Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7778343

VIDEO ANOMALY DETECTION IN SURVEILLANCE VIDEOS USING RESNET18+DEEPLABV3 ARCHITECTURE

S. Naveen Kumar
Research Scholar, Dr. M.G.R. Educational and Research Institute, Chennai, India
sk.naveenkumar@gmail.com

Dr. R. Shoba Rani
Professor, Department of Computer Science and Engineering,
Dr. M.G.R. Educational and Research Institute, Chennai, India
shobarani.cse@drmgrdu.ac.in

Abstract: *The popularity of video anomaly detection may be attributed to its use in security cameras. A variety of plausible abnormalities may be captured by surveillance cameras. Convolutional neural networks (CNNs) have recently found great success in the field of video analysis, and as a result, more and more AD algorithms are incorporating CNNs into their own to improve processing speed, efficiency, and detection accuracy. In this research, we use a novel approach to anomaly identification in video surveillance by means of a unique algorithm. To improve performance and achieve high accuracy, we employed the Resnet18+DeeplabV3 anomaly detection method enhanced with PSO. We show how our methods can automatically indicate the difference between regular and suspicious actions using real-world video footage, which may help with security monitoring. The proposed method shows much better performance with 91.2% accuracy which is higher that present methodologies.*

Keyword: Video anomaly detection, Resnet18, DeeplabV3, PSO.

I. INTRODUCTION

Video surveillance is an attractive and significant area of study in both the commercial world and the academic world. Train stations, airports, military bases, and shopping malls are just some of the places where safety concerns have prompted a surge in interest in video monitoring among academics. Finding out whether anything suspicious is going on at the location is vital since it is difficult to develop detection systems without information about a target activity & the scene. Video surveillance is often employed in crowded areas in order to detect and analyze irregularities in a chaotic setting. Any incident that deviates from the norm is considered an anomaly. It might be challenging or even appear impossible to model and interpret the results of the unusual scenes. Since anomalies can be defined as activities that depart from the known patterns, spotting them in video surveillance may be a difficult task. In certain contexts, such as a gun club, an unusual occurrence, such as a shooting, may be commonplace. Even though shooting is often considered deviant activity, it is normalized among shooting clubs. In contrast, certain actions may seem typical at first glance but are really rather unusual when seen in a specific context. [1].

The problem of anomaly detection (AD) is one to the most fundamental ones in computer vision, and it has been extensively investigated in many different fields. The essential concept is how to recognize out-of-the-ordinary data that differs substantially from the norm. An anomaly might be called an outlier or a unique observation depending on the context or data being examined. Anomaly detection methods and their applications may be found in many fields, including those dealing with fraud detection, video monitoring, healthcare, security checks, and defect identification. To detect flaws and outliers, AD is commonly utilized in the area of high-speed rail safety inspection. More and more applications based on the AD task are being used in train safety inspection to enhance detection efficiency, decrease detecting costs, and enable intelligent detection [2]. This is because of the critical role train safety inspection plays in ensuring the secure and reliable operation of high-speed trains. When compared to what is considered to be typical or normal, an anomaly stands out as a substantial departure. These anomalous occurrences provide a significant challenge to the efficiency of VAD (Automated video anomaly detection) models [3] since they are both rare and varied, but also highly contextual and often ambiguous.

For obvious reasons, public spaces are being outfitted with surveillance cameras. With the use of AI and machine learning, security cameras can now automatically recognize and identify objects or events of interest. To detect anomalies in videos means to locate them in time and place. Anomaly detection in the workplace, in the realm of security, and elsewhere are only a few examples [4]. Since low resolution with dynamic background, modeling the crowd behavior, occlusion between individuals, illumination changes, or random variations in a crowd all pose significant challenges to the research community, anomaly detection in public surveillance systems is of great interest in computer vision. Modern public safety, security, monitoring of group activities, sports analysis, & visual surveillance all rely heavily on automatic anomaly detection. It would be much easier to make sound judgments on safety and emergency control if automated surveillance systems could swiftly identify odd and hazardous circumstances in a crowded setting. Therefore, surveillance systems are crucial for public safety, security, & statistics reasons in complex and congested locations such as busy streets, political demonstrations, public festivities, airports, railway stations, and retail malls [5].

The two primary approaches to determining video quality are the subjective and objective approaches. Humans have a great visual system, making subjective diagnosis more accurate. Human resources are costly; thus, their use is limited to areas such as analyzing large-scale video surveillance systems. Subjective diagnosis has been shown to be ineffective, and manual surveillance monitoring is time-consuming. Recent years have seen a focus on objective diagnosis as a means to address these drawbacks. Imaging and video-based technology are both examples of objective diagnostic methods. The image-based method relies on a single still picture from the camera to make the anomaly prediction, while the video-based approach use a brief video clip to make the anomaly detection. Methods based on moving pictures are often more reliable than those based on still pictures. The analysis of motion [5, 6] may be used to the identification of static images [6, 7].

The popularity of video anomaly detection may be attributed to its use in security cameras. While the price of installing surveillance systems has dropped dramatically in recent years, human involvement is still needed to spot crimes like assault, theft, and vandalism. Creating sophisticated algorithms for the video anomaly detection is necessary [7] due to the high

expense of extra human work and lost time.Convolutional neural networks (CNNs) have recently found great success in the field of video analysis, and as a result, more and more AD algorithms are incorporating CNNs into their own to improve processing speed, efficiency, and detection accuracy. In this research, we use a fresh approach to anomaly identification in video surveillance by means of a unique algorithm. To improve performance and achieve high accuracy, we employed the Resnet18+DeeplabV3 anomaly detection method.

II. RELATED WORKS

The use of surveillance footage for diagnostic purposes has been extensively researched. Here, we go into what we learned about video anomaly detection methods.

For the purpose of outdoor video surveillance, Wei Niu et al. [8] provided a framework for identifying and tracking people's movements. The following are some of the benefits of their study: The sophisticated control and fail-over methods they provide to low-level motion detection techniques like frame differencing & feature correlation make activity detection as well as tracking more reliable. Without the need to construct complex Markov chain, hidden Markov model (HMM), or coupled hidden Markov model (CHMM), they propose an efficient representation of human activities that allows recognition of different interaction patterns among a group of people based on simple statistics computed on the tracked trajectories. They showed how their methods work by applying them to real-world video footage, where they successfully distinguished between typical and suspicious activities in a parking lot.

Entering, using a terminal, opening a cabinet, picking up a phone, etc. are all examples of human activities that may be recognized by the method Douglas Ayers et al. [9] suggested. Low-level methods such as skin detection, tracking, & scene change detection are used by the system. We have been able to correctly identify these motions in several sequences, even those involving more than one individual executing the same motions.Their system can identify these movements thanks to stored information about the space. The'states' and 'transitions' between these states are what our system uses to mimic action recognition. The location of a person, the detection of a change in the scene, or the presence of a monitored item may all trigger a change in status. Key frames are extracted from video sequences, thereby serving as content-based video compression, and the system also generates written descriptions of detected behaviors. Multiple video sequences have been used to successfully test the system. In this study, we offer some of the most important findings. The principles given here may be used to improve automated security.

The work of Gwanyong Park et al.[10] It was hypothesized in this research that thermal and visual photographs of building envelopes may be used to identify thermal anomalies automatically. In order to identify the wall domain, a deep learning-based picture segmentation model was employed, using earlier research utilizing wall temperature data on distribution as a point of comparison. It achieved an accuracy of mIoU 0.842 in categorizing walls—the primary aim for detecting thermal anomaly—after being trained using visible pictures of the building envelopes as part of a segmentation model. When looking at thermal pictures with various walls or anomalies, a multimodal distribution was suggested, and from that, a algorithm for anomaly identification was developed. Using picture segmentation as an example, a quantitative evaluation of the method was performed.

The method of driver monitoring proposed by OkanKopuklu et al. [11] is based on open set recognition. To achieve this goal, we produce and release the Driver Anomaly Detection (DAD) data sets, a video-based benchmark dataset including types of anomalous actions that have not been seen before in its test set. Since some of the aberrant activities in this dataset have never been observed before, it is imperative that we are able to identify them and separate them from regular driving. In order to extend the embedding of normal driving that has been learnt, which can then be utilized to identify anomalous activities in the test set, we offer a contrastive learning strategy.

The anomaly gap & scene gap between simulated and real-world settings were addressed by a unique PFMF proposed by Zuhao Liu et al. [12]. The proposed system consists of an anomaly classifier, a domain classifier, and a mapping adaption branch to produce unknown anomalies of unbounded kinds and close the scene gap, respectively. Our method introduces a new paradigm in using virtual datasets, which may be used to forego time-consuming anomaly gathering in the real world. The proposed PFMF achieves state-of-the-art results across three benchmark datasets, or the ablation research demonstrates the efficacy of our model's individual components.

Scene-dependent video anomaly detection was the topic of work by Shengyang Sun et al. [13], who introduced a hierarchical semantic comparison approach. Incorporating scene-aware autoencoders & motion augmentation into our hierarchical semantic contrasted learning architecture, our proposed model shows great promise on both scene-independent & scene-dependent VAD. Three independent experiments using both publicly available and custom-built datasets confirm the efficacy of our approach.

According to Bo Quan and colleagues [14]. The deep learning-based technique for semantic picture segmentation performs well and has several potential uses. In this study, we propose a refined version of the DeepLabV3 model by fusing numerous shallow features for more accurate classification and segmentation using the Unified Neural Network (U-Net). This allows us to take use of the ASPP structure's ability to expand the receptive field. Improving semantic segmentation by addressing the issue of sample imbalance, the DICE loss or BCE loss function set may be used to great advantage. Through experimental comparison on the tough road segmentation dataset, the approach presented in this study confirms its efficacy by successfully decreasing the misdetection or road truncation in scene road segmentation, outperforming deeplabv3 and U-Net in the process.

The general strategies for FER using ResNet compositions were given by Pratyush Shukla et al. [15]. The most popular and straightforward methods are those that handle ResNet variations like ResNet-18 and ResNet-50. They may be used in whole machine learning projects and provide good outcomes. Using the sobel-operator yields unique outcomes and is feasible with the ResNet variations. They provide remarkable results when used to extract edges from complicated pictures and are therefore very valuable. The combination of ResNet with Attention Mechanism & Deformable Convolution, which is based on the principle of employing a dynamic kernel shape, is the state-of-the-art method. Other methods, such as the 3D Inception-ResNet technique and ResNet paired using Heart Rate Variability analysis, are novel and may be used in practical contexts for the first time.

Summary:

- The model's performance was enhanced as a consequence of higher-quality features and enhanced anomaly detection methods.
- TensorFlow Lite is a deep learning framework optimized for mobile devices, and it recommends DeepLabV3+ as a base image segmentation model owing to its great performance and speed.
- Deep learning-based anomaly detection in surveillance videos may extract more useful information and better categorize anomaly kinds.

III. PROPOSED METHOD

In this article, we introduced a unique method for detecting anomalies in video surveillance. We improved performance and detection accuracy by using a Resnet18+DeeplabV3 algorithm for the anomaly detection. Using our method, typical abnormalities in surveillance video images including brightness anomaly, blur, occlusion, etc. can be detected with only a single frame, and the diagnostic issue is transformed into a multiclass job. We suggest a method of frames extraction that involves taking still images from each movie. Our method is image-based, but it can be easily applied to a video sequence by extracting frames at random and putting them into our network one by one. Our suggested Resnet18+DeeplabV3 architecture will identify anomalies in the collected and processed frames. The conceptual structure of the suggested procedure is shown in figure1.

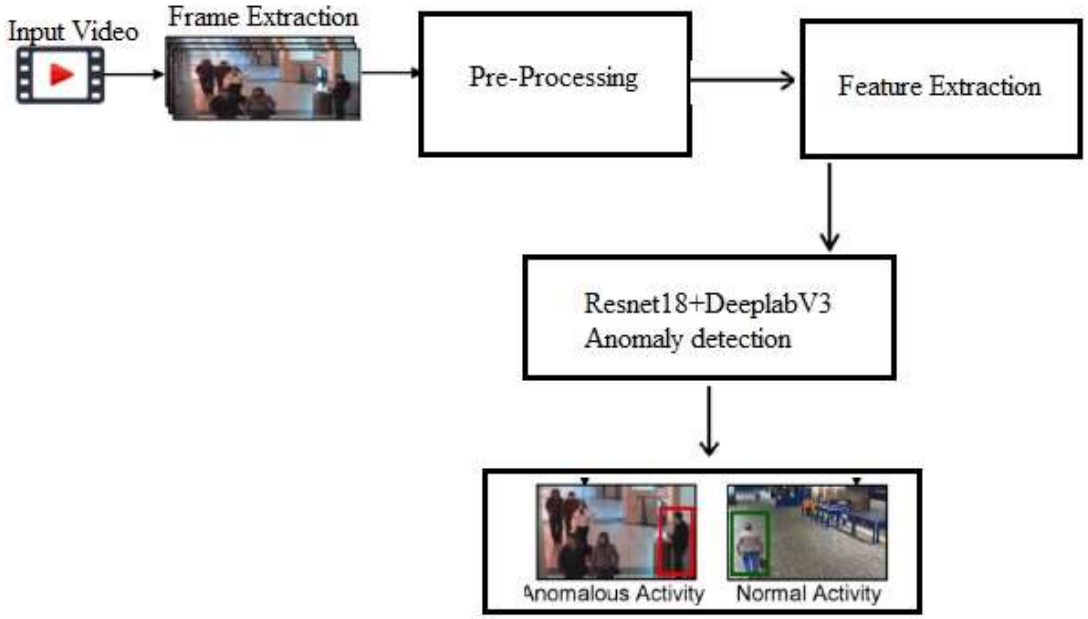


Figure1. Block Diagram of the Proposed method

i) Input Video

In this case, a digital camera or security camera was used to record the footage. After undergoing a number of enhancement procedures, the film is adapted so that it fits the requirements. The file might be an MP4 or AVI. After the video has been input, preprocessing must take place. Closed-circuit television, or CCTV, is a kind of video surveillance that records its surroundings. A continuous stream of pictures produced by the camera is sent into the

system is shown in figure2. The visual representation of a person may tell us a great deal about them. The videos used here are of 5 to 15 seconds duration with 30 frames/sec with a constant frame width of 1280 and height of 720 pixels. With size not more than 10MB rate.

Figure2. Video records

ii) Frame Extraction

The first stage is to edit the video footage. This requires taking the time to manually grab every frame from each video. After receiving video input, we scale each frame to 240 x 320 pixels. Multiple frames work together to create the video. FFmpeg is a tool for handling many forms of multimedia files. Using FFmpeg, we were able to split the movie into individual images. There will be 30 frames shown every second. The figure3 shows the frames that is extracted from the video.

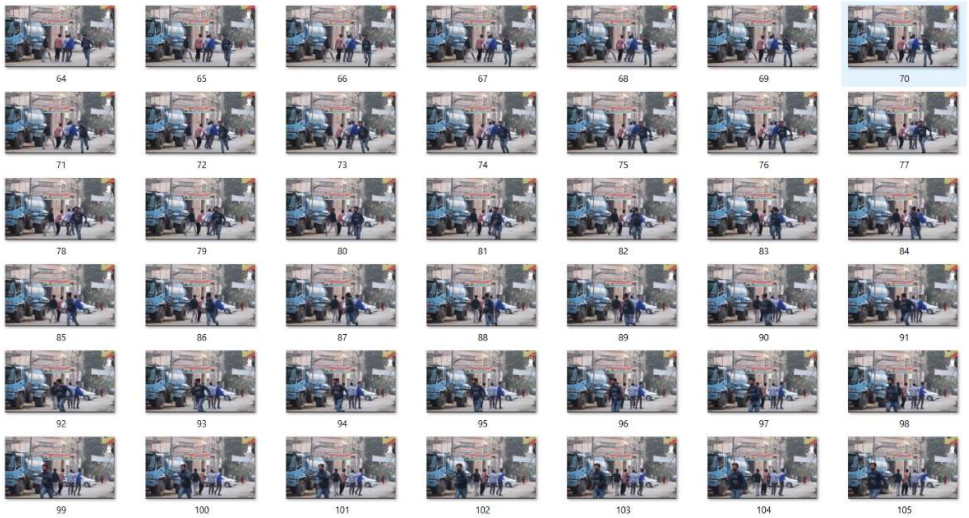


Figure3. Frames

iii) Pre-processing

Both sets of data underwent pre-processing in which color frames were taken from a stationary video camera and then run through a median filter to remove unwanted noise. One of the most well-known types of filters based on order statistics is the median filter. Due to the fact that it does not produce new unreasonable pixel values like when the filter crosses an edge, edge preservation is vital during noise reduction [5]. To choose which pixel to use as a replacement, the median filter sorts all neighboring pixels numerically using eq. 1.

$$I(x, y) = median\{f(s, t)\} \quad (1)$$

Medians often contain the value at the pixel location,(x,y). Following NR, we used Histogram Equalization (eq. 2) to fine-tune the contrast.

$$Sk = T(Rk) = (l-1)\sum_{j=0}^{k} \Pr(rj) \quad (2)$$

For each pixel in the input picture with intensity r, the resulting intensity level after intensity mapping is s, where k= 0, 1, 2,..., l 1. The variable r represents the intensity of the input image, while Pr symbolizes the probability density function (PDF) of r, as illustrated by using subscript r on p. The equalized output picture was created by transforming each input pixel with intensity Rk to a matching output pixel with level Sk.

iv) Feature Extraction

In this work, a videoswin transformer model is used to extract features; this model was trained on large-scale data sets like Kinetics or ImageNet. We can get higher-quality feature extraction by using a pre-trained model. Self-attention in traditional transformer models is computed with regard to all components, however this is computationally costly in the case of pictures. To address this problem, the swin transformer creates windows inside pictures and performs the necessary self-attention calculations within these bounds. Now it can more effectively acquire the self-attention value for all of the pictures by sliding the window across them.
The initial stage in feature extraction is framing the films (with dimensions H by W, for simplicity). Now we have a video for feature extraction, where each frame includes RGB channels, as represented by the set T of these frames. When we multiply the batch size, N, by the number of channels, C, we get the input dimension, N C T H W [7].

v) Resnet18+DeeplabV3 Anomaly detection

The identification of anomalous occurrences in general is given significant weight in the field of video anomaly detection. It's not unjust since single-type abnormal event identification often uses the same methods as broad abnormal event detection. Each anomalous positive video in the anomaly detection dataset contains several anomalous frames. All the clips in a typical negative video are going to be depressing. Therefore, common sense suggests that we should simply provide favorable training information for recognition of action [4].
Recently, deep learning has seen tremendous growth in the field of computer vision, with impressive successes in areas such as face recognition, picture categorization, behavior identification, and backdrop restoration. The primary distinction between deep learning and conventional approaches is that the former may eliminate the need for tedious human feature construction in favor of automated feature learning across several datasets. When it comes to data analysis and processing, deep learning may act like a human neural network. It can handle intricate surveillance scenarios thanks to a complicated model it can create to enhance characteristics' expression abilities (and hence their generalization powers). Typically, it has a unique deep structure & can correctly learn the intricate mapping between the inputs and the outputs [6].

K.M. He et al. presented ResNet in 2015 as a solution to the disappearance/explosion of gradients and performance loss due to depth. Furthermore, the ResNet is competent in classifying images. Through the use of residual units, ResNet is capable of constructing very deep networks [16, 17]. The ResNet-18 architecture is engrained in the area of computer vision & object identification [15] due to its eighteen deep layers.

$$x_{l+1} = f[xl + f(xl, kt)] \quad (3)$$

Here, f is the operator, F is the residual function, k is the convolution kernel, and xl and xl+1 are the input & output of the l-th residual unit.

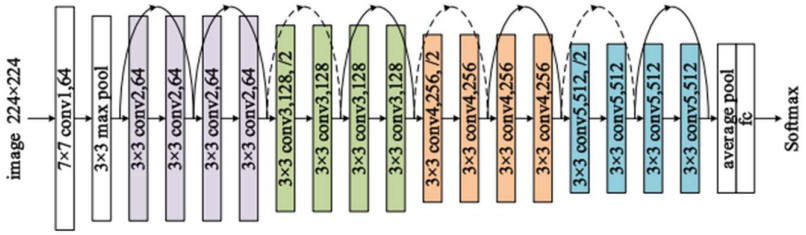


Figure4. ResNet-18 Architecture

Figure 4 shows that ResNet's input picture size is 224x224, and that its convolution kernel size for the first convolution layer is 7x7 and that it is 3x3 for the other layers. After complete connection is used to produce an eigenvector from the final convolution layer's average pooled feature map, the classification probability is normalized using Softmax. For a given number of filters, the convolution layer will always produce a feature map of the same size [19].

The Deeplab-V3 network is supported by Xception. Deeplab-V3 network incorporates Encoder-Decoder Structure, a feature map generated in the Encoder, and uses concatenation for the next step of fusion to accomplish the fusion of shallow feature details and a deep semantic features [14]. Xception's residual connection mechanism similar to ResNet substantially speed up the convergence of Xception, resulting in better deep feature extraction and more powerful performance.

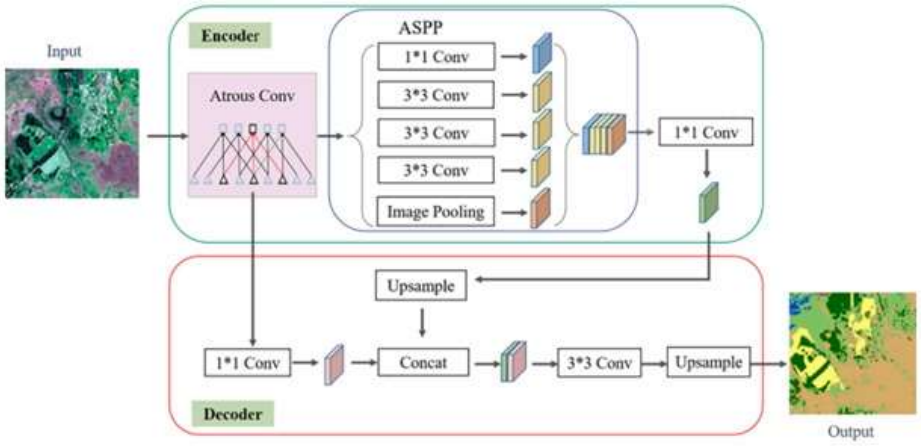


Figure5. DeepLabV3 Architecture.

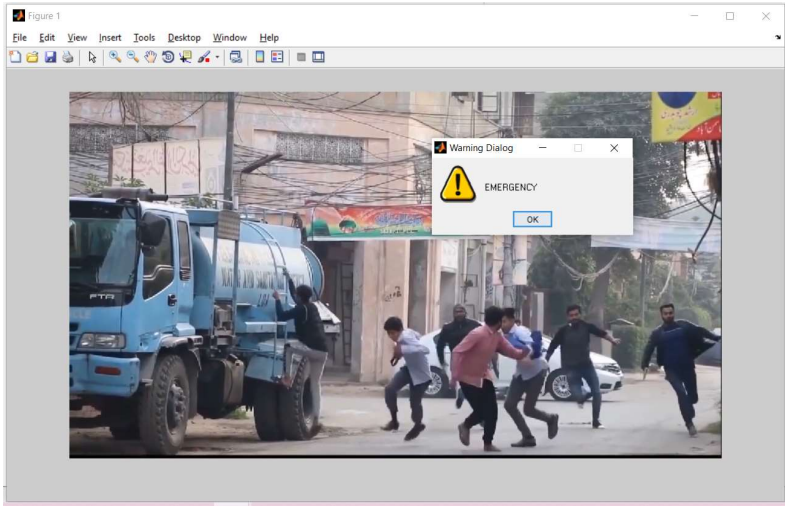
To extract an image's characteristics, DeepLabV3+ uses a convolutional layer with a decreasing and increasing number of weights, as seen in figure 5. After the input image's

essential properties have been stripped away by the encoder network, the decoder network will progressively restore the data in order to make an estimate of the object type for each pixel. For precise region-boundary estimation, the decoder additionally directly uses low-level characteristics. Reduce the number of weights in the model by using the ASPP (Atrous Spatial Pyramid Pooling) structure or depthwise separable convolution [18, 10] to create a deep network with low computational requirements.

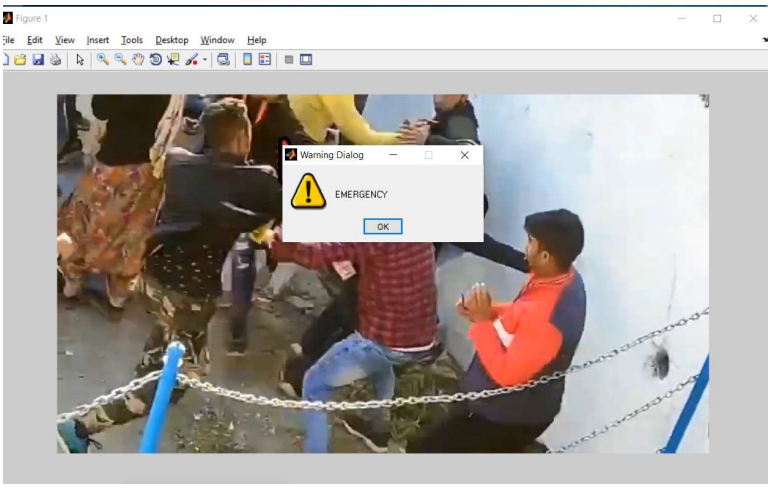
Because of its shallowness, ResNet-18 achieves results comparable to those of other ResNets while retaining a greater number of low-scale characteristics. Image features are extracted first using the convolution layer's feature extraction procedure, and then their size is decreased using a maximum pool layer to avoid over fitting and accelerate the computation time. Finally, the completely connected layer receives the recovered feature information for classification [20], and the residual structure is used to extract the high-level features in order to avoid gradient explosion. To this end, we use a pre-trained version of ResNet-18 as the feature extractor (encoder) for our system model and employ depth-wise separable convolution in DeepLabV3 to minimize the model's dimensionality.

vi) Results

The identification of anomalies in about 250 short films included testing on around 200 movies with anomalous activity and 50 videos with regular motion. Videos range in length from 5 to 20 seconds with a frame rate of 30 frames per second. Online footage was obtained and edited down to a more manageable length using a video editor program.



(a)



(b)

Figure6. Anomaly detected (a) people chasing and (b) gang fight

The figure6 shows the output result that shows the warning when Anomaly is detected. The figure6 (a) shows that the people are chasing someone as an anomaly and figure6 (b) shows the warning when a fight breaks out as an Anomaly.

where recall quantifies how many pixels from the moving item were properly identified, and accuracy quantifies how many pixels from the moving object were really detected. We may characterize them as

$$Recall = TPR = \frac{TP}{FN+TP} \quad (4)$$

$$\text{Precision} = \frac{TP}{FP+TP} \quad (5)$$

Although accuracy is a common measure in classification issues, it is seldom used to evaluate anomaly detection methods. We provide it for completeness' sake, but it will not factor into our decisions.

$$Accuracy = \frac{TP + TN}{TN + FP + TP + FN} \quad (6)$$

Table I. Frame-wise confusion matrix for five sample videos

Video	Methodology	TP	TN	FP	FN
Video 1	PSO-CNN	174	41	14	9
	Resnet18-Deeplabv3	179	50	5	4
Video 2	PSO-CNN	229	86	21	27
	Resnet18-Deeplabv3	244	93	14	12
Video 3	PSO-CNN	112	114	18	9
	Resnet18-Deeplabv3	116	121	11	5
Video 4	PSO-CNN	366	137	17	28
	Resnet18-Deeplabv3	378	146	8	16

Video 5	PSO-CNN	276	99	15	25
	Resnet18-Deeplabv3	287	108	6	14

Table II. Frame-wise validation parameters for five sample videos

	Video	Precision	Recall	F1-score	Accuracy	Specificity
Video 1	PSO-CNN	92.55	95.08	93.80	90.34	74.55
	Resnet18-Deeplabv3	93.65	96.72	95.16	92.44	78.18
Video 2	PSO-CNN	91.60	89.45	90.51	86.78	80.37
	Resnet18-Deeplabv3	93.28	92.19	92.73	89.81	84.11
Video 3	PSO-CNN	86.15	92.56	89.24	89.33	86.36
	Resnet18-Deeplabv3	89.06	94.21	91.57	91.70	89.39
Video 4	PSO-CNN	95.56	92.89	94.21	91.79	88.96
	Resnet18-Deeplabv3	97.13	94.42	95.75	93.98	92.86
Video 5	PSO-CNN	94.85	91.69	93.24	90.36	86.84
	Resnet18-Deeplabv3	96.22	93.02	94.59	92.29	90.35

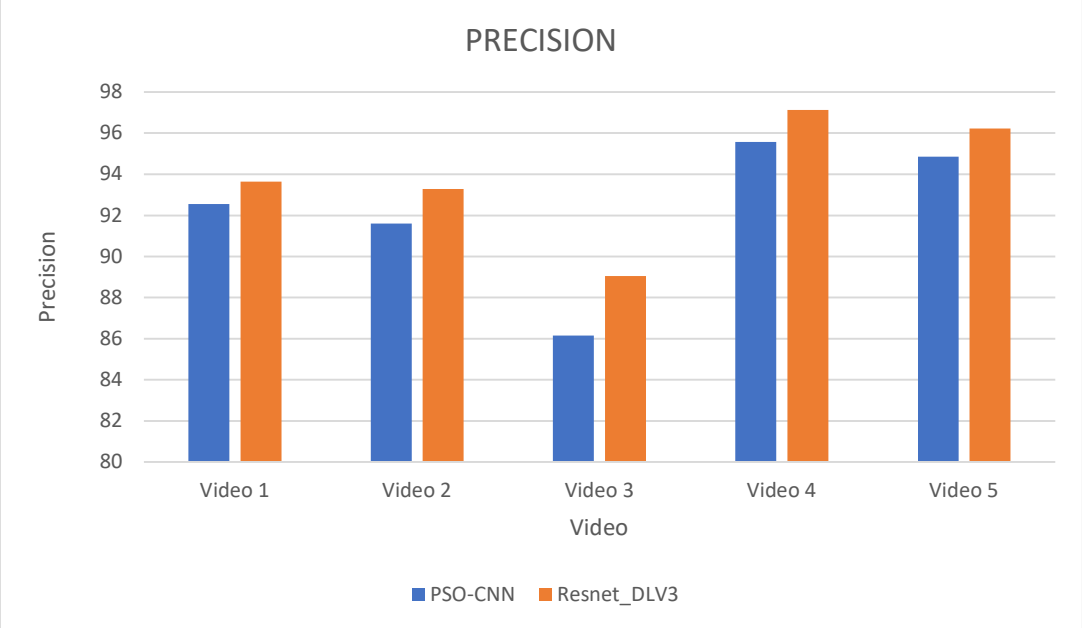


Figure7. Precision value for five sample videos

Figure7 shows the Precision value for five sample videos when using PSO-CNN and Resnet18-Deeplabv3 method. From the Precision result the Resnet18-Deeplabv3 shows high Precision value the PSO-CNN. The highest Precision value using Resnet18-Deeplabv3 is 97.13%.

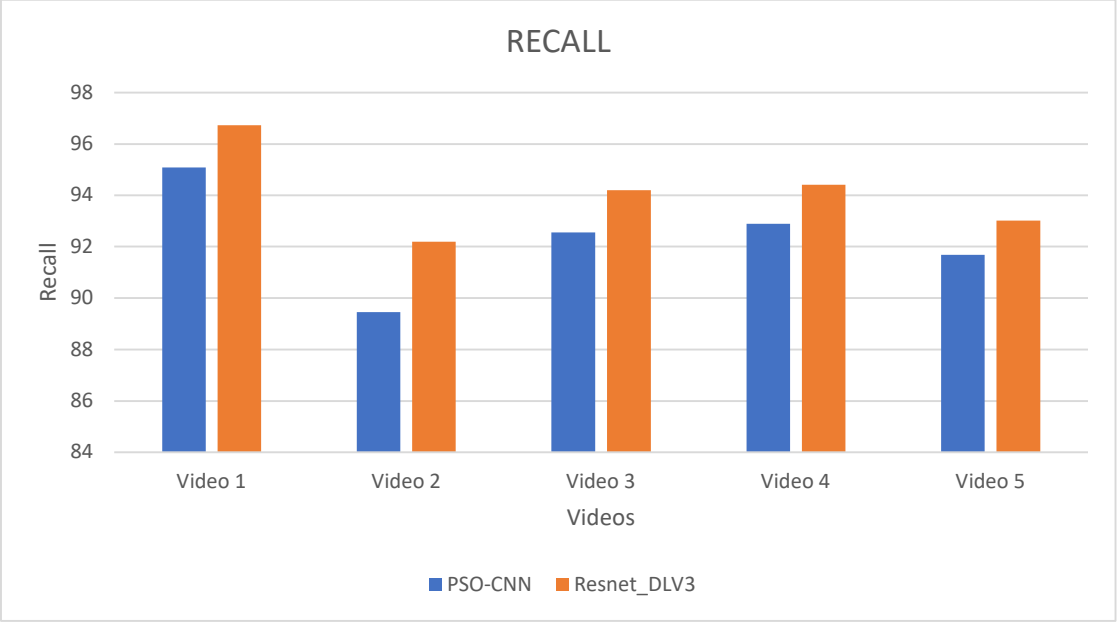


Figure8. Recall value for five sample videos

Figure 8 depicts the Recall value for five different videos using the PSO-CNN and Resnet18-Deeplabv3 methods. The Resnet18-Deeplabv3 has a high Recall value when compared to the PSO-CNN. Resnet18-Deeplabv3 had the greatest Recall value of 96.72%.

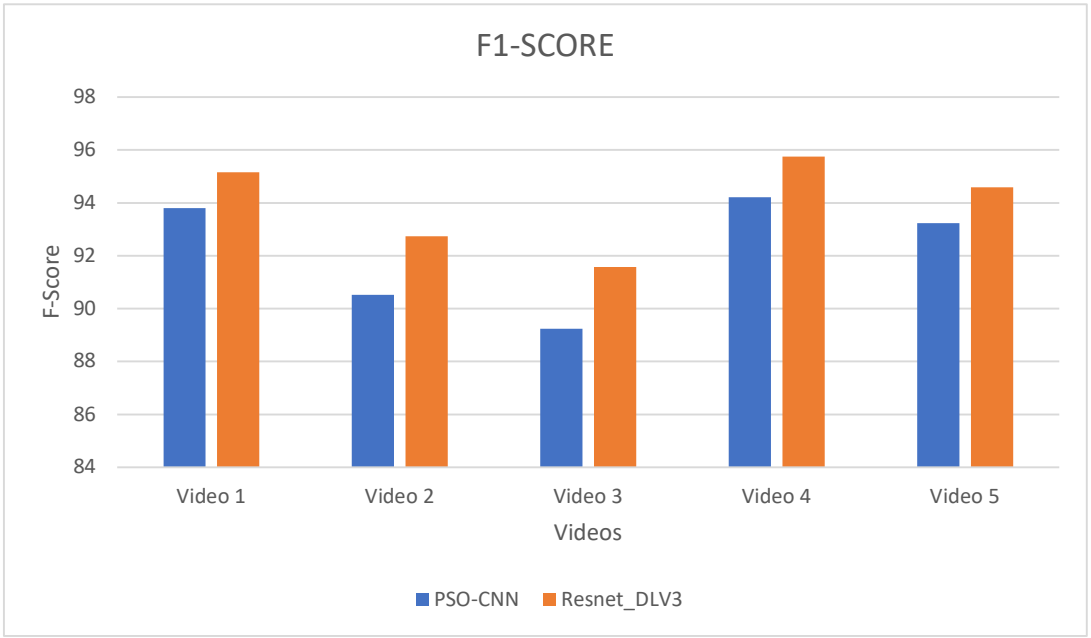


Figure9. F1-score value for five sample videos

Figure9 shows the F1-score value for five sample videos when using PSO-CNN and Resnet18-Deeplabv3 method. From the F1-score result the Resnet18-Deeplabv3 shows high F1-score value the PSO-CNN. The highest F1-score value using Resnet18-Deeplabv3 is 95.75%.

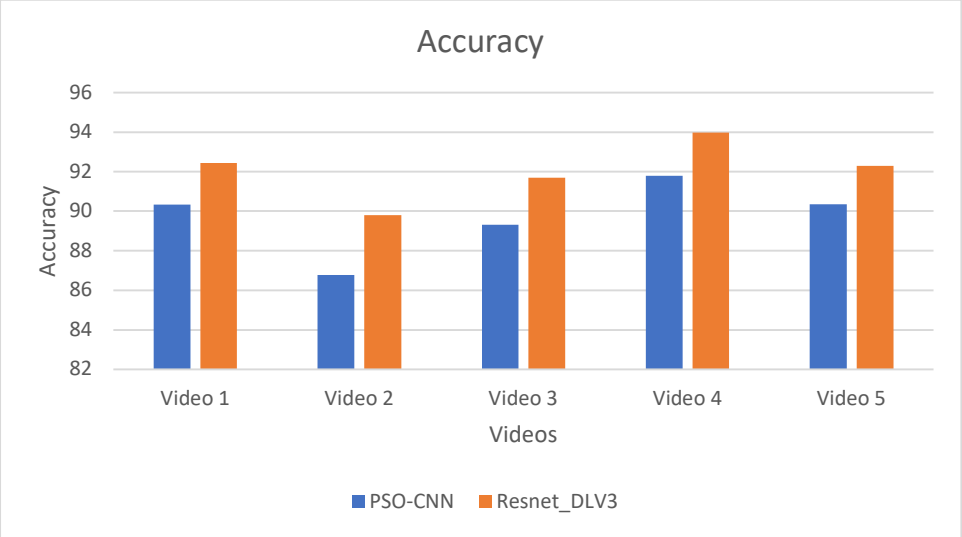


Figure10. Accuracy value for five sample videos

Figure 10 depicts the Accuracy value for five different films using the PSO-CNN and Resnet18-Deeplabv3 methods. According to the Accuracy result, the Resnet18-Deeplabv3 has a high Accuracy value for than PSO-CNN. Resnet18-Deeplabv3 has the highest Accuracy value of 93.98%.

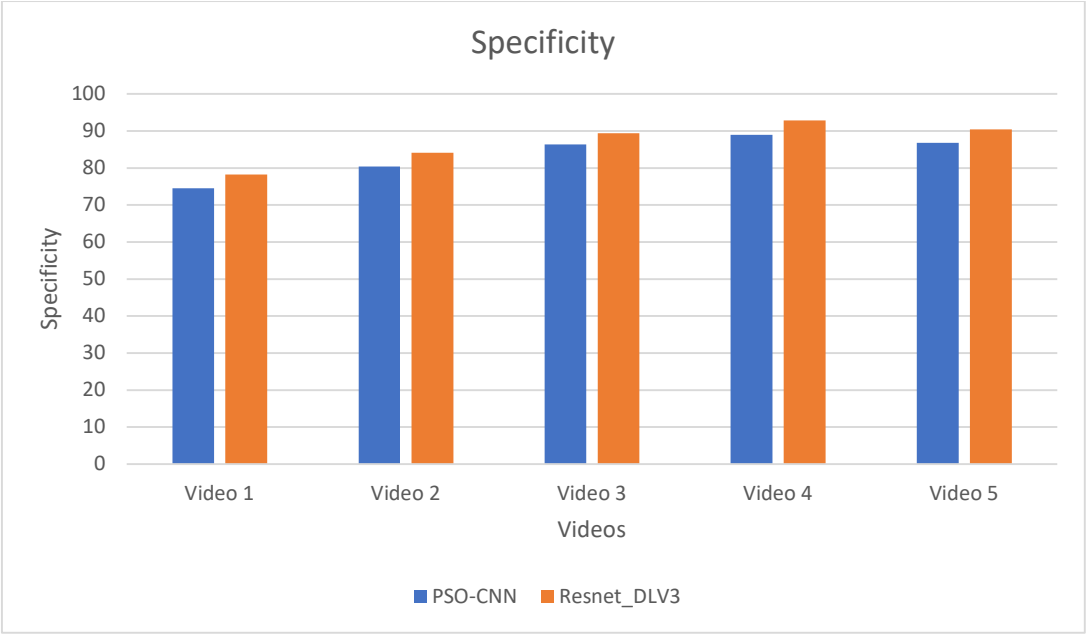


Figure11. Specificity value for five sample videos

Figure11 shows the Specificity value for five sample videos when using PSO-CNN and Resnet18-Deeplabv3 method. From the Specificity result the Resnet18-Deeplabv3 shows high Specificity value the PSO-CNN. The highest Specificity value using Resnet18-Deeplabv3 is 92.86%.

Table III. Confusion matrix with respect to videos

Video	TP	TN	FP	FN
PSO-CNN	181	19	8	42
Resnet18-Deeplabv3	185	15	7	43

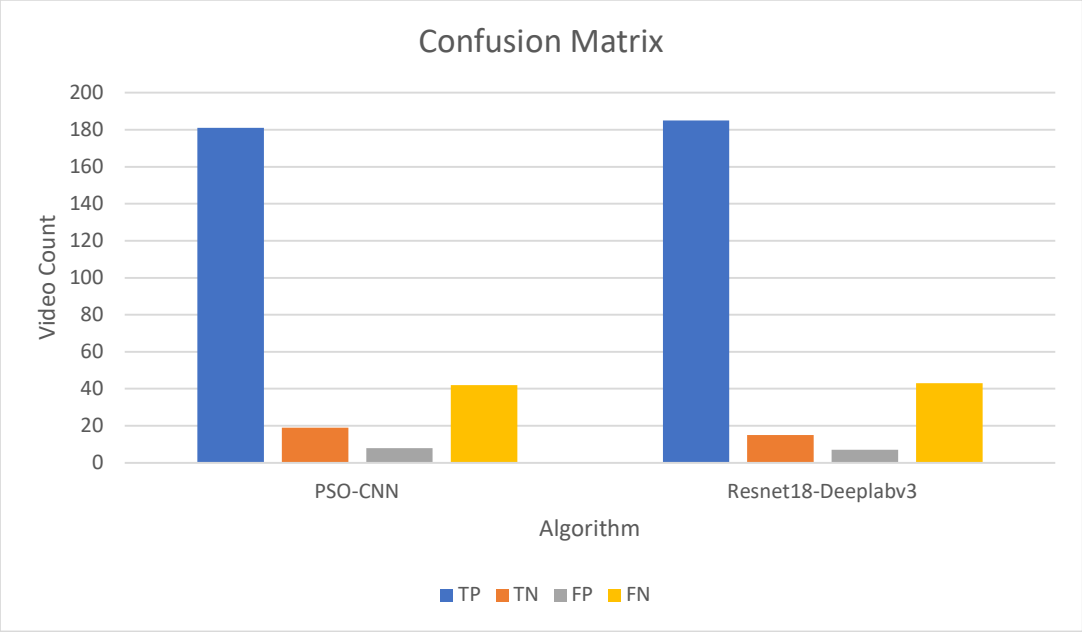


Figure 12. Comparison of confusion matrix parameters of existing and proposed

Table IV. Validation parameters with respect to videos

Video	PRECISION	RECALL	FSCORE	Accuracy	Specificity
PSO-CNN	90.50	95.77	93.06	89.20	68.85
Resnet18-Deeplabv3	92.50	96.35	94.39	91.20	74.14

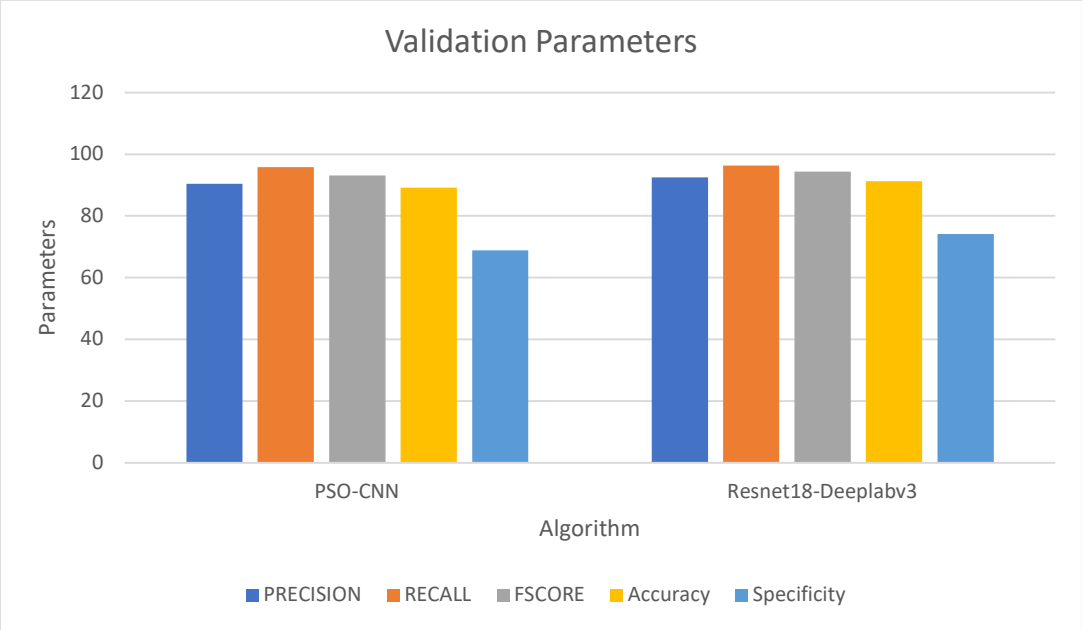


Figure 13. Comparison of validation parameters of PSO-CNN and Resnet18 _ deeplabV3

Figure 12 and 13 shows the confusion matrix of proposed work I and II where the proposed work II with Resnet18 and Deeplabv3 shows a slight higher performance in all aspects compared to PSO CNN

Table V. Comparison of accuracy oef existing and proposed methods

Method	**Accuracy**
I3d-resnet50 +CNN	78.78
online Kalman filtering (OKF)	81.85
PCA + ANN	77.78
IMTSL+CNN	82
RPCA-MFTSL AND PSO-CNN	86
RPCA-MFTSL and Resnet 18-Deeplabv3+	91.2

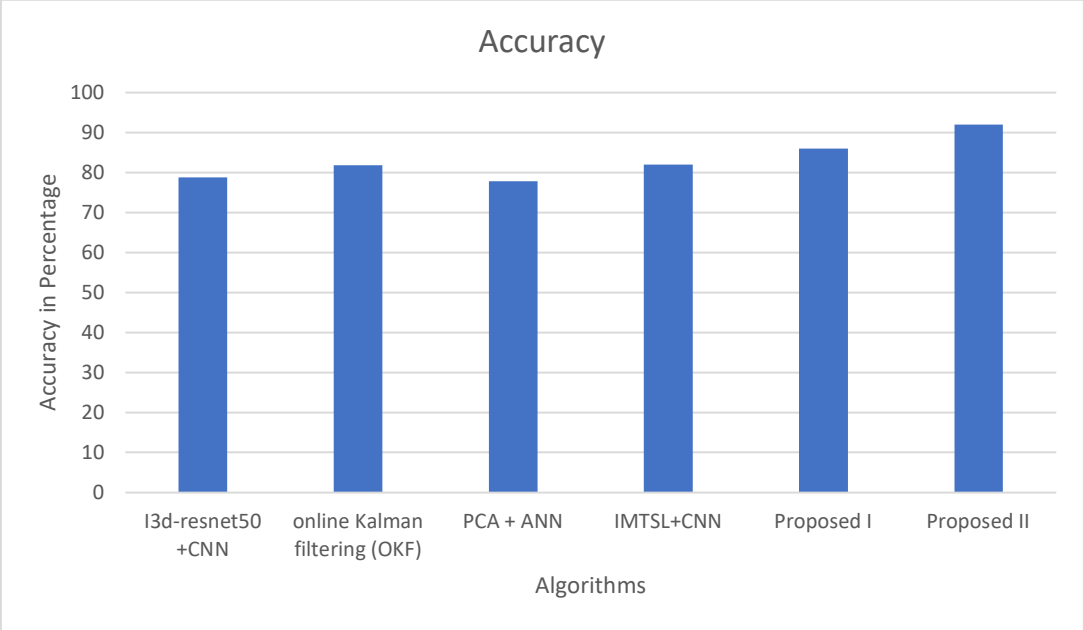


Figure14. Accuracy results using various algorithms

Figure14 shows the Accuracy value using various algorithms methods. From the Accuracy results our proposed method RPCA-MFTSL and Resnet18-Deeplabv3 shows high accuracy value than other methods The highest accuracy value using RPCA-MFTSL and Resnet18-Deeplabv3 is 92.04%.

IV. CONCLUSION

Video surveillance is a useful tool for modern organizations of all sizes to monitor productivity, keep an eye on employees, and prevent theft and other illegal activities. In this research, we implement a unique algorithm into video surveillance systems in order to identify anomalies. To improve performance and achieve high accuracy, we employed the Resnet18+DeeplabV3 anomaly detection method. Using real-world video footage, we show how our methods can automatically tell the difference between typical and unusual activities, a useful tool for security cameras. Based on the accuracy results, our proposed methods using Resnet18-Deeplabv3 have a higher accuracy value than other approaches. The best accuracy value we obtained using Resnet18-Deeplabv3 is 91.2 %. For future research, we are going to improve our anomaly detection to get more performance and higher accuracy using other methods.

V. REFERENCES

1. Qasim, M., &Verdu, E. (2023). Video anomaly detection system using deep convolutional and recurrent models. *Results in Engineering*, *18*, 101026.
2. Wang, Z., Zhang, Y., Luo, L., & Wang, N. (2022). AnoDFDNet: a deep feature difference network for anomaly detection. *Journal of Sensors*, *2022*.
3. Tur, A. O., Dall'Asen, N., Beyan, C., & Ricci, E. (2023). Exploring Diffusion Models for Unsupervised Video Anomaly Detection. *arXiv preprint arXiv:2304.05841*.

4. Tan, W., & Liu, J. (2023, February). Detection of fights in videos: A comparison study of anomaly detection and action recognition. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V* (pp. 676-688). Cham: Springer Nature Switzerland.
5. Abdullah, F., Javeed, M., & Jalal, A. (2021, November). Crowd Anomaly Detection in Public Surveillance via Spatio-temporal Descriptors and Zero-Shot Classifier. In *2021 International Conference on Innovative Computing (ICIC)* (pp. 1-8). IEEE.
6. Chen, F., Wang, W., Yang, H., Pei, W., & Lu, G. (2022). Multiscale feature fusion for surveillance video diagnosis. *Knowledge-Based Systems*, *240*, 108103.
7. Deshpande, K., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2023, April). Anomaly detection in surveillance videos using transformer based attention model. In *Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part VII* (pp. 199-211). Singapore: Springer Nature Singapore.
8. Niu, W., Long, J., Han, D., & Wang, Y. F. (2004, June). Human activity detection and recognition for video surveillance. In *2004 IEEE international conference on multimedia and expo (ICME)(IEEE Cat. No. 04TH8763)* (Vol. 1, pp. 719-722). IEEE.
9. Ayers, D., & Shah, M. (2001). Monitoring human behavior from video taken in an office environment. *Image and Vision Computing*, *19*(12), 833-846.
10. Park, G., Lee, M., Jang, H., & Kim, C. (2021). Thermal anomaly detection in walls via CNN-based segmentation. *Automation in Construction*, *125*, 103627.
11. Kopuklu, O., Zheng, J., Xu, H., &Rigoll, G. (2021). Driver anomaly detection: A dataset and contrastive learning approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 91-100).
12. Liu, Z., Wu, X. M., Zheng, D., Lin, K. Y., & Zheng, W. S. (2023). Generating Anomalies for Video Anomaly Detection With Prompt-Based Feature Mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 24500-24510).
13. Sun, S., & Gong, X. (2023). Hierarchical Semantic Contrast for Scene-aware Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22846-22856).
14. Quan, B., Liu, B., Fu, D., Chen, H., & Liu, X. (2021, August). Improved deeplabv3 for better road segmentation in remote sensing images. In *2021 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)* (pp. 331-334). IEEE.
15. Shukla, P., & Kumar, M. Explicating ResNet for Facial Expression Recognition.
16. Guo, M., & Du, Y. (2019, October). Classification of thyroid ultrasound standard plane images using ResNet-18 networks. In *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)* (pp. 324-328). IEEE.
17. Schurischuster, S., &Kampel, M. (2020, November). Image-based classification of honeybees. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1-6). IEEE.
18. Liu, M., Fu, B., Xie, S., He, H., Lan, F., Li, Y., ... & Fan, D. (2021). Comparison of multi-source satellite images for classifying marsh vegetation using DeepLabV3 Plus deep learning algorithm. *Ecological Indicators*, *125*, 107562.

19. Ou, X., Yan, P., Zhang, Y., Tu, B., Zhang, G., Wu, J., & Li, W. (2019). Moving object detection method via ResNet-18 with encoder–decoder structure in complex scenes. *IEEE Access*, *7*, 108152-108160.
20. Zhang, H., & Wang, F. (2022, May). Fault identification of fan blade based on improved ResNet-18. In *Journal of Physics: Conference Series* (Vol. 2221, No. 1, p. 012046). IOP Publishing.