Anomaly Detection in Crowded Scenes by SL-HOF Descriptor and Foreground Classification

Siqi Wang, En Zhu, Jianping Yin
College of Computer
National University of Defense Technology
Changsha, China
Email: wangsiqui10c@gmail.com, {enzhu, jpyin}@nudt.edu.cn

Fatih Porikli
College of Engineering and Computer Science
Australian National University
Canberra, Australia
Email: fatih.porikli@anu.edu.au

Abstract—With the widespread use of surveillance cameras, massive video data analysis has become an extremely labor-intensive work. In this paper, we propose an efficient approach to detect video anomaly in crowded scenes based on Spatialized Localized Histogram of Optical Flow (SL-HOF) descriptor and foreground classification. For motion description, the new SL-HOF descriptor can not only preserve classic HOF descriptor’s favorable capability of characterizing the motion velocity and direction of foreground in crowded scene, but also depicts the spatial distribution of optical flow, which implicitly encodes the structure and local motion information of foreground objects in videos. SL-HOF is shown to significantly outperform other classic video descriptors. To further boost the performance of anomaly localization, we then introduce Robust PCA based foreground classification to discriminate anomalous foreground texture. Instead of computationally expensive approaches like \( l_1 \)-norm Sparse Coding, we adopt classic one-class SVM (OCSVM) to model normal video events and detect outliers (anomaly). Our experiments on the challenging UCSD datasets show our approach can achieve state-of-the-art results when compared to existing video anomaly detection methods.

I. INTRODUCTION

Video anomaly detection, which plays a center role in smart video surveillance technology, has drawn increasing interest from academia and industry due to its substantial potential in liberating human beings from long-time tedious work in monitoring possibly as many as hundreds of surveillance screens. Various applications of video anomaly detection can be found in realms such as public security and city management.

Possibilities always come with challenges. The most fundamental challenge in video anomaly detection lies in the definition of “normal/abnormal event” is not as straightforward as the “face” or “pedestrians”. Early work like [1] attempted to explicitly describe certain “abnormal” events or behaviors. However, such methods are severely limited to certain occasions, not to mention that prior knowledge of anomaly is often unavailable. Therefore, recent works tend to consider video anomaly detection as an ”outlier detection” problem [2]. In such methods, only normal video events are modeled and ”anomaly” is considered to be those events diverging significantly from normal events, the idea of which is also followed in this paper. Another key challenge comes from video representation. Object tracking and trajectory analysis [3] [4] [5] [6] seem to be a natural idea to capture high level feature and it works well in uncrowded scenes. Unfortunately, tracking based methods perform poorly in crowded scenes due to frequent occlusions and complex foreground motion. Thus, robust low level features are proposed to deal with crowded scene anomaly detection. Mahadevan et al. [7] proposed a popular Multi-scale Histogram of Optical Flow (MHOF) descriptor. Roshkhari et al. [9] represent spatio-temporal video volume by classic HOG descriptor, while Zhao et al. [10] combined 3D HOG and HOF for video representation in anomaly detection. Proposed by Kratz et al. [11], 3D gradient is adopted by [12] [13]. The third challenge is video event modeling. Sparse Coding is a representative category among those methods. Cong et al. [8] reconstructed a new event by a dictionary, which consists of representative normal events, by solving a \( l_1 \)-norm optimization problem. Lu et al. [12] learned multiple small fixed-size sparse combinations to enable a high-speed detection process. In addition to Sparse Coding, Antic et al. [14] extracted a set of foreground hypotheses to jointly explain all foreground pixels in testing videos. Chen et al. [15] proposed to use Gaussian Process Regression (GPR) and hierarchical feature representation to detect video anomaly. Zhang et al. [13] modeled appearance by Support Vector Data Description (SVDD), while Saligrama et al. [16] and Mahadevan et al. [7] both adopted probabilistic models to describe normal video events.

In this paper, we address video anomaly detection from crowded scenes based on the proposed SL-HOF and foreground classification. The rest of paper is organized as follows: In Sec. II, we discuss how to represent video event with SL-HOF descriptor in terms of motion and analyze the underlying reasons why SL-HOF based representation is effective. In Sec. III we introduce foreground classification as a supplement to SL-HOF based representation to model video foreground appearance on pixel intensity level.. Robust PCA is used to extract video foreground and generate textural features of foreground objects in normal video events. Finally, both SL-HOF features and textural features of normal video events are modeled by OCSVM. Experiments and results on UCSD datasets are shown in Sec. IV. Sec. V concludes the paper.
II. SL-HOF DESCRIPTOR

In this section, optical flow and classic Histogram of Optical Flow (HOF) descriptor is briefly reviewed. Then we present the SL-HOF video descriptor and explain the reasons of SL-HOF’s effectiveness.

A. Optical flow and HOF

Optical flow is calculated by estimating the motion velocity and direction of each pixel from two consecutive video frames. Optical flow has been widely used in video anomaly detection due to its powerful capability of describing motions in video. To represent motion, HOF descriptor is a frequently used descriptor. To calculate HOF, the optical flow magnitude of each pixel is voted into a D-bin histogram (See Fig. 1) as a HOF feature.

B. SL-HOF

Classic HOF is not discriminative enough to detect different complex anomaly, therefore SL-HOF is proposed. The calculation process of a SL-HOF video representation is shown as Fig. 2: Firstly, each video frame is partitioned into non-overlapping M x N patches with equal size, and patches at the same spatial location in consecutive d video frames are stacked as a spatio-temporal cuboid, which is a standard practice in video anomaly detection. Each spatio-temporal cuboid is assumed to be a video event with one or several foreground objects. d is usually small in SL-HOF (e.g. d = 5 in our configuration). This aims to characterize the foreground motion in a small temporal interval, in which the foreground does not suffer from drastic change in spatial location. Then, a spatio-temporal cuboid is partitioned spatially into m x n local 3D regions. Unlike cell-based 3D HOG [17], we do not partition the spatio-temporal cuboid temporally because it will dampen the motion statistics in each region and increase feature dimension. Next, histograms of optical flow are calculated from each region rather than the entire cuboid, which differentiates SL-HOF from HOF. Finally, all histograms are concatenated to obtain a SL-HOF representation \( \mathbf{v} = [\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_m^T] \) as the motion summarization of the cuboid (or video event).

The operations done by SL-HOF descriptor are simple. However, our experiments show SL-HOF descriptor works surprisingly well in video anomaly detection (See Sec. IV-B). The following two properties of SL-HOF contribute to its sound performance: First of all, SL-HOF can depict the distribution of optical flow in video foreground objects, which implicitly encode the structural information of foreground objects into its vector representation. Consider an example of a walking man (normal event) and a man in the wheelchair (abnormal event) in Fig. 3. When the wheelchair moves at a close speed and direction to the walking man, HOF descriptor can be easily fooled since classic HOF extracted from the entire spatio-temporal cuboid wipes out the spatial location information of foreground when calculating histogram. By contrast, SL-HOF yields completely different feature vectors because the spatial distributions of optical flow for man and wheelchair are totally different, which is caused by their different structure. To be more specific, strong optical flow histograms (by "strong" we mean at least one bin of the histogram has a large vote value) are found in region 2, 6 and 10 for the walking man, while strong optical flow histograms are observed in region 3, 6, 7, 8, 10, 11, 12 for man in the wheelchair (See Fig. 3). Therefore, localized HOF extraction can preserve the rough location information of optical flow during calculating histograms. Consequently, SL-HOF can implicitly embed the structural information of foreground objects through optical flow’s spatial distribution.

Secondly, SL-HOF can capture local motion information.
Fig. 4: Local motion information embedded in SL-HOF based video representation.

of foreground objects. Local motion information can be important for discriminating different foreground objects, which is illustrated by Fig. 4: A man (normal event) and a skater (abnormal event) both heading towards right. Two objects share very similar structure and appearance, and it will be difficult to discriminate them when the skater share a close speed with the walking man. When representing both objects with HOF descriptor, the difference will be minor since almost all pixels in both objects are moving towards right and their optical flow are accumulated into the same bin (The first bin in this example). However, it should be noted that the local motion of each body part is not consistent when human is walking. For example, human’s two legs advance alternatively when walking. As shown in Fig. 4, the man’s supporting leg in region 11 almost remain static while the other leg in region 10 is rapidly moving, thus leading to different histograms in region 10 and 11: region 10 can observe a strong optical flow histogram while region 11 cannot. However, since the skater moves as a whole on skateboard, both region 10 and 11 can observe a strong optical flow histogram. SL-HOF can capture such difference in local motions. Besides, in crowded scenes with severe occlusions, localized histogram extraction enables SL-HOF to describe the local motions of individual object parts while HOF obviously cannot. Thus, SL-HOF is more discriminative than HOF or MHOF. SL-HOFs extracted from the same spatial location are described by OCSVM.

III. FOREGROUND CLASSIFICATION

In addition to motion anomaly, there exists appearance or texture anomaly that can be directly classified by eyes from the foreground of a single video frame, e.g., a bicycle on the pavement. Since SL-HOF has been used to characterize the motion of video foreground, we propose foreground classification as a supplement to classify anomalous foreground texture. The procedure is given in Fig. 5: The first step is to perform Robust PCA [18] on video sequences to obtain a sparse matrix $E_t$ for video frame $f_t$. Afterwards, $p_{i,j}$, the probability that pixel $f_{i,j}$ belongs to foreground, is estimated by $p_{i,j} = 2/\exp(-\lambda \cdot (E_{i,j})^2) - 1$, where $\lambda = 1/N_p$, $N_p$ is the total number of pixels on a frame. Next, the probability map is binarized by threshold 0.5 to obtain the foreground. To generate normal texture samples, a fixed-size bounding box is generated by taking each foreground pixel as the box center. To filter out redundant boxes, only those boxes with $>30\%$ pixels to be foreground are preserved. Non-maximum suppression and sampling are adopted to further reduce the number of boxes. Remaining boxes are selected as texture samples and described by HOG descriptor. Texture samples with centers at the same spatial location (patch) are collected for training OCSVM, which is applied to discriminate anomaly from testing texture samples. To control false alarms, the decision threshold of OCSVM is lowered.

IV. EXPERIMENTS

In this section, we test the proposed approach on the most commonly-used UCSD datasets with crowded scenes. In Sec. IV-A, we introduce the adopted UCSD ped1 and ped2 datasets and the configuration of experiments. Sec. IV-B compares the proposed SL-HOF descriptor with other classic video descriptors. Sec. IV-C demonstrate the effect of combining SL-HOF representation and foreground classification. Detection results and comparison with other state-of-the-art methods on are given in Sec. IV-D. Equal Error Rate (EER), ROC Curve and Area under the Curve (AUC) under two commonly-used evaluation criteria, frame-level and pixel-level criteria from [7], are adopted for method comparison. All experiments are carried out under MATLAB 2015b on a PC with 32 GB RAM and 3.90 Ghz Intel i7 4790 processor.

A. Datasets and Experimental Configuration

UCSD ped1 dataset consists of 34 training video sequences and 36 testing video sequences with 200 $158 \times 238$ pixel

Fig. 5: Foreground extraction and texture samples generation.
frames per volume. UCSD ped2 dataset contains 16 training sequences and 12 testing video sequences with $240 \times 360$ video frames, and the numbers of which range from 120 to 180. All of the training volumes merely include normal events such as pedestrians walking on the pavement, while testing volumes contain abnormal events such as a skater and a vehicle on the pavement in crowded or uncrowded scenes.

The detection configuration is as follows: A local patch on a video frame is set to be $10 \times 10$, with consecutive $D = 5$ patches are stacked into a spatio-temporal cuboid. Features extracted from cuboids are assembled into the spatio-temporal basis from [8] to incorporate neighboring spatial and temporal correlation. For SL-HOF, the cuboid is partitioned into $7 \times 8$ regions to yield best performance, and PCA is adopted to reduce SL-HOF feature dimension. The video frames are resized into three scales for detection: $120 \times 180$, $100 \times 150$ and $80 \times 60$ for UCSD ped1 dataset, and $180 \times 270$, $120 \times 180$, $100 \times 150$ for UCSD ped2 dataset. Foreground classification is conducted on the original scale with patch size $21 \times 21$ and bounding box size $20 \times 10$. Both SL-HOF features of spatio-temporal cuboids and HOG features of texture samples are described by OCSVM [19] with Gaussian kernel. The parameters $\nu$ and $\sigma$ are selected from $2^{-12}, 2^{-11}, \ldots, 2^0$ and $2^{-12}, 2^{-11}, \ldots, 2^{12}$ by 10-fold cross-validation, respectively.

**B. Descriptor Comparison**

In this section, the proposed SL-HOF descriptor is compared with the following frequently-used video descriptors in anomaly detection: MHOF [8], 3D HOG [20], 3D HOG+HOF [10], 3D Gradient [12]. For convenience, a single scale ($100 \times 150$) detection is performed on UCSD ped1 dataset with frame-level evaluation criteria using different descriptors. Other configurations are the same as that in Sec. IV-A. The ROC curves obtained by different descriptors are given in Fig. 6 and the EERs are summarized in Tab. I: SL-HOF can improve detection performance significantly by approximately 10% EER reduction, and AUC is improved by 9% to 15%.

**C. Combination of SL-HOF and Foreground Classification**

In this section, we show the effect of combining SL-HOF with foreground classification (FC). Detection with SL-HOF only and detection with SL-HOF and FC are performed on UCSD ped1 dataset under more precise pixel-level evaluation criteria. As shown in Fig. 7 and Tab. II, foreground classification can enhance the anomaly localization performance by detecting texture anomaly, despite generating slightly more false alarms.

**D. Method Comparison**

In this section, we compare the proposed approach with state-of-the-art approaches in literature. For UCSD ped1 dataset, the following approaches are compared: Gaussian Process Regression (GPR) based method [15], Sparse Combination Learning (SC) [12], Mixture of Dynamic Texture (MDT) [7], Sparse Reconstruction Cost (SRC) [8], Social Force (SF) [21], Social Force and MPPCA (SF+MPPCA), Dense STC [9] and Adam et al. [22]. EERs, ROC Curves and AUCs of each method are compared under both frame-level and pixel-level criteria (See Fig. 8, Fig. 9, Tab. III). From Tab. III, our approach yields comparable results with state-of-the-art methods under frame-level criteria while it evidently outperforms other methods under more precise pixel-level criteria.

As to results on UCSD ped2 dataset, the following state-of-the-art approaches are listed for comparison: Spatio-temporal Composition (STC) [23], motion and appearance cue (Zhang et al. [13]), Mixture of Dynamic Texture (MDT) [7], MPPCA

---

**TABLE I: Comparison of video descriptors.**

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>EER</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHOF</td>
<td>29%</td>
<td>76.45%</td>
</tr>
<tr>
<td>3D HOG</td>
<td>31%</td>
<td>73.98%</td>
</tr>
<tr>
<td>3D HOG+HOF</td>
<td>29%</td>
<td>76.28%</td>
</tr>
<tr>
<td>3D Gradient</td>
<td>33%</td>
<td>70.87%</td>
</tr>
<tr>
<td>SL-HOF</td>
<td>21%</td>
<td><strong>85.73%</strong></td>
</tr>
</tbody>
</table>

**TABLE II: Combining SL-HOF with FC.**

<table>
<thead>
<tr>
<th></th>
<th>EER</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL-HOF</td>
<td>37%</td>
<td>63.57%</td>
</tr>
<tr>
<td>SL-HOF+FC</td>
<td>35%</td>
<td><strong>64.35%</strong></td>
</tr>
</tbody>
</table>
In this paper, we propose a simple but efficient approach to detect anomaly from crowded scenes based on SL-HOF.

![Frame-level evaluation on UCSD ped1](image1)

![Pixel-level evaluation on UCSD ped1](image2)

![Frame-level evaluation on UCSD ped2](image3)

![Pixel-level evaluation on UCSD ped2](image4)

**TABLE III:** Detection Results on UCSD ped1 dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>EER(frame)</th>
<th>AUC</th>
<th>EER(pixel)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL-HOF+FC</td>
<td>18%</td>
<td>87.45%</td>
<td>38%</td>
<td>64.35%</td>
</tr>
<tr>
<td>SC</td>
<td>15%</td>
<td>91%</td>
<td>40.3%</td>
<td>63.8%</td>
</tr>
<tr>
<td>GPR</td>
<td>23.8%</td>
<td>88.8%</td>
<td>37.3%</td>
<td>63.3%</td>
</tr>
<tr>
<td>SF</td>
<td>31%</td>
<td>67.5%</td>
<td>79%</td>
<td>19.7%</td>
</tr>
<tr>
<td>SF+MPPCA</td>
<td>32%</td>
<td>67%</td>
<td>71%</td>
<td>21.3%</td>
</tr>
<tr>
<td>MDT</td>
<td>25%</td>
<td>81.8%</td>
<td>55%</td>
<td>44.1%</td>
</tr>
<tr>
<td>SRC</td>
<td>19%</td>
<td>86%</td>
<td>54%</td>
<td>46.1%</td>
</tr>
<tr>
<td>Dense STC</td>
<td>16%</td>
<td>89%</td>
<td>58%</td>
<td>41.7%</td>
</tr>
<tr>
<td>Adam et al.</td>
<td>38%</td>
<td>65%</td>
<td>76%</td>
<td>13.3%</td>
</tr>
</tbody>
</table>

**TABLE IV:** Detection Results on UCSD ped2 dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>EER(frame)</th>
<th>AUC</th>
<th>EER(pixel)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL-HOF+FC</td>
<td>9%</td>
<td>95.07%</td>
<td>19%</td>
<td>81.04%</td>
</tr>
<tr>
<td>STC</td>
<td>13%</td>
<td>92%</td>
<td>26%</td>
<td>-</td>
</tr>
<tr>
<td>Zhang et al.</td>
<td>-</td>
<td>90%</td>
<td>-</td>
<td>73.7%</td>
</tr>
<tr>
<td>MDT</td>
<td>25%</td>
<td>85%</td>
<td>55%</td>
<td>-</td>
</tr>
<tr>
<td>MPPCA</td>
<td>30%</td>
<td>77%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SF+MPPCA</td>
<td>36%</td>
<td>71%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Adam et al.</td>
<td>42%</td>
<td>63%</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

is achieved by classic OCSVM instead of more complex event modeling approaches like Sparse Coding, which often calls for higher computation due to involving $l_{2,1}$ and $l_1$-norm optimization. It demonstrates the effectiveness of our video representation again. Compared with other methods, our approach takes a simple implementation but can achieve comparable or higher detection performance. Examples of detected video abnormal events on UCSD ped1 and ped2 are shown in Fig. 12 and Fig. 13, and we can see different categories of video anomaly can be well detected.

**V. CONCLUSION**

In this paper, we propose a simple but efficient approach to detect anomaly from crowded scenes based on SL-HOF.
Fig. 12: Detected anomaly in ped1: Biker, wheelchair, vehicle and skater.

Fig. 13: Detected anomaly in ped2: Biker, skater, vehicle and walking man with bike.

descriptor and foreground classification. The proposed SL-HOF descriptor can capture the spatial distribution of optical flow and local motion information embedded in video foreground objects, which leads to a higher discriminative power than classic video descriptors in video anomaly detection. Foreground classification is proposed to enhance anomaly localization performance by detecting texture anomaly in video frames. The proposed approach yields state-of-the-art results on the challenging UCSD datasets.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Project No. 60970034, 61170287, 61232016).

REFERENCES