Abnormal Event Detection Using HOSF

Shwu-Huey Yen, Chun-Hui Wang
Department of Computer Science and Information Engineering
Tamkang University, New Taipei City, Taiwan 25137
Email: 105390@mail.tku.edu.tw, 802410018@s02.tku.edu.tw

***Abstract*—In this paper a simple and effective crowd behavior normality method is proposed. We use the histogram of oriented social force (HOSF) as the feature vector to encode the observed events of a surveillance video. A dictionary of codewords is trained to include typical HOSFs. To detect whether an event is normal is accomplished by comparing how similar to the closest codeword via z-value. The proposed method includes the following characteristic: (1) the training is automatic without human labeling; (2) instead of object tracking, the method integrates particles and social force as feature descriptors; (3) z-score is used in measuring the normality of events. The method is testified by the UMN dataset with promising results.**

***Keywords*—normality; crowd; social force (SF); histogram of oriented social force (HOSF); z-value**

I. Introduction

Today, with technology advances and life-quality enhancements, there have been more and more surveillance cameras installed in our living environment for security reasons. The detection of abnormal behaviors is an important research issue in surveillance system and computer vision [1-3], as well as in other research region such as applications of elderly care [4]. One of the most challenging tasks is how to automatic detect and analyze individual activities in video, especially in crowded scenes.

Some public spaces such as train stations, hospitals, theaters, schools or others, where the surrounding camera monitors are mainly monitoring the scene for suspicious behavior and are logging events. By combining those video data and computer vision techniques, it could bring effective performance for surveillance system to reduce manpower cost and to avoid human negligence due to fatigue; these smart surveillance systems could also capture the alert which was filtered through amount data, so that it helps monitoring staffs make an appropriate response and disposal in short time to achieve the aim of safety precaution and crime monitoring.

Conventional methods for behavior analysis or identification are mostly based on the tracking of the interested objects. However, these methods are too dependent on individual characteristics such as shape, color and movement trajectory. Considering the environmental instabilities, unavoidable occlusions, or cluttered backgrounds, the reliable tracking would be a major challenge. Therefore, many studies have focused on the approaches which inherited from macroscopic and microscopic models such as optical flow and social force [5] to process the event detection via statistical models.

In this work, we propose an efficient abnormal event detection system in crowded scene based on social force model which considers human interactive relationships and concepts of social influence. The proposed method includes the following contributions: (1) the training is automatic instead of human labeling; (2) instead of object tracking, the method integrates particles and social force as feature descriptors which well adapted in both crowded or few people scenes; (3) a simple z-score is used in measuring the normality of events. Due to computation simplicity, the normality detection can be real-time implemented once the training is finished.

II. Related Work

In recent years, several researches have been proposed in the literature on behavior analysis of video sequence and unusual event detection in computer vision. As technology advances, many approaches of detection in surveillance system have been proposed [1-3]; in particular, analyzing trajectory of individual is the most extensive and intuitive, such well-developed methods as meanshift [6], HOG [7] or motion patterns and appearance [8]. However, in crowd videos, the reliability of detection and tracking still is a major challenge [9].

Since the difficulties of human tracking in crowded scenes, some researchers use motion or motion-related feature descriptors [10-12] to model crowd behaviors. Adam et al. [10] used local histogram of optical flow as low-level feature. Kratz et al. [11] proposed to use spatio-temporal gradients in a Gaussian mode. Mahadevan et al. [12] proposed a dynamic texture model which combines the appearance and dynamics in crowded scenes. Dynamic texture is very effective in expressing the dynamic background such as ocean wave or crowded scene, however, it pays the price in training cost.

The crowd behaviors are prone to be blind and infectious. Consequently, it is essential to model crowd interaction patterns for understanding or predicting their behaviors. In general, the crowd behaviors studies could be classified as follow: (1) macroscopic models, which analyze the density and velocity of whole crowd; (2) microscopic models, which focus on individual behavior motivation; (3) hybrid models, which consider the overall behavior as well as individual by mixing both methods above. These models can be applied to crowd simulation, crowd management, disaster management, exit design of buildings or others [9]. Wang et al. [13] gave a complete evaluation of methods on behavior recognition.

Mehran et al. [5] adopted an optical-flow-based social force model and particle advection scheme to analyze the interaction forces between individuals and surrounding environments.

After that, some studies based on social force model have been proposed in succession to detect crowd behaviors as well [14-16]. Inspired by [5] and the difficulties of tracking in crowds, we adopt social force model and use histogram of oriented social force (HOSF) as the feature descriptor to accomplish abnormal detection in crowded scene.

III. Proposed Method

The proposed method includes training and testing phases. For a given sequence, the beginning T frames are used to establish the dictionary D. The rest of frames, so called events, are for testing. An event (frame) is normal if it has occurred several times in the previously seen data (learned dictionary D). The overall proposed system is depicted in Fig. 1 and the details are explained in the following.

A. Active Particles (AP)

To begin, pyramid Lucas-Kanade [17] method is used to calculate optical flow of pixels over the image sequence and a corresponding optical flow sequence is acquired. Similar to [5], we analyze the crowd videos via particles movement. A 3D grid of particles (with a spacing of 3 pixels) is placed over the optical flow sequence along temporal and spatial axes; therefore, each particle is made of 3×3×3 pixels and its optical flow is defined as the median of these 27 optical flows to reduce noises. We define a particle as "active" if its motion magnitude is more than a threshold Th_{motion}. The active particles (AP) will be the feature points, and the inactive ones will be treated as background and ignored. Herein, we will use "particle frame" to indicate a frame consists of particles instead of pixels, which is 1/3 of original width/height/length of a given video.

B. Feature Descriptor

Once feature points are located, we define the HOSF feature descriptor.

1) Estimation of Social Force: The social force of an AP P_i is the sum of interaction forces among P_i and its neighboring APs. Thus, we modify the equation from [16] as in (1).

$$F_i^{soc} = \sum_{P_j \in N(P_i)} f_{i,j}^{soc} \qquad (1)$$

with

$$f_{i,j}^{soc} = m_j exp\left[-\left(\frac{r_{i,j}-d_{i,j}}{b}\right)\right] v_{i,j}\ , \qquad (2)$$

where m_j is the mass of particle j (let $m_j = 1$), $r_{i,j}$ is the distance between particles; $d_{i,j}$ is sum of radius of two particles; $v_{i,j}(=v_j - v_i)$ is the direction vector of influence force which is obtained by the difference of optical flows of APs P_i and P_j. N is the set of neighboring APs of P_i. The neighboring area, $N(P_i)$, is set as $(2k+1) \times (2l+1)$ centered at P_i and $b = \sqrt{k^2+l^2}$ is the largest distance. Equation (2) represents the influence force is proportional to distance between two particles, mass and velocity of two particles.

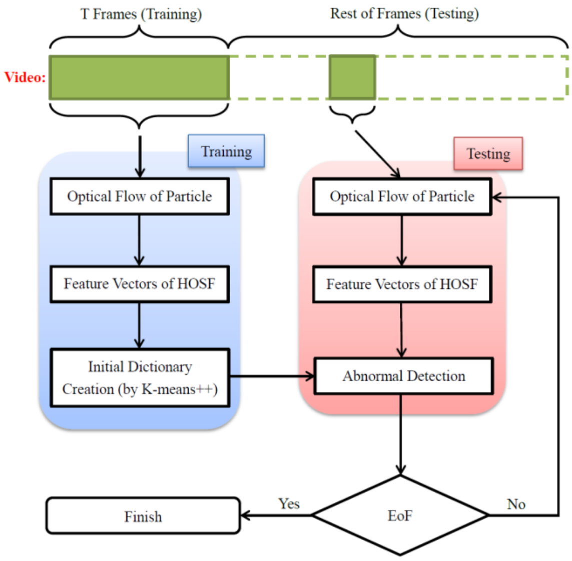


Fig. 1. Flow chart of the proposed method.

2) Histogram of Oriented Social Force (HOSF): After the calculation of social forces for all APs in a particle frame, we use cuboids to encode histogram of oriented social force (HOSF). For each AP P_i in the particle images, a cuboid of size $2n \times 2n \times 2m$ centered on P_i is divided into 8 sub-cuboids each size is $n \times n \times m$ as shown in Fig. 2. For each sub-cuboid, an 8-bin (0, ±π/4, ±π/2, ±3π/4,π) histogram of the social force magnitude is built. Finally, for each AP P_i, a 64-dimensional HOSF feature vector of P_i is obtained by concatenating eight social force histograms with norm normalized to be 1. Similar to the histogram of oriented gradient (HOG) [7], HOSF not only remains the relation of spatial and temporal, but tolerates some drifting via sub-cuboids.

C. Dictionary Creation

After all HOSFs of APs are collected from the beginning T/3 particle frames, we use K-means++ [18] to cluster feature vectors into K classes (K=100 in the experiment). For each class k, the codeword c_k is defined as the mean HOSF vectors. The dictionary D is made of the codeword c_k of K classes and it would be the basis to identify whether a given frame is normal. Besides the codewords, the following information of D are also kept as depicted in TABLE I.

D. Event Detection

In testing phase, we only perform the normality test for a particle frame if it has enough number of APs (at least 3) and claim normal for those do not. An event E_t is defined as the set of HOSF vectors with $E_t = \{f_t(P_1), \ldots, f_t(P_{M(t)})\}$, where $f_t(P_i)$ is the HOSF vector of AP P_i in the particle frame t. In determining normality, we compare each feature vector $f_t(P_i)$ with the most similar codeword in D and measure the distance via z-score as in (3).

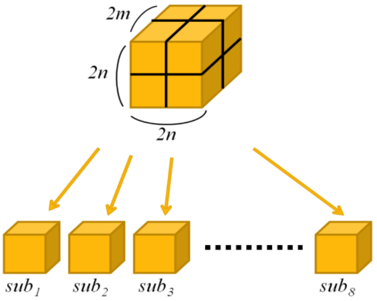


Fig. 2. A sample for a split of the cuboid.

1) Z-Score Value: For a HOSF $f_t(P_i)$ of the AP P_i in the particle frame t, assume c_k is the closed codeword for $f_t(P_i)$, the z-score value is estimated as

$$z(P_i) = \left| \frac{\|f_t(P_i)-c_k\|-\mu_k}{\sigma_k} \right|. \qquad (3)$$

Equation (3) is to measure whether it is common that the distance between $f_t(P_i)$ and c_k comparing to the distances between the rest of elements in the same cluster and c_k. A small z-value implies $f_t(P_i)$ is a typical normal pattern; otherwise, it may be an unusual pattern or a noise. E_t is claimed as an abnormal event if (4) is satisfied:

$$\frac{1}{M(t)}\left(\sum_{1\le i\le M(t)} \hat{z}(P_i)\right) \ge Th_N \qquad (4)$$

with

$$\hat{z}(P_i) = \begin{cases} z(P_i), & z(P_i) < Th_N \\ \underset{P_j\in N(P_i)\cup\{P_i\}}{median}\left(z(P_j)\right), & z(P_i) \ge Th_N \end{cases}, \qquad (5)$$

where $M(t)$ is the number of APs in frame t, N is the set APs within the 8-neighbor of P_i and Th_N (1.3 is used in the experiment) is an user-defined normality threshold. Since an unusual event is exhibited as a blob of unusual APs, we use (5) to reduce the noise. Finally, we consider the average z-values of an event to determine the normality of a particle frame t according to (4).

2) Temporal Smoothing: To further reduce noises, we apply temporal smoothing on the detection results. Due to fact that the duration of abnormal events would be several seconds in the sequence, a mode operation of (2*r*+1) detected particle frames is applied on the center particle frame. The "mode" smoothing is triggered on the particle frame *t* whenever there is a normality transition from frames (*t*-1) to *t*. The operation takes the majority detection results in frames (*t*-*r*), ..., *t*, ..., (*t*+*r*) and assign to the frame *t*. (a) shows the detection results up to (*t*+2) where green/red indicating detected as normal/abnormal and gray indicating not yet processed. As observed, a normality transition occurs on frame *t*. Thus, a smoothing operation is triggered and the final detected result on frame *t* is modified to be "normal" since this is the majority normality results.

TABLE I. THE INFORMATION IN THE DICTIONARY D

notation	description
c_k	The codewordof cluster k
μ_k, σ_k	mean and standard deviation of the distances fromc_k to each data in cluster k
n_k	size of cluster k
N	total number of data in D, i.e., $N = n_1 + \ldots + n_k + \ldots + n_K$

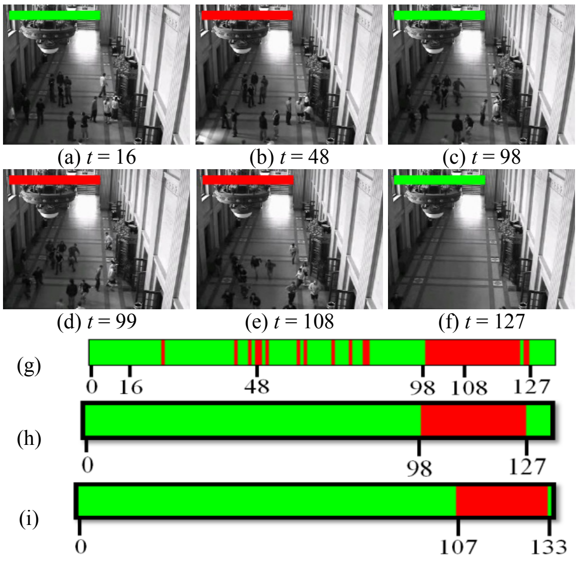
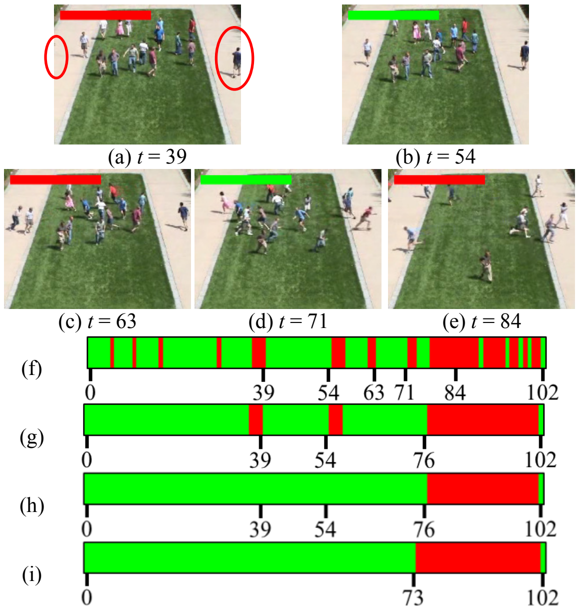

(a) *t* = 16 (b) *t* = 48 (c) *t* = 98
(d) *t* = 99 (e) *t* = 108 (f) *t* = 127

Fig. 3. Test result of test_1; (a)-(f) are the frames of the detection results at time *t*, (g) is the complete detection result (without temporal smoothing), (h) is after temporal smoothing is applied, and (i) is the ground truth provided by UMN dataset and their normality transitions are about 1/3 second delay than ours..

IV. EXPERIMENTAL RESULTS

To testify the proposed method, two video cliques, as in Fig. 3 & 4, of public available UMN dataset [19] are tested. In the training phase, for each scenario, frames of the beginning 5~10 seconds are used for training and the rest are for testing where the frame rate is 30 per second.

Fig. 3 (a)-(f) exhibit some of the detection result of test_1 where green/red indicating normal/abnormal result. In 3(b), the detection result was mistakenly to be abnormal due to a door open suddenly (on the lower left corner). As observed, people are about to run on *t* = 98 and start running on next frame, and people are evacuated (the hallway is empty) on *t* = 127. Our method correctly detected on these frames. 3(g) shows the complete test result without temporal smoothing. There are some false alarms around frame 40s to 90s. However, after temporal smoothing with *r* = 2, as shown in (h), the final detected result are quite accurate. Comparing to the ground truth of UMN shown in 3(i), frames 107 and 133 were the first and the last frame labeled as abnormal which is 9 and 6 frames apart from ours result. But, as observed from the video, people start running on frames 98 and the hallway remains empty from frame 127. Besides, 9 frames and 6 frames apart mean less one third of a second. We think that less than a second difference should be accepted as correct detection.



(a) $t = 39$ (b) $t = 54$

(c) $t = 63$ (d) $t = 71$ (e) $t = 84$

Fig. 4. Test result of test_2; (a)-(e) are the frames of the detection results at time t, (f) is the complete detection result (without temporal smoothing), (g) & (h) are after temporal smoothing with r=2 and r=3, respectively; (i) is the ground truth provided by UMN dataset.

Fig. 4 presents the test result of test_2. Our method mistakes (a) as abnormal due to the sudden appearing and disappearing on the areas red circled. As shown in (c), people start running with panic at $t = 63$. (g) & (h) are the results after temporal smoothing with $r = 2$ & 3, respectively.

V. Conclusion

A normality detection method was proposed for crowd scenes. We first define active particle (AP) using optical flow. From an AP, the HOSF feature descriptor affected by its 3D neighbors is obtained. A dictionary D is constructed by mean vectors, so called codewords, from clustered HOSF in training phase. If the HOSF of an AP has occurred quite often, then distance between its HOSF and the most similar codeword should be small. Based on this, the z-value on distance between the HOSF and the nearest codeword is used to determine the normality.

In our experiments, most of results are satisfactory and it is time efficiency. However, parameters Th_N in normality threshold in (4) and r in temporal smoothing are data dependent. In the future, we would study more on adaptive parameters.

Acknowledgement

The authors would like to express their gratitude for the grant support from the National Science Council (NSC 102-2221-E-032-061) of Taiwan, R.O.C.

References

[1] T.B. Moeslund, A. Hilton and V. Kru¨ger, "A Survey of Advances in Vision-Based Human Motion Capture and Analysis," Computer Vision and Image Understanding, vol. 104, no. 2, pp. 90–126, 2006.

[2] P.K. Turaga, R. Chellappa, V.S. Subrahmanian and O. Udrea, "Machine Recognition of Human Activities: A Survey," IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 11, pp. 1473–1488, 2008.

[3] O.P. Popoola and K. Wang, "Video-Based Abnormal Human Behavior Recognition—A Review," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 42, pp. 6, 865–787, 2012.

[4] J. Yin, Q. Yang and J.J. Pan, "Sensor-Based Abnormal Human-Activity Detection," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 8, pp. 1082–1090, 2008.

[5] R. Mehran, A. Oyama and M. Shah, "Abnormal Crowd Behavior Detection using Social Force Model," Computer Vision and Pattern Recognition, pp. 935–942, 2009.

[6] G.R. Bradski, "Computer Vision Face Tracking for Use in a Perceptual User I nterface," IEEE Workshop on Applications of Computer Vision, pp. 214–219, 1998.

[7] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," Computer Vision and Pattern Recognition, vol. 1, pp. 886–893, 2005.

[8] P. Viola, M.J. Jones and D. Snow, "Detecting Pedestrians using Patterns of Motion and Appearance," International Journal of Computer Vision, vol. 63, no. 2, pp. 153–161, 2005.

[9] S. Saxena, F. Brémond, M. Thonnat and R. Ma, "Crowd Behavior Recognition for Video Surveillance," International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 970–891, 2008.

[10] A. Adam, E. Rivlin, I. Shimshoni and D. Reinitz, "Robust Real Time Unusual Event Detection using Multiple Fixed-Location Monitors," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 3, pp. 555–560, 2008.

[11] L. Kratz and K. Nishino, "Anomaly Detection in Extremely Crowded Scenes using Spatio-Temporal Motion Pattern Models," Computer Vision and Pattern Recognition, pp. 1446–1453, 2009.

[12] V. Mahadevan, W. Li, V. Bhalodia and N. Vasconcelos, "Anomaly Detection in Crowded Scenes," Computer Vision and Pattern Recognition, pp. 1975–1981, 2010.

[13] H. Wang, M.M. Ullah, A. Kläser, I. Laptev and C, Schmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition", British Machine Vision Conference, pp. 1–11, 2009.

[14] Y. Zhang, L. Qin, H. Yao and Q. Huang, "Abnormal Crowd Behavior Detection Based on Social Attribute-Aware Force Model," IEEE International Conference on Image Processing, 2012.

[15] R. Raghavendra, A.D. Bue, M. Cristani and V. Murino, "Optimizing Interaction Force for Global Anomaly Detection in Crowded Scenes," IEEE Workshop on Modeling, Simulation and Visual Analysis of Large Crowds, pp. 136–143, 2011.

[16] M. Luber, J.A. Stork, G.D. Tipaldi and K.O. Arras, "People Tracking with Human Motion Predictions from Social Forces," International Conference on Robotics and Automation, pp. 464–469, 2010.

[17] J.-Y. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the Algorithm," Intel Corporation, Microprocessor Research Labs, 2000.

[18] D. Arthur and S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027–1035, 2007.

[19] Unusual crowd activity dataset of University of Minnesota, available from http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi