International Journal of Safety and Security Engineering
Vol. 13, No. 6, December, 2023, pp. 1143-1151
Journal homepage: http://iieta.org/journals/ijsse

A Robust Multi Descriptor Fusion with One-Class CNN for Detecting Anomalies in Video Surveillance

K. Chidananda[1*], A.P. Siva Kumar[2]

[1] Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Ananthapuramu 515002, A.P., India
[2] Department of Computer Science and Engineering, JNTUA College of Engineering, Ananthapuramu 515002, A.P., India

Corresponding Author Email: chida.koudike@gmail.com

https://doi.org/10.18280/ijsse.130618

Received: 19 October 2023
Revised: 27 November 2023
Accepted: 15 December 2023
Available online: 25 December 2023

Keywords:
Video Anomaly Detection (VAD), computer vision, machine learning, feature extraction, dimensionality reduction, One-Class CNN, Multiple Feature Descriptor, spatiotemporal information

ABSTRACT

In the domain of computer vision and machine learning, Video Anomaly Detection (VAD) has emerged as a pivotal area of inquiry, particularly relevant to security, surveillance, and video analytics. The extraction of pertinent features from video data constitutes a foundational aspect of VAD, enabling the discernment of anomalous patterns and structures. Features such as motion, texture, shape, and aesthetics are extracted and tailored according to the exigencies of the application and the intrinsic properties of the video content. The amalgamation of multiple features is often requisite for refining the accuracy of anomaly detection systems. Given the inherent high dimensionality of video data, dimensionality reduction techniques have been employed to mitigate computational demands and enhance the precision of the anomaly detection process. The present study delineates a novel approach centered on the deployment of a One-Class Convolutional Neural Network (CNN). This network is exclusively trained on normal events to establish a baseline representation of typicality. During the evaluation phase, the network is tasked with predicting the normality or abnormality of new video segments against this established norm. Moreover, this work introduces a novel fused feature descriptor, referred to as the Multiple Feature Descriptor (MFD), which is designed to encapsulate the spatiotemporal attributes of video data effectively. The proposed methodology has been subjected to rigorous testing against publicly available datasets, where it has demonstrated superior performance, outstripping numerous contemporary state-of-the-art methods in both accuracy and computational efficiency.

1. INTRODUCTION

Video Anomaly Detection (VAD) has cemented itself as an indispensable task within the realms of computer vision and surveillance [1]. The principal objective of VAD is the automated differentiation of abnormal actions or behaviors within surveillance footage. Anomalies of interest span a broad spectrum, from acts of theft and vandalism to emergent situations such as fires or accidents. The criticality of VAD is anchored in its utility for bolstering public safety, fortifying security measures, and preempting criminal activity. The capability of VAD systems to identify and flag anomalous events in real-time confers upon security operatives the advantage of prompt response to potential threats, thereby upholding public safety [1]. Moreover, the scope of VAD transcends public security, finding relevance in domains such as industrial monitoring, traffic management, and medical diagnostics, where the early detection of irregularities is paramount. In light of the proliferation of video surveillance systems across a plethora of public spaces-airports, train stations, shopping centers, and more-the demand for automated systems capable of vigilant anomaly detection has surged. The traditional approach of manually sifting through surveillance footage is not only labor-intensive and costly but also prone to reliability issues, highlighting the imperative for sophisticated automated anomaly detection systems. These systems are paramount for the realization of surveillance that is both efficient and effective [2].

The significance of Video Anomaly Detection (VAD) in the field of video analytics has escalated, driven by its potential for the automatic identification of abnormal occurrences and potential hazards within public spaces [3]. The automated analysis of surveillance footage stands as a cornerstone for activity tracking and recognition, a function that becomes critical in environments where human surveillance is limited. Despite this, the precise detection of anomalies within video streams presents a considerable challenge, particularly due to the complexity involved when subjects traverse the fields of view of multiple cameras [4]. VAD systems have been extensively implemented on a global scale, augmenting public safety by enabling the detection of irregular activities such as altercations, theft, vehicular mishaps, and other criminal acts [5]. The urgency for robust and effective VAD methodologies has intensified in parallel with the expanding reliance on automated surveillance systems. Illustrative examples of these applications are depicted in Figure 1, showcasing frames from

various datasets that are analyzed within this study.

In the realm of video surveillance, the efficacy of anomaly detection systems is paramount for ensuring public safety and security. Figure 1 illustrates eight instances of anomalies identified in various surveillance scenarios, each highlighting the critical nature of such systems [6]. Examples a and b, sourced from the UCSD dataset [7], depict anomalies in the form of a truck inappropriately traversing a footpath and a pedestrian straying onto a lawn, respectively. Examples c and d, from the Avenue dataset [8], portray a person engaged in the act of throwing an object and another individual carrying a suspicious bag. The fifth instance, example e from the MDVD dataset [9], captures a vehicle parked incorrectly, while example f from the same dataset illustrates an altercation between individuals. The final examples, g and h, extracted from the ShanghaiTech dataset [10], present a person seizing a suitcase and traffic moving on a pavement, respectively.

The process of anomaly detection is underpinned by the critical role of feature selection and extraction. It is through the discernment of features such as motion, texture, shape, and color that relevant information is culled from video frames. The ability of anomaly detection algorithms to distinguish between normal and abnormal behavior hinges on the robustness of these features. Techniques such as optical flow, histogram of oriented gradients (HOG), local binary patterns (LBP), and deep neural networks (DNNs) are employed for feature extraction. Thus, the judicious selection and accurate extraction of features are indispensable to the performance of Video Anomaly Detection systems, directly influencing their accuracy and efficiency.

Variability in the dimensionality of Video Anomaly Detection (VAD) datasets is often observed, influenced by the nature of the dataset and the plethora of features extracted from video frames. High-dimensional data are typified in VAD datasets, reflective of the extensive volume of frames and their extracted features. The number of features, contingent upon the extraction methodology employed and the video's inherent attributes, can significantly differ. Furthermore, certain datasets may exhibit increased dimensionality due to the complexity or heterogeneity of the anomalies they contain. Therefore, the dimensionality is a pivotal factor in the selection and application of feature extraction and dimensionality reduction techniques.

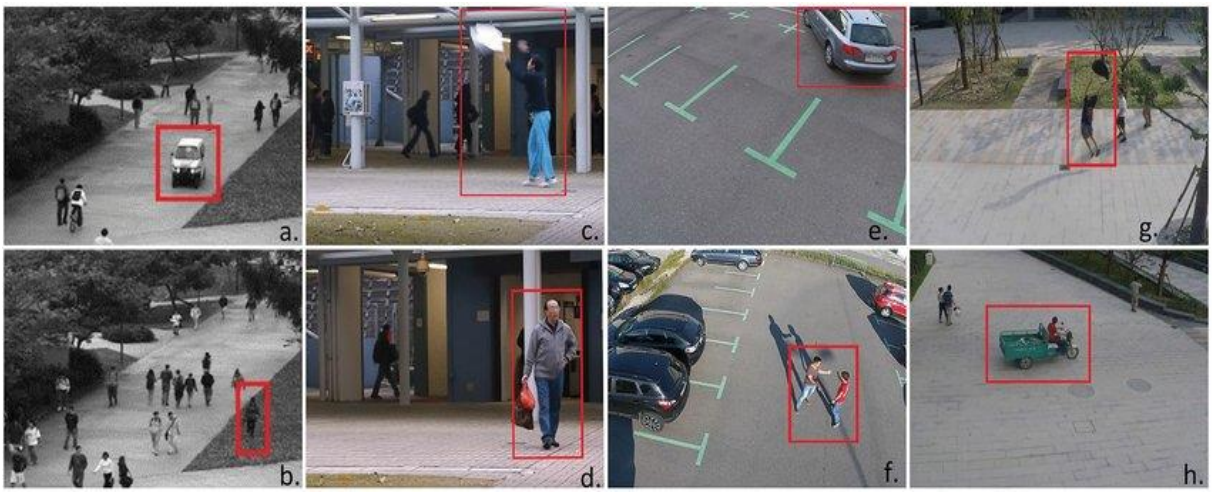

Figure 1. An illustration of a single surveillance video frame [6]

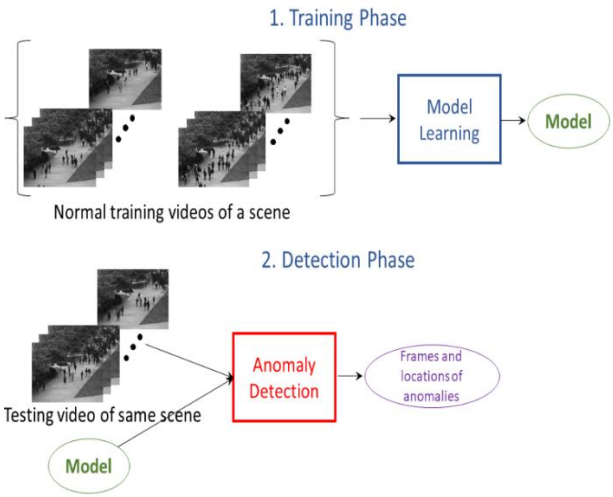

Figure 2. Implementation of proposed work

In the current investigation, a novel approach employing a multi-descriptor fusion alongside a One-Class Convolutional Neural Network (OC-CNN) [11] has been introduced for the purpose of anomaly detection within video surveillance frameworks. The OC-CNN, a variant of the Convolutional Autoencoder (CAE), is exclusively trained on regular samples, enabling it to reconstruct the normal patterns present in the input data. The objective of the OC-CNN is to encapsulate the quintessential structure of normal patterns, thereby utilizing this knowledge to identify anomalies. A fused feature descriptor, combining Histogram of Oriented Gradients (HOG) [12] and Local Binary Patterns (LBP) [13], has been employed to detect anomalous events in surveillance footage effectively. Termed HOG-LBP, this descriptor amalgamates gradient and textural features to garner robust and distinctive feature representations. Specifically, HOG encapsulates the image's gradient information, while LBP is responsible for textural detail. The synergy of these descriptors enhances the robustness and discriminative power of the feature representation, beneficial for anomaly detection. The methodology presented in this paper unfolds in two distinct phases, as delineated in Figure 2. The training phase incorporates the HOG-LBP with the OC-CNN to foster the learning of normal patterns. The ensuing sections of the paper are organized as follows: Section II delves into the related

work, providing a comprehensive background. A detailed exposition of the HOG-LBP feature descriptor is presented in Section III. Section IV delineates the implementation details, and the paper concludes with Section V, summarizing the study's findings.

2. RELATED WORK

Dollár et al. [14] presents an approach to recognize human behaviors in video sequences using sparse spatio-temporal features. The authors propose a method that selects key spatio-temporal interest points in the video sequence and extracts local descriptors around these points. The spatio-temporal characteristics of the behavior are then represented using the descriptors. On the KTH human action dataset, the suggested method is assessed and found to perform at the cutting edge at the time of publishing. A deep learning (DL) method for identifying anomalous occurrences in films is suggested by Xu et al. [2]. Based on how much the films deviate from the learnt usual behavior, the learned representations are then utilized to categorize the movies as either normal or anomalous. The method is assessed using various datasets and contrasted with other cutting-edge techniques. The findings demonstrate that the suggested method performs superior than the other approaches in terms of accuracy and is able to identify a variety of abnormal events.

Zhu and Newsam [15] propose a new feature extraction method for VAD based on motion-awareness. The authors argue that conventional methods that rely solely on appearance-based features may fail to detect anomalies in videos with complex motion patterns. To report this issue, they have proposed a novel feature extraction approach that combines appearance-based features with motion-aware features. The motion-aware features are obtained by analyzing the optical flow patterns in the video, and they provide a complementary representation of the motion information that is not captured by the appearance-based features. The findings of the authors' comprehensive tests on a number of benchmark datasets show that the suggested approach beats cutting-edge techniques in terms of detection precision and resilience to complicated motion patterns.

In order to extract both spatial and temporal information from video frames, Ullah et al. [16] suggested. ‘s technique combines CNN features with bi-directional LSTM networks. Before feeding the bi-directional LSTM network, the CNN features are first retrieved from the video frames. By processing the video sequence both forward and backward, the bi-directional LSTM network learns to recognize the temporal connections between the frames. A probability distribution across the potential action classes is the network's ultimate output. The suggested technique performs at the cutting edge when tested against numerous benchmark datasets for action recognition. The results show how well bidirectional LSTM networks and CNN features work together to capture spatial and temporal information in video sequences. The method may be used in robotics, human-computer interaction, and video surveillance.

A strategy that makes use of DL methods to find anomalous occurrences in films was put out by Chong and Tay [17]. The technique uses a spatiotemporal autoencoder to recognize the characteristics of typical occurrences in a video and identify abnormal ones. Training and testing are the two aspects of the approach. The spatiotemporal autoencoder is trained on data from typical occurrences during the training phase in order to recognize the characteristics of typical events. The trained spatiotemporal autoencoder is utilized to identify anomalous occurrences during the testing phase. The reconstruction error is determined after the autoencoder has processed a set of frames. The sequence is labelled abnormal if the reconstruction error surpasses a specific threshold.

An approach for summarizing movies from IoT surveillance networks using deep CNNs and hierarchical weighted fusion is suggested by Muhammad et al. [18] Naik and Thimmaiah [19], Mohamed et al. [20], Bangare and Patil [21]. The suggested solution seeks to retain high accuracy in video summarization while lowering the high computational cost of processing video data in IoT surveillance networks. The technique employs hierarchical weighted fusion to choose significant frames for summarization after using a pre-trained CNN to mine features since from video frames.

By studying the temporal regularities of the sequences, Hasan et al. [22] provide a technique for identifying abnormalities in video sequences. A DNN is used in the technique, which trains itself to anticipate the subsequent frame in a video series providing the previous frames. The network can spot anomalies when the predicted frame considerably differs from the actual frame since it was trained on regular video sequences. To anticipate the next frame in a video series, a deep neural network with an encoder-decoder architecture is trained. The encoder network, which encodes the input frames into a lower-dimensional representation, is a CNN. The decoder network is a deconvolutional neural network that reconstructs the input frames from the encoded representation. To reduce the discrepancy between the expected and real frames, the network is trained.

Current advances in "computer vision, machine learning, and DL have enabled the development of robust and accurate VAD systems. These systems use various techniques, such as feature extraction, dimensionality reduction, and anomaly detection algorithms, to analyse and identify abnormal events or behaviours in surveillance videos. There have been several recent advances in VAD, including:

a. **Deep Learning:** Deep neural networks have shown remarkable performance in detecting anomalies in video surveillance. They can automatically learn complex features and representations from raw data, allowing for more accurate detection of anomalies.
b. **Unsupervised Learning:** Autoencoders and generative adversarial networks (GANs) are examples of unsupervised learning approaches that have been used to identify abnormalities in video surveillance without the requirement for labelled data.
c. **Multi-modal Fusion:** Combining information from multiple modalities, such as video, audio, and depth, can progress the accurateness of anomaly detection in complex scenes.
d. **Transfer Learning:** Transfer learning techniques have been used to transfer knowledge learned from one dataset to another, improving the performance of anomaly detection in datasets with limited labelled data.
e. **Graph Convolutional Networks (GCNs):** GCNs have been used to learn spatiotemporal dependencies among the video frames, allowing for more accurate detection of anomalies in complex scenes.
f. **Attention Mechanisms:** Attention mechanisms have

been used to selectively focus on important regions of the video frames, refining the precision of anomaly detection.

These recent advances have shown promising results in enlightening the exactness of VAD and have opened new avenues for research in this field. Whereas the usage of multiple descriptors like HOG and LBP will be an added advantage for the well-known ON-CNN kind of architectures. This paper has considered the advantages of prominent dimensionality reduction techniques like HOG and LBP and made it as fusion technique for reducing the high dimension. One class convolutional neural network is also a prominent technique, which could be easily classifying that the action in the video frame is of normal or abnormal action.

3. MFD AND OC-CNN

Multiple Feature Descriptors are commonly used in VAD to capture different aspects of the video data. This approach involves combining multiple types of features to obtain a more robust and comprehensive representation of the video data. Examples of feature descriptors that can be combined include motion features, texture features, shape features, and color features. By combining Multiple Feature Descriptors, the resulting feature vector can capture a more diverse set of characteristics, allowing for better detection of anomalies in complex and dynamic scenes. The fusion of feature descriptors can be done at different levels, such as early fusion where features are concatenated before classification, or late fusion where multiple classifiers are used to combine the results of each feature descriptor.

HOG and LBP are two prevalent feature descriptors considered in computer vision and image processing tasks, including VAD (Figure 3 and Figure 4). HOG is a feature descriptor that calculates the gradient orientation and magnitude in local image regions. The gradient orientations are quantized into histograms to create a feature vector representation of the local image region. HOG has been shown to be effective in detecting human shapes and movements in surveillance videos, which can be useful for detecting abnormal events. LBP is a texture descriptor that compares the intensity values of pixels with their surrounding pixels to represent the local structure of an image. The LBP operator outputs a binary code that represents the local texture pattern of the image. LBP has been applied in VAD to capture texture variations and irregularities in the video frames, which can be indicative of abnormal events. Both HOG and LBP have been used as single feature descriptors or combined with other feature descriptors to improve the performance of VAD systems. Though, it is significant to note that the effectiveness of these feature descriptors can vary depending on the specific characteristics of the video dataset and the anomaly types that are being detected. HOG and LBP are both feature descriptors used for image and video analysis.

The mathematical form of HOG can be represented as follows:

- Image gradient computation: Compute the gradients of the picture in the x and y direction using Sobel operators.
- Histogram computation: Divide the image into small cells (typically 8×8 or 16×16 pixels) and calculate the gradient magnitude and orientation for each pixel. Compute the histogram of orientations for each cell.
- Block normalization: Combine the histograms of neighboring cells into blocks (typically 2×2 or 3×3 cells) and normalize the block by L2-norm to reduce the influence of illumination variations.

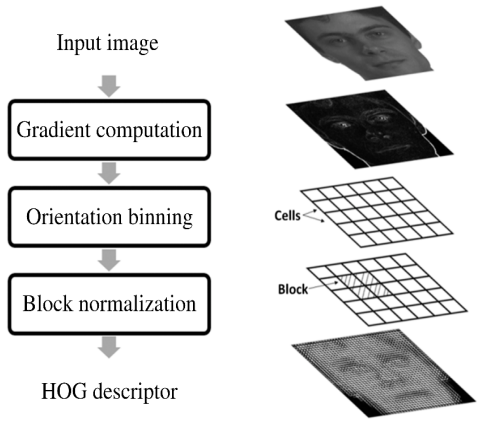

Figure 3. HOG descriptor

The mathematical form of LBP can be represented as follows:

- Image patch extraction: Extract a square patch from the image centered around each pixel.
- Thresholding: Compare the concentration of each pixel in the patch to the strength of the central pixel. Assign a binary value of 1 if the pixel value is >= the central pixel value, and 0 otherwise.
- Pattern coding: Convert the binary pattern into a decimal number and create a histogram of the frequency of occurrence of each pattern in the image.

$$LBP(gp_x, gp_y)\sum_{p=0}^{P-1} S(gp-gc)\times 2^p$$

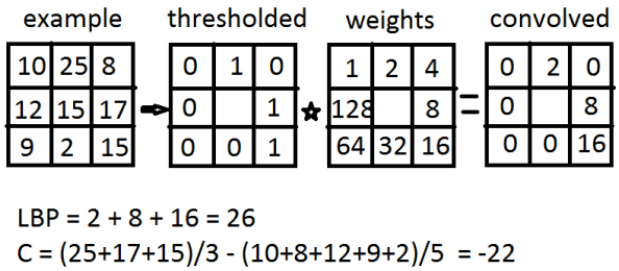

Figure 4. Evaluation of LBP

Overall, both HOG and LBP are powerful and effective feature descriptors for VAD, with their mathematical forms providing a clear framework for understanding their computation.

A framework for VAD based on DL is called One-Class CNN. It is designed to recognise any abnormalities that depart from the regular patterns it has trained to recognise in a video. One-Class CNN only needs normal data for training, in contrast to conventional anomaly detection techniques that need labelled data for both normal and abnormal occurrences. One-Class CNN has the ability to simulate the typical patterns of the data during training and recognise any deviations from

those patterns during testing. One-Class CNN extracts spatiotemporal properties from video input using a 3D convolutional neural network (CNN). By reducing a reconstruction error between the input frames and the model's output frames, the model learns to recognise the typical patterns of the data by ingesting video frames. While testing, the model looks for abnormalities by comparing the input frames' reconstruction error to a predetermined threshold. The model marks the input frames as anomalous if the error rises over the threshold. One-Class CNN has shown good results in a range of anomaly detection applications, including seeing unusual occurrences in surveillance footage, spotting flaws in production methods, and spotting abnormalities in x-ray pictures (Figure 5).

Using OC-CNN, features are extracted from the input frames and converted into a lower-dimensional representation using the encoder component of the network (Figure 6). The network's decoder component is then utilized to rebuild the input frames using the features that were retrieved. Only normal samples are utilized to train the network during the training phase, and the loss function is created to reduce the discrepancy between the reconstructed frames and the original normal frames. The frame is labelled as an anomaly if the difference is greater than a certain limit. The benefit of OC-CNN is that it can identify abnormalities unsupervisedly and does not need any previous knowledge of them. In numerous benchmark datasets, it has been shown that OC-CNN performs better than other cutting-edge approaches for detecting video anomalies. Its processing efficiency makes it appropriate for real-time applications.

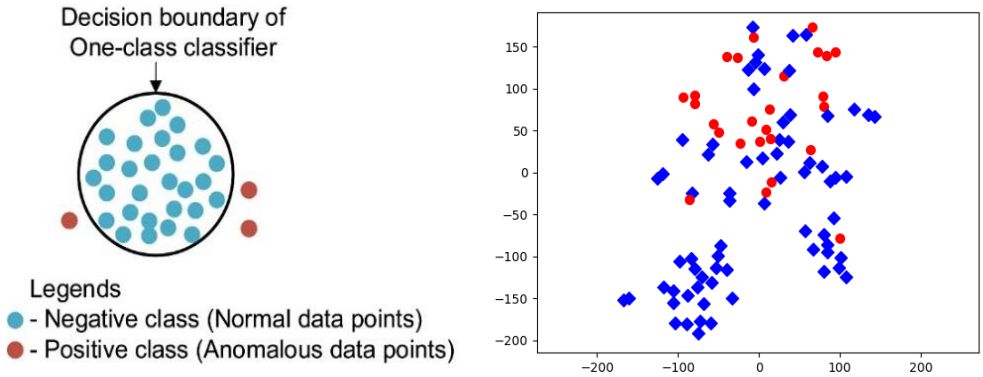

Figure 5. Decision boundary classifier with One-Class

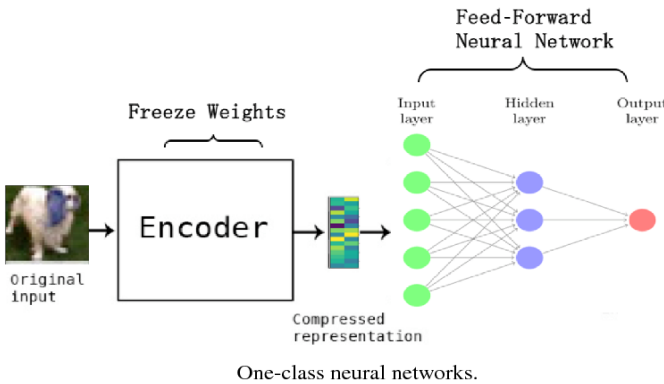

Figure 6. Neural networks with One-Class

4. IMPLEMENTATION AND RESULTS

VAD is a vital task in video investigation systems that aims to identify abnormal events or behaviors in each scene and overall proposed architecture has shown in Figure 7. The detection of such events is a challenging task that requires the extraction of discriminative features from the video data and the use of appropriate classification algorithms. There has been an increase in interest in the creation of effective and efficient algorithms for VAD due to the accessibility of video data, developments in CV, and ML approaches. However, the implementation of such algorithms can be a complex and time-consuming task, requiring careful consideration of factors such as feature selection, data pre-processing, model training and evaluation, and deployment in real-world systems. In this context, it is important to have a clear understanding of the different stages involved in the implementation of a VAD system, as well as the available tools and resources for facilitating this task (Figures 8-10). Implementing VAD typically involves the following steps:

- Data collection: Collecting video data from cameras or other sources.
- Pre-processing: Pre-processing the video data by removing noise, resizing, and converting to the desired format.
- Feature extraction: Extracting relevant features from the pre-processed video data using techniques like HOG, LBP, or CNNs.
- Training: Training a model on the extracted features to identify normal and abnormal behaviour in the video data.

- Testing: Testing the trained model on new video data to detect anomalies.
- Post-processing: Post-processing the results to eliminate false positives and refine the anomaly detection.
- Visualization: Visualizing the results to provide actionable insights to the end-users.

To extract HOG features from a video, the following steps are taken:

a. The video is divided into small cells.
b. For each cell, the gradients in the x and y directions are calculated.
c. The gradients are then quantized into a set number of orientations (usually 9) and the magnitude of the gradient is added to the corresponding orientation bin.
d. The orientation histograms for each cell are concatenated to form a single feature vector.

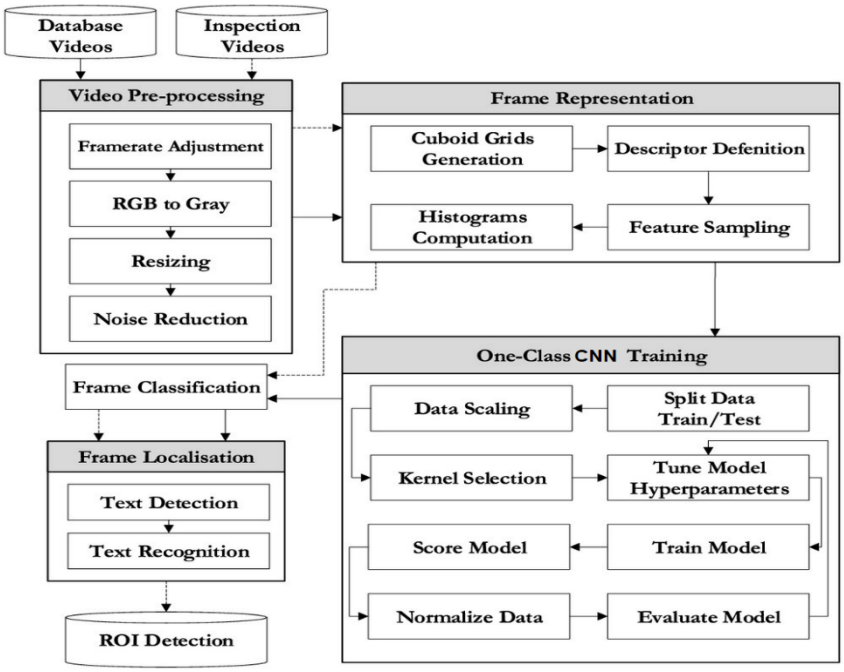

Figure 7. Overall Architecture of the proposed method

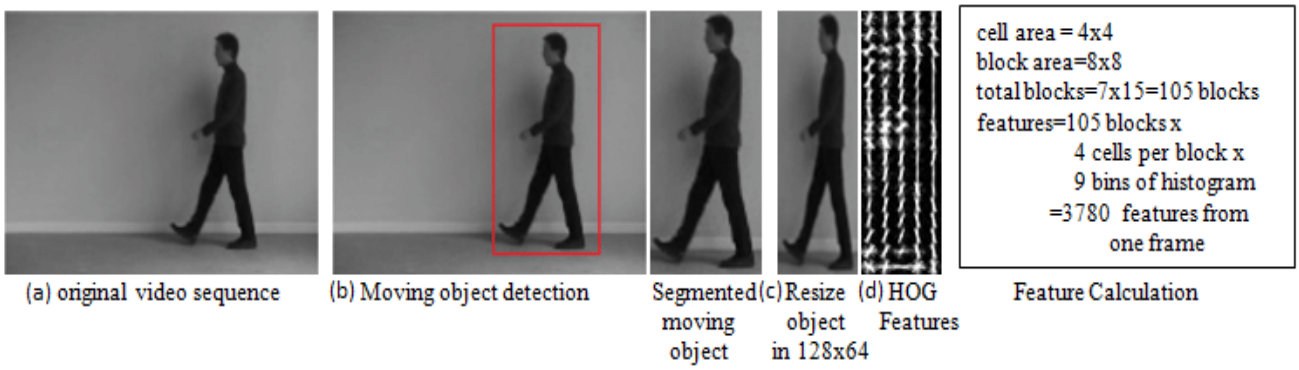

Figure 8. Process of feature calculation

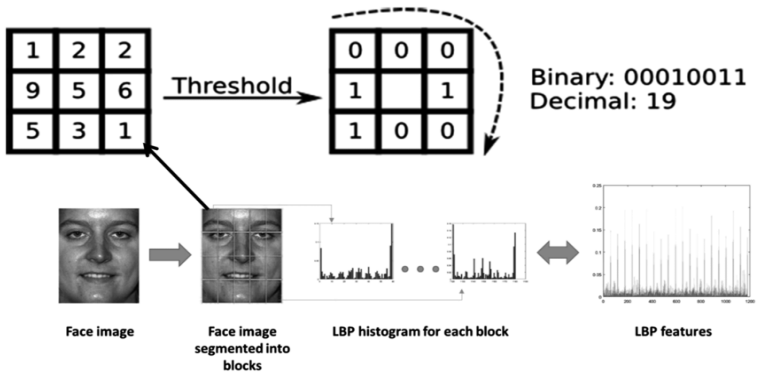

Figure 9. Calculation of LBP threshold values

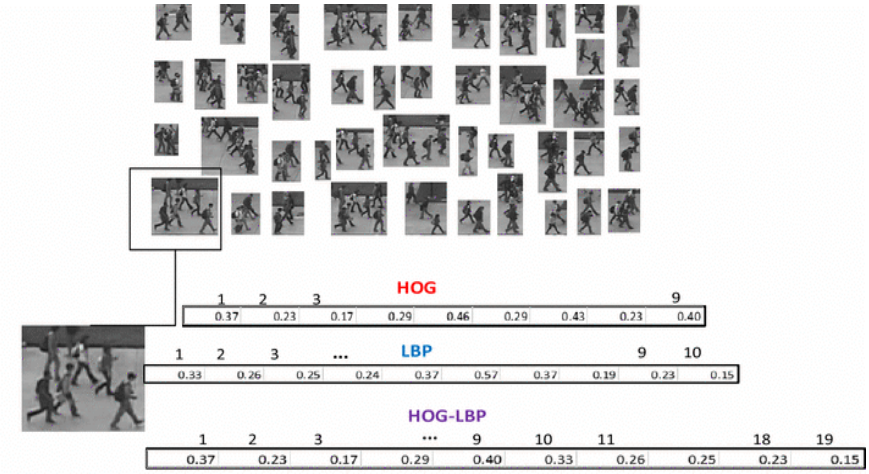

Figure 10. Video segmentation into small cells

To extract LBP features from a video, the following steps are taken:

a. The video is divided into small cells.
b. For each cell, a binary pattern is generated by comparing the intensity of the central pixel to the intensity of its neighbors.
c. The binary pattern is converted to a decimal value.
d. The decimal values for each cell are used to form a histogram, and the histograms for each cell are concatenated to form a single feature vector.

To extract LBP features from a video, the following steps are taken:

The video is divided into small cells.

a. For each cell, a binary pattern is generated by comparing the intensity of the central pixel to the intensity of its neighbors.
b. The binary pattern is converted to a decimal value.
c. The decimal values for each cell are used to form a histogram, and the histograms for each cell are concatenated to form a single feature vector.

Both HOG and LBP are popular feature descriptors for VAD because they are computationally efficient and provide robust representations of the underlying visual content in videos.

The below Figure 11 and Figure 12 are the output generated from the proposed approach on Avenue and UCSD datasets.

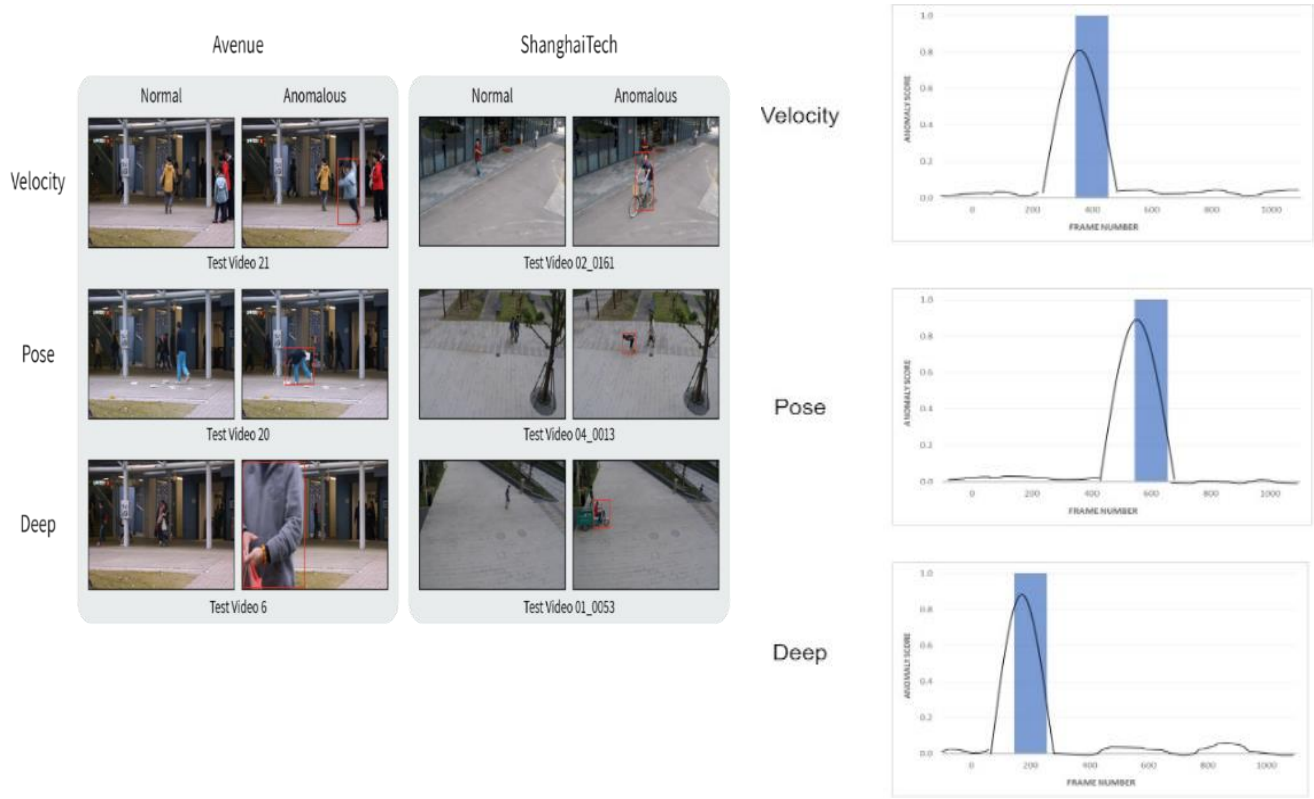

Figure 11. Results of the proposed approach implemented on Avenue & ShanghaiTech dataset

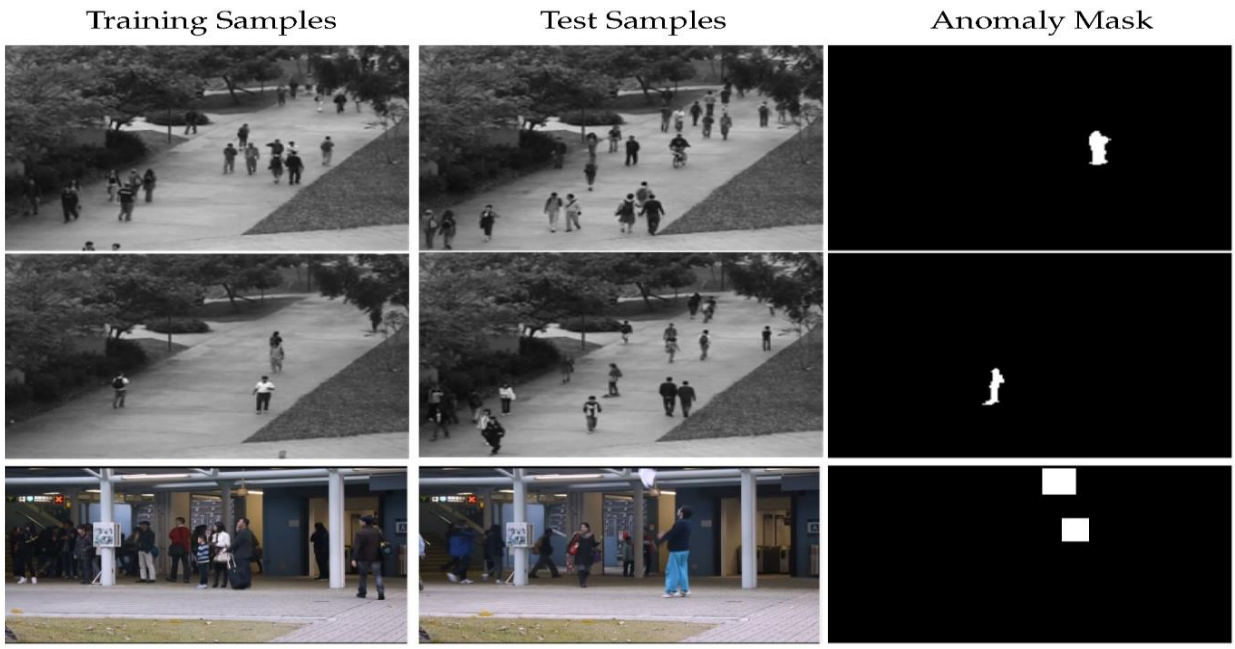

Figure 12. Results of the proposed approach implemented on UCSD Dataset identified with the ROI of respective anomaly

The proposed method is evaluated using the Accuracy metric and the attained results we given in the below Table 1. Also, the proposed approach is compared with existing systems and projected the comparison in the below Table 2.

Table 1. Evaluation of proposed method on Accuracy Metric

S. No	Dataset Name	Accuracy Achieved
1	Avenue	97.78
2	ShanghaiTech	96.88
3	UCSD	98.34

Table 2. Comparison with existing approaches

S. No	Author	Method Used	Accuracy
1	Dollár et al. [14]	PCA	86.89
2	Xu et al. [2]	ICA	91.24
3	Zhu and Newsam [15]	t-SNE	90.32
4	Ullah et al. [16]	UMAP	94.22
5	Chong and Tay [17]	PCA with HOG	96.78
6	Proposed Method	HOG-LBP	98.56

5. FUTURE DIRECTIONS

Some potential future directions for research on VAD using HOG-LBP and One-Class CNN could include:

- Investigating the use of other feature extraction methods in combination with HOG and LBP, such as DL-based methods like recurrent neural networks (RNNs).
- Exploring the use of more advanced techniques for anomaly detection, such as generative models like variational autoencoders (VAEs) or adversarial networks.
- Conducting experiments on larger and more diverse datasets to evaluate the generalization capabilities of the proposed approach.
- Addressing the issue of class imbalance in anomaly detection by exploring techniques such as oversampling or undersampling, or using more advanced methods like cost-sensitive learning or ensemble methods.
- Developing methods for real-time or near real-time anomaly detection in videos, which would require optimizing for computational efficiency and minimizing the latency between video frames.
- Investigating the use of transfer learning, where models trained on one dataset are fine-tuned on another dataset, to improve the performance of the anomaly detection system.

6. CONCLUSION

In conclusion, VAD is an important application of computer vision and has been the focus of research for many years. Two popular feature descriptors used in VAD are HOG and LBP, which capture local spatial and texture information. One class CNN is another popular approach for anomaly detection that uses a deep autoencoder and a feed-forward network. Both HOG-LBP and One class CNN have shown promising results in detecting anomalies in videos. The robust multi-descriptor fusion with One-Class Convolutional Neural Network (OC-CNN) for detecting anomalies in video surveillance is to enhance the accuracy, adaptability, and reliability of anomaly detection systems in diverse and dynamic surveillance scenarios. By integrating Multiple Feature Descriptors, such as Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP), and leveraging the power of OC-CNN for One-Class classification, the goal is to create a comprehensive and effective anomaly detection framework. Nevertheless, the individual application and the features of the dataset will determine which strategy is most suited. Additionally, the calibre and volume of training data have a significant impact on how well these strategies work. In conclusion, VAD using HOG-LBP and One class CNN is a promising research field with the potential to enhance public safety and security in a variety of applications. the main objective of combining robust multi-descriptor fusion with OC-CNN is to create an advanced

anomaly detection system that excels in adaptability, efficiency, and accuracy, thereby contributing to the improvement of video surveillance technologies for enhanced security and situational awareness.

REFERENCES

[1] Cheng, K.W., Chen, Y.T., Fang, W.H. (2016). An efficient subsequence search for video anomaly detection and localization. Multimedia Tools and Applications, 75: 15101-15122. https://doi.org/10.1007/s11042-015-2453-4

[2] Xu, D., Yan, Y., Ricci, E., Sebe, N. (2017). Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding, 156: 117-127. https://doi.org/10.1016/j.cviu.2016.10.010

[3] Chaaraoui, A.A., Climent-Perez, P., Flórez-Revuelta, F. (2013). Silhouette-based human action recognition using sequences of key poses. Pattern Recognition Letters, 34(15): 1799-1807. https://doi.org/10.1016/j.patrec.2013.01.021

[4] Cui, X., Liu, Q., Gao, M., Metaxas, D.N. (2011). Abnormal detection using interaction energy potentials. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, Colorado Springs, CO, USA, pp. 3161-3167. https://doi.org/10.1109/CVPR.2011.5995558

[5] Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G. (2019). Graph Convolutional Label Noise Cleaner: Train a Plug-and-play Action Classifier for Anomaly Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1237-1246.

[6] Patrikar, D.R., Parate, M.R. (2022). Anomaly detection using edge computing in video surveillance system: review. International Journal of Multimedia Information Retrieval, 11: 85-110. https://doi.org/10.1007/s13735-022-00227-8

[7] Li, W., Mahadevan, V., Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(1): 18-32. https://doi.org/10.1109/TPAMI.2013.111

[8] Lu, C., Shi, J., Jia, J. (2013). Abnormal event detection at 150 FPS in MATLAB. In 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, pp. 2720-2727. https://doi.org/10.1109/ICCV.2013.338

[9] Citamak, B., Caglayan, O., Kuyu, M., Erdem, E., Erdem, A., Madhyastha, P., Specia, L. (2021). MSVD-Turkish: A comprehensive multimodal video dataset for integrated vision and language research in Turkish. Machine Translation, 35: 265-288. https://doi.org/10.1007/s10590-021-09276-y

[10] Liu, W., Luo, W., Lian, D., Gao, S. (2018). Future frame prediction for anomaly detection - a new baseline. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 6536-6545. https://doi.org/10.1109/CVPR.2018.00684

[11] Oza, P., Patel, V.M. (2019). One-Class convolutional neural network. IEEE Signal Processing Letters, 26(2): 277-281. https://doi.org/10.1109/LSP.2018.2889273

[12] Dalal, N., Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, pp. 886-893. https://doi.org/10.1109/CVPR.2005.177

[13] Brahnam, S., Jain, L.C., Nanni, L., Lumini, A. (Eds.). (2013). Local Binary Patterns: New Variants and Applications. Springer Publishing Company, Incorporated. https://doi.org/10.1007/978-3-642-39289-4

[14] Dollár, P., Rabaud, V., Cottrell, G., Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, pp. 65-72. https://doi.org/10.1109/VSPETS.2005.1570899

[15] Zhu, Y., Newsam, S. (2019). Motion-aware feature for improved video anomaly detection. arXiv preprint arXiv:1907.10211.

[16] Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., Baik, S.W. (2017). Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access, 6: 1155-1166.

[17] Chong, Y.S., Tay, Y.H. (2017). Abnormal event detection in videos using spatiotemporal autoencoder. arXiv:1701.01546. https://doi.org/10.48550/arXiv.1701.01546

[18] Muhammad, K., Hussain, T., Tanveer, M., Sannino, G., de Albuquerque, V.H.C. (2020). Cost-effective video summarization using deep CNN with hierarchical weighted fusion for IoT surveillance networks. IEEE Internet of Things Journal, 7(5): 4455-4463. https://doi.org/10.1109/JIOT.2019.2950469

[19] Naik, A.J., Thimmaiah, G.M. (2021). Detection and localization of anamoly in videos using fruit fly optimization-based self organized maps. International Journal of Safety and Security Engineering, 11(6): 703-711. https://doi.org/10.18280/ijsse.110611

[20] Mohamed, S., Hassan, A.M., Aslan, H.K. (2021). IoT modes of operations with different security key management techniques: A survey. International Journal of Safety and Security Engineering, 11(6): 641-651. https://doi.org/10.18280/ijsse.110604

[21] Bangare, P.S., Patil, K.P. (2022). Study and analysis of various authentication and authorization for IoT devices: A challenging overview. International Journal of Safety and Security Engineering, 12(2): 209-216. https://doi.org/10.18280/ijsse.120209

[22] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S. (2016). Learning temporal regularity in video sequences. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 733-742. https://doi.org/10.1109/CVPR.2016.86