Received: October 1, 2024. Revised: November 4, 2024. 335

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

YOLO v7 - Distance Intersection of Union for Detecting Objects and Anomalies in Video Surveillance

Kiran Kalla[1]* **Jaya Suma Gogulamanda[2]**

[1]*Department of Computer Science Engineering,*
Jawarharlal Nehru Technological University, Kakinada, Andhra Pradesh – 533003, India
[2]*Department of Information Technology, JNTUGV, Vizianagaram, Andhra Pradesh – 535003, India*
* Corresponding author's Email: Kallakiran1974@gmail.com

Abstract: Surveillance videos are essential for crime prevention and public safety. However, defining abnormal events remains challenging, which hinders their effectiveness and limits the use of supervised techniques. The existing method have difficulty in accurately detecting and tracking the objects, that minimizes the detection and tracking performance. In this research, the You Only Look Once v7 (YOLO v7) - Distance Intersection of Union (D-IoU) and Earth Mover Distance (EMD) approach is proposed to detect and track the objects and anomalies in video surveillance. The D-IoU loss function is used in the YOLO v7 model improves the precision of bounding box prediction by considering the distance between centres of the bounding box, which is useful for the accurate localization of object. Then, the features are extracted by using the Inception V3 approach that extracts the meaningful features that help differentiate the anomalies in the detected object. To detect the anomalies, the EMD method is used which effectively detects the anomalies. The YOLO v7 – D-IoU and EMD approach obtained 96.1% accuracy on UCSD Ped 1 datasets and 98.8% accuracy on UCSD Ped 2 dataset. The proposed method showed effective performance when compared to conventional methods like Three-Dimensional Convolutional Neural Network (3D-CNN).

Keywords: Bounding box, Distance intersection of union, Earth mover distance, Video surveillance, You only look once v7.

1. Introduction

Automatic video surveillance act as the initial phase in various Artificial Intelligence (AI) applications developed for tracking human crowds and analyze the behavior of crowd [1-3]. The devices of automatic surveillance quickly detect critical and unusual situations in crowded environments enabling accurate and timely decisions for emergency and safety controls [4]. Therefore, the system of surveillance is significant in crowded and complex environments such as busy streets, train stations and airports to detect and control panic behavior triggered by violent events, ensuring public security and safety [5, 6]. The growth of Closed-Circuit Television (CCTV) cameras has expanded exponentially due to vertical and horizontal extension of surveillance coverage and increased usage in both industrial and urban fields [7]. However, observing human behaviours is impossible and ridiculous for precise analyze and evaluation of each video stream [8]. Numerous computer-vision-based initiatives have been developed recently to track and detect the objects in surveillance videos and enables to analysis their behavior [9].

The analysis of anomaly behavior is the recent research area in computer vision with essential applications including the automatic detection of avoidance and panic behaviors, violent and chaotic or riots crowd behaviour [10]. The performance of advanced surveillance architectures has faced challenges in correctly identifying abnormal changes in images, To address this, different intelligent surveillance techniques have been developed, offering much more efficient solutions to detect such changes [11, 12]. Two phases of typical diverging motion patterns are implemented such as circular and

straight motions and divergence of crowd occurs when the crowd departs from a circular or straight walking way [13,14]. The main concentration of anomaly detection established the normal behaviour patterns of users and detected the intrusions by comparing and matching patterns with monitoring systems [15]. The YOLO v7 used for object detection have improved accuracy and precision while compared with YOLO v4 and YOLO v5 in different object detection tasks. YOLO v7 combined with high speed, accuracy, effective computation makes it significant for object detection like video surveillance and object detection.

The significant contributions of the research are described below.

- The You Only Look Once v7 (YOLO v7) - Distance Intersection of Union (D-IoU) and Earth Mover Distance (EMD) method is proposed to detect and track the object and anomalies in video surveillance, which effectively detects and tracks the objects and anomalies in video frames.
- The D-IoU based loss function is used in the YOLO v7 model which effectively improves the prediction performance of bound boxes for object detection.
- The Inception v3 based feature extraction method is used to extract the deep features, which is useful to differentiate the normal anomalies in the video frames.

This research is organized as follows: Section 2 examines the literature review of existing research. Section 3 presents the process and details of object tracking and anomaly detection. Section 4 provides the results and the discussion and conclusion of this research paper are given in Section 5.

2. Literature review

Abdullah [16] presented the Improved Watershed Transform (IWT) and then the Conditional Random Field (CRF) for attaining multi-object segmented frames through processing the pixel-level labeling. The Social Force technique was implemented for extracting the contextual architecture of the environment through combination of selected specific histogram optical stream and inner force method. The multi-object classification was processed through Feature Pyramid Network (FPN) and incorporated the contextual architecture of the environment. Jaccard similarity was used for deciding to detect abnormal and identify unusual objects. But, the presented method was ineffective in complex crowded frames because of size depended object detection.

Niaz [17] suggested the Three-Dimensional Convolutional Neural Network (3D-CNN) with autoencoder, with architecture of encoder-decoder which learned the spatiotemporal representation and reconstructed the input by latent space. The skip connections between the encoder and decoder blocks facilitated to transmission of data across different scales of feature representations, improving the redeveloping process and enhancing overall performance. The structure includes the modules of spatial attention which highlight the significant regions in input, enabling enhanced anomaly detection. However, the suggested method did not accurately detect the human behaviours, which minimizes the method's effectiveness.

Le and Kim [18] developed the spatial and temporal branch in defined network which exploited both spatial and temporal data efficiently. The network includes structure of residual autoencoder, contains deep CNN based encoder and multiple phase channel attention depended decoder, trained in unsupervised manner. The temporal shift technique was utilized to exploit temporal features, and the dependency of context was extracted through modules of channel attention. But the method didn't fine-tune and train the detector, limiting the method's accuracy.

Gayal [19] introduced the Hierarchical based Social hunting optimization tuned Deep Convolutional Neural Network (HiS-Deep CNN) to detect the anomaly in videos. The enhancement of detection classifier was depended on training algorithm, the HiS which was developed depended on integrated characteristics from the timber wolf and Ateles Geoffrogis search agents. But introduced method was computation complex and needed huge training.

Ilek and Dener [20] implemented the effective frame level Video Anomaly Detection (VAD) technique depended on Transfer Learning (TL) and Fine-Tuning (FT) method. The anomalies were detected by CNN based DL algorithms and it was trained through TL and FT algorithms. But the method was difficultly in accurately detecting the objects, that minimizes the detection and tracking performance.

3. Proposed method

The effective DL based methods are developed to detect and track the objects and anomalies in video surveillance. The dataset used in this research are

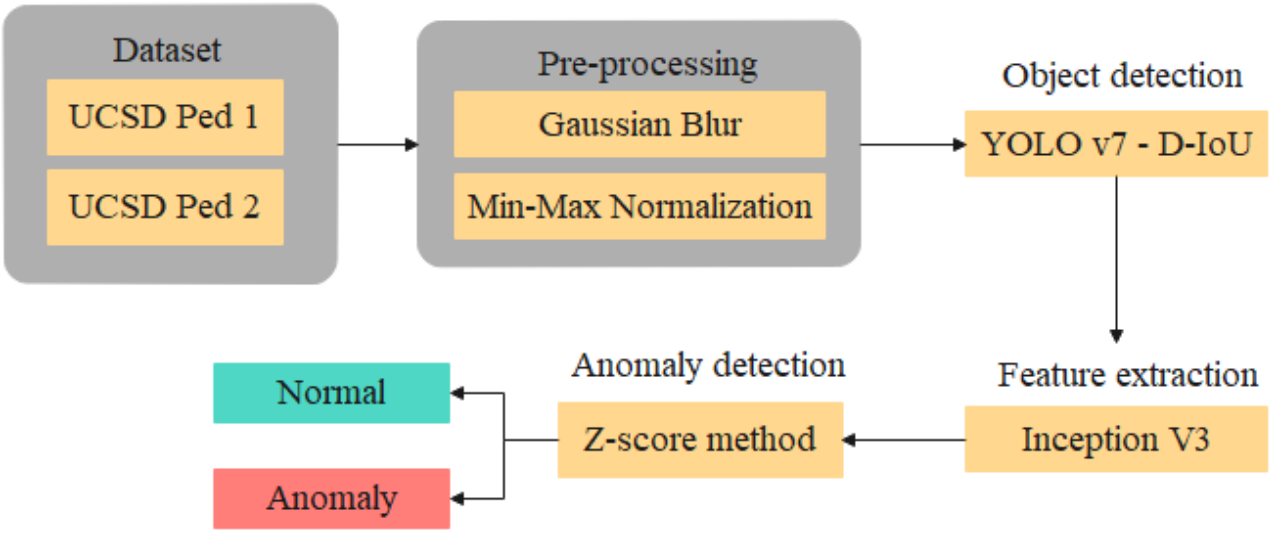

Figure. 1 Process of object and anomaly detection

UCSD Ped 1 and UCSD Ped 2 datasets and it is pre-processed by using Gaussian Blur filter and Min-Max normalization. Then, the object is detected and tracked by using the YOLO v7 - D-IoU. Next, the meaningful deep features are extracted by using the Inception V3 network and at last the anomalies are detected by using the EMD method. Fig. 1 represents the process of object detection and anomaly detection.

3.1 Dataset

The datasets used for object tracking and anomaly detection are University of California San Diego (UCSD) datasets, which have a set of video sequences [21], CHUK [] and Shanghai Tech []. These sequences are used for object tracking and anomaly detection. The detailed explanation is described as follows.

3.1.1. UCSD Pedestrian 1

This dataset includes 14,000 frames separated from 70 video sequences. This dataset is separated into 34 video sequences for training and 36 video sequences for testing. This dataset has 40 abnormal events like little carts, bikers and mini trucks for abnormal event detection in surveillance of videos. Fig. 2 represents the sample images in UCSD Ped 1 dataset.

3.1.2. UCSD Pedestrian 2

This dataset includes 4,560 frames separated from 28 video sequences, in that 16 video sequences are used for training and 12 video sequences are used for testing. This dataset particularly concentrates on anomalies in bikers with 12 occurrences of events.

Moreover, this dataset is used to evaluate the abnormal events involved in bicycles. Fig. 3 represents the sample images in the UCSD Ped 2 dataset.

3.1.3. CUHK Avenue dataset

This dataset records the pedestrian movements in Chinese University of Hong Kong (CUHK). This dataset is attained by fixed camcorder with size of 360×640 pixels and frame rate of 25 FPS. There include 15 fragments, every fragment includes approximate of 2 mins.

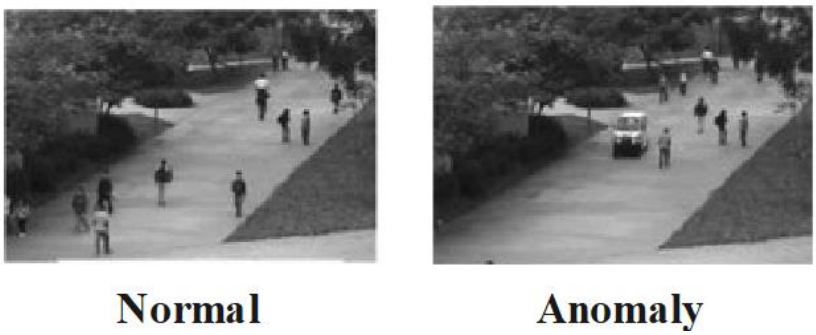

Figure. 2 Sample images in UCSD Ped 1 dataset

Figure. 3 Sample images in UCSD Ped 2 dataset

3.1.4. Shanghai Tech dataset

This dataset is the comprehensive resource of abnormal event detection in the video surveillance, totally 317,198 frames from 437 sequence of videos. In that, 330 video sequences are training set and 107 video sequences are testing set. This dataset includes 130 abnormal events.

3.2 Pre-processing

The video frames in the dataset are given as input to the pre-processing phase to enhance the quality of image frames. The pre-processing techniques used in this research are Gaussian Blur and Min-Max normalization. The detailed explanation of pre-processing techniques is explained below.

- Gaussian Blur - This filter technique is used to minimize the noise and smoothen the image. This process supports minimising noise and fewer variations in video frames that enhance the robustness of the model. Smoothens the edges and minimizes the high-frequency noise and is useful for detecting and tracking objects in video surveillance.
- Min-Max Normalization - This process adjusts the pixel values to a certain range [0,1] which standardize the input for object detection and tracking. This process ensures the pixels are in uniform scale which fastens the training process and causes good performance. This helps the model to converge faster during the training and enhances the stability of the learning process.

3.3 Object detection and tracking

The pre-processed image is given as input to the detection phase to detect and track the objects for anomaly detection. In this research, the YOLO v7 method with Distance Intersection of Union (D-IoU) loss function is used which effectively detects and tracks the objects in pre-processed images. The YOLO v7 is majorly enhanced and provide correct detection and tracking performance without maximizing the inference and execution costs. It provides a faster and more robust network architecture, providing an effective technique for combining features, improved the performance of object recognition, stabilizing loss function and optimizing label handling and training efficiency. Compared to DL algorithms, the YOLO v7 utilizes less computation hardware. It is trained much fastly without utilization of pre-trained weights. The YOLO v7 architecture has many changes like compound scaling, Extended Efficient Layer Aggregation Network (EELAN), bag of freebies along planned and reparametrize convolution, coarseness in auxiliary loss and fineness to lead loss.

3.3.1. EELAN

It is the backbone of YOLO v7 architecture, the method of “expand, shuffle and merge cardinality” is employed. The structure of EELAN enables more efficient learning while preserving the actual gradient path.

3.3.2. Compound Model Scaling for Concatenation Based methods

Method scaling aims to adjust key characteristics of the model to produce versions that are adaptable to different application needs. Various parameters for scaling can’t assigned independently and considered in conventional algorithms by utilizing the architectures of concatenation. For example, maximising the method’s depth impacts the proportion between input of transition layer and output channels, which resulted in less utilization of hardware by the method. The compound scaling algorithm allows method to keep components which have initial phase and continuously take the optimum design.

3.3.3. Planned reparametrize convolution

The planned Reparametrize convolution is the technique to improve the method after training. That extended process of training yields good inference results. The method and module phase combine two types of reparameterization, which are utilized to finalize the methods. This is another type of convolutional block known as RepConv, similar to ResNet, as it includes two identical connections with 1×1 filters. This prevents the similar connection while the convolutional layer along concatenation or residual is utilized for replacing reparametrize convolution. This integrates the D-IOU with penalty term which considered distance between centre of predicted and ground truth boxes.

3.3.4. D-IoU loss function

This loss function integrates the IoU with penalty term which considered distance among centre of predicted and ground truth boxes. This process increases the sensitivity of loss function for spatial arrangement of bounding boxes leads to better accuracy in localization. The mathematical formula for D-IoU loss function is given in Eq. (1).

$$D - IoU\ loss\ = 1 - IoU + \frac{\rho^2(b_p, b_{gt})}{c^2} \quad (1)$$

In the above Eq. (1), the b_p represents the bounding box and the b_{gt} represents the bounding box of ground truth, the $\rho^2(b_p, b_{gt})$ represents Euclidean distance among centre point of predicted bounding box and bounding box of ground truth. The c represents the diagonal length of small enclosing box which covers predicted and ground truth boxes. This process improves the precision of bounding box prediction through considering distance among center of box that is useful for correct localization of object. This supports to achieve good convergence during training through giving much informative loss function

3.4 Feature extraction

The detected image is given as input to feature extraction phase, which extracts the meaningful features to detect the anomalies. In this research, the Inception V3 technique is used to extract the meaningful features. The technique has learned rich representation for high range of images. The Inception v3 network's input is image which has 3 various kinds of inception modules. This module enabled the automatic learning of information without manual process. The module of grid size reduction resolves the issue of feature bottlenecks and execution overhead, attaining object recognition through utilizing the softmax function. The most significant feature of the network is the dividing of large 2D convolutional kernel to 2 little 1D convolutional kernels. This enhances network performance and improves the execution speed when minimizing computation cost. Additionally, network decomposed the symmetric convolutional kernels to asymmetric convolution kernels like dividing the convolutional kernels of 3×3 to 1×3 and 3×1 convolutional kernels. The kernel of the deconvolutional method stores the huge number of parameters and fastens execution when minimizing overfitting. For resolving the issue of feature representations and huge execution, 2 modules of Grid Size reduction are incorporated among every 3 Inception modules for minimizing feature map size through utilizing parallel two-phase architecture. By using the Inception V3 based feature extraction method, it extracted the deep features from low-level to high-level features. These extracted features are given as input to the anomaly detection phase to detect the anomalies.

3.5 Anomaly detection

The statistical distances measure the separation between two statistical objects and when accompanied by the property of symmetry which is called metric. In field of anomaly detection, the Z-score value is employed to compare query observation for extracted patterns from the normal samples. Following evaluation of distance with threshold, the anomaly is detected. The statistical distance, Earth Mover Distance (EMD) called as Wasserstein metric is employed in image domain to compared with 2 probability distributions majorly depended on executing statistical distances among two signatures. The signature has lists of pairs and its mathematical formula is given in Eq. (2).

$$S = \{(x_1, m_1).(x_2, m_2) \ldots (x_n, m_n)\} \quad (2)$$

In the above Eq. (2), x_i represents certain features and the m_n represents their mass. Consider 2 signatures, P and Q that has m and n clusters and their mathematical formula is given in Eqs. (3) and (4).

$$P = \{(p_1, w_{p1}).(p_2, w_{p2}) \ldots (p_m, w_{pm})\} \quad (3)$$

$$Q = \{(q_1, w_{q1}).(q_2, w_{q2}) \ldots (q_m, w_{qm})\} \quad (4)$$

In the above Eqs. (3) and (4), the $p_i(q_i)$ represents representative cluster and the $w_{pi}(w_{qi})$ represents cluster weight. The major aim is to identify the flow of matrix $F = [f_{i,j}]$, in that $f_{i,j}$ represents flow among p and q, the whole cost is reduced with their relevance constraints and its mathematical formula is given in Eqs. (5) to (9).

$$min \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j} \quad (5)$$

$$f_{i,j} \geq 01 \leq i \leq m, 1 \leq j \leq n \quad (6)$$

$$\sum_j f_{i,j} \leq w_{pi}, 1 \leq i \leq m \quad (7)$$

$$\sum_i f_{i,j} \leq w_{qi}, 1 \leq j \leq n \quad (8)$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} = min \{\sum_{j=1}^{m} W_{pi} . \sum_{j=1}^{n} W_{qi}\} \quad (9)$$

This optimization is solved through linear programming. It is depended on solving type of transportation issue. The *F* flow is measured, next, EMD is referred to as process normalized through whole flow and its mathematical formula is given in Eq. (10).

$$EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j}} \quad (10)$$

4. Experimental results

The performance of YOLO v7 – D-IoU and EMD approach is simulated with MATLAB 2020 b environment with metrics of accuracy, f1-score, recall, precision, Area Under the Curve (AUC) describes total 2D area below the Receiver Operating Characteristic (ROC) and Equal Error Rate (EER). The mathematical formula for metrics is described in Eq. (11) to (14),

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$Recall = \frac{TP}{TP+FN} \quad (13)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (14)$$

In the above equations, the TP is True Positive that referred the quantity of correctly classified abnormal frames. The TN is True Negative that referred as quantity of correctly classified normal frames. The FP is False Positive that referred as quantity as quantity of normal frames that are misclassified. The FN is False Negative that referred as quantity of abnormal frames which are misclassified.

In table 1, the performance of the feature extraction method is evaluated with different metrics on UCSD Ped 1 and UCSD Ped 2 datasets. The existing techniques considered to evaluate the performance of Inception V3 based feature extraction approach are DenseNet, EfficientNet, VGG-16 and VGG-19. The Inception V3 based feature extraction approach obtained 96.1% accuracy, 95.5% precision, 95.2% recall and 95.3% f1-score on UCSD Ped 1 dataset. The Inception V3 based feature extraction approach obtained 98.8% accuracy, 98.2% precision, 97.4% recall and 97.8% f1-score on UCSD Ped 2 dataset.

In table 2, the performance of Inception V3 based feature extraction approach is evaluated with metrics of AUC and EER on two datasets. The existing techniques such as DenseNet, EfficientNet, VGG-16 and VGG-19 are considered to evaluate the performance of Inception V3 based feature extraction approach. The Inception V3 based feature extraction

Table 1. Performance of Inception V3 based feature extraction approach

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
UCSD Ped 1				
DenseNet	93.7	93.2	92.9	93.0
EfficientNet	94.3	94.0	93.4	93.8
VGG-16	95.0	94.6	94.1	94.3
VGG-19	95.5	95.1	94.5	94.8
Inception V3	96.1	95.5	95.2	95.3
UCSD Ped 2				
DenseNet	95.2	94.5	94.1	94.2
EfficientNet	95.8	95.0	94.2	94.7
VGG-16	96.1	95.4	94.8	95.1
VGG-19	96.7	96.3	95.9	96.5
Inception V3	98.8	98.2	97.4	97.8

Table 2. Performance of Inception V3 based feature extraction approach

Methods	AUC (%)	EER
UCSD Ped 1		
DenseNet	92.8	17.95
EfficientNet	93.2	17.54
VGG-16	93.7	16.89
VGG-19	94.3	16.21
Inception V3	95.6	15.54
UCSD Ped 2		
DenseNet	94.7	18.87
EfficientNet	95.2	18.24
VGG-16	95.8	17.85
VGG-19	96.5	17.43
Inception V3	97.2	16.33

method obtained 95.6% AUC and 15.54 EER on UCSD Ped 1 dataset. The feature extraction method obtained 97.2% AUC and 16.33 EER on UCSD Ped 2 dataset.

In table 3, the performance of the classification method is evaluated with different metrics on UCSD Ped 1 and UCSD Ped 2 datasets. The existing techniques considered to evaluate the performance of YOLO v7 – D-IoU and EMD approach are Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Region based CNN (RCNN) and YOLO v5. The YOLO v7 – D-IoU and EMD approach obtained 96.1% accuracy, 95.5% precision, 95.2% recall and 95.3% f1-score on UCSD Ped 1 dataset. The Inception V3 based feature extraction approach obtained 98.8% accuracy, 98.2% precision, 97.4% recall and 97.8% f1-score on UCSD Ped 2 dataset.

In table 4, the performance of the classification approach is evaluated with metrics of AUC and EER on two datasets. The existing techniques considered to evaluate the performance of YOLO v7 – D-IoU and EMD approach are RNN, CNN, RCNN and YOLO v5. The YOLO v7 – D-IoU and EMD approach obtained 95.6% AUC and 15.54 EER on UCSD Ped 1 dataset. The YOLO v7 – D-IoU and EMD approach obtained 97.2% AUC and 16.33 EER on UCSD Ped 2 dataset. The below figure 4 and 5 represents the confusion matric for UCSD ped 1 and UCSD ped 2 datasets respectively.

4.1 Comparative analysis

The performance of YOLO v7 – D-IoU and EMD approach is compared with existing methods of IWT-CRF [16], 3DCNN based autoencoder [17], HiS-Deep CNN [19] and Attention based residual autoencoder [18] on both datasets of UCSD Ped 1, UCSD Ped 2, CHUK Avenue and Shanghai Tech datasets with accuracy, AUC and EER metrics.

The YOLO v7 – D-IoU and EMD approach obtained 96.1% accuracy, 96.7% AUC and 15.54 EER on UCSD Ped 1 dataset. The YOLO v7 – D-IoU and Z-score approach obtained 98.8% accuracy, 97.2% AUC and 16.33 EER on UCSD Ped 2 dataset. Table 5 describes the comparative analysis of YOLO v7 – D-IoU and EMD approach.

4.2 Discussion

The results of the YOLO v7 - D-IoU and EMD approach are evaluated with four datasets of UCSD Ped 1, UCSD Ped 2, CHUK Avenue and Shanghai Tech. The existing methods considered to evaluate the performance are DenseNet, VGG-16, VGG-19, Efficient Net, RNN, RCNN, CNN and YOLO v5 with metrics of accuracy, f1-score, precision, recall, AUC and EER. Additionally, the performance of YOLO v7 - D-IoU and EMD method is compared with IWT-CRF [16], 3DCNN based autoencoder [17], HiS-Deep CNN [19] and Attention based residual autoencoder [18] on datasets of UCSD Ped 1, UCSD

Table 3. Performance of classification approach

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
UCSD Ped 1				
RNN	94.1	93.0	92.6	92.8
CNN	94.6	93.8	93.0	93.4
RCNN	95.0	94.3	93.6	93.9
YOLOv5	95.4	94.8	94.1	94.4
YOLO v7 – D-IoU and EMD	96.1	95.5	95.2	95.3
UCSD Ped 2				
RNN	96.4	95.9	95.3	95.7
CNN	97.0	96.5	96.0	96.3
RCNN	97.5	97.0	96.5	96.8
YOLOv5	98.2	97.6	97.2	97.4
YOLO v7 – D-IoU and EMD	98.8	98.2	97.4	97.8

Table 4. Performance of classification approach

Methods	AUC (%)	EER
UCSD Ped 1		
RNN	93.2	18.4
CNN	93.8	17.9
RCNN	94.3	17.2
YOLOv5	95.1	16.5
YOLO v7 – D-IoU and EMD	95.6	15.54
UCSD Ped 2		
RNN	94.9	16.41
CNN	95.6	16.32
RCNN	96.1	16.89
YOLOv5	96.8	17.48
YOLO v7 – D-IoU and EMD	97.2	16.33

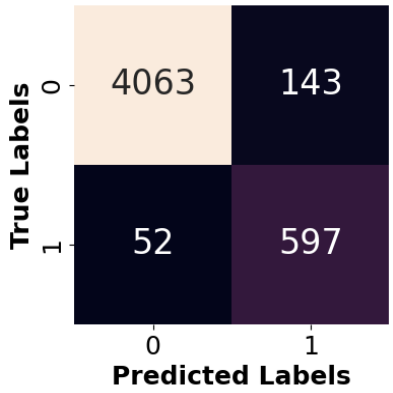

Figure. 4 Confusion matrix for UCSD Ped 1 dataset

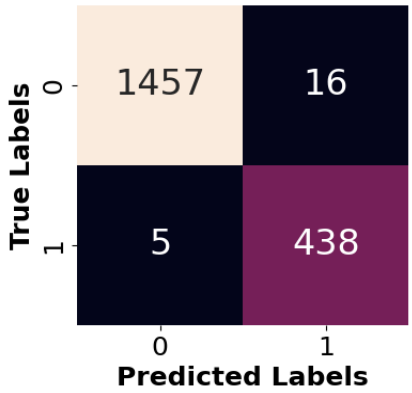

Figure. 5 Confusion matrix for UCSD Ped 2 dataset

Table 5. Comparative Analysis

Datasets	Methods	Accuracy (%)	AUC (%)	EER
UCSD Ped1	IWT-CRF [16]	94.3	NA	NA
	3DCNN based autoencoder [17]	NA	94.6	NA
	HiS-Deep CNN [19]	NA	74.73	17.39
	Proposed YOLO v7 – D-IoU and EMD	96.1	95.6	15.54
UCSD Ped2	3DCNN based autoencoder [17]	NA	96.7	NA
	Attention based residual autoencoder [18]	NA	97.4	NA
	HiS-Deep CNN [19]	NA	79.54	18.21
	Proposed YOLO v7 – D-IoU and EMD	98.8	97.2	16.33
CHUK Avenue dataset	IWT-CRF [16]	93.7	NA	NA
	3DCNN based autoencoder [17]	NA	84.7	NA
	Attention based residual autoencoder [18]	NA	86.7	NA
	HiS-Deep CNN [19]	NA	83.04	16.36
	Proposed YOLO v7 – D-IoU and EMD	95.67	91.23	15.02
Shanghai Tech dataset	3DCNN based autoencoder [17]	NA	74.8	NA
	Attention based residual autoencoder [18]	NA	73.6	NA
	HiS-Deep CNN [19]	NA	83.69	15.66
	Proposed YOLO v7 – D-IoU and EMD	94.78	89.95	14.32

Ped 2 datasets and CHUK Avenue and Shanghai Tech. The YOLO v7 - D-IoU and EMD approach is proposed to detect and track the objects and anomalies in video surveillance. The D-IoU loss function is used in the YOLO v7 model, which improves the precision of bounding box prediction by considering distance between the centre of the box that is useful to correct localization of the object. The, features are extracted using Inception V3 based feature extraction model to extract the meaningful features that help differentiate the anomalies in detected objects. To detect the anomalies, the EMD method is used which effectively detects the anomalies.

5. Conclusion

In this research, the YOLO v7 - D-IoU and EMD approach is proposed to detect and track the objects and anomalies in video surveillance. The D-IoU loss function is used in the YOLO v7 model, which improves the precision of bounding box prediction by considering distance among the centre of the box that is useful for correct localization of the object. The features are extracted using Inception V3 based feature extraction model that captures meaningful features to differentiate the anomalies in detected objects. To detect the anomalies, the EMD method is used which effectively detects the anomalies. The YOLO v7 – D-IoU and Z-score approach obtained 96.1% accuracy on UCSD Ped 1 datasets and 98.8% accuracy on UCSD Ped 2 dataset. The proposed method shows effective performance when compared to conventional methods like 3D-CNN. In future, different DL based approaches can be used to detect the anomalies in video surveillance.

Notations

Notations	Descriptions
b_p	Bounding Box
b_{gt}	Bounding Box of Ground Truth
$\rho^2(b_p, b_{gt})$	Euclidean distance among center point of predicted bounding box and bounding box of ground truth
x_i	Certain features
m_n	Mass
P and Q	Signatures
m and n	Clusters
$p_i(q_i)$	Representative cluster
$w_{pi}(w_{qi})$	Cluster weight
$f_{i,j}$	Flow among p and q

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] F. Abdullah and A. Jalal, "Semantic segmentation based crowd tracking and anomaly detection via neuro-fuzzy classifier in smart surveillance system", *Arabian Journal for Science and Engineering*, Vol. 48, No. 2, pp. 2173-2190, 2023.

[2] Y. Yang, Z. Fu, and S.M. Naqvi, "Abnormal event detection for video surveillance using an enhanced two-stream fusion method", *Neurocomputing*, Vol. 553, p. 126561, 2023.

[3] V.A. Kotkar and V. Sucharita, "Fast anomaly detection in video surveillance system using robust spatiotemporal and deep learning methods", *Multimedia Tools and Applications*, Vol. 82, No. 22, pp. 34259-34286, 2023.

[4] D. K. Sampath, and K. Kumar, "Abnormal Crowd Behaviour Detection in Surveillance Videos Using Spatiotemporal Inter-Fused Autoencoder", *International Journal of Intelligent Engineering & Systems*, Vol. 16, No. 6, pp. 470-481, 2023, doi: 10.22266/ijies2023.1231.39.

[5] A. A. Hamid, S. A. Monadjemi, and B. Shoushtarian, "ABDviaMSIFAT: Abnormal Crowd Behavior Detection utilizing a Multi-Source Information Fusion Technique", *IEEE Access*, 2024.

[6] Y. Liu, J. Liu, K. Yang, B. Ju, S. Liu, Y. Wang, D. Yang, P. Sun, and L. Song, AMP-Net: Appearance-Motion Prototype Network Assisted Automatic Video Anomaly Detection System", *IEEE Transactions on Industrial Informatics*, Vol. 20, No. 2, pp. 2843-2855, 2024.

[7] S. Altowairqi, S. Luo, P. Greer, and S. Chen, "Efficient Crowd Anomaly Detection Using Sparse Feature Tracking and Neural Network", *Applied Sciences*, Vol. 14, No. 9, p. 3928, 2024.

[8] M. M. Ali, "Real-time video anomaly detection for smart surveillance", *IET Image Processing*, Vol. 17, No. 5, pp. 1375-1388, 2023.

[9] T. Alafif, A. Hadi, M. Allahyani, B. Alzahrani, A. Alhothali, R. Alotaibi, and A. Barnawi, "Hybrid classifiers for spatio-temporal abnormal behavior detection, tracking, and recognition in massive Hajj crowds", *Electronics*, Vol. 12, No. 5, p. 1165, 2023.

[10] M. H. Sharif, L. Jiao, and C. W. Omlin, "Deep crowd anomaly detection by fusing reconstruction and prediction networks", *Electronics*, Vol. 12, No. 7, p. 1517, 2023.

[11] I. Fomin, Y. Rezets, and E. Smirnova, "Anomaly Detection on Video by Detecting and Tracking Feature Points", *Engineering Proceedings*, Vol. 33, No. 1, p. 34, 2023.

[12] H. Li, J. Chen, X. Sun, C. Li, and J. Chen, "Multi-memory video anomaly detection based on scene object distribution", *Multimedia Tools*

and Applications, Vol. 82, No. 23, pp. 35557-35583, 2023.

[13] M.R. Bhuiyan, J. Abdullah, N. Hashim, F.A. Farid, and J. Uddin, "Hajj pilgrimage abnormal crowd movement monitoring using optical flow and FCNN", *Journal of Big Data*, Vol. 10, p. 86, 2023.

[14] H. Huang, B. Zhao, F. Gao, P. Chen, J. Wang, and A. Hussain, "A Novel Unsupervised Video Anomaly Detection Framework Based on Optical Flow Reconstruction and Erased Frame Prediction", *Sensors*, Vol. 23, No. 10, p. 4828, 2023.

[15] K. Masood, M.M. Al-Sakhnini, W. Nawaz, T. Faiz, A.S. Mohammad, and H. Kashif, "Identification of anomaly scenes in videos using graph neural networks", *Comput Mater Continua*, Vol. 74, No. 3, pp. 5417-5430, 2022.

[16] F. Abdullah, M. Abdelhaq, R. Alsaqour, M.H. Alatiyyah, K. Alnowaiser, S.S. Alotaibi, and J. Park, "Context aware crowd tracking and anomaly detection via deep learning and social force model", *IEEE Access*, Vol. 11, pp. 75884-75898, 2023.

[17] A. Niaz, S.U. Amin, S. Soomro, H. Zia, and K.N. Choi, "Spatially Aware Fusion in 3D Convolutional Autoencoders for Video Anomaly Detection", *IEEE Access*, 12, pp. 104770-104784,2024.

[18] V.T. Le and Y.G. Kim, "Attention-based residual autoencoder for video anomaly detection", *Applied Intelligence*, Vol. 53, No. 3, pp. 3240-3254, 2023.

[19] B.S. Gayal and S.R. Patil, "Detection and localization of anomalies in video surveillance using novel optimization based deep convolutional neural network", *Multimedia Tools and Applications*, Vol. 82, No. 19, pp. 28895-28915, 2023.

[20] E. Dilek and M. Dener, "Enhancement of Video Anomaly Detection Performance Using Transfer Learning and Fine-Tuning", *IEEE Access*, Vol. 12, pp. 73304-73322, 2024.

[21] UCSD Ped dataset: https://www.kaggle.com/datasets/aryashah2k/ucsd-pedestrian-database