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

Traitement du Signal
Vol. 38, No. 6, December, 2021, pp. 1623-1635
Journal homepage: http://iieta.org/journals/ts

A Deep Learning Based System for the Detection of Human Violence in Video Data

Muhammad Shoaib[1*], Nasir Sayed[2]

[1] Department of Computer Science, CECOS University of IT and Emerging Sciences, Peshawar 25000, Khyber Pakhtunkhwa, Pakistan
[2] Department of Computer Science, Islamia College Peshawar, Peshawar 25120, Khyber Pakhtunkhwa, Pakistan

Corresponding Author Email: mshoaib@cecos.edu.pk

https://doi.org/10.18280/ts.380606

Received: 28 October 2021
Accepted: 2 December 2021

Keywords:
violence detection, deep learning, convolutional neural network, image classification object localization

ABSTRACT

The number of security cameras positioned within the surrounding area has expanded, increasing the demand for automatic activity recognition systems. In addition to offline assessment and the issuance of an ongoing alarm in the case of aberrant behaviour, automatic activity detection systems can be employed in conjunction with human operators. In the proposed research framework, an ensemble of Mask Region-based Convolutional Neural Networks for key-point detection scheme, and LSTM based Recurrent Neural Network is used to create a deep neural network model (Mask RCNN) for recognizing violent activities (i.e. kicking, punching, etc.) of a single person. First of all, the key-points locations and ground-truth masks of humans in an image are selected using the selected region; the temporal information is extracted. Experimental results show that the ensemble model outperforms individual models. The proposed technique has a reasonable accuracy rate of 77.4 percent, 95.7 percent, and 88.2 percent, respectively, on the Weizmann, KTH, and our custom datasets. As the proposed effort applies to industry and in terms of security, it is beneficial to society.

1. INTRODUCTION

With the rise in public demand for security and image processing technologies, intelligent surveillance research is gaining traction, with many researchers attempting to process surveillance video data using computers' powerful computing capacity for enhanced security control. The overwhelming amount of video captured from our large-scale surveillance cyber-physical systems should be reviewed today to ensure the public's need for protection. Human observers-based security systems, on the other hand, are ineffective. Missed alerts are typical due to personnel's limited ability to constantly track surveillance footage, necessitating the introduction of automatic alerting strategies for suspicious activity such as war. The proposed method would also serve as a basis for e-physical systems that track and regulate physical entities [1]. The behaviourist approach may employ computer vision to detect and describe a target's motion as a vital step toward behaviour identification. High-level individual structure and low-level picture information methods are the two sorts of behavioural analysis approaches. Point representations of the body, 2D models of the human body [2], and 3D models of the human body [3-5] are all examples of information employed in greater techniques.

Furthermore, to obtain more detail, this approach often requires multi-viewpoint video or 3D cameras. As a result, the model becomes more complex and computationally intensive. Moreover, according to the infancy of individual action recognition methodologies, the program's stability is impossible to ensure. Motion trajectories [6], shape characteristics, texture features [7], optical flow characteristics, and other image information are employed in behavioural analysis approaches focused on low-level picture information. This method, characterized by simplicity and minimal complexity, specifies the target's behaviours from a macro perspective. Therefore, the performance is better in real-time processing as compared to advanced human structure-based approaches. Unfortunately, combat is an example of atypical action that puts citizens' lives in jeopardy and has a negative effect on society.

Consequently, putting in place real-time, accurate fighting behaviour recognition to aid security agents and ensure their protection is crucial. Furthermore, optical flow information derived from video sequence pixel shifts has strong spatial and temporal properties, and it is commonly used in video processing to explain target motion patterns [8]. Thus, in recent years, optical flow-based battle detection has gotten much coverage.

Some researchers have been able to detect fights by adjusting the threshold of optical flow features. Following that, researchers focused on battle detection using machine learning algorithms. Some descriptor- and cutoff point detection methods have been presented based on the retrieved optical flow information. Liu et al. [9] combined texture and optional flow features to increase recognition rates in congested scenes. The magnitude-based weighted direction histogram can reduce noise-induced directional confusion to some extent, and its entropy can be utilized to characterize motion confusion. It does not, however, account for changes in optical flow energy throughout a war. Akti et al. [10] employed attention layer-based LSTM RNN to detect fighting scenes in surveillance videos to detect fighting activities.

The shift in optical flow energy of the target object may be used to detect anomalous activity. However, it does not

differentiate between battle and fleeing [9]. Liu et al. [9] successfully reduced the rate of anomaly detection misclassification by using motion and shape features to detect war. However, it continues to make mistakes when judging session conduct. Ren et al. [11], Huang and Chen [12] proposed a novel idea for function definition by using mean and variance to distinguish fights based on the amplitude and direction of the optical flow statistics. Although the threshold approach detection method is quick and simple to use, it performs badly in videos with many shooting views, and it is difficult to avoid misclassification of activities like running and chatting. Several academics have also investigated machine learning models to recognize fighting and distinguish it from other activities.

Gao et al. [13] introduced a new features extraction method (oriented violent flow), which was combined with two state-of-the-art classifiers, i.e. SVM and AdaBoost, to develop a violence recognition system in video data. Simple action like punching, on the other hand, have a poor identification rate in this form. Yang et al. [14] proposed a real-time method based on optical flow histograms, in which the scale descriptor is of optical flow direction (HOFO), and histogram of rotation invariant feature descriptors are computed and detected using an SVM classifier. However, this approach has a problem with session error judgments. When the difference in optical flow is taken into account, the result is higher. Lejmi et al. [15] Spatio-temporal features are extracted from violent and non-violent videos; the labeled sequential features are loaded into an LSTM classification model, which is then used to recognize eight different types of violence and non-violence. However, this approach is less successful in crowded conditions, so Mahmoodi and Salajeghe [16] recently proposed the histogram of optical flow size and direction (HOMO) as a new function descriptor to enhance current violence detection. Xu et al. [17] proposed a motion activating information retrieval technique from surveillance videos for localization guidance using optical flow maps. Febin et al. [18] proposed a cascaded violence detection system for action recognition based on motion boundary SIFT (MoBSIFT) and motion filtering. Machine learning models have shown to be effective in various situations, but most overlook the importance of multiple views in battle detection in real-world scenarios. Existing algorithms are limited in solving misjudgment in running, overtaking, and other circumstances with motion characteristics close to combat. By examining the motion characteristics of battles, we propose two descriptions for resilience throughout video collection and misjudgment of non-fighting behaviours: motion direction inconsistency (Modi) and weighted motion direction inconsistency (WMoDI). For motion area tagging, this paper also employs a deep learning system. To improve the accuracy and robustness of multiple views, we integrate three existing descriptors and two unique descriptors to calculate the final features of six statistical characteristics using these descriptors. The proposed descriptors are sufficiently discriminative to differentiate between combat and non-fighting actions in videos with many shooting views according to experimental findings on the CASIA action dataset and the interaction dataset. The approach also has some precision benefits and can accurately distinguish fights in several datasets.

The traditional machine learning approach for recognizing human actions relies heavily on human observation to extract visual features. A large amount of human experience and background knowledge constrains it. Most of these algorithms work well when only using the exact dataset from a single experiment. Many digital video clips are available on the Internet containing violent content, such as YouTube. Our humans cannot meet the demand for labeling all the data composed of violent content and independently extracting them with higher accuracy. Machines must learn how to extract characteristics of human behavior. The proposed model is implemented using the resnet 101 CNN model; the first and second descriptors are used by the regional proposal network that uses the CNN features extracted by the resnet 101, the third descriptor is used to extract features from the Ground truth mask, while descriptor four is used to find the patterns which are used by the boundary box regression layer and finally the six descriptors along with LSTM Layer is used for recognizing the human activities.

The proposed deep learning-based violence detection model is trained on two benchmarks (KTH and Weizmann) and a custom-developed dataset. The performance of the models is evaluated using various evaluation metrics; the proposed model achieved an average accuracy on the validation set: Weizmann=77.4, KTH=95.7%, and Custom dataset=88.2%.

In section 2 literature review, the literature on action recognition and violent interaction detection (vid) with the overall literature review summary has been discussed. In the Methodology section, in preprocessing stage, the person position is identified, each video frame is selected and extracted with a time interval of 0.1 seconds and resized to a resolution of 128×128. The phase in the methodology section is key points selection model is discussed, and in the final stage of the methodology section, the LSTM model is discussed, which is responsible for detecting violent scenes. Section four is about results and discussion; first, various benchmark and custom-developed datasets are discussed. After that, the training effectiveness with a time scheme is discussed. A detailed experimental procedure is performed to evaluate the performance of the proposed scheme, and finally, the results achieved by the proposed model are compared with the model from the state-of-the-art.

2. LITERATURE REVIEW

Many models have been developed to detect various human behaviours, including violent behaviour. The surge in interest in this field has resulted in a slew of studies that have been incorporated into several surveys. The researches [19-25] present more comprehensive vision-based activity recognition domain surveys. Moeslund et al. [24] propose a new hierarchical structure with four phases: initialization, tracking, pose estimation, and recognition. Furthermore, Badi [20] only considers whole-body movements, leaving out Aggarwa and Cai's [19] work on gesture recognition. Analysis, tracking, and recognition of body structure are discussed [21]. Low, medium and high-level vision duties are referred to as action potential, activity recognition, and action recognition. Gavrila [22] discussed detection and recognition strategies in two and three dimensions. In a recent study [26], methods for recognizing violence in surveillance footage were emphasized. Recent optical flow studies on violence detection and coherent movement descriptors for transferring items has gotten plenty of press. They diagnosed human interest via the usage of optical waft histograms in each horizontal and vertical guideline as movement descriptors [27]. Wang and Schmid [28] hired optical waft to generate densely tracked sampled

places that have been later applied to symbolize human activities. It is a powerful technique for movement recognition. However, it calls for specialized hardware and is computationally expensive. ViF, a singular descriptor for real-time crowd violence detection utilizing optical flows, changed into recommended by Hassner et al. [29]. ViF, on the alternative hand, did now no longer utilize the optical float course characteristic. Yuan Gao had resolved this hassle in the study by Gao et al. [13] by growing a singular characteristic extraction technique referred to as Violent Oriented Flows (OViF). Instead, it uses the advantage of variable motion importance facts in statistical motion guidelines.

Some researchers have evolved deep neural community-primarily based models totally that figuring out violent moves in the video, considering that deep convolutional neural networks are getting greater outstanding withinside the area of vision-primarily based movement detection [30-33]. FightNet changed and constructed through Zhou et al. [33] to portray the complex photograph. FightNet changed into skilled to cooperate spatial and temporal networks using three styles of input images: RGB, optical float, and acceleration. Dong et al. [31] describe a state of affairs wherein a little work is accomplished on someone to hit upon non-public violence. This study employs a 3-circulate deep neural community. Two extraordinary turbulent characteristics styles, including transferring gadgets and accelerating flow maps, are retrieved from video collection without any pre-processing throughout the first flows. The 1/3 circulate is an encoding approach primarily based totally on Long Short Term Memory (LSTM). Finally, all three streams are mixed using rating-stage fusion, and the quality fact rating for violent films is determined. Sudhakaran and Lanz [32] furnished a variation to the above approach, and thus he uses a series of convolutional layers to extract body-stage capabilities, which the use of ConvLSTM had then aggregated. The convolutional neural community evaluates neighbourhood motion withinside the video together with the ConvLSTM. The version decreases the chance of overfitting through encoding modifications withinside the video the use of adjacent body differences. Another approach proposed through Baccouche et al. [30] is immediately researching training image samples. It is a -step neural-primarily based version that classifies a hard and fast of human sports using apriori modelling. The ConvNets are, to begin with, prolonged to a few dimensions, in which they self-research the capabilities of interest. These self-discovered capabilities are then hired in the collection to train the LSTM based RNN model, which similarly classifies it withinside the two stages. There may be an overhead of training each range independently in the approach mentioned above, which may be prevented by growing a one-step version. A single-step 3D-ConvNet-LSTM framework can reduce computing charges by simply training the proposed model once. The following format is used for the remainder of the paper: Segment 3 digs into the details of the suggested Mask RCNN with key-points detection model and architecture, as well as the LSTM model employed in our study. In section 4, we give experimental results on our dataset and a standard dataset utilizing several deep model approaches.

2.1 Action recognition

In recent years, action recognition has gotten much attention as a research subject in video analysis.

To obtain the final prediction performance, action recognition algorithms previously used hand-crafted features such as scale-invariant feature transform (SIFT) [34] histogram of directed gradients (HOG) [35], enhanced dense trajectories (IDT) [28], and multiclass support vector machines for classifying various human actions [36] (SVM). IDT outperforms all of the hand-crafted functions. Deep learning algorithms have considerably improved the efficiency of action recognition in recent years. The test-driven development (TDD) algorithm was proposed by Wang et al. [37]. By substituting standard features in the IDT algorithm with deep CNN-based features, two normalization approaches were developed: Normalization of spatial and temporal dimensions and channel normalization. Single RGB images and streaming pictures are used as sources to CNN and merged features to model Spatial-temporal features simultaneously in the two-stream approach [38, 39]. To boost the model's expressiveness, the TSN method14 recommends a two-stream oriented multi-segment approach. By establishing optical flow networks to replace the pre-extraction of optical flow in the video, Zhu et al. [40] incorporated optical flow networks to the optical flow branch of the two-stream model. This resulted in a considerable increase in recognition speed. The recognition accuracy determines the speed. To learn the global information in the original input data, Wang et al. [41] created a nonlocal block. The C3D approach [42] is utilized to accomplish direct action recognition, which considerably improves identification time by modelling the Spatio-temporal properties of the input frame series. Tran et al. [43] examined the performance of common network designs for action recognition tasks and discovered that three-dimensional (3D) convolution outperformed two-dimensional (2D) convolution. This year, the authors introduced the R(2 + 1)D structure after one-dimensional (1D) convolution was widely employed for channel shift in deep learning. The R(2 + 1)D structure of 3D convolution is split into 2D spatial convolution and 1D temporal convolution, yielding additional nonlinear layers for better outcomes on action recognition datasets and simpler optimization Decomposing 3D convolution into 2D spatial convolution and 1D temporal convolution, according to Zhao et al. [44], will result in high recognition accuracy. 1D convolution, but on the other hand, makes an underlying assumption: that characteristics with different time steps be matched, which aids 1D regularization in aggregate features in the very same location. Motion detection suffers greatly as a result of this assumption. TrajectoryNet [44] is described in this study. TrajectoryNet can compensate again for distortion of objects moving and the significant changes caused by video's persistence of movement, allowing visual features to be grouped along motion routes.

This method delivers considerable improvements in performance on two big datasets by directly replacing the 1D temporal convolution layer in Separable-3D (S3D) [44] with a trajectory convolution layer. On this assumption, the Efficient Convolutional Network Online Video Understanding (ECO) method [45] offers two recognition stages: collecting 2D features from the input frame sequence and stacking the appropriate channels of all 2D feature maps as input to a 3D CNN.

2.2 Violent interaction detection (VID)

VID distinguishes among structures that depend on guide traits and strategies that depend on deep gaining knowledge of and are in the direction of motion recognition. The IDT [28]

capabilities are the maximum, not unusual, place of the guide function-primarily based techniques. Serrano Gracia et al. [46] brought new traits received from movement trajectories among consecutive frames to differentiate competitive behaviours. EY Fu and his colleagues [47, 48] hired optical flow akin to the two-flow approach utilized in movement recognition to get mobility traits. Y Gao et al. provided Violence-orientated flow [13] as a brand new function extraction technique for reading dynamic value alternate records for VID tasks (OViF). VID distinguishes between systems that rely on manual characteristics and methods that rely on deep learning and are closer to action recognition. The IDT [28] features are the most common of the manual feature-based techniques. Wang and Schmid [28] used violent crow textures to detect violent activity in the input video population, displaying [49] population density using a grayscale co-generation matrix (GLCM) method. Based on the distribution of the optical flow field, Zhou et al. [50] segmented the motion region as the input video and used bag-of-words (BoW) [51] to extract two low-level aspects of the segmented motion region's Spatio-temporal characteristics. The final detection results were categorized using SVM after the features were coded to reduce redundant features. Mostly on fields hockey set of data and several other two linked data sets, the approach does have a high detection accuracy. Deep CNN-based feature extraction approaches have become more widespread as hardware platforms, and deep learning algorithms have improved. TSN's first two patterns were combined with TSN [52] and accelerated fields [53]. Likewise, Xia et al. [54] created a new frame-difference modality to complement the original RGB modality before combining the effects. Because it only uses RGB frames, this technique will boost recognition speed while maintaining recognition efficiency. Furthermore, most of the videos in the associated dataset are cropped from movie snippets or web videos, with few actual scene footage. Fu et al. [55] introduce a new cross-species learning approach that is computationally simple and effectively tackles the limits of human conflict movies by exploiting natural similarities between people and animals and the overall learning mechanism to overcome this problem.

In recent years, various techniques for detecting violent activities from video data have been proposed in the literature. The reviewed literature examines a variety of cutting-edge violence detection techniques. According to the reviewers, violence detection methods are classified into three categories based on the classification technique used: traditional machine learning, support vector machines (SVM), and deep learning. The feature extraction and target detection techniques are also discussed for each method. In addition, the datasets and video features used in the techniques are discussed, as they are critical in the recognition process. The steps of the studied method have been presented in the architecture diagram for better understanding. The overall findings have been discussed, which may aid in the identification of future research opportunities in this field.

3. METHODOLOGY

This section will clarify the proposed methodology at the granularity stage, with a complete deep neural network used to classify a single person's violent activities (kicks, punches). The video will be captured by the camera and then fed into the updated video.

In the Mask-RCNN [56] model, a single person recognized the presence in the picture. When a person is detected in a video, the keypoint detection model receives a continuous collection of 15 frames, which extracts the person's shape and deep spatial features. The LSTM [57] classifier uses Spatio-temporal features to predict violent events, and these key-points and the person's form (mask) are fed into it. The proposed approach is divided into subsections, which are depicted in Figure 1.

3.1 Pre-processing

Machine learning approaches do not rely on hand-crafted functions and instead use training data to recognise objects of interest automatically. These methods are incapable of simultaneously detecting and recognizing behaviour on inputs of any size. As a consequence, the process must be divided into several stages. The person's position must first be identified. After that, the activity can be identified. The key stage of the proposed approach is identifying a single individual in the input video, which guarantees that the subsequent stages of violent activity detection are accurate. In this segment, only those sequences that include an individual are considered for person detection instead of processing the entire video.

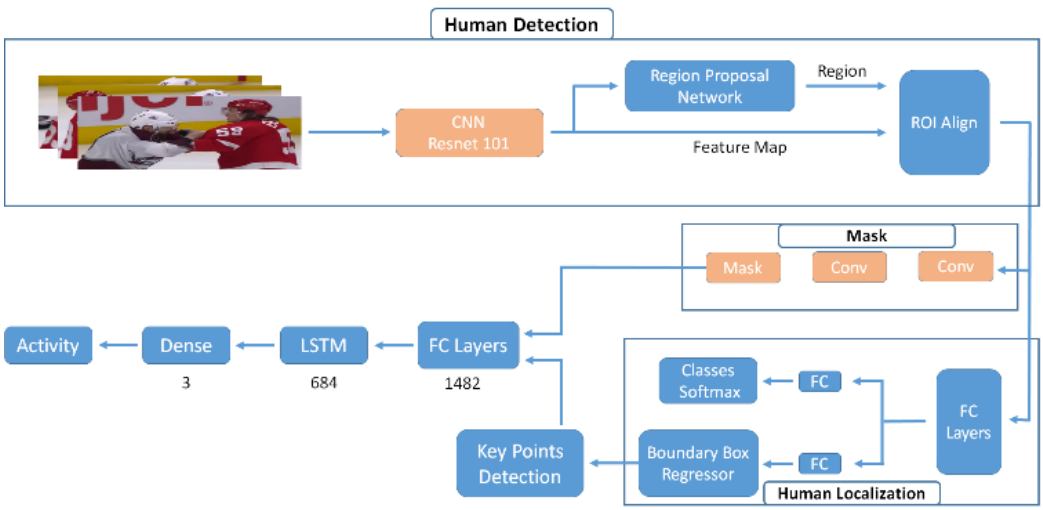


Figure 1. Depict the proposed single-person violence detection system's architecture

The Mask-RCNN model is fed the video stream, which has been modified to only detect single individuals. Semantic segmentation is a state-of-the-art technique used for automatic multi-object detection in video and images, including the context. Each video frame is selected and extracted at 0.1 s intervals, calibrated to ensure uniformity (128,128). The next step is to locate the moving person's region in a video frame. As a result, each video frame is passed to the state-of-the-art ResNet 101 CNN model [58], which extract a high dimension of learning features. The presence (or absence) of the object of interest and the score of its selected object about that selected region are then predicted using these extracted features by a region proposal network (RPN). A pooling layer for the Region of Interest is added to minimize the size of all selected region proposals to the same size.

The uniform size recommendations are then transmitted to fully connected layers, which identifies and outputs all objects during the first instance of the observed individual in our example. After that, a joint intersection set (IoU) with ground truth boxes is computed for each predicted region. Estimate the fragmentation masks for each entity belonging to a specified area using the RoI based on the IoU values. Each area receives a mask of 28×28 inches, which is then improved for interpretation. The bounding boxes represent the anticipated categories of the model and the probability for specific categories and the masks of the people that were recognized. In our proposed work, we changed the Mask-RCNN, and at the first instance of the person, the category is detected by the model. It was created to detect numerous occurrences of diverse objects from 81 different categories (including background).

On certain dataset samples, Figure 2 shows an example of individual detection using the original and upgraded models. The modified Mask RCNN finds just one instance of the first observed individual in a crowd, as shown in Figure 2.

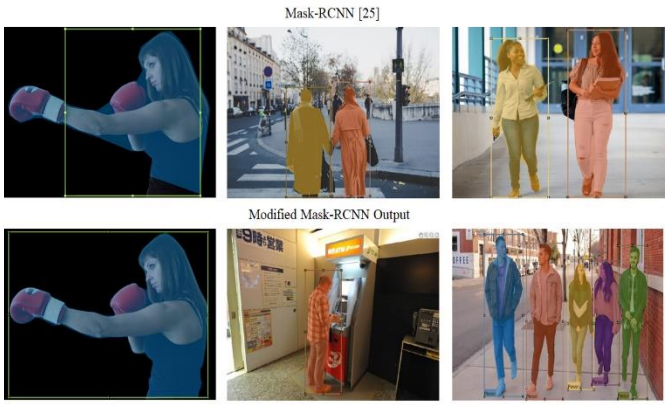


Figure 2. Mask-RCNN and its variants multiple and single people in a video sample were detected using mask-RCNN, respectively

3.2 Key-point detection

The entire control system, as seen in Figure 3, uses the bounding boxes obtained in the previous module to locate 17 human joints. 15 of the 17 interest sites (as shown in Figure 3) are centred on key-spots of the hands and legs, which are deeply engaged in human activity.

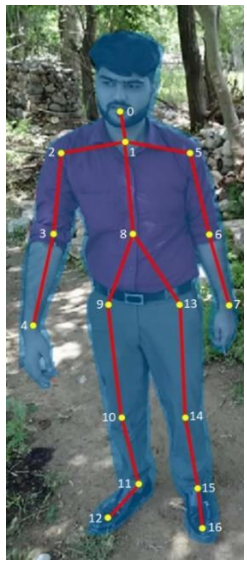


Figure 3. Visualization of key-points

Since not all key-points in videos can be identified due to obstacles such as clothing or occlusions, a combination of key-points can aid in action differentiation. A single frame's input (in this case, joints) is saved as: *[j0_r, j0_l, j1_r, j1_l, j2_r, j2_l, j3_r, j3_l, j4_r, j4_l, j5_r, j5_l, j6_r, j6_l, j7_r, j7_l, j8_r, j8_l, j9_r, j9_l, j10_r, j10_l, j11_r, j11_l, j12_r, j12_l, j13_r, j13_l, j14_r, j14_l, j15_r, j15_l,j16_r,j16_l].* The 17 joint p and q location key-points were stored in JSON format after the model was run on a single frame. Only the r and l positions for each frame were saved from the JSON file, and the shape of the individual structure was masked onto that pointing. This is used to build a database for that frame's active category numbers, joint 2D locations, and masked regions.

When a single joint is not located, it has the coordinates [0.0,0.0]. The center of the enclosing box d, which contains the body from the previous module is established as $d_c \in R^2$, width d_w and height d_h:d=(d_c,d_w,d_h). Furthermore, by translated through using box center dc as well as scaling by both the box size (d_w, d_h), the joint I is normalized by d, which would be portrayed by $N(i{:}d)$.

$$N(i{:}\,d) = \begin{pmatrix} \frac{1}{dw} & 0 \\ 0 & \frac{1}{dh} \end{pmatrix} (-dc) \tag{1}$$

Thus, the pose in the localized coordinate system may be predicted using Eq. (1) while the representation of joints is normalized $N(i{:}d)$.

$$z *= N^{-1}(\Psi(N(J);\theta)) \tag{2}$$

where, $\Psi\,(x;\,\theta)$ € R2k denotes the mathematical function that is used to retreat the image matrix j to a vector containing normalized poses and k symbolizes the parameters of a model. This is focused on the Ψ of the [59] architecture of the model, which consists of seven layers, each of which is comprised of a linear transformation that is followed by a nonlinear transformation. The conclusion layer generates the regression's target value, in this case 34 joint dimensions.

3.3 Long short term memory (LSTM)

The proposed method will be an excellent choice for detecting behaviour because it considers spatial and temporal aspects. To encode temporal variation, recurrent neural networks (RNN) are required. The detection technique should have the ability to hold together the final dimensions features and how they fluctuate over time to recognize violent behaviour.

The 15×38 input is passed through the time distribution layers to isolate the time-step data of each spreading layer, which would otherwise be transformed to the form of 1×38 size vector, where the time-step is represented by 15, and the value 38 represents the 17 joint key points from the previous module's human mask detection. After that, they are divided by half and quarter of the input vector dimension, then multiplied by alternate filtering and time distribution layers. The LSTM (Long Term Short Term Memory) layer would then be used. The LSTM layer aims to obtain temporal information to transfer the data to the samples, each of which has 15 time steps 1482 features. For the LSTM, each layer generates secret unit information, and each layer has 684 units. The data is passed to a three-layer dense output layer, which outputs one of the three groups.

4. RESULTS AND DISCUSSION

4.1 Database

At the start of the experimental and discussion section, the description of the benchmark and our own (Custom) created dataset are discussed; the benchmark datasets used for evaluating our proposed violence detection system are KTH [60] and Weizmann [61]. KTH dataset (n=20) 25 participants in the database have performed in four scenarios: outdoors, outdoors with various images scale, outdoors scene with subjects cloths and some indoors scenes.

Walking, jogging, boxing, waving, and applauding are only a few of the many human actions.

There are actually 2391 sequences in the KTH database. Each sequence was shot with a fixed camera against a constant backdrop at a rate of 25 frames per second. These sequences have been downscaled to a spatial resolution of 160×120 pixels and are, on average, 4 seconds long.

4.2 Weizmann

The Weizmann [61] dataset includes bending, jump raising, jumping, jumping in position, Jogging, side leaping, skipping, strolling, one-handed waving, and two-handed waving are examples of one and two different wavings. Each process is carried out by nine objects, each with 90 low-resolution (180144) photos.

4.3 Custom dataset

Indoors, dressed up in a variety of outfits (see Figure 4). For assessment, we created a video database of two types of single violent human actions (punching, kicking) performed twice by 20 different individuals in different episodes: different background and lighting environments, both indoors and outdoors, and different angles indoors, outdoors, and outdoors. There are 273 videos in the database right now, 90 of which are boxing films, 90 of which are kicking videos, and 93 of which are non-violent videos. During training, each 0.1 s long frame was extracted from the images, yielding 40,423 frames.

Furthermore, every two seconds of video (equal to 15 frames) was tallied as a clip, yielding 2,694 total video clips. The entire dataset was split into training and test sets using an 80:20 ratio. The dataset is cross validated using the hold-out cross validation scheme with a percentage of 70:30%, the selection of 70% training data and 30% validation data is made randomly. All the videos are recorded with a resolution of 1920×1080 pixels using a *Canon XA15*. Each frame of the video has been given a name. The classifier is used to optimize the parameters of each operation after it has been trained on the training set. The test collection yields the granted recognition results.

4.4 Effectiveness of the models and comparisons

Several experiments were carried out to assess the efficacy of the proposed model about the dataset supplied in this section. Experiments on extracting deep features were carried out using the Caffe toolkit. To test the models' outputs, we evaluated the classification accuracy and loss of the 3DCNN and CNN + LSTM models with the proposed model on three different datasets, including standard datasets such as KTH, Weizmann, and Custom datasets. Table 1 demonstrates the classification accuracy and loss of 3DCNN, CNN-LSTM, and the proposed model on three independent datasets

Compared to our dataset and Weizmann, all three approaches perform well on the KTH dataset, as shown in the table. Because our dataset includes motions performed at various angles in various scenarios involving light shifts, it becomes more intricate, whereas the Weizmann dataset has multiple visually comparable activities. Based on the findings from these datasets, our proposed approach will effectively identify and localize violent and non-violent events. It was shown that for every 15 consecutive frames, including changes in key point coordinates and changes in the mask produced improves isolation and computation performance. This process specifies that highlighting particular motion patterns in a very short period is needed to achieve high precision.

4.5 Training effectiveness with time

On the KTH, Weizmann, and Custom datasets, Figures 5-11 display model training plots, the 3D-CNN, CNN + LSTM techniques' validation accuracy and loss graphs, respectively. Figure 7 shows the validity accuracy and lack of the proposed version over time using datasets from KTH, Weizmann, and our own. We will see how the performance of the modern-day version may be more desirable via way of means of the usage of the proposed version. The proposed version achieves 93.4 percentage accuracy at the KTH dataset, with a validation lack of 0.23 and an easy mastering charge of 0.005, which is better than the modern-day country of the art. The derived version becomes additionally evaluated on its dataset, with an accuracy of 86.5 percentage and a validation lack of 0.37. All techniques have commenced converging at the KTH dataset.

Figure 4. A sample frame from the collected dataset, which includes a variety of behaviors and circumstances

Table 1. Compares the accuracy and error of the suggested methods on the 3DCNN, CNN-LSTM, Weizmann, KTH, and Custom Datasets

Dataset	Weizman		KTH		Custom Dataset	
Method	**Accuracy**	**Loss**	**Accuracy**	**Loss**	**Accuracy**	**Loss**
3DCNN[42]	65.7%	0.81	82.2%	0.53	73.6%	0.60
CNN-LSTM[32]	71.3%	0.57	89.0%	0.44	84.1%	0.52
Proposed Method	**77.4%**	**0.51**	**95.7%**	**0.21**	**88.2%**	**0.34**

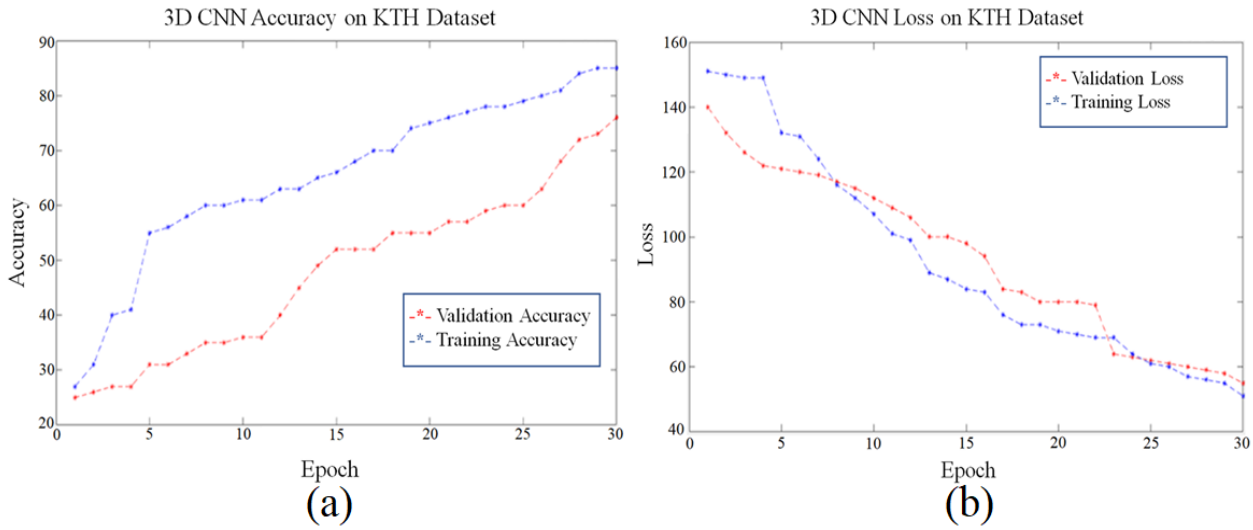


Figure 5. (a) Training and validation accuracy achieved on KTH dataset using the 3D-CNN model, (b) Training and validation loss achieved on KTH dataset using the 3D-CNN model

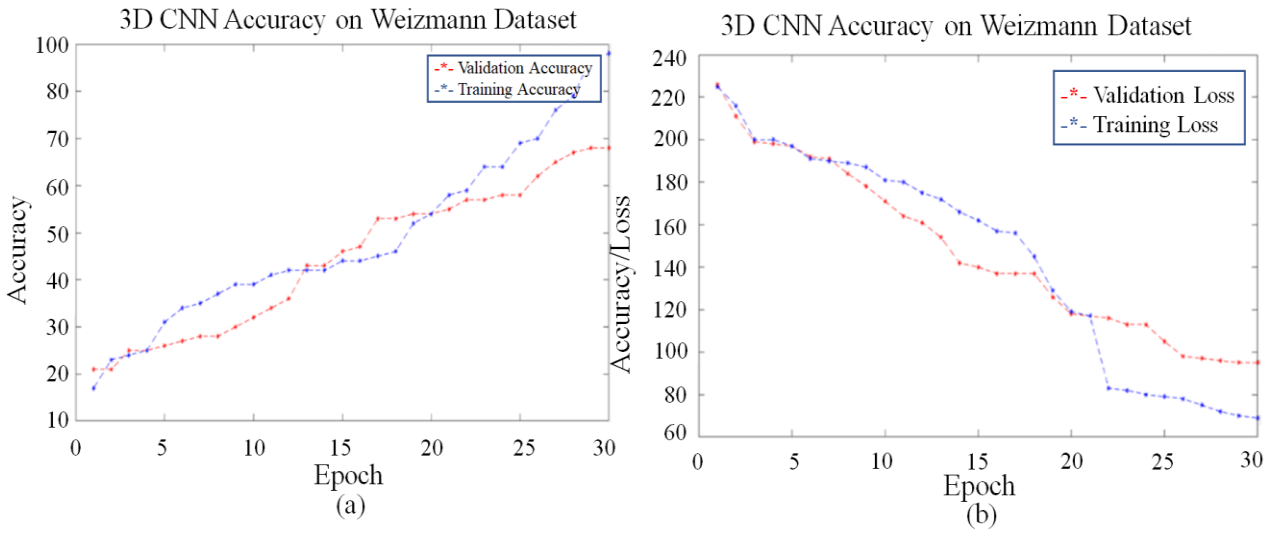


Figure 6. (a) Training and validation accuracy achieved on Weizmann dataset using the 3D-CNN model, (b) Training and validation loss achieved on Weizmann dataset using the 3D-CNN model

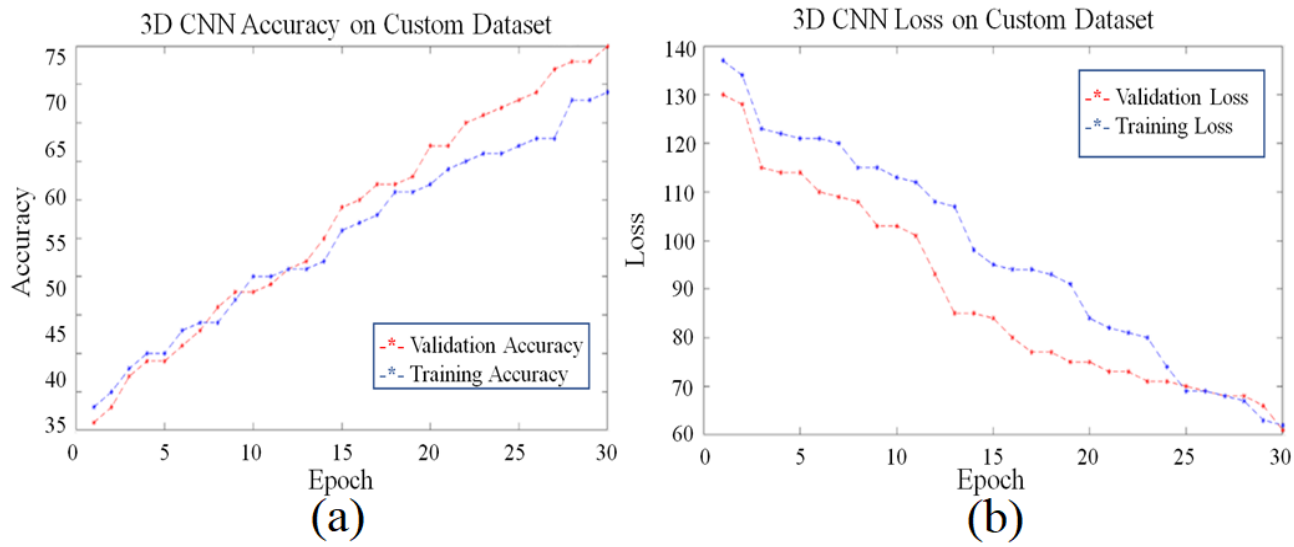


Figure 7. (a) Training and validation accuracy achieved on Custom dataset using the 3D-CNN model, (b) Training and validation loss achieved on Custom dataset using the 3D-CNN model

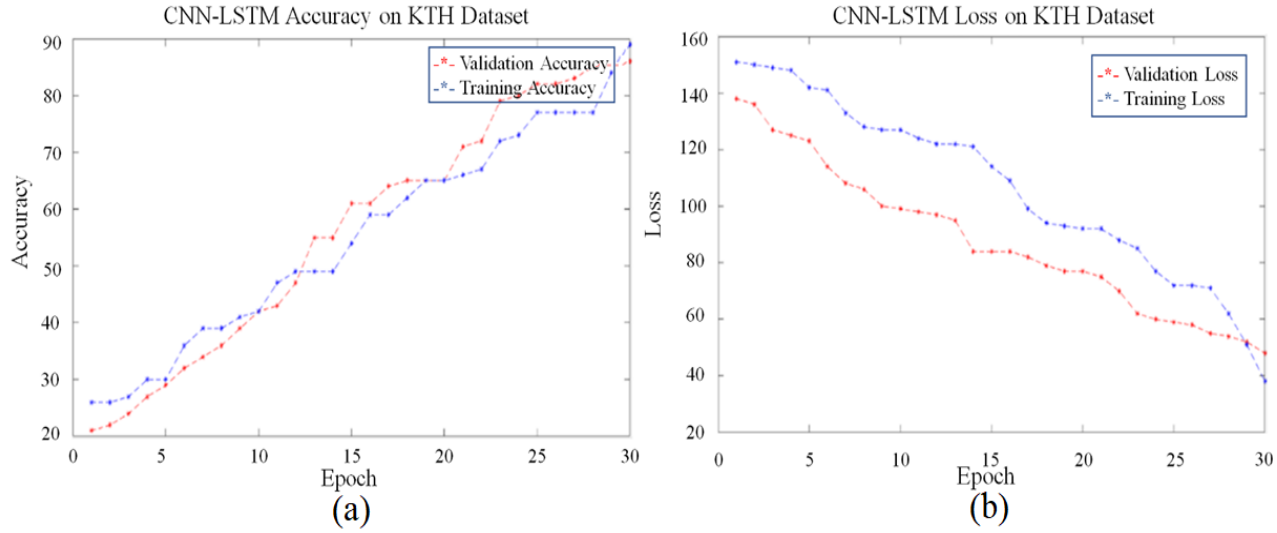


Figure 8. (a) Training and validation accuracy achieved on KTH dataset using the CNN-LSTM model, (b) Training and validation loss achieved on KTH dataset using the CNN-LSTM model

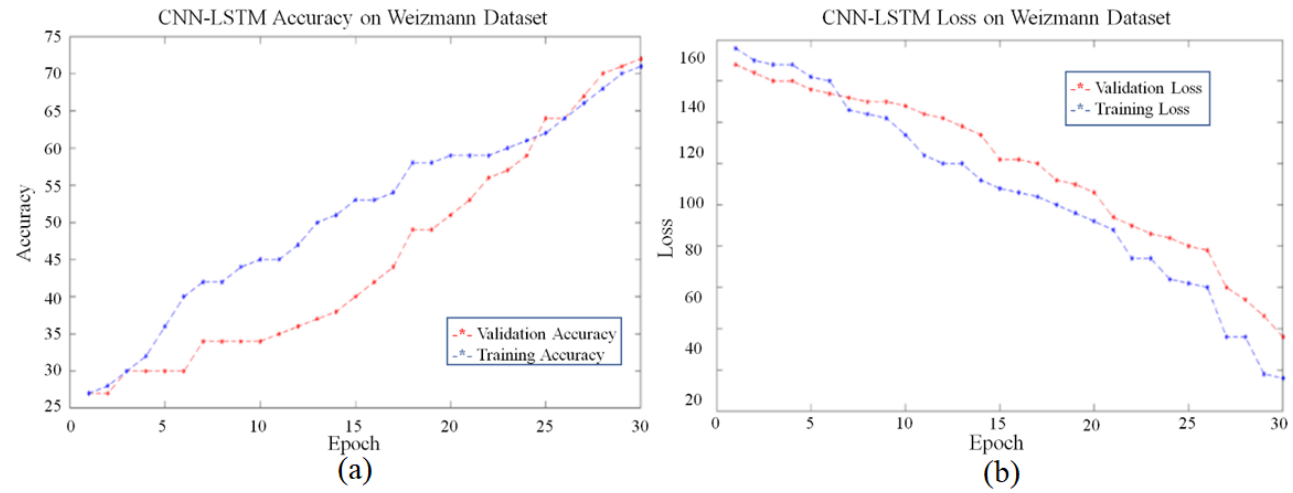


Figure 9. (a) Training and validation accuracy achieved on Weizmann dataset using the CNN-LSTM model, (b) Training and validation loss achieved on Weizmann dataset using the CNN-LSTM model

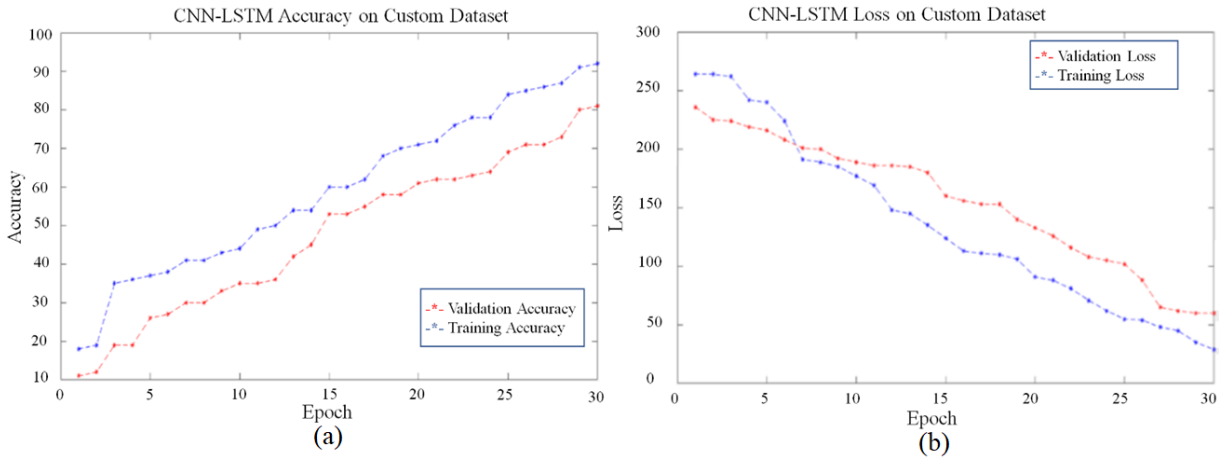


Figure 10. (a) Training and validation accuracy achieved on Custom dataset using the CNN-LSTM model, (b) Training and validation loss achieved on Custom dataset using the CNN-LSTM model

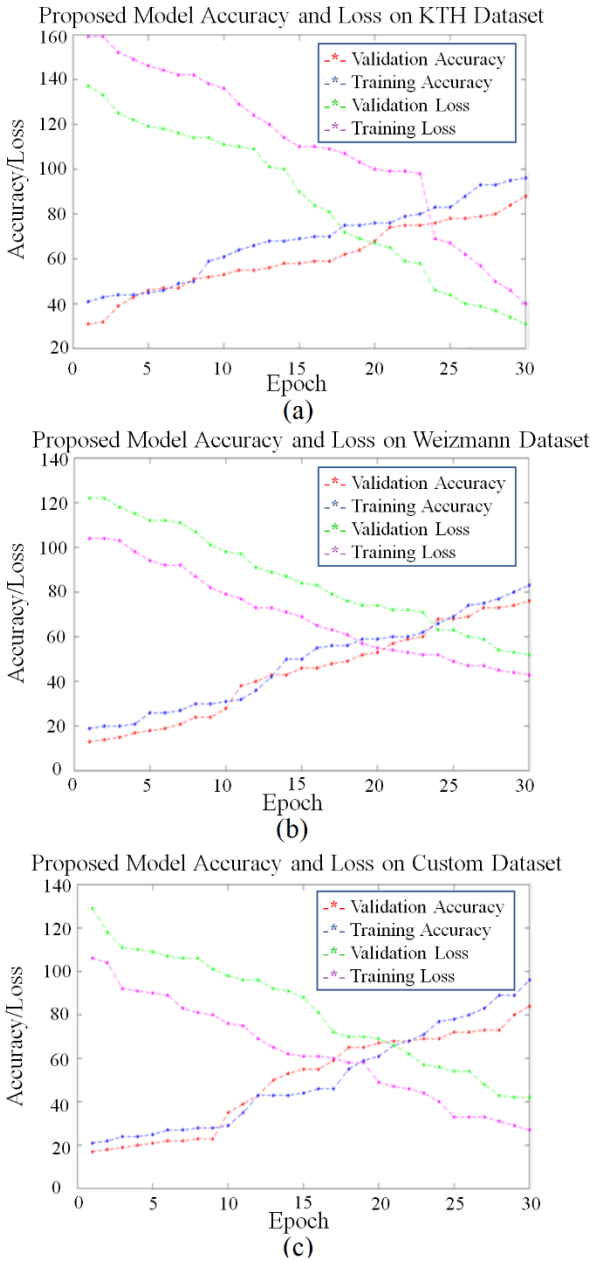


Figure 11. Proposed model training and validation accuracies along with their losses achieved on KTH (a), Weizmann (b), and Custom Datasets (c)

4.6 The projected favorable and negative review scores

Figure 12 shows the confusion matrices of the generated algorithms on the three datasets. The KTH dataset's confusion matrices show that the suggested method correctly differentiates groups with relatively similar stances. Because some action groups (such as running and walking) differ primarily in their temporal properties, the classification must consider this. There are activities in the Weizmann dataset that have identical movements but no space (i.e., just leg movements) and face temporal classification challenges, such as skipping and running. In the proposed dataset, nonviolent practices were correctly classified, but boxing and kicking were incorrectly classified due to inaccurate identification of key points in the back angle. The result of the suggested model is shown in Figure 13. These generated films were merely rebuilt from the generated frames based on the expected labels for display purposes.

On the developed custom dataset, Figure 10 shows the accuracy and loss achieved by the proposed model. On KTH and Weizmann, we also show the proposed model's validation accuracy and loss. As can be seen, performance is improved by using the proposed model rather than the existing model. The proposed model achieves the highest accuracy of 82.2 percent on a custom dataset, with a validation loss of 0.34 and a basic learning rate of 0.005, which is higher than state-of-the-art methods.

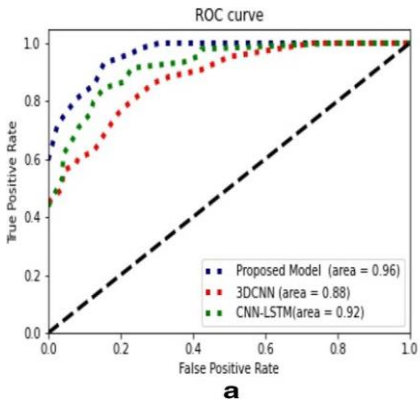
4.7 AUC (Area under the Curve) and ROC (Receiver Operator Characteristic) curves

An AUC-ROC (receiver operator characteristic) curve may reveal the output of a gadget getting to know the version classifier. The ROC plots the proper tremendous rate (TPR) vs the false positive rate (FPR) at diverse levels (FPR). It is thought that an excessive TPR will correctly classify a tremendous classification. The AUC (location below the curve) is a ROC curve description representing the classifier's potential to categorize. The AUC of the proposed version's ROC curve is honestly more than the AUC of the 3DCNN and CNNLSTM ROC curves, as illustrated in Figure 13 (a).

We have also applied a v/s all technique to multi-class classification using binary AUC-ROC curves. Figure 13 (b) shows the multi-category ROC curves for the proposed model. As can be shown, the nonviolent category has an AUC of 1,

meaning that it is easily observable compared to other categories. Both boxing and kicking have an AUC of greater than 0.8, meaning that the separability metric is equivalent in both sports (Figure 14). Consequently, the proposed model outperforms other models in the dataset when it comes to classifying positive categories.

Figure 12. The proposed model's confusion matrix, based on datasets from KTH, Weizmann, and our own



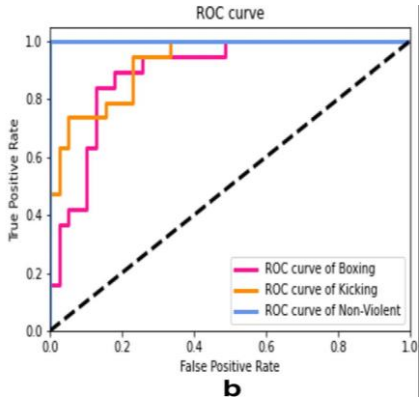


Figure 13. The area of the receiver operating characteristic curve

Figure 14. Depict how the proposed model correctly and incorrectly classifies various violent actions

5. CONCLUSIONS

This study investigates the significance of a single-person aggressive behaviour detection system that makes use of a selective pre-training model to detect hostile behaviour in order to detect hostile behaviour in the workplace. Increased accuracy of existing cutting-edge object detection systems can be leveraged to achieve single-person detection by improving their precision. It is integrated with an LSTM technique to mine the Spatio-temporal structure employed to account for the performed behaviour to comprehend it better and improve its accuracy. We test the effectiveness of multiple algorithms for recognizing violent behaviour in videos using our own custom dataset, which we have created. Following the outcomes of the experiments, it was discovered that the methodology has good performance and can give results comparable over a wide range of datasets. This is encouraging. While the theory's computational efficiency and adaptability are its most major advantages, it may also be applied to increase the accuracy of a wide range of deep learning-based violent scene identification applications. The proposed methodology may be improved in the future in order to handle more complex data, such as numerous individuals working with a large number of occluded areas, and it is feasible that this will be accomplished.

REFERENCES

[1] Cardenas, A.A., Amin, S., Sastry, S. (2008). Secure control: Towards survivable cyber-physical systems. In 2008 The 28th International Conference on Distributed

Computing Systems Workshops, Beijing, China, pp. 495-500. https://doi.org/10.1109/ICDCS.Workshops.2008.40
[2] Ghazal, S., Khan, U.S., Saleem, M.M., Rashid, N., Iqbal, J. (2019). Human activity recognition using 2D skeleton data and supervised machine learning. IET Image Processing, 13(13): 2572-2578.
[3] Ding, W., Liu, K., Belyaev, E., Cheng, F. (2017). PT US CR. Pattern Recognition.
[4] Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X. (2019). Fast and robust multi-person 3D pose estimation from multiple views. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7792-7801.
[5] Wang, X., Yang, L.T., Song, L., Wang, H., Ren, L., Deen, M.J. (2020). A tensor-based multiattributes visual feature recognition method for industrial intelligence. IEEE Transactions on Industrial Informatics, 17(3): 2231-2241. https://doi.org/10.1109/TII.2020.2999901
[6] Sandifort, M.L., Liu, J., Nishimura, S., Hürst, W. (2018). An entropy model for loiterer retrieval across multiple surveillance cameras. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pp. 309-317. https://doi.org/10.1145/3206025.3206049
[7] Villán, A.F. (2017). Facial attributes recognition using computer vision to detect drowsiness and distraction in drivers. ELCVIA: Electronic Letters on Computer Vision and Image Analysis, 16(2): 25-28. https://raco.cat/index.php/ELCVIA/article/view/336172.
[8] Chen, H., Ye, S., Nedzvedz, O.V., Ablameyko, S.V. (2018). Application of integral optical flow for determining crowd movement from video images obtained using video surveillance systems. Journal of Applied Spectroscopy, 85(1): 126-133.
[9] Liu, J., Qiu, L., Gao, E. (2016). Human abnormal behavior detection based on region optical floe energy. In 6th international conference on Electronic, Mechanical, Information and Management, No. 1050-1057.
[10] Aktı, Ş., Tataroğlu, G.A., Ekenel, H.K. (2019). Vision-based fight detection from surveillance cameras. In 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, pp. 1-6. https://doi.org/10.1109/IPTA.2019.8936070
[11] Ren, J., Lin, C., Liu, Q., Obaidat, M.S., Wu, G., Tan, G. (2018). Broadcast tree construction framework in tactile internet via dynamic algorithm. Journal of Systems and Software, 136: 59-73. https://doi.org/10.1016/j.jss.2017.11.020
[12] Huang, J.F., Chen, S.L. (2014). Detection of violent crowd behavior based on statistical characteristics of the optical flow. In 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Xiamen, China, pp. 565-569. https://doi.org/10.1109/FSKD.2014.6980896
[13] Gao, Y., Liu, H., Sun, X., Wang, C., Liu, Y. (2016). Violence detection using oriented violent flows *,**. Image and Vision Computing, 48: 37-41. https://doi.org/10.1016/j.imavis.2016.01.006
[14] Yang, Z., Zhang, T., Yang, J., Wu, Q., Bai, L., Yao, L. (2013). Violence detection based on histogram of optical flow orientation. In Sixth International Conference on Machine Vision (ICMV 2013), Vol. 9067, p. 906718. https://doi.org/10.1117/12.2051390
[15] Lejmi, W., Khalifa, A.B., Mahjoub, M.A. (2020). A novel spatio-temporal violence classification framework based on material derivative and LSTM neural network. Traitement Du Signal, 37(5): 687-701. https://doi.org/10.18280/ts.370501
[16] Mahmoodi, J., Salajeghe, A. (2019). A classification method based on optical flow for violence detection. Expert Systems with Applications, 127: 121-127. https://doi.org/10.1016/j.eswa.2019.02.032
[17] Xu, Q., See, J., Lin, W. (2019). Localization guided fight action detection in surveillance videos. In 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, pp. 568-573. https://doi.org/10.1109/ICME.2019.00104
[18] 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
[19] Aggarwal, J.K., Cai, Q. (1999). Human motion analysis: A review. Computer Vision and Image Understanding, 73(3): 428-440. https://doi.org/10.1006/cviu.1998.0744
[20] Badi, H. (2016). A survey on recent vision-based gesture recognition. Intelligent Industrial Systems, 2(2): 179-191. https://doi.org/10.1007/s40903-016-0046-9
[21] Bobick, A.F. (1997). Movement, activity and action: the role of knowledge in the perception of motion. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 352(1358): 1257-1265. https://doi.org/10.1098/rstb.1997.0108
[22] Gavrila, D.M. (1999). The visual analysis of human movement: A survey. Computer Vision and Image Understanding, 73(1): 82-98. https://doi.org/10.1006/cviu.1998.0716
[23] Krüger, V., Kragic, D., Ude, A., Geib, C. (2007). The meaning of action: A review on action recognition and mapping. Advanced Robotics, 21(13): 1473-1501. https://doi.org/10.1163/156855307782148578
[24] Moeslund, T.B., Hilton, A., Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 104(2-3): 90-126. https://doi.org/10.1016/j.cviu.2006.08.002
[25] Poppe, R. (2010). A survey on vision-based human action recognition. Image and Vision Computing, 28(6): 976-990. https://doi.org/10.1016/j.imavis.2009.11.014
[26] Naik, A.J., Gopalakrishna, M.T. (2017). Violence detection in surveillancevideo-a survey. International Journal of Latest Research in Engineering and Technology (IJLRET), 11-17.
[27] Lertniphonphan, K., Aramvith, S., Chalidabhongse, T.H. (2011). Human action recognition using direction histograms of optical flow. In 2011 11th International Symposium on Communications & Information Technologies (ISCIT), pp. 574-579. https://doi.org/10.1109/ISCIT.2011.6089701
[28] Wang, H., Schmid, C. (2013). Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, pp. 3551-3558.
[29] Hassner, T., Itcher, Y., Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence,

RI, USA, pp. 1-6. https://doi.org/10.1109/CVPRW.2012.6239348
[30] Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A. (2011). Sequential deep learning for human action recognition. In International Workshop on Human Behavior Understanding, pp. 29-39. https://doi.org/10.1007/978-3-642-25446-8_4
[31] Dong, Z., Qin, J., Wang, Y. (2016). Multi-stream deep networks for person to person violence detection in videos. In Chinese Conference on Pattern Recognition, pp. 517-531. https://doi.org/10.1007/978-981-10-3002-4_43
[32] Sudhakaran, S., Lanz, O. (2017). Learning to detect violent videos using convolutional long short-term memory. 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-6. https://doi.org/10.1109/AVSS.2017.8078468
[33] Zhou, P., Ding, Q., Luo, H., Hou, X. (2017). Violent interaction detection in video based on deep learning. In Journal of Physics: Conference Series, 844(1): 012044. https://doi.org/10.1088/1742-6596/844/1/012044/
[34] Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
[35] Yu, J., Song, W., Zhou, G., Hou, J.J. (2019). Violent scene detection algorithm based on kernel extreme learning machine and three-dimensional histograms of gradient orientation. Multimedia Tools and Applications, 78(7): 8497-8512. https://doi.org/10.1007/s11042-018-6923-3
[36] Acar, E., Hopfgartner, F., Albayrak, S. (2013). Violence detection in hollywood movies by the fusion of visual and mid-level audio cues. In Proceedings of the 21st ACM International Conference on Multimedia, pp. 717-720. https://doi.org/10.1145/2502081.2502187
[37] Wang, L., Qiao, Y., Tang, X. (2015). Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4305-4314. https://doi.org/10.1109/CVPR.2015.7299059
[38] Simonyan, K., Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199. https://arxiv.org/abs/1406.2199.
[39] Wang, L., Xiong, Y., Wang, Z., Qiao, Y. (2015). Towards good practices for very deep two-stream convnets. arXiv preprint arXiv:1507.02159. https://arxiv.org/abs/1507.02159.
[40] Zhu, Y., Lan, Z., Newsam, S., Hauptmann, A. (2018). Hidden two-stream convolutional networks for action recognition. In Asian Conference on Computer Vision, pp. 363-378. https://doi.org/10.1007/978-3-030-20893-6_23
[41] Wang, X., Girshick, R., Gupta, A., He, K. (2018). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794-7803.
[42] 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
[43] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6450-6459.
[44] Zhao, Y., Xiong, Y., Lin, D. (2018). Trajectory convolution for action recognition. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 2208-2219.
[45] Zolfaghari, M., Singh, K., Brox, T. (2018). Eco: Efficient convolutional network for online video understanding. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 695-712.
[46] 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
[47] Fu, E.Y., Leong, H.V., Ngai, G., Chan, S. (2015). Automatic fight detection based on motion analysis. In 2015 IEEE International Symposium on Multimedia (ISM), Miami, FL, USA, pp. 57-60. https://doi.org/10.1109/ISM.2015.98
[48] Fu, E.Y., Leong, H.V., Ngai, G., Chan, S.C. (2017). Automatic fight detection in surveillance videos. International Journal of Pervasive Computing and Communications, 13(2): 130-156. https://doi.org/10.1108/IJPCC-02-2017-0018
[49] Lloyd, K., Rosin, P. L., Marshall, D., Moore, S.C. (2017). Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures. Machine Vision and Applications, 28(3-4): 361-371. https://doi.org/10.1007/s00138-017-0830-x
[50] Zhou, P., Ding, Q., Luo, H., Hou, X. (2018). Violence detection in surveillance video using low-level features. PLoS One, 13(10): e0203668. https://doi.org/10.1371/journal.pone.0203668
[51] Fei-Fei, L., Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, Vol. 2, pp. 524-531. https://doi.org/10.1109/CVPR.2005.16
[52] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In European Conference on Computer Vision, pp. 20-36. https://doi.org/10.1007/978-3-319-46484-8_2
[53] 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. https://doi.org/10.1088/1742-6596/844/1/012044
[54] Xia, Q., Zhang, P., Wang, J., Tian, M., Fei, C. (2018). Real time violence detection based on deep spatio-temporal features. In Chinese Conference on Biometric Recognition, pp. 157-165. https://doi.org/10.1007/978-3-319-97909-0_17
[55] Fu, E.Y., Huang, M.X., Leong, H.V., Ngai, G. (2018). Cross-species learning: A low-cost approach to learning human fight from animal fight. In Proceedings of the 26th ACM international conference on Multimedia, pp. 320-327. https://doi.org/10.1145/3240508.3240710
[56] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2020). Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 386-397. https://doi.org/10.1109/TPAMI.2018.2844175

[57] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404: 132306. https://doi.org/10.1016/j.physd.2019.132306

[58] 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.

[59] Toshev, A., Szegedy, C. (2014). DeepPose: Human pose estimation via deep neural networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1653-1660. https://doi.org/10.1109/CVPR.2014.214

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

[61] Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R. (2005). Actions as space-time shapes. In Tenth IEEE International Conference on Computer Vision (ICCV'05), pp. 1395-1402. https://doi.org/10.1109/ICCV.2005.28