International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 35, No. 3, June, 2021, pp. 255-263
Journal homepage: http://iieta.org/journals/ria

A Framework for Anomaly Classification Using Deep Transfer Learning Approach

Ruchi Jayaswal*, Manish Dixit

Madhav Institute of Technology and Science, M.P., Gwalior 474005, India

Corresponding Author Email: ruchi.jayaswal23@mitsgwalior.in

https://doi.org/10.18280/ria.350309

Received: 2 April 2021
Accepted: 9 June 2021

Keywords:
human behavior dataset21(HBD21), deep learning, Xception model, long short-term memory (LSTM), transfer learning, fine tune

ABSTRACT

Over the last few years, surveillance CCTV cameras have rapidly grown to monitor human activities. Suspicious activities like assault, gun violence, kidnapping need to be observed in public places like malls, public roads, colleges, etc. There is a need for such a surveillance system that automatically recognizes human behavior, such as violent and non-violent actions. Action recognition has become an active research topic for researchers within the computer vision field. However, the human behavior recognition community has mainly focused only on regular actions like walking, running, jogging, etc. Though, detecting behavior in anomaly subjects like assault violence, gun violence, or general aggressive behavior has been comparatively less research in these specific events due to a lack of datasets and algorithms. Thus, there is an increasing demand for datasets to develop abnormal behavior algorithms that can classify anomaly actions. In this paper, the novel dataset is proposed named Human Behavior Dataset 2021 (HBD21). There are four categories of videos available in this dataset: Assault violence, Gun violence, Sabotage violence, and Normal events. This proposed dataset contains a total of 456 videos. Each video has the same length of each category. This paper aims to make a robust surveillance system framework with the help of a deep transfer learning approach and proposed a novel hybrid model. In this view, the current research work is categorized into three phases. Firstly, the preprocessing technique is applied to enhance the brightness of videos, and for resizing then, frames are extracted from each video. Secondly, the transfer learning-based Xception model is used to extract relevant features from frames. The third phase is a classification of behaviors in which a modified LSTM technique is applied. The model is trained using LSTM on the HBD21 dataset. Moreover, using proposed methods on the HBD21 dataset, the accuracy is obtained 97.25% overall.

1. INTRODUCTION

From the last few years, outperformed infrastructure growths are observed in security-related problems worldwide. Still, there is an increased demand for security; video-based surveillance has become a vital topic for researchers for analysis. As with the increase of global problems, sometimes the problems have turned into a pandemic situation. Moreover, due to this condition, people become more aggressive. According to FBI data [1], violence is increased in these pandemic situations. Since there is a need for an intelligent video surveillance system that detects violence in public places in real-time applications, prevention, Diagnosis, and operation that have contributed to the intelligent video analysis competencies can develop real and consistent video surveillance applications. In computer vision, the classification of violent and non-violent actions can be done using surveillance camera videos. The CCTV cameras are fixed at different public safety places such as schools, shopping malls, colleges, public roads, hospitals, banks, markets, streets, etc., to capture human actions. The system analyzes the behavior of humans, whether their actions are normal and abnormal. Activity recognition may be of direct use in real-life applications such as assault, fighting, gun violence, burglary, and many more. Large-scale surveillance systems aim to be deployed in significant safety concern areas such as institutions and prisons for alerting via alarm authorities to prevent dangerous situations.

Earlier, the human behavior surveillance system was more dependent on the traditional method or the human operator. But nowadays, the surveillance system dependent on automated systems because of better efficiencies and reliability, and in terms of security, these automated systems are beneficial to detect activities. Many researchers have applied CNN, 3D CNN, traditional methods, deep learning techniques to make an automated intelligent surveillance system. This work mainly focuses on the challenging task of detecting violent and non-violent actions in videos. This paper introduces a novel dataset, HBD21, and applies a new technique to classify abnormal and normal behaviours automatically. Transfer learning is used in this paper to detect the human behaviors videos. Several researchers design many new algorithms to recognize human behaviors, which will be described in the literature review. Over the last decade, several techniques for unusual activity recognition are proposed.

1.1 Basic concepts: Human behavior detection

The basic block diagram of abnormal or suspicious human behavior recognition is shown in Figure 1. Several steps are followed to develop an intelligent surveillance system. In the abnormal activity recognition procedure, the initial part is to segment a entire video into several frames. The second step is preprocessing, in which background elimination, resizing, brightness adjustment, noise removal steps are processed. The third step is feature extraction by various techniques. Earlier, the feature extraction technique is done by machine learning techniques such as LBP, HOG, Optical flow, and many more. Nevertheless, the advancement in these techniques goes well, just like the Deep learning approach becomes the active method for all researchers because it generated the relevant features. The CNN model, MobileNet [2], VGG19 [3], Inception [4], ResNet [5], Xception [6], etc., are deep learning models through which we can train the models according to application requirements. Lastly, applied classification techniques like SVM, decision tree, CNN model, LSTM [7], RCNN [8], etc., to classify the anomaly actions from the frames. And then generate an alarm if any suspicious is found in the video at run time. Many researchers suggested different behavior detection techniques to improve the system's efficiency, accuracy, and performance. The following section will discuss research work that has been done in violence detection.

This paper is segmented into four main sections. Section II discusses the literature survey related to violence detection and human behavior recognition. Section III presents the proposed research methodology of this work. Section IV discusses the experimental setup and outcomes, and Section V presents the conclusion and future work.

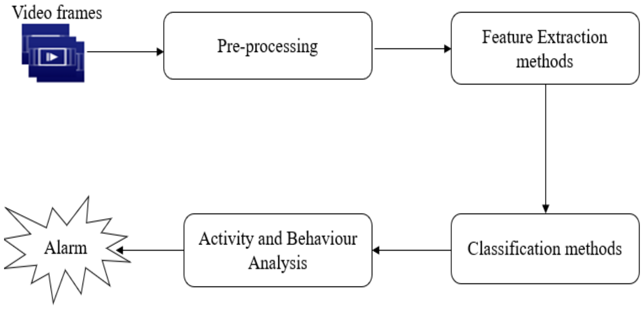


Figure 1. The framework of human behavior recognition

2. LITERATURE SURVEY

Several researchers [9-11] have done their work in the human behavior recognition field. Abnormality was previously described as explosions and blood flow [12, 13], but these cues always missed notice or alarm. Lately, the author [14] suggested a feature set that provides strong multimodal audio and visual signals by first combining the audio-visual features and generating multimodal fusion patterns. However, because the audio channel is not available in real life, audio-based techniques are always limited.

Later researchers mainly focus on the visual features and recognize abnormal actions by computer vision technologies. Many researchers used machine learning techniques for feature extraction such as Improved Dense Trajectories (iDT) [15], Space-Time Interest Points (STIP) [16], Orientation Histogram of Optical Flow (OHOF) [17] used in action recognition.

Recently, features are learned by deep learning models and recognize the abnormal actions successfully. With the use of GPU, parallel processing and collection of a large-scale training set of images are possible. Sabokrou et al. [18] states that using a cascading 3D deep neural network provides a cubic patch technique for detecting and pointing out aberrant activities in complicated video scenes. Many more models like the Deep belief network proposed by Erfani et al. [19] to obtain generic causal features and SVM are used to classify anomaly detection.

Xu et al. [20] suggested a human action recognition system using a machine learning approach. Optical flow is used for movement feature expression then the CNN model is used to extract relevant features. Lastly, SVM is applied to classify and recognize the behaviors of humans. The author used two benchmark databases for experiment purposes, namely Weizmann and KTH. The Weizmann dataset involves ten variant actions: bend, run, jump, pjump, jack, walk, skip, side, wave1, wave2, and the KTH dataset contains six actions: walking, running, jogging, boxing, handwaving, hand-clapping. The overall accuracy on Weizmann and KTH datasets are 93.5% and 89.9%, respectively. To improve the accuracy, the computation time of the system has also increased.

Tay et al. [21] suggested an approach for abnormal behavior detection using the CNN model. RGB frames are picked up manually, and then apply 3*3 moving average filter to remove noise. Further, all frames are assigned labels and trained with CNN layers. Five benchmark databases have been taken for experiment purposes: CMU, UTI, PEL, Hockey Fighting dataset, and WED datasets. Each datasets have distinct background scenarios, such as indoor areas, movie scenes, game fields, etc. The author experimented with two segments. The first experiment involves categorizing the frames into binary classes -unusual and usual actions. The second experiment is to classify the frames into six categories: kicking, pushing, punching as unusual actions, and handshaking, hugging and pointing as usual actions. At a learning rate of 0.001 and 100 epochs, the average accuracy is 98.02% of all datasets.

Wang and Xia [22] presents a new technique for anomaly classification using a deep learning approach. He is utilized two SDAEs (Stack denoising autoencoder) to learn motion and appearance features. Bag of words are used to describe behavior, and the Agglomerative Information Bottleneck approach is utilized to minimize feature dimensions. A sparse representation is used to enhance the abnormal behavior detection accuracy. The author uses two benchmark datasets, namely BEHAVE and BOSS. The Behave dataset includes walking together, splitting, running together, following, meeting as normal behavior, ignoring, chasing, and fighting as abnormal behavior. The BOSS dataset includes three normal videos, and abnormal actions, including grabbing a cell phone, fighting, grabbing the newspaper, harassing, fainting, and panicking, are involved. The AUC on BEHAVE dataset is 0.986 and on the BOSS dataset is 0.982.

The work in anomaly detection was on public datasets like KTH [23], Weizmann [24], UCSD, Behave, Boss datasets. They all contain around six to eleven actions, including fighting, running, pedestrian walkaway, non-fight actions. Some researchers used synthesized datasets for violence recognition. Still, there is a lack of a realistic dataset in behavior recognition.

Various papers based on machine learning and deep learning with their datasets and performance analysis are shown in Table 1. The authors described the violence detection techniques using different methods.

Table 1. Violence detection techniques using different methods

Year	Author	Observation Methodology	Dataset Used	Activities Involved	Performance Analysis
2020	Amrutha et al. [25]	Authors classified the videos into suspicious actions (Students using mobile phones, fighting, fainting) and normal actions (walking, running). VGG-16 model is used to feature extraction. LSTM network is used to train the model.	KTH, CAVIAR, YouTube video and around 300 videos of Suspicious and normal actions are captured from CCTV in college.	Students using mobile phones, Fighting, fainting in class, walking, running	In the initial ten epochs, the accuracy of the training phase is 76%. This accuracy is improved as it increases with the number of epochs. Total accuracy is obtained at 87.15%.
2020	Nayak et al. [26]	This paper presents a brief survey on video anomaly detection. Comparative studies on deep learning methods in anomaly detections are summarized.	Publicly available datasets for videos anomalies are discussed	Discuss abnormal and normal actions	In-Depth analysis in the existing deep learning methods for abnormal detection.
2019	Febin et al. [27]	In this paper, the author proposed a cascaded detection method named MoBSIFT based on motion boundary SIFT, Histogram of optical flow, and movement filtering. Filtered only violent actions and avoid the non-violent frames at the time of feature extraction.	Hockey and Movies datasets are used.	Fight and non-Fight actions	Time complexity is reduced using the proposed method. Time taken to detect violent actions is 0.257 sec/frames. The accuracy on hockey and movies dataset to detect fight action is 90.2% and 91%.
2019	Ramzan et al. [28]	This review paper summarized the violence detection techniques based on traditional machine learning, SVM, and deep learning approaches.	SBU Kinect interaction, Hockey, Movies, Behave, Caviar, KARD, MediaEval, UCF 101 are discussed.	Violent and non-violent discussed	The author presents state-of-the-art research in violence detection techniques.
2019	Ullah et al. [29]	The author proposed a three-staged end-to-end framework. In the first phase, people are detected using the CNN model to avoid unwanted frames; then, in the second stage, filtered frames are fed into the 3D CNN model where spatiotemporal features are fed to the softmax classifier for final prediction are extracted. In the last stage, to optimize the model, the OPENVINO toolkit is used.	Violent Crowd, Violence in Movies dataset, Hockey Fights datasets are used	Fight and non-Fight actions	Accuracies are obtained on three datasets by changing their learning rate and the number of iterations. After comparison, the outcome states that the sliding window performs better compared to the Support Vector Machine.
2019	Dandage et al. [30]	The authors suggested a method for violence detection by Deep Learning. Features are extracted using the CNN model, and then Faster RCNN is used to detect a person in the video; and lastly, LSTM is applied to identify the violence and send the alert if an anomaly is detected.	The model used three different datasets.	Violence and non-violence activities	CNN and LSTM are together to work in work and increases the accuracy at a certain margin.
2018	Khaleghi and Moin [31]	This paper presents a novel method for anomaly detection. In the first stage, preprocessing is done in which it removes background based on the most occurrence of frequency amid video frame patches. At the second stage, features are extracted and detect the anomaly actions. Lastly, trained data is fed to the deep classifier, linear classifier, and autoencoders to evaluate the final decision.	UCSD dataset is used.	Pedestrian walkaway	The equal error rate at frame level by the proposed method is 14% and 25%.
2017	Zhou et al. [32]	The author proposed a ConvNet and FightNet model for violent interaction detection. The image acceleration field is introduced to extract motion features. Each video is framed as RGB images then the optical flow field is calculated by using frames and acceleration field. Lastly, Fightnet is trained with three types of input modalities.	Violent interaction dataset is proposed using four datasets UCF-101, HMDB51 and take each video examples of the Hockey and Movies dataset.	From HMDB51 dataset- hit, kick, punch actions and from UCF101 dataset- Punch, Sumo Wrestling, Boxing Speed Bag, Boxing Punching Bag actions	The result from all modalities on the violent interaction dataset is 97.06%.
2016	Bilinski and Bremond [33]	The author suggested an extension of Improved Fisher Vectors (IFV) to represent a video using local and their Spatio-temporal positions. Further, the sliding window approach is applied for violence detection.	Violent-Flows, Hockey Fight, Movies datasets are used.	Crowd violence, fight, and non-fight actions	Accuracy on the violent- flows dataset is 96.4%, on the Hockey fight dataset is 93.7%, and on the Movies, the dataset is 99.5%.
2015	Serrano Gracia et al. [34]	Features are extracted from motion blobs to distinguish fight and non-fight sequences.	Movies, Hockey, and UCF 101 datasets are used.	Fight and non-fight, Punch Sumo actions	This proposed method is faster, with accuracies ranging from 70% to 98%, depending on the dataset.

3. PROPOSED METHODOLOGY

The proposed work will use videos collected from the camera to classify abnormal and normal activities in videos. In this work, there are two main modules presented for the methodology. The first module is the extraction and learning of the features by the Xception model. While the second module is the detection and classification of abnormal and normal activities by the LSTM classifier.

However, Pre-processing step is also a necessary step for good accuracy of the system. The dataset is divided into two parts like all other deep learning models: The train and test phases. The flow of research methodology is presented in Figure 2.

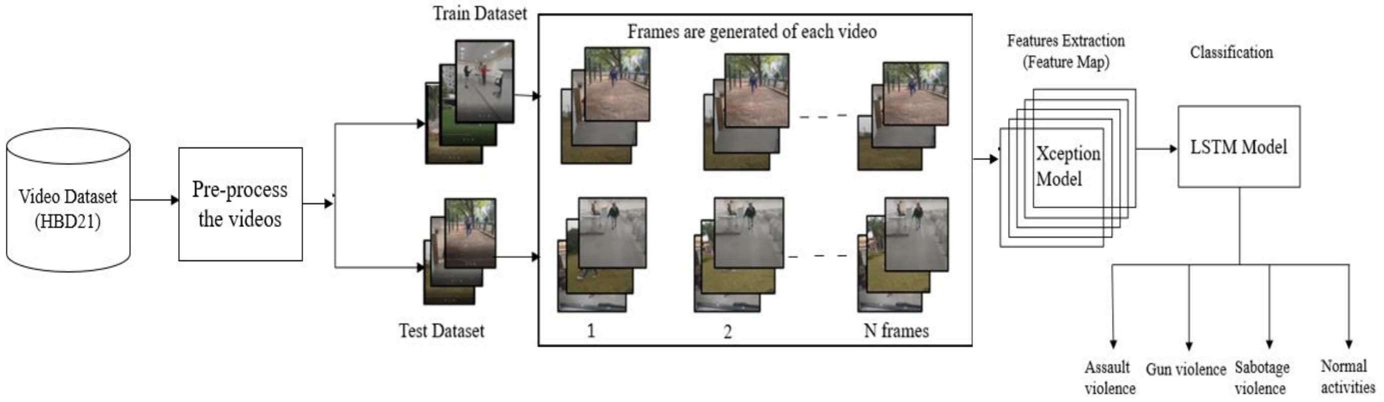


Figure 2. A framework for human action recognition

The proposed method is described in a stepwise manner.

Step 1: Load the HBD21 video dataset.

Step 2: Preprocess the dataset, i.e., increase the brightness of the video by adding 1.4 values in each pixel of video.

Step 3: Dataset is divided into 70% training and 30% testing video dataset.

Step 4: Frames are generated for each video file.

Step 5: Extract the features of frames using the transfer learning Xception model.

Step 6: Now, classify the features of videos with the help of the LSTM classifier.

Let's system takes a video stream(V^i) of length L, which has the same frame size a*b*c. A substream of specified sequence length(S.L) that defined the number of samples to be processed is segmented out from the video stream and passed to the action recognition module, where S.L.⊆L. The features are extracted from each video sequence and then use the LSTM network to distinguish the human actions. The entire process is repeated for each sub-stream of specified sequence length until the whole video stream is traversed. The algorithm depicts the whole workflow of the proposed work in the algorithm paradigm.

Algorithm: Anomaly action recognition process

Input: Sequence Lenght (S.L.), Image Size a*b*c

Initialize: N_t = Null

V=$\{v_1^t, v_2^t, v_3^t, v_4^t,\dots v_{S.L.}^t, v_1^{t+1}, v_2^{t+1}, v_3^{t+1}, v_4^{t+1}, \dots, v_{S.L.}^{t+m}\}$

Output: Final Outcome O

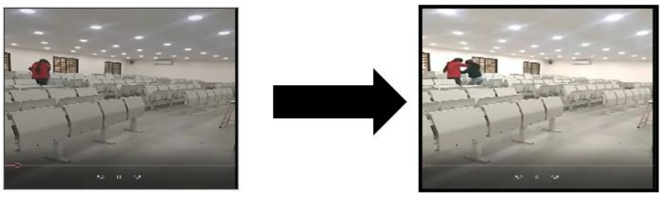
```
while V is not fully traversed, do
  Initialize tracker with V[0]
  For i=t→(t+m) do
    Obtain tracking outcomes, T, in frame
    For $v_j^i$ in $\{v_1^i, v_2^i, v_3^i, \dots v_{S.L.}^i\}$ do
      features = xception($v_j^i$, size = (299,299,3))
      $x_t$ = features
      $n_t$ = LSTM($x_t$, $n_{t-1}$)
      $n_{t-1} = n_t$
      T = $n_t$
    end for
  Add T to O
  End for
End While
```

As illustrated in the presented algorithm, feature extraction and classification are the two main modules for structuring a practical human behavior recognition system. Thus, this section explores the involved deep learning-based techniques for building an expert monitoring system which is described as follows:

3.1 Preprocessing of video dataset

The first step of this work is video enhancement. With the help of this technique, the brightness of videos is increased. This technique helps improve the accuracy of the system with the use of an image enhancement technique. By adding the value 1.4 in the video files, the brightness of videos is increased, as shown in Figure 3. Due to this operation, the performance of the system is improved.

Now, the next step is to extract the full-frame sequences of entire video files. The frames are generated for each video are 30 frames/second. For the proposed dataset, each video's length is 9 seconds, so total frames are generated of a single video ranging from 270 to 300 sequences of frames. So, we have a total of 456 videos, and approximately 1,27,680 frames are generated from the entire dataset.



(a) (b)

Figure 3. Sample of (a) original video (b) enhanced video

3.2 Feature extraction process using Xception model

In addition to the augmentation and frames generation process, now the features are extracted using Xception [6] deep learning model. The input image size of the model is 299*299. It extends the InceptionV3 architecture, which replaces the standard Inception modules with depth-wise separable convolutions. The top-1 accuracy and top-5 accuracy of the Xception model are 0.79 and 0.945, and it has the maximum accuracy among VGG-16, ResNet-152, and InceptionV3 models.

Xception is a modified architecture that mainly depends on two components [35]. The first is Depth wise Separable convolution, and the second component is shortcuts between convolution blocks as in ResNet. In this network, there are 71 layers. The Xception architecture is segmented into three major segments as shown in Figure 4; Entry Flow, Middle Flow, and Exit Flow. In the architecture, there are 36 convolution layers organized into 14 modules. Except for the first and last modules, all other modules are surrounded by linear residual connections. As can say, when trained on the ImageNet dataset, Collet [6], the Xception architecture is a linear stack of depthwise separable convolutions with residual connections. The features of videos frames are extracted using the Xception model as it gives the best features to classify this dataset's activities.

3.3 Transfer learning approach

With the help of a transfer learning approach, a model can be trained through a simple modification from one application to a new application [36]. The Xception model has a pool of parameters that need to be learned. At the initial step of training, these parameters are often set at random, resulting in a very large initial error for the network, which can easily results in low convergence and overfitting issues. In order to cope up with this issue, a supervised pretraining tactic of transfer learning based on the feature choice is introduced. The objective is to extract common feature representations in the Source and Target domain and then use these feature extraction to perform knowledge transfer. The Source Domain and Target Domain of transfer learning is specifically defined:

$$\mathrm{S(d)} = \{\mathrm{x}, \mathrm{P(x)}\} \tag{1}$$

$$\mathrm{T(d)} = \{\mathrm{x}, \mathrm{P(x)}\} \tag{2}$$

where, S(d) is the source domain, and T(d) is the target domain; the minimal probability distributions for the feature vector and the feature vector in one domain, respectively, are x and P(x). Imagenet recognition is the Source Task S(d) in the Source Domain, while Behavior Recognition is the Target Task T(d) in the Target Domain. The network parameters in the target task are established using the pre-trained model obtained from the Source task, and the feature information is transferred from the Source Domain to the Target Domain. Transfer learning is a technique that takes the existing knowledge to address different but related contexts. The last layer of the Xception model is removed, and the model can fine-tune according to the model's requirement. This model generates virtuous features on this dataset than other models.

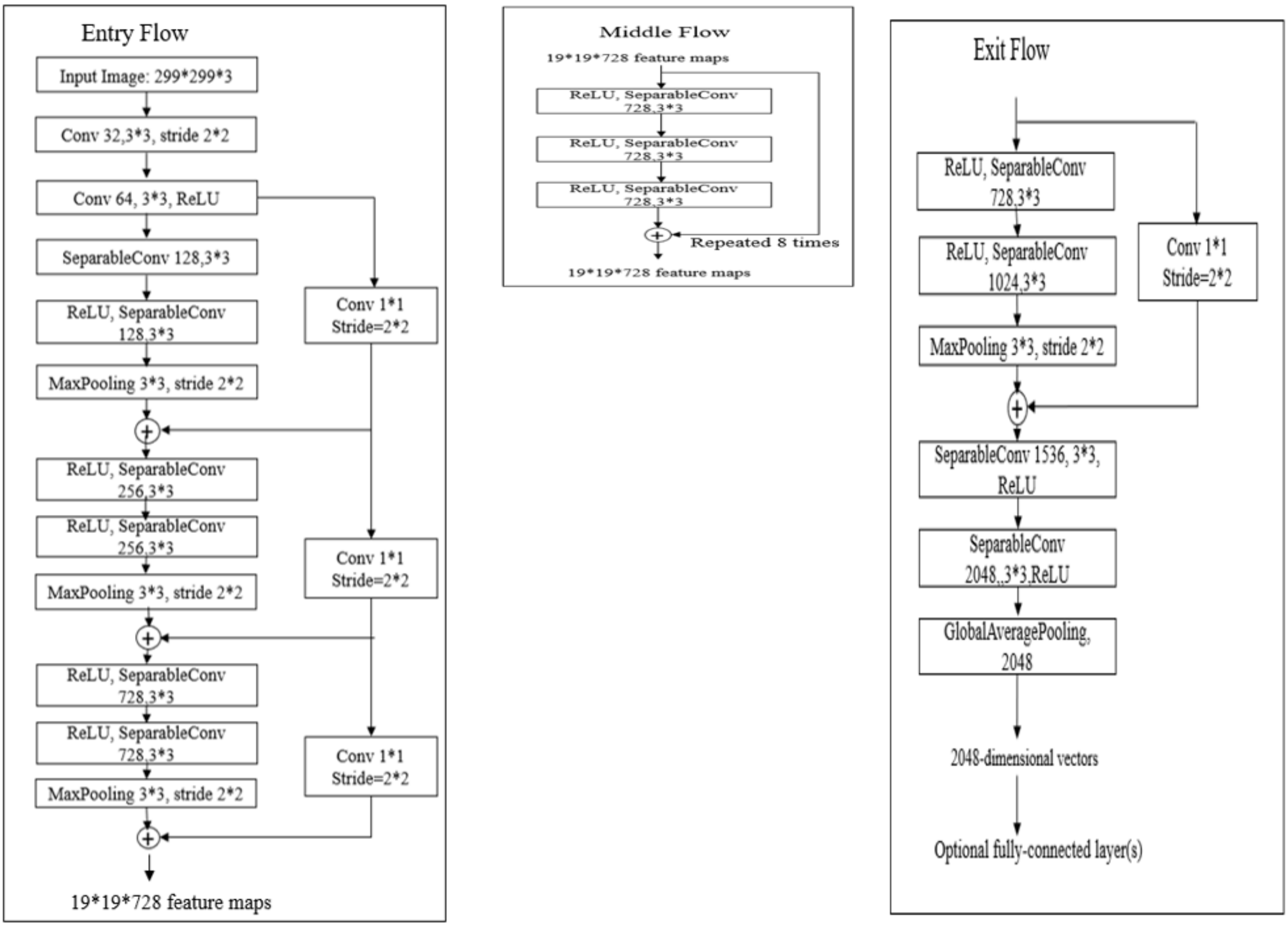

Entry Flow
Input Image: 299*299*3
Conv 32,3*3, stride 2*2
Conv 64, 3*3, ReLU
SeparableConv 128,3*3
ReLU, SeparableConv 128,3*3
Conv 1*1 Stride=2*2
MaxPooling 3*3, stride 2*2
ReLU, SeparableConv 256,3*3
ReLU, SeparableConv 256,3*3
Conv 1*1 Stride=2*2
MaxPooling 3*3, stride 2*2
ReLU, SeparableConv 728,3*3
ReLU, SeparableConv 728,3*3
Conv 1*1 Stride=2*2
MaxPooling 3*3, stride 2*2
19*19*728 feature maps
Middle Flow
19*19*728 feature maps
ReLU, SeparableConv 728,3*3
ReLU, SeparableConv 728,3*3
ReLU, SeparableConv 728,3*3
Repeated 8 times
19*19*728 feature maps
Exit Flow
ReLU, SeparableConv 728,3*3
ReLU, SeparableConv 1024,3*3
Conv 1*1 Stride=2*2
MaxPooling 3*3, stride 2*2
SeparableConv 1536, 3*3, ReLU
SeparableConv 2048,,3*3,ReLU
GlobalAveragePooling, 2048
2048-dimensional vectors
Optional fully-connected layer(s)

Figure 4. The Xception architecture [6]

3.4 Detection component (Activity Classification) by LSTM classifier

After the feature extraction process, the features vector is fed to LSTM [37] architecture to learn or train the model. LSTM is a kind of Recurrent Neural Network. RNN is the most promising model for abnormal behavior recognition in real-time videos. The goal of long-short-term memory (LSTM) is to use short-term memory to solve the problem of long-term dependence. Because the CNN model is not suitable for temporal variation, LSTM is utilised for temporal characteristics in succussive frames. This model consists of three Gates. Input Gate Forget Gate and Output Gate. The equations of all gates in LSTM models are represented in Eqns. (3), (4), (5).

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{3}$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{4}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{5}$$

where, i_t represents input gate, f_t indicated forget gate, o_t represents output gate, σ, w_x, $h_{t\text{-}1}$, x_t, b_x, h_t indicates the sigmoid function, weight for the respective gate (x) neurons, the output of the past LSTM block (at timestamp t-1), input at the current timestamp, biases and current observation for the respective gates(x), respectively.

The information to the cell state can be increased and decreased using an LSTM.

- A Forget Gate decides what sort of new information to store.
- An Input Gate determines which values in the cell will be updated.
- The results of an Output Gate are determined by the cell state.

The cell state, candidate cell state, and final output are represented by Eqns. (6), (7), (8).

$$\breve{c}_t = tanh(w_c[h_{t-1}, x_t] + b_c) \tag{6}$$

$$c_t = f_t * c_t - 1 + i_t * \breve{c}_t \tag{7}$$

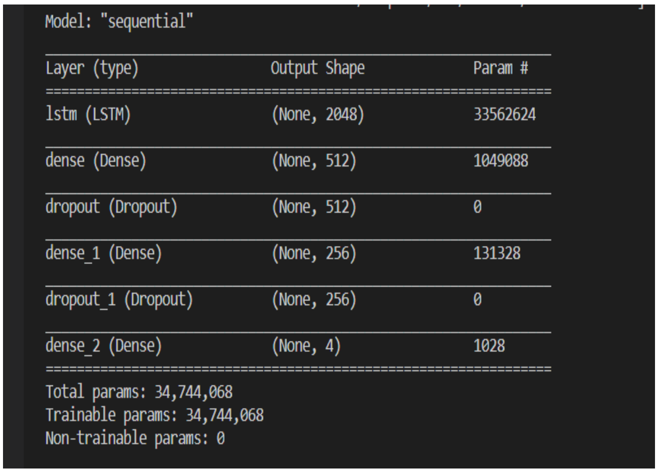
$$h_t = o_t * \tanh\,(c^t) \tag{8}$$

where, c_t stands the cell state at timestamp(t), $\breve{c}_t$ represents a candidate for cell state at timestamp(t), w_c denote to weight matrices and tanh represents tanh layer in deep recurrent network.

The modified LSTM model takes the 2048 features set for classification. An LSTM layer receives the feature map as an input. The number of classes in the dataset is equal to the number of neurons in the last layer. In this dataset, there are four classes. Thus the number of neurons here is four.

Furthermore, we have added one more dense layer, and a dropout layer is added for better results. LSTM layer's output is then sent via a dropout layer to a 4-unit Fully Connected (FC) dense layer with a ReLU activation function through the dropout layer. Before being categorized by a 4-class softmax function, the output from the dense layer is transferred to the dense layer. The learning rate is changed by 0.0001 to improve the result, and an ADAM optimizer is used. The system classifies the videos as abnormal behavior; three categories (Assault-violence, Gun-violence, Sabotage-violence) and normal behavior such as walking, running, jogging, exercise videos are present.

The graphical form of the LSTM model is shown below (Figure 5).



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 2048)              33562624
_________________________________________________________________
dense (Dense)                (None, 512)               1049088
_________________________________________________________________
dropout (Dropout)            (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 256)               131328
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 4)                 1028
=================================================================
Total params: 34,744,068
Trainable params: 34,744,068
Non-trainable params: 0
```

Figure 5. Graphical representation of LSTM model

4. EXPERIMENTAL ANALYSIS AND OUTCOMES

4.1 Proposed dataset

There are two types of actions for human behavior recognition: Abnormal and Normal actions. The proposed dataset Human Behaviour Dataset 2021(HBD21) mainly focused on abnormal activities. This dataset comprises abnormal action categories and normal actions. This dataset is prepared where natural daylight and artificial light conditions have been adopted to facilitate the recognition process. HBD21 dataset contains a total of 456 annotated videos of abnormal and normal actions. This dataset is captured in experimental labs, ground, parks, classes and prepared by 30 people. Moreover, this video dataset contains several objects like a bat, gun, racket, bottle, etc. In abnormal actions, there are three categories of violent actions.

The first category is assault, the second gun violence, and the third is sabotage violence actions videos present. In normal actions, there is no category divided further. However, running, jogging, doing exercise, walking videos are present in normal activities. The format of all videos is in '.mp4'. Table 2 describes the number of videos present in each category of the dataset. Figure 6 shows the sample video files of all categories of HBD21. The HBD21 dataset is available at Mendeley Data - Human Behavior dataset (HBD21) [38].

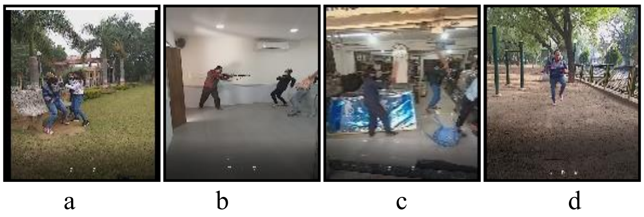


Figure 6. Samples of HBD21 dataset a) Assault violence, b) Gun-violence, c) Sabotage-violence, d) Normal actions

Table 2. Description of videos dataset (HBD21)

Categories of videos	Video's Description	Resolution	Number of videos	Total videos
Abnormal activity	Assault violence	640*480	104	348
	Gun violence	640*480	144	
	Sabotage violence	640*480	100	
Normal activity	Normal actions	640*480	108	108
	Total videos			456

4.2 Experimental setup

This proposed work is implemented in Windows 10, core i7 with 4GB Nvidia 1650 Ti GTX GeForce graphics GPU, 16GB RAM system. Python open-source software is used to implement this work, and TensorFlow, Keras libraries must be installed with python to execute the work. CUDA toolkit is considered to speed up the computation.

4.3. Results and discussion

The performance of the proposed work is calculated the accuracy of the model.

Accuracy metrics [39, 40] can be reported as True Positive (TP), signifying the suspect behavior for the detection of abnormal activity; a False Negative (FN) applies to the unusual action classification as usual; a False Positive (FP) represents the classification of a usual activity as unusual, and a True Negative (TN) represents non-suspicious action classifies correctly [41].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

With the help of this metric, the performance of the system can evaluate. Results are shown below.

Table 3. The comparison of accuracy of the suggested model with other models on four categories of HBD21

Model	Assault violence	Gun violence	Normal Actions	Sabotage violence
VGG19+LSTM	0.71	0.72	0.85	0.80
Inception+LSTM	0.88	0.84	0.88	0.87
Xception+LSTM	0.91	0.92	0.89	0.90
Transfer learning with Xception+ LSTM model (Proposed model)	**0.98**	**0.96**	**0.98**	**0.97**

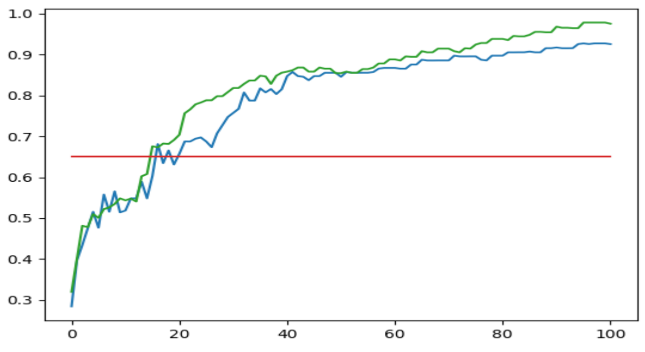


Figure 7. Training accuracy Vs Validation accuracy with benchmark

Table 4. Comparison of recognition accuracy of existing similar work with proposed work

Model	Dataset	Accuracy
Amrutha [1]	Synthesized dataset, CAVIAR, KTH, You-tube videos	87.15%
Peipei Zhou [9]	Chapter 1 HMDB51, UCF101, HOCKEY, Movies	97.06%
Proposed Model	**HBD21**	**97.25%**

The accuracy of work is achieved at 100 iterations, batch size was 16, and the learning rate was 0.0001. The proposed methodology (Transfer Learning with Xception and modified LSTM) worked well at the proposed dataset (HBD21). It is observed from Table 3 that the proposed transfer learning with Xception + LSTM network consistently outperforms the other models on the same datasets. It can conclude that the Transfer learning extracts spatial characteristics, while the LSTM model extracts temporal features, which enhance the accuracy score. Figure 7 depicts the training accuracy and validation accuracy with a (red line) benchmark-setting 0.65 value. The green line and blue line show the training and validation accuracy, respectively. It shows the graph changing as the number of epochs increases. Table 4 shows the comparison between the existing work from the proposed work. It can be observed that our model works well among both methods. The novelty of this paper is to generate an HBD21 dataset in the field of human behavior recognition and give a deep learning-based solution to detect abnormal and normal actions.

5. CONCLUSION

This paper introduced a new hybrid transfer learning model for abnormal and normal video classification on a novel HBD21 dataset. These categories of human behavior are very crucial for society's security. This research work is beneficial for police departments because they will get updated quickly and easily without human involvement.

The benefit of this study is that deep learning techniques are used in learning and detecting components. Transfer learning deep learning techniques make the computation fast and fine-tune the Xception model according to dataset categories. LSTM is used for classifying the classes of the dataset. Adding the dense and dropout layers and learning rate set to 1e-4 of the LSTM model that improves the classification results. The accuracy is achieved 92% upto 100 iterations on the HBD21 dataset.

It is possible to develop a novel technique and add a new component to this model for better results for more improvement. An optimization algorithm can be applied at the time of feature extraction for appropriate feature selection so that a good classification of actions can be done.

ACKNOWLEDGEMENT

The authors would like to thank the Madhav Institute of Technology and Science college to grant permission to use the labs and sports material to record the videos. Furthermore, the authors would like to thank the people involved in these videos for preparing the dataset.

REFERENCES

[1] Overview of Preliminary Uniform Crime Report, https://www.fbi.gov/news/pressrel/press-releases/overview-of-preliminary-uniform-crime-report-january-june-2020, accessed on January–June, 2020.

[2] Phan, H., He, Y., Savvides, M., Shen, Z. (2020). Mobinet: A mobile binary network for image classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3453-3462.

[3] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[4] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[5] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[6] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251-1258.

[7] Wang, X., Miao, Z., Zhang, R., Hao, S. (2019). I3d-lstm: A new model for human action recognition. In IOP Conference Series: Materials Science and Engineering, 569(3): 032035. https://doi.org/10.1088/1757-899X/569/3/032035

[8] Liang, M., Hu, X. (2015). Recurrent convolutional neural network for object recognition. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3367-3375. https://doi.org/10.1109/CVPR.2015.7298958

[9] Chong, Y.S., Tay, Y.H. (2015). Modeling representation of videos for anomaly detection using deep learning: A review. arXiv preprint arXiv:1505.00523.

[10] Popoola, O.P., Wang, K. (2012). Video-based abnormal human behavior recognition—A review. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(6): 865-878. https://doi.org/10.1109/TSMCC.2011.2178594

[11] Suresh, A.J., Visumathi, J. (2020). Inception ResNet deep transfer learning model for human action recognition using LSTM. Materials Today: Proceedings. https://doi.org/10.1016/j.matpr.2020.09.609

[12] Clarin, C., Dionisio, J., Echavez, M., Naval, P. (2005). DOVE: Detection of movie violence using motion intensity analysis on skin and blood. PCSC, 6: 150-156.

[13] Chen, L.H., Hsu, H.W., Wang, L.Y., Su, C.W. (2011). Violence detection in movies. In 2011 Eighth International Conference Computer Graphics, Imaging and Visualization, pp. 119-124. https://doi.org/10.1109/CGIV.2011.14

[14] Derbas, N., Quénot, G. (2014). Joint audio-visual words for violent scenes detection in movies. In Proceedings of International Conference on Multimedia Retrieval, pp. 483-486. https://doi.org/10.1145/2578726.2578799

[15] Wang, H., Kläser, A., Schmid, C., Liu, C.L. (2013). Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 103(1): 60-79. https://doi.org/10.1007%2Fs11263-012-0594-8

[16] Laptev, I. (2005). On space-time interest points. International Journal of Computer Vision, 64(2): 107-123. https://doi.org/10.1007/s11263-005-1838-7

[17] Zhang, T., Yang, Z., Jia, W., Yang, B., Yang, J., He, X. (2016). A new method for violence detection in surveillance scenes. Multimedia Tools and Applications, 75(12): 7327-7349. https://doi.org/10.1007/s11042-015-2648-8

[18] Sabokrou, M., Fayyaz, M., Fathy, M., Klette, R. (2017). Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing, 26(4): 1992-2004. https://doi.org/10.1109/TIP.2017.2670780

[19] Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition, 58: 121-134. https://doi.org/10.1016/j.patcog.2016.03.028

[20] Xu, H., Li, L., Fang, M., Zhang, F. (2018). Movement human actions recognition based on machine learning. International Journal of Online Engineering, 14(4). https://doi.org/10.3991/ijoe.v14i04.8513

[21] Tay, N.C., Connie, T., Ong, T.S., Goh, K.O.M., Teh, P.S. (2019). A robust abnormal behavior detection method using convolutional neural network. In: Alfred R., Lim Y., Ibrahim A., Anthony P. (eds) Computational Science and Technology. Lecture Notes in Electrical Engineering, vol 481. Springer, Singapore. https://doi.org/10.1007/978-981-13-2622-6_4

[22] Wang, J., Xia, L. (2019). Abnormal behavior detection in videos using deep learning. Cluster Computing, 22(4): 9229-9239. https://doi.org/10.1007/s10586-018-2114-2

[23] Schuldt, C., Laptev, I., Caputo, B. (2004). Recognizing human actions: a local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, pp. 32-36. https://doi.org/10.1109/ICPR.2004.1334462

[24] Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R. (2007). Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(12): 2247-2253. https://doi.org/10.1109/TPAMI.2007.70711

[25] Amrutha, C.V., Jyotsna, C., Amudha, J. (2020). Deep learning approach for suspicious activity detection from surveillance video. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 335-339. https://doi.org/10.1109/ICIMIA48430.2020.9074920

[26] Nayak, R., Pati, U.C., Das, S.K. (2020). A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing, 106: 104078. https://doi.org/10.1016/j.imavis.2020.104078

[27] Febin, I.P., Jayasree, K., Joy, P.T. (2020). Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm. Pattern Analysis and Applications, 23(2): 611-623. https://doi.org/10.1007/s10044-019-00821-3

[28] Ramzan, M., Abid, A., Khan, H.U., Awan, S.M., Ismail, A., Ahmed, M., Mahmood, A. (2019). A review on state-of-the-art violence detection techniques. IEEE Access, 7: 107560-107575. https://doi.org/10.1109/ACCESS.2019.2932114

[29] Ullah, F.U.M., Ullah, A., Muhammad, K., Haq, I.U., Baik, S.W. (2019). Violence detection using spatiotemporal features with 3D convolutional neural network. Sensors, 19(11): 2472. https://doi.org/10.3390/s19112472

[30] Dandage, V., Gautam, H., Ghavale, A., Mahore, R., Sonewar, P.A. (2019). Review of violence detection system using deep learning. International Research Journal of Engineering and Technlogy (IRJET), 6(12): 1899-1902.

[31] Khaleghi, A., Moin, M.S. (2018). Improved anomaly detection in surveillance videos based on a deep learning method. In 2018 8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN), pp. 73-81. https://doi.org/10.1109/RIOS.2018.8406634

[32] Zhou, P., Ding, Q., Luo, H., Hou, X. (2017). Violent interaction detection in video based on deep learning. Journal of Physics: Conference Series, 844(1): 012044.

[33] Bilinski, P., Bremond, F. (2016). Human violence recognition and detection in surveillance videos. In 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 30-36. https://doi.org/10.1109/AVSS.2016.7738019

[34] Serrano Gracia, I., Deniz Suarez, O., Bueno Garcia, G., Kim, T.K. (2015). Fast fight detection. PloS One, 10(4): e0120448. https://doi.org/10.1371/journal.pone.0120448

[35] Venu, S.K. (2020). An ensemble-based approach by fine-tuning the deep transfer learning models to classify pneumonia from chest X-ray images. arXiv preprint arXiv:2011.05543.

[36] Li, Z., Ye, J. (2019). Abnormal behavior recognition based on transfer learning. In Journal of Physics: Conference Series, 1213(2): 022007. https://doi.org/10.1088/1742-6596/1213/2/022007

[37] Arifoglu, D., Bouchachia, A. (2017). Activity recognition and abnormal behaviour detection with recurrent neural networks. Procedia Computer Science, 110: 86-93. https://doi.org/10.1016/j.procs.2017.06.121

[38] Ruchi, J., Manish, D. (2021). Human Behavior dataset (HBD21). Mendeley Data, 1. https://doi.org/10.17632/xh9pgb3w8c.1

[39] Tripathi, R.K., Jalal, A.S., Agrawal, S.C. (2018). Suspicious human activity recognition: A review. Artificial Intelligence Review, 50(2): 283-339. https://doi.org/10.1007/s10462-017-9545-7

[40] Jayaswal, R., Jha, J. (2017). A hybrid approach for image retrieval using visual descriptors. In 2017 International Conference on Computing, Communication and Automation (ICCCA), pp. 1125-1130. https://doi.org/10.1109/CCAA.2017.8229965

[41] Jayaswal, R., Dixit, M. (2020). Comparative analysis of human face recognition by traditional methods and deep learning in real-time environment. In 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT), pp. 66-71. https://doi.org/10.1109/CSNT48778.2020.9115779