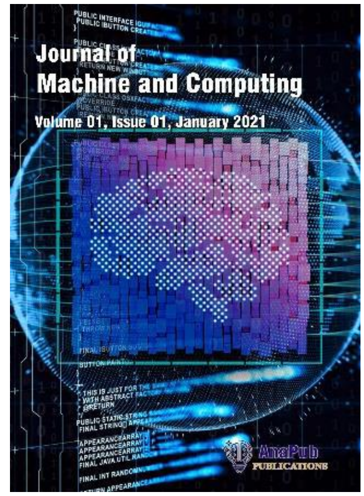
Journal Pre-proof



Detection and Recognition of Suspicious Multi-Task Human Action Identification from Preloaded Videos Using CCTV Stationary Cameras

Pavankumar Naik and Srinivasa Rao Kunte R

DOI: 10.53759/7669/jmc202505095

Reference: JMC202505095

Journal: Journal of Machine and Computing.

Received 16 May 2024

Revised form 06 December 2024

Accepted 20 March 2025

Detection and Recognition of Suspicious Multi-Task Human Action Identification from Preloaded Videos Using CCTV Stationary Cameras

Pavankumar Naik [1]*, & R. Srinivasa Rao Kunte [2]
[1] Research Scholar, Department of Computer Science & Engineering
Institute of Engineering & Technology, Srinivas University, Mangalore, India.
Orcid ID: 0000-0002-4295-0309; E-Mail ID: pavanraj.cse@gmail.com

[2] Research Professor, Department of Computer Science & Engineering
Institute of Engineering & Technology, Srinivas University, Mangalore, India.
Orcid ID: 0000-0002-5062-1505; E-Mail ID: kuntesrk@gmail.com

Abstract

Even more emphasis has been made on the use of video surveillance for sighting suspicious activities in the common places. As with other retrospective investigations, forensic investigations and riot inspections have normally required the use of automated offline video processing systems. However, development in the area that attempts at real time event detection has not been very impressive. Thus, the present work aims at developing a framework for processing raw video data gathered by a stationary colour camera within a given area to allow for real-time analysis of the observed activities. The suggested strategy begins with the acquisition of Object-level data by following and identifying objects and people in the scene via blob matching in real-time. Temporal features of those blobs are used to semantically characterize behaviours and events in terms of object and interobject motion attributes. A number of behaviours that are pertinent to public safety, such as lounging, gatherings, fainting, fighting, stealing, abandoned objects, occlusion, Abuse, Arrest and other activities available on UCF crime dataset. Were selected for the purpose of the demonstration of this method. The conclusions suggested in the work are based on experiments performed with currently easily accessible libraries.

Keywords: Suspicious Activity recognition, Loitering, Human Activity, behavior recognition, fainting, fighting, meeting, blob matching, CCTV Video Processing and occlusion.

I. INTRODUCTION

Today's world relies heavily on several programmes that help with many aspects of life. Video surveillance systems are one of the important application [1]. These systems are important so as to preserve the security of the persons. Thus, the goal of this project is to reduce the effects of riots and security force positioning.

Video surveillance systems compare one frame to the other to look for suspicious activities. They can be mechanical involving the use of security guards to perform surveillance, however, this is sophisticated, expensive and has high probability of causing accidents. The best scenario is to obtain a completely autonomous or video recording system.

Recognition of moving objects in videos is one of the simplest problems in videos analysis as illustrated in the fig. 1.

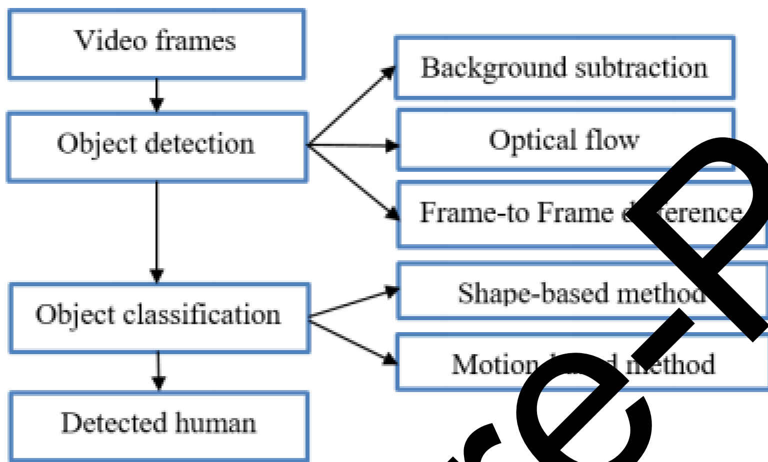


Fig.1 Fundamental steps for detecting moving objects in video.

The objects are recognized against still backgrounds; particular algorithms are used for that. These techniques include the “Optical Flow Method [3],” the “Background Subtraction Method [2],” and the “Frame-To-Frame Difference Method [4]. “ Among the above stated methods, the “Background Subtraction Method” is used in this study to detect the moving objects from the static background. Also because of several environmental factors such as illumination, glare, and cast shadows, this project encounters several challenges. Hence, object segmentation is a challenging problem, and it cannot be solved without the help of effective surveillance system. These problems are dealt with by way of morphological techniques in order to negate noise. Two methods are used to classify items: there is a method referred to as “motion-based classification” that sorts objects based on temporal information and another called “shape-based classification” which sorts basing on the spatial details.

It is an automated system that notifies the monitoring workers of undesirable behaviour depending on the configuration of the user. When it comes to constructing fully automated behaviour recognition, there is a set of problems that must be overcome. First, it is necessary to define and follow objects of interest in a scene like people and baggage. The second component is the formation of a regular procedure to characterize incidents. IT contexts such as fighting are complex because they are rife with multiple potential results. Labeling them is generally not easy.

What advancement are the findings of this study? It can detect the behavior at any abstraction level, unlike the most scientific works that use the machine learning approach to outline suspicious behavior This is a novel semantics-based approach as compared to the most existing academic work that uses machine learning to detect complicated action behavior down to pixels. Also, our processing system operates on real time. Indeed, most of the indicated work components can be described as not unique or revolutionary, yet our efficient integration of these aspects has a large worth. Although this integration is an area neglected in prior research, it must be done to obtain precise high-level conclusions.

Machine learning on the other hand employed reliable datasets for training and testing that can be expensive and is another disadvantage. It is rather challenging to obtain such a type of data especially regarding anomalous behaviours; however, this information is essential for setting threshold and parameters of classifiers. This approach is completely different and deduces its strategies from human thinking and logic, and thus does not entail any training. This second method, according to our thinking, is relatively more feasible and practical as compared to the said method above. It offloads from the system the need to prescribe extremely specific characteristics of learning such as matters having to do with the pruning constants of the decision trees that are often hard to set and even when set often require the services of a technician. From the perspective of the parameters used, it can be stated that the semantic approach uses more comprehensible and significant parameters than the other one. The basic technique of background subtraction that is applied in the current work involves the elimination of the foreground blobs from each picture. These blobs comprise of the live and non-live elements in a scene together with semantic information regarding events seen. Indeed, one gob can encompass an enormous quantity of neighboring or masked items. The conclusions are then made to split, track as well as categorize the items speaking to the essence of blob extraction. Last, the system predicates on the events which are exceptions and categorizes them.

1.1 DETECTION METHODS

a) Object Detection

The first component in visible surveillance system is the motion detection. Some of the approaches used to find moving items within a scene are the "Frame-to-Frame Difference Method" by [4], the "Background Subtraction Method" by [2] and the "Optical Flow or Movement Method" [illegible]. In the case of using the background models, motion detection main purpose is to separate moving objects from the stationary parts of the picture.

i) Frame-To-Frame Difference Method

This technique is performed by evaluating the differences in pixel intensity between two sequential frames of the captured photos to determine moving objects – persons or cars [4]. Since this strategy targets a system where there is movement, this type of value must be essential and vital because it boosts performance under dynamic contexts. Indeed, it could be considered as a somewhat less complicated version of what is known as the "Background Subtraction Method.

ii) Background Subtraction

With a view of isolating moving areas in a video with a fixed background, the method of background subtraction [2] is generally applied. It is 'wake aware' in terms of lighting changes and also pixel workable. The variables expressed a person's pixel coordinates from one step to the next; the centroid algorithm reflected the distance.

$$\text{Distance} = \sqrt{((a2 - a1)^2 + (b2 - b1)^2)}$$

Where

A2 = Earlier Pixel Position

A1: Pixel Position in Width

B2 = Earlier Pixel Position

B1: Pixel Position in Width

A moving object's speed can be calculated by squaring the centroid's path length with the frame rate of the video. The equation is:

$$\text{Velocity} = \frac{Distance\ Traveled}{Frame\ Rate}$$

iii) Optical Movement

Optical movement is another movement that deals with real-time technique for characterising the features through the distribution of velocity and object in the image. Though the employed method is successful, it is prone to noise and calls for the use of specialised instruments. This makes them prone to make errors especially if they are set in areas with constantly fluctuating light or complex background. Additionally, specialized equipment increases the probability of limiting the system's availability and raises the price of implementation.

b) Object classification

Contrast enhancement, object categorization [5] is used or partitioning specific areas from moving blobs [6]. The identification of gestures in videos is a frequent research area these days. Man recognition in videos has been the area of active research for the past several years.

For segmentation, "Background subtraction" and the "Gaussian Mixture Model"have been used in many two-dimensional video art works. Also, the ''spatio-temporal bag of features'' is employed by some researchers to predict action. In classification they use what they referred to as "non-linear support vector machine". The two sub-phases of the classification phase include a motion-based classification and a shape-based classification.

i) SHAPE-BASED CLASSIFICATION

Shape classification offers many descriptions like boxes and blob areas. It is noteworthy that R. T. Collins et al [7] applied common classification pattern in order to detect moving objects in films. Their method involves using a neural network classifier to partition the moving objects into sections. The input characteristics included the display of photos that were a combination of both low and high representation of form categorization parameters. The categorization method was used on all frames as well as on each moving blob within the frames, conclusions were shown in diagrams. To increase the outcome's reliability, temporal consistency was strictly adhered to during the classification process.

ii) MOTION-BASED CLASSIFICATION

The analysis of motion, which was discussed in [8], is important for separating mobile objects such as vehicles and less rigid objects like human beings using characteristics that belong to moving objects only. Apart from investigating the distinctive characteristics of targets, some studies explore temporal aspects to improve classification's efficiency and performance in changing contexts.

2. LITERATURE REVIEW

Many behaviors are involved in behavior recognition and for each of them, different detection mechanisms are needed. For example, the approaches that focus on such aspects of the crowd rather than behavioral characteristics are required for analysis, for instance, the movement of the crowd [9]. Due to the fact that short term human movements are relatively easier and cyclic for example; in gymnastic exercises [10], gestures [11] and space-time structures [13] different detection algorithms using body models [12] and space-time structures [13] must be employed.

Thus, the primary purpose of this article is to develop a method for the automatic identification of suspicious behaviour in the public space. They include; Loitering [14] abandoned objects and fights [15]. The actions that take time and the behaviours normally encompass a number of players. Therefore, practice challenges emerge concerning trajectory identification, identity tracking, and classification of objects.

This is one of the key issues affecting the discipline since the published works mostly just focus on the type of behaviour in question, and not its relative flexibility. For example, simple background subtraction techniques are used in solutions for the " abandoned luggage detection " [16], [17]. This method suffices the identification only of stationary foreground objects like walking, running, etc but is incapable of detecting behaviours such as fighting, loitering, etc. Furthermore, it is quite clear that many research works that present a general flow for behaviour identification pay much attention to the implementation approach but provide only an outline.

The behaviour analysis often employs the Grammar-based detection technique where situation- and temporal state-transition-based techniques such as HMM [6] and temporal random forest [3] have been considered. However, such machine learning techniques employed have some intrinsic constraints as far as classifiers are concerned, as well as activities they are presumably to classify. However, every classifier has its own advantages and/or disadvantages; for instance, the parallel and subevents are hard to be discovered by HMMs. Additionally, there is no extensively labelled dataset for training is another problem, especially when handling an enormous number of features and dynamism related to activities such as fighting.

While semantics-based recognition differs from conventional learning methods, it provides for a clearer description of events and peoples' understanding is improved. This method can accept both manual and trainable event definitions [18] and define the flows in plain language. The CASE (Case Frame Representation) paradigm that was developed by Fillmore in 1968 [19] allows formulating assertions in the natural language using case frames including agents, predicates, places, and objects. Hakeem et al [19] extended the interval algebra with temporal logic to enhance the CASE model, which resulted in the CASEE model [20] that allowed the modelling of activity with non-sequential sub event. Subsequently, Hakeem, and Shah have presented another CASE representation in [19], wherein a learning based probabilistic transition model replaces the tree like structure.

Like what we have done, Fernandez et al. [21] recently suggested a multilevel architecture, WFAM, with a knowledge taxonomy, which can be classified into the method that only depends on the logical description of the events and does not need training. They apply a different form from ours, what is called fuzzy metric-temporal Horn logic, to cope with the uncertainty. However, an integration of low-level feature processing evidences remained insufficient in their study. This gap is important since their experimental outcomes contain a comprehensive list of descriptors, entities, and events besides a variety of human motion, such as the positions of the body, motion, and face recognition. That is why the low-level processing is not complete in order to identify such behaviours as kicking the vending machines.

The work that we have carried out builds up on this simple and appropriate approach for real-time performance described in [2]. About this method, there is no need to train or learn as is the case with Fernandez et al. [21]. The events that Fuentes and Velastin [2] categorize as events in a transportation setting are position, trajectory, and split/merge events of a lower level. These descriptors define the foundation of the semantic approach, which we are going to employ. We present our analysis of all the used techniques and employed features, starting with the object detection and leading up to the detection of suspicious activity, as opposed to the approach described in [2].

Table 1: List of Methods used for human activity recognition from video data.

Methods	Description	Typical Accuracy	References
Spatiotemporal 3D Convolutional Networks (3D CNNs)	Leverages spatio-temporal features from video data for activity recognition.	85%	[Qiu et al., 2022][22]
Long Short-Term Memory (LSTM) Networks with Attention	A type of RNN that incorporates attention mechanisms to remember long-term dependencies.	87%	[Tang et al., 2023][23]
Graph Convolutional Networks (GCNs)	Models relationships and interactions between entities in a graph structure for activity recognition.	82%	[Chen et al., 2023] [24]
Transformer Models	Uses self-attention mechanisms to capture long-range dependencies in sequential data.	8[illegible]%	[Arnab et al., 2021][25]
Temporal Convolutional Networks (TCNs)	Employs convolutional layers to model temporal dependencies in sequential data.	83-90%	[Gülçehre et al., 2022]
Spatiotemporal Autoencoders	Uses autoencoders to learn spatiotemporal features for anomaly detection.	80-90%	[Chen et al., 2023]
Hybrid Deep Learning Models	Combines different types of neural	87-92%	[Wu et al., 2021]

	networks (e.g., CNNs and RNNs) for improved performance.		
Generative Adversarial Networks (GANs)	Uses a generator and a discriminator to learn robust feature representations for anomaly detection.	82-92%	[Sultani et al., 2022]
Multi-Stream Networks	Combines multiple streams of information (e.g., RGB, optical flow) for comprehensive analysis.	88%	[[illegible]in et al., 202[illegible]]
Recurrent Convolutional Networks (RCNs)	Integrates convolutional layers with recurrent layers to capture spatial and temporal features.	8[illegible]%	[Yue et al., 2023]

Challenges in detecting suspicious human activity in CCTV videos:Identification of the main issues relating to the use of CCTV videos in identifying suspicious human activities is as follows:

Environment Variability: The CCTV videos involves people under different light, mode of viewing, and with many persons in the back ground and thus distinguishing a improper activity is difficult.

Scale and Resolution: That is why, CCTV cameras can cover a large territory and the sizes as well as resolution of the objects, which is being observed, can be different. Security is one of the areas where the best algorithm must be implemented for the detection of the malicious activity in the same way at different scales and resolutions.

Information received from CCTV cameras should be processed in real-time so that any suspicious movement can be attended to in a favorable way. However, the real time processing of data of such a volume even limited to video data only poses computational challenges.

Complex Interactions: Interactions of things and human may entail different actions and relations within the same level, This means that human activities in this level are compound activities. Specifically, estimating deviations or abnormalities in the population within crowded places is difficult because of occlusions and overlapping trajectories.

Anomaly Detection: Anomaly detection is implemented in general for the purpose of the normal and the abnormal behaviour distinction. However, it becomes rather complicated most of the times to tell how an abnormality is defined and where to get labelled data for training of the anomaly detection models.

Privacy Concerns: When analyzing CCTV entails observation of people in public areas then the questions of privacy are touched. Particularly, much care and ethical measure should be put in this case so as to balance the human rights to data privacy and the societal security needs for surveillance.

Limited Labelled Data: The training datasets for the suspicious human activity in CCTV video is often smaller and less diverse with other related training datasets. From the above discussion, it is understood that for getting better accuracy and model robustness level, sufficient labelled data are required most of the time that is not always available.

Adaptability to Context: This is because behaviour pattern may differ from one place to another due to multifaceted culture of the various regions. Therefore, the models that have high versatility are beneficial and those which can address the cultural factors that affect the decision making.

2.2 Scope of the work:

Algorithmic Development: This opens a wonderful chance to create progressive algorithms which can analyze CCTV videos to identify suspect human activity.

A Integration of Multiple Sensors: For instance, fusing with video from other sources like infrared, motion, and audio sensors would increase surveillance systems’ detection ability and redundancy.

Machine Learning and AI: Thus, deep learning, reinforcement learning, transfer learning and other such machine learning and artificial intelligence techniques can be employed to detect suspicious activities with better efficiency.

Real-time Monitoring systems: Designing methods for real-time observing enhanced with the capability to analyze CCTV footage in real time and alert security personnel when certain undesirable activities are observed can hugely impact various aspects of security.

Privacy-Preserving Technologies: Many privacy issues regarding CCTV surveillance can be solved with the help of privacy-preserving technologies, which include anonymization of data, encryption of data, and differential privacy while the suspicious activity identification issues still can be solved effectively.

Benchmark Datasets and Evaluation criteria: Perhaps, creating common sets of reference data and assessment guidelines to identify suspicious behavior in video surveillance footage could enhance the reproducibility of the researchers' results.

Topics like the use of CCTV surveillance and rules of responsible use can minimize the violations of people's right to privacy and encroachment on civil liberties. Solving these problems, one can establish more efficient and accurate algorithms for identifying suspicious human actions in a video surveillance system with the improvement of the security and performance associated with public safety.

3. Methodology

In this section we will discuss more details the methods of blob matching and how they can be used to spot abnormal human activities in CCTV tapes. The process examined consists of a number of stages, ranging from the basic ones like object detection to the final stage of activity analysis, which is also indispensable for the recognition of suspicious behaviors.

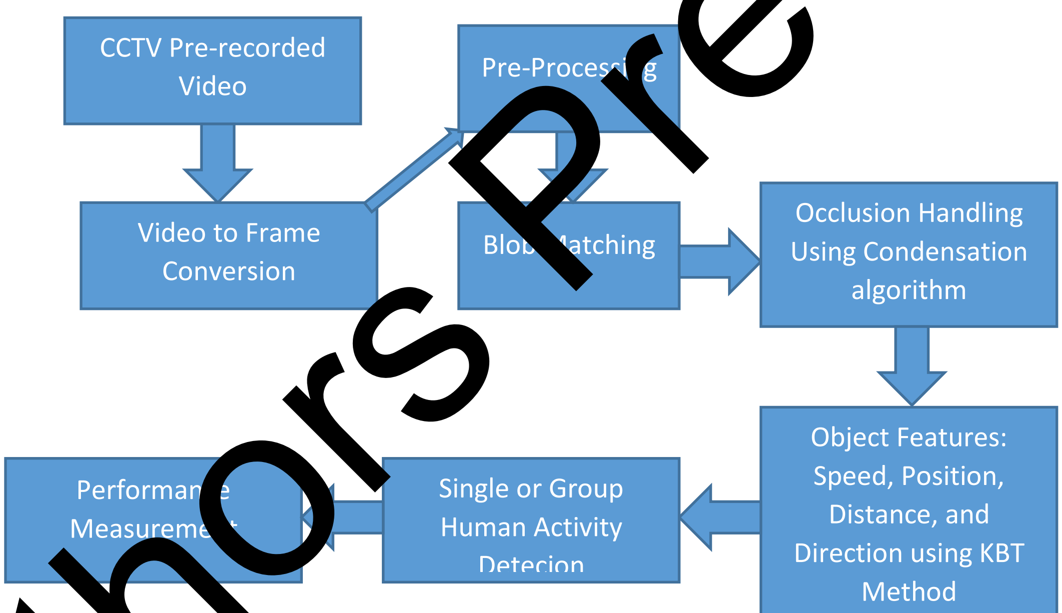


Fig 1: Proposed System for Single or Group of Suspicious Human Activity Detection

Video to Frame Conversion: Input video from CCTV prerecorded are divided and captured as n number of frames per second and each of those captured frames helps in carrying out an image analysis.

Pre-processing: To perform the image processing, they first involve the preprocessing of the input video frames received. The purpose of this preprocessing is to eradicate several types of noise from these photos; however, the predominant and most common type is the salt and pepper noise. This sort of noise looks as if it is random white and black dots scattered throughout the photos. Also, during pre-processing, it involves issues such as erasing random pixels and enhancing the quality of images. Medians filters are used for removal of the noise, and that is related to the fact that it identifies noisy pixels, and change it to the mean value of the neighbouring pixels. This phase is essential in order to obtain high-quality images that will be free from artefacts and ensure the foundation for subsequent work. Since the noise is filtered, each of the individual fames that are preprocessed are then forwarded to blob matching.

Blob Matching for Suspicious Activity Detection

1. Initialization

Frame Captured: They work on extract frames as an input.

Grayscale Conversion: This will help in decreasing the computation time and therefore frames should be converted to grayscale.

2. Background Subtraction

Goal: Find people involved in motion (maybe, it is a suspicious human).

Static Background Modeling: Perform a Gaussian Mixture Models (GMM) on the background to create a model of it.

Foreground Extraction: Reduce the current frame with background model and get moving objects. Gaussian Mixture Models (GMM) by which pixel intensity over time is modeled for differentiating between the foreground and the background.

3. Blob Detection

Goal: Identify & segment different blobs, which are present in the moving objects of the foreground.

Thresholding: Then apply simplified binary threshold which divides moving objects from the background of the foreground mask.

Morphological Operations: Apply dilation and erosion to the blob detection, this stage would help you to eliminate some noise and also sharpening the blobs shape.

Methods for Determining Threshold Values

1. Otsu's Method:

In Otsu's method, the threshold searching is performed piece-meal without the need for manual intervention and the threshold value is chosen in such a way that the variance within the available classes, particularly the background and foreground classes, is maximized.

```
import cv2
import numpy as np
# Load the image
image = cv2.imread('frame.jpg', cv2.IMREAD_GRAYSCALE)
```

```
# Apply Otsu's thresholding
ret, thresh = cv2.threshold(image, 0, 255, cv2.THRESH_BINARY + cv2.THRESH_OTSU)
# Display the threshold value
print("Otsu's threshold value:", ret)
```

2. Mean or Median-based Thresholding:

The best method of determining the threshold of the image is by using the mean or the median of the intensity values of the image.

```
mean_val = np.mean(image)
ret, thresh = cv2.threshold(image, mean_val, 255, cv2.THRESH_BINARY)
```

3. Histogram Analysis:

With reference to histogram analyze the image to find out the threshold value which is equivalent to background & foreground.

```
import matplotlib.pyplot as plt
# Compute the histogram
hist = cv2.calcHist([image], [0], None, [256], [0, 256])
```

4. Adaptive Thresholding:

Adaptive thresholding should be used to set the threshold different for different regions of the corresponding image.

```
adaptive_thresh = cv2.adaptiveThreshold(image, 255,
cv2.ADAPTIVE_THRESH_GAUSSIAN_C, cv2.THRESH_BINARY, 11, 2)
```

Another method of thresholding is the adaptive thresholding that is applicable when there are contrasting shadows in the picture. Unlike in the global thresholding where one threshold value for the whole image is determined, in adaptive thresholding each pixel's threshold value is computed from its own local neighborhood. Gaussian-weighted-Sum: As for the recall parameter, when using cv2. ADAPTIVE_THRESH_GAUSSIAN_C, the threshold value with this method is the sum of the pixel under consideration and its neighbors and weights given to the neighbor pixel are high if they are close to the pixel under consideration. cv2. THRESH_BINARY means that if the intensity level of the pixel is greater than the threshold it is equalled to 255 else it is equal to 0. Other options include cv2. THRESH_BINARY_INV, cv2. THRESH_TRUNC, cv2. THRESH_TOZERO, and cv2. The floating variables THRESH_TOZERO_INV which specify the pixel and intensity values based on the threshold. Block Size (11): The number of neighboring pixels considered in each case, these were 11X11 local neighborhood around a specific pixel.

Constant (2): Minus the obtained values to calculate the mean or weighted sum; then, fixed amount to create the threshold for a specific pixel.

The optimum method is to decide the blob detection threshold value for the particular CCTV video footage and its application based not in the form of a single number, but as a range of applicable values. Starting with adaptive methods like Otsu's method or adaptive thresholding can then be used to obtain good starting points. It is crucial always to fine-tune and do the testing in the real operational field in a bid to get the best results.

4. Feature Extraction

Goal: Obtain the attributes that are distinctive in identifying each blob completely.

Shape and Size: Find out the blob area, the circumferences of the blob as well as the ratio of the blob's width to its breadth.

Bounding Box: Find out how large the bounding box should be.

Color and Texture: Analyze the histograph and texture of the blob's colour distribution.

5. Blob Matching and Tracking

Goal: This feature allows tracking the same blobs from one frame to another to determine the movement and the behaviour of the objects.

Matching Criteria: Act on spatial distance, size, shape, color, and the motion trajectory of the blobs to assign the blobs to the frames. Particle filter also used to estimate future position of a blob and optimize according to the new arriving data.

Handling Occlusions: Apply the proximity calculation when estimating positions in short-term occlusions to generate accurate predictions.

Split and Merge Events: Identify when blobs are split into many smaller blobs or when two or more blobs combine into one and adjust the tracking.

6. Behavior Analysis

Goal: In this step, identify by using several algorithms and heuristics, patterns of tracked blobs motion and their interactions with environment are analyzed to recognize dangerous activities.

Pattern Recognition: Even describe patterns of movement and behavior that are contrary to the usual or customary ones.

Examples of Suspicious Activities: Stalling, sudden movements, running and racing into areas they are not supposed to be in.

Contextual Analysis: Usefulness of the results increases if the context in which it has been performed is taken into consideration (for instance, location, time).

Example: Analyzing if a person is loitering near an ATM at unusual hours.

Occlusion Handling Using Condensation algorithm

The Condensation algorithm, or "Conditional Density Propagation," is an algorithm used in computer vision for tracking the objects whose state can be described by probability density functions. Unlike most tracking algorithms which may just be tracking one estimate of the state, the use of the Condensation algorithm means that there are many state hypotheses held to accommodate ambiguity.

In the general case, when it is necessary to determine the object boundary depending on the change in the object's shape and appearance due to occlusion, changes in the viewpoint or some deformations at different frames using the Condensation algorithm, we can reveal the problem of effective management of such cases.

```
# Initialize particles
particles = initialize_particles()
weights = initialize_weights()

for each frame in video:
    # Prediction step
    predicted_particles = []
    for particle in particles:
        predicted_particle = predict(particle)
        predicted_particles.append(predicted_particle)
        # Observation step
    for i, particle in enumerate(predicted_particles):
        likelihood = compute_likelihood(particle, frame)
        weights[i] = likelihood

    # Update step
    weights = normalize(weights)

    # Resampling step
    particles = resample(predicted_particles, weights)

    # Estimate object boundary from particles
    estimated_boundary = estimate_boundary(particles)

    # Display or process the estimated boundary
    display(estimated_boundary, frame)
```

KBT Method: Kernel-based tracking also called as Mean Shift algorithm is a most efficient method of tracking the object in the video sequences. For the purpose of enhancing the resemblance between the target model and the candidate regions that are present in the subsequent frames, this strategy relocates a search window. Although it can be very easily understood, it has immense potential in real-time scenarios and even helps in direction, speed, and distance of objects in that frame.

Single or group Human activity detection: To find the matching of visual activities of individual or group behaviours, the Visual Feature Matching (VFM) method is used. Visual feature matching is a powerful method for human activity detection in video sequences. By detecting and matching keypoints on human figures, the method can track motion patterns and recognize activities. Where this helps in find the key points of the object based on which pattern analysis will be done by tracing the key points of the object example a walking activity might be recognized by alternating movements of leg keypoints.

The activities in the video are recognised and semantically characterised using the keypoints that have been gathered and sent into the system. The preceding results, which record the features of the supplied behavioural object and semantic scene pair, are updated every unit time. Based on the record's contents, a set of precise defined conditions for a unique action of interest are tested. If the exact requirements in the existing problem statement are met, the behaviour is detected. Individual interest-related behaviours are discussed below, along with relevant instances. This includes Abuse, Arrest, Arson, Assault, Burglary, Explosion, Normal Video, Road Accident, Robbery, Shooting, Shoplifting, Stealing and Vandalism. Performance of such a system can be conveniently measured by utilizing three major performance metrics: Accuracy, Sensitivity (Recall), and Specificity. Accuracy specifies how frequently the system successfully detects both the suspicious and the non-suspicious activities. High accuracy indicates that the system is good overall, but it does not inform us about individual errors (false positives or false negatives). Sensitivity refers to how well the system is able to detect real crimes without failing to detect them. High Sensitivity implies fewer false negatives (FN), meaning nearly all criminal activity is detected. Low Sensitivity implies the system is failing to detect real crimes, which can be risky. Specificity is a measure of how effectively the system resists false alarms by accurately labeling normal activities. High Specificity implies less FP (false positives), so that normal activity is not falsely identified as crime. Low Specificity implies excessive false alarms with unnecessary responses. Below mentioned table represents the equation for find the performance metric on Accuracy, Sensitivity & Specificity.

Metric	Equation	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Measures overall correctness of the system.

Sensitivity (TPR)	$\frac{TP}{TP + FN}$	Measures how well crimes are detected.
Specificity (TNR)	$\frac{TN}{TN + FP}$	Measures how well normal activities are correctly classified.

Where:

- **TP (True Positives)** → correctly predicted suspicious activity.
- **TN (True Negatives)** → correctly predicted normal activity.
- **FP (False Positives)** → incorrectly predicted suspicious activity when it was normal.
- **FN (False Negatives)** → incorrectly predicted normal activity when it was suspicious.

Table 1 represented the accuracy, sensitivity and specificity for UCF Crime Dataset.

Activity	Trained Data	Test Data	Accuracy	Sensitivity (TPR) (%)	Specificity (TNR) (%)
Abuse	200	75	98.3	97.99	97.95
Arrest	200	75	98.7	98.21	98.34
Arson	200	75	[illegible]	98.08	98.11
Assault	200	75	[illegible]	98.12	98.04
Burglary	200	75	98.4	98.15	97.97
Explosion	200	75	98.7	98.45	98.31
Fighting	200	75	98.8	98.58	98.32
Normal Videos	200	75	99.1	98.64	98.76
Road Accidents	200	75	98.9	98.52	98.51
Robbery	200	75	98.8	98.39	98.39
Shooting	200	75	98.7	98.49	98.26
Shoplifting	200	75	98.8	98.31	98.26
Stealing	200	75	98.7	98.25	98.34
Vandalism	200	75	98.8	98.54	98.35

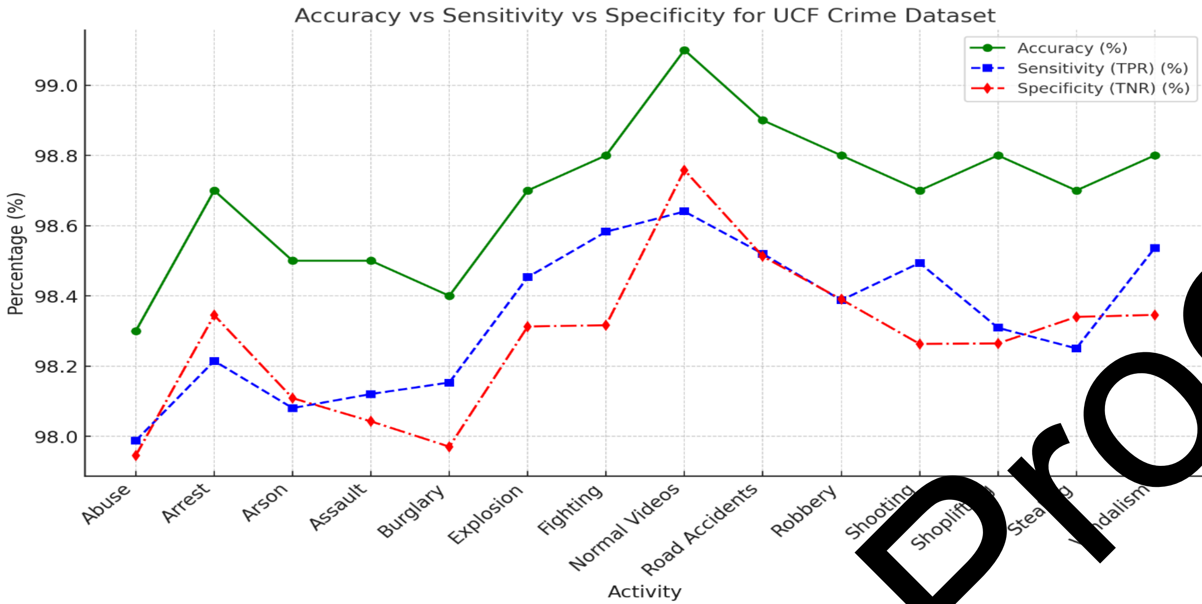


Fig.2 represents the graphical analysis of Accuracy, Specificity and Sensitivity on UCF Crime Dataset.

Results

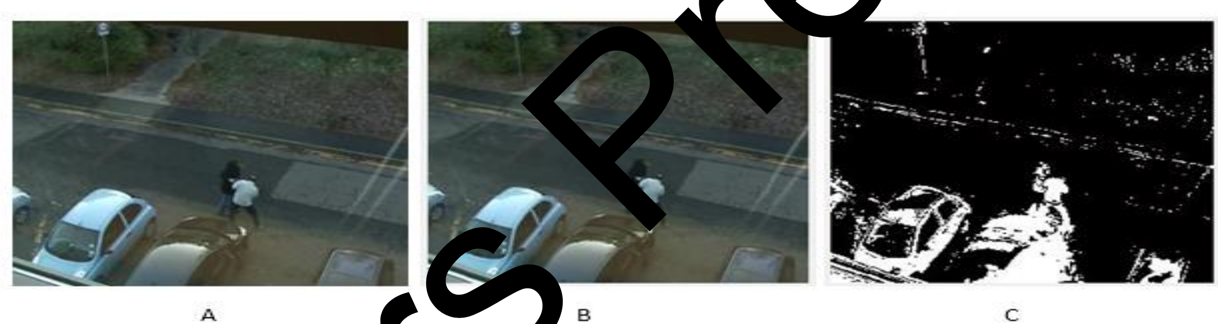
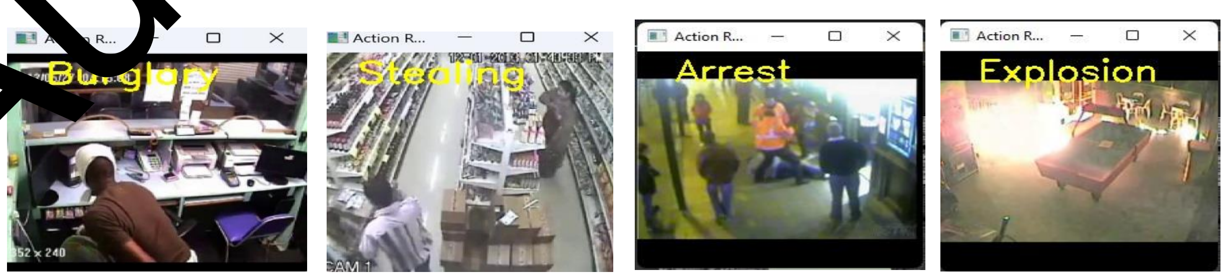


Fig.2 Represents the one of the frame of the CCTV video

In Fig.2, (A) represent Original frame extracted from CCTV video, (B) Represents Preprocessed Frame and (C) represent Background Subtracted Image. Similarly all other frames are processed and key features are preserved from all the frames and finally the suspicious activity will be recognized efficient after processing the frames with all the steps which are included in Fig.1. UCF Crime Dataset results are as shown below in fig.2.1.



(A) (B) (C) (D)

(E) (F) (G) (H)

(I) (J)

Fig2.1 – A) Burglary B) Stealing C) Arrest D) Explosion E) Fighting F) Road Accident G) Robbery H) Abuse I) Assault J) Shooting

Even proposed system is capable to detect the group activities as Loiter & Abandoned, Meet, Walk Together with Occlusion, Abandoned, Faint, and Walk Together without Occlusion, Meet, Fight and Meet, Fight & Loiter. In addition to that system is capable to detect the Objects like Gun and Knife.

4. Comparative Study with Existing Models

This study evaluates the accuracy of the proposed models and compares them with other deep learning models. There's still limited research on using the UCF-Crime dataset for anomaly detection. Table 2 highlights the accuracy scores of both the proposed models and other deep learning approaches for detecting abnormal and suspicious activity in the UCF-Crime dataset. From the table2 We can see that Resnet50 & ConvLSTM model has a accuracy of 81.71%, comparative to ResNet18, ResNet34 and ResNet50 with SRU the accuracy is high for ResNet50 with SRU for UCF Crime Dataset. 2DCNN model gives a result of 87.55% and ConvGru-CNN model gives accuracy of 82.22%, comparative to all these models our proposed model gives a results as 98.7% with minimum improvement of 7%.

Table 2: Performance Analysis Matrix

Method	Accuracy
ResNet50 and ConvLSTM	81.71 %
ResNet18+ SRU	89.08 %
ResNet34 +SRU	90.09 %
ResNet50 + SRU	91.64 %
2D-CNN and ESN	87.55 %
ConvGRU-CNN	82.22 %

Proposed Model	98.73 %

5. Conclusion

It is a structured procedure of detecting, tracking, and analyzing the moving objects in the CCTV videos to detect the suspicious human activity Blob matching. If each step is improved and the latest methods are applied, the efficiency and effectiveness of recognition of the forbidden activities can be significantly increased and, therefore improve the surveillance and security levels. Few of the Challenges are Shifting of lights and shadows, the changing of the weather conditions can complicate also the background subtraction and thus the blob detection plays a vital role to extract the key features, Several cases of people interference make blob matching and behavior analysis difficult especially where many individuals are involved and In the case of long term occlusions where an object is completely lost novel approaches have to be taken to manage them.

Future Scope

Blending the process with blob matching and deep learning models to improve the models and results obtained from them. Example: Combined with the use of CNNs to detect the objects in the initial frames and later using the more conventional blob matching to track the objects. Using the combination of two and more strategies (for example, using GMM for background subtraction and deep learning for behavior analysis). Edge Computing: Deploying edges to handle greater compute and render various algorithms which always require real-time performance.

Reference

[1] Ihaddadene, N., & Djeraba, C. (2008, December). Real-time crowd motion analysis. In *2008 19th International Conference on Pattern Recognition* (pp. 1-4). IEEE.

[2] Divya J, M.E, G. H, Pune, Prof. Dr. R.S.Bichkar (2015)," *Automatic Video Based Surveillance System for Abnormal Behaviour Detection".* IJSR, Vol: 4 Issue: 7, pp 1743- 1747.

[3] R.Naveen Kumar and Dr.S.Chandrakala (2016*),"DETECTING AGGRESSIVE HUMAN BEHAVIOR IN PUBLIC ENVIRONMENTS "*Department of computer science and engineer, Rajalakshmi Engineering College Chennai, Tamil nadu. ISSN: 0976-1353 Volume 22 Issue 2.

[4] C.Srinivas Rao, P.Darwin (2012), "Frame Difference and Kalman Filter Techniques for Detection of Moving Vehicles in Video Surveillance", Vol. 2, Issue 6, pp.1168-1170, (IJERA).

[5] Mohamed Benouis, Senouci Mohamed, Redouane Tlemsani, Lotfi Mostefai (2006), "Gait recognition based on model-based methods and deep belief networks", International Journal of Biometrics8(3/4):237 DOI:10.1504/IJBM.2016.082598.

[6] Christopher M. Bishop, M. Jordan, J.Kleinberg, B. Schölkopf (2006),“Pattern Recognition and Machine Learning” Springer.

[7] R.T. Collins and et al (2000), “A system for video surveillance and monitoring”. Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

[8] Srinivasa Rao Chalamala and Prasanna Kumar (2016), “A Probabilistic Approach for Human Action Recognition using Motion Trajectories”, IEEE, India, pp 185-190.

[9] N. T. Siebel and S. J. Maybank, “The ADVISOR visual surveillance system,” in Proc. ECCV Workshop ACV, 2004, pp. 103–111.

[10] Z. Zhang, T. Tieniu, and H. Kaiqi, “An extended grammar system for learning and recognizing complex visual events,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 2, pp. 240–255, Feb. 2011.

[11] D. Demirdjian and C. Varri, “Recognizing events with temporal random forests,” in Proc. Int. Conf. Multimodal Interfaces, Cambridge, MA, 2009, pp. 293–294.

[12] D. Weinland, R. Ronfard, and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” Comput. Vis. Image Underst., vol. 115, no. 2, pp. 224–241, Feb. 2011.

[13] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as space–time shapes,” in Proc. 10th IEEE ICCV, 2005, vol. 2, pp. 1395–1402.

[14] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs, “Detection of loitering individuals in public transportation areas,” IEEE Trans. Intell. Transp. Syst., vol. 6, no. 2, pp. 167–177, Jun. 2005.

[15] S. Blunsden, E. Andrade, and R. Fisher, “Non parametric classification of human interaction,” in Proc. 3rd Iberian Conf. Pattern Recog. Image Anal., Part II, Girona, Spain, 2007, pp. 347–354.

[16] P. Fatih, “Detection of temporarily static regions by processing video at different frame rates,” in Proc. IEEE Conf. AVSS, 2007, pp. 236–241.

[17] A. Singh, S. Sawan, M. Hanmandlu, V. K. Madasu, and B. C. Lovell, “An abandoned object detection system based on dual background segmentation,” in Proc. 6th IEEE Int. Conf. AVSS, 2009, pp. 352–357.

[18] A. Hakeem and M. Shah, “Learning, detection and representation of multi-agent events in videos,” Artif. Intell., vol. 171, no. 8/9, pp. 586–605, Jun. 2007.

[19] C. J. Fillmore, The Case for Case, 1967. [Online]. Available: http://linguistics.berkeley.edu/~syntax-circle/syntax-group/spr08/fillmore.pdf

[20] J. F. Allen, “Maintaining knowledge about temporal intervals,” Commun ACM, vol. 26, no. 11, pp. 832–843, Nov. 1983.

[21] Ashwin Shenoy, M., and N. Thillaiarasu. "Enhancing temple surveillance through human activity recognition: A novel dataset and YOLOv4-ConvLSTM approach." Journal of Intelligent & Fuzzy Systems Preprint (2023): 1-16.

[22] Qiu, Z., Yao, T., & Mei, T. (2022). Learning spatiotemporal features with 3D convolutional networks. IEEE Transactions on Multimedia.

[23] Tang, J., Lin, K., & Su, Y. (2023). Attention mechanisms in LSTM networks for action recognition. Neural Networks.

[24] Thillaiarasu, N., and Ashwin Shenoy. "Enhancing Security Through Real-Time Classification of Normal and Abnormal Human Activities: A YOLOv7-SVM Hybrid Approach." IAENG International Journal of Computer Science 51.8 (202[illegible]).

[25] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. International Conference on Computer Vision.