An Automatic Violence Detection System using Deep Convolutional Network

Muhsin Khan*, Dr. Muhammad Qasim Khan*, Dr. Fazal Malik*, and Muhammad Suliman*

*Department of Computer Science, Iqra National University
Peshawar, KPK, Pakistan

Abstract- In the past few years, due to its highly beneficial applications, the exploration of violence detection has become an increasingly pertinent subject in the realm of computer vision, leading to the development of numerous proposed solutions by researchers. This work presents a Transfer learning approach for violent scene detection, with redundant frames removal from video frames for classification into violent and non-violent categories. Two benchmark datasets are used for validating the proposed technique for abnormal event's detection; inception v3, a state-of-the-art Convolutional Neural Network (CNN) composed of 316 layers is used. The model is pre-trained with a benchmark ImageNet dataset' which includes 1000 classes. In the first phase, redundant video frames from the dataset are removed using an unsupervised learning approach. To detect whether the two frames are similar or not; features from both frames are extracted, and Euclidean distance is computed. If the distance between the frames is small, then the image is considered redundant and is dropped from the dataset. In the second phase, transfer learning is performed to detect violent frames. In transfer learning, the CNN model is updated by removing the last three layers and adding a fresh layer to the network. Hold out cross-validation technique is applied to partition the dataset into train and test sets, data is randomly divided into 70- 30 percent where 70% data is selected for training and 30% is selected for testing. The redundant frame removal method helps in efficient training and validation by reducing the size of the dataset. The proposed model performed much better than state-of-the-art approaches and achieved an accuracy of 78.61% for the real-life violence dataset and 73.86% for the Hockey Fight Dataset. The accuracy can be further increased by choosing the best optimal threshold value for key frame selection and tuning the epoch and other hyper parameters.

Index Terms- Computer Vision, Automatic Violence Detection System, Convolution-al Neural Networks (CNN), Transfer Learning, Action Recognition.

I. INTRODUCTION

Addressing violence detection is a prevalent concern in recent studies focusing on human-to-human action recognition, particularly crucial in surveillance videos. The significance of detecting violence in videos lies in enhancing public safety and preventing crimes within smart cities. For effective implementation in real-world surveillance scenarios, the recognition of violence must be both swift and accurate to facilitate timely intervention. Typically, CCTV footage is reviewed hours or days after an incident, serving its purpose in legal proceedings but seldom contributing to real-time crime prevention or response. The responsibility of monitoring extensive CCTV footage predominantly rests on a limited number of security personnel. Due to fatigue, worker boredom, and breaks in observation, human oversight becomes unreliable [1]. Enterprises, governmental bodies, and law enforcement agencies have been driven to employ extensive surveillance systems to identify hazardous environments and adeptly respond to violent interactions, given the surge in criminal activities. The intricate nature and requisite consideration of specific features make automatic violence detection an ongoing and formidable challenge for the majority of security systems [1].

Apart from CCTV footage, Violence Detection Systems can be used for content analysis over the Internet. In the modern world, it is very important to filter media data and has many useful applications. A massive amount of data is uploaded to the internet for cloud-based storage and processing. Media can be in the form of video, audio, animation, or in some other form. This data can be labeled and grouped manually by experts in the field. The labeled data can be used to train machine learning models that will be able to automatically detect various suspected anomalies in digital media content. Automatic video analysis systems have a lot of different useful applications such as parental control, rating and ranking of a video, filtering data of interest for the user, and avoiding uploading inappropriate videos to the internet. Action recognition technology is an interesting and challenging research area in machine learning, and the task can be broadly divided into three classes:

1) Action recognition in the wild where the actors are performing some action in a cluttered background, all background objects are also included in recognition of a particular action.
2) Skeleton-based, depth images and disparity information from the Microsoft Kinect device are collected, and the skeleton image of a human is used to recognize human actions.
3) Human segmentation is another method, in which first the human image from the video frames is segmented, and is then passed to an action recognition model. The model then classifies the action of the human and assigns a class label to the video frame.

Human action recognition has also been performed using embedded sensors. These are real-time special-purpose computer systems designed to recognize specific human actions. But such a

complicated task cannot be achieved efficiently and accurately due to the low processing speed of such sensors.

Training of a machine learning model is done with many classes, where each class usually has thousands of video frames. The system must be trained on a Graphics Processing Unit (GPU) based system for efficient training; once the training is complete, the model must be validated using a validation set; if the model achieves satisfactory results, it is then deployed to the embedded system. To identify human actions in real time, a camera must be connected to the embedded system to capture an image of a human and validate the video frame for human action recognition.

Video processing using deep learning can help develop quick analysis systems to monitor patient's health conditions, monitor prisoners' actions, public safety, and surveillance systems. Traditional machine learning models work very well for a few actions, but with a large amount of data and more classes in a dataset, it is very difficult if not impossible to accomplish this task with state-of-the-art accuracy [1]. For the recognition of hundreds of actions, a deep learning-based action recognition system must be developed. Convolutional Neural Networks (CNN) can recognize thousands of classes efficiently and accurately. But creating a state-of-the-art CNN model is not an easy task; it requires several machine learning, and programming skills. In video surveillance scenarios where violence detection is applicable, extends to various contexts, encompassing both indoor and outdoor settings, such as buildings, traffic areas, and even on police body cameras. Such scenarios are illustrated in the in Figure 1 adopted from [2], pictures are taken from different sources, and we combine them into one picture to show different scenarios of violence.

Fig. 1. Various situations where real-time violence detection is relevant, along with corresponding scenes of violence that need to be identified: A) Video surveillance within indoor spaces B) Video surveillance in traffic environments C) Utilization of police body cameras. [2].

A. *Factors Affecting VDS Performance*

One challenge in action recognition lies in the real-time classification of human activity immediately following the occurrence of the action. This challenge becomes more pronounced when dealing with surveillance videos, influenced by several significant factors. These factors encompass the reduced quality of surveillance footage, inadequate or absent lighting, and the lack of contextual information that could facilitate the detection and classification of violent versus non-violent actions. Consequently, the difficulty in action detection is heightened by poor-quality surveillance videos and the absence of contextual information. The complexity further intensifies when attempting to differentiate between violence and non-violence in real-time due to the diverse factors involved in developing a system for real-time violence detection. One specific challenge within video recognition based systems is the need to filter out surplus frames, considering that most videos operate at a frame rate of 30 frames per second. This frame rate might be high or low, depending on the use of the different video encoding standards used by video cameras. But even with a typical 30 FPS, videos contain redundant information. Simple ways of removing such redundant information have been proposed, such as selecting every 5th frame, or selection of odd or even frames. But these methods have not been proven to be very effective [7].

Beyond the challenges posed by poor video quality, violence has the potential to occur in diverse settings and at any time of the day. Consequently, the solution must be resilient enough to effectively detect violence under varying conditions. Another issue is that the majority of the studies regarding VDS are done with hand-crafted features or conventional approaches. Such features are non-effective in terms of time, and a huge amount of similar redundant features. The redundant features not only decrease the accuracy but also affect algorithm effectiveness by generating overfitting problems when training the classifier.

Shape, color, and other low-level features may also be not measured robustly for action recognition. Many conventional features lack invariance, and the system's accuracy is influenced by certain image operations, such as changes in brightness, scale, rotation, and conversion. Additionally, some approaches prove ineffective in classification tasks. In our proposed work outlined in this paper, we have successfully addressed both of these challenges.

B. *Deep Learning and CNN*

Deep learning, a sub-discipline of machine learning employing neural network architecture, involves multiple computing layers to capture non-linear representations of data at a sophisticated level of complexity. Deep Learning and Machine Learning have been used in diverse fields such as medical [17] [18], computer networks [21], student performance prediction [20], Google stock prediction [22] [23], and software engineering [19]. CNN, a specific form of deep learning inspired by the human visual cortex system, is specifically designed for the analysis of video and image data. Recurrent Neural Networks (RNNs) are commonly used in computer vision, bio-informatics, and Natural Language Processing (NLP) for supervised learning problems, achieving greater accuracy with improved machine performance [4]. These models' key benefits are the ability to automatically create features from the data, letting pattern recognition systems depend less on heuristics that are manually created. CNNs are feed-forward neural networks where the flow is in a straight and one direction, while the RNN models can transfer data to a network's next as well as previous node. Millions of neurons

exist in a human neural system linked to send and receive signals, the nodes in CNN behave as a neuron, where the vertices that join the nodes work as dendrites. The CNN works by stacking several layers of neurons. Each layer is responsible for gathering and feeding information about a small region; that's why it is called a Feed Forward Neural Network. Finally, the output is produced through a classification layer, which, during training, slightly adjusts the weights of the functions in the last layers to minimize classification errors. This process is referred to as back-propagation, as it involves making adjustments instead of retraining the system from the beginning with these new modifications [3] [4], it pushes the error back into the network. Since the 1990s, CNN models achieved greater accuracy in recognition of handwriting, and facial recognition among other classification problems. Convolutional neural networks are special models that were developed for image and video data, Alex-net was the first model that won the ImageNet challenge [46]. ImageNet is a benchmark dataset used to validate a CNN model and has 1000 classes. CNN belongs to the deep learning class, in deep learning classification models a lot of data is required to train the model. CNN, RNN, and other deep learning models don't perform well on a small dataset. The CNN model has two parts, and each part is responsible for a particular task. Like other neural network models, CNN also has an input layer, some hidden layers, and a final classification output layer. The CNN model that is available in MATLAB is pre-trained on the ImageNet dataset. CNN training calculates the optimal weights for the extracted image features. In transfer learning, the old data such as features, weights, labels, and bias values are removed from the network by removing the last three layers. Three new layers, namely fully connected, softmax, and classification layers, are introduced into the inception-v3 models. Other information such as parameters and hyper parameters are tuned accordingly. An optimization algorithm is used to get the optimal learning rate for the CNN model. By calculating the most optimal weights for the image features, choosing optimal learning is important. Given below are some standard layers that are used in almost all CNN models: • Convolutional Layer • Pooling layer • Normalization • ReLU • Sigmoid • Fully Connected • Softmax • Classification etc.

C. *Foreground Detection*

In computer vision, the foreground detector is the basic kind of tracking method. The foreground detector algorithm consists of these steps: background image selection as a reference image; threshold value initialization for subtraction operations; subtraction of the image from the background image; finally extraction of the apparent foreground areas. There may be frames composed of redundant information in a video series with a high FPS rate; the foreground detector technique may be useful for extracting redundant content from the video. The algorithm presented in [5] selects an image as the foreground and isolates the foreground images to serve as the reference image.

D. *Optical Flow*

Optical flow is a moving object's motion pattern in a video series. It is a robust tool for detecting boundaries and surfaces in a video file; recognition takes place due to the motion in the video scene [5]. In simple terms, optical flow is the moving objects' velocities, which are measured with the help of brightness patterns in an image. Video frames are used to measure moving objects' motion by discrete image displacements and instantaneous image velocities.

E. *Inception-v3*

Inception-v3 is a CNN model trained on more than a hundred thousand images on the Image-Net dataset. Image- Net is a benchmark dataset having 1000 different classes. Inception-v3 is a 316-layer pre-trained neural network model composed of two parts. A featuring extraction layers from 1 to 313 with a convolutional operation and the last three i.e., 314-316 are classification layers which include fully connected followed by soft max and a classification layer [6].

F. *Transfer learning*

To take a CNN model, already learned on any dataset, scratch off the last three layers, and remake or retrain the model with another dataset having different weights and classes is called transfer learning. For the sake of this research work, we removed the last three layers of the Inception-v3 model (i.e., 314 to 316) and have added new layers for training and recognition of violent and non-violent scenes. Computers need to generate patterns and follow or use them to develop significance from a series of images and understand various concepts. Retraining a learned model is to identify distinct higher-level features. As a result, accuracy is improved, and the training time is decreased. As training requires so much time and resources, it is a valuable technique [6].

II. LITERATURE REVIEW

Audio feature extraction, the utilization of MoSIFT descriptors for spatiotemporal analysis and the incorporation of optical flow motion vectors define state-of-the-art approaches in violence detection within surveillance videos. Despite achieving promising results with nearly 85% accuracy, current methods exhibit drawbacks such as low precision, high memory requirements, and substantial computational costs. These limitations make them unsuitable for real-world applications, especially in surveillance, where the demand is for both high accuracy and swift results.

Authors in [3] suggested a computer vision methodology that achieves a detection accuracy of up to 95% for specific datasets and delivers real-time results. This approach combines established computer vision and image processing techniques with advanced deep learning algorithms. Study on violence recognition in videos is rare, compare to other sorts of video analysis tasks. An action recognition system via dense trajectories and a descriptor is introduced in [10], this method computes motion boundaries based on optical flow information to identify a trajectory for the motion. Further enhancements have been made by incorporating adjustments for camera motion and employing Fisher Vector encoding in [11], emerging as state-of-the-art at its time, this approach involves aggregating multiple traditional descriptors into a bag of features. The classifier then determines how these descriptors interact to define each type of action, with the ultimate goal of automatically extracting features from videos. Another method suggested a 3D-based CNN for action detection, utilizing several consecutive video frames as input for the network, capturing motion information. They achieved comparable results to using dense trajectories but with frames of significantly lower quality, approximately one-fourth of the original resolution. In the exact

work of action recognition, most of the work is done on low-level features. The typical methodology consists of the feature extraction around interest points, for example, gradients, optical flows, intensities, and some additional local features [12]. Some previous research work has used threshold values for audio and visual features.

They measured the amplitude and energy for the auditory features of the acoustic signal, similarly, quick variations in the whole entropy have also been used. Which is based on the calculation of dynamic activity on visual features to recognize fast activities such as pixel color thresholds for blood detection [13]. Other methods proposed an acoustic approach for sensing basic audio events, for instance blasts shootings, car breakings, engines, etc. They employ Hidden Markov Models (HMM) training to identify target sound events, subsequently modeling associations among various events using Gaussian mixture models to extract more intricate semantic context [14]. Those previous approaches, though, rely on precise events and observed each one independently. One technique that simplifies and attempts to classify and recognize human actions is proposed by [11]. They eschewed the Bag of Visual Words (BoVW) methodology, opting instead for low-level features like Space-Time Interest Points (STIP) and Motion SIFT (MoSIFT), a variant of the SIFT image descriptor tailored for non-audio contexts. This involved introducing a histogram of optical flows that represents local motion. For each video, these features were employed to generate a bag of words, subsequently classified using Support Vector Machine (SVM). Another approach based on the Bag of Visual Words, utilizing local spatiotemporal features, was employed to classify video scenes as either violent or normal. Several STIP-detected descriptors were hard-coded to leverage spatiotemporal information, creating a bag of features for each frame. A linear support vector machine (SVM) was then trained to categorize the videos, yielding reasonably improved results for certain activities, with an accuracy reaching almost 77% [16]. Every approach underscores the importance of incorporating spatiotemporal features and motion in the detection of violence.

However, each study presents results on diverse datasets using distinct metrics. Additionally, the variability in theories regarding violence hinders a direct comparison with current methods. While there has been a great deal of work concerning action recognition, automatic human action analysis is still a growing field. Some research works making use of unique sounds that signify specific varieties of violence or use of some descriptors made for general actions and merging them in order to detect sudden changes in motion [24]. Some studies combine a number of descriptors in a bag of visual words (BoVW) method [16] [24]. While emphasizing the importance of utilizing spatiotemporal information to detect violence by combining audio features with motion from the video stream, they continue to rely on hand-crafted descriptors to construct a highly subjective model. Machine-learning procedures, particularly CNNs, have achieved outstanding results in image-related tasks and video classification [24]. Across various challenging datasets, these networks outperformed earlier approaches, including the BoVW scheme. Typically, these networks have been examined and refined within the action recognition domain [25]. Many architectures are used for combining the spatiotemporal information, each of which has its specific way of carrying the motion information, resulting in numerous features for the classification task. Recently, in the specific task of violence recognition these convolutional networks have been used, showing favorable results [26]. Generally, these approaches combine features taken out from the neural network with some hand-crafted features in the expectation to solve the matter without knowing its influences. These approaches are still far from the remarkable results in the classification task of images.

Numerous violence detection methods outlined in this section hinge on the detection of a point of interest to produce a vector of features encompassing pertinent regions within a frame. Additionally, they also necessitate the utilization of algorithms for motion tracking. While some of these works demonstrate good accuracy, they have trade-off in terms of efficiency, such as time complexity. For any violence detection method to be deemed suitable for real-time applications, such as real time video surveillance, it must possess the capability to detect crimes almost instantaneously [9].

III. METHODOLOGY

In this work, a novel framework for the detection of violent frames is proposed. The proposed framework has five phases, as shown in Figure 2. All of these phases are explained here in this section.

A. Datasets

In this work, we have chosen two different datasets to test the effectiveness of our proposed method for real time violence detection. The details of these datasets are briefly discussed next

1. Hockey Fight Dataset:

The hockey fight dataset contains 1000 videos taken from the National Hockey League (NHL), and a moving camera captures all the images. Half of them are classified as fighting (500 clips) and another half are labeled as non-fighting. Each clip includes approximately 40 frames with 360 x 288 of resolution [44]. Figure 3 shows the clips included in the Hockey Fight Dataset

Fig. 3. Hockey Fight Dataset Clips.

2. Real-Life Violence Situation Dataset:

The benchmark dataset for violence detection is used; it consists of 2000 videos, 1000 violent and 1000 normal scenes. All the videos are obtained from online video platforms; includes street fights among people. The videos are 1280 x 720 pixels of resolution, 24 bits RGB with frame rate of 25 FPS. The footage also contains a frequency of 44 kHz, with [45].

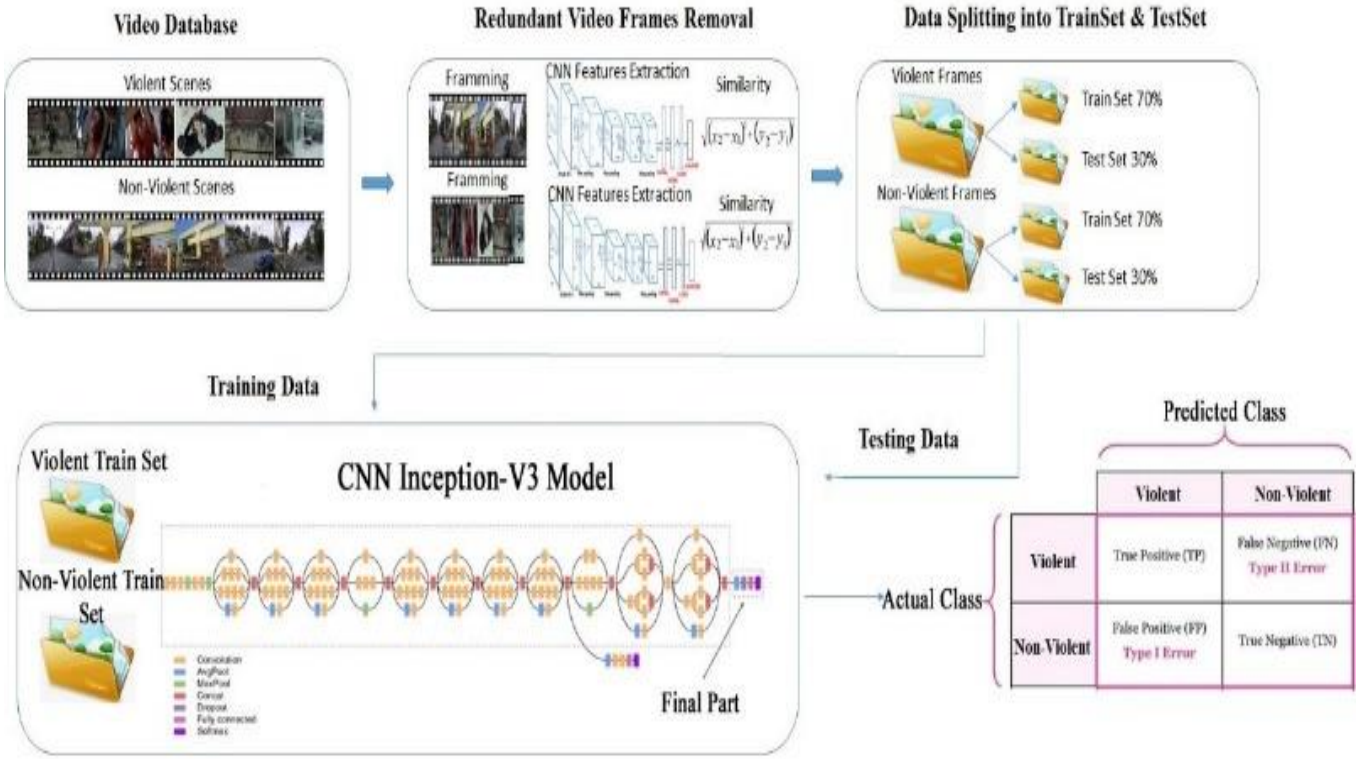

Fig. 2. Methodology of the Proposed Framework

In Figure 4 some of the clips which are included in real life violence situation are given.

Fig. 4. Real-Life Violence Situation Dataset Clips.

B. Redundant Video Frames Removal

This research endeavors to employ an efficient learning based method for extracting image features in computer vision, with the aim of developing an innovative approach for eliminating redundant frames in videos. Various feature detectors and descriptors are assessed for their suitability in recognizing Content-Based Feature Retrieval (CBFR) within a video. Conventional methods typically rely on the scale invariant feature transform (SIFT) for selecting interest points and extracting features from those points. In this study, we propose substituting SIFT features with the inception-v3 CNN feature extraction. Unsupervised and supervised classification methods are compared to showcase differences in performance.

The system's efficiency and accuracy are enhanced by employing a similarity measure technique between two video frames to quantify the degree of similarity. This measure indicates how much the two frames resemble each other, with closeness measured as the distance between dimensions describing object features. A minimal difference in similarities suggests greater frame/image identity, while a more considerable distance implies lesser similarity. The Euclidean distance, a standard measurement method, is utilized to determine the distance for selecting or discarding a frame from a video. The redundant video frames are identified and eliminated through the Content-Based Image Retrieval technique (CBIR), also known as image-based search. CBIR, a computer vision technology, utilizes a feature descriptor and a similarity measure technique. The videos from both classes are initially converted into frames to detect redundant video frames using a predefined threshold value. To extract features using the inception-v3 model, the images must be resized to match the model's input size, which requires RGB images of dimensions 299 x 299. Various state-of-the-art similarity measures, such as Euclidean distance, hamming, cosine similarity, etc., are employed in the process.

C. Primary Consideration of similarity

The Euclidean distance, also known as the Euclidean metric, is the ordinary straight-line distance between two points in Euclidean space. This distance metric transforms Euclidean space into a metric space, and the corresponding norm is called the Euclidean norm. In older literature, this metric is sometimes referred to as the Pythagorean metric. The L2 norm or L2 distance is a simplified term used interchangeably with the Euclidean norm. The distance between the L2 norm or L2 distance is measured. Figure 5 shows the step-by-step process of obtaining the non-redundant frames from a video. Two different video frames X, Y are similar if the distance between X, and Y is V=15.

Feature_1 = [13, 24, 35, 16, 57, 32, 53, 31, 25, 26, 37] (1)

$$Feature\_2 = [26, 12, 17, 5, 25, 16, 11, 22, 36, 21, 17] \quad (2)$$

$$V = Feature\_1 - Feature\_2 \quad (3)$$

$$V\_2 = V * V\,T \quad (4)$$

$$Distance = (V\_2) \quad (5)$$

Fig. 5. Redundant video frames removal using Siamese CNN model.

The second phase detects and removes similar video frames. In high FPS videos, most of the video frames consist of redundant information. So an unsupervised method is developed to drop the redundant video frames from the dataset. CNN based learning features with Euclidean distance are used to select optimal video frames.

D. Validation

The third phase is to split the dataset into train and test sets using hold-out cross-validation techniques. For big image datasets and large CNN models, the hold-out method performs the best when there is complex data. Other methods (such as k-fold and leave-one-out, etc.) makes it very time inefficient, mainly in cases where we need to ease the over-fitting problem.

Parameters Values	**Parameters Values**
E-Poch 50 and 100	E-Poch 50 and 100
Batch-Size 64	Batch-Size 64
Learning Rate 0.0001	Learning Rate 0.0001
Optimization SGDM	Optimization SGDM
Tool Boxes DIP, CV and DL	Tool Boxes DIP, CV and DL
Datasets Real-life Violence Situations	Datasets Real-life Violence Situations
and Hockey Fight	and Hockey Fight
Similarity Measure Euclidean Distance	Similarity Measure Euclidean Distance
Cross Validation 70/30 percent	Cross Validation 70/30 percent

TABLE I: METHODS AND PARAMETERS

E. Transfer Learning

The fourth phase is to perform transfer learning in the inception-v3 CNN model. And in the final phase, the performance of the proposed model is evaluated. Instead of handcrafted features, features are extracted through the inceptionv3 CNN model and then transfer learning is used for the classification task with the help of the last 3 layers.

F. Evaluation Methods

This section reports and discusses the results of our proposed scheme with the help of standard classification performance evaluation matrices commonly used in machine learning. These matrices are accuracy, recall, precision, and F-measure. For each dataset, different parameters are used; the goal is to identify optimal parameters for training the model. The selected parameters for classifying video frames into violent and non-violent are given in Table I. The accuracy, precision, recall, and F1-Score or F-Measure of our proposed violence detection system are calculated with the help of a confusion matrix whose format is shown in Figure 6. In this figure, TP stands for True Positive, FP means False Positive, FN stands for False Negative, and TN stands for True Negative.

	Class A	**Class B**
Class A	TP Rate	FP Rate
Class B	FN Rate	TN Rate

Fig. 6. Confusion Matrix for Performance evaluation.

The primary metric of significance is the overall accuracy, calculated by dividing the number of correctly identified images by the total number of images. Additionally, the top-1 error rate is provided, representing the error rate within the class with the lowest overall accuracy. It is customary to report this metric alongside overall accuracy [48] [49] [50]. Each network is accompanied by the confusion matrix of the best-trained model (selected based on accuracy on the validation set), along with a graph depicting the accuracy for each training iteration. This visual representation facilitates the comparison of networks sometimes, it can reveal trends that other metrics might miss. Each row of pixels in the confusion matrix illustrates how frequently a particular class appeared, while each column indicates how often a specific class was predicted by the network. Correct predictions align with the diagonal. A dark blue pixel signifies that a guess of that type was never made for the class represented by that row, while a yellow pixel indicates a 50% prediction rate. A dark red pixel denotes a 100% prediction rate. An ideal network, correctly classifying every image, would be dark red along the diagonal and dark blue elsewhere.

Additionally, a graph displaying overall accuracy for every 40 training iterations is included in Figure 7. This is to give the reader a sense of how transfer learning affects the convergence of a network with transfer learning.

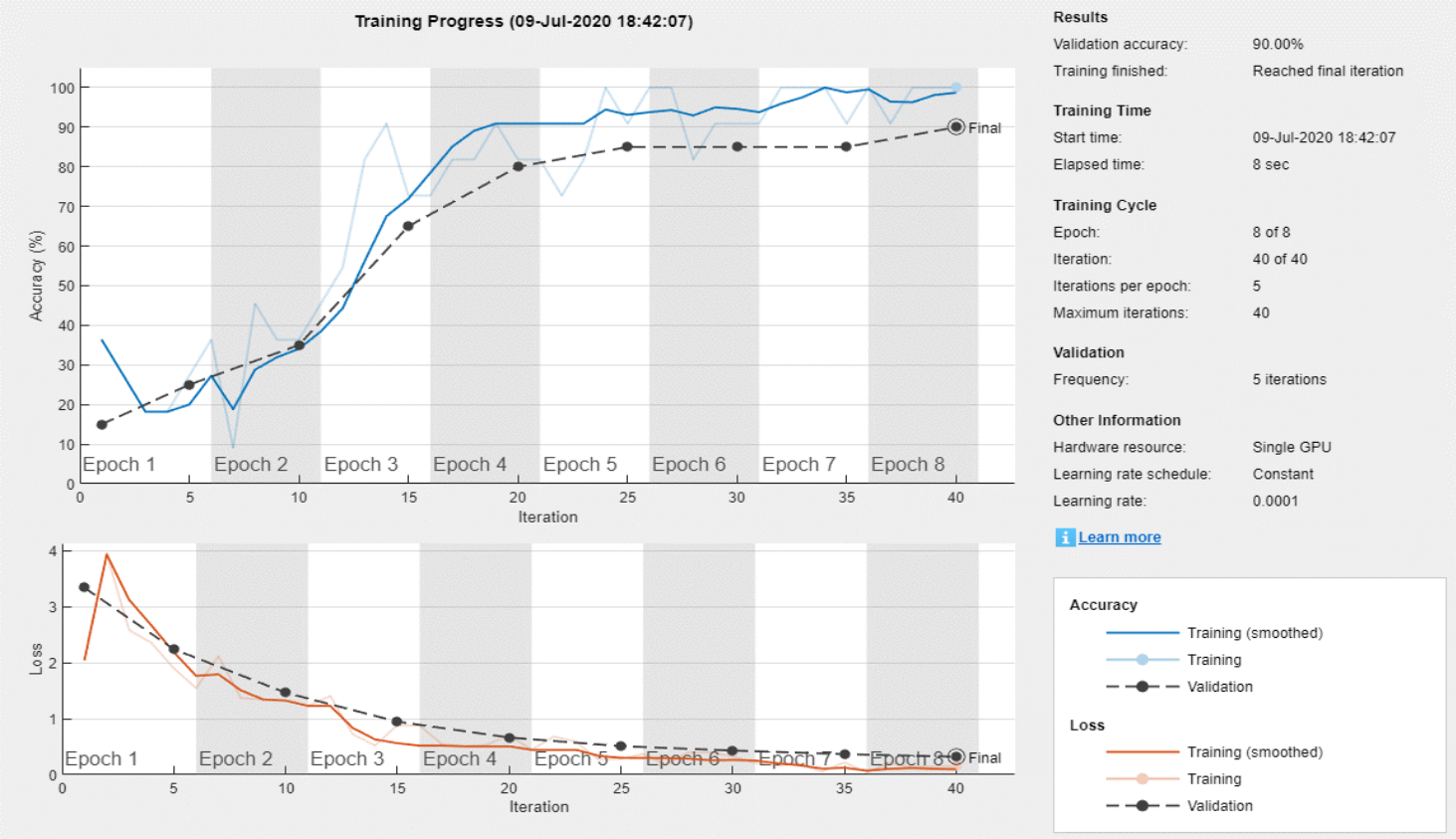

Fig. 7. Training Progress accuracy and loss with 40 Iterations.

IV. RESULTS AND DISCUSSIONS

A. *Real-Life Violence Dataset*

The real-life Violence dataset has been tested with two different epoch values, i.e., 50 and 100. The values of the confusion matrix for both configurations have been shown in Table II and the corresponding performance matrices values in Table III. These tables suggest that doubling the epoch from 50 to 100 results in a slight increase in performance.

E-Poch = 50	**Batch-Size = 64**	**Optimizer = SGDM**
	Violent	Non-Violent
Violent	85.27 14.73	85.27 14.73
Non-Violent	37.52 62.48	37.52 62.48
E-Poch = 100		
	Violent	Non-Violent
Violent	80.31 19.69	80.31 19.69
Non-Violent	23.09 76.91	23.09 76.91

TABLE II: CONFUSION MATRIX VALUES REAL-LIFE VIOLENCE DATASET.

Metric	**Values**	
Epoch	50	100
Accuracy	73.87	78.61
Recall	69.44	77.66
Precision	85.27	80.31
F-Measure	76.54	79.21

TABLE III
PERFORMANCE MATRICES REAL-LIFE VIOLENCE DATASET.

B. *Hockey Fight Dataset*

For the Hockey Fight Dataset, we repeat the same process. The values of the confusion matrix for epochs 50 and 100 have been shown in Table IV, and the corresponding performance matrices values in Table V. Performance comparison of both the datasets is graphically shown in Figure 8. In this figure Accuracy, Recall

E-Poch = 50	**Batch-Size = 64**	**Optimizer = SGDM**
	Violent	Non-Violent
Violent	73.48	26.52
Non-Violent	38.75	61.25
E-Poch = 100		
	Violent	Non-Violent
Violent	85.27	14.73
Non-Violent	37.55	62.45

TABLE IV
CONFUSION MATRIX VALUES FOR HOCKEY FIGHT DATASET.

Metric	**Values**	
Epoch	50	100
Accuracy	73.87	78.61
Recall	69.44	77.66
Precision	85.27	80.31
F-Measure	76.54	79.21

TABLE V
PERFORMANCE MATRICES OF THE HOCKEY FIGHT DATASET.

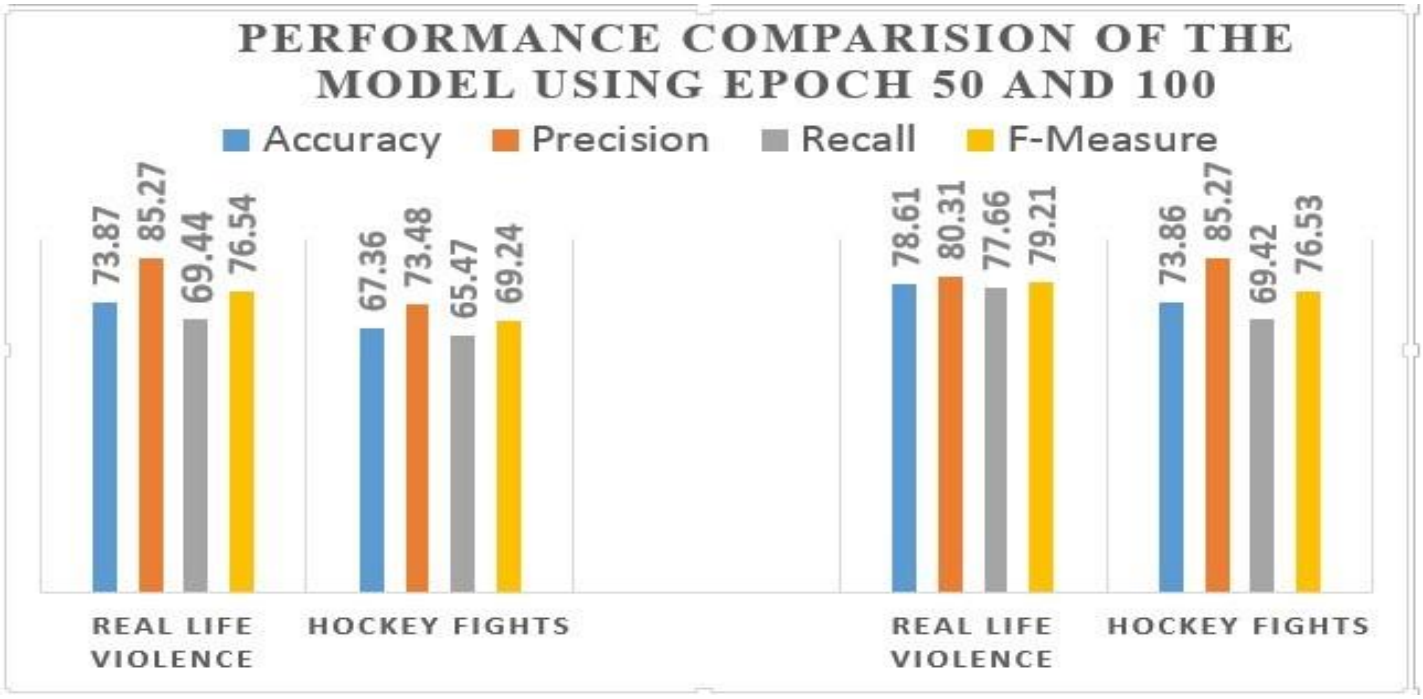

Fig. 8. Performance evaluation of both datasets in Bar-chart

Precision, and F-measure with 50 epochs is shown on the left side while the performance of the model on 100 epochs is shown on the right side of the chart.

From these results, it can be concluded that increasing the epoch value may increase or decrease the performance of our scheme. Overall, 100 epoch values achieved better results for both datasets as compared to 50 epochs, but this may not work for every dataset. It is also noteworthy that the performance gain by increasing the epoch value for both datasets used in this research is not the same. The performance gain for the Real Life violence dataset is more considerable in comparison to the performance gain in the Hockey Fight Dataset. This might be due to the fact that the videos in the two datasets used in this research have different resolutions, lighting, and environment. The Hockey fight dataset videos were recorded with a moving camera, while the videos in the real-life violence dataset are mostly recorded with static cameras, both indoors and outdoors. Therefore, we can conclude that the proposed approach works best for static videos. However, that may change when there are more classes in the data set or using some different instead of transfer learning or for the removal of redundant frames

C. Results Comparison with State of the Art

This section compares the performance results of our proposed scheme with that of state-of-the-art schemes proposed in the literature in terms of accuracy. Table VI summarizes the results of different state-of-the-art methods compared to our proposed model, graphically represented in Figure 9. From this table and the figure, it can be observed that all state-of-the-art models considered uses a combination of 4 different datasets and different classifier algorithms. Therefore, the accuracy of all methods varies. Some methods are performing extremely well for one dataset, but their accuracy drops sharply when the dataset changes. This means that these methods work well for limited situations. SVM with Ada boost in [51] didn't have low accuracy in comparison but performed consistently well when the dataset is changed. CNN transfer learning [55] without removing the redundant frames' performance is also not remarkable because of the overfitting problem and also the results are not consistent. The proposed method uses CNN and Siamese. In comparison to the literature, our proposed method performed consistently for both datasets. Classifier is working very well for both datasets which means the model has learned rich features, therefore the accuracy is decent on both of the datasets. The proposed method is remarkable because it is working well even if we changed the dataset. Some state-of-the-art methods are working extremely well for a few datasets, but are still not stable for other dataset.

V. CONCLUSION

Violence detection is a very important area of real time human activity recognition and surveillance. Timely, robust, and accurate voice detection is critical and can save potentially human lives. This research has been proof of the concept to remove the redundant video frames from the dataset, the redundant video frames are removed using an unsupervised learning algorithm. In order to find the similarity between two sequential video frames, CNN-based features are extracted, and using Euclidean distance which is a state-of-the-art similarity measure technique the distance is calculated. Two images with less similarity score show that video frames have more similarity, using a threshold scheme the similar video frame is considered as

redundant and removed from the dataset. For the detection of violence in a video frame, transfer learning is performed in the

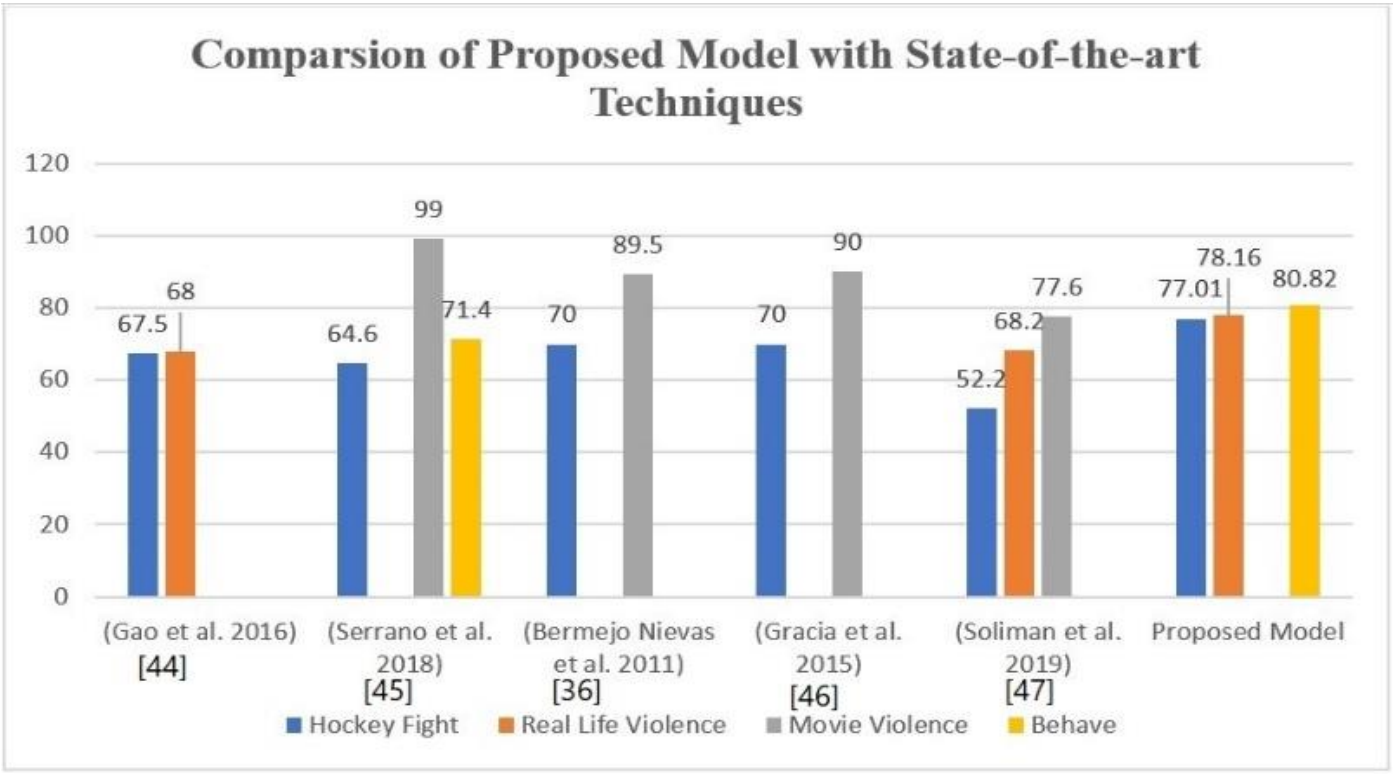

Fig. 9. Comparison of Proposed Model with some State-of-the-art Models using Bar Graph.

Method	Dataset	Classifier	Max Accuracy
[44]	Hockey Fight, Movie violence	Bag-of-Words + STIP, and MoSIFT.	89.5%, 70.00%
[51]	Hockey Fight, Real Life violence	SVM + Adaboost	67.50%, 68.00%
[52]	Hockey Fight, Behave	2D CNN, Hough Forest	64.6%, 71.4%
[53]	Hockey Fight, Movie Violence	SVM	70.0%, 89.5%
[54]	Hockey Fight, Movie Violence	Random Forest	70%, 90%
[55]	Hockey Fight, Real Life Violence Situation.	CNN (Transfer Learning)	52.2%, 68.2%
Proposed Model	Hockey Fight Real Life Violence Situation	Siamese + CNN	77.01%, 78.16%

TABLE VI
DETAILED ACCURACY WITH DIFFERENT DATASETS AND CLASSIFIER.

inception-v3 CNN model. First, the model is customized by removing old weights, labels, features, and biases that are residing in the last three layers of the inception-v3 pre-train model. Hold-out cross-validation is applied to split the dataset into training-set 70% and 30% test set. The proposed violence detection system is validated using benchmark violence datasets. The proposed model shows higher accuracy for classifying frames into violent and nonviolent scenes, also with the removal of redundant frames the training and validation process becomes very efficient.

REFERENCES

[1] Yuan GAO, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. Violence detection using oriented violent flows. Image and Vision Computing, 48:3741, 2016..

[2] Eneim, M. (2016). An Intelligent Method for Violence Detection in Live Video Feeds (Doctoral dissertation, Florida Atlantic University).

[3] I. P. F. K. Jayasree and P. Theresa, "Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm," Pattern Anal. Appl., no. 0123456789, 2019.

[4] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human ac-tions: A local svm approach. In Pattern Recognition, 2004. ICPR 2004. Pro-ceedings of the 17th International Conference on, volume 3, pages 32-36. IEEE, 2004.

[5] F. Gong et al., "A Real-Time Fire Detection Method from Video with Multifeature Fusion," vol. 2019, 2019.

[6] N. S. Altman. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician, 46(3):175185, 1992.

[7] Javed A. Aslam, Raluca A. Popa, and Ronald L. Rivest. On Estimating the Size and Con dence of a Statistical Audit. In Proceedings of the USENIX Workshop on Accurate Electronic Voting Technology, EVT'07, pages 8-8, Berkeley, CA, USA, 2007. USENIX Association.

[8] Sandra Avila, Daniel Moreira, Mauricio Perez, Daniel Moraes, Isabela Cota, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. RE-COD at MediaEval 2014: Violent Scenes Detection Task. In Working Notes Proc. MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, 2014.

[9] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In European conference on computer vision, pages 404-417. Springer, 2006.

[10] Christopher M Bishop. Pattern recognition and machine learning, volume 60. Springer, 2012.

[11] Leo Breiman. Bagging Predictors. Machine Learning, 24(2):123-140, 1996.

[12] Mingyu Chen and Alexander Hauptmann. MoSIFT: Recognizing human actions in surveillance videos. Research Showcase at Carnegie Mellon University, 2009.

[13] Guangchun Cheng, Yiwen Wan, Abdullah N. Saudagar, Kamesh Namuduri, and Bill P. Buckles. Advances in Human Action Recognition: A Survey. CoRR, abs/1501.05964, 2015.

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. Machine Learn-ing, 20(3):273-297, 1995.

[15] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. Background and foreground modeling using nonparametric kernel density for visual surveillance. Proc. IEEE, 90, 2002.

[16] Yoav Freund, Robert Schapire, and N Abe. A short introduction to boosting. Japanese Society for Artificial Intelligence, 14(771-780):1612, 1999.

[17] H. Ali Palwasha Zeb, Arbab Muhammad Qasim, Muhammad Qasim Khan, "Classification of Acute Myeloid Leukaemia using Deep Learning Features", The Science Tech 3 (4 Oct-Dec 2022), 84-98 2023. Online Availabe at https://journals.qurtuba.edu.pk/ojs/index.php/tst/article/view/753

[18] H.Ali, Arbab Muhammad Qasim, Palwasha zeb, Muhammad Qasim Khan,"Automatic Detection and Classification of Acute Lymphoblastic Leukemia Using Convolution Neural NetworkThe ScienceTech 3 (4 Oct-Dec 2022), 59-72. Online Availabe at https://journals.qurtuba.edu.pk/ojs/index.php/tst/article/view/686

[19] Rubab Wafa, Muhammad Qasim Khan, Fazal Malik , Akmalbek Bobomirzaevich Abdusalomov, Young Im Cho and Roman Odarchenko, "The Impact of Agile Methodology on Project Success, with a Moderating Role of Person's Job, Fit in the IT Industry of Pakistan". Journal of Applied Sciences, VOLUME 12, NUMBER 21, Year 2022, ISSN 2076-3417,DOI: 10.3390/app122110698

[20] Shah Hussain, Muhammad Qasim Khan, "Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning". Annals of Data Science (2021) DOI: 10.1007/s40745-021-00341-0

[21] Muhammad Qasim Khan, Steinar Hidle Andresen, Muhammad Inam Ul Haq, "Handover Architectures for Heterogeneous Networks using the Media Independent Information Handover(MIH)". Computing and Informatics. 2016, vol. 35 (1) (Impact factor 0.542). Available online. http://147.213.75.17/ojs/index.php/cai/article/view/1558

[22] K. Ullah, Muhammad Qasim Khan"GOOGLE STOCK PRICES PREDICTION USING DEEP LEARNING". In Proceedings of IEEE International Conference on System Engineering and Technology (ICSET 2020).DOI: 10.1109/ICSET51301.2020.9265146.

[23] Muhammad Qasim Khan, "Signaling Storm Problems in 3GPP Mobile Broadband Networks, Causes, and Possible Solutions: A Review". In Proceedings of IEEE International Conference on Computing, Electronics & Communications Engineering 2018 (iCCECE '18) University of Essex, Southend, UK. DOI: 10.1109/iCCECOME.2018.8658708.

[24] J. Yu, W. Song, G. Zhou, and J. Hou, "Violent scene detection algorithm based on kernel extreme learning machine and three-dimensional histograms of gradient orientation," 2018.

[25] G. Li, H. Song, and Z. Liao, "An Effective Algorithm for Video-Based Parking and Drop Event Detection," vol. 2019, 2019.

[26] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Realtime de-tection of violent crowd behavior. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, pages 1-6. IEEE, 2012.

[27] M Vrigkas, C Nikou, and IA Kakadiaris. A Review of Human Activity Recognition Methods. Front. Robot. AI 2: 28. doi: 10.3389/frobt, 2015.

[28] H. Wang, A. Klaser, C. Schmid, and C. L. Liu. Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis, 103, 2013.

[29] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable key points. In 2011 International conference on computer vision, pages 2548-2555. IEEE, 2011.

[30] J. Liu, J. Yan, M. Tong, and Y. Liu. A Bayesian framework for 3D human motion tracking from the monocular image, in IEEE International Conference on Acoustics, Speech and Signal Processing (Dallas, TX: IEEE). 13981401, 2010.

[31] David G Lowe. Distinctive image features from scale-invariant key points. International Journal of Computer Vision, 60(2):91-110, 2004.

[32] V. I. Morariu and L. S. Davis. Multi-agent event recognition in structured scenarios. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011.

[33] Mark Nixon. Feature Extraction & Image Processing. Academic Press, 2008.

[34] M. Pantic and L. Rothkrantz. Towards an ect-sensitive multimodal human-computer interaction. In IEEE, Special Issue on Multimodal Human-Computer Interaction, Invited Paper, 2003.

[35] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in rst-person camera views. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2847-2854, June 2012.

[36] Trevor Hastie; Robert Tibshirani; Jerome Friedman. The Elements of Statistical Learning. Springer, 2 edition, 2008.

[37] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. IEEE Trans. Pattern Anal. Mach. Intell, 29, 2007.

[38] Neil Robertson and Ian Reid. A general method for human activity recognition in video. Computer Vision and Image Understanding, 104(2):232-248, 2006.

[39] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network," pp. 1–15, 2019.

[40] M. T. Khan et al., "Smart Real-Time Video Surveillance Platform for Drowsiness Detection Based on Eyelid Closure," vol. 2019, 2019.

[41] Ismael Serrano Gracia, Oscar Deniz Suarez, Gloria Bueno Garcia, and Tae-Kyun Kim. Fast ght detection. PLoS ONE, 10(4):1-19, 04 2015.

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. ArXiv preprint arXiv: 1212.0402, 2012.

[43] Tao Zhang, Zhijie Yang, Wenjing Jia, Baoqing Yang, Jie Yang, and Xiangjian He. A new method for violence detection in surveillance scenes. Multimedia Tools and Applications, pages 1-23, 2015.

[44] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno Garc a, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In International Conference on Computer Analysis of Images and Patterns, pages 332-339. Springer, 2011.

[45] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno Garc´ıa, and R. Sukthankar, "Violence detection in video using computer vision techniques," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6855 LNCS, no. PART 2, pp. 332–339, 2011, doi: 10.1007/978-3-642-23678-5 39.

[46] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "applied sciences cover the voilence: A Novel Deep-Learning-Based Approach Towards Violence-Detection in Movies", 2019.

[47] Pang-Ning Tan. Introduction to Data Mining, Section 4, pages 145-205, 2006.

[48] Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani. An Introduction to Statistical Learning, pages 316-321. Springer, 2013.

[49] Jing Wang and Zhijie Xu. Spatio-temporal texture modelling for realtime crowd anomaly detection. Computer Vision and Image Understanding, 144:177-187, 2016.

[50] Long Xu, Chen Gong, Jie Yang, Qiang Wu, and Lixiu Yao. Violent video detection based on MoSIFT feature and sparse coding. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3538-3542. IEEE, 2014.

[51] Tao Zhang, Wenjing Jia, Baoqing Yang, Jie Yang, Xiangjian He, and Zhong-long Zheng. MoWLD: a robust motion image descriptor for violence detection. Multimedia Tools and Applications, pages 1-20, 2015.

[52] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "violence detection using Oriented VIolent Flows," Image Vis. Comput., vol. 48–49, no. 2015, pp. 37–41, 2016, doi: 10.1016/j.imavis.2016.01.006.

[53] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight Recognition in Video Using Hough Forests and 2D Convolutional Neural Network," IEEE Trans. Image Process., vol. 27, no. 10, pp. 4787–4797, 2018, doi: 10.1109/TIP.2018.2845742.

[54] I. S. Gracia, O. D. Suarez, G. B. Garcia, and T. K. Kim, "Fast fight detection," PLoS One, vol. 10, no. 4, pp. 1–19, 2015, doi: 10.1371/journal.pone.0120448.

[55] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence Recognition from Videos using Deep Learning Techniques," Proceeding of - 2019 IEEE 9th Int. Conf. Intell. Comput. Inf. Syst. ICICIS 2019, pp. 80–85, 2019, doi: 10.1109/ICICIS46948.2019.9014714.

AUTHORS

First Author – Muhsin Khan, Department of Computer Science, Iqra National University, Peshawar, KPK, Pakistan

Second & Correspondence Author – Dr. Muhammad Qasim Khan, Department of Computer Science, Iqra National University, Peshawar, KPK, Pakistan,

Third Author – Dr. Fazal Malik, Department of Computer Science, Iqra National University, Peshawar, KPK, Pakistan

Fourth Author – Muhammad Suliman, Department of Computer Science, Iqra National University Peshawar, KPK, Pakistan

.