a rectangular box in figure 1.

2.2 STD Formulation

STD is obtained from STTM and is explained below. In order to consider location specific anomalies, the STTM, as illustrated in figure 1, is divided into 2x2 blocks. Then a histogram with non-uniform quantization for bin values as shown in figure 2 is generated for each of 4 blocks of STTM to record the variation at different times. After histograms with non-uniform quantization are obtained for each block, the bins of lower value range (the first few bins) of the histograms are removed as they represent the background motion. Finally, the remaining bins of the 4 histograms are concatenated to form the descriptor. The purpose of using non-uniform quantization is to give some emphasis on motion with slight variations. Such emphasis is achieved using the A-law compression [15]. The compression parameter, A value of the A-law is tuned to attain best performance.

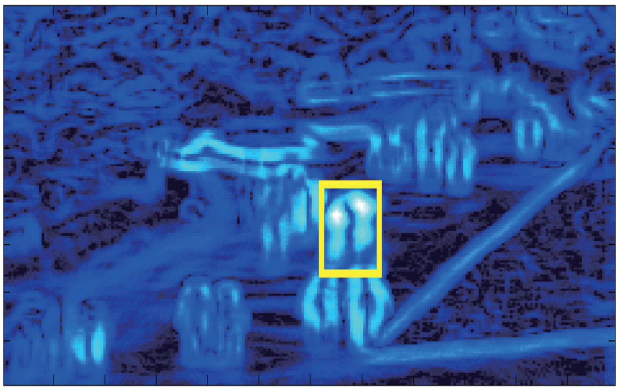

Figure 1. STTM representation of image

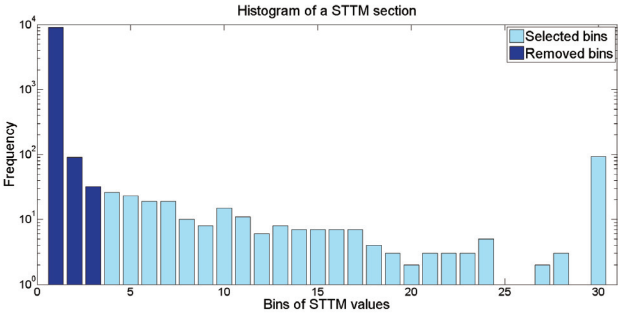

Figure 2. Histograms of a STTM section

3 Simulational Experiments

To test the performance of the descriptor, the UCSD dataset [3] is selected. In this dataset, anomalies occur naturally and are not staged. UCSD dataset contains 2 sets of pedestrian videos. The first set of the videos from the front view consists of 34 training and 36 test videos while the second set of videos consists of side view where 16 training and 12 test videos are available. The training videos only contain footage of normal event which is pedestrian walking. Anomalies in the test videos include bikers, skaters, small carts, wheel chairs, etc. For all videos, a STTM is computed and STD is generated to represent every frame.

To optimize the result, two experiments are conducted where parameters of Gaussian Kernel and STD are tuned. A simple mixture of Gaussian Hidden Markov Model (HMM) is used in both experiments to classify the scenes. Mixture of Gaussian HMM is used widely in abnormal human activity detection [1] and has given good results in many cases.

The proposed STD can be easily trained using the mixture of Gaussian HMM to capture the abnormalities of the scene that are unconstrained during training. Mixture of Gaussian HMM is mainly defined by the number of Gaussian mixture (M) and number of states (Q). In both two experiments, M is set to be 1 and Q is set to 2 for simplicity. Note that the overall distribution of STD for each frame must be close to Gaussian [16] in order for the mixture of Gaussian HMM to be effective.

In our experiments, the log likelihood obtained for the testing videos sequence using the Gaussian mixture HMM is compared to the ground truth provided from the UCSD dataset. Since the log likelihood is a scalar, a threshold value need to be chosen to find the optimum detection. By varying the threshold, a Receiver Operating Characteristic (ROC) curve can be obtained. The optimum point for maximum detection for such condition will be the equal error rate (EER) point where the ROC curve intercepts with the EER line as shown in figure 3 and 4.

3.1 STD Parameter Validation

In our experiment 1, the parameters of the Gaussian kernel that are used to generate STTM are tuned. The spatial variance, σ_i^2 and temporal variance, τ_i^2 of the Gaussian kernel parameters are varied to test for different conditions. After the STTM is obtained, a single histogram is computed and upper portion of the histogram are taken as the STD. STD for each frame are then trained using a mixture of Gaussian HMM. Similarly, STD for the test videos are produced and tested using the mixture of Gaussian HMM. The log likelihood for each frame is compared with the ground truth to obtain the ROC curve and EER can be found from the ROC curve. By comparing the EER obtained for different STTM parameters, the best parameters are found. The best performance with the lowest EER for pedestrian 1 and 2 dataset is respectively 32.93% and 39.5% when the STTM is generated using parameters, σ_i^2=5 and τ_i^2=1.

In experiment 2, after the STTM is computed as explained above, other parameters of STD are adjusted. The A value of A-law compression are varied to optimize the descriptor performance. After the optimal A value is found, the total number of bins in the histograms is varied. The total number of removed bins at the lower end is changed as well to create different STD. The final set of the bins is taken to be the optimal STD. STD is then modeled using a mixture of Gaussian HMM where the log likelihood obtained is compared to the ground truth to get the ROC curve.

As shown in figure 3 and 4, A value used by the A-law compressor effects the ROC curve and the EER found. In the best case, the selected A value of 717 used in A-law compressor results in achieving optimal EER of 32.42 % in pedestrian 1 dataset while optimal EER of 28.45 % is obtained for pedestrian 2 dataset using a A value of 24. The effect of using A-law compressor is more significant in pedestrian 2 dataset as shown in figure 4 where the ROC curves deviate a lot for the best and worst case values of A.

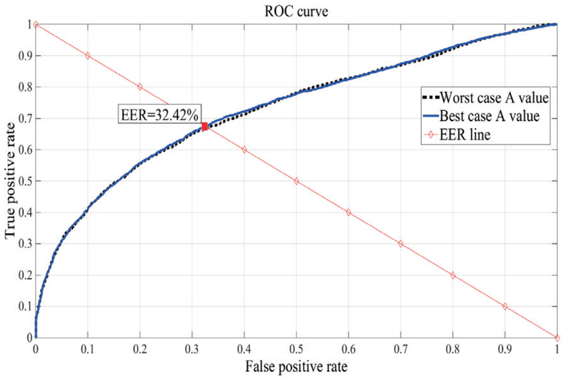

Figure 3. Pedestrian 1 ROC curve for different A value

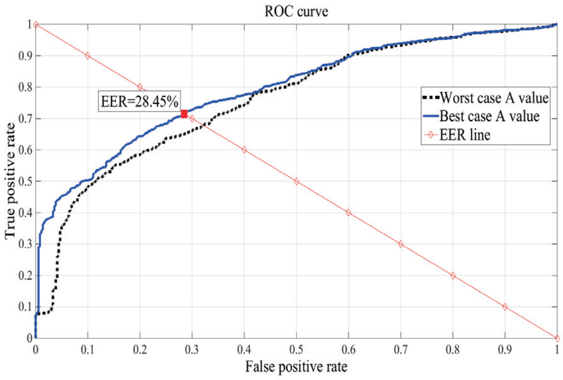

Figure 4. Pedestrian 2 ROC curve for different A values

3.2 Results and evaluation

An important criterion in evaluating the performance of the algorithm is to compare the EER value. Table 1 shows the EERs of different descriptors. The mixture of dynamic texture (MDT) for temporal case [5] has the best performance compared to all other methods while our proposed method comes second. This is followed by other methods including Mixture of Principle Component Analysis (MPPCA) [17], Social Force (Force Flow) [18], MDT for spatial case [5] and Local Motion Histogram (LMH) [19]. An initial evaluation has been performed to compare the computational speed of MDT and our proposed STD. The evaluation results are shown in table 2. A dedicated computer with Intel Core 2 Duo CPU T5800 @2 GHz, 4GB RAM and 64 bit Linux system is used in this evaluation. When running in MATLAB environment, MDT algorithm takes 114.343 seconds for every test frame while our proposed algorithm only requires 5.145 seconds. MDT's slow computation speed rise from the application of multiple state space model in the formulation of the descriptor while our proposed work is faster as it works by extracting only minimal amount of information in spatial temporal domain.

Table 1. EER results

Descriptor	Ped1	Ped2	Average
MDT-temp.	22.9%	27.9%	25.4%
Proposed STD	32.4%	28.5%	30.5%
MPPCA	35.6%	35.8%	35.7%
Force flow	36.5%	35.0%	35.8%
MDT-spat.	43.8%	28.7%	36.3%
LMH	38.9%	45.8%	42.4%

Table 2. Computational cost

Descriptor	Time taken per frame
MDT-temp	114.343s
Proposed STD	5.145s

4 Conclusion

In this paper, we proposed a novel abnormal human activity descriptor, STD, that can describe the variation of texture in spatial and temporal domain. The experiments illustrate the performance of STD when modeled with mixture of Gaussian HMM. Compared with other algorithm using the UCSD dataset, we showed that our descriptor has relatively low EER and through the use of trace and determinant in computing STTM, we achieved lower computational cost. As our future work, we plan to evaluate the computational speed of other descriptors and experiment with other datasets.

5 Acknowledgments

This work was supported by Malaysia Ministry of Higher Education Fundamental Research Grant Scheme, ERGS/1/2013/TK02/MUSM/03/1.

References

[1] O. Popoola and K. Wang, "Video-based abnormal human behavior recognition;a review," *in IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 6, pp. 865–878, Nov 2012.

[2] M. Thida, Y. Yong, P. Climent-Prez, H.-l. Eng, and P. Remagnino, "A literature review on video analytics of crowded scenes," in *Intelligent Multimedia Surveillance*, P. K. Atrey, M. S. Kankanhalli, and A. Cavallaro, Eds. Springer, 2013, pp. 17–36.

[3] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 1975–1981.

[4] M. J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *in Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75 – 104, 1996.

[5] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, Jan 2014.

[6] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 3449–3456.

[7] Y. Wang, K. Huang, and T. Tan, "Abnormal activity recognition in office based on r transform," in *IEEE International Conference on Image Processing (ICIP)*, vol. 1, Sept 2007, pp. 341–344.

[8] A. Chan and N. Vasconcelos, "Mixtures of dynamic textures," in *Tenth IEEE International Conference on Computer Vision*, vol. 1, Oct 2005, pp. 641–647.

[9] J. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *in Pattern Recognition Letters*, vol. 48, pp. 70 – 80, 2014.

[10] I. Laptev and T. Lindeberg, "Space-time interest points," in *Ninth IEEE International Conference on Computer Vision,*, Oct 2003, pp. 432–439 vol.1.

[11] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceeding of the 4th Alvey Vision Conference*, 1988, pp. 23.1–23.6.

[12] W. Förstner and E. Gülch, "A fast operator for detection and precise location of distinct points, corners and centres of circular features," *in Proceeding of the ISPRS intercommission conference on fast processing of photogrammetric data*, pp. 281–305, 1987.

[13] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proceeding of the British Machine Vision Conference (BMVC)*, 2009, pp. 124.1–124.11.

[14] A. P. Witkin, "Scale-space filtering: A new approach to multi-scale description," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 9, Mar 1984, pp. 150–153.

[15] A. Gersho, "Principles of quantization," *in IEEE Transactions on Circuits and Systems*, vol. 25, no. 7, pp. 427–436, Jul 1978.

[16] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *in IEEE Proceedings*, vol. 77, no. 2, pp. 257–286, Feb 1989.

[17] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009, pp. 2921–2928.

[18] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009, pp. 935–942.

[19] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, March 2008.