Advanced Research in Electrical and Electronic Engineering
p-ISSN: 2349-5804; e-ISSN: 2349-5812, Volume 11, Issue 1; January-March, 2024, pp. 5-11

http://www.krishisanskriti.org/Publication.html

Abnormal Event Detection in Social Surveillance System

Mrs. Swathika[1], Logha Mathivanan[2] and Raagavi Durai Raj[3]

[1]Assistant Professor(Sr.G), Department of Artificial Intelligence and Data Science
Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India
[2]Department of Artificial Intelligence and Data Science
Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India
[3]Department of Artificial Intelligence and Data Science
Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India
E-mail: [1]swathikap@mepcoeng.ac.in, [2]mathilogha19112002ai@mepcoeng.ac.in, [3]raagavi20_ai@mepcoeng.ac.in

Abstract—*Monitoring systems play a crucial role in detecting anomalies across diverse real-world scenarios. This paper introduces a novel approach to anomaly detection in surveillance videos using deep learning techniques. Our method addresses the challenge of outlier annotation in training videos, which can be time-consuming and impractical. Instead, we propose a comprehensive multiple-instance ranking loss function that learns outliers from weakly labeled training videos, where labels are assigned at the video level rather than the clip level. In our approach, normal and abnormal videos are treated as "bags," while video segments serve as "instances" in the multi-instance learning (MIL) paradigm. We develop a deep anomaly classification model that automatically detects high anomaly scores for anomalous video clips. Additionally, we incorporate sparsity and temporal smoothness constraints into the loss function to enhance outlier localization during training. Moreover, we explore spatiotemporal context modeling techniques to leverage both spatial and temporal information, capturing the dynamics of scenes over time. Through extensive experiments, we demonstrate the effectiveness of our approach in detecting anomalies in surveillance videos, achieving promising results in terms of accuracy and localization. Our proposed method offers a practical and efficient solution for anomaly detection in monitoring systems, alleviating the need for labor-intensive labeling processes.*

Keywords—*Abnormal Event Detection, Social Surveillance, Deep Learning, Neural Networks.*

INTRODUCTION

In the era of digital surveillance, the proliferation of video cameras in public and private spaces has produced an unprecedented amount of visual data. This increase highlights the critical need for efficient and accurate anomaly detection systems to ensure safety and security. Traditional monitoring systems rely heavily on manual monitoring, which is fraught with challenges such as human error and the impracticality of constant attention. In addition, the accurate labeling of anomalies in large datasets of monitoring material is both time-consuming and cumbersome, which is a major obstacle to the development of automatic anomaly detection systems.

To address these issues, this paper presents a novel approach to anomaly detection in surveillance videos using deep learning techniques. The core of our method is the use of a multi-case sequence loss feature, which innovatively learns to detect outliers from weakly labeled training videos. This approach avoids the need for precise frame-by-frame annotation. Instead, tags are defined at the video level and video clips are treated as "occurrences" in "bags", a method that greatly reduces the annotation burden. Our technique further differentiates itself by adding loss function sparsity and temporal smoothness constraints, which improves the model's ability to find anomalies effectively. Integrating spatiotemporal context modeling uses both spatial and temporal information, capturing the dynamic nature of scenes and facilitating the understanding of anomalies over time. Through rigorous testing, our method showed promising results, providing a significant improvement in both accuracy and localization of detected anomalies.

This method's introduction signals a major breakthrough in surveillance technology by offering an effective and workable way to discover anomalies while reducing the difficulties associated with labor-intensive labeling procedures and manual monitoring. The development of automated solutions to improve security measures and operational efficiency in various contexts is crucial, given the growing demand for robust monitoring systems.This work heralds a significant advancement in surveillance technologies, offering a viable and efficient solution to the prevalent issues of manual monitoring and exhaustive annotation. It lays the groundwork for the next generation of monitoring systems, capable of enhancing security measures and operational efficiencies across various domains. While this study primarily concentrates on the technological aspects and implementation of our anomaly detection system, it also sets the stage for future exploration into the integration of natural language processing to generate contextual captions for identified

anomalies, promising to further the interpretability and utility of surveillance systems.

LITERATURE SURVEY

Automatic detection of human activity in surveillance videos is changing with the advancement of deep learning methods. Traditionally, the industry has relied on labor-intensive feature engineering techniques to extract relevant information from the raw video input. However, the advent of convolutional neural networks (CNN) has changed this approach, allowing direct processing of raw input data and greatly improving the efficiency and accuracy of feature recognition tasks. The paper "3D Convolutional Neural Networks for Human Action Detection" [1] makes an important contribution to this emerging landscape by introducing a new 3D CNN architecture specifically adapted for action detection tasks. Using 3D convolutions to capture both spatial and temporal dependencies of video sequences, the proposed model excels at extracting complex motion patterns that are crucial for viewing human activity. In particular, the model's ability to generate multiple data channels from input frames and integrate them into a comprehensive feature representation increases its effectiveness in handling complex real-world scenarios such as airport surveillance videos. In addition, the authors' innovative approach to validating model results and combining predictions from different models further improves the robustness and generalizability of the proposed framework. Through extensive testing and evaluation of the underlying methods, the paper shows a significant improvement in performance recognition performance, reinforcing its position as a major contributor to the field. The insights gathered from this paper provide a strong foundation for further research and innovation in deep learning-based activity detection systems, promising continued progress in real-world monitoring applications.

In the field of abnormal event detection, significant advances have been made in surveillance video, which has prompted innovative methods to improve detection accuracy and efficiency. One such innovative approach is presented in the article "Detection and Localization of Abnormal Events through Coming Event Prediction" [2]. This paper introduces an ingenious technique called adversarial event prediction (AEP), which fundamentally changes the paradigm of anomalous event detection using an event prediction framework. Using samples of common events, AEP builds a predictive model that detects correlations between current and future events during training. The proposed adversarial learning mechanism further enhances the ability of AEP to learn discriminative representations to predict future events by limiting the learning of past event representations. Careful testing using various datasets such as UCSD-Ped, CUHK Avenue, Subway, and UCF-Crime has fully demonstrated the effectiveness of AEP in anomaly detection, showing its superiority over existing state-of-the-art methods [2].This important work highlights the potential of adversarial learning methods for anomaly event detection and provides a new perspective on the challenges of anomaly detection in surveillance video. Avoiding the need for additional information, such as optical flow or separate samples of abnormal events during training, AEP represents a significant departure from conventional abnormal event detection approaches. In addition, the paper's comprehensive performance analysis and comparative evaluations against existing methods confirm that the proposed mutual learning framework is effective in building the best models for normal event prediction and anomaly detection [2].Experiences gathered from the Anomalous Events article. Detection and Localization by Competing Event Prediction" pave the way for future research projects aimed at further improving and expanding the capabilities of adversarial learning methods for abnormal event detection. As the field evolves, the integration of innovative methods such as AEP promises to improve the reliability and trustworthiness of monitoring systems, contributing to the development of real-world applications that include safety, security and public infrastructure [2].

In addressing the critical challenge of weakly-supervised video anomaly detection, the paper titled "Contrastive Attention for Video Anomaly Detection"[3] introduces an innovative approach to enhance detection performance amidst the prevalent issue of data imbalance, where anomaly instances are vastly outnumbered by normal instances. Traditional methods, often framed within a multiple instance learning paradigm, tend to overlook this imbalance, leading to less effective anomaly localization. This work pioneers the development of a novel, lightweight anomaly detection model designed to leverage an abundance of normal video data, thereby refining the classifier's ability to discriminate between normal and anomalous events effectively.The core innovation of this study lies in its contrastive attention module, which not only enhances the prediction of anomalous segments but also ingeniously generates a "normal" feature representation from anomalous videos. This is achieved by encouraging the model to misclassify these converted features, thereby refining the detection of actual anomalies. Additionally, to counteract the selection of persistent normal segments that may be mistakenly identified as anomalies, the authors introduce an attention consistency loss. This loss function utilizes the classifier's high confidence in recognizing normal features to guide the attention mechanism towards more accurate anomaly detection.Through extensive experimentation across multiple large-scale datasets, including UCF-Crime, ShanghaiTech, and XD-Violence, the proposed model demonstrates a significant improvement in frame-level anomaly detection performance, as evidenced by its superior Area Under the Curve (AUC) metrics compared to existing state-of-the-art methods. The release of the model's codebase promises to facilitate further research and development within the field, encouraging the adoption and adaptation of contrastive attention mechanisms in video anomaly detection tasks. The insights and methodologies presented in "Contrastive Attention for Video Anomaly Detection"[3] mark

a significant advancement in the quest for more effective and efficient anomaly detection in video data, setting a new benchmark for future explorations in the domain.This contribution is especially relevant to the broader context of anomaly detection research, as it addresses a common yet challenging problem in video surveillance and monitoring systems. The innovative approach of employing contrastive attention to mitigate data imbalance and enhance anomaly detection capabilities provides a valuable reference point for subsequent studies aiming to refine and extend anomaly detection techniques.

To address the challenge of weakly supervised video anomaly detection, "Weakly Supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning" [4] proposes an improved approach that advances the level of multiple learning (MIL) in this field. . Conventional MIL frameworks often fail to distinguish rare abnormal clips from common normal cases in video, especially when the abnormalities are subtle and have little deviation from normal. This difficulty is compounded by general control of key temporal relationships between video clips, which can be the key to accurate outlier detection.The introduction of efficient temporal feature learning (RTFM) is an important step forward to mitigate these challenges. RTFM is designed to refine the detection process by using feature scope learning, which skillfully detects positive cases (abnormal events) and improves the robustness of the method against the bias of dominant negative cases found in abnormal videos. This new method not only addresses the inherent limitations of imbalance and the subtlety of anomalies, but also integrates extended circuits and self-aware mechanisms. These improvements are essential to capture both long- and short-range temporal dependencies, enabling more accurate learning of feature sizes.The performance of RTFM is underlined by extensive experimental validation on four benchmark datasets: ShanghaiTech, UCF-Crime, XD-Vilence and UCSD-Peds. In particular, the use of the UCF crime data in the evaluation process closely aligns with our research objectives, further enhancing the relevance of the findings to our own research. The results are convincing, showing that the RTFM-enhanced MIL model not only outperforms several state-of-the-art methods by a significant margin but also significantly improves fine anomaly resolution and sampling efficiency. The contribution of the work is therefore twofold: it presents a theoretically sound method to improve anomaly detection in weakly observed video and provides empirical evidence of its superiority in handling complex anomaly detection tasks on various datasets. This study sets a new benchmark for the field and provides a strong foundation for future research aimed at improving the accuracy and reliability of anomaly detection systems.

The paper "Real-World Anomaly Detection in Surveillance Video"[5] presents a comprehensive approach to anomaly learning using both normal and abnormal videos. One of the main challenges in anomaly detection is the tedious task, which can be very time-consuming, of marking abnormal parts or clips in training videos. To solve this problem, the authors propose a new methodology based on a deep multi-case classification framework, where training labels (abnormal or normal) are determined at the video level instead of the clip level. This innovative approach enables efficient learning of anomalies without precise labels, which greatly reduces the labeling burden.The proposed method processes normal and abnormal videos such as bags and video clips in a multi-level learning (MIL) framework. By automatically learning a deep anomaly classification model, the system can predict high deviations in abnormal videos, which facilitates effective anomaly detection. In addition, the introduction of sparsity and temporal smoothness constraints in the sequential loss function improves the model's ability to accurately locate anomalies during training, which further improves detection performance. In addition to proposing a new method, the paper presents an extensive dataset. . . contains 128 hours of real surveillance video. This dataset, consisting of 1,900 long and uncut videos, includes 13 realistic anomalies such as fights, traffic accidents, robberies and thefts, as well as common activities. This dataset has two main purposes: general anomaly detection, where anomalies are grouped and compared with normal activities, and anomaly detection, where each of the 13 anomalies is detected separately. Experimental results demonstrate the effectiveness of the proposed MIL method for anomaly detection and show significant improvements over the state-of-the-art. In addition, the authors provide an overview of the challenges arising from the material and identify opportunities for further research in this area. The availability of this complex dataset opens up opportunities to develop more efficient anomaly detection systems and underlines the importance of continuous innovation in this field.

PROPOSED MODEL.

Pretrained 3-D ConvNet

The proposed anomaly detection framework incorporates a pretrained 3D Convolutional Neural Network (3D ConvNet) [1] as a foundational element. Specifically, the model integrates the C3D-v1.0 feature extractor, a variant of the Convolutional 3D (C3D) architecture, renowned for its efficacy in video analysis tasks. The C3D architecture is specified in the given Fig. 1. The C3D-v1.0 feature extractor comprises eight convolutional layers, five max-pooling layers, and three fully connected layers. Input to the model consists of a spatiotemporal cube of video frames, with dimensions (3, 16, 112, 112), representing RGB channels, 16 frames and a spatial resolution of 112x112 pixels. The final layer output is tailored for anomaly detection, manifesting as a single unit with a sigmoid activation function, providing an anomaly score. The trained nature of the 3D ConvNet is advantageous, leveraging knowledge from a large-scale dataset, such as Sports-1M, to enhance the model's ability to discern normal and anomalous patterns. This pre-trained 3D ConvNet forms a critical component of the proposed anomaly detection model,

contributing to its robustness and effectiveness in surveillance system applications.

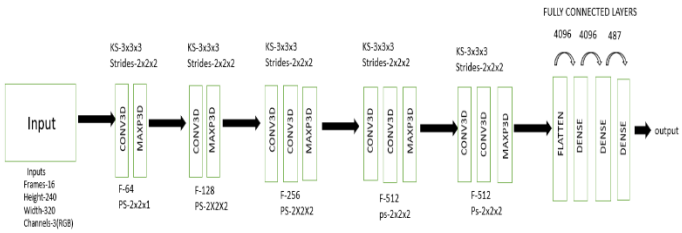

Fig. 1: The C3D Architecture used for feature extraction

For ease of integration, the C3D-v1.0 feature extractor's architecture "Fig. 1", and pre-trained models can be accessed from the relevant repository, ensuring reproducibility and facilitating further research in the field.

- *Convolution Operation*: Given input volume X and a filter W, the convolution operation produces an output volume O as follows:

$$O(i,j,k) = \sum_{a,b,c} X\,(i+a, j+b, k+c) \bullet W(a,b,c) \qquad (1)$$

- *Max-Pooling Operation*: The max-pooling operation selects the maximum value from a set of values in a pooling window.

$$O(i,j,k) = max_{a,b,c} X(2i+a, 2j+b, 2k+c) \qquad (2)$$

- *Fully Connected Layer*: The fully connected layer computes the weighted sum of inputs with learnable weights and adds a bias term.

$$Y = softmax(W \bullet X + b) \qquad (3)$$

The above mentioned equations 1,2 and 3 perform the convolution operations.The proposed model leverages the Convolutional 3D (C3D) architecture Fig. 1, for effective anomaly detection in surveillance systems. The C3D model is characterized by its eight convolutional layers, five max-pooling layers, and three fully connected layers. The convolutional layers employ 3D filters, allowing simultaneous spatiotemporal feature extraction. Input to the model consists of a sequence of video frames forming a spatiotemporal cube with dimensions (3, 16, 112, 112), representing RGB channels, 16 frames, and a spatial resolution of 112x112 pixels. The final output, tailored for anomaly detection, is a single unit with a sigmoid activation function, yielding an anomaly score. Training involves pre-training the model on normal behavior, followed by fine-tuning a dataset comprising both normal and anomalous examples. The loss function for training is binary cross-entropy.

MULTIPLE INSTANCE LEARNING

A well-defined optimization function is utilized in conventional supervised classification scenarios employing support vector machines (SVM), assuming the availability of labels for positive and negative examples. The classifier is trained to minimize a specific hinge loss, considering the feature representation of instances (images or video segments) denoted by φ(x), where x represents an image patch or a video segment, yi represents the label of each example, b is a bias, k is the total number of training examples, and w is the classifier to be learned. However, in supervised anomaly detection, obtaining accurate temporal annotations for video segments is challenging due to its time-consuming and labor-intensive nature.

To address this challenge, Multiple Instance Learning (MIL) [2] is introduced, relaxing the requirement for precise temporal annotations. In MIL, only video-level labels indicating the presence or absence of anomalies in the entire video are needed. Positive videos, containing anomalies, are represented as positive bags (Ba), where different temporal segments serve as individual instances within the bag, while negative videos are denoted by negative bags (Bn), where instances in the bag do not contain anomalies. The objective is to optimize the classifier by considering the maximum scored instance in each bag, thus mitigating the need for detailed instance-level annotations.

This paradigm shift allows for a more flexible approach in the context of anomaly detection, as MIL alleviates the necessity for precise temporal annotations and adapts to the inherent ambiguity in anomaly labeling within video sequences. The formulation considers bag-level labels, making the learning process less dependent on labor-intensive temporal annotation processes, which is particularly advantageous in real-world applications where obtaining accurate temporal information is challenging.

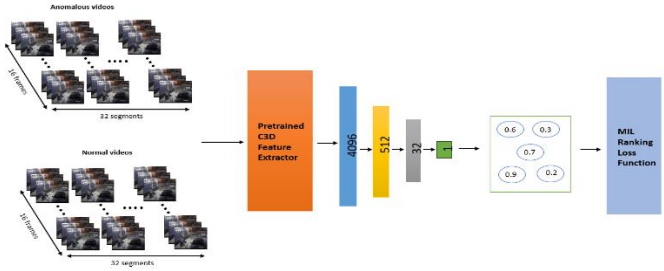

Fig. 2: The MIL Framework Architecture

In traditional supervised classification scenarios employing support vector machines (SVM), labels for both positive and negative instances are readily available, and the classifier is trained using the optimization function:,

$$min_w \frac{1}{k} \sum_{i=1}^{k} \max\big(0, 1 - y_i(\omega \bullet \phi(x) - b)\big) + \frac{1}{2}||\omega||^2 \qquad (4)$$

where 1 represents the hinge loss, yi is the label of each example, ϕ(x) denotes the feature representation of an image patch or video segment, b is a bias, k is the total number of training examples, and w is the classifier under training. Achieving a robust classifier in this context necessitates

precise annotations for both positive and negative examples. In the realm of supervised anomaly detection, the classifier requires temporal annotations for each segment in videos. However, the process of obtaining temporal annotations for videos is arduous and time-consuming.

MIL alleviates the necessity for accurate temporal annotations, departing from the assumption of precise temporal information. In the MIL framework "Fig. 2",, the precise temporal locations of anomalous events within videos remain unknown. Instead, only video-level labels that indicate the presence or absence of an anomaly across the entire video are required. A video containing anomalies is assigned a positive label, while a video devoid of anomalies is labeled as negative. Subsequently, a positive video is represented as a positive bag (Ba), wherein distinct temporal segments serve as individual instances within the bag, denoted as (p1, p2,...,pm), with 'm' being the count of instances in the bag. The assumption is made that at least one of these instances contains the anomaly. Correspondingly, a negative video is represented as a negative bag (Bn), where temporal segments within the bag constitute negative instances (n1, n2,...,nm). Importantly, in the negative bag, none of the instances contain an anomaly. Given the absence of precise information (i.e., instance-level labels) for positive instances, the optimization of the objective function involves considering the maximum scored instance within each bag:

$$min_{\omega} \frac{1}{z}\sum_{j=1}^{z} \max\left(0,1 - Y_{\mathcal{B}_J}\left(max_{i\in\mathcal{B}_j}(\omega.\phi(x_i)) - b\right)\right) + \frac{1}{2}||\omega||^2 \quad (5)$$

Where YBj denotes the bag-level label, z represents the total number of bags, and all other variables remain consistent with Equation 1.

DEEP MIL RANKING MODEL

Defining anomalous behavior poses a considerable challenge [5], given its subjective nature and substantial variation from person to person. The assignment of 1/0 labels to anomalies is non-trivial, and the scarcity of sufficient anomaly examples often leads to treating anomaly detection as a low-likelihood pattern detection problem instead of a classification one.In our proposed methodology, we reframe anomaly detection as a regression problem, aiming for anomalous video segments to exhibit higher anomaly scores than their normal counterparts. A direct approach involves employing a ranking loss that promotes elevated scores for anomalous video segments relative to normal segments, expressed as:

$$f(\gamma_a) > f(\gamma_n) \quad (6)$$

Here, Va and Vn denote anomalous and normal video segments, while f(Va) and f(Vn) signify the corresponding predicted anomaly scores within the range of 0 to 1, respectively. This ranking function is expected to perform effectively when segment-level annotations are available during the training phase.

However, in the absence of annotations at the video segment level, the utilization of Equation 3 becomes impractical. Instead, we introduce a novel multiple-instance ranking objective function:

$$max_{i\in\mathcal{B}_a} f(\gamma^i{}_a) > max_{i\in\mathcal{B}_n} f(\gamma^i{}_n) \quad (7)$$

where the maximum is computed over all video segments within each bag. Unlike enforcing ranking on every instance within the bag, our approach mandates ranking solely on the two instances with the highest anomaly scores in the positive and negative bags, respectively. The segment associated with the highest anomaly score in the positive bag is likely the true positive instance (anomalous segment), while the segment with the highest anomaly score in the negative bag closely resembles an anomalous segment but is, in fact, a normal instance. This challenging negative instance is viewed as a hard instance, capable of generating a false alarm in anomaly detection. Equation 4 is designed to drive positive instances and negative instances farther apart in terms of anomaly scores. Our hinge-loss formulation for the ranking loss is expressed as:

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max\left(0,1 - max_{i\in\mathcal{B}_a} f(\gamma^i{}_a) + max_{i\in\mathcal{B}_n} f(\gamma^i{}_n)\right) \quad (8)$$

This formulation aims to encourage a clear separation between positive and negative instances based on their anomaly scores, providing an effective strategy for anomaly detection in the absence of detailed segment-level annotations.

However, a drawback of the aforementioned loss function is its oversight of the inherent temporal structure within anomalous videos. In real-world scenarios, anomalies often manifest for brief durations, leading to sparse scores for instances (segments) within the anomalous bag, signifying that only a few segments may contain the anomaly. Additionally, considering that a video comprises a sequence of segments, the anomaly score should exhibit smooth variations between these video segments. To address these considerations, we introduce constraints for both sparsity and smoothness on the instance scores. Consequently, the loss function is refined as follows:

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max\left(0,1 - max_{i\in\mathcal{B}_a} f(\gamma^i{}_a) + max_{i\in\mathcal{B}_n} f(\gamma^i{}_n)\right) + \lambda_1 \sum_i^{n-1} (f(\gamma^i{}_a) - f(\gamma^{i+1}{}_a))^2 + \lambda_2 \sum_i^n f(\gamma^i{}_a) \quad (9)$$

where λ 1denotes the temporal smoothness term, and λ 2 represents the sparsity term. In this MIL ranking loss, the error is back-propagated from the maximum scored video segments in both positive and negative bags. Through training on an extensive set of positive and negative bags, the network is anticipated to learn a generalized model capable of predicting high scores for anomalous segments within positive bags (refer to Figure 2). Ultimately, our comprehensive objective function is defined as:

$$L(W) = l(\mathrm{B}_a, \mathrm{B}_n) + \lambda_3 ||W||^2 \quad (10)$$

where W signifies the model weights.

Formation of Bags: Each video undergoes subdivision into an equivalent number of non-overlapping temporal segments, and these segments serve as instances within bags. For each video segment, we extract 3D convolution features [1], a choice made for its computational efficiency and demonstrated effectiveness in capturing both appearance and motion dynamics crucial for video action recognition.

ANOMALY DETECTION

Anomalies, within the scope of surveillance systems, represent instances that deviate significantly from expected or "normal" behavior. In video analysis and security monitoring, detecting anomalies is crucial as these deviations often signal potential threats, safety hazards, or irregular activities. These anomalies can manifest in various forms, including suspicious movements, unexpected actions, or irregular patterns not conforming to the established norm.

DATASET

Recognizing the constraints inherent in existing datasets utilized for anomaly detection methodologies, a concerted effort was undertaken to curate a pioneering, extensive dataset meticulously crafted to serve as the litmus test for evaluating our novel approach. The dataset used in our paper was University of Central Florida (UCF) Crime video dataset. This meticulously assembled repository comprises a wealth of long-form, unaltered surveillance videos, each meticulously selected to encapsulate an exhaustive array of 13 distinctive real-world anomalies. Among these anomalies are occurrences of Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism, It also includes one more class which contains normal videos , meticulously chosen due to their significant ramifications on public safety, thereby ensuring a holistic representation of diverse anomalies commonly encountered within surveillance settings.

Methodical Video Curation Process:

The systematic curation process of this exceptional dataset commenced with an intensive training regimen for ten annotators, each equipped with varying levels of expertise in the realm of computer vision. Employing sophisticated text-based search strategies on prominent platforms such as YouTube and LiveLeak, a meticulous search was conducted encompassing a myriad of search queries for each anomaly, spanning linguistic variations and cultural contexts. Stringent criteria were meticulously applied to filter out videos that did not meet the rigorous standard set forth. Instances of manual editing, prank videos, non-CCTV sourced content, news snippets, handheld camera captures, and compilations were meticulously sieved out, ensuring the exclusive retention of unadulterated surveillance footage displaying explicit anomalies. This rigorous curation methodology culminated in the procurement of a corpus of 950 unedited, authentic surveillance videos capturing explicit anomalies, paralleled by an equal count of 950 normative videos, aggregating to a total of 1900 videos within the dataset.

Elaborate Annotation Procedures

While the core of our anomaly detection methodology necessitated video-level labels for training, the efficiency evaluation demanded a more intricate approach. This involved the meticulous annotation of temporal extents delineating the onset and cessation of each anomalous event within the testing videos. Multiple annotators were tasked with meticulously scrutinizing and labelling these temporal nuances within the same videos, culminating in a composite amalgamation and averaging of annotations provided by divergent annotators. This exhaustive annotation endeavor spanned across several months, comprising intensive collaborative efforts towards dataset completeness and accuracy.

Dichotomy of Training and Testing Set Distribution

The resulting dataset underwent a meticulous partitioning, segregating into two distinct subsets: a

Fig. 3 . Input video of the MIL Framework before detecting Anomaly

Fig. 4 . Output of the MIL Framework after detecting Anomaly comprehensive training set comprising 800 normative and 810 anomalous videos, juxtaposed against a rigorous testing set housing the remaining 150 normative and 140 anomalous videos. Both subsets intricately encompassed the full spectrum of 13 anomalies, adorning various temporal instances within the video corpus. Notably, a subset of videos within the dataset showcased multitudinous anomalies, adding layers of complexity to the evaluation process.A meticulous analysis of video length distribution within the training set, frame counts, and anomaly prevalence within the testing videos is depicted through detailed visualization in Fig. 3, 4, and 5, respectively.

EXPERIMENTS

Methodological Implementation Specifics

Within our anomaly detection framework, intricate visual features are extracted from the FC6 layer of the C3D network [36]. Preceding feature derivation, intricate manipulations involve resizing each video frame to dimensions of 240 × 320 pixels, accompanied by a standardized frame rate of 30 fps. The meticulously computed C3D features for discrete 16-

frame video clips undergo a stringent l2 normalization. Feature abstraction further entails computing the average of all 16-frame clip features within a specific video segment, resulting in high-dimensional (4096D) feature representations. These intricate features are subsequently fed into a carefully constructed 3-layer FC neural network, replete with layers comprising 512, 32, and 1 unit(s) correspondingly. Intermediately situated layers are invigorated by ReLU activation, while the terminal layer employs Sigmoid activation. Optimization takes shape through the meticulous employment of the Adagrad optimizer, initialized with a learning rate of 0.001. Harmoniously intertwined within the network architecture is the judicious application of dropout regularization at a rate of 60% between FC layers.

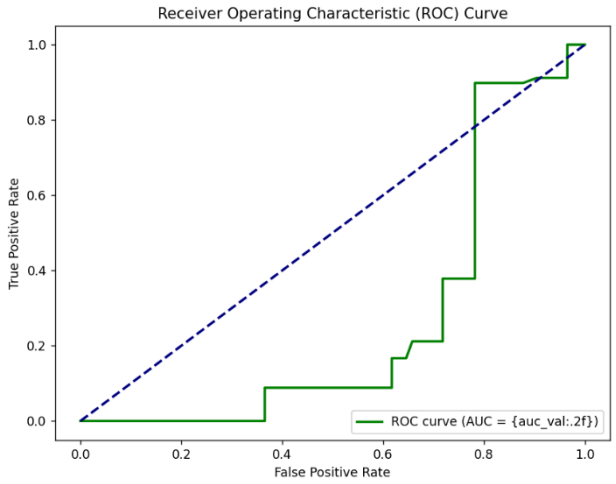

Fig. 5: The ROC Curve for the Anomaly detected video

Thorough Evaluation and Critical Assessment

Unveiling the dynamics behind model training unveils the intrinsic ability of our proposed approach to predict anomaly locations sans the crutch of segment-level annotations. The evolutionary trajectory of anomaly scores over the iterative training process unveils the network's remarkable acuity in accurately discerning between anomalous and normative segments. Moreover, the meticulous analysis of false alarm rates when scrutinizing normal videos underscores our approach's robustness, spotlighting a palpable diminution in false alarm rates vis-a-vis alternative methodologies. This substantiates the irrefutable necessity of training models with a judicious blend of both anomalous and normative video data for the fortification of robust anomaly detection systems.In the working code of our proposed model initially we have visualized the ROC – AUC curve for the anomaly detected video Fig. 5 . The Output video brings the red bounded box whenever the predicted anomalous score crosses 0.4 which was the fixed threshold. Thus visualization of the anomalous part was enhanced as shown in Fig. 4 which is the output video and Fig. 3 denotes the frames of the input video.

Probing Anomalous Activity Recognition Endeavors

The dossier extends beyond anomaly localization to embrace the challenging realm of anomalous activity recognition. Subdividing event-labeled video subsets, we venture into the domain of activity recognition experiments employing established methodologies like the C3D feature extractor which was used to extract spatio-temporal features in our paper.The outcomes, however, unveil a disheartening performance due to the multifarious challenges inherent in protracted untrimmed surveillance videos marked by low resolutions, pronounced intra-class variations, mutable viewpoints, shifting illumination, and ambient background noise. This narrative underscores the singular complexity and intrinsic intricacies defining our dataset in the domain of anomalous activity recognition.

REFERENCES

[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4489–4497.

[2] Jongmin Yu,Younkwan Lee,Kin Choong Yow,Moongu Jeon,Witold Pedrycz,”Abnormal Event Detection and Localization via Adversarial Event Prediction” in Proc. IEEE.Trans on Neural Networks and Learning Systems,2022.

[3] Shuning Chang,Yanchao Li,Shengmei Shen,Jiashi Feng“Contrastive Attention for Video Anomaly Detection” in Proc.IEEE Trans on Multimedia,Vol. 24,2022.

[4] Yu Tian,Guansong Pang,Yuanhong Chen,Rajvinder Singh,Johan W.Verjans,Gustavo Carneiro, “Weakly-supervised Video Anomaly Detection with Robust TemporalFeature Magnitude Learning” in 2021.

[5] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6479–6488

[6] G. Pang, C. Shen, H. Jin, and A. van den Hengel, “Deep weakly supervised anomaly detection,” 2019, arXiv:1910.13601.

[7] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y.-W. Tai, “Memory-attended recurrent network for video captioning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8339–8348.

[8] Y. Gal and Z. Ghahramani, “A theoretically grounded application of dropout in recurrent neural networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 29, 2016, pp. 1019–1027

[9] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In ICCV, 2013.