Computers, Materials & Continua
DOI:10.32604/cmc.2022.017522
Article

Tech Science Press

Automatic Unusual Activities Recognition Using Deep Learning in Academia

Muhammad Ramzan[1,2,*], Adnan Abid[1] and Shahid Mahmood Awan[1,3]

[1]School of Systems and Technology, University of Management and Technology, Lahore, 54782, Pakistan.
[2]Department of Computer Science and Information Technology, University of Sargodha, Sargodha, 40100, Pakistan
[3]School of Electronics, Computing and Mathematics, University of Derby, Derby, United Kingdom
[*]Corresponding Author: Muhammad Ramzan. Email: F2017288014@umt.edu.pk
Received: 02 February 2021; Accepted: 25 April 2021

Abstract: In the current era, automatic surveillance has become an active research problem due to its vast real-world applications, particularly for maintaining law and order. A continuous manual monitoring of human activities is a tedious task. The use of cameras and automatic detection of unusual surveillance activity has been growing exponentially over the last few years. Various computer vision techniques have been applied for observation and surveillance of real-world activities. This research study focuses on detecting and recognizing unusual activities in an academic situation such as examination halls, which may help the invigilators observe and restrict the students from cheating or using unfair means. To the best of our knowledge, this is the first research work in this area that develops a dataset for unusual activities in the examination and proposes a deep learning model to detect those unusual activities. The proposed model has been named Automatic Unusual Activity Recognition (AUAR), which employs motion-based frame extraction approaches to extract key-frames and then applies advanced deep learning Convolutional Neural Network algorithm with diverse configurations. The evaluation using standard performance measures confirm that the AUAR model outperforms the already proposed approaches for unusual activity recognition. Apart from evaluating the proposed model on the examination dataset, we also apply AUAR on Violent and Movies datasets, widely used in the relevant literature to detect suspicious activities. The results reveal that AUAR performs well on various data sets compared to existing state-of-the-art models.

Keywords: Deep learning; unusual activities; examination; CNN; surveillance; human activity recognition

1 Introduction

Traditional surveillance requires manual observation to identify unusual activities, which is tedious and prone to error activity. The use of cameras for surveillance is growing exponentially. Surveillance cameras capture a huge volume of video data. Observation of human behavior and categorizing actions is very subjective in different situations. Based on this reason, observational activity can be classified into a normal or abnormal/unusual activity. Regular activities can be

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

categorized as the usual or daily activities performed by humans, such as hand waving, eating, sitting, standing, walking, etc. Unusual activities are different from normal routine activities and vary in specific situations, known as suspicious activities. A lot of work has been done for the detection of suspicious activities in different situations such as observing an abandoned object, theft, health monitoring of patients at the hospital and home (i.e., Fall) [1], road accidents [2], traffic rules violation [3], driver drowsiness [4,5] etc. Nowadays, terrorist activities [6] are happening in crowded and sensitive places such as religious places like mosques, churches, educational institutes, airports, bus stations, government buildings, and shopping malls. The terrorists target such places, hence detecting any suspicious activities or any orphan suitcase around such places has gained utmost importance in the current era, which can be automatically classified as a suspicious activity.

Human activity detection is an important research area in image processing and video analysis [7]. Human activity recognition from still images or video sequences is a challenging task due to various reasons such as deformation, viewpoint variation, illumination changes, background clutter, partial occlusion, and scale variation. Vision-based activity detection systems generally consist of stages such as video/image preprocessing, key-frame extraction, feature extraction, classification, and activity detection.

In particular, the traditional methods of invigilation during the examination to detect unfair means require manual observation of students. An invigilator cannot monitor all the students and may lose attention over time, allowing pupils to engage in cheating activities [8]. Thus, there is a need for automated and intelligent video-based suspicious activity detection systems that may help analyses, detect and minimize unwanted acts resulting in unfair means. However, less work is done for the automatic detection of suspicious/ unusual activities for invigilation during an academic examination that is limited to a few activities and uses handcrafted features and hard-coded algorithms for detection [9].

Deep learning-based Human activity recognition (HAR) is an active research area that plays a vital role in monitoring people's daily life behavior and recognizing activities in a crowded scene and the critical regions through video surveillance. The significant benefit of deep learning is its ability to perform automatic feature extraction and learning compared to conventional vision-based methods. Deep learning models' strength makes it possible to perform automatic high-level feature extraction, and representation learning thus achieves high performance in many areas. Deep learning based on Convolutional Neural Networks (CNNs) has been widely adopted for the video-based human physical activity recognition task. This research automatically analyses and detects cheating activities during examination through videos using deep learning techniques.

This research presents a deep learning-based model name Automated Unusual Activity Recognition (AUAR) to detect unusual activities, including cheating and malpractice during the examination. The proposed system extracts key-frames based on human motion from a video sequence/stream; deep learning model 2D and 3D CNN used for classification task to detect suspicious activities. Furthermore, we have also created a data set for unusual activity recognition during the examination. Thus, the main research contributions of this paper are as follows:

- The dataset has been created for the examination of unusual activity detection systems. For the data set processing, data labelling has been carried out by expert annotators. The dataset is freely available for research purposes.
- We propose to utilize a motion-based Key-frame extraction method to extract only salient frames from a video sequence.
- We proposed to utilize 2D and 3D CNN architectures to detect suspicious activities.

- The research work evaluation using standard evaluation measures prove that the proposed model AUAR outperforms the existing approaches.

The rest of the paper is organized as follows: Section 2 reviews the related work. Section 3 explains the proposed suspicious activity detection system. Section 4 discusses the empirically-based results, and Section 5 provides the conclusion and future work.

2 Related Work

Unusual human activity detection is an important research area in the field of image processing and video analysis. Tracking and understanding objects' behavior through videos has been a research focus for an extended period due to its essential role in human-computer interaction and surveillance. Various algorithms and approaches have been used to detect suspicious objects in public and crowded places in the last decade. Many researchers have been explored the activity recognition problem in different domain. There are two primary activity recognition approaches discussed so far: vision-based [10] and sensor-based activity recognition.

The advancement of image representation approaches and classification methods in vision-based activity recognition literature follows the research trajectory of local, global, and depth-based activity representation methods. Other approaches that being discussed in the literature for human activity detection can be categorized as video-based [11], fuzzy-based [12], trajectory-based [13], hierarchically based [14], data mining based, and color histogram-based suspicious movement detection and tracking [15]. The unusual activity detection process is typically composed of four steps, scene segmentation, feature extraction, monitoring, and human behaviour detection from the video streams.

The vision-based activity recognition literature follows the research trajectory of local, global and depth-based activity representation approaches. Wang et al. worked on patient condition recognition, elder people caring and human fall detection (in hospitals and at homes) for their assistance using surveillance video based on PCANet. Babiker et al. [15] present a human activity recognition system based on feature extraction analysis methods. The author uses two types of feature extraction approaches, the Harris corner detector and blob analysis features. A multi-layer perceptron feeds forward neural network used as a classifier for human activity recognition on KTH and Weizmann datasets. A.K. [16] using the frame deviation method is used to extract key-frames. For feature classification, a Random Forest algorithm is used. Wiliam et al. [17] proposed a contextual information based automatic suspicious behavior detection system. An inference algorithm use for decision making by combining information about the context and learned system knowledge as behavior is suspicious or not. The proposed approach is tested on the CAVIAR dataset and Z-Block dataset. Roy and Om [18] work on suspicious and violent activity detection using the HOG feature extractor and SVM classifier. The trained SVM classifier classifies activities as violent and non-violent, such as kicking, punching and fighting. Other main approaches that are discussed recently for human activity detection are Fuzzy based [19] Trajectory-based [20], Hierarchical based [21].

There are very few articles in the literature that address detecting suspicious activity during examination through video datasets to facilitate invigilators efficient conduction of exams. The authors in [22] provide a framework to monitor student activities during examination by detecting face region using Haar features, detecting hand contact and hand signalling as cheating activities, and raising an alert. Works on the detection of suspicious activity during academic offline examination. This work is divided into three modules; impersonation checks using a PCA-based

face recognition method, detecting such facial malpractices in which students get involved in a conversation with another, and identifying illegal materials or gadgets.

The recent years have shown significant development in the field of deep learning. Deep learning achieves excellent performance and recognition accuracy in various areas such as pattern recognition, image/object recognition, natural language processing, speech recognition, etc. A potential advantage of deep learning models over vision-based methods is their ability to perform automatic feature extraction and learn by examples using machine learning. The computer vision-based methods involve handcrafted low-level features (e.g., colour, edges, corners, contrast) for classification. In contrast, deep learning correlated to A.I. often abstract high level (e.g., Shapes, contours, depth information) feature from a low level, thus achieving high accuracy for classification tasks. The recent advancement in deep learning makes it possible to recognize an activity through video surveillance. Hassan et al. [23] proposed a smartphone-based HAR approach with inertial sensors. The authors use triaxial accelerometers and gyroscope sensors for efficient feature extraction, and Deep Belief Network (DBN) is used for the classification task to recognize the physical activity of humans. The experiments were performed on ANN, SVM and DBN classifiers and showed 89.06%, 94.12% and 95.85% accuracy. Sabokrou et al. [24] propose a detection and localization of anomaly in crowded scenes in video datasets. Authors use cubic patch-based methods and use a cascade of classifiers. These classifiers are divided into two steps, a deep 3D stack auto-encoder for identifying normal cubic patches and then using a complex deeper 3D CNN. The authors compare the proposed method's performance with other researchers' work on UCSD and UMN benchmark datasets. Limin Wang et al. [25] in this paper, the author has used the Temporal Segment Network for action recognition from videos on limited training data. This approach was tested on the HMDB51 and UCF101 dataset and had obtained 69.4% and 94.2% performance gain. Ramachandran et al. present a framework for unusual human activity detection, tracking and features extraction using CNN. The extracted features are then fed into Multi-class Support Vector Machine (MSVM) for classification and detection of suspicious activities. The experiments were performed on a standard dataset and achieved 95% accuracy. Jalal et al. proposed a method to recognize human interaction in an outdoor environment using a Multi-feature algorithm with CNN. The proposed method is evaluated on the BIT-Interaction dataset and recognize eight complex activities. The experimental results show 84.63% recognition accuracy.

Computer vision and Deep learning base HAR is an active research area that plays an essential role in monitoring people's daily life behavior and recognizing activities in a crowded scene and the critical regions through video surveillance [26]. However, less work is done to detect suspicious activities during an examination that is limited up to a few activities and uses hard-coded algorithms for detection. This domain's previous work only involves computer vision-based, handcrafted features and hard-coded algorithms for detecting each category of unusual activity. There is no machine learning involved in classification and detection [27]. Senthilkumar et al. [28] work was to establish a system for evaluating and identifying suspicious behavior in a classroom setting. The system structure consists of three parts to control the student's actions during the study exam. First, the student's face region is identified; secondly, the student's hand contact detection and thirdly, the student's hand signal. Tab. 1 show the already existing techniques related to Unusual Activity detection.

Table 1: Unusual activity detection techniques

Reference	Domain	Technique	Activities	Dataset	Accuracy/Results
Senthilkumar and Narmatha (2016) [28]	Computer Vision	Viola-Jones Algorithm, the grid for motion and Convex hull	Cheating activities during the exam	Self-Created Dataset	Positive Predictive Value 98.7%, 73%78.9%
Chen Wang et al. (2017) [29]	Computer Vision	spatio-temporal sparse	Anomaly Detection	UCSD	AUC Score 0.7477
Labiba Gillani et al. [30]	Deep Learning	-H2O Aunocoder -Probabilistic neural network	Activity recognition and Anomaly detection	CASAS	90% Accuracy
Malik Ali et al. [31]	Deep Learning	Yolo	Patient Monitoring by Abnormal Human Activity Recognition	Created own dataset	96.8%. Accuracy
Nadeem Iqbal et al. (2019) [32]	Machine Learning	HOG-SVM	Abnormal Activity Recognition	Created own dataset	98.02% Accuracy
Babiker et al. (2017) [33]	Deep learning	Multi-layer perceptron feeds forward N.N.	Running, boxing, bending etc	KTH and Weizmann datasets.	For three scenarios 98.9, 93 and 90%
Devi et al. (2017) [34]	Computer Vision	PCA, Region of Interest (ROI)	Impersonate checking, facial malpractice, cheating material	Created own dataset	
Sabokrou et al. (2017) [35]	Deep learning	3D Convolutional Neural Networks	Crowd anomaly detection and localization	UCSD and UMN dataset	Frame-level, 9.1%, pixel-level, 15.8%, AUC 99.6% and EER 2.5
Hassan et al. (2018) [36]	Deep learning	Deep Belief Network (DBN)	Human physical activities	A standard dataset from UCI	Overall accuracy 90.85 %
Jalal et al. (2019) [37]	Deep Learning	Multi features with CNN	Eight different human interactions like boxing, fight, kick, push and bow etc	BIT-Interaction dataset	Recognition Accuracy is 84.63%

3 Proposed Research Methodology

The proposed method uses deep learning to classify key-frames of a video sequence in normal and unusual activities. Fig. 1 shows a comprehensive framework of the proposed system showing the steps of the proposed research method.

3.1 Datasets

In this research, three datasets have been used for empirical analysis. First is the own created dataset, Examination Unusual Activity (EUA) and two more standard published dataset, Violent

flow ("Crowd Violence\Non-violence Database." https://www.openu.ac.il/home/hassner/data/violentflows/) and Movies (Movie and Hockey datasets. "https://figshare.com/articles/figure/_Movie_and_Hockey_datasets_/1375015").

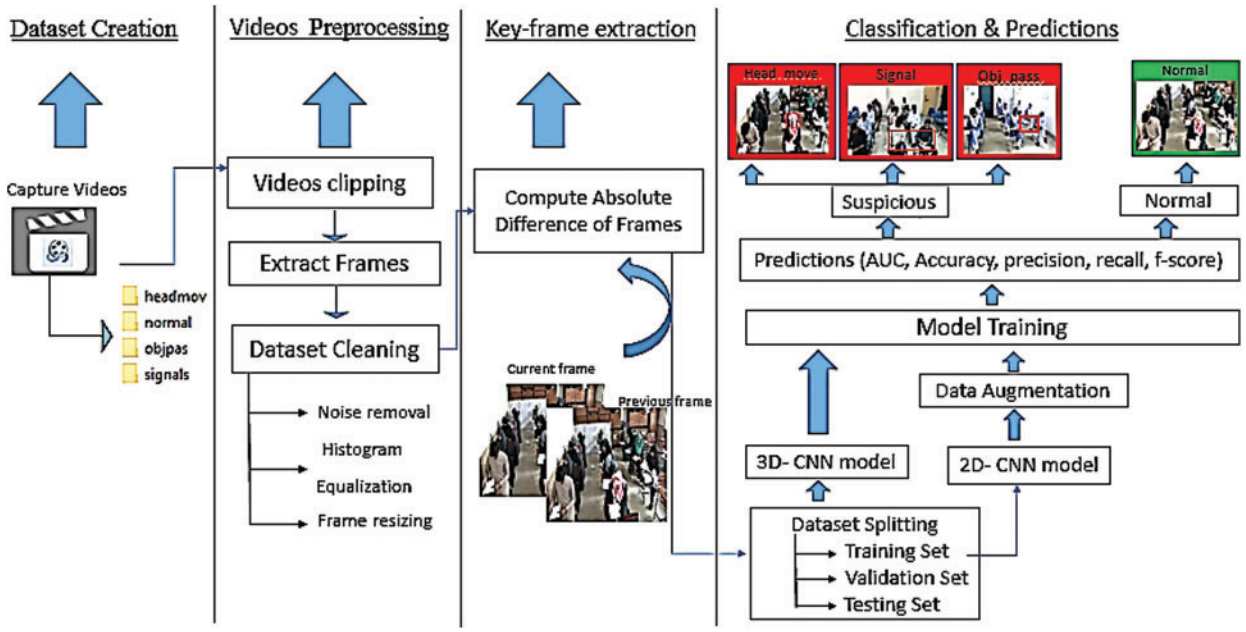

Figure 1: The framework diagram of the proposed research approach

3.1.1 EUA Dataset

There is no standard dataset available in the domain of academic examination invigilation for suspicious activities detection. To address this issue, we have developed our dataset to evaluate the proposed AUAR System. This system is proposed for Unusual Activity recognition in Academic settings. EUA Dataset has been created to classify the activities into normal and abnormal. The suspicious activity of cheating has been detected using three activities: head movement, object passing, and signalling.

For dataset preparation, the videos are captured with the help of university students studying in the Computer Science & Information Technology Department, University of Sargodha, Pakistan. The DSLR camera used for the acquired video with the 20.1 megapixels resolution, the number frames per second, is 29, and the frames' size is 1440 × 1080. All video clips are preprocessed and saved in mp4 format. Each category has 100+ video clips, and the dataset contains a total of 510 videos.

There is vast variability in the dataset as multiple students conduct activities and multiple camera perspectives document these activities. At a fixed point with a static context, events are captured. The next measure for the preparation of the dataset is the validation of this dataset. Some quality metrics define quality, variance, lightness, hue, saturation, and quantity for this dataset. Fig. 2 shows a few glimpses of the frames/image for these three activities of the EUA dataset. The frequency of different categories of activities in the prepared dataset has been presented in Tab. 2. The table shows that the data set comprises of 550 videos, with three

categories of unusual activities and one category of usual activity videos. Whereas Tab. 3 presents the characteristics of the DSLR camera used for recording these videos.

Figure 2: Sample frames from EUA dataset

Table 2: Statistics of the EUA dataset

Activities	Class labels	No. of videos
Head movement	0	160
Object passing	1	120
Signalling	2	125
Normal	3	145

Table 3: Characteristics of DSLR camera

Camera characteristics	
Resolution	1440 × 1080
Frame rate	29fps
Color space	RBG
Video format	MP4

3.1.2 Violent and Movies Dataset

Two other benchmark datasets named as movie dataset and violent-flow dataset have also been used in this research. These datasets have been included in this research because these datasets have been widely used in similar articles addressing the unusual activity recognition

problem.Between two (or a few) people, the Hockey data set was used to evaluate methods for classifying videos as violent or non-violent. The collection includes 1,000 clips divided into five parts, each with 100 violent and non-violent scenes. The dataset of Violent Flows consists of 246 real-life video in which both violent and non-violent scenes are included. The purpose and motivation for including these datasets in this research are to evaluate the proposed model's effectiveness on a variety of datasets to gauge its general applicability.

3.2 Video Preprocessing

After the video capturing process, long-duration videos are converted into short clips of three-second duration each. According to classes of unusual activities, we convert every video into .mp4 format. For video preprocessing, the Gaussian filter is used for noise removal, and histogram equalization is performed on frames of video. After preprocessing, extracted frames are resized into 128 × 128.

3.3 Key Frame Extraction

For the detection of unusual activity from a video sequence, there is a need to extract key-frames that consist of an unusual series of actions. In this dataset, each video consists of a sequence of frames at the rate of 30 fps (http://www.imctv.com/pdf/ipcamera/IP_Surveillance_Design_Guide.pdf.). The frames are very similar to each other, so information in a video sequence is highly redundant, so only a few frames are required that contain meaningful information. These frames are usually called key-frames. Several techniques exist for key-frame extraction, such as colour histogram, histogram difference, frame difference, correlation, entropy difference, etc. In this research work, we apply the motion-based key frame extraction method.

For key-frames extraction, first, downsampling of all the videos is carried out by selecting a skipping factor equal to three for consecutive frames. The skipping factor helps to eliminate redundant frames. We applied a motion-based key frame extraction method [38]. In this method, we take the pixel-wise absolute difference between two consecutive frames.

Threshold value T is calculated by using T= (mean of absolute difference + standard deviation of absolute difference)

$$(absdiff)_f = absdiff\,(Cf_{i+1}, Pf_i) \tag{1}$$

Where Pf_i represents the previous frame and Cf_{i+1} as a current frame in the above equation. Then, we compute the average difference of the matrix obtained in Eq. (1).

$$Avg_{diff} = Avg\left((absdiff)_f\right) \tag{2}$$

If the Avg_{diff} exceeds a pre-defined Threshold (T), then the current frame is selected as a key-frame or skipped otherwise.

$$KF_i = if \begin{cases} Avg_{diff} > T\ \ key\ frame \\ Avg_{diff} < T\ \ Not\ a\ key\ f\text{rame} \end{cases} \tag{3}$$

We update the frames as prev_frame = curr_frame and repeat the whole process. Our key-frame extraction algorithm extracts 11,500 frames for fame level classification, and a sequence of 20 structures out of 550 video clips for video level classification is obtained. Some sample key-frames extracted for four classes of the EUA dataset are shown in Fig. 3.

Figure 3: Key-frames (a) head_move (b) signal (c) object_pass (d) normal

3.4 CNN Architecture

The Convolution Neural Network (CNN) is one of the most widely used architectures among the deep learning architectures for events or activity recognition and automatic feature extraction. This research examines the CNN model on two different architectures: 2D- CNN for Frame level and 3D-CNN for video level detection. The main difference between these two architectures is that the 2D model learns only spatial information using a single frame as input. In contrast, the 3D model learns both space and time information from a video sequence by using the sequence/stack of frames as input.

3.4.1 Features Extraction Using CNN

CNN uses 3 × 3 filters for automatic feature extraction. CNN model trains based on these self-learned features. The output after applying filters is known as a feature map. The first few layers of the network may detect simple features like lines, circles, edges. The network combines these findings in each layer and continually learns more complex concepts as we go deeper and deeper into the network's layers.

3.4.2 2D-CNN Architecture

The proposed 2D model learns only spatial information by using a single frame as input. In the CNN model, convolutional layers perform feature extraction. The number of convolutional layers depends on the complexity of the problem. As we increase the training samples, we need more convolutional layers to capture the feature map by applying kernels of varying sizes. The pooling layer aims to reduce the feature map's size and the number of parameters extracted through convolutional layers. Also, ignoring minor details such as translational, rotational invariance and focusing on the bigger picture (maximum activation). The researchers previously used many techniques to perform pooling operation such as Max pooling, Global average pooling, stochastic pooling [39], etc. The performance analysis shows that Max pooling performs better and extensively used in research than other techniques. A fully connected layer connects every neuron from the Max-pooled layer to every four output neurons. In this research study, the number of convolutional and pooling layers is selected based on training data experiments. We have performed different experiments to compare several different approaches to convolutional and pooling layers and choose the best approach. The proposed 2D-CNN model configuration is as described in Tab. 4.

The 2D-CNN architecture consists of five Convolution layers, the input layer has the shape (128,128,3), and kernel size 3 × 3 with pooling layers of kernel 2 × 2 batch normalization layer is also added after each Convolution layer for normalizing input data, and Relu activation

function is used. All the neurons in the ReLu function do not activate at the same time. After convolutional layers, a flattened layer is added, then add two fully connected dense layers followed by a dropout of 0.5 followed by a softmax output layer consist of four (as equal to a number of classes) neurons. The data is split into three parts after model construction: 70% training, 15% validation, and 15% test set. A training dataset is a set of samples used during the learning process. In comparison, the validation dataset is a set of examples used for parameter selection and independent from training. For the performance evaluation, a Test set is used.

Table 4: Convolutional model configuration

Layers	Conv1	Pool 1	Conv2	Pool 2	Conv3	Pool 3	Conv4	Pool 4	Conv5	Pool 5	FC1	FC2
Kernel	3 × 3	2 × 2	3 × 3	2 × 2	3 × 3	2 × 2	3 × 3	2 × 2	3 × 3	2 × 2	–	–
Stride	1 × 1	1 × 1	1 × 1	1 × 1	1 × 1	1 × 1	1 × 1	1 × 1	1 × 1	1 × 1	–	–
Neurons/ Channel	32		32		64		64		128		512	512

The deep learning models require a large amount of data for training. Suppose there is not enough dataset available for training on the CNN model that affects the model's performance and accuracy. Hence, the solution to this research problem is *Data Augmentation*. The data augmentation is increasing the dataset by using different methods to help deep learning models learn diversity in the dataset, prevent overfitting and produce better results. The data augmentation process includes the following parameters, e.g., horizontal or vertical flip, width shift, rescale and rotation range. Then we generate batches of data for training up to 50 epochs with batch size 20 to fit in RAM and process easily.

3.4.3 3D-CNN Architecture

We moved from 2D-CNN a frame-level classification model to the 3D-CNN classification model for video action recognition. 2D-CNN achieve tremendous success in the image recognition domain. Increased complexity and dimensionality of 3D-CNNs has limited the work on video analysis and recognition [40]. The flow diagram of the proposed 3D-CNN architecture is presented in Fig. 4.

The proposed 3D-CNN model takes an input sequence of 20 frames having (20,128,128,3) dimensions the sequence of 20 frames input to 3D-CNN architecture for training. The network consists of 4 Conv3D and MaxPooling3D layers followed by one fully connected layer with a 0.5% dropout and dense SoftMax output layer. The configuration details of the 3D-CNN model are described in Tab. 4.

In the proposed model, 3D Convolution and pooling layers are used to preserve the space and time information to learn special and temporal features from a video sequence to learn representations better. The 3D-CNN model layers were selected after extensive experiments by increasing/decreasing layers in the model and fine-tune hyperparameters of the model; thus, the configurations are optimized that give the best results in terms of accuracy and loss.

After configuring model layers, the model is compiled with the cost/loss function "categorical_crossentropy" and the optimization function "RMSprop" with a learning rate of 0.001. The

input videos are split into 80% training and 20% test videos dataset. The model is fit on training and validation dataset splits for 30 epochs.

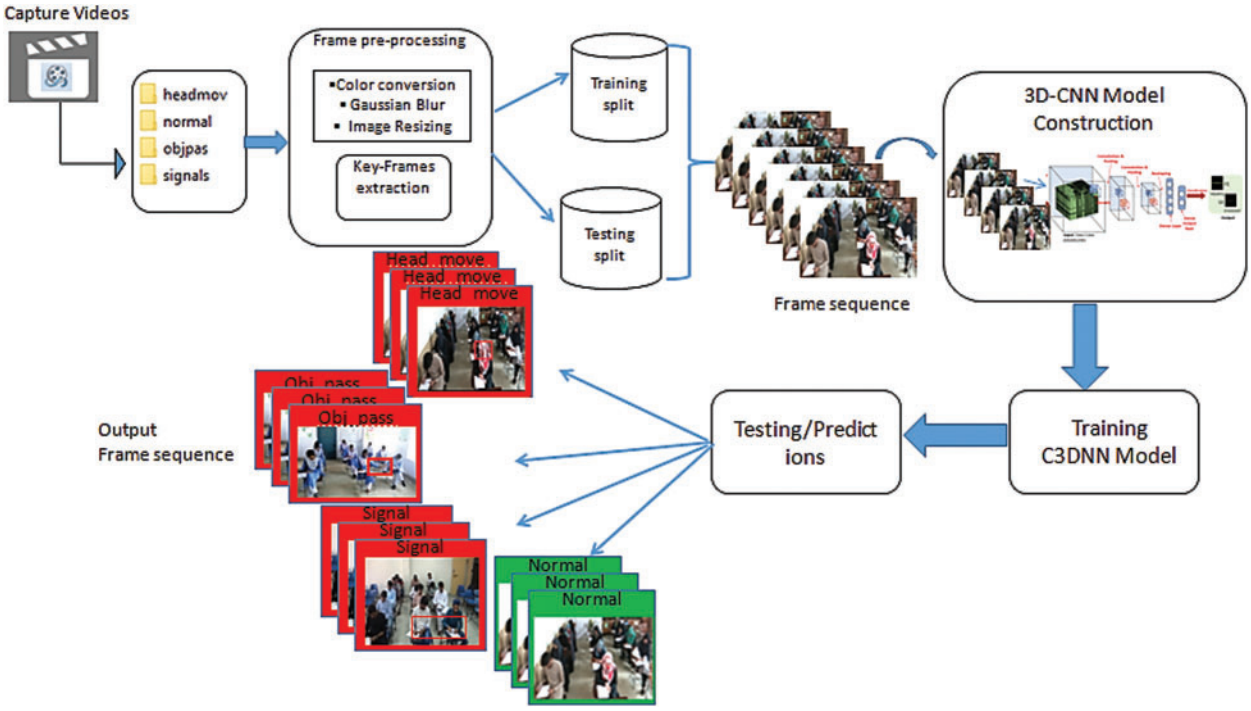

Figure 4: Flow diagram of 3D CNN

4 Experimental Results and Discussion

In this section, we evaluate the performance of our proposed system on the EUA dataset. The standard dataset for AUAR during Examination does not exist. In earlier studies, the authors have created their video dataset (containing only a few videos) to evaluate the proposed method. The experiments are performed on Google Colab a free web-based cloud service that provides Tesla K80 GPU with 12 GB RAM and TPU to train and process deep learning models.

The proposed research work is divided into two implementation domains: the first one covers 2D-CNN, while the second domain implements 3D-CNN. We performed the experiments on two different dataset settings: spatial domain (frames/image level) and space-time (sequence of frames/video) level activity detection. In this research, we evaluated the performance of deep learning approaches based on AUROC (Area Under the Receiver Operating Characteristics)

4.1 Evaluation of CNN Architectures

The 2D-CNN model uses the AUAR dataset consists of 1725 testing frames of 4 classes. While the 3D-CNN model takes 110 test videos having a sequence of 20 frames per video clip. We load the trained model; evaluate the performance of the CNN models on test split, and see how well our model learns to generalize actions. The model takes the test split as input and predicts the class label for each frame according to conventional and non-conventional activities compared to ground truth labels. The experimental results of the 2D-CNN model on the test dataset show

77%, and 3D-CNN models show 73% accuracy. In Fig. 5 2D-CNN shows 0.94, and in Fig. 6 3D-CNN shows 0.91 micro and macro average ROC curve. The ROC shows the probability curve for each action class according to probability scores calculated by the model. AUC for each class, as shown in the figure representing how the model learns to distinguish between each category of unusual activities.

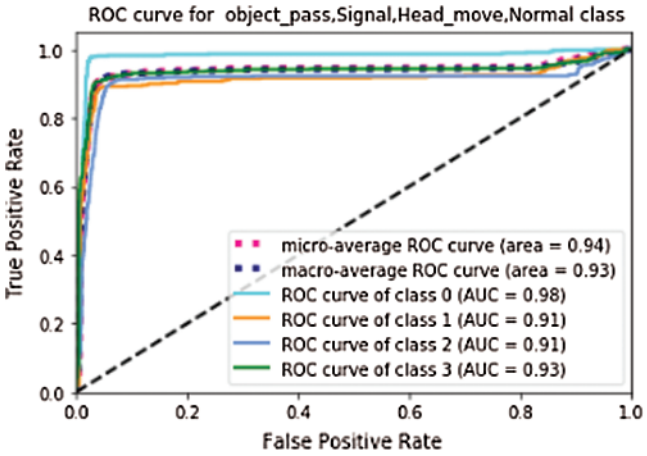

Figure 5: ROC for 2D-CNN

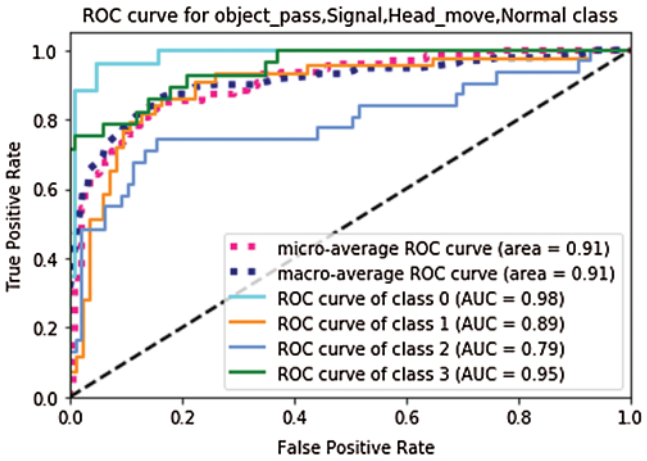

Figure 6: ROC for 3D-CNN

4.2 Comparison of Deep Learning Models

In Tabs. 5 and 6, we summarise the results of deep learning CNN models based on the performance evaluation matrix of AUROC and Classification Report for four EUA dataset classes. Tab. 6 presents a comparative analysis based on the classification report for three unusual activities and one normal class, and the overall accuracy of the proposed system using test datasets.

Table 5: Comparison of CNN models based on AUROC

Evaluation measure	AUROC (Area Under the Receiver Operating Characteristics)	
Classes	Frame level detection results	Video level detection results
	2D-CNN	3D-CNN
Head move	98%	98%
Object passing	91%	89%
Signals	91%	79%
Normal	93%	95%
Micro average ROC	94%	91%

Table 6: Comparison of CNN models based on classification report

Classes	2D-CNN			3D-CNN		
	Precision	Recall	F-score	Precision	Recall	F-score
Head move	85	89	87	75	79	77
Object passing	70	70	70	79	66	72
Signals	69	70	69	77	75	76
Normal	81	75	78	65	75	70
Overall accuracy	77%			73%		

4.3 Comparative Analysis of the Proposed Model with Standard Dataset

The proposed method is evaluated on Movie and Violent flow datasets to analyze the CNN model's performance on these two standard datasets considered benchmarks in the relevant studies. The videos are preprocessed, and key-frames are extracted for normal and unusual behavior and input to CNN architecture. The proposed model performance was evaluated with a standard dataset based on classification accuracy with another state-of-the-art technique. Tab. 7 shows that the proposed method AUAR can better classify unusual behaviours compared to existing techniques.

Table 7: Comparison of classification results on standard datasets

Classifier	Violent flow dataset	Movie dataset
Improved fisher vectors [26]	96.4%	99%
ConvLSTM [41]	94.57%	100%
2D-CNN [42]	–	99%
Substantial derivative [43]	85.43%	96.89%
Proposed method	97.34%	100%

5 Conclusion

This article presents a novel deep learning-based unusual activity detection model in the examination hall. The proposed deep learning model is based on CNN. It outperforms existing models used for unusual activity recognition that uses computer vision and hardcoded algorithms to detect unusual activities during the examination. Apart from proposing the model, we have also developed a video dataset for unusual examination hall activities. The performance of the proposed research work is evaluated on a frame-level consisting of 11500 Key-frames, and the video level consists of 550 video clips of 4 different classes. We have used AUROC as an evaluation matrix. The detection results of deep learning models show excellent performance on our developed dataset. The accuracy of deep learning models for frame-level is higher than the video level due to limited video dataset and GPU resources. The proposed CNN models show an optimized accuracy on our unusual activity dataset regardless of dataset complexity and resource limitations. Apart from this examination dataset, we evaluated the proposed model on two other widely-used datasets, including the Violent-flow dataset and Movie dataset for unusual activity recognition. The proposed model outperformed existing models for all three datasets.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Elouni, H. Ellouzi, H. Ltifi and M. ben Ayed, "Intelligent health monitoring system modeling based on machine learning and agent technology," *Multiagent and Grid Systems*, vol. 16, no. 2, pp. 207–226, 2020.

[2] B. Fernandes, M. Alam, V. Gomes, J. Ferreira and A. Oliveira, "Automatic accident detection with multi-modal alert system implementation for ITS," *Vehicular Communications*, vol. 3, no. 1, pp. 1–11, 2016.

[3] S. Asadianfam, M. Shamsi and A. Rasouli Kenari, "Big data platform of traffic violation detection system: Identifying the risky behaviors of vehicle drivers," *Multimedia Tools and Applications*, vol. 79, no. 33–34, pp. 24645–24684, 2020.

[4] M. Ramzan, H. U. Khan, S. M. Awan, A. Ismail, M. Ilyas *et al.,* "A survey on state-of-the-art drowsiness detection techniques," *IEEE Access*, vol. 7, no. 1, pp. 61904–61919, 2019.

[5] M. Ramzan, S. M. Awan, H. Aldabbas, A. Abid, M. Farhan *et al.,* "Internet of medical things for smart D3S to enable road safety," *International Journal of Distributed Sensor Networks*, vol. 15, no. 8, pp. 1– 10, 2019.

[6] M. I. Uddin, N. Zadda, F. Aziz, Y. Saeed, A. Zeb *et al.,* "Prediction of future terrorist activities using deep neural networks," *Complexity*, vol. 2020, no. 1, pp. 1– 16, 2020.

[7] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail *et al.,* "A review on state-of-the-art violence detection techniques," *IEEE Access*, vol. 7, no. 1, pp. 107560–107575, 2019.

[8] A. Arinaldi and M. I. Fanany, "Cheating video description based on sequences of gestures," in *5th Int. Conf. on Information and Communication Technology, ICoIC7*, Melaka, Malaysia, pp. 1–8, 2017.

[9] A. Sargano, P. Angelov and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," *Applied Sciences*, vol. 7, pp. 110, 2017.

[10] M. Babiker, O. O. Khalifa, K. K. Htike, A. Hassan and M. Zaharadeen, "Harris corner detector and blob analysis featuers in human activty recognetion," in *2017 IEEE Int. Conf. on Smart Instrumentation, Measurement and Applications, ICSIMA*,Putrajaya, Malaysia, pp. 1–5, 2018.

[11] S. R. Ke, H. L. U. Thuc, Y. J. Lee, J. N. Hwang, J. H. Yoo *et al.,* "A review on video-based human activity recognition," *Neurocomputing*, vol. 2, no. 2, pp. 1–23, 2013.

[12] S. Abdelhedi, A. Wali and A. M. Alimi, "Fuzzy logic based human activity recognition in video surveillance applications," *Advances in Intelligent Systems and Computing*, vol. 427, no. 1, pp. 227–235, 2016.

[13] H. A. Abdul-Azim and E. E. Hemayed, "Human action recognition using trajectory-based representation," *Egyptian Informatics Journal*, vol. 16, no. 2, pp. 187–198, 2015.

[14] M. Fazli, K. Kowsari, E. Gharavi, L. Barnes and A. Doryab, "Hierarchical human activity recognition using neural networks," *Neural Network*, vol. 2, no. 7, pp. 1–13, 2020.

[15] S. Kamal, A. Jalal and D. Kim, "Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM," *Journal of Electrical Engineering and Technology*, vol. 11, no. 6, pp. 1857–1862, 2016.

[16] Q. Lei, J.-X. Du, H.-B. Zhang, S. Ye and D.-S. Chen, "A survey of vision-based human action evaluation methods," *Sensors*, vol. 19, no. 19, pp. 4129, 2019.

[17] A. Wiliem, V. Madasu, W. Boles and P. Yarlagadda, "A suspicious behaviour detection using a context space model for smart surveillance systems," *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 194–209, 2012.

[18] P. K. Roy and H. Om, "Suspicious and violent activity detection of humans using HOG features and SVM classifier in surveillance videos," in *Advances in Soft Computing and Machine Learning in Image Processing,* Berlin, Germany: Springer Verlag, vol. 730, no. 1, pp. 277–294, 2018.

[19] S. Abdelhedi, A. Wali and A. M. Alimi, "Fuzzy logic based human activity recognition in video surveillance applications," *Advances in Intelligent Systems and Computing*, vol. 427, no. 1, pp. 227–235, 2016.

[20] B. Boufama, P. Habashi and I. S. Ahmad, "Trajectory-based human activity recognition from videos," in *Int. Conf. on Advanced Technologies for Signal and Image Processing*, Fez, Morocco, pp. 1–5, 2017.

[21] M. Fazli, E. Gharavi k. Kowsari, L. Barnes and A. Doryab, "HHAR-net: Hierarchical human activity recognition using neural networks," in *Intelligent Human Computer Interaction: 12th Int. Conf., IHCI*, Daegu, South Korea, pp. 1–6, 2020.

[22] Z. Li, Z. Zhu and T. Yang, "A multi-index examination cheating detection method based on neural network," in *Proceedings—Int. Conf. on Tools with Artificial Intelligence, ICTAI*, NY, USA, pp. 575–581, 2019.

[23] M. M. Hassan, M. Z. Uddin, A. Mohamed and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Future Generation Computer Systems*, vol. 81, no. 1, pp. 307–313, 2018.

[24] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, vol. 172, no. 1, pp. 88–97, 2018.

[25] Y. Fang, R. Zhang, Q. F. Wang and K. Huang, "Action recognition in videos with temporal segments fusions," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. I, no. 1, pp. 244–253, 2020.

[26] D. R. Beddiar, B. Nini, M. Sabokrou and A. Hadid, "Vision-based human activity recognition: A survey," *Multimedia Tools and Applications*, vol. 79, no. 41–42, pp. 30509–30555, 2020.

[27] D. T. Nguyen, T. D. Pham, N. R. Baek and K. R. Park, "Combining deep and handcrafted image features for presentation attack detection in face recognition systems using visible-light camera sensors," *Sensors*, vol. 18, no. 3, pp. 699, 2018.

[28] T. Senthilkumar and G. Narmatha, "Suspicious human activity detection in classroom examination," in *Computational Intelligence, Cyber Security and Computational Models*. Singapore: Springer, pp. 99–108, 2015.

[29] C. Wang, H. Yao and X. Sun, "Anomaly detection based on spatio-temporal sparse representation and visual attention analysis," *Multimedia Tools and Applications*, vol. 76, no. 5, pp. 6263–6279, 2017.

[30] L. G. Fahad and S. F. Tahir, "Activity recognition and anomaly detection in smart homes," *Neurocomputing*, vol. 423, no. 1, pp. 362–372, 2021.

[31] M. A. Gul, M. H. Yousaf, S. Nawaz, Z. Ur Rehman and H. Kim, "Patient monitoring by abnormal human activity recognition based on CNN architecture," *Electronics*, vol. 9, no. 12, pp. 1993, 2020.

[32] N. Iqbal, M. M. Saad Missen, N. Salamat and V. B. S. Prasath, "On video based human abnormal activity detection with histogram of oriented gradients," in *Handbook of Multimedia Information Security: Techniques and Applications*. Cham: Springer, pp. 431–448, 2019.

[33] M. Babiker, O. O. Khalifa, K. K. Htike, A. Hassan and M. Zaharadeen, "Harris corner detector and blob analysis featuers in human activty recognetion," in *2017 IEEE Int. Conf. on Smart Instrumentation, Measurement and Applications, ICSIMA 2017*, Sakaka, SA, pp. 1–5, 2018.

[34] G. Devi, G. Suvarna and S. Chandini, "Automated video surveillance system for detection of suspicious activities during academic offline examination," *International Journal of Computer and Information Engineering*, vol. 11, no. 12, pp. 1265–1271, 2017.

[35] M. Sabokrou, M. Fayyaz, M. Fathy and R. Klette, "Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992–2004, 2017.

[36] M. M. Hassan, S. Huda, M. Z. Uddin, A. Almogren and M. Alrubaian, "Human activity recognition from body sensor data using deep learning," *Journal of Medical Systems*, vol. 42, no. 6, pp. 1–8, 2018.

[37] A. Jalal, M. Maria and A. S. Hasan, "Multi-features descriptors for human activity tracking and recognition in Indoor-outdoor environments," in *16th Int. Bhurban Conf. on Applied Sciences and Technology*, Islamabad, Pakistan, IEEE, pp. 371–376, 2019.

[38] M. Huang, H. Shu and J. Jiang, "An algorithm of key-frame extraction based on adaptive threshold detection of multi-features," in *Proceedings of the International Symp. on Test and Measurement*, China, Hong Kong, pp. 149–152, 2009.

[39] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," in *1st Int. Con. on Learning Representations*, Scottsdale, United States, pp. 1–5, 2013.

[40] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[41] M. Majd and R. Safabakhsh, "Correlational convolutional LSTM for human action recognition," *Neurocomputing*, vol. 396, no. 1, pp. 224–229, 2020.

[42] V. D. Hoang, D. H. Hoang and C. le Hieu, "Action recognition based on sequential 2D-CNN for surveillance systems," in *Proceedings: IECON, 2018—44th Annual Conf. of the IEEE Industrial Electronics Society*, Washington, DC, USA, pp. 3225–3230, 2018.

[43] S. Mohammadi, H. Kiani, A. Perina and V. Murino, "Violence detection in crowded scenes using substantial derivative," in *2015 12th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Karlsruhe, Germany, pp. 1–6, 2015.