Advanced video anomaly detection using 2D CNN and stacked LSTM with deep active learning-based model

Anoopa S *, Salim A, Nadeera Beevi S
Dept. of Computer Science, College of Engineering Trivandrum,India
Dept. of Computer Applications, TKM College of Engineering,Kollam, India
APJ Abdul Kalam Technological University
**Corresponding author: mailanoopas@gmail.com*

Abstract

Around the world, the video surveillance system has gained wide acceptance and astonishing growth due to its broad applications. The surveillance system has become a paramount tool and benchmark for analyzing the harmony and safety of society. Anomaly detection and its associated applications play a key role in the integrity of the system. The aim of anomaly detection is to find rare and sparse occurrences of events from videos. Developing an accurate and time-efficient system is still remains challenging due to the dynamic nature of a nomalies. The deep learning-based end-to-end system with full use of both spatial and temporal features from the input videos is proposed. The model combines the use of 2DCNN and Stacked LSTM to extract frame-level features through an anisotropic Gunnar Farneback Optical Flow algorithm. The system is evaluated on the benchmarked datasets namely UCSD Ped1 and UCSD Ped2, and it achieves an AUC of 95% and 94% respectively. The experimental results indicate that the proposed method is superior to state-of-the-art algorithms.

Keywords: Active learning; end to end ; 2DCNN; Stacked LSTM; Gunnar Farneback Optical Flow;

1. Introduction

The demand for Intelligent Video Surveillance Systems is expanding due to the rapid growth of urbanization and industrialization. Anomaly Detection is one of the interesting and challenging areas of research in the present era. Any divergence from expected and customary behaviour is referred to as an anomaly in the system. The importance of anomaly detection is the detection and tracking of moving objects, traffic monitoring, loss prevention, monitor operations and outdoor perimeter security. In the present scenario, the occurrence of Crowd Anomalies are frequent in public places such as railway stations, roads, stadiums and other public places. In this context it is very important to improve the anomaly detection methods in overcrowded areas to ensure public safety. Crowd Behaviour Analysis plays an important role in implementing an Automatic Video Surveillance system for detecting violence, crime and attacks in both public and private areas. This analysis helps evaluating how people behave in large groups and to retrieve useful information from crowded videos. Designing a general-purpose model is very difficult because of the subjective and context-dependent nature of anomalies (T.Li *et al.*, 2015). The crowd behaviour analysis pipeline contains different aspects like detection stages, tracking stages, feature extraction stages and crowd behaviour classification stage. The classification stage seeks to recognize specific behaviour and abnormal patterns in video based on the extracted attributes. The extracted features should be added to provide meaningful information on crowd behaviour. Different approaches have been developed for detecting abnormal events in crowded scenes and usually, they are based on similarity-based models(Y. Cong 2013;C. Piciarelli 2008; Kratz 2009). Such models came under Supervised approaches and learned from the changes of motion and appearance.

In recent years, several real-time approaches based on deep learning and classical methods have been adopted (S. D. Bansod 2020; R. Nawaratne 2019). When these two approaches are analysed it is noted that deep learning

networks show a better result than traditional classical methods by combing basic and complicated structures to learn efficient compact data representations. However, deep learning methods require a large quantity of data to work well and are quite expensive to train complicated data. Another issue is that this strategy with only spatial information or only temporal information is not suitable for a variety of crowded scenarios (Bendali-Braham 2021). As a result, creating a general-purpose model for crowd anomaly detection is quite challenging. An end to end adaptive system for crowd anomaly detection and its localization is proposed. Firstly, the input video is split into specific consecutive frames and pre-processing is performed. Background subtraction is performed on pre-processed frames which helps to isolate the foreground of an image for subsequent processing. The optical flow vector helps to obtain the apparent motion of objects in video frames and related depth information. These frame-level features are fed into our model for anomaly detection and localization.

1.1 Contributions

1. An improved end to end model with 2D Convolutional Neural Network with Stacked LSTM is devised.
2. A novel approach for obtaining the apparent motion of an object and depth information using anisotropic diffusion filter based optical flow is proposed.
3. The proposed model is evaluated by two publically available benchmarked datasets and the result shows that our proposed method outperforms the other state-of-art approaches.

The rest of the study is structured as follows: Section 2 describes the detailed study of related works in this domain. The proposed method is detailed in section 3. Section 4 presents experimental results and discussions. The conclusion of the paper with a plan for future scope is drawn in section 5.

2. Related Works

The existing deep learning algorithms used for video anomaly detection are categorized into five categories in terms of their training and learning frameworks. These five categories are Supervised (M. Sabokrou 2018; S. Lin 2019), Unsupervised (D. Xu 2015; M. Sabokrou 2015), Semi-Supervised (S. Akcay 2018: L. Ruff 2020), Training less models (Y. Yuan 2015; A. D. Giorno 2016) and Active learning-based models (T. Pimentel 2016; Y. Liu 2020).

2.1 Supervised Models

In supervised models, video anomaly detection is considered as a binary classifier and it requires both normal and abnormal data for training. Well-defined abnormal activities and a balanced dataset are the two constraints of this model. However, in most cases, it is impossible to clearly define the video due to ambiguous nature, sparse occurrence, evolutionary quality, and data imbalance problems (R. Chalapathy *et al*., 2019). (X.Cui *et al*., 2011) developed a learned Support Vector Machine for the detection and localization of anomalies. (Biswas *et al*., 2016) proposed a unified network for detecting both local and global anomalies from sparse and dense crowds. (Mahadevan *et al*., 2010) developed a learning Mixture of Dynamic texture (MDT) for finding both spatial and temporal abnormalities. These models are based on Spatio-temporal features and appearance-motion descriptors. Recently, different deep learning-based models are developed for the detection of anomalies in crowded scenes. (Sabokrou *et al*., 2018) developed a fully connected neural network for the fastest anomaly detection. (Singh *et al*., 2020) proposed an Aggregation of Ensembles (AOE) method which employs an ensemble of fine-tuned CNN based on the idea that different CNN architectures have different semantic representations of the crowd. This technique uses transfer learning concepts to eliminate the need for training from scratch. (M. Ravanbakhsh *et al*., 2017) developed Generative Adversarial Network (GAN)to learn an internal representation of the normality of crowd behaviour. Later (Wang *et al*., 2018) introduced a shallow generative neural network with a more accurate and powerful learning capacity. This method helps to reduce the loss by using the feature between the encoder and decoder. (Sabin *et al*., 2021) introduced a novel approach to reduce the loss and improve the accuracy of the predicted class by adding Long Short Term Memory (LSTM)with optical features. Supervised approaches give

better results if the class labels and class boundaries are well known. The major challenges in the supervised approaches are data scarcity and class distribution imbalance problems.

2.2 Unsupervised Models

Unsupervised anomaly detection methods cannot be directly applied to the system as there is no clear idea about the values of output. It is required to discover the most relevant features used for discriminating different samples.(D.Xu *et al*., 2015) proposed a deep learning-based double fusion scheme by integrating both appearance and motion. (N.Li *et al*., 2021) proposed an unsupervised statistical framework based on Spatio-temporal configuration and multiple scale analysis. All the methods under this category fail to detect complex anomalous behaviour if the normal and abnormal frame distributions are not clearly differentiated.

2.3 Semi-Supervised Models

Methods in this category use a small amount of labelled data during training. Most of the methods under this category are modelled with autoencoder, which produces high reconstruction costs for abnormal activities and low error for normal activities. (M.Hassan *et al*., 2016) proposed an autoencoder to identify irregularity in videos which helps to detect past and future regular motion from a single frame.(W.Liu *et al*.,2018) developed a new method for future frame prediction with temporal and spatial constraints. (E.Hatimaz *et al*.,2016) introduced a semantic search interface with optical flow features that are used to detect abnormal crowd behaviour. Sometimes these approaches cause data scarcity issues due to the imbalance between labelled and unlabelled data.

2.4 Training-less Models

Training-less models are based on the external domain knowledge strategy (A. A. Sodemann *et al*., 2012). No training is needed with labelled and unlabelled data. (Zhao *et al*.,2011) implemented an online detection method for detecting abnormal events in video with dynamic sparse coding. The basic idea of this approach is to build a sparse coded dictionary for representing knowledge. Yuan et al. proposed an online anomaly detection method by constructing a Structural Context Descriptor that helps to represent the interaction between individuals (Y. Yuan *et al*., 2015). Models under this category are free from bias and overfitting, but unable to adapt to the dynamic nature of anomalies.

2.5 Active Learning-based Models

In the supervised, unsupervised and semi-supervised anomaly detection approaches, the model is only trained with an offline normal dataset and there is no updation with new data. So these methods are not suitable for effective real-time-time applications (J. Varadarajan *et al*., 2017). These issues are solved by using the Active learning approach in deep neural networks. (Nawaratne *et al*., 2019) developed a Spatio-temporal learning network with fuzzy aggregation to get better real-time performance. Recently (Y.Liu *et al*., 2020) proposed GAN (Generative Adversarial Network) to directly generate outliers by solving the issue curse of dimensionality .

The above analysis shows that the performance of the fully supervised and semi-supervised methods are constrained by data scarcity and class distribution imbalance problems, whereas the unsupervised models fail to clearly differentiate normal and abnormal samples. On the other hand training less models are free from bias, but not able to handle the sparse occurrence of scenes. These flaws are overcome by introducing a deep active learning-based end-to-end system for accurate anomaly detection and localization.

3. Proposed Methodology

The proposed method makes use of different deep learning networks for accurate anomaly detection and localization. It is motivated by observing flaws in the previous work, particularly about false negatives and lack of experiment design. A novel end-end model with 2D Convolutional Neural Network and Stacked LSTM is used

for the accurate detection and localization of crowd anomaly by capturing both spatial and temporal features. The work flow of the system is depicted in Figure 1.

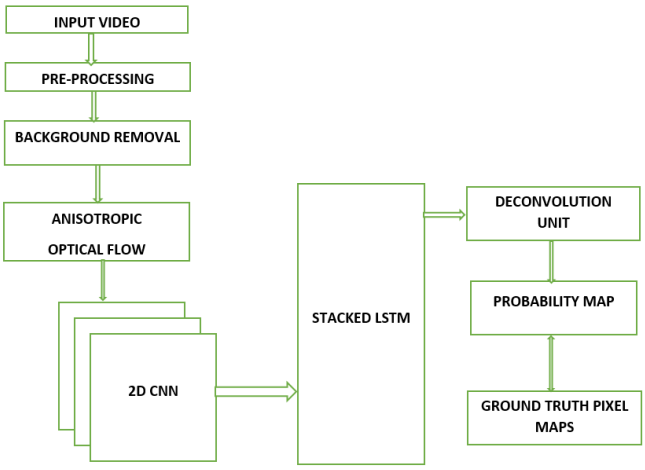


Fig. 1. Architecture of the System

3.1 Pre-processing

The raw video is divided into different consecutive frames and each frame is resized into a size of 128×128. Pixels values in the images are resized from 0 to 1 to achieve normalization. The dimension of every frame is converted to grayscale to reduce complexity.

3.2 Background Removal

Background Subtraction is a technique that helps to extract the foreground region of an image for further processing. It is a widely used method to detect moving objects from the sequence of frames. An Adaptive Gaussian Mixture Model (AGMM) helps to remove the background region which deals with illumination changes and dull movements (Zivkovic *et al*., 2006). Every pixel in the edge is represented as a combination of Gaussian distributions. The pixel value at a time t is represented as $\theta(t)$. The Bayesian decision function Y is used to divide pixels into two groups: background(BG) and Foreground(FG)regions.

$$Y = \frac{p(BG)/\theta(t)}{p(FG)/\theta(t)} \tag{1}$$

Initially, we have no information about how foreground and background region appear in a frame.As a result we assume that the probability of both foreground and background pixels are same.

$$p(BG) = p(FG) \tag{2}$$

If all the foreground objects are considered as a uniform distribution, then

$$p(\theta) = T_F \tag{3}$$

and the pixel is considered as background region if

$$p(\theta) > T_B \tag{4}$$

where T_B and T_F are threshold values which are not fixed values. The different threshold value is selected for each pixel and these thresholds are adapted by time based on the spatial variations in illuminations.

3.3 Optical Flow

Optical flow is one of the commonly used methods to obtain the apparent motion of an object from video frames. It helps to obtain depth information in relation to object direction and speed. Lucas Kanade algorithm is a widely adopted method for estimating the movement of interesting patterns from scenes(Xie *et al*., 2019). But this technique could not be applied to the video frames, where the motion is greater than one pixel between two consecutive video frames. This limitation is resolved by using an anisotropic diffusion filter Gunnar Farneback Optical Flow algorithm to minimize the error between two consecutive frames and preserve the direction of objects in the video (Farneback 2001; Gharsallah 2017). It assumes that displacement between consecutive frames is negligibly small or constant. The first step is to apply an anisotropic diffusion filter to preserve the edges and boundaries. The diffusion filter is defined by

$$D(I) = \frac{1}{||u||^2 + 2(u^2)}(I * I^T + u^2 I) \tag{5}$$

where I is the gradient of an image in frame 1 and u is a parameter to control the smoothness. The Optical flow vector v is defined as

$$v = \begin{bmatrix} v_x \\ v_y \end{bmatrix} \tag{6}$$

Let $v_x(x, y)$ and $v_y(x, y)$ be the flow vector of frames which is calculated as follows

$$v_x(x, y) = ax + by + c \tag{7}$$

$$v_y(x, y) = dx + ey + f \tag{8}$$

where x and y represent image coordinates and a, b, c, d, f are the parameters which are computed using Gaussian least square approximation. The spatiotemporal vector v is represented in matrix form as follows

$$v = Sp \tag{9}$$

$$S = \begin{bmatrix} x & y & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x & y & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}$$

$$p^T = \begin{bmatrix} a & b & c & d & e & f & 1 \end{bmatrix} \tag{11}$$

3.4 2D Convolutional Neural Network

The 2D CNN Model is used to identify frame-level features and extract spatial high level features. The 2D convolution is obtained by combing the 2D kernel into the cube generated by stacking consecutive frames together. Hence, the motion information is easily obtained by constructing a feature map in convolutional layers. The proposed 2DCNN architecture contains four convolutional layers and two max-pooling layers. The first and second convolutional layer has 4 filters with a kernel size of 3×3 and the corresponding pooling layer has a kernel size of 2×2. The second and third convolutional layer has 8 filters with kernel size 3×3. The operation performed in the convolutional layer is defined by

$$F(i, j) = (I * K)(i, j) \tag{12}$$

where F is the output feature map, I represent the input matrix and K denotes the 2D kernel filter. All the convolutional layers have a relu activation function and each layer is modelled as time distributed layer to reduce the complexity. ReLu function $f(x)$ is represented as

$$f(x) = max(0, x) \tag{13}$$

Finally, flatten layer is added to convert the output of the final pooling layer into a sequence of vectors. Table 1 shows the details of 2D ConvNet architecture.

Table 1. Structure of 2D ConvNet Architecture

LAYERS	INPUT	KERNAL	OUTPUTS
CONV1	128×128×1	4, 3×3	128×128×4
CONV2	128×128×4	4, 3×3	128×128×4
MAXPOOL2	128×128×4	2×2	64×64×4×
CONV3	64×64×4	8, 3×3	64×64×8
CONV4	64×64×8	8, 3×3	64×64×8
MAXPOOL2	64×64×8	2×2	32×32×8
FLATTEN1	32×32×8	-	8192

3.5 Stacked LSTM

Long Short Term Memory networks are a type of recurrent neural network which helps to learn long term dependencies and shorten the time between frame processing. LSTM model helps to capture frame-level temporal features and successfully solve vanishing gradient problems (Hochreiter *et al*., 1997). In the proposed work, we are using Stacked LSTM with Adam optimization to reduce overfitting and to improve the accuracy of classification. The Stacked LSTM is an expansion of LSTM that has multiple hidden layers with numerous cells in each layer. Each layer in the Stacked LSTM outputs a sequence of vectors that will be utilized as an input to the next LSTM layers. This structure of the hidden layer helps to capture the information from data at different scales. The general architecture of stacked LSTM is described in Figure 2. The output values of hidden layers are updated at each step as shown below.

$$h_t = o_t tanh(c_t) \tag{14}$$

$$o_t = \sigma(w_{x0}x_t + w_{h0}h_{t-1} + b_0) \tag{15}$$

$$c_t = f_t c_{t-1} + i_t tanh(w_{xc}x_t + w_{hc}h_{t-1} + b_1) \tag{16}$$

$$f_t = \sigma(w_{xf}x_t + w_{hf}h_{t-1} + b_2) \tag{17}$$

$$i_t = \sigma(w_{xi}x_t + w_{hi}h_{t-1} + b_3) \tag{18}$$

where x_t is the input vector, σ is the sigmoid activation function,W and b represents weight vector and bias,h_t and h_{t-1} are the current and previous state respectively. In order to reduce the overfitting problem, initialize the cell state value $c_0 = 0$ and $N(\mu, \sigma)$ using normal distribution (Sabih *et al*., 2021). The LSTM generates the probability map which helps to find anomalies in the frame.

3.6 Training

The entire network is trained from end to end using active learning approaches. The weight vectors of the Stacked LSTM networks are tuned with the backpropagation algorithm using gradient descent optimization. The 2D CNN network weights are updated using simple gradient optimization. This procedure ensures that CNN retrieved features are directly relevant to the Stacked LSTM sequence classification. The Active Learning steps used in the system is described below.

1. Train the model on the labelled data. (Supervised learning).
2. Evaluate the model with unlabelled data.

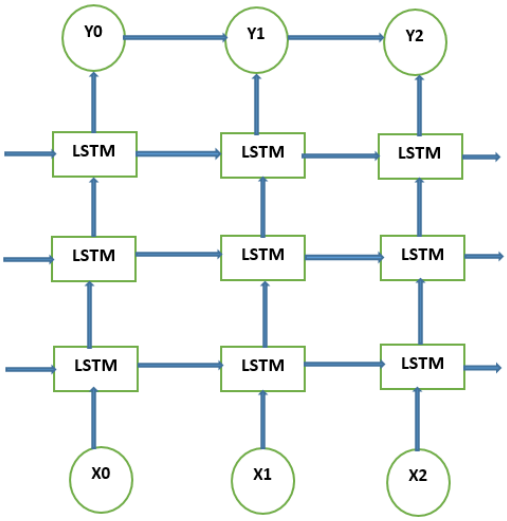


Fig. 2. Architecture of Stacked LSTM

3. Based on this evaluation, choose the valuable samples to be labelled.
4. These valuable data samples are labelled and added to the labelled list.
5. Retrain the model using a new dataset.
6. Repeat the above steps until the performance of the model is stable.

The Adaptive Moment Estimation (Adam) algorithm with a learning rate of 0.001 and momentum of 0.99 are used for loss optimization. Adam is a combined RMSprop and Stochastic Gradient Momentum algorithm with an adaptive learning rate. The learning rate of each network weight is adjusted with the help of both the first and second moments of the gradient. Initially, both the first and second-moment values are set to zero. Moving average parameters are used to update the moment values, and weight updating is done based on these values. The procedure is continued until the weight is converged. The steps for weight updation are described in algorithm 1.

Algorithm 1 Algorithm for Weight Updation

- Step 1: Set the $first_{moment}$ and $second_{moment}$ values to zero.
- Step 2: Repeat steps 3 to 6 until the weight converges.
- Step 3: Compute $dx = computegradient(x)$.
- Step 4: Compute $first_{moment} = beta_1 first_{moment} + (1 - beta_1)dx$.
- Step 5: Compute $second_{moment} = beta_2 second_{moment} + (1 - beta_2)dx$.
- Step 6: $W_{new} = W_{old} - \frac{\rho(first_{moment})}{(second_{moment}}$)

where $beta_1$ and $beta_2$ are moving average parameters and their values are 0.9 and 0.999 respectively. The loss function L is calculated using the binary cross entropy loss method and is given by equation (13).

$$L = -y_i log(y_i) + (1 - y_i) log(1 - (y_i)) \tag{19}$$

where (y_i) is the predicted value and y_i is the probable class value of x_i. The loss function is regularized with L2 regularizer to avoid overfitting. The new loss function is given by the equation

$$Loss_n ew = L + \gamma ||w||^2 \tag{20}$$

Where γ is the hyper parameter that is used to tune the regularization.

4. Results and Discussion

The performance of the proposed method is evaluated by conducting an experiment on publically available datasets UCSD Ped 1 and UCSD Ped 2. A comprehensive analysis of both quantitative and qualitative indexes are performed to compare the performance of the proposed system with different state-of-the-art methods. The system is implemented using Keras framework and experiments are conducted on Intel Xeon W-2145(8C 3.70GHz,11Mb), 32GB DDR4-2666rg with NVIDIA Quadro V100 GPU.

4.1 Evaluation Criteria

The performance of an anomaly detection model can be measured using three different criteria such as frame level, pixel-level and dual pixel level anomalies. The frame level criteria means crowd anomaly is measured at the frame level and is used to verify the accuracy of abnormal crowd anomaly events detection. A frame is detected as an anomaly even if at least one pixel in the frame is found as abnormal (M. Sabokrou *et al*., 2017). In the pixel level criteria, the anomaly is measured at pixel level and then the accuracy of anomaly localization is calculated. If a frame is considered as anomalous when at least 40 percentage of true anomalous pixels are overlapped with Ground Truth(GT) labels (C. Lu *et al*., 2013). This criterion helps to evaluate the accuracy of both anomaly detection and localization. The frame level criteria is used for evaluating the performance of the proposed model.

4.2 Evaluation Metric

The following three quantitative metrics are used for comparing the performance of proposed method with other state-of-the-art algorithms.

1. Area under the Curve (AUC): The area under the ROC curve helps to quantify the accuracy of the crowd anomaly detector.
2. Equal Error Rate (EER): It helps to calculate the percentage of misclassified frames when the true positive rate equals to false-negative rate.
3. Error Detection Rate (EDR): It is the ratio of number of anomaly detected to the total number of anomalies. Hence, it is used for pixel-level evaluation.

4.3 Experimental Setup

All raw video frames are divided into different consecutive frames and resized into 128×128. An anisotropic diffusion filter based on Gunnar Farneback optical flow maps for each frame is generated from spatiotemporal maps with a size 128×128×3. The Adam optimizer with a learning rate of 0.0001 is used to change the parameters during the training stage of our model.

4.4 Dataset

The publically available benchmarked datasets UCSD Ped 1 and Ped 2 dataset, generated with the help of a stationary camera mounted at pedestrian walkways are used in the experiment (Li *et al*., 2014). It is divided into 2 subsets Ped1 and Ped 2 and its crowd density ranges from sparse to crowded. The UCSD Ped1 contains 34 training videos and 36 testing videos. The content in the videos is a group of people walking in a public area. Bikes, Skaters, tiny carts and individuals walking on a sidewalk are the common anomalous events. Both training and testing sequences have approximately 200 frames of dimension 158×238. The UCSD Ped 2 contains 16 training videos and 12 testing videos. Each sequence has approximately 120 to 180 frames of dimension 360×240. Table 2 shows the dimensions of the datasets.

Table 2. Dimension of Datasets

DATASET	TRAINING SET	TESTING SET	FRAME DIMENSION
UCSD PED 1	34	36	238×158
UCSD PED 2	16	12	360×240

4.5 Performance Analysis

The performance of the proposed method is compared with other state-of-the-art. The proposed method accurately detects anomalous events like vehicles, skaters, bikers, etc from the crowded scenes. The effective detection of anomalies in both datasets using the proposed model is shown in Figure 3. Figure 4 shows the ROC Curve on UCSD Ped1 and UCSD ped2. The obtained ROC curve is closer to the top left corner, which indicates better performance of our model.

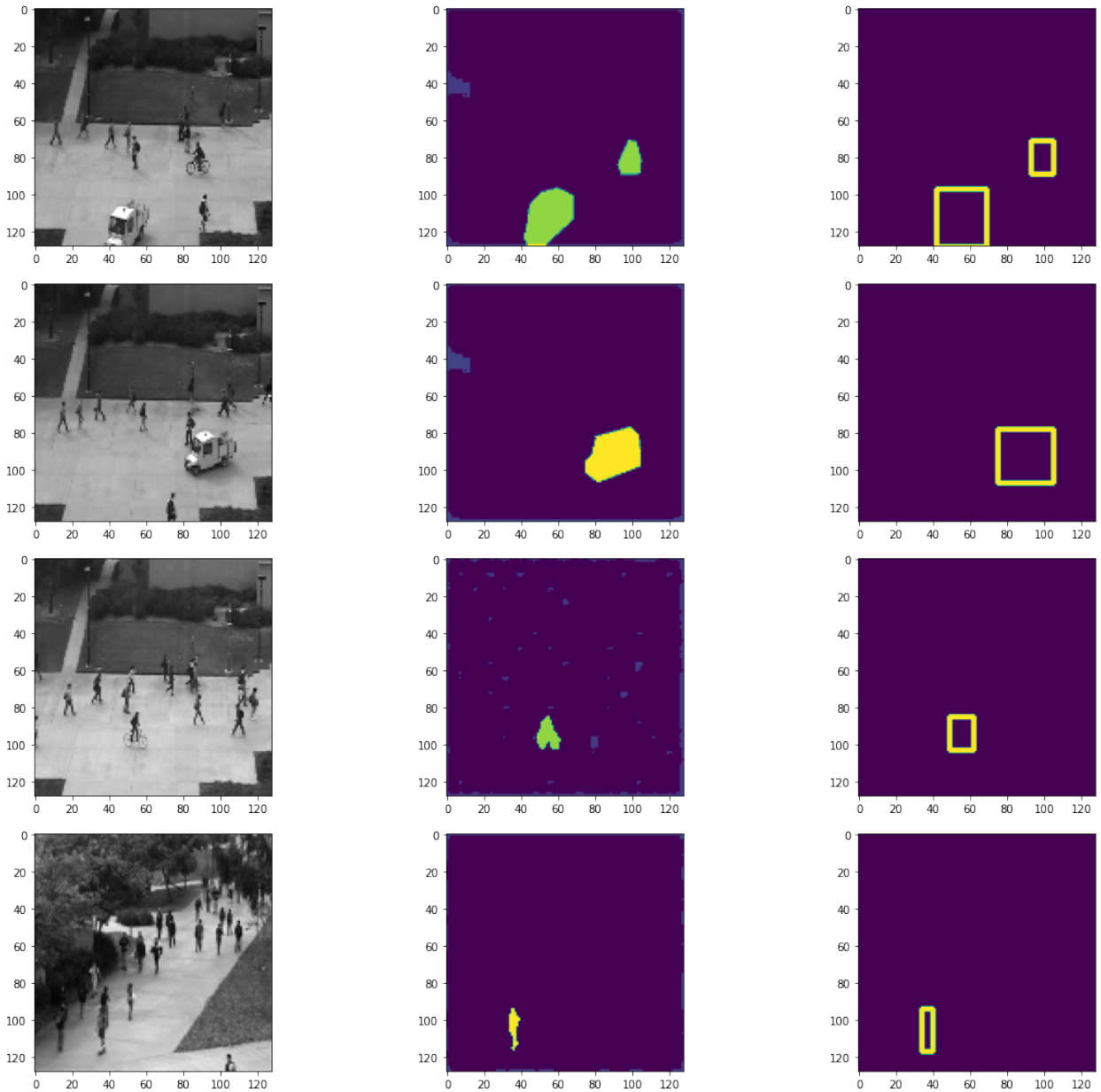
Table 3 demonstrates the comparison of AUC values of the proposed method with other state- of-the-art techniques S2-VAE (Wang *et al*., 2018), MDT (S. D. Bansod *et al*., 2020), Conv-AE (M. Hasan *et al*., 2016), (Nawarante *et al*., 2019),SCG-SF (Chu *et al*., 2019),ST-CaAE (Li *et al*., 2020) and (M.Sabokrou *et al*., 2018). The results show that the proposed method achieves 95 % AUC on ped1 and 94% AUC on the ped2 dataset which is higher than other methods. The higher value of AUC indicates high performance and accuracy of detection for a given test dataset. Table 4 demonstrates the comparison of EER and EDR of the proposed method with other state-of-the-art methods. To calculate the percentage of misclassified frames, an Equal Error Rate metric is used. EER value of our method is 16.7 and 16.8 which indicate the misclassification rate of the proposed system is lesser than other state-of-the-art methods. The Error Detection Rate(EDR) of our method is 83.23 which indicates the anomaly can be clearly detected and localized. Compared with other state-of-the-art methods, our method eliminates false-negative detection by capturing spatio-temporal high-level features. Figure 5 shows the model accuracy of both UCSD Ped1 and Ped2 datasets and Figure 6 shows the comparison of model test loss between basic LSTM and Stacked LSTM on both datasets UCSD Ped1 and UCSD Ped 2. It is clearly understood that in the case of basic LSTM testing loss of the model is higher than training loss. While in the case of Stacked LSTM testing loss is comparatively lesser than training loss and these graphs show that our model is free from overfitting and an optimal model with a validation accuracy of 98.86%.

Table 3. Comparison of AUC score with state-of-the-art methods.

METHOD	UCSD PED 1	UCSD PED 2
MDT (S. D. Bansod *et al*., 2020)	0.818	0.829
Conv-AE (M. Hasan *et al*., 2016)	0.750	0.850
(Nawarante *et al*., 2019)	0.752	0.911
(M.Sabokrou *et al*., 2018)	-	0.925
SCG-SF (Chu *et al*., 2019)	0.909	0.902
S2-VAE (Wang *et al*., 2018)	0.94	-
ST-CaAE (Li *et al*., 2020)	0.90	0.92
PROPOSED METHOD	0.95	0.94

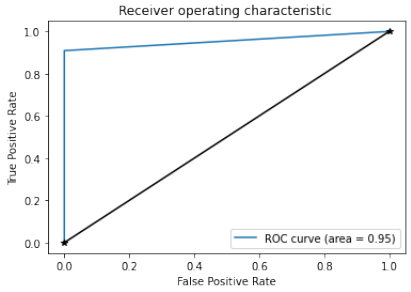
Table 4. Comparison of EER and EDR of the proposed method with other state-of-the-art methods

METHOD	UCSD PED 1		UCSD PED 2	
	EER	EDR	EER	EDR
MDT (S. D. Bansod *et al.*, 2020)	25	45	25	46
Conv-AE (M. Hasan *et al.*, 2016)	27.9	-	21.7	-
(Nawarante *et al.*, 2019)	29.9	-	8.9	-
PROPOSED METHOD	16.7	83.23	16.8	83.24

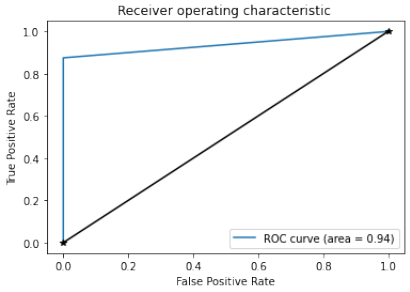


(a) (b) (c)

Fig. 3. Anomaly detection results on UCSD PED Datasets, (a) original frames. (b) frame showing detected anomalies and (c) localization of anomalies.

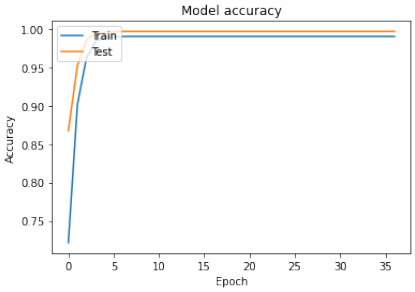


UCSD Ped1: ROC Curve

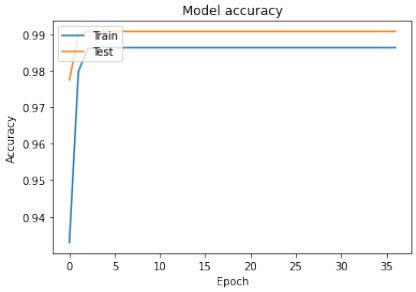


UCSD Ped2: ROC Curve

Fig. 4. ROC curve of UCSD Ped Dataset

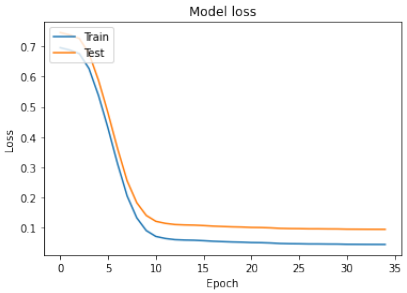


UCSD Ped1: Accuracy

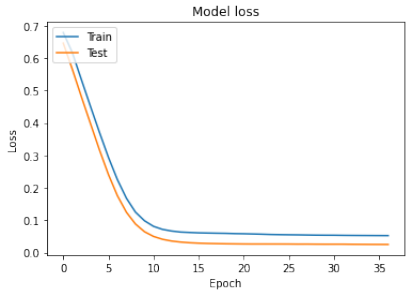


UCSD Ped2: Accuracy

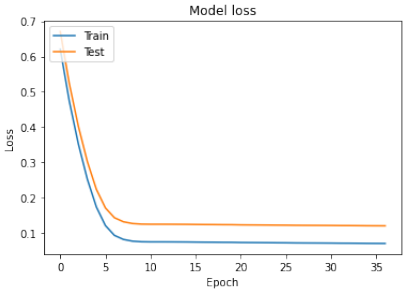
Fig. 5. Model Accuracy of UCSD Ped1 and UCSD Ped2 dataset



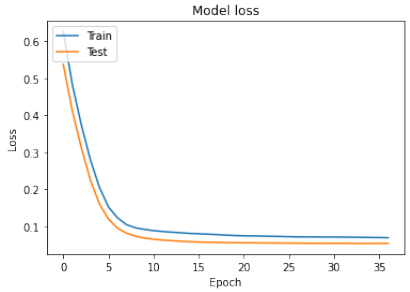
UCSD Ped1: Model Loss



UCSD Ped1: Model Loss



UCSD Ped2: Model Loss
(a)Basic LSTM



UCSD Ped2: Model Loss
(b) Stacked LSTM

Fig. 6. Comparison of model test loss between basic LSTM and Stacked LSTM on both datasets UCSD Ped1 and UCSD Ped 2

5. CONCLUSION

A novel deep active learning-based end-to-end method is developed for detecting and locating abnormal events in surveillance video. The architecture of the proposed method is simple and it effectively captures both spatial and temporal features. The Stacked LSTM combined with 2D CNN helps to reduce the overfitting problem and to improve the accuracy of detection by reducing false prediction. The proposed method produced competitive results of AUC 95% 94% respectively on two publically available benchmarked datasets and validation accuracy of 98.86%, which shows a competitive performance than the existing state-of-the-art methods. In future, the performance of the presented system can be improved with the help of more datasets by finding a solution for both motion position artifacts.

References

A. A. Sodemann, M. P. Ross, B. J. Borghetti(2012)A review of anomaly detection in automated surveillance *IEEE Trans. Syst. Man, Cybern. B, Cybern.* **42**(6) (1257–1272.)

A. D. Giorno, J. A. Bagnell, M. Hebert (2016). A discriminative framework for anomaly detection in large videos *ECCV.* (334–349)

B. Zhao, L. Fei-Fei, E. P. Xing (2011) Online detection of unusual events in videos via dynamic sparse coding *CVPR* (3313–3320.)

Bendali-Braham, Mounir (2021). Recent trends in crowd analysis: A review *Machine Learning with Applications* **4** (100023)

C. Lu, J. Shi, and J. Jia(2013) Abnormal event detection at 150 FPS in matlab *in Proc. IEEE Int. Conf. Comput. Vision* (2720–2727.).

C. Piciarelli, C. Micheloni, G. Foresti (2008). Trajectory-based anomalous event detection *IEEE Trans. Circuits Syst. Video Technol.***18**(11) (1544–1554.)

Chu, W., Xue, H., Yao, C. and Cai, D.(2018) Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos. *Pattern Anal. Mach. Intell.*(246-255)

D. Xu, E. Ricci, Y. Yan, J. Song, N. Sebe(2015). Learning deep representations of appearance and motion for anomalous event detection *BMVC* (8.1–8.12.)

E. Hatirnaz, M. Sah, C. Direkoglu(2020)A novel framework and concept-based semantic search interface for abnormal crowd behaviour analysis in surveillance videos *Multimed. Tools. Appl* (1–39.)

Farneback, Gunnar(2001) Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. Vol. 1. IEEE*

Gharsallah, Mohamed B., and Ezzedine B. Braiek(2017) New anisotropic diffusion method to improve radiographic image quality *Kuwait Journal of Science 44.3*

Hochreiter, Sepp, and Jürgen Schmidhuber(1997) Long short-term memory *PNeural computation 9.8* (1735-1780).

J. Varadarajan, R. Subramanian, N. Ahuja, P. Moulin, J.-M. Odobez(2017) Active online anomaly detection using dirichlet process mixture model and gaussian process classification *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (615–623.)

Kratz, K. Nishino, (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models *CVPR* (1446–1453.)

L. Ruff, R. A. Vandermeulen, N. G¨ornitz, A. Binder, E. M¨uller, K.-R. M¨uller, M. Kloft (2020). Deep semisupervised anomaly detection *ICLR*

Li, N., Chang, F. and Liu, C.(2020) Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes *IEEE Transactions on Multimedia, 23*(203-215)

Li, W., Mahadevan, V., Vasconcelos, N(2014) Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell. (2014). https://doi.org/10.1109/TPAMI.2013.111*

M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, L. S. Davis(2016) Learning temporal regularity in video sequences 155*CVPR* (733–742.)

M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, N. Sebe (2017) Abnormal event detection in videos using generative adversarial nets *ICIP* (1577–1581)

M. Sabokrou, M. Fathy, M. Hoseini, R. Klette (2015). Real-time anomaly detection and localization in crowded scenes *CVPR* (56-62.)

M. Sabokrou, M. Fayyaz, M. Fathy, R. Klette(2017) Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes *IEEE Trans. on Image Processing 26 (4)* (1992–2004.).

M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, R. Klette, (2018) . Fully convolutional neural network for fast anomaly detection in crowded scenes, *Comput. Vis. Image Und* **172** (88-87.)

N. Li, X. Wu, D. Xu, H. Guo, W. Feng(2021) Spatio-temporal context analysis within video volumes for anomalous-event detection and localization 155*The Visual Computer* (309–319.)

R. Chalapathy, S. Chawla (2019) Deep learning for anomaly detection: A survey *arXiv preprint arXiv* (1901.03407)

R. Nawaratne, D. Alahakoon, D. De Silva, X. Yu (2019). Spatiotemporal anomaly detection using deep learning for real-time video surveillance *IEEE Trans. Ind. Informal* **16**(1) (393–402.)

S. Akcay, A. Atapour-Abarghouei, T. P. Breckon (2018). emi-supervised anomaly detection via adversarial training *ACCV* (622–637.)

S. Biswas, V. Gupta (2016) Abnormality detection in crowd videos by tracking sparse components *Machine Vis. Apps.* **28**(1-2)(35–48)

S. D. Bansod, A. V. Nandedkar, (2020). Crowd anomaly detection and localization using histogram of magnitude and momentum *The Visual Comput.* **36**(3) (609–620.)

S. Lin, H. Yang, X. Tang, T. Shi, L. Chen (2019). Social mil: Interaction-aware for crowd anomaly detection *AVSS, IEEE* (1-8.)

Sabih, Mohammad, and Dinesh Kumar Vishwakarma(2021) Crowd anomaly detection with LSTMs using optical features and domain knowledge for improved inferring *The Visual Computer* (1-12.)

Singh, Kuldeep (2020) Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets *Neurocomputing 371* (188-198)

T. Li, H. Chang, M. Wang, B. Ni, R. Hong, S. Yan, (2015). Crowded scene analysis: A survey. *IEEE Trans. Circuits Syst. Video Technol.* **25**(3) (pp. 367–386.)

T. Pimentel, M. Monteiro, A. Veloso, N. Ziviani (2016). Deep active learning for anomaly detection *arXiv preprint arXiv.* (1805.09411)

V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos (2010) Anomaly detection in crowded scenes *CVPR* (1975–1981)

W. Liu, W. Luo, D. Lian, S. Gao(2018) Future frame prediction for anomaly detection–a new baseline *CVPR* (6536–6545.)

Wang, T., Qiao, M., Lin, Z., Li, C., Snoussi, H., Liu, Z. and Choi, C(2018) Generative neural networks for anomaly detection in crowded scenes *IEEE Transactions on Information Forensics and Security 14.5* (1390-1399)

X. Cui, Q. Liu, M. Gao, D. N. Metaxas(2011) Abnormal detection using interaction energy potentials *CVPR* (3161–3167)

Xie, Shaoci, Xiaohong Zhang, and Jing Cai.(2019) Video crowd detection and abnormal behavior model detection based on machine learning method *Neural Computing and Applications 31.1* (175-184.)

Y. Cong, J. Yuan, J. Liu, (2013). Abnormal event detection in crowded scenes using sparse representation *Pattern Recognition* **46**(7) (1851–1864)

Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang, X. He (2020) Generative adversarial active learning for unsupervised outlier detection *IEEE Transactions on Knowledge and Data Engineering* **32**(8)(1517–1528)

Y. Yuan, J. Fang, Q. Wang (2015). Online anomaly detection in crowd scenes via structure analysis *EEE Trans. Cybern.* (548–561)

Zivkovic, Zoran, and Ferdinand Van Der Heijden(2006) Efficient adaptive density estimation per image pixel for the task of background subtraction *Pattern recognition letters 27.7* (773-780.)

Submitted: 08/03/2022
Revised: 28/05/2022
Accepted: 05/06/2022
DOI: 10.48129/kjs.splml.19159