Abnormal Event Detection in Urban Surveillance Videos Using GAN and Transfer Learning

Ali Atghaei, Soroush Ziaeinejad, and Mohammad Rahmati
Department of Computer Engineering and Information Technology
Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran
Email: {atghaei, ziaeinejad, rahmati}@aut.ac.ir

Abstract—**Abnormal event detection (AED) in urban surveillance videos has multiple challenges. Unlike other computer vision problems, the AED is not solely dependent on the content of frames. It also depends on the appearance of the objects and their movements in the scene. Various methods have been proposed to address the AED problem. Among those, deep learning–based methods show the best results. This paper is based on deep learning methods and provides an effective way to detect and locate abnormal events in videos by handling spatio-temporal data. This paper uses generative adversarial networks (GANs) and performs transfer learning algorithms on pre-trained convolutional neural network (CNN) which result in an accurate and efficient model. The efficiency of the model is further improved by processing the optical-flow information of the video. This paper runs experiments on two benchmark datasets for AED problem (UCSD Peds1 and UCSD Peds2) and compares the results with other previous methods. The comparisons are based on various criteria such as area under curve (AUC) and true positive rate (TPR). Experimental results show that the proposed method can effectively detect and locate abnormal events in crowd scenes.**

Index Terms—**Deep learning, event detection, generative adversarial network, machine learning, neural networks, transfer learning.**

I. Introduction

Using smart surveillance cameras has recently become very popular. Compared to human monitoring systems, these smart cameras have more consistent behavior and can provide higher accuracy and quicker response. The higher efficiency of these systems has attracted researchers and developers active in developing automated surveillance systems [1]. By considering the widespread use of automated surveillance systems in different applications, it is expected that computer vision–based systems will be able to automatically process a large amount of visual information. Abnormal Event Detection (AED) in videos is one of the most popular computer vision topics. Because of the nature and subjective definitions of 'abnormality' (or 'anomaly'), AED is a very challenging and content-dependent problem [2]. with its major problems addressed, AED can be used for a wide variety of applications such as crowd analysis, subway stations and urban pathways surveillance, summarization of surveillance videos, and smart home monitoring. In the case of using AED for crowd analysis, this system is expected to understand the crowd behavior in a public place for a period of time and inform human agents if it observes an unusual event in the video.

An event is considered abnormal if it is unlikely or unexpected to occur. In statistics, 'anomaly' is defined as an unusual behavior in a distribution or an outlier data point in a data space. Anomaly detection systems are trained with a normal dataset and construe an outlier as abnormal behavior. These systems usually try to jointly detect and locate abnormal behavior. Locating means detecting the location of an abnormal event by showing the abnormal pixels of each frame. For AED to have high accuracy and quick response, an effective data representation method is required. Data space in AED is spatio-temporal with both appearance and movement involved in the process. Researchers use various methods to locate abnormal parts of a video file. The most popular method is gridding, which splits a sequence of frames to smaller fixed-size 3D patches by applying a fixed grid on the frames [3].

Crowd scene analysis has multiple challenges such as occlusion, shadowing, and overlapping of moving objects. Different algorithms have been proposed to overcome these challenge but all of them have their advantages and disadvantages. There are several methods to understand the movement information of scene elements. These methods are able to precisely model the direction and speed of each individual object. However, these methods are usually very time consuming. Besides, the perspective distortion of urban surveillance videos adds to the complexity of the problem as it causes different scale and movement patterns based on object locations and camera position. Because of different lighting situations and subtle difference between normal and abnormal cases, an accurate discriminative model is needed to detect abnormal patches and frames. Various machine learning models such as deep neural networks require a large amount of labeled data. However, AED is an unsupervised learning problem for which gathering a large amount of labeled data is a time consuming and arduous task [2].

Several methods are proposed for AED problem. Trajectory-based methods give a highly accurate model by using the movement of scene entities. Recently, some methods such as Histogram of Gradients (HoG) or Histogram of Optical Flows (HOF) are used to model spatio-temporal properties of the videos. In these methods, an entity

is considered abnormal if the model has never observed its movement pattern before. However, these methods are not efficient for crowd scenes because of their high time complexity and also the problem of moving object occlusion [4], [5]. Researchers have published several papers to address the AED problem. A considerable portion of these published research is evaluated with UCSD Peds1 and UCSD Peds2 datasets [6].

Researchers had used traditional machine learning methods before creating deep learning models. Although these methods are still in use, they have become limited to specific applications. An HMM model using combined dynamic texture as the feature set [6], social force method using spatio-temporal data filtering [7], sparse representation method [8], optical flow clustering method [9], bag-of-visual-word model to represent images [10], and a GMM model using 3D gradient images [11] are examples of these traditional models.

Another class of machine learning models is deep learning networks. Recent research activities commonly use these models to solve complex problems. Researchers also use deep models to address the AED problem. A one-class SVM model using optical flow features which extracted by an auto-encoder network is proposed in [3], and a one-class SVM model using feature set which is extracted by a pre-trained deep model is proposed in [2].

This paper uses a new method to address the AED problem based on two essential steps: Spatio-temporal features processing and motion analysis using optical flow images. Fig. 1 shows the block diagram of the learning phase of the proposed method which has three main parts: pre-processing, training the GAN, and analyzing the optical flow images. The testing phase includes a pre-processing step which is then followed by a spatio-temporal representation. The representation is then calculated and fed to the trained discriminator network as an input. In the final step, abnormal patches of all the frames are determined and an output image is created to show the appearance-motion abnormalities of each frame.

The rest of this paper is organized as follows: Section 2 describes our feature extraction methods and different types of analys that we use. This section also shows the block diagram of the training and testing phases. Section 3 explains the use of transfer learning by importing a part of VGG16 network and Section 4 reports the experimental results and comparisons with state-of-the-art methods. Finally, Section 5 concludes the paper.

II. APPEARANCE-MOTION AND MOTION-ONLY ANALYSES

To detect abnormalities in videos, researchers often analyze either appearance features or motion features. However, processing both appearance and motion features is usually necessary because this combination may include additional important information.

In this work, we take advantages of both appearance and motion features. In the proposed model, the input which is a sequence of original frames moves in two different paths. One path (straight path shown in Fig. 1) is appearance-motion analysis which extracts 3D spatio-temporal features. Another path (downward path shown in Fig. 1) is motion-only analysis which extracts abnormal directions or speeds using optical flow images.

A. Pre-processing of Input Images

Pre-processing of the data is typically the first step to be taken prior to using any learning model. In this step, certain image processing tasks are applied to the frames to resolve their appearance challenges and prepare them to enter the learning model. Examples of the tasks that can be done in the pre-processing step are histogram equalization, foreground extraction, edge detection of foreground objects, obtaining spatio-temporal representations, and patch extraction. A 3D structure is formed to jointly consider the motion features and the appearance features. Fig. 2 shows the spatio-temporal representation of three frames chosen by picking every other frame in a set of consecutive frames.[1] The spatio-temporal representation of frame number t is

$$D_t = <I_t, I_{t\text{-}2}, I_{t\text{-}4}>, \tag{1}$$

where I_t is the edge image of frame number t.

B. Patch Extraction From Appearance-Motion Representation

Frame patches are determined by adding a grid overlay to the appearance-motion representation of the frames. Among all patches, those that have at lease a minimum amount of foreground pixels are chosen to avoid noises and useless data. These patches are gathered from video frames and inputted to the network for learning and testing processes. Fig. 3 shows the output of the patch selection unit.

C. Objects Movement Analysis

In this work we focus on the hue and intensity of the output of Gunner-Farneback optical flow algorithm [12]. The hue shows the path direction and the intensity shows the speed of moving objects in a video sequence.

Fig 4 shows the block diagram of the testing phase of the proposed method. Sorted histograms of direction and speed intensities are calculated in all video frames. Motion abnormalities are then detected based on the sorted histograms. The first 5% of hue/intensity values with the lowest frequencies in the sorted histogram are considered as abnormal. As a result, motion abnormalities are obtained from this step. The result of this analysis, together with that of the appearance-motion analysis can effectively detect and locate abnormal patches in the video.

---

[1]Using every other frame instead of using three consecutive frames better shows the movement of the objects.
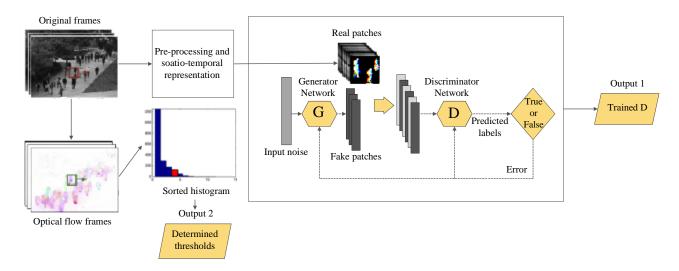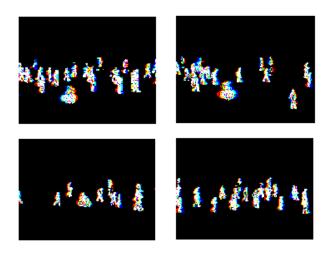
Fig. 1. Training phase of the proposed method.

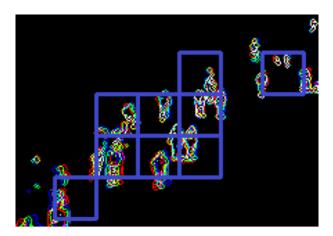
Fig. 2. Spatio-temporal representation of three frames.

Fig. 3. Patch selection from spatio-temporal representation.

III. TRANSFER LEARNING TO BENEFIT FROM VGG16

Data acquisition is one of the main challenges of every machine learning problem. Also, most of the machine learning methods- especially deep learning–based methods- are very time consuming. One of the potential solutions is to transfer a pre-trained model with a specific data domain to a related but different domain without the need for re-learning or providing new datasets. For instance, a model that learned to detect cars in a video sequence can detect unseen trucks without re-learning procedure. This concept is referred to "transfer learning" that introduced in [13].

In this work we use GAN that introduced in [14] to learn the normal data distribution using the normal appearance-motion patches. In each step, the generator outputs patches and gives them to a discriminator that uses a pre-trained VGG16 network. The discriminator then guesses the originality of its input patches. The final error of the discriminator is calculated based on the accuracy of its detections and returned to both the generator and the discriminator. Thus, the generator improves its fake generation accuracy and the discriminator improves its ability to distinguish between real and fake images.

In this work, the first six layers of a pre-trained VGG16 network are used to obviate the need for training more than 10 million learnable parameters and gathering a large amount of various data. Fig. 5 shows the results of this network after 16000 iterations. At first, the generator generates random noises. However, after 16000 iterations the generated patches become very similar to pedestrians in the real patches. This means that the generator outputs are improving.

IV. EXPERIMENTAL RESULTS

The efficiency of the proposed method is evaluated with AUC and EER criteria. The comparison between the proposed method and the state of the art shows that the
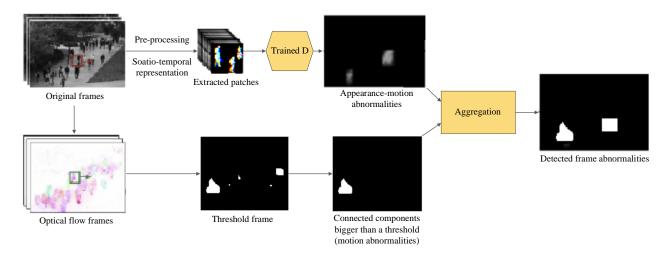
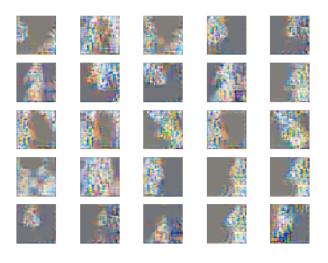Fig. 4. Testing phase of the proposed method.

Fig. 5. The generator outputs after 16000 iterations.

proposed method is very effective in detecting and locating abnormalities in the videos.

Python programming language and Keras module are used to implement the GAN subnetworks. The computer used in this experiment has NVIDIA GeForce 1050 GPU, 32 GB of RAM, and an Intel Core i7 CPU running at 3.1 GHz. To evaluate the model, roc_curve function of Python sklearn module is used to calculate the area under the ROC curve. The generator has four convolutional layers and the discriminator has two parts: The first part is a portion of VGG16 from the input to the pooling layer of the fourth block. This part is constant and unlearnable. The second part consists of two fully-connected learnable layers which are concatenated to the first part.

A. The utilized datasets: UCSD Peds1 and UCSD Peds2

UCSD Peds dataset includes two subsets: UCSD Peds1 and UCSD Peds2. Both show a crowded pedestrian zone where bikers, skaters, and carts are considered abnormal entities. This dataset has 50 training and 48 testing video

samples. The difference between the two subsets is their frame size and the angle of the camera. In UCSD Peds1, the camera is placed on a high altitude. UCSD Peds1 has perspective distortion and a resolution of 238*158 pixels. In UCSD Peds2 the camera is placed on a lower altitude. It has a resolution of 360*240 pixels and NO perspective distortion.

B. Model Evaluation Criteria

There are two types of evaluation for AED models and approaches: Frame-level evaluation and pixel-level evaluation. In frame-level evaluation, the model accuracy is calculated based on the abnormal frame detection without localizing the abnormality. On the contrary, pixel-level evaluation localizes abnormalities and checks the model output with the pixels of ground-truth images. If a model can correctly detect at least 40% of abnormal pixels, its detection is considered true.

The area under the ROC curve which is drawn based on true positive rate (TPR) and false positive rate (FPR) is a common criterion to evaluate and compare the performance and accuracy of different models. TPR is calculated as

$$\text{TPR} = \frac{TP}{P} = \frac{TP}{TP + FN}, \tag{2}$$

where TP is the number of frames that are correctly detected as abnormal frames and FN is the number of frames that are incorrectly considered as normal frames. Equation (2) can also be used in the pixel-level mode but instead of the number of frames, the number of pixels should be counted. FPR is calculated as

$$\text{FPR} = \frac{FP}{N} = \frac{FP}{FP + TN}, \tag{3}$$

where FP is the number of frames that are incorrectly detected as abnormal frames and TN is the number of frames that are correctly considered as normal frames.

Fig. 6 shows the Equal error rate (EER) which is a point in the ROC at the intersection of the curve and a
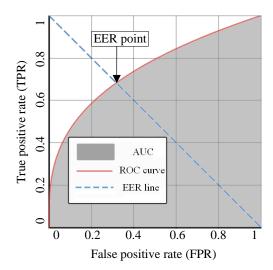
Fig. 6. EER and ROC.

Fig. 7. Detected abnormality with UCSD Peds1.

line that traverses from (0,1) and (1,0). False negative rate (FNR) is the proportion of positives which yield negative test outcomes that calculated as:
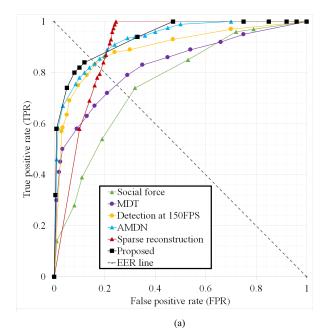
$$\text{FNR} = \frac{FN}{P} = \frac{FN}{TP + FN}, \qquad (4)$$

At EER, FPR and FNR are equal. A lower EER shows that the algorithm is more accurate with less error. Another important criterion is time complexity. An algorithm is more attractive to use in different applications if its overall execution time is sufficiently short.

C. Results with UCSD Peds1 and UCSD Peds2

After training the model with the training data, the model is tested with testing data and the system is evaluated with the above-mentioned criteria by using the ground-truth of the dataset. Fig. 7 shows an instant of testing the system when it is detecting a biker as an abnormality.

Figs. 8(a) and 9(a) show that compared to the state-of-the-art AED methods, the proposed method is more effective
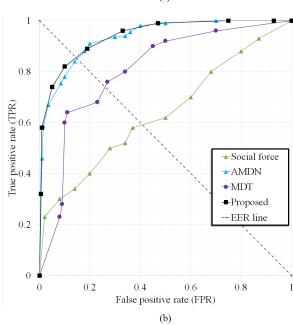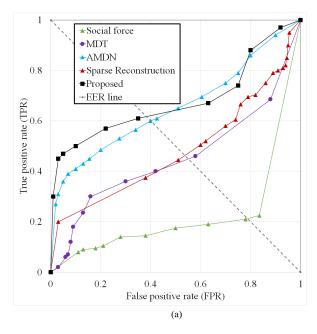

(a)

(b)

Fig. 8. Frame-level ROC comparison of the proposed method with state of the art. (a) when UCSD Peds1 is used and (b) when UCSD Peds2 is used.

in both frame-level and pixel-level evaluations as it has a smaller EER and a larger AUC.[2]

Fig. 9(a) shows that the overall accuracy and AUC of all methods in the pixel-level evaluation are less than those in the frame-level evaluation. This is because the pixel-level evaluation is a more stringent evaluation: In addition to evaluating the ability of an AED method in detecting the abnormalities of the frames, it also evaluates the ability of the AED method in locating the abnormalities. This results in larger EER and smaller AUC values as compared to

---

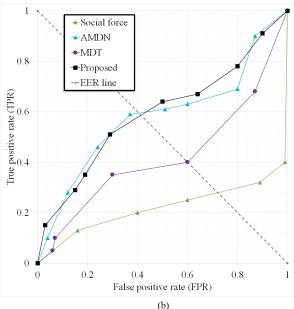[2]ROC curves of the previous methods are derived from [15].

Fig. 9. Frame-level ROC comparison of the proposed method with state of the art. (a) when UCSD Peds1 is used and (b) when UCSD Peds2 is used.

frame-level evaluation. Table I lists the values of AUC and EER of each method and further validates the effectiveness of the proposed method.

The main reason for the superiority of the proposed method is that is uses both appearance and motion analyses. With UCSD Peds1, detection of motion abnormalities is important: Skaters, carts, and bikers usually move faster than pedestrians. Also with more challenging abnormalities such as low-speed bikers, appearance analysis of the proposed method is useful in achieving effective results. The proposed method also has remarkable results with UCSD Peds2 dataset. Table II shows the superiority of the

proposed method over the state-of-the-art methods using EER comparison.

TABLE I
EER AND AUC COMPARISONS WITH UCSD PEDS1.

Method	Frame-level EER	Frame-level AUC	Pixel-level EER	Pixel-level AUC
MDT [7]	25%	81%	58%	44%
Social force [8]	31%	68%	79%	20%
AMDN [3]	16%	92%	40%	67%
Proposed	**14%**	**93%**	**36%**	**73%**

TABLE II
FRAME-LEVEL AND PIXEL-LEVEL EER COMPARISONS WITH UCSD PEDS2.

Method	Frame-level EER	Pixel-level EER
MDT [7]	24%	54%
Social force [8]	42%	80%
AMDN [3]	16%	42%
Proposed	**15%**	**17%**

D. Time Analysis

For GAN to be sufficiently accurate, its learning process requires a large number of repetitions. The time complexity for both training and testing phases is one of the major problems in most deep neural networks including GAN. However, we decrease the time complexity of the GAN by using transfer learning. In this work, VGG16 network is used to extract the appearance features. VGG16 is a well-known pre-trained network which is trained with ImageNet dataset. ImageNet dataset contains more than 14 million images from more than 21800 categories. VGG16 is trained by a large number of various data. Thus, its feature extractor is general enough and we can use its first layers as the feature extractor of our method. To show the effectiveness of the proposed method, we noted the execution times of different parts of our method. Major execution times that should be analyzed in the learning process of the proposed method are pre-processing time, optical flow extraction time, 3D representation time, and classification time. Table III shows these execution times for the proposed method.

TABLE III
EXECUTION TIME OF EACH TASK FOR ONE FRAME.

Pre-processing	Optical flow extraction	3D representation	Classification	Total
0.001 s	0.020 s	0.001 s	0.290 s	0.312 s

V. CONCLUSION

In this paper, a new GAN-based method is proposed to address AED problem. This method uses both appearance-motion and motion-only representations of the input data. Therefore, it effectively detects various abnormalities in shape, skeleton, speed, and direction. The proposed method

transfers the knowledge of a pre-trained CNN (VGG16) to its discriminator CNN to solve this unsupervised problem. This knowledge transfer makes the training phase of the proposed method highly efficient. Experimental case studies are carried out to compare the proposed method with the state of the art. The experiments are based on UCSD Peds1 and UCSD Peds2 datasets.

The results show the effectiveness of the proposed method as it has a lower EER and higher AUC as compared to other methods. In addition to the performance studies, time complexity study is also carried out which shows that the proposed method is sufficiently time effective. With the utilized personal computer, the proposed method can detect and locate abnormalities of a frame in a short time. By using GPU oriented codes and with the advancements of the computers, the proposed method will also be applicable for real-time AED with higher video frame rates (e.g., 30+ fps).

REFERENCES

[1] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[2] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, vol. 172, pp. 88–97, 2018.

[3] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, 2017.

[4] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011.

[5] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2054–2060, 2010.

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

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

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

[9] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2112–2119, 2012.

[10] M. Javan Roshtkhari and M. D. Levine, "Online dominant and anomalous behavior detection in videos," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2611–2618, 2013.

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

[12] G. Farneback, "Two-frame motion estimation based on polynomial expansion," *Scandinavian conference on Image analysis*, pp. 363–370, 2003.

[13] L. Y. Pratt, "Discriminability-based transfer between neural networks," pp. 204–211, 1993.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[15] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1577–1581, 2017.