A Two-Stage System for Surveillance Video Summarization and Unsupervised Abnormal Event Detection in Educational Institutions

M. E. ElAlmi[1,*], M. M. Lotfy[2], M. M. Ghoniem[3]

[1]Prof. of Computer and Information System, Faculty of Specific Education, Mansoura University, Egypt

[2]Demonstrator of computer teacher preparation Department, Faculty of Specific Education, Mansoura University, Egypt

[3]Lecturer of computer teacher preparation Department, Faculty of Specific Education, Mansoura University, Egypt

Emails: moh_elalmi@mans.edu.eg; mariamlotfy@mans.edu.eg; m_ghonem@mans.edu.eg

Abstract

Surveillance cameras play a pivotal role in educational institutions. They monitor the educational process, detect violations, and protect students from potential injuries or dangers. Continuous recording generates a massive amount of video data. Human observers spend significant time and effort reviewing the footage. Reviewing aims to detect and quickly address abnormal events. Abnormal events are rare in educational environments. Observers may become bored during continuous monitoring. This may cause fatigue and loss of attention. To overcome these challenges, this paper proposes an intelligent system that combines summarization and abnormal event detection in surveillance video. It is divided into two stages: The first stage starts with the extraction of static, feature-based key frames that highlight the video's most significant content. In the second stage, Convolutional Autoencoder (CAE) network used to detect abnormal events from the key frames generated by the summary stage. The proposed system produces two separate videos: a general summary and a dedicated abnormal events video sent to the relevant individuals. The proposed system was tested on some benchmark datasets. The experimental results demonstrated that the proposed system was effective in reducing browsing time and effort, as well as in detecting abnormal events within an educational context.

Keywords: Static summarization; Surveillance video; Convolutional Autoencoder (CAE); Abnormal Events

1. Introduction

The use of surveillance cameras has become increasingly prevalent in both public areas such as roads, banks, and airports and private places such as companies and schools to increase security.

The use of surveillance cameras has become a major factor in increasing security in educational institutions. Cameras help protect property and detect cases of abnormality, violence, or criminal behavior. These cases may affect students, teachers, and workers. Such incidents can hinder the educational process [1]. Classroom observation videos are used to improve and measure the efficiency of teachers' performance, improve the quality of teaching, and increase classroom interaction. They allow for observation of actions and activities that cannot be adequately captured through images or audio recordings. They also provide greater flexibility to return to the video at any time and the ability to fast-forward, rewind, zoom in, etc.

Surveillance camera recordings generate huge amounts of videos. This requires sufficient resources such as personnel, time, and storage devices to process all these clips and detect any abnormal activity. Human monitors need to stay focused when reviewing videos for long hours that can lead to boredom and fatigue especially when monitoring multiple cameras at once. To overcome these challenges, video summaries can be used to reduce the time of reviewing the entire video by extracting only the priority information from the original video [2].

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

Video summarization is the process of reducing video sequences to a select few important key frames that capture the essential visual semantics, while removing redundant or overly similar frames.

Video summarization techniques aim to create concise representations of videos. Keyframe summarization involves extracting significant frames that capture the most important moments. Skimmed video summarization compiles short video segments along with their corresponding audio. These methods produce a condensed version of the original footage [3].

To detect abnormal events, unexpected events or behaviours are estimated relative to the expected context of events from the video. Mostly, continuous monitoring with great effort and attention by a human, which is a difficult task because these events occur rarely compared to normal events, does this traditionally unintelligently.

This paper provides an intelligent system to summarize and detect abnormal events with surveillance cameras to to enhance both the effectiveness and efficiency of security surveillance in educational institutions. The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 details the proposed system, and Section 4 discusses the application and experimental results. Finally, Section 5 provides the conclusion.

2. Related Work

This section reviews prior research across three interconnected themes: (1) video summarization (2) Abnormal event detection (3) Video Summarization and Abnormal Event Detection.

2.1 Video Summarization

Yarrarapu et al (2024) [4] had developed an effective video summarization framework using the MobileNetSSD algorithm to extract key frames based on objects of interest (OoI) and applied temporal analysis to reflect the sequence of events. The framework was tested on the TVSum and SUMMe datasets and proved to be effective and fast, yielding clear and accurate summaries.

Javaid et al (2023) [5] had proposed a two-stage approach for underwater video summarization YoloV3 model for object detection and Perceived Motion Energy (PME) approach for keyframe extraction. It was tested on a public dataset. The results demonstrated how the approach assisted in guiding marine researchers.

Köprü and Erzin (2022) [6] had developed a video summarization model with affective information by training their continuous emotion recognition model (CER-NET) using the RECOLA dataset. They suggested affective video summarization architectures (AVSUM) and explored attention mechanisms. The AVSUM-GRU model showed remarkably better results in F-score and face recall compared to earlier methods in experiment tests on the TvSum dataset.

2.2 Abnormal Event Detection

Paulraj and Vairavasundaram (2025) [7] had designed a model that included C-LSTM networks to embed temporal dependencies and Swin Transformers for learning spatial features with the aim of weakly supervised abnormal event detection in videos. Experimental results on benchmark datasets proved that ST-HTAM outperformed state of the art methods in reducing false alarms and improving detection accuracy.

Mumtaz et al (2024) [8] had developed a dynamic frame-skipping method and Net model to detect abnormal actions at varying speeds. Tested on UCF-Crime and Shanghai Tech datasets, it achieved AUC scores of 86.1% and 99.87%, outperforming existing methods with high F1 scores for activities like explosions, accidents, robbery, and stealing.

Ingle and Kim (2022) [9] had proposed the MSD-CNN model for detecting guns and knives in CCTV videos. Trained on normal and abnormal evet frames, experimental results on public datasets showed high accuracy for detecting guns and knives, but limited performance for other abnormal event not included in the training.

Saponara et al (2021) [10] had used the YOLOv2 CNN for the detection of fire and smoke from CCTV footage. Training on indoor and outdoor datasets enabled detection with better performance than other fire and smoke detection methods.

2.3 Video Summarization and Abnormal Event Detection

Jung et al (2024) [11] had developed a video analysis tool (VIDEX) for surveillance footage investigation by combining object and abnormal event detection within a user-friendly MVVM-based interface. It uses YOLOv5

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

and abnormal event detection models to generate indexed summaries. VIDEX proved valuable in criminal investigations, providing enhanced processing accuracy.

Sabha and Selwal (2023) [12] had presented a method for video summarizing that focuses on face mask non-compliance in order to identify COVID-19 protocol violations. Using a refined ResNet-50 network, CoSumNet achieved good identification accuracies in both seen and unseen contexts after being trained on the "Face Mask Detection ~12K Images Dataset" and validated using CCTV footage.

Muchtar et al (2022) [13] had proposed one deep learning framework for both abnormal event detection and summarization. Abnormal event is detected and the regions of interest are extracted by the system through 3D convolutional networks and blob analysis. Experimental results showed that key information was preserved while competitive summarization performance was obtained.

Although video summarization, abnormal event detection, and their combination have been the subject of prior research, the suggested system improves on this integration by effectively detecting abnormal events using keyframes that are extracted during summarization, offering a quicker and less computationally demanding alternative to analysing the entire video.

3. The Proposed System

The proposed system is divided into two primary stages: The first stage begins with static video summarization to extract its key frames. In the second stage, abnormal events are detected as shown in block diagram 1.

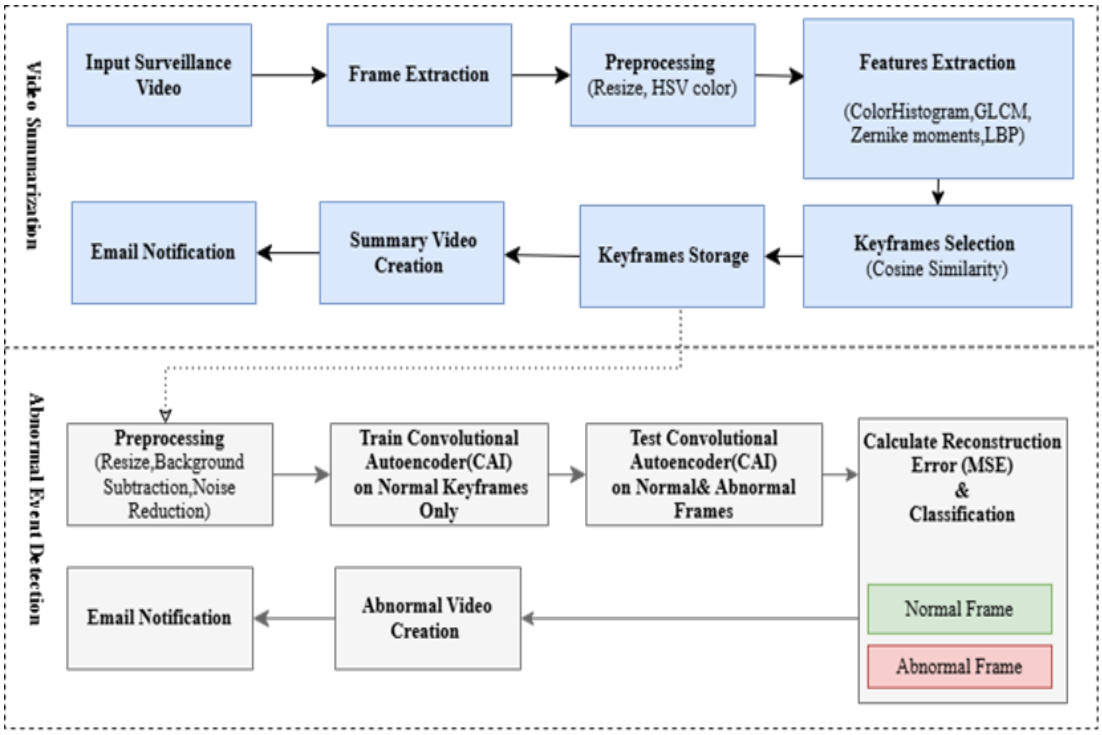

Figure 1. Proposed system for video summarization & abnormal event detection

3.1 Video Summarization

The first stage generates a static video summary by condensing input videos into representative keyframes. As shown in figure (1), the workflow consists of four stages: Pre-processing, Feature extraction, Keyframe extraction and Mailing the summary part of video.

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

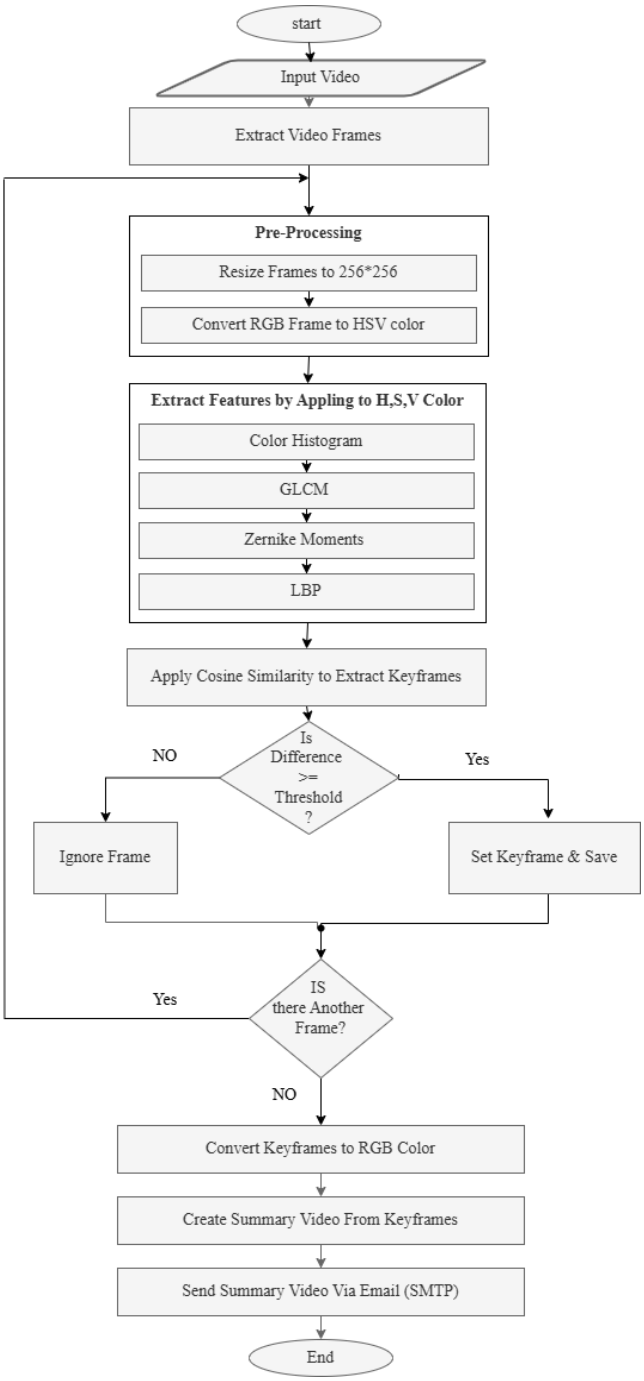

Figure 2. First stage (Video Summarization)

3.1.1 Pre-processing

The initial step in the pre-processing stage is to extract frames from the input video by choosing every second frame using a systematic sampling technique. It sacrifices computational efficiency for sufficient temporal coverage of video content. All frames are reduced to a uniform size of 256×256. Downsizing frames to this uniform size reduces computational complexity and accelerates processing without compromising valuable visual information because key details remain identifiable at this scale. The RGB frames are then translated to Hue, Saturation, and Value (HSV) color space. Saturation and Hue or Color, and Value or brightness are divided up by HSV, which allows an elimination of issues like shadows or uneven lighting. By separating the factors of brightness, the method is less sensitive to changes in lighting, leading to more even feature extraction across a broad range of environments.

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

3.1.2 Feature extraction

To identify the key information to examine, the pre-processed frames are feature-extracted in the next step. Since static video summarization relies on these features to highlight important events, feature selection is crucial. This aligns with how human beings process visual information by focusing on essential characteristics like color, texture, and shape. In extracting them, it has four large features: color histograms, GLCM texture features, Zernike moment shape features, and LBP texture features.

A. Color histogram

A color histogram is a mapping function of each color to its count. Colors are represented on the x-axis, and pixel values on the y-axis. Such a representation provides a general statistical description of an image's color content with extra chromatic information beyond what grayscale histograms can represent, hence supporting target recognition.

Color histograms in photography and image processing quantify color variation by counting the pixels within specific ranges of color. They can both record the full color model of an image and express it as a statistical representation of the unbroken color continuum. This excellent description is well suited to object detection since it facilitates matching histogram signatures even when one does not know the object's location or orientation [14].

Color spaces that divide hue, saturation, and brightness (like HSV or HSL) or channels (like red, green, and blue) can be used to build color histograms. They offer global summary of overall color distribution as well as a localized perspective when computed within the segmented regions, more accurately describing spatial variations.

Also, the effectiveness of histogram comparison algorithms like the Histogram Intersection method largely depends on the color space applied. For instance, using HSV or CIELab values may enhance performance since their components are orthogonal to one another. Color histograms are further invariant to translation, scaling, and rotation transformations, meaning that they form a valid resource in most image processing tasks.

In the current implementation, all HSV color channels (Hue, Saturation, Value) are divided into 16 equal bins, forming a 1×16 feature vector for each channel. Concatenating all three channels produces a 1×48 feature vector representing global color distribution.

B. GLCM texture features

A commonly applied statistical technique for texture analysis is the Gray Level Co-occurrence Matrix (GLCM) that determines the spatial relationship between pairs of pixels in an image. Texture analysis via GLCM determines key features of the surface structure of an image well and thus discriminates between textures such as rough, smooth, or bumpy by analyzing Gray-level co-occurrence variability.

The GLCM is constructed by calculating how often an intensity pixel (i) appears in a specific spatial relationship with a pixel of intensity (j) and forming a matrix in which the matrix element p(i,j) is the probability of co-occurrence [15].

characterized by the following equation, each element in the GLCM matrix is the number of co-occurring pixel pairs sharing the same intensities at some distance (s) and direction of (θ) in the image matrix.

$$GLCM(i,j)_{\theta} = |\{((p_1,p_2)|\, I(p_1)) = i,\ I(p_2) = j\}| \quad (1)$$

Haralick initially suggested 14 texture features with the help of the GLCM to characterize the image texture properties. Subsequent studies widened the feature set to 21 consisting of other measures like cluster shade, cluster prominence, and other measures of information.

The comprehensive nature of GLCM, which encapsulates both first-order and second-order statistics, underscores its enduring relevance as a texture feature extraction method. The descriptive statistics derived from the GLCM, known as second-order statistics, provide a robust measure of texture essential for applications in medical imaging, remote sensing, and image segmentation.

After computing the GLCM, five statistical features—energy, contrast, homogeneity, entropy, and correlation—are extracted to analyze texture.

Energy evaluates how uniformly the GLCM's elements are distributed. A high-energy value occurs when the matrix elements are uniform. The formula is:

$$Energy = \sum_i \sum_j P(i,j)^2 \quad (2)$$

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

Contrast, alternatively termed inertia, measures the intensity variation between pixels. The formula is:

$$Contrast = \sum_i \sum_j (i-j)^2 * P(i,j) \quad (3)$$

Homogeneity quantifies how close the distribution of elements in the GLCM is to its diagonal. The formula is:

$$Homogeneity = \sum_i \sum_j \frac{P(i,j)}{1+(i-j)^2} \quad (4)$$

Entropy quantifies the complexity of the texture. The formula is:

$$Entropy = \sum_i \sum_j P(i,j) * log\big(P(i,j)\big) \quad (5)$$

Correlation quantifies how closely two pixels are related. The formula is:

$$Correlation = \frac{\sum_i \sum_j [ij * P(i,j)] - \mu_x * \mu_y}{\sigma_x * \sigma_y} \quad (6)$$

Where P(i,j) represents the element of the co-occurrence matrix at position (i,j), μx and μy denote the mean values of the row and column weights, respectively, while σx and σy are the corresponding standard deviations. The variables i and j refer to the indices of the co-occurrence matrix.

In the current implementation, GLCM is computed across four directions (0°, 45°, 90°, 135°) for each HSV channel. From each matrix, five statistical features—energy, contrast, homogeneity, entropy, and correlation— are extracted and then averaged across the four directions. Concatenating results from H, S, and V produces a 1×15 feature vector.

C. Zernike moment shape features

Zernike moments provide a powerful and efficient method for extracting shape features from digital images. By transforming the image from Cartesian to polar coordinates and computing the moments over the unit circle, they reduce redundancy while preserving rotation invariance. Leveraging invariance to rotation, translation, and scaling, Zernike moments capture both global and local shape variations effectively lower orders represent the overall structure, while higher orders reveal subtle, localized details. Integrating these moments into a spatial pyramid framework further strengthens their descriptive power by offering a hierarchical feature representation.

The computation of Zernike moments relies on an orthogonal set of Zernike polynomials defined over the unit circle, which minimizes data redundancy and guarantees rotation invariance. The Zernike moment of order n and repetition mmm for an image function f(x,y) can be mathematically defined as follows [16]:

$$Z_{n,m} = \frac{n+1}{\pi} \sum \ \in \text{ unit disc } \sum \ f(\sigma,\theta) \ V^*_{n,m}(\sigma,\theta) \quad (7)$$

where the Zernike polynomial of high order in polar form is expressed as:

$$V_n^m(\rho,\theta) = R_n^m(\rho)\, e^{im\theta} \quad (8)$$

Here, n is the order of the moment while mmm denotes its repetition under the restriction that $n \geq 0$, $|m| \leq n$, and $(n - |m|)$ is even. This definition, when restricted to the unit disk (i.e., $x2+y2 \leq 1$), permits lower-order moments to capture the global shape properties while the higher-order moments capture the finer details.

In the current implementation, Zernike moments of order up to 10th are calculated for every HSV channel, resulting in 25 moments per channel. Concatenating channel-wise results in a 1×75 feature vector representing rotational-invariant shape features.

D. LBP texture features

Local Binary Pattern (LBP), is a popular texture descriptor in computer vision because of its easy computation and ability to extract stable features for uses like texture analysis, object classification, and remote sensing. Its insensitivity to varying illumination conditions also makes it a desirable candidate for image search and classification.

The fundamental idea of LBP is to match each pixel in an image against the pixels that are around it and describe these relationships in the form of binary patterns. The matching is normally performed within a 3×3 block, where the center pixel is matched against its eight neighbors, and the result is described as an 8-bit binary number in a specific order, such as clockwise.

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

In LBP, once the center pixel g(c) and the neighboring pixels g(p) gray values are considered, the local texture feature is computed through the function S [17]:

$$S(p,c) = \begin{cases} 1, if\ g(p) \geq g(c) \\ 0, if\, g(p) < g(c) \end{cases} \quad (9)$$

$$\text{lbp}(c) = \sum_{i=1}^{8} 2^{i-1} S(p_i, c) \quad (10)$$

This function determines the binary code for each neighboring pixel based on its relative intensity compared to the center pixel, and this code subsequently forms the basis for the overall LBP code calculation.

In the current implementation, every HSV channel is subjected to uniform Local Binary Patterns (LBP) with radius 1 and 8 neighborhood points. Each channel generates 59 uniform patterns, and concatenation across channels produces a 1×177 feature vector.

3.1.3 Keyframe extraction

Using the feature set that were taken out of the frames in the previous stage, the cosine similarity measure is used to compare them to extract only the key frames and reduce redundancy.

- Cosine Similarity

Cosine similarity is widely applied in text similarity detection, image recognition, and machine learning, especially in high-dimensional spaces. Recent research has confirmed its efficiency and robustness across various data analysis tasks, underscoring its practical value in modern applications.

Cosine similarity is a metric used to assess the alignment between two data vectors by calculating the cosine of the angle between them. It is computed using the following formula [18]:

$$\text{Cosine Similarity}(A,B) = \frac{A \cdot B}{||A||\,||B||} \quad (11)$$

In this equation, A and B stand for the data vectors in this equation, A·B is the dot product of these vectors, and ‖A‖ and ‖B‖ indicate their Euclidean norms (magnitudes).

The resulting similarity score is a number between -1 and 1, where 1 denotes perfect similarity, 0 denotes orthogonality (no similarity), and -1 shows complete opposition.

Finally, the extracted keyframes, processed in HSV color space are converted back to their native RGB representation.

3.1.4 Mailing the summary part of video

A summary video is generated from the extracted keyframes. This reduces storage capacity and review time and effort for observers. This condensed video is sent to the host using the SMTP protocol. SMTP (Simple Mail Transfer Protocol) operates as a standardized communication protocol that ensures reliable email delivery across the internet.

3.2 Abnormal Event Detection

In the second step, abnormal events are detected from the extracted key frames in several steps: (1) Pre-processing (2) Applying the Convolutional Autoencoder network (3) Converting the abnormal frames to video and sending it to the relevant individuals as shown in figure (2).

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

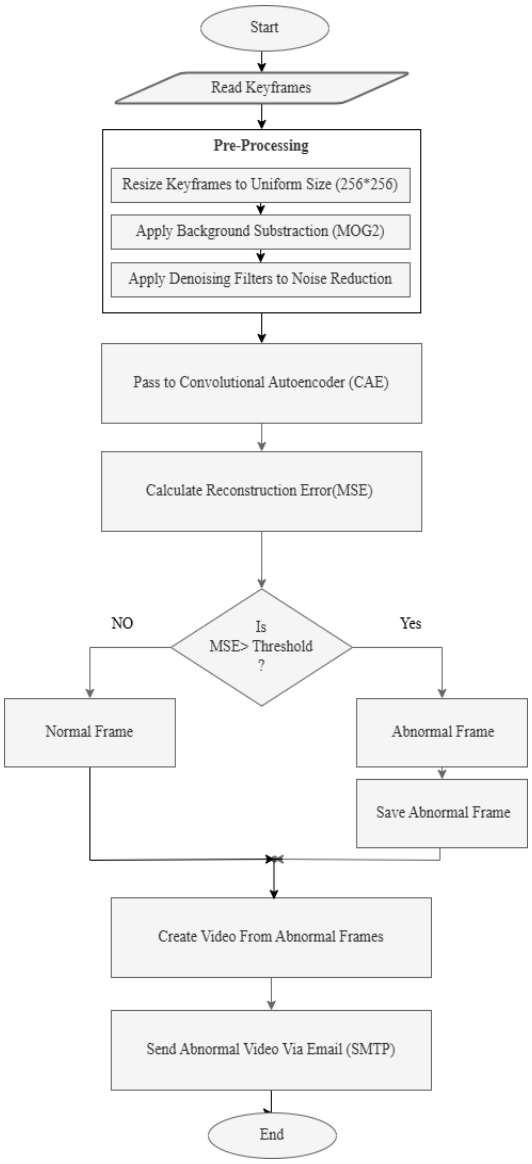

Figure 3. Second stage (Abnormal Event Detection)

3.2.1 Pre-processing

Preprocessing is an essential step to enhance the accuracy and efficiency of detecting abnormal events. This stage consists of several sub steps:

A. Keyframe Extraction:

The keyframes generated in the previous video summarization stage are retrieved. These keyframes effectively represent the most important content of the video, allowing for a more efficient process compared to working with the entire video frames. By focusing on keyframes, the computational cost is significantly reduced, while retaining the most relevant information.

B. Frame Size Uniformity:

To facilitate accurate comparison, the frame sizes are standardized. This step, also applied during video summarization, gives all the frames the same size, which is necessary for feature extraction and proper abnormal event detection.

C. Background Subtraction:

In scenes, abnormal events are usually related to dynamic objects. Hence, background subtraction has been used to take the attention towards dynamic objects by removing static backgrounds. It applied the MOG2 method, which

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

models every pixel using one of the several Gaussian distributions. The method separates from the background the objects moving on the scene, hence gaining a better detection of unusual events.

It begins by creating a mask where moving objects are white and the background is black. The resulting mask is then converted to RGB to be consistent with the original image. For consistency, the original image is also resized to the mask's size. Finally, a bitwise AND is carried out between the re-sized original image and the RGB mask, where colours of moving objects are retained while the static background is actually eliminated.

The above method eliminates the false positive caused by shadows or reflections, quite often reported in surveillance systems. In this system, the focus lies solely on moving objects, and that higher accuracy is given to detected abnormal events.

D. Noise Reduction:

Finally, noise in the image is minimized for clearer visibility. In the real video input, noise and interference can obscure the very information that the detection algorithms need to detect the anomaly.

To get over this issue, a denoising method is applied to remove all the noise and interferences within the video frames. Therefore, the noise effectively accentuates the detail of useful features so that moving objects and interesting information are more highly perceived and processed.

Hence, the system performance of anomaly identification is enhanced with fewer false positives and more precisely detecting the unusual events that take place in the course of surveillance footage.

3.2.2 Convolutional Autoencoder Network (CAE)

CAEs are one of the most widely used architectures in unsupervised abnormal event detection for surveillance videos since they are able to learn invariant spatial features from image sequences without using any labelled abnormal examples. Indeed, CAEs are especially good at modelling normal patterns within the surveillance data by specializing on the reconstruction of input frames. When unusual or abnormal events are given, the reconstruction error is larger, which is an effective indicator for abnormal event detection. Frames of abnormal events contain motion patterns that are significantly different from the motion patterns of normal frames.

CAEs leverage shared weights and spatial locality to learn local spatial features and successfully detect subtle abnormal events in surveillance videos. Their ability to extract two-dimensional structures and local information distinguishes them from traditional autoencoders, making them a strong choice for this task.

A convolutional autoencoder consists of two main components [19]:

- **Encoder:** Responsible for compressing the input image into a lower-dimensional latent space through a sequence of convolutional layers.
- **Decoder:** Reconstructs the original image from the latent representation using deconvolutional (or transposed convolution) layers.

The reconstruction loss function, which guides training, is given by:

$$e(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 + \lambda(w)^2 \qquad (12)$$

where λ represents the regularization parameter, and w^2 denotes the weights.

The encoding process for the k-th feature map can be mathematically expressed as follows:

$$h_k = \sigma(x * w_k + b_k)$$

where * indicates a 2D convolution operation, σ represents the activation function (commonly ReLU), and bk denotes the bias.

The decoding (reconstruction) operation is:

$$y = \sigma\left(\sum_{k\in H} h_k * w_k' + c\right) \qquad (13)$$

where wk′ is the flipped version of the learned weights and c is the bias per input channel. The initial layers focus on extracting low-level features, while the subsequent layers capture higher-level features such as motion and appearance.

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

Following the training phase, where the convolutional autoencoder learns to reconstruct normal event images, the algorithm proceeds to detect abnormal event in unseen data. This involves calculating the reconstruction error, specifically the Mean Squared Error (MSE), between the input image Ii and its reconstructed counterpart Ir, as defined by the equation:

$$MSE = \frac{1}{MN}\sum_{M,N}[I_i(m,n) - I_r(m,n)]^2 \qquad (14)$$

Here, M and N denote the image dimensions (rows and columns, respectively). An average Mean Squared Error (MSE) is then calculated from the training data and used as a threshold to distinguish between normal and abnormal events.

During testing, images containing both normal and abnormal events are processed, and their reconstruction errors are computed. If an image's MSE exceeds the established threshold, it is classified as abnormal.

Finally, to assess the system's performance, the Abnormality Score is computed and used to determine the Area Under the Curve (AUC). The abnormality score is calculated as:

$$\text{Abnormality score} = \frac{e(x) - \min e(x)}{\max e(x)} \qquad (15)$$

Where, x represents the reconstructed output image, and e(x) denotes the reconstruction error.

The ReLU activation function is employed because of its computational efficiency. Finally, the Adam optimizer is employed with an initial learning rate of 0.001, a batch size of 64, and up to 100 training epochs.

3.2.3 Mailing the Abnormal part of video

Frames classified as abnormal are converted into a video. Due to the small number of abnormal frames, the frame rate is reduced to a minimum and each frame is repeated multiple times to focus on the review process. The generated video is sent to the relevant personnel using SMTP to speed up the decision-making process.

4. Experimental results

This section provides a comprehensive evaluation of the proposed system, which integrates keyframe extraction for video summarization and abnormal event detection in surveillance footage. The system's performance is measured using quantitative metrics to demonstrate its effectiveness in capturing essential visual information, minimizing video redundancy, and accurately detecting abnormal events across diverse scenarios.

The experiments in this paper were performed in Python on a system equipped with an 8th Gen Intel® Core™ i5-8250U CPU at 1.60 GHz and 16 GB RAM.

4.1 Performance measures

To evaluate the effectiveness of the proposed video summarization method, three standard metrics are used: precision, recall, and F1-score. These metrics compare the selected summary frames with the ground truth by considering true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

4.1.1 Precision

Precision reflects the proportion of selected frames that are actually relevant. In other words, it indicates how accurately the method identifies important frames. It is calculated as:

$$Precision = \frac{TP}{TP+FP} \qquad (16)$$

A higher precision value suggests fewer irrelevant frames have been mistakenly included in the summary (i.e., fewer false positives).

4.1.2 Recall

Recall measures the method's ability to capture all the relevant frames. It represents the fraction of important frames from the ground truth that have been correctly identified in the summary:

$$Recall = \frac{TP}{TP+FN} \qquad (17)$$

A high recall implies that the summarization method misses fewer important frames (i.e., fewer false negatives).

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

4.1.3 F1-Score

Calculated as the harmonic mean of precision and recall, the F1-score serves as a balanced metric, effectively summarizing performance when precision and recall differ:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (18)$$

This score ranges between 0 and 1, with values closer to 1 indicating stronger overall performance.

Where:

- True Positive (TP): Frames correctly identified as relevant, matching those in the ground truth.
- False Positive (FP): Frames incorrectly marked as relevant but not present in the ground truth.
- False Negative (FN): Ground truth frames that were missed by the summarization method.
- True Negative (TN): Frames correctly identified as irrelevant. However, in the context of video summarization, TN is often not emphasized, since the primary goal is to detect and retain important frames.

After extracting keyframes, the second stage in the proposed system aims to detect normal and abnormal frames in surveillance camera footage. The effectiveness of the proposed abnormal event detection method is assessed using the Area Under the Curve (AUC) metric. This metric captures the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR), reflecting the overall detection performance.

It is computed by plotting TPR and FPR values on a two-dimensional graph across different threshold settings. The area under the curve formed by these points yields the AUC value. A value approaching 1 indicates near-perfect performance in distinguishing between normal and abnormal frames.

AUC is generally utilized in classification problems where classes are balanced.

The TPR and FPR are derived from the confusion matrix, based on the following formulas [20]:

$$TPR = \frac{TP}{TP+FN} \qquad (19)$$

$$FPR = \frac{FP}{FP+TN} \qquad (20)$$

Where:

TP and TN represent the numbers of correctly classified abnormal and normal frames, respectively, while FP and FN refer to normal and abnormal frames that were misclassified.

4.2 Results and analysis

Evaluation of the proposed method was conducted on the VSUMM dataset [21], containing 50 videos sourced from the Open Video collection. Each video is in MPEG-1 format, with a frame rate of 30 fps and a resolution of 352 × 240 pixels. They are colored and include audio. The dataset covers a variety of genres such as documentary, educational, historical, ephemeral, and lecture videos with video lengths between 1 and 4 minutes, amounting to a total of approximately 75 minutes of content.

The dataset also contains 250 manually produced summaries, contributed by 50 users. Each user summarized five distinct videos, providing five unique summaries for each video from different individuals.

The performance of the proposed method was assessed using different thresholds of Cosine Similarity (τ) to optimize keyframe extraction. The threshold value of τ = 0.90 consistently delivered the best performance in terms of Precision, Recall, and F1-Score.

The results were obtained by comparing the summaries generated with those provided by five users for each video. Precision, Recall, and F1-Score were computed for every comparison, and the final values represent the average scores across all user summaries.

Table 1 presents the performance of the proposed method on five sample videos (V21 to V25) from the VSUMM dataset.

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

Table 1: Evaluation Metrics of the Proposed Method on VSUMM Dataset[21] Samples

Video name	Precision (%)	Recall (%)	F1-Score (%)
V21	47.80	66.10	55.48
V22	75.10	92.30	82.82
V23	79.60	91.10	84.96
V24	82.80	89.40	85.97
V25	87.20	90.40	88.77

From table1, the proposed method achieves high precision and recall across most videos. For instance, in V25, the system proposed in this work scores a precision of 87.20% and a recall of 90.40%, which presents excellent keyframe selection. The F1-score also reaches 88.77%, which illustrates a well-balanced ability of the system to locate relevant frames and not unnecessary ones, further validating the proposed approach to summarize the significant content of each video.

Various threshold values (0.80, 0.85, 0.90, and 0.95) were tried to determine the most suitable setting for keyframe selection. Based on the experiment, using $\tau = 0.90$ yields the optimal balance between recall and precision in order to have highly relevant selected keyframes with low redundancy.

In the second stage, the proposed system detects normal and abnormal frames in surveillance videos. The experiments were conducted on two well-known datasets: CUHK Avenue and UCSD [22], including both Ped1 and Ped2.

Each dataset consisted of both training and testing videos. The training phase was conducted using videos that contained only normal events. The proposed system's performance was assessed using test videos that included normal and abnormal events.

The CUHK Avenue dataset was collected from the CUHK campus and consists of 37 video sequences (16 normal for training and 21 abnormal for testing), each around two minutes long. It contains 30,652 frames recorded at 640×360 resolution, 25 FPS, and RGB format from a single-scene perspective. Normal events include pedestrian activities, while abnormal events include unusual actions (such as running and throwing objects), wrong directions (e.g., walking against traffic), and abnormal objects (such as carrying bicycles).

The UCSD dataset includes two subsets: Ped1 and Ped2, recorded by fixed, elevated grayscale cameras overlooking pedestrian walkways. Ped1 videos have a resolution of 238×158, and Ped2 videos have 360×240, both captured at 10 FPS in 8-bit grayscale. Ped1 contains 34 training and 36 testing video clips, while Ped2 consists of 16 training and 12 testing clips. each comprising 200 frames. Normal events involve pedestrian walking, whereas abnormal events include non-pedestrian objects and unusual trajectories (bikers, skaters, carts, grass-walking).

Both datasets reflect outdoor scenes with dense pedestrian activity, closely resembling university campus environments. Their focus on abnormal events, including vehicles, makes them well suited for evaluating abnormal event detection methods in smart campus settings.

Examples of video frames from these datasets are shown in Figure 3.

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

Figure 3. Samples of public datasets: UCSD Ped1, Ped2 and CUHK Avenue

Evaluation of the proposed system was conducted on the CUHK Avenue and UCSD (Ped1 and Ped2) datasets using the AUC metric. Table 2 displays the metrics obtained with and without keyframe extraction.

Table 2: Evaluation Metrics of the Proposed System With and Without Keyframe Extraction

No	Dataset Name		AUC (%)	
	Frames		Keyframes	All Frames
1	CUHK Avenue		88.2	87.5
2	UCSD	Ped1	91.5	89.6
		Ped2	92.9	91.7

As shown in table 2, the proposed system demonstrated good performance in detecting abnormal events on the Ped2 and Ped1 datasets, achieving AUC scores of 92.9% and 91.5%, respectively, using extracted keyframes compared to using all frames. Furthermore, obtaining keyframes from the previous summarization stage reduced execution time and computational complexity, improving the overall system efficiency.

4.3 Comparison and Discussion

This section compares the proposed system with state of the art methods. The VSUMM dataset is used for summarization as illustrated in table 3, while the CUHK Avenue and UCSD (Ped1 and Ped2) datasets are used for abnormal event detection as illustrated in table 4.

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

Table 3: Evaluation of the Proposed System in Comparison with State-of-the-Art Methods on VSUMM Dataset Samples

Methods	Avg Precision (%)	Avg Recall (%)	Avg F1-Score (%)
Hannane et al. [23]	43.8	80.5	56.7
Bendraou et al. [24]	55	56	55
Sreeja & Kovoor [25]	70	71	71
The Proposed System	78	79	78.496

Table 4: Comparison Between the Proposed System and State-of-Art Methods

Methods	AUC (%)		
	UCSD		CUHK Avenue
	Ped1	Ped2	
Hasan et al. [26]	81	90	70.2
Luo et al. [27]	75.5	88.1	77
Ionescu et al. [22]	68.4	82.2	80.6
Liu et al. [28]	83.1	95.4	85.1
Cruz-Esquivel &Guzman Zavaleta[29]	71.1	87.3	83.3
The Proposed System	91.5	92.9	88.2

According to the results, the proposed system effectively summarized key content in surveillance videos compared to other methods on the VSUMM dataset. It also achieved the highest performance in abnormal event detection on the CUHK Avenue, Ped1, and Ped2 datasets, with AUC scores of 88.2%, 91.5%, and 92.9%, respectively. However, Liu et al. [28] obtained a higher AUC score than the proposed method on the Ped2 dataset.

5. Conclusion

A two-stage intelligent system is proposed in this paper for analysing educational surveillance videos, integrating summarization with abnormal event detection. In the first stage, static summarization, which leveraged several descriptors-color histograms, GLCM texture metrics, Zernike moment shape properties and local binary pattern (LBP) properties-produced a concise representation of the most important data from the original video. Evaluation was performed on the VSUMM dataset, achieving an accuracy of 78%, a recall of 79%, and an F1 score of 78.5%. These results indicate that the system is able to preserve useful content while review time is significantly reduced. The second phase consisted of a convolutional autoencoder (CAE) that processed the main frames from the summarization phase to detect abnormal events. On the CUHK Avenue and UCSD (Ped1 and Ped2) datasets, testing gave AUC values of 88.2%, 91.5%, and 92.9%, respectively, with high accuracy of detection testified. By integrating them, the proposed system improves safety and operational efficiency in educational settings by lowering viewing time and human effort while sending out prompt, focused alerts for any unusual activity.

Funding: "This work received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

References

[1] M. M. Elzain, "Surveillance Cameras Technologies at Imam Abdul Rahman Bin Faisal University," *Arab Journal for Scientific Publishing (AJSP)*, 2023.

[2] M. Malekar, "Detecting criminal activities of surveillance videos using deep learning," *Int. J. Sci. Res. Comput. Sci.*, vol. 7, no. 1, 2021.

[3] A. Kushwaha, M. Khare, R. M. Bommisetty, and A. Khare, "Human activity recognition based on video summarization and deep convolutional neural network," *The Comput. J.*, vol. 67, no. 8, pp. 2601-2609, 2024.

[4] M. Yarrarapu, N. Leelavathy, and D. Haritha, "Efficient video summarization through MobileNetSSD: a robust deep learning-based framework for efficient video summarization focused on objects of interest," *Multimedia Tools Appl.*, pp. 1-26, 2024.

[5] M. Javaid, M. Maqsood, F. Aadil, J. Safdar, and Y. Kim, "An efficient method for underwater video summarization and object detection using YoLoV3," *Intell. Autom. Soft Compute*, vol. 35, no. 2, 2023.

[6] B. Köprü and E. Erzin, "Use of affective visual information for summarization of human-centric videos," *IEEE Trans. Affective Comput.*, vol. 14, no. 4, pp. 3135-3148, 2022.

[7] S. Paulraj and S. Vairavasundaram, "Transformer-enabled weakly supervised abnormal event detection in intelligent video surveillance systems," *Eng. Appl. Artif. Intell.*, vol. 139, p. 109496, 2025.

[8] A. Mumtaz, A. B. Sargano, and Z. Habib, "AnomalyNet: a spatiotemporal motion-aware CNN approach for detecting anomalies in real-world autonomous surveillance," *The Visual Comput.*, pp. 1-22, 2024.

[9] P. Y. Ingle and Y. G. Kim, "Real-time abnormal object detection for video surveillance in smart cities," *Sensors*, vol. 22, no. 10, p. 3862, 2022.

[10] S. Saponara, A. Elhanashi, and A. Gagliardi, "Real-time video fire/smoke detection based on CNN in antifire surveillance systems," *J. Real-Time Image Process.*, vol. 18, pp. 889-900, 2021.

[11] J. Jung, S. Park, H. Kim, C. Lee, and C. Hong, "Artificial intelligence-driven video indexing for rapid surveillance footage summarization and review," in *Proc. Thirty-Third Int. Joint Conf. Artif. Intell.*, pp. 8687-8690, Aug. 2024.

[12] A. Sabha and A. Selwal, "CoSumNet: A video summarization-based framework for COVID-19 monitoring in crowded scenes," *Artif. Intell. Med.*, vol. 139, p. 102544, 2023.

[13] K. Muchtar, M. R. Munggaran, A. Mahendra, K. Anwar, and C. Y. Lin, "A unified video summarization for video anomalies through deep learning," in *Proc. 2022 IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, pp. 1-4, 2022.

[14] J. Chaki, N. Dey, J. Chaki, and N. Dey, "Histogram-based image color features," in *Image Color Feature Extraction Techniques: Fundamentals and Applications*, pp. 29-41, 2021.

[15] A. R. Zubair and O. A. Alo, "Grey level co-occurrence matrix (GLCM) based second order statistics for image texture analysis," *arXiv preprint arXiv: 2403.04038*, 2024.

[16] R. Mahum, A. Irtaza, M. Nawaz, T. Nazir, M. Masood, S. Shaikh, and E. A. Nasr, "A robust framework to generate surveillance video summaries using combination of Zernike moments and R-transform and deep neural network," *Multimedia Tools Appl.*, vol. 82, no. 9, pp. 13811-13835, 2023.

[17] Z. Sedaghatjoo, H. Hosseinzadeh, and B. S. Bigham, "Local Binary Pattern (LBP) optimization for feature extraction," *arXiv preprint arXiv: 2407.18665*, 2024.

[18] F. Tessari, K. Yao, and N. Hogan, "Surpassing cosine similarity for multidimensional comparisons: Dimension insensitive Euclidean metric (DIEM)," *arXiv preprint arXiv: 2407.08623*, 2024.

[19] H. Wang, Z. Han, X. Xiong, X. Song, and C. Shen, "Enhancing yarn quality wavelength spectrogram analysis: A semi-supervised anomaly detection approach with convolutional autoencoder," *Machines*, vol. 12, no. 5, p. 309, 2024.

[20] E. Şengönül, R. Samet, Q. Abu Al-Haija, A. Alqahtani, B. Alturki, and A. A. Alsulami, "An analysis of artificial intelligence techniques in surveillance video anomaly detection: A comprehensive survey," *Applied Sciences*, vol. 13, no. 8, p. 4956, 2023.

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025

[21] S. E. F. De Avila, A. P. B. Lopes, A. Luz Jr, and A. A. Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognit. Lett,* vol. 32, no. 1, pp. 56-68, 2011.

[22] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2895-2903, 2017.

[23] R. Hannane, A. Elboushaki, and K. Afdel, "Efficient video summarization based on motion SIFT-distribution histogram," in *Proc. 13th Int. Conf. Comput. Graph. Imaging Visualization (CGiV)*, pp. 312-317, Mar. 2016.

[24] Y. Bendraou, F. Essannouni, and A. Salam, "From local to global key-frame extraction based on important scenes using SVD of centrist features," *Multimedia Tools Appl.*, vol. 78, no. 2, pp. 1441-1456, 2019.

[25] M. U. Sreeja and B. C. Kovoor, "A multi-stage deep adversarial network for video summarization with knowledge distillation," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 8, pp. 9823-9838, 2022.

[26] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 733-742, 2016.

[27] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, pp. 439-444, Jul. 2017.

[28] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection – A new baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 6536-6545, 2018.

[29] E. Cruz-Esquivel and Z. J. Guzman-Zavaleta, "An examination on autoencoder designs for anomaly detection in video surveillance," *IEEE Access*, vol. 10, pp. 6208-6217, 2022.

DOI: https://doi.org/10.54216/FPA.200207
Received: January 05, 2025 Revised: March 09, 2025 Accepted: May 25, 2025