OBJECT-ORIENTED ANOMALY DETECTION IN SURVEILLANCE VIDEOS

Xiaodan Li, Weihai Li, Bin Liu, Qiankun Liu, Nenghai Yu

School of Information Science and Technology, University of Science and Technology of China
Key Laboratory of Electromagnetic Space Information, the Chinese Academy of Sciences

ABSTRACT

Detecting and localizing anomalies in surveillance videos is an ongoing challenge. Most existing methods are patch or trajectory-based, which lack semantic understanding of scenes and may split targets into pieces. To handle this problem, this paper proposes a novel and effective algorithm by incorporating deep object detection and tracking with full utilization of spatial and temporal information. We propose a new dynamic image by fusing both appearance and motion information and feed it into object detection network, which can detect and classify objects precisely even in dim and crowd scenes. Based on the detected objects, we develop an effective and scale-insensitive feature, named histogram variance of optical flow angle (HVOFA), together with motion energy to find abnormal motion patterns. In order to further discover missing anomalies and reduce false detected ones, we conduct a post-processing step with abnormal object tracking. The proposed algorithm outperforms state-of-the-art methods on standard benchmarks.

***Index Terms*—** Anomaly detection, Dynamic image, Object detection, HVOFA, Tracking

1. INTRODUCTION

Discovering abnormal behaviors or events in surveillance videos is of high demand and great importance for public security. With the rapid development of computer vision technologies, automatic anomaly detection has been attracting continuous attention [1, 2, 3]. However, anomaly detection is a challenging problem due to its highly-scene-related property [4]. Most works [5, 6] approaching anomaly detection adopt the following steps. In the training phase, features of normal training samples are extracted. A reference model is then fitted on these features. During testing phase, if features of the input data cannot fit the reference model well, they are considered as anomalies.

Existing approaches for anomaly detection can be roughly categorized into patch-based or trajectory-based methods [7]. Patch-based methods avoid the process of object detection and extract features such as the histogram of gradient (HOG) [8] and histogram of optical flow (HOF) [9] from image patches [10, 11, 12]. However, processing all patches with fixed strides is quite time-consuming. Sabokrou *et al.* [13] proposed to model discriminative patches around interest points. Nevertheless, it lacks semantic understanding of scenes and may split single target into pieces, which is not reasonable for analyzing behaviors. As to trajectory-based methods, Ma *et al.* [14] and Piciarelli *et al.* [15] tried to use visual tracking to collect the trajectories of normal objects and learn a normal trajectory model, which can detect distant objects and get global motion information. However, tracking all the targets is quite time-consuming.

In this paper, we propose a new object-oriented anomaly detection algorithm by incorporating deep object detection and anomalies tracking with full utilization of spatial and temporal information. In real surveillance videos, objects are difficult to be recognized just by their appearance due to fuzziness. So we firstly propose a new dynamic image, which is the fusion of angle, magnitude of optical flow and intensity of the input image, to extract foreground objects with object detection network. The utilization of object detection can not only provide positions, but also the classes of targets, which helps to find appearance anomalies. As for motion anomalies, most existing hand-crafted features such as HOF are not universal for depth-of-field. So we propose a new scale-insensitive feature named histogram variance of optical flow angle (HVOFA) to analyse behaviors. Besides, location anomalies that people or objects appear in inactive region are considered in this work. Speciffically, we mine the active region according to the training data and judge if the object appears in it. Since object detection may miss some distant targets and the proposed HVOFA is a local feature, we utilize object tracking method to track the abnormal candidates instead of all the targets to find missing anomalies and extract their full trajectories as global features to filter out false detected candidates, which is more efficient. The main contributions of this paper are as follows:

- We propose a new dynamic image and a deep dynamic image object detection network, which can greatly improve the detection performance in surveillance videos.

*Corresponding author: whli@ustc.edu.cn. This work is supported by the National Natural Science Foundation of China (Grant No. 61371192), the Key Laboratory Foundation of the Chinese Academy of Sciences (CXJJ-17S044) and the Fundamental Research Funds for the Central Universities (WK2100330002, WK3480000005).

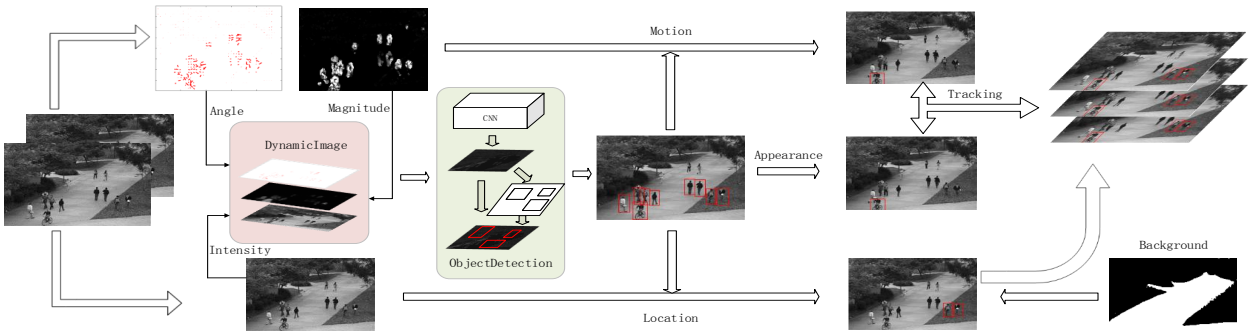


Fig. 1. Framework of the proposed algorithm.

- We propose a new feature named HVOFA, which is effective and scale-insensitive.
- We propose a novel object-oriented anomaly detection algorithm with object detection and tracking, which greatly promotes the performance by making full use of both spatial and temporal information. As far as we know, this is the first time deep object detection and tracking network are utilized for anomaly detection. The experimental results demonstrate the superiority of our algorithm to state-of-the-art approaches.

2. THE PROPOSED METHOD

In this section, we illustrate the proposed algorithm in detail. As shown in Fig. 1, our algorithm consists of three modules. Firstly, a dynamic-image-based object detection is performed to extract objects. Secondly, object category, HVOFA, and motion energy are extracted to detect appearance and motion anomalies. Meanwhile, location anomalies are detected with the background model. Lastly, tracking is utilized on the detected abnormal candidates to find missing targets who can not be detected by object detection. We also conduct a post-processing step to remove false positive anomalies according to the extracted trajectories. Pixels belonging to abnormal candidates will be assigned with corresponding abnormal scores to get the final scoremaps.

2.1. Dynamic Image Object Detection

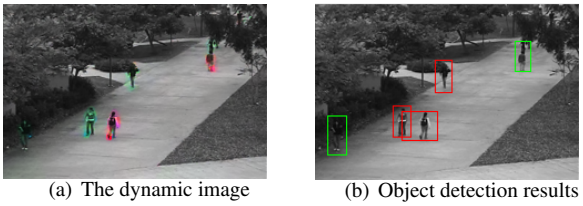
Most existing algorithms extract foreground patches or clusters simply according to the optical flow [12]. These methods are time-consuming and lack semantic understanding of scenes. Thanks to the rapid development of deep learning technologies in computer vision, we propose to adopt deep object detection network to extract objects effectively.

However, existing object detection algorithms suffer from bad lighting conditions and visual quality of surveillance videos. As shown in Fig. 2(a), objects are hard to be recognized when their appearances are similar to the background. Consequently, existing object detection methods may not work well. To make objects distinguishable, we propose a new dynamic image to fuse appearance and motion information together. The proposed dynamic image can be directly fed into existing object detection algorithms with almost no changes on network structures. Specifically, we first calculate the optical flow of the input image. Then we assign the angle and magnitude of the optical flow as the first and second channels of the dynamic image. The intensity of the original image is the third channel. The structure of the dynamic image is like an Hue-Saturation-Intensity (HSI) image:

$$\begin{cases} H & = \text{Angle of the optical flow,} \\ S & = \text{Magnitude of the optical flow,} \\ I & = \text{Image intensity.} \end{cases} \quad (1)$$

To visualize the dynamic image, we treat it as an HSI image and transform it into a RGB image as shown in Fig. 2(a), where the colors are caused by object motion.

We adopt the simple but accurate and efficient region-based fully convolutional network (RFCN) with ResNet-101 pre-trained on ImageNet [16] for oject detection. The model is finetuned on some other labeled traffic surveillance videos[1]. Fig. 2(b) shows an example of detection. The objects in green bounding boxes which cannot be detected from the original RGB image can now be well detected, which indicates that using the proposed dynamic image can obtain a better detection results in dim scenes.



(a) The dynamic image (b) Object detection results

Fig. 2. Dynamic image for object detection. (a) The transformed dynamic image. (b) Object detection results from the dynamic image. Objects in green bounding boxes cannot be detected from the original RGB image.

[1]https://drive.google.com/open?id=0B4ucRlpkNQn-Wm8ydWZJc0kyX0E

2.2. Appearance and Motion Anomaly

Based on detected objects, we extract features and judge whether their appearance or motion are abnormal .

Appearance Anomaly. As a by-product of object detection, we can obtain the object categories and the corresponding confidence scores. We check whether each detected object belongs to normal classes. If not and the confidence s_c is beyond 0.9, it will be regarded as anomaly. The confidence score will be regarded as the anomaly score s_a.

Motion Anomaly. Except for appearances, objects can also be abnormal due to illegal motion. For example, running is not allowed in some scenarios like hospitals. To detect such motion anomalies, we define the so called motion energy to reflect the speed of object motion:

$$E_m = \sum_{i=1}^{N} v_i^2/N, \tag{2}$$

where N is the number of pixels of an object, v_i is the magnitude of optical flow at each pixel.

Besides, some objects may be deceptive, for example, a slow-moving person on a skateboard may be classified as a normal pedestrian according to the appearance and motion speed. To deal with such cases, some hand-crafted features like HOF have been proposed. However, HOF is calculated with weights according to magnitudes, which is sensitive to depth-of-field and requires further processing [17]. So we develop a new effective and scale-insensitive feature called histogram variance of optical flow angle (HVOFA), which counts the frequency of different directions just by angles of optical flow. The HOFA feature is defined as:

$$\text{HOFA} = [f_1, f_2, \ldots, f_B], \quad \sum_{i=1}^{B} f_i = N, \tag{3}$$

where B is the number of directions in HOFA, f_i is the number of pixels of a direction, N is the number of pixels of an object. Accordingly, HVOFA is defined as:

$$\text{HVOFA} = \sum_{i=1}^{B} (f_i - \bar{f})^2 \leq (\sum_{i=1}^{B} f_i)^2 - N^2/B, \tag{4}$$

where $\bar{f} = N/B$. The above inequality is derived with Cauchy inequality. In crowd scenes, rigid objects like cars and skateboards have larger HVOFA than pedestrians, even when they share the same mean value of optical flow. The reason is that pedestrians have more body region motion. Accordingly, if HVOFA of an object is larger than others, the target is more likely to be rigid and abnormal.

If both two features differ from other objects greatly, the target is more likely to be abnormal. In crowd scenes, we set the rule that if the HVOFA s_v and motion energy E_m of an object are larger than thresholds τ_1, τ_2 respectively, we calculate its motion anomaly score s_m as:

$$s_m = s_v + E_m, \quad s_v \geq \tau_1 \text{ and } E_m \geq \tau_2. \tag{5}$$

2.3. Location Anomaly

To deal with location anomaly like walking on the grass, a background model is firstly established with principle component analysis which considers background and foreground as a low-rank matrix and a sparse error matrix respectively [18]. We use successive frames to extract a serials of foregrounds and combine them to extract the active region. Instead of simply regarding objects with most of their pixels not in active region as anomalies, we treat objects whose lower part of the body is not in the active region as anomalies and the location anomaly score s_l is set 1. This can effectively avoid false positive detections such as objects standing next to lawns.

(a) Ped1 (b) Ped2

Fig. 3. Detected anomalies

Overall. After previous steps, we get scoremaps for each frame by assigning pixels belonging to detected abnormal candidates value with s using Eq.(6):

$$s = \omega_1 s_a + \omega_2 s_m + \omega_3 s_l, \tag{6}$$

where $\omega_1, \omega_2, \omega_3$ are weighing coefficients. We adopt a piecewise weighting scheme here. If the abnormal candidate is detected by location-anomaly mechanism, $\omega_1 = 1$ is set to be 1 automatically and ω_2, ω_3 are set to be 0; If it's detected by appearance-anomaly module, $\omega_2 = 1$; Otherwise, $\omega_3 = 1$.

2.4. Post-processing with Tracking

In video anomaly detection, distant objects are too small to be detected. To deal with this problem, TCCF [19] is utilized to track abnormal candidates detected by previous steps to pick up missing targets, which is more efficient than tracking all objects. Besides, HVOFA and motion energy are local motion features, which may cause some false positive detections such as walking with big movements. So we get global motion features of abnormal candidates to remove them according to their trajectories. Specifically, if the target is first detected in frame i before tracking and the coordinate of its center is (x_i, y_i), we define the offset as:

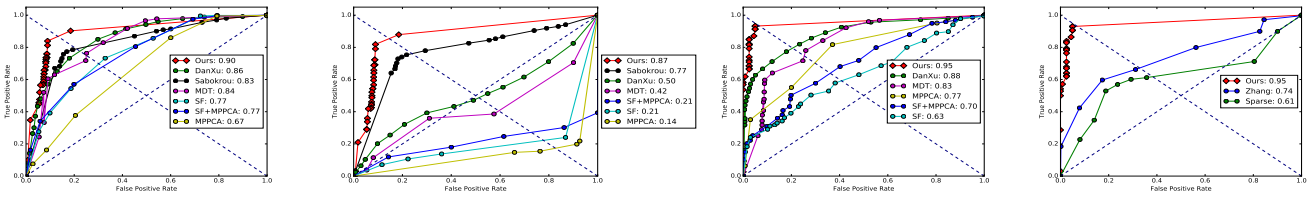
$$\text{offset} = \sqrt{(x_{i+2} - x_{i-2})^2 + (y_{i+2} - y_{i-2})^2}. \tag{7}$$

We remove abnormal candidates whose offsets are smaller than the offset threshold τ_3.

The tracked targets will get the same score with the corresponding initial targets for tracking to get the final scoremaps, which are used for normal/abnormal classification using Eq.(8):

$$pixel(i,j) = \begin{cases} abnormal & \text{if } s(i,j) > \text{threshold} \\ normal & \text{Otherwise} \end{cases}, \tag{8}$$

where i, j indicates the location of the pixel.



(a) Frame Level on Ped1 (b) Pixel Level on Ped1 (c) Frame Level on Ped2 (d) Pixel Level on Ped2

Fig. 4. ROC curves and AUC (As far as we know, only [20, 21] have provided pixel-level ROC curves for Ped2)

3. EXPERIMENTS

3.1. Datasets and Experimental Setting

Datasets. To demonstrate the effectiveness of the proposed framework, we compare our algorithm with several state-of-the-art methods on the challenging benchmark dataset UCSD[2]. As shown in Fig. 3, UCSD contains two scenes: Ped1 and Ped2. The majority of moving objects in this dataset are pedestrians. Non-pedestrian objects and pedestrians with anomalous motion are considered as anomalies.

Parameter Setting. The thresholds for HVOFA and motion energy τ_1, τ_2 are set to be 0.07, 2.0 for Ped1 and 0.547, 0.8 for Ped2 respectively. The offset threshold τ_3 is set to be 8.25.

Evaluation. We use two criteria for evaluation: receiving operating characteristic equal error rate (ROC-EER) and area under curve (AUC) [13]. Higher AUC and lower EER mean a better performance. For frame level evaluation, true positive means in a truly anomalous frame, there's at least one pixel is judged as abnormal. For pixel level, if more than 40% of truly anomalies pixels are detected, it will be treated as true positive [1], which requires accurate localization.

3.2. Results and Discussion

Component Analysis. We conducted several experiments to validate the effectiveness of each part of our algorithm. As shown in Table 1, each part of the proposed algorithm has promoted the performance in different scenes, especially the proposed RFCN with Dynamic image (D-RFCN) and tracking. And this also proves the generality of the proposed algorithm as well as the developed features. Ped2 has no location anomalies, background model makes no difference.

Comparison with State-of-the-art Methods. As shown in Fig. 4 and Table 1, the proposed method provides improved results as comparing to other state-of-the-art methods on both scenes. What's more, our performances on pixel level are almost the same with that on frame level, validating that the proposed method can detect and localize anomalies precisely. We also tested our algorithm on Scene02 of ShanghaiTech Campus dataset [22]. The AUC/EER for frame level is 0.85/0.19, while it is 0.71/0.33 for Conv-AE [23].

[2]http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

Table 1. EER (%) of Frame and Pixel Level on UCSD

Method	Ped1		Ped2	
	Frame	Pixel	Frame	Pixel
SF [24]	31	76	42	80
MPPCA [25]	40	82	30	71
SF+MPPCA [1]	32	71	36	72
Dan Xu [26]	22	42	20	-
Conv-AE [23]	27.9	-	21.7	-
Cascade [13]	**9.1**	15.8	8.2	19
RFCN	46.2	46.8	19.1	21.8
RFCN + T*	33.2	35.3	11.3	14.5
D-RFCN + T*	23.7	26.2	10.5	10.7
D-RFCN + T* + V* + E*	15.5	17.8	6.6	6.9
D-RFCN + T* + V* + E* + B*	13.1	**14.5**	**6.6**	**6.9**

* T: tracking, V: HVOFA, E: energy, B: background model

Our proposed algorithm provides three advantages, which ensures the good performance. First, object detection with deep networks can learn semantic information of videos, which helps avoiding the split of moving targets and rendering a precise localization. Meanwhile, the proposed dynamic image incorporates motion information with appearance, which promotes the accuracy of detection further. Second, HVOFA can distinguish abnormal patterns from normal ones with only angle of optical flow, which makes it scale-insensitive. Third, tracking makes the best use of temporal information, which reduces the false positive and missing targets. The combination of object detection and tracking helps to achieve a good performance since multi-frame detection based mechanism can handle drift in tracking.

4. CONCLUSION

In this paper, we put forward an object-oriented anomaly detection and localization algorithm for surveillance videos. We proposed a new dynamic image for object detection and improved the detection accuracies in real surveillance videos. We also proposed an effective and scale-insensitive feature HVOFA for motion anomalies. Besides, we proposed a background model to detect location anomalies. Lastly, tracking is utilized to pick up missing anomalies and reduce false positive candidates. Extensive experiments were conducted and demonstrated the effectiveness of each part of the proposed algorithm, as well as the superior overall performance.

5. REFERENCES

[1] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR*. IEEE, 2010, pp. 1975–1981.

[2] Yachuang Feng, Yuan Yuan, and Xiaoqiang Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, 2017.

[3] Asimenia Dimokranitou, *Adversarial Autoencoders for Anomalous Event Detection in Images*, Ph.D. thesis, Purdue University, 2017.

[4] Borislav Antić and Björn Ommer, "Video parsing for abnormality detection," in *ICCV*. IEEE, 2011, pp. 2415–2422.

[5] Jefferson Ryan Medel and Andreas Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," *arXiv preprint arXiv:1612.00390*, 2016.

[6] Roberto Leyva, Victor Sanchez, and Chang-Tsun Li, "Video anomaly detection with compact feature sets for online performance," *TIP*, 2017.

[7] Serhan Coşar, Giuseppe Donatiello, Vania Bogorny, Carolina Garate, Luis Otavio Alvares, and François Brémond, "Toward abnormal trajectory and event detection in video surveillance," *TCSVT*, vol. 27, no. 3, pp. 683–695, 2017.

[8] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005, vol. 1, pp. 886–893.

[9] Navneet Dalal, Bill Triggs, and Cordelia Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*. Springer, 2006, pp. 428–441.

[10] Tan Xiao, Chao Zhang, and Hongbin Zha, "Learning to detect anomalies in surveillance video," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1477–1481, 2015.

[11] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette, "Real-time anomaly detection and localization in crowded scenes," in *CVPR*, 2015, pp. 56–62.

[12] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *CVIU*, vol. 156, pp. 117–127, 2017.

[13] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette, "Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes," *TIP*, vol. 26, no. 4, pp. 1992–2004, 2017.

[14] Ke Ma, Michael Doescher, and Christopher Bodden, "Anomaly detection in crowded scenes using dense trajectories," .

[15] Claudio Piciarelli and Gian Luca Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recognition Letters*, vol. 27, no. 15, pp. 1835–1842, 2006.

[16] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *NIPS*, 2016, pp. 379–387.

[17] Ziping Zhu, Jingjing Wang, and Nenghai Yu, "Anomaly detection via 3d-hof and fast double sparse representation," in *ICIP*. IEEE, 2016, pp. 286–290.

[18] John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *NIPS*, 2009, pp. 2080–2088.

[19] Qiankun Liu, Bin Liu, and Nenghai Yu, "Tccf: Tracking based on convolutional neural network and correlation filters," in *ICIG*. Springer, 2017.

[20] Ying Zhang, Huchuan Lu, Lihe Zhang, and Xiang Ruan, "Combining motion and appearance cues for anomaly detection," *Pattern Recognition*, vol. 51, pp. 443–452, 2016.

[21] Yang Cong, Junsong Yuan, and Ji Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognition*, vol. 46, no. 7, pp. 1851–1864, 2013.

[22] Weixin Luo, Wen Liu, and Shenghua Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," *ICCV, Oct*, 2017.

[23] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis, "Learning temporal regularity in video sequences," in *Proceedings of the CVPR*, 2016, pp. 733–742.

[24] Ramin Mehran, Alexis Oyama, and Mubarak Shah, "Abnormal crowd behavior detection using social force model," in *CVPR*. IEEE, 2009, pp. 935–942.

[25] Jaechul Kim and Kristen Grauman, "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in *CVPR*. IEEE, 2009, pp. 2921–2928.

[26] Dan Xu, Rui Song, Xinyu Wu, Nannan Li, Wei Feng, and Huihuan Qian, "Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts," *Neurocomputing*, vol. 143, pp. 144–152, 2014.