IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-8, Apr 2019 E-ISSN: 2347-2693

Suspicious Activities and Anomaly Detection in Surveillance Video Using Multiple Instance Learning Techniques

M. Petchiammal Baby[1*], T. Ratha Jeyalakshmi[2]

[1]Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-12, Tamilandu, India
[2]Department of computer applications, Sri Sarada College for women, Tirunelveli -11, India

Corresponding Author: sasibaby08@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si8.4853 | Available online at: www.ijcseonline.org

Abstract— Surveillance videos are proficient to detain a diversity of sensible anomalies. In this work, we advise to find out anomalies by comparing both normal and irregular videos. To remain on away from annotating the irregular segments or clips in training videos, which is very time overwhelming, we recommend to learn anomaly during the deep multiple case position framework by stage averaging weakly labeled direction videos, i.e. the training labels are at video level instead of clip-level. In our approach, we think normal and anomalous videos as bags and video segments as instances in multiple instance learning (MIL), and mechanically learn a deep anomaly location form that predicts high anomaly scores for anomalous video segments. Furthermore, we begin sparsity and temporal softness constraints in the ranking loss function to improved localize anomaly during training. We also set up a new large-scale first of its kind dataset of 128 hours of videos. It consists of 1900 long and untrimmed real-world surveillance videos, with 13 practical anomalies such as fighting, road accident, burglary, break-in, etc. as well as normal activities. This dataset can be used for two tasks. First, general irregularity detection considering all anomaly in one group and all normal activities in another group. Second, for recognizing each of 13 abnormal actions. Our investigational consequence clarify that our MIL performance for anomaly detection achieves significant development on anomaly detection act as compared to the state-of-the-art Techniques. We present the consequences of several current deep learning baselines on anomalous action recognition. The low detection presentation of these baselines finds that the dataset taken is very hard and opens extra opportunities for opportunity work.

Keywords—Multiple instance learning, anomaly, dataset, surveillance video.

I. INTRODUCTION

Surveillance cameras are more and more being used in communal places e.g. streets, intersections, banks, shopping malls, etc. to augment public safety. However, the monitoring potential of law enforcement agency has not kept pace. The result is that there is a glaring shortage in the consumption of surveillance cameras and an unworkable ratio of cameras to human monitors. One critical task in video surveillance is detecting anomalous events such as traffic accident, crimes or illegal activities. Generally, anomalous events rarely occur as compared to normal activities. Therefore, to alleviate the waste of labour and time, developing intelligent computer vision algorithms for mechanical video anomaly discovery is a pressing need. The goal of a practical irregularity detection system is to timely signal an activity that deviates normal patterns and identify the time window of the occurring anomaly. Therefore, anomaly detection can be careful as coarse level video accepting, which filters out anomalies from normal patterns. Once an anomaly is detect, it can further be categorize into one of the specific activities using classification techniques. A small step towards addressing anomaly detection is to develop algorithms to detect a detailed anomalous event, for instance violence detector and traffic accident detector . However, it is obvious that such solutions cannot be comprehensive to detect other anomalous events, therefore they render a limited use in carry out.

Motivation and contributions.

While the previous mentioned approach are appealing, they are based on the supposition that any pattern that deviates from the learned regular patterns would be calculated as an anomaly. However, this guess may not hold true because it is very difficult or impracticable to define a normal event which takes all promising normal patterns/behaviours into account . More outstandingly, the border line between normal and anomalous behaviours is often ambiguous. In addition, under pragmatic circumstances, the same actions could be a normal or an abnormal performance under different conditions. Therefore, it is argued that the training data of normal and abnormal events can help an anomaly

recognition system learn better. In this paper, we proposition an anomaly detection algorithm using weakly labelled education videos. That is we only know the video-level labels, i.e. a video is normal or contains abnormality somewhere, but we do not know where. This is intriguing because we can easily annotate a large number of videos by only assigning video-level labels. To formulate a weakly-supervised learning approach, we resort to multiple instance learning (MIL). Specifically, we propose to learn anomaly through a deep MIL framework by treat regular and anomalous surveillance videos as bags and short segments/clips of each video as instances in a bag. Based on training videos, we unconsciously learn an anomaly ranking model that predicts high anomaly scores for abnormal segments in a video. During testing, a long untrimmed video is divided into segments and fed into our deep set of connections which assigns anomaly score for each video segment such that an anomaly can be detect. In synopsis, this paper makes the follow contributions.

We proposition a MIL solution to anomaly recognition by leveraging only weakly labelled training videos. We intention a MIL ranking loss with sparsity and silkiness constraints for a deep knowledge network to learn anomaly scores for video segments. To the best of our knowledge, we are the first to formulate the video anomaly detection trouble in the context of MIL.

We introduce a large-scale video anomaly detection dataset consisting of 1900 real-world surveillance videos of 13 different abnormal events and normal behavior captured by surveillance cameras. It is by far the largest dataset with more than 15 times videos than obtainable anomaly datasets and has a total of 128 hours of videos.

Investigational results on our new dataset show that our projected method achieve superior practice as contrast to the state-of-the-art anomaly recognition approach. • Our dataset also serves a challenging benchmark for activity appreciation on untrimmed videos due to the complication of activities and large intra-class variation. We there results of baseline methods, C3D and TCNN, on recognize 13 dissimilar abnormal activities.

II. RELATED WORK

Datta et al. [11] predictable to detect human hostility by exploit motion and limbs course of people. Anomaly detection is one of the most difficult and long location difficulty in computer vision. For video surveillance applications, there are several attempts to detect violence or aggression in videos.

Kooij et al. [16] in work video and audio data to detect aggressive actions in surveillance videos. Any aggressive actions that held in the videos are detected by Neural Networks.

Gao et al. [15] proposed violent flow descriptors to detect violence in crowd videos. This helps to spot out the violence in the crowded areas.

Mohammadi et al. [17] planned a new behavior heuristic based approach to categorize violent and non-violent videos. Beyond violent and non-violent patterns intolerance, authors in proposed to use tracking to model the ordinary motion of people and detect deviation from that normal motion as an anomaly. knowledge to rank is an active investigate area in machine learning. These approaches mainly focused on humanizing relative scores of the items instead of personality scores.

Joachims et al. [18] presented rank-SVM to advance retrieval excellence of search engines. Bergeron et al. designed an algorithm for solving numerous instance ranking troubles by means of subsequent linear encoding and demonstrated its application in hydrogen abstraction problem in computational chemistry.

III. PROPOSED ANOMALY DETECTION

The planned approach begin with in-between surveillance videos into a fixed numeral of segments for the duration of training. These segments make instances in a bag. Using both encouraging (anomalous) and downbeat (normal) bags, we train the anomaly detection model using the future deep MIL ranking loss.

Multiple Instance Learning methods

In average supervised classification troubles using support vector machine, the labels of all encouraging and pessimistic examples are available and the classifier is educated using the subsequent optimization function:

$$\min w \; 1 \; k \; Xk \; i=1 \; 1 \; 1 \; z \; \}| \; \{ \; \max(0, \; 1 - yi(w.\varphi(x) - b)) + 1 \; 2 \; kwk \; 2 \; , \quad (1)$$

where 1 is the hinge loss, yi represents the label of each model, φ(x) denotes attribute demonstration of an figure patch or a video division, b is a bias, k is the total number of preparation examples and w is the classifier to be learned. To learn a robust classifier, accurate annotations of positive and negative examples are needed. In the context of supervised abnormality detection, a classifier needs sequential annotations of each segment in videos. However, obtaining temporal annotations for videos is time consuming and laborious. MIL relaxes the assumption of having these truthful temporal annotations. In MIL, precise temporal locations of abnormal events in videos are unknown. Instead, only video-level labels indicating the presence of an anomaly

in the whole video is needed. A video containing anomaly is labelled as positive and a video without any anomaly is labelled as pessimistic. Then, we characterize a optimistic video as a positive bag Ba, where different chronological segment make entity instances in the bag,

(p 1 , p2 , . . . , pm),

where mis the numeral of instance in the bag. We assume that at least one of these instances contains the abnormality. Similarly, the downbeat video is denoted by a pessimistic bag, Bn, where temporal segments in this bag form unresponsive instances (n 1 , n2 , . . . , nm). In the unconstructive bag, none of the occurrence contain an irregularity. Since the exact in sequence (i.e. instance-level label) of the optimistic instances is unknown, one can optimize the objective meaning with respect to the highest scored instance in each bag :

min w 1 z Xz j=1 max(0, 1−YBj (max i∈Bj (w.φ(xi))−b)) +kwk 2 , (2)

where YBj denotes bag-level label, z is the entirety numeral of bags, and all the other variables.

Deep MIL Ranking Model:

Abnormal activities is difficult to define accurately, since it is quite skewed and can vary largely from person to person. Further, it is not obvious how to assign 1/0 labels to anomaly. Moreover, due to the unavailability of sufficient examples of anomaly, anomaly detection is regularly treated as low likelihood pattern detection instead of classification problem. In our future approach, we pose anomaly detection as a regression problem. We want the anomalous video segments to have higher anomaly scores than the ordinary segments. The in a straight line forward move toward would be to use a position loss which encourages high scores for abnormal video segment as compare to normal segments, such as:

f(Va) > f(Vn), (3)

where Va and Vn represent abnormal and normal video segments, f(Va) and f(Vn) represent the corresponding predict scores, respectively. The above ranking purpose should work well if the segment-level annotations are known during preparation. However, in the absence of video segment level annotations, it is not possible to use Eq. 3. Instead, we propose the following several instance ranking purpose function:

max i∈Ba f(V i a) > max i∈Bn f(V i n), (4)

where max is taken over all video segments in each bag. Instead of enforcing ranking on every instance of the bag, we enforce ranking only on the two instances having the maximum abnormality score in that order in the optimistic and pessimistic bags. The division corresponding to the highest anomaly score in the positive bag is most likely to be the true positive instance (anomalous segment). The segment corresponding to the maximum anomaly score in the negative bag is the one looks most similar to an anomalous segment but actually is a normal instance. This unconstructive instance is considered as a hard instance which may generate a false alarm in anomaly uncovering. By using Eq. 4, we want to push the positive instances and negative instances far apart in terms of anomaly score. Our position loss in the hinge-loss formulation is therefore given as follows:

l(Ba, Bn) = max(0, 1 − max i∈Ba f(V i a) + max i∈Bn f(V i n)). (5)

One restriction of the above loss is that it ignore the underlying temporal structure of the abnormal video. First, in real-world scenarios, irregularity often occurs only for a short time. In this case, the scores of the instance (segments) in the anomalous bag should be sparse, representative only a few segment may contain the anomaly. Second, since the video is a sequence of segments, the anomaly score should vary smoothly between video segments. Therefore, we enforce temporal smoothness between anomaly scores of temporally adjacent video segments by minimize the difference of scores for adjacent video segments. By incorporate the sparsity and silkiness constraints on the instance scores, the loss function becomes

l(Ba, Bn) = max(0, 1 − max i∈Ba f(V i a) + max i∈Bn f(V i n)) +λ1 1 z }| { (nX−1) i (f(V i a) − f(V i+1 a))2 + λ2 2 z }| { Xn i f(V i a), (6)

where 1 indicates the temporal smoothness term and 2 represents the sparsity term. In this MIL ranking loss, the error is back-propagated from the greatest scored video segments in both positive and negative bags. By teaching on a large numeral of optimistic and negative bags, we expect that the network will learn a generalized model to predict high scores for anomalous segments in positive bags.

our complete objective function is given by

L(W) = l(Ba, Bn) + kWkF , (7)

where W represents model weights.

Bags Formations:

We partition each video into the equal numeral of non-overlapping temporal segments and use these video segments as bag instance. Given each video segment, we extract the 3D convolution features [36]. We use this feature representation due to its computational efficiency, the evident capability of capturing appearance and motion dynamics in video action gratitude.

IV. IMPLEMENTATION DETAILS

We extract visual skin tone from the fully connected (FC) layer FC6 of the C3D network . Before compute features, we re-size each video frame to 240 × 320 pixels and fix the frame rate to 30 fps. We compute C3D features for every 16-frame video clip followed by l2 normalization. To obtain features for a video segment, we take the average of all 16-frame clip features within that segment. We input these features (4096D) to a 3-layer FC neural network. The first FC layer has 512 units followed by 32 units and 1 unit FC layers. 60% dropout regularization is used between FC layers. We experiment with deeper networks but do not observe better detection accuracy. We use ReLU activation and Sigmoid activation for the first and the last FC layers respectively, and employ Adagrad optimizer with the initial learning rate of 0.001. The parameters of sparsity and silkiness constraints in the MIL ranking loss are set to $\lambda1=\lambda2 = 8 \times 10{-}5$ for the best performance. We divide each video into 32 non-overlapping segments and consider each video slice as an instance of the bag. The number of segments (32) is empirically set. We also experiment with multi-scale overlapping temporal segments but it does not affect detection accuracy. We randomly select 30 positive and 30 negative bags as a mini-batch. We compute gradients by reverse mode automatic differentiation on computation graph using Theano. Specifically, we identify set of variables on which loss depends, compute gradient for each variable and obtain final gradient through chain rule on the computation graph. Each video passes through the set of associates and we get the score for each of its temporal segment. Then we compute loss as shown in Eq. 6 and Eq. 7 and back-propagate the loss for the whole batch.

Evaluation Metric

Following previous works on anomaly detection , we use frame based receiver operating characteristic (ROC) curve and corresponding area under the curve (AUC) to evaluate the performance of our method. We do not use equal error rate (EER) as it does not measure anomaly correctly, specifically if only a small portion of a long video contains anomalous behaviour.

Table 1. A comparison of anomaly datasets. Our dataset contains larger number of longer surveillance videos with more realistic anomalies.

	# of videos	Average frames	Dataset length	Example anomalies
UCSD Ped1 [27]	69	201	4 min	Bikers, small carts, walking across walkways
UCSD Ped2 [27]	29	160	4 min	Bikers, small carts, walking across walkways
Subway Entrance [3]	1	121,730	1.4 hours	Wrong direction, No payment
Subwa Exit [3]	1	64,899	1.4 hours	Wrong direction, No payment
Avenue [28]	36	800	20 min	Run, throw, new object
UMN [2]	4	1200	4 min	Run
BOSS [1]	10	4000	25 min	Harass, Disease, Panic
Ours	1800	7237	120 hours	Abuse, arrest, arson, assault, accident, burglary, fighting, robbery

Comparison with the State-of-the-art:

We compare our method with two state-of-the-art approaches for anomaly detection. Lu et al. proposed dictionary based approach to learn the normal behaviours and used reconstruction errors to detect anomalies. Following their code, we extract 7000 cuboids from each of the normal training video and compute gradient based features in each volume. After reducing the feature dimension using PCA, we learn the dictionary using sparse representation. Hasan et al. proposed a fully convolutional feed forward deep auto-encoder based approach to learn local features and classifier. Using their implementation, we train the network on normal videos using the temporal window of 40 frames. Similar to, reconstruction error is used to measure anomaly. We also use a binary SVM classifier as a baseline method. Specifically, we treat all anomalous videos as one class and normal videos as another class. C3D features are computed for each video, and a binary classifier is trained with linear kernel. For testing, this classifier provides the probability of each video clip to be anomalous.

The binary classifier results demonstrate that traditional action recognition approaches cannot be used for anomaly detection in real-world surveillance videos. This is because our dataset contains long untrimmed videos where anomaly mostly occurs for a short period of time. Therefore, the features extracted from these untrimmed training videos are not discriminative enough for the anomalous events. In the experiments, binary classifier produces very low anomaly scores for almost all testing videos. Dictionary learnt by is

not robust enough to discriminate between normal and anomalous pattern. In addition to producing the low reconstruction error for normal portion of the videos, it also produces low reconstruction error for anomalous part. Hasan learns normal patterns quite well. However, it tends to produce high anomaly scores even for new normal patterns. Our method performing significantly better than demonstrates the effectiveness and it emphasizes that training using both anomalous and normal videos are indispensable for a robust anomaly detection system.

Analysis of the Proposed Method Model training:
The underlying assumption of the proposed approach is that given a lot of positive and negative videos with video-level labels, the network can automatically learn to predict the location of the anomaly in the video. To achieve this goal, the network should learn to produce high scores for anomalous video segments during training iterations. Figure 8 shows the evolution of anomaly score for a training anomalous example over the iterations. At 1,000 iterations, the network predicts high scores for both anomalous and normal video segments. After 3,000 iterations, the network starts to produce low scores for normal segments and keep high scores of anomalous segments. As the number of iterations increases and the network sees more videos, it automatically learns to precisely localize anomaly. Note that although we do not use any segment level annotations, the network is able to predict the temporal location of an anomaly in terms of anomaly scores.

False alarm rate:
In real-world setting, a major part of a surveillance video is normal. A robust anomaly detection method should have low false alarm rates on normal videos. Therefore, we evaluate the performance of our approach and other methods on normal videos only.. Our approach has a much lower false alarm rate. than other methods, indicating a more robust anomaly detection system in practice. This validates that using both anomalous and normal videos for training helps our deep MIL ranking model to learn more general normal patterns.

Anomalous Activity Recognition Experiments:
Our dataset can be used as an anomalous activity recognition benchmark since we have event labels for the anomalous videos during data collection, but which are not used for our anomaly detection method discussed above. For activity recognition, we use 50 videos from each event and divide them into 75/25 ratio for training and testing3 . We provide two baseline results for activity recognition on our dataset based on a 4-fold cross validation. For the first baseline, we construct a 4096-D feature vector by averaging C3D features from each 16-frames clip followed by an L2-normalization. The feature vector is used as input to a nearest neighbor classifier. The second baseline is the Tube Convolutional Neural Network .

which introduces the tube of interest (ToI) pooling layer to replace the 5-th 3d-max-pooling layer in C3D pipeline. The ToI pooling layer aggregates features from all clips and outputs one feature vector for a whole video. Therefore, it is an end-to-end deep learning based video recognition approach. The quantitative results i.e. confusion matrices and accuracy These stateof-the-art action recognition methods perform poor on this dataset. It is because the videos are long untrimmed surveillance videos with low resolution. In addition, there are large intra-class variations due to changes in camera viewpoint and illumination, and background noise. Therefore, our dataset is a unique and challenging dataset for anomalous activity recognition.

Table 2. Activity recognition results of C3D and TCNN .

Method	C3D	TCNN
Accuracy	22.0	25.4

V. CONCLUSION

We propose a deep learning approach to detect real world anomalies in surveillance videos. Due to the complexity of these realistic anomalies, using only normal data alone may not be optimal for anomaly detection. We attempt to exploit both normal and anomalous surveillance videos. To avoid labor-intensive temporal annotations of anomalous segments in training videos, we learn a general model of anomaly detection using deep multiple instance ranking framework with weakly labeled data. To validate the proposed approach, a new large-scale anomaly dataset consisting of a variety of real-world anomalies is introduced. The experimental results on this dataset show that our proposed anomaly detection approach performs significantly better than baseline methods. Furthermore, we demonstrate the usefulness of our dataset for the second task of anomalous activity recognition.

REFERENCES

[1]http://www.multitel.be/image/researchdevelopment/research-projects/boss.php.
[2]Unusual crowd activity dataset of university of Minnesota.Inhttp://mha.cs.umn.edu/movies/crowdactivity-all.avi.
[3] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixedlocation monitors. TPAMI, 2008.
[4] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In NIPS, pages 577–584, Cambridge, MA, USA, 2002. MIT Press.
[5] B. Anti and B. Ommer. Video parsing for abnormality detection. In ICCV, 2011.
[6] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. ´ NetVLAD: CNN architecture for weakly supervised place recognition. In CVPR, 2016.

[7] A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In CVPR, 2008.
[8] C. Bergeron, J. Zaretzki, C. Breneman, and K. P. Bennett. Multiple instance ranking. In ICML, 2008.
[9] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM Comput. Surv., 2009.
[10] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas. Abnormal detection using interaction energy potentials. In CVPR, 2011.
[11] A. Datta, M. Shah, and N. Da Vitoria Lobo. Person-onperson violence detection in video data. In ICPR, 2002.
[12] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solv- ´ ing the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 89(1):31–71, 1997.
[13] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition, 48(10):2993–3003, 2015.
[14] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res., 2011.
[15] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu. Violence detection using oriented violent flows. Image and Vision Computing, 2016.
[16] J. Kooij, M. Liem, J. Krijnders, T. Andringa, and D. Gavrila, Multi-modal human aggression detection. Computer Vision and Image Understanding, 2016.
[17] S. Mohammadi, A. Perina, H. Kiani, and M. Vittorio. Angry crowds: Detecting violent events in videos. In ECCV, 2016.
[18]T.Joachims. Optimizing search engines using clickthrough data. In ACM SIGKDD, 2002.

Author Profile

M. Petchiammal@Baby is currently pursuing full time Ph.D in Computer Science at Sri Sarada College for Women, Tirunelveli. She did her PG in the same institution in 2017 and completed M.Phil (Computer Science) from Manonmaniam Sundaranar University, Tirunelveli in 2018. Her current field of interest is Video Analysis with Image Processing Techniques. Her areas of interest include Cloud Computing and Big Data.

