ANOMALY DETECTION IN CROWDED SCENE VIA APPEARANCE AND DYNAMICS JOINT MODELING

Xiaobin Zhu[1], *Jing Liu*[1], *Jinqiao Wang*[1], *Yikai Fang*[2], *Hanqing Lu*[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]System Research Center, Nokia Research Center
[1]{xbzhu,jliu, jqwang,luhq}@nlpr.ia.ac.cn, [2]ykfang@gmail.com

ABSTRACT

In this paper, we propose a novel solution of anomaly detection in crowd scene by jointly modeling appearance and dynamics of motion. First, a novel high-frequency feature based on optical flow (HFOF) is introduced. It can well capture the dynamic information of optical flow. Besides, we adopt the other two types of features, namely multi-scale histogram of optical(MHOF), and dynamic textures (DT). MHOF reserves the motion direction information, while DT captures appearance variant property. The three types of features can complement each other in modeling crowd motions. Finally, multiple kernel learning (MKL) is adopted to train a classifier for anomaly detection. Experiments are conducted on a publicly available dataset of escaping scenarios from University of Minnesota and a challenging dataset from Internet. The results of comparative experiments show the promising performance against other related work.

Index Terms— Anomaly detection, Wavelet transform, Multiple kernel learning, High-frequency, Dynamic texture

1. INTRODUCTION

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. It is an important and challenging problem that has been researched within diverse application domains. Anomaly detection in crowded scenes presents unique challenges because it requires monitoring an excessive number of individuals and their activities, and retaining structural information regarding the entire scene.

Anomaly detection is an active area of research on its own. Various approaches have been proposed, for both crowded and non-crowded scenes. In the following, we will focus on the case of crowded scene. For the work relevant to the crowded scenes can be broadly divided into two categories, according to the type of scene representation adopted. One category is object-based approach, which considers the group as a collection of individuals [1][2][3]. To understand crowd behavior, segmentation, detection, or tracking should be performed in advance. In simple scene, such approaches can achieve good performance. However, in crowed scenes, object occlusions can severely affect the accuracy of segmentation or tracking, which can heavily affect the performance of detection. Additionally, the computational cost will also be tremendous when various objects exist. The other category adopts motion representations regarding to the whole scene that avoid tracking. The most popular is based on optical flow, or spatio-temporal gradients [4][5][6][7]. Mehranet al. [4] modeled crowd behavior using a simplified social force model. This method adopts optical flow measures of interaction within crowds, which are combined with a Latent Dirichlet Allocation (LDA) for anomaly detection. However, LDA is based on a finite vocabulary of discrete words, which will result in the rich motion information lost in word quantization. Kratz and Nishino [5] propose a spatial-temporal model. Firstly, it extracts the temporal and spatial gradient characteristics of scenarios. Then a distribution-based Hidden Markov Model(HMM) is used to describe the motion transitions in the local video regions. This method works only for single kind of normal behavior type. With the change of the types of normal behavior, the detection rate of the abnormal behaviors will heavily decrease. C. Yang et al. [6] propose a novel algorithm for abnormal event detection based on the sparse reconstruction cost for multi-level histogram of optical flows. However, the multi-level histogram of optical flow individually cannot well reflect the dynamics of crowd individually.

Traditional approaches always focus on the spatio-temporal domain with limited feature representations, while the change in frequency domain is often neglected. Based on this consideration, we extract the optical flow in spatio-temporal bricks. Then, wavelet transform is applied. The high-frequency information is adopted to characterize the dynamic characteristics of motion, which is referred to as the high-frequency spatial-temporal feature based on optical flow(HFOF). Besides, in order to detect anomaly in extremely crowd scenes, we combine HFOF with two other types of features, namely Multi-Scale Histogram of Optical Flow (MHOF), and dynamic texture (DT). The MHOF can well keep the motion direction information, while dynamic texture can well keep appearance variant feature. Finally, an effective classifier is

trained based on multiple kernel learning (MKL) using above three types of features. The main contribution can be summarized as the following:(1) A novel HFOF feature is proposed to describe the dynamics of motions; (2) Multiple features are jointly combined to model the appearance and dynamics of motions in crowded scenes; (3) Multiple kernel learning is used to train a classifier for anomaly detection in crowded scenes.

2. METHODOLOGY

We propose a new method for anomaly detection in crowd scene. The framework is summarized in Fig. 1. First, we divide the videos into clips of T frames without overlapping. Then we partition every clip into a collection of local spatio-temporal 3D bricks with equal size $N \times M \times T$. Robust optical flow is computed using the method in [8], then averaged across the T frame in individual brick. Multi-scale Histogram of optical flow (MHOF) and High-frequency information (HFOF) is computed based on the averaged optical flow in every brick. Dynamic texture (LBP-TOP) is computed based on the intensities of pixels. The three types of feature in bricks belong to one clip are concatenated individually. Finally, we train a classification using multiple kernel learning by jointly exploring the above three types of feature for detecting abnormal events.

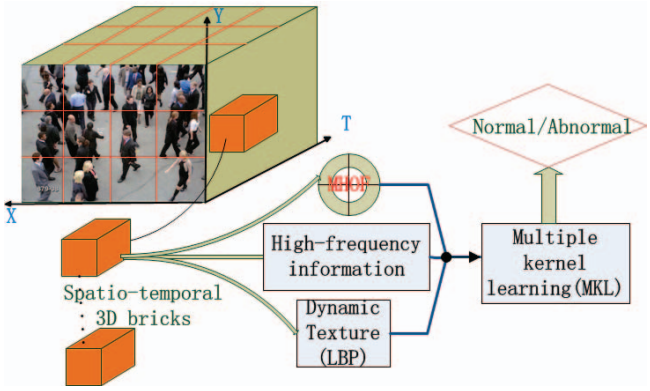

Fig. 1. The framework of our algorithm.

2.1. High-Frequency Feature

Feature in high-frequency can well capture the dynamic information of the transformed feature. we use the high-frequency information of optical flow as a more discriminative representation to characterize the motion dynamics in crowded scenes. We can verify the observation from a simple example shown in Fig. 2. We can see that the distribution of the high-frequency information in case of normal events is more smooth and weaker (with less non-zero values in high-frequency domain) than that of the abnormal case.

In the following, we will present the details for the high-frequency feature extraction. Let $O_{avg}(x, y)$ denotes the averaged optical flow across T direction in 3D brick. Through a single-level two-dimensional discrete wavelet transform for $O_{avg}(x, y)$, we can get an approximate coefficient matrix cA, and detail coefficient matrices cH, cV, and cD. The cA describes the approximate information. The cH describes the dynamic characteristics in horizontal direction. The cV describes the dynamic characteristics in vertical direction, and the cD describes the dynamic characteristics in diagonal direction. This process can be expressed as Eq.(1):

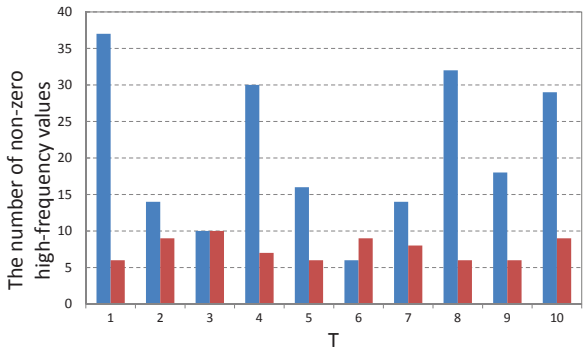

Fig. 2. The abnormal scene(Blue color); The normal scene(Red color). T refers to brick length in terms of frames.

$$[cA, cH, cV, cD] = f_w[O_{avg}(x, y)] \quad (1)$$

where f_w is the function of wavelet transformation. In this paper, we adopt bior3.7 wavelets in wavelet family [9]. During wavelet transform, we compute the two directional optical flow separately. In addition, we neglect the coefficients blow a threshold for filtering out the noise. We use the mean of coefficients and the number of non-zero coefficients in cH, cV, and cD, to characterize high-frequency feature. Lets $HB_{ci} \in R^{12}$ denotes one high-frequency feature for brick i in clip c, where $i \in \{1, 2, ..., k\}$(k is the number of bricks in a clip in our paper). Then the high-frequency feature of clip c is organized as: $H_c = [HB_{c1}, HB_{c2}, ..., HB_{ck}] \in R^{12}$, where $c \in \{1, 2, ..., K\}$(K is the total clip number of training data in our paper).

2.2. Multi-scale HOF

In order to describe the motion direction information, we adopt a motion feature descriptor called Multi-Scale Histogram of Optical Flow [6](MHOF). The MHOF has $D = 8$ bins including two scales. The smaller scale uses the first 4 bins to denote 4 directions with motion magnitude $r < T_r$; the bigger scale uses the next 4 bins corresponding to $r > T_r$(T_r is the magnitude threshold). The MHOF not only describes the motion direction information, but also preserves the motion energy information.

Lets $MB_{ci} \in R^D$ denotes a MHOF feature for brick i in clip c,where $i \in \{1, 2, ..., k\}$. Then the MHOF feature of clip

c is organized as: $M_c = [MB_{c1}, MB_{c2}, ..., MB_{ck}] \in R^D$, where $c \in \{1, 2, ..., K\}$.

2.3. Dynamic Texture

Dynamic textures (DT) are sequences of images of movement that exhibit spatio-temopral stationary properties. Recent research [10] has been shown that dynamic texture is very suitable for unusual event detection in crowded scenes which is extended from the traditional spatial texture into the temporal domain. Let $P(x_c, y_c, t_c)$ be the centre pixel in a spatio-temporal neighborhood. The volume LBP(VLBP) is defined as the joint distribution of the intensities of $3 * P + 3$ pixels on the current frame, t_c, the previous frame, $t_c - L$, and the next frame, $t_c + L$ in,

$$VLBP(x_c, y_c, t_c) = \sum_{i=0}^{3P+1} f(p_q - p_c)2^i \tag{2}$$

$$f(x) = \begin{cases} 1, & x \geq 0 \\ 0, & otherwise \end{cases} \tag{3}$$

where P is the number of neighbors in each frame, L is the temporal interval, $p_q$is the neighbor pixels's intensities, and p_c is the centre pixel intensity.

In order to reduce the total number of patterns, [11] further simplifies this model, only calculating the local binary patterns from three orthogonal planes (LBP-TOP). All the three planes, only XY contains rich appearance features. Because HFOF and MHOF have contained rich motion information, in our algorithm, LBP-TOP of XY plane is extracted from each brick. Different from [10], we use pixel variance in the 2D surface as input. In this plane, we use the 4 pixel neighborhood. As a result, it contains 2^4 local binary patterns. Lets $DB_{ci} \in R^{16}$ denotes dynamic texture feature for brick i in clip c, where $i \in \{1, 2, ..., k\}$, then the dynamic texture feature of clip c is organized as: $D_c = [DB_{c1}, DB_{c2}, ..., DB_{ck}] \in R^{16}$, where $c \in \{1, 2, ..., K\}$.

2.4. MKL for Anomaly detection

MKL refers to the process of learning a kernel machine with multiple kernel functions or kernel matrices. Recent research efforts on MKL have shown that learning SVMs with multiple kernels not only increases the accuracy but also enhances the interpretability of the resulting classifiers [12]. Thus, to jointly explore the above three types of features, we adopt the MKL algorithm to learn a classifier for abnormal detection.

The MKL formulation is to find an optimal way to linearly combine the given kernels. Suppose we have a set of base kernel functions $\{kl\}_{m=1}^{M}(M = 3$ in our algorithm). An ensemble kernel function k is then defined as:

$$kl(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{M} \beta_m kl_m(\mathbf{x}_i, \mathbf{x}_j), \beta \geq 0 \tag{4}$$

Consequently, and often-used MKL model from binary-class data $\{(\mathbf{x}_i, y_i) \in \pm 1\}_{i=1}^{N}$ is:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i kl(\mathbf{x}_i, \mathbf{x}) + b \tag{5}$$

3. EXPERIMENTS

Our algorithm is tested on two datasets: the publicly available dataset from University of Minnesota, and a more challenging dataset which has been collected from websites including youtube.com and ThoughtEquity.com. Details are shown in the following sessions.

3.1. Evaluations on UMN Dataset

The dataset comprises the videos of 11 different scenarios of an escape event. The videos are captured in 3 different indoor and outdoor scenes. Fig. 3 are some selected frames of these scenes. Each video clip starts with an initial part of normal behaviors and ends with sequences of abnormal behaviors.

We follow the same setup as in [4]. We also take optical flow features as the baseline in [4]. In addition, for illustrating the effectiveness of HFOF, we conduct experiments using feature MHOF combined with DT(denoted as MHOF+DT), and using individual HFOF. The ROC in Fig. 5 shows the experimental results. The results illustrate that the proposed method outperforms these state-of-art methods. In addition, we can conclude that HFOF is a high discriminative feature for abnormal detection.

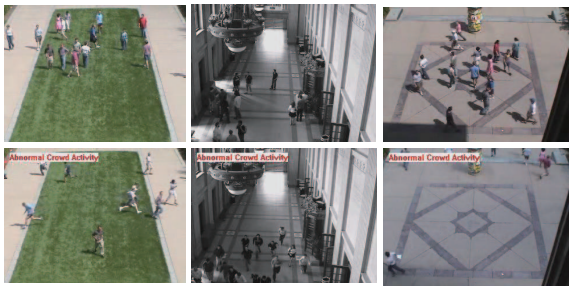

Fig. 3. Top line: samples from normal events; Bottom line: samples from abnormal events.

3.2. Evaluations on Web Dataset

To further evaluate the effectiveness of our algorithm, we conduct experiments on a more challenging dataset, collected from website including Youtube.com and ThoughtEquity.com. The dataset contains 16 sequences of normal crowd scenes such as pedestrian walking, marathon running, and 10 scenes of abnormal scenes such as people fighting, escaping. Fig. 4 is some selected frames of these scenes.

Fig. 4. Top line: samples from normal events; Bottom line: samples from abnormal events.

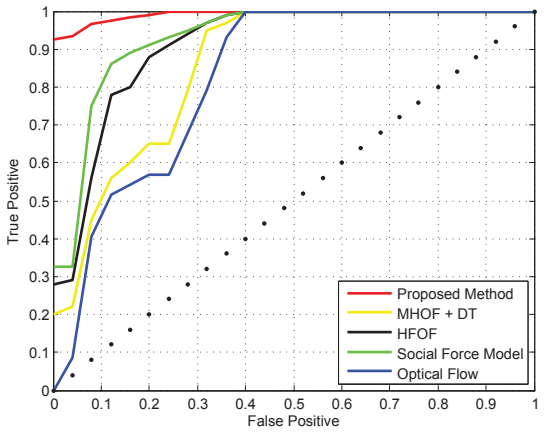

Fig. 5. The ROCs for abnormal detection on UMN dataset.

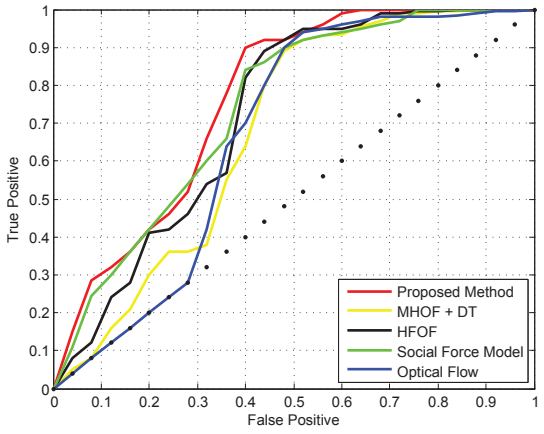

Fig. 6. The ROCs for abnormal detection on web dataset.

We compare our method with the optical flow based method and social force model [4]. We also conduct experiments using MHOF combined with DT, and using individual HFOF. Fig. 6 is the results. It shows our method outperforms these state-of-art methods in such complex scenes. Also, we can conclude that HFOF is a high discriminative feature for abnormal detection. It comes from the fact that our method do not only consider the variance of appearance, but also the dynamics of motion.

4. CONCLUSIOIN

We propose a novel anomaly detection algorithm in crowded scene. In our algorithm, three highly discriminative features, namely HFOF, MHOF, and DT, are extracted together to depict the appearance and the dynamic information of motion in crowded scene. And a classification model based on multiple kernel learning is trained based on above three features. Experiments conducted on the UMN dataset and a challenging web dataset demonstrated the effectiveness of our method, and it is competitive with the state-of-the-art methods.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 60835002, 60905008, and 61070104).

6. REFERENCES

[1] M.T. Chan, A. Hoogs, J. Schmiederer, and M. Petersen, "Detecting rare events in video using semantic primitives with hmm," in *ICPR*, 2004, pp. 150–154.

[2] H. Dee and D. Hogg, "Detecting inexplicable behavior," in *BMVC*, 2004, pp. 477–486.

[3] H.T. Peter, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu, "Unified crowd segmentation," in *ECCV*, 2008, pp. 691–704.

[4] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *CVPR*, 2009, pp. 935–942.

[5] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *CVPR*, 2009, pp. 1446–1453.

[6] C. Yang, J.S. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR*, 2011, pp. 3449–3456.

[7] S.D. Wu, B.E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *CVPR*, 2010, pp. 2054–2060.

[8] M.J. Black and P. Anandan, "The robust estimation of multiple motions: parametric and piecewise-smooth flow fields," in *CVIU*, 1996, pp. 75–104.

[9] A. Cohen, I. Daubechies, and J.C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Comm. Pure Appl. Math.*, vol. 45, no. 5, pp. 485–560, 1992.

[10] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR*, 2010, pp. 2361–2369.

[11] T. Ojala and M. Pietikainen, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[12] S. V. N. Vishwanathan, Z. Sun, N. Theera-Ampornpunt, and M. Varma, "Multiple kernel learning and the smo algorithm," in *NIPS*, 2010, pp. 1975–1981.