International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 05, ISSN No. 2455-2143, Pages 21-30
Published Online September 2024 in IJEAST (http://www.ijeast.com)

INTELLIGENT SURVEILLANCE SYSTEM

Abdullah Asim, Hammad Javed
Department of Computer Science
FAST NUCES, Lahore

Abstract— **Watchful surveillance is one of the crucial aspects when it comes to ensuring security in any surroundings. Be it a shopping mall, a densely populated residential area or even busy roads vigilant surveillance is inevitable to ensure security and safety of citizens. This work focuses on the design and development of an Intelligent Surveillance System. It aims to enhance the security measures by identifying abnormal behavior like fight and fire in real-time scenarios. The proposed framework for fight detection includes 3 major functionalities/ components. i.e., preprocessing, Feature extraction and anomaly detection. Preprocessing is applied to optimize the data and get efficient results from the model. After Preprocessing this frame stream is passed to Mobilenet (A variation of CNN developed by google and using pretrained weights of imagenet) and Bi-LSTM that detect the fight in it. The fire use case has also been handled by using a CNN. The work done highlights the effectiveness of the proposed framework in accurately detecting anomalies, and contribution to enhanced security in surveillance systems.**

Keywords—**Neural Network, GRU, CNN, LSTM, SVM, RNN, ANN.**

I. INTRODUCTION

Surveillance is one of the most important measures needed for the assurance of security in any area. Security is needed anywhere where there is some population living. A watchful surveillance can be regarded as the one in which security personnel are notified immediately as soon as some potentially hazardous event occurs. Currently we are relying on security cameras and manual surveillance to detect any abnormal behavior which may be threatening. The problem with such surveillance as in normal security cameras is that they are unable to notify the relevant authorities timely and automatically. But with the advancements in machine learning and computer vision we can shift this trend of surveillance from manual to automated intelligent surveillance systems. These systems will be able to detect an anomaly in real time scenarios and resultantly give an alert to take required security measures.

CCTVs have been used since a long time to enforce security, but they are not very effective because they require continuous monitoring. Manually monitoring the CCTV footage is very difficult due to the large amount of video data. Reviewing the surveillance footage to find anomalies is also a tedious task and will require a lot of time. Detecting anomalies becomes even more difficult when the area is crowded. Even if an anomaly is found it will be too late. Demand for real time anomaly detection has increased due to the rapid increase in the number of video surveillance cameras installed in cities and towns. This project aims to develop a surveillance system which invokes an alert whenever an anomaly is detected in a real time scenario. This way the relevant security personnel can act swiftly and efficiently to encounter any potentially dangerous situation. This work will be dealing with the two use cases which are fight and fire.

The first and most important use case is to detect fighting on streets. Fighting has become very common in populated localities.[1]. There is a need to detect these fighting and inform the relevant authorities to minimize the casualties and stop these acts.

A further use case is the detection of fire with the aim to prevent any causality or economic loss. In the United States, a fire department responds to a fire incident every 23 seconds. Moreover in 2020 the United States bore a total loss of $8.1 billion out of which $822.6 million were in total property loss [2]. So, a watchful and timely alert can be vital in prevention of such losses.

A. Project Scope

In this project we will be using Computer Vision techniques to train two models for detection of fight and fire. We will be using a supervised learning approach. Our work is currently divided into 4 phases.

In the first phase we will be working on the dataset. Apart from using the available data on the internet we will also prepare our own dataset to improve the accuracy of our model. We will prepare our own dataset by searching surveillance videos from YouTube and from movies. In the second phase we will be working on training different models on our prepared dataset. In the third phase we will evaluate the performance of different models and search for the best one. In the last phase we will be working on a proper way of notifying once the anomaly is detected. Main technologies to be used are OpenCV for processing the video data, TensorFlow for implementing the deep learning models like CNN, BiLSTM.

B. Purpose Work

Public safety is the main purpose we wish to achieve in our project by developing a system which can correctly classify any abnormal behavior which is potentially threatening. The

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 05, ISSN No. 2455-2143, Pages 21-30
Published Online September 2024 in IJEAST (http://www.ijeast.com)

purpose of this project is to develop a model which can identify anomalies in a video with improved accuracy and can assist and ease the process of surveillance by passing an immediate alert whenever an anomaly is detected. The model will be trained to classify fight and fire use cases. It will also optionally classify the presence of weapons or fire in a surrounding. The benefit we want to achieve from this project is improved surveillance which is crucial for security of any surroundings.

C. Intended Audience

The audience that will be using our model/surveillance system is basically anyone or anyplace where security needs to be ensured for public safety. It can be shopping malls, a densely populated locality, busy roads or even educational institutions where it immediately needs to be determined if someone is fighting somewhere within the premises of that educational institute. The user of the model will be the relevant authorities which are responsible for security of the respective area. The model will ultimately be used for the protection and safety of the people in the area where it is implemented.

II. LITERATURE REVIEW

A detailed literature review written using various research papers.

A. Real-world Anomaly Detection in Surveillance Videos [3]

a) Summary of the research item:

In this research work the authors have created their own data set through youtube videos, real life leaks. This data set includes 13 real life anomalies i.e Abuse Arrest, Fighting. In this research work videos are firstly resized to 240 x 320 pixels and the fixing frame rate to 30fps. Then features from the videos are extracted using fully connected layer FC6 of the C3D network. In this research work videos are first divided into the segment and then two different sections/bags are created one for anomaly videos and other for normal videos. In Deep MIL ranking model ranking is not applied on every instance of the bag but instead it is applied only on those instances of the bag that have high anomaly scores.

In this research work anomaly detection is posed as a regression problem. where anomalous videos should have high anomaly scores. Before moving toward feature extraction of classifying videos there is some pre-processing that includes resizing each video frame to 240 x 320 pixel and frame rate is fixed at 30fps. C3D Features for every 16 frames of videos are computed, for the computer feature for the video to take an average of 16frame clip features.

b) Critical analysis of the research item

In this research work based on real world surveillance videos a new, diverse data set was created for training purposes because videos through surveillance cameras have a lot of light and weather change. Its performance was increased due to explicit motion information. For better use of sequential information at prediction stage LSTM was used by BI-LSTM is much more beneficial in this case.

c) Relationship to the proposed research work

This research work is directly related to anomalies detection in surveillance videos. According to the authors it includes 13 real world anomalies including fight detection which is the main purpose of our work. Data set created in this research work can be used for our testing purpose because the data set contains not only YouTube videos but also real time leak videos.

B. Fight Detection in Video Sequences Based on Multi-Stream Convolutional Neural Networks [4] a) Summary of the research item:

In this research work a deep learning approach based on multistream and high-level hand craft descriptors to detect fights in the videos. Firstly, video frames are passed through the feature generator algorithm. Output of each generator is then given as an input to a modified VGG so that a weight vector can be calculated.

This research work is using the methodology of multiteam architecture. Due to insufficient amount of information in dataset each VGG-16 model has to go through different datasets. According to the authors, they have used the VGG-16 model because it is simple to understand, less deep than other models and follow the classical coevolutionary approach.

b) Critical analysis of the research item

The evaluation metrics used in this research paper was based on already existing models so that it can easily be compared. In this research paper outlier classification is balanced by learning the high-level features individually.

The data set used in this research work contains the Hockey dataset and Movie fight dataset. It is not trained over the surveillance cameras videos so it may not work. VGG-16 is a simple model but due to many parameters it is not the best option for this type of work.

c) Relationship to the proposed research work

This research work is related to fight detection. The methodology and techniques used in these research papers are close to techniques that we have proposed in our research work. Firstly, this research work talks about the fight detection in video that is related to our proposed aim of Fight detection in surveillance cameras. Another relation is that they have used VGG-16 for feature detection but we can keep the same methodology and change this model with Bi-LSTM.

C. Learning to Detect Violent Videos [5]

a) Summary of the research item:

In this research work three different datasets, A combination of Convolutional Neural Network and Convolutional LSTM is

used. Frame level features from a video are first extracted through Convolutional Neural Network. Then these frame level features are aggregated using LSTM with convolutional gates. This research work showed that the Recurrent Neural Network capable of encoding localized spatio-temporal changes is better for detecting violence in the video. Deep Neural networks trained on difference of frames perform better then the one trained on raw frames.

b) Critical analysis of the research item

In this research the method is evaluated on three datasets and resulted in improvement in performance as compared to state of art methods. Instead of just training the model on some raw frames' accuracy and performance of the model is improved by training it over the difference of frames. In this research work convolutional Neural Network is used instead of traditional fully connected Neural Network because it gives better video representation as compared to LSTM.

In this research work three different datasets are used but the size of the Hockey dataset was larger than the other two dataset so this model will perform better if tested over the hockey dataset.

c) Relationship to the proposed research work

Fight detection is an important use case for our proposed work and this research paper is also related to violence detection in videos. The methodology and technique used in this research paper is similar to our proposed one i.e., first using CNN for extracting the features from frames and then using Conv LSTM for classification. We can also use these datasets to train our models not only on surveillance videos but also on movies' fight videos.

D. Fight detection in Hockey videos [6]

a) Summary of the research item:

This research work is related to fighting in hockey sport videos using blur and random transform and Convolutional Neural Network. Firstly, local motion within video frames is extracted through local motion. frames with fight scenes are detected using a pre-trained deep learning model VGG-Net.

In this research work Feed Forward Neural Network and Convolutional Neural Network both are used. In Feed Forward Neural Network the violent action in a video can be identified through large acceleration or deceleration. It can be calculated using motion blur. In the case of Convolutional Neural Network, the first 3D convolutional layer is used to extract features from input data, then it is passed to a pooling layer that reduces the complexity of data. Then the last layer (Fully connected layer) like feed forward neural network is used to get the desired output.

b) Critical analysis of the research item:

In this research work the fight detection is done through blur and radon transform with the help of Feed Forward Neural Network. Its performance was also improved using pre-trained 22VGG and CNN. With simple Feed Forward the performance is recorded as high as 56.00% whereas after tuning of VGG16Net the performance improved to 75.00%.

The complete training and testing of this model is done on hockey dataset so this performance and accuracy is only possible in case of Hockey videos while in case of surveillance video its performance will differ.

c) Relationship to the proposed research work:

The main objective of our proposed model is to detect fights in surveillance videos and this research paper is also related to fight detection but in just Hockey sports videos. In this research paper hockey dataset is used for training and testing purposes we can also use this dataset to enhance performance of our model. This research work is linked with our proposed method because of the methodology and technique of feature extraction using CNN.

E. Fight detection in Hockey videos [7]

a) Summary of the research item:

In this research work CNN is used for feature extraction while Bi-LSTM is applied as a classification model once the features have been extracted by CNN. The input to CNN is not a video, rather uniform sampling is used, and 5 or 10 frames are extracted from each video sequence, resized to the input size of architecture and then sent to CNN for feature extraction. Various CNN's are used in experimentation such as VGG16, Xception and a specially trained Fight-CNN which has the similar architecture as the Exception with the difference that it has two fully connected layer before the classification layer and the classification layer is mapped onto two classes. For classification Bi-LSTM is used as it has the ability to learn the context of past and current information. The input to Bi-LSTM are the extracted features obtained from Fight-CNN discussed before. Once the forward learning is done a backward pass is processed and, in this way, both past and future information is taken into consideration while making a decision on classification. An attention layer is used which determines how much each output should be affected by other inputs. The purpose of the attention layer is to get the extracted feature vector and focus on only the significant data in it i.e., it forms a new vector of features which contain only those features which are significant. It does so by checking their attention matrix and their relationship with other inputs. This newly formed vector of features is passed to classification layers.

b) Critical analysis of the research item:

In this research work the use of Bi-LSTM for classification is one of the key factors in improving the accuracy of the classification. The inherent ability of Bi-LSTM to learn the past and current context of the video sequence allows for better classification. Furthermore, the use of a modified Xception model i.e., a Fight-CNN and an attention layer has also proven to do better while extracting the right features

from frames. Dividing the video into frames and passing them as input rather than the whole sequence in one go is also beneficial.
But one of the drawbacks in this work is that the model is intensely trained on the hockey dataset with about 800 videos. So, when it is tested on a hockey fighting video it's accuracy is as high as 97% but when it is tested in a real time scenario i.e. a fighting scene in a public places like a restaurant, a street or a shopping mall etc. it's accuracy falls to 72% which is somewhat because of the fact that the model was over trained for hockey dataset and less trained for such real time scenarios outside hockey.

c) Relationship to the proposed research work:
The methodology and techniques used in the above discussed research paper are very close to what we have proposed in our research work. First of all, the research work talks about fight detection in surveillance cameras which is exactly our proposed aim for the research work. Fighting detection is one of the key use cases on which we will train our model. Moreover, as it is the case with the discussed research work, we also intend to train these models in a way that they are able to perform well for surveillance cameras footage. Furthermore, one of the main relationships to the proposed research work is the use of CNN for feature extraction and Bi-LSTM for classification. Last but not the least we are also interested in processing the video as frames and use them as input to the CNN.

F. Detection of Real-world Fights in Surveillance Videos [8]

a) Summary of the research item:
In this research work the authors have used a newly collected dataset of CCTV-Fight footages. The dataset is different from what has been used in many other research works discussed here. It is collected from YouTube using different keywords such as CCTV-Fight, Mugging, i.e., ical, Violence , Surveillance e.t.c.50% data is used for training, 25% for validation and 25% for testing which is selected randomly. The crux of the methodology used in the research work is to extract the features using CNN (2D and 3D) and local interest-points and then pass on those useful and relevant features to a classifier like LSTM and SVM. The final step in the method includes generation of segments by aggregating the predictions from previous steps.
In the feature extraction step, three approaches are discussed. Firstly, VGG16 architecture (2D CNN) is used in which two models are generated, one for RGB data of frames and other for temporal stream (stack of optical flows). C3D architecture (3D CNN) is applied on a sequence of 16 frames and correlation between sequence of frames can be obtained. In local interest points local feature detector named Temporal Robust Features (TORF) is used which extract low level spatial temporal features from the video. For classification prediction is made on a frame or snippet level and is given a confidence score. For the 2D CNN approach prediction is done by the CNN classification layer. For 3D CNN based approach LSTM is used as a classifier. Parameters and architecture were determined by using grid searching while performance of LSTM was evaluated using validation split. For the TORF based solution a simple linear SVM is used for prediction. Cross validation and grid search is used during the training phase of SVM. In the final step smoothing and then aggregation is applied to each previous snippet and then a final segment is generated.

b) Critical analysis of the research item:
The most invaluable contribution of this literature is the dataset formed and used. Normally most of the literature related to surveillance videos use the hockey dataset and Hollywood movies fight dataset which does not fully meet the needs of a good dataset. The dataset used here contains real footage of CCTV cameras gathered from YouTube. Also, the average length of videos in the dataset is 2 minutes for fight and 45 seconds for non fight while the range of videos for fight is 5 seconds to 12 minutes while it is 3 seconds to 7 minutes for non fight. This duration and range is far better than what other datasets offer. Another good thing is that explicit motion information is used which has shown a positive impact on the performance of the model shown by the results in the research work.
One of the drawbacks, however, is that at the classification level the sequential information (frame/snippets) has not been used properly because of the use of LSTM. Use of Bi-LSTM here would have provided better results as it takes into account the contextual information of past and future for its prediction. Also, there can be improvements made to detect early fight motion which is crucial in fighting detection.

c) Relationship to the proposed research work:
Fighting detection is a key use case for our proposed research work and this paper directly deals with that use case. The dataset used in the discussed research work can be used in our model for training, validation or testing. As we require CCTV footage of fighting to implement our model this dataset is useful for us. The methodology used here also resembles the approach we will be using (Frames as Input, Feature extraction and processing, then a classifier). Furthermore, different techniques of feature extraction are discussed which are similar to what we are supposed to use i.e., using CNN (VGG16, C3D) for feature extraction. The classifier used is LSTM and SVM. Although we are more inclined to use Bi-LSTM but still the use of LSTM and SVM for experimentation is helpful in giving us a broader view of the performance of various models.

G. Automatic Fight Detection in Surveillance Videos [9]

a) Summary of the research item:
The approach used in the research work focuses on detection of fast moving "objects" or motion regions. The research work

aims to build a model for low resolution videos with noise and in such case action recognition or a human body part recognition is not easy to do. So instead, a motion analysis approach is used where first of all optical flows are computed for two consecutive frames, motion information is extracted from it. The second step involves noise removal from optical flows generated. As the dataset contains videos from surveillance cameras hence there are many noise factors including background movements and environmental light changes. So, it's vital to remove these noises before any further processing. After that type of motion is detected. In this step the author has categorized motion into different types based on their number of motion regions, average size of motion region, moving directions etc. Once the motion type is determined, features are extracted from it to pass them onto the machine learning model for classification. The three types of attributes collected in this step are motion magnitude, motion acceleration and motion region attraction. Then the machine learning algorithm SVM is used to classify the extracted features as fight or non fight.

b) Critical analysis of the research item:

The dataset used in the research work has a good variety and is taken from real CCTV footage as opposed to the hockey dataset used in most of the literature related to fight detection in a video. One more good thing about the research work is the amount of work they put in pre-processing the frames before actually extracting the features from it. As we know that videos from cameras may contain varying environments and surroundings with many noisy factors such as background movements or light changes etc. It is vital to clean your data before processing it for use in the ML model. This approach is more robust as it does not rely on high level action recognition. A drawback in the research work is the inability of the model to efficiently classify the simulated fights videos. The accuracy of such datasets is lower than what other research works have been able to achieve. Furthermore, early fight detection and detection from different angles of videos has not been discussed as well.

c) Relationship to the proposed research work:

The approach used here in this research work to focus on the fast-moving motion regions instead of detecting the whole action or a human body part makes it possible for the models to be robust which is exactly what we are trying to achieve in our research work. We want to build a model which is not only accurate in detection of fights but is also low in consumption of resources and is applicable in low quality surveillance videos as well. So, this research gives us the pathway of how to build this feature in our model. Also, the various types of datasets used here including real fight dataset and simulated fight dataset are beneficial in checking if our model can perform well for both the scenarios or even better if it can distinguish between the two i.e., a real fight or a simulated fight.

H. Efficient Violence Detection in Surveillance [10]

a) Summary of the research item:

In this research work the proposed algorithm consists of three main steps. In the first step spatial features are extracted using time distributed U-Net. Second step involves the extraction of temporal features while the last step is where classification is performed. The input to the U-Net is 30 frames snippet of 1second of the video. Once it receives the frame it extracts spatial features from it in a sequential time distributed manner using Mobile net V2 as an encoder. This newly formed sequence of features is then passed onto the next stage where temporal features are obtained using LSTM. Finally, all this information is used by the two-layer classifier Mobile net V2 which classifies the events as either violent or nonviolent.

Three different varieties of datasets are used in this work namely hockey dataset, movies fight dataset and RW-2000 dataset by Cheng et al. The RW-2000 dataset contains the actual footage of CCTV cameras where 1000 videos are violent, and 1000 videos are non violent out of which only 1600 videos are chosen. Whereas hockey and movie fights dataset are the same widely used dataset in the field of anomaly detection. The model takes about 4 million parameters. Cross validation with five-fold was performed for experimentation and accuracy, precision and F1 score was calculated for each dataset separately.

b) Critical analysis of the research item:

One of the main strengths of this research is the use of the RW2000 dataset. The RW-2000 dataset is vital because of two reasons. Firstly, it is a huge dataset with 200model stop000 violent and 1000 non violent). Secondly this dataset provides real CCTV footage of fights in various scenarios such as in daylight or at night, in a crowded or deserted area etc.. These variations are invaluable if we want to train a model for real time detection of fighting. Another strength of the work is the fact that the model they built is a lightweight model (in terms of number of parameters used) astoo models containing LSTM, Xception,C3D etc. presented in literature.

As far as weaknesses are concerned, the model could not surpass the accuracy of Xception and Bi-LSTM approach in movie and hockey fight dataset. Another limitation is the use of LSTM in extraction of temporal features where Bi-LSTM could have been useful due to its ability to provide contextual information of both past and future frames with respect to the current frame.

c) Relationship to the proposed research work:

The discussed research work is about fighting detection in a surveillance video which is one of our main use cases in the proposed research work. The dataset RW-2000 used in the research paper can be of real importance as it contains videos of varying environments and surroundings. So, it can be a useful resource. Another key link to our proposed research is the fact that this research work provides a comparative view of different approaches that have been used till now in the field

of anomaly detection in video surveillance. This helps us build a better understanding of what we are supposed to use in order to improve the accuracy of our model.

I. Literature Review Summary

Real-world Anomaly Detection in Surveillance Videos	Deep Multiple instances Ranking framework, Binary SVM Classifier C3D, TCNN	Accuracy of C3D = 23.0 Accuracy of TCNN = 28.4
Fight Detection in Video Sequences Based on Multi-Stream Convolutional Neural Networks	CNN (VGG16), SVM	Accuracy: 89% for hockey dataset 100% for movie fight dataset
Learning to Detect Violent Videos	Convolutional Neural Network (CNN) Recurrent Neural Network (RNN) Convolutional Long Short-Term Memory,	Accuracy = 97.1±0.55%
Fight Detection in Hockey Videos	1) Feed Forward Neural Network 2) Deep learning model (VGG16-Net)	Accuracies of 56.00% and 75.00% achieved
Vision-based Fight Detection From Surveillance Cameras	CNN, Bi-LSTM	Accuracy = 72%
Detection of Real-World Fights in Surveillance Videos	Local Interest points, CNN (VGG16 and C3D), LSTM , SVM	mAP: 79.5% F-measure: 75.9%
Automatic Fight Detection in Surveillance Videos	Optical Flow for motion analysis Support Vector Machine Classifier (SVM)	Corrected Classify rate (CCR) = 78.6%
Efficient Violence Detection in a surveillance video	U-Net,Mobilenet V2, LSTM	Accuracy: 82.0 ± 3% for RWF-2000 dataset 96.1 ± 1% for hockey fights dataset 99.5 ± 2% for movie fights dataset

TABLE I. LITERATURE REVIEW SUMMARY

J. Conclusion

After going through the literature related to the fight detection in a video or in general anomaly detection, we have come across different approaches to solving the problem at hand. The approaches include applying optical flow for motion analysis, using CNN specially VGG16 and C3D has been used in much research works for feature extraction, and then for classification we have come across models like LSTM, Bi-LSTM, SVM etc. All these techniques and approaches have their own pros and cons and from what we have read so far, we conclude that the use of CNN and Bi-LSTM has shown better results as compared to other methods.

Another big challenge faced in this domain is the availability of a quality dataset. By quality dataset we mean a collection of real-life fight footage of CCTV cameras covering various environments and surroundings such as videos in bright daylight or at night with low street lights, videos with much activity in background or with no activity or motion in background except for the fight itself. Variety of dataset is one of the main challenges we need to cater to. In most of the research work the accuracy obtained for hockey and movie

fight datasets is quite high but when it comes to real CCTV footage of fights the accuracy significantly drops. The accuracy of hockey fights dataset and movie fight dataset is not much relevant as it does not provide the variety of fight scenarios and surroundings which real CCTV videos can provide. One more factor which is evident in most of the research work is the preprocessing or cleaning of the dataset. Removing noise from the videos or cropping them to be fit for our use is a vital factor in achieving a good, improved accuracy.

III. PROPOSED APPROACH AND METHODOLOGY

In this chapter we will discuss the approach and methodology we will be using in our system. Detailed description of each step in the pipeline of models will be the point of talk in the given chapter. More specifically we can split the pipeline into four main steps, details of which will be provided below.

A. Data preprocessing and input

The first step in our proposed methodology is to preprocess the data to make it suitable to be passed as input to the next step. In our methodology we aim to process the video frame by frame. First, frames are extracted from the video and some preprocessing is applied on them which will include resizing the frames height and width to a standard size of 64 each and then the frames will be normalized. Frames per second (fps) value is decided to be 20 for the project.

After these preprocessing steps the next step would be to compute the optical flow of the frames. We are using an optic flow-based approach rather than RGB based approach because it has shown better results in past works. Moreover, as we aim to develop a system which can be installed in surveillance cameras for the detection of fighting, it is essential to keep the computational cost as low as possible. Therefore, instead of detecting the human body part like arms, wrists, etc. to detect fights, we prefer to use a motion analysis-based approach where we capture the rapid motion region using optic flow. The main task optic flow performs is motion analysis. It detects the areas in the frames where there is higher motion and then highlights those areas. For each frame we compute its optic flow and then this flow will go as input into the feature extraction model which is the next step in the pipeline. The optic flow method requires more computing power than normal so for surveillance purposes where such computing power is not available we used the approach without optic flow.

For the fire detection the model is trained on the image dataset of fire.The input to model is the image or frame of a fire/no fire video. The main preprocessing done is the resizing of the frames height and width to a standard size of 64 each and then the frames will be normalized.

B. Feature Extraction Model (Transfer Learning)

In this step we are using the transfer learning approach. A pre-trained CNN model called mobilnet is used to extract the features from the optic flow frame given as input to the model. Mobilnet is trained on the readily available and popular dataset ImageNet. Mobilnet has proven to be the most effective model in our research work. We previously used simple CNN (not pre-trained) for feature extraction, but it did not prove to be as effective as mobilnet.

Mobilnet on the other hand improved our accuracy by around 8%. A frame is first passed on to the time-distributed 2D convolution layer where the relu activation function is used. The spatial feature of the frame is extracted at the convolutional layer with the induction of the time distributed wrapper layer; we were able to apply the same layer of CNN to each frame independently. Then max pooling is applied to the resulting vector obtained from the convolution layer. This process will be repeated several times and through experiments we will determine how many times convolution and max pooling should be done. After convolution and max pooling is done the vector of extracted features is finally flatten into a 1D array which is then passed on to the next layer in the pipeline i.e., classification layer.

For fire detection, after the first Step preprocessed images are passed to CNN that is trained on our dataset of images. Using the above distributed wrapper layer, we were able to apply the same number of layers on each image. Then this CNN will extract the features images and then this features vector will be converted to a 1D array that will be passed for the predictions.

C. Classification Model

In this step the vector of extracted features obtained from MobileNet is processed by a classification model to predict the output of the frame. At each time-step spatial features are fed to the classification model to ensure the spatiotemporal sequence modeling. For classification we have proposed to use BiLSTM due to its ability to capture the time sequence patterns without suffering from the problems like vanishing gradient in case of a simple RNN. BiLSTM makes use of logic gates and has this ability to forget information that is old or irrelevant by using the forget unit. We will use 32 layers in the BiLSTM model. A flattened feature vector from CNN is passed to BiLSTM and a prediction is made for each individual frame separately at each time step by BiLSTM. Detection of fight from videos will require us to capture both spatial as well as temporal features so by making use of CNN and then BiLSTM we will be able to capture both these features of the frames.

The predictions of each individual frame by BiLSTM are then carried out to the final step in our pipeline i.e., the output layer.

For the prediction of fire or no fire we use the spatial and temporal features collected using CNN. After separation of

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 9, Issue 05, ISSN No. 2455-2143, Pages 21-30
Published Online September 2024 in IJEAST (http://www.ijeast.com)

features this feature is passed to two dense layers that will analyze the information.

D. Output Layer

The main purpose of this layer is to produce a single label prediction of the entire video. We use the predictions done at each time step by BiLSTM to predict the final label of the video. In the output layer we apply a dense layer with the softmax activation function. Softmax computes the probability for each time-step prediction. We then take the average of these probabilities across all frames and choose the label which is the most probable one.

In this way, the previous individual frame predictions are combined to make a one final prediction of the whole video which is the final output of our system.

For fire use cases, the prediction of a fire video is done by separating the frames of the video and then doing a prediction on each of the frames separately. These predictions are then combined to make one final prediction for the whole video.

C. Architectural Design of LRCN approach

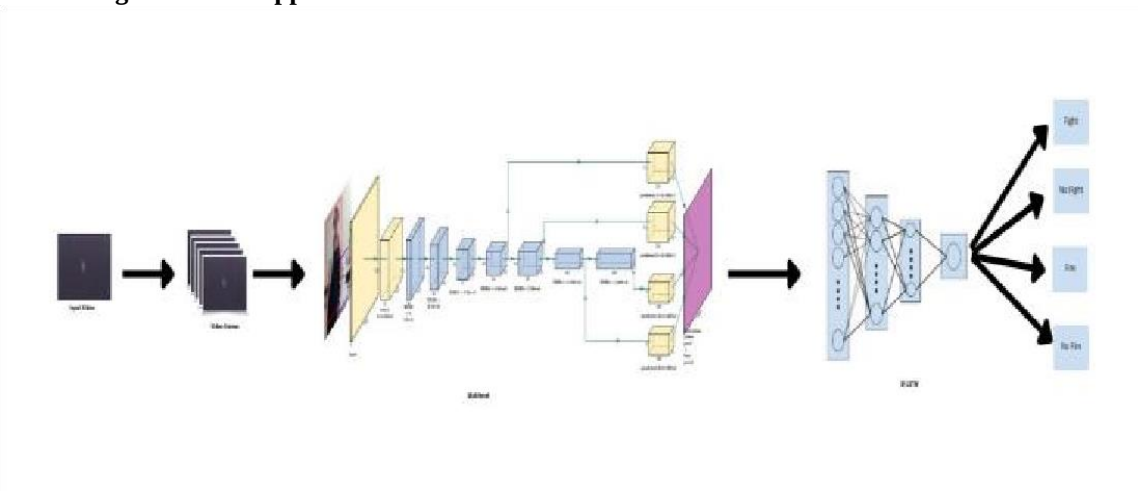

Figure 1
https://nitheshsinghsanjay.github.io/

E. Conclusion

In the above-mentioned methodology, we proposed to process the video frame by frame and compute optic flow of each frame.These are then passed to MobileNet where they are processed independently to extract spatial features. Max pooling is then applied, and the features vector is flattened to 1D array. These vectors of features are passed to BiLSTM at each time step where prediction is made for each frame. These predictions are then combined after applying a dense layer with softmax to get a final single output label for the entire video.

For fire detection somewhat similar CNN architecture is used with convolution and max pooling layers applied three times followed by flattening and dense layer.

There is only one main module which will receive video frames as input and will predict where fighting is present in the video or not.

IV. EXPERIMENTATION AND RESULTS

In our experiment, dataset was split in 80 % for training and 20% for testing. The model is trained with callback mechanisms to decide on optimal number of epochs. Batch size of 8 is used. Each video's width and height were set to 64 x 64 and 20 frames from each video were picked.

Dataset was trained with 4 different approaches:

- Without optical flow and without transfer learning
- Without optical flow and with transfer learning
- With optical flow and without transfer learning
- With optical flow and with transfer learning
- CNN based approach for Fire Use Case

For transfer learning Mobilnet pre trained on imagenet dataset is used.

A. Dataset used:

The first and foremost step of building any neural network is the collection of datasets which fits well to the requirements and satisfies the constraints of quantity and quality.

Dataset was one of the main challenges we faced while implementing both the use cases. For fighting we needed a dataset that depicts real life surveillance footage and can adhere to the varying levels of light, surroundings, and crowd. For the fire use case we could not find any video dataset, so we shifted towards using the image dataset.

In our experiment. For fire case we have trained our model on the dataset of images. We currently have collected a data set

of 925 images out of which 432 contain fire and 493 do not contain the fire. We obtained dataset from different sources which include Kaggle, research papers and our own custom dataset.We have collected this dataset from github, kaggle. The total dataset for the fighting case we used to train our model contained 2480 videos (1240 of fighting and 1240 of nonfighting).

We obtained dataset from the following source:

- "Surveillance fight videos" dataset [7] which comprises 300 videos of 2 seconds each out of which 150 videos are fight videos and the other 150 videos are no fight videos.
- 2000 videos from Kaggle [11]
- Our own dataset of 500 videos

B. Our own Custom Dataset:

Since our aim is to develop a system which can work for surveillance footage in real time, we needed a dataset which has variable surrounding environments and factors in its videos. The surveillance fight videos dataset provides videos which are taken in varying backgrounds of day and night, crowded areas or deserted locations, varying locations such as streets, restaurants and even airplanes. So, such a dataset can be crucial for us in developing our system.

But this dataset lacks the quantity of videos we need to train our model well so we made our own dataset by searching for videos on YouTube with terms like violence, fight etc. This dataset was then trimmed to fit our model's requirement.

C. Results:

Approach	Accuracy
Optic Flow and Transfer Learning (Fighting Detection)	82.66%
Optic Flow and No Transfer Learning (Fighting Detection)	77.42%
No Optic Flow and No Transfer Learning (Fighting Detection)	75.4%
No Optic Flow and Transfer Learning (Fighting Detection)	84.68%
Fire Detection (CNN Approach)	84%

TABLE II: LITERATURE REVIEW SUMMARY

V. CONCLUSION AND FUTURE WORK

For our fight use case one of the main challenges was to get the quality dataset in abundance. We not only acquired dataset from various sources like research papers, github, kaggle etc. but also made our own dataset of 500 videos (taken from youtube surveillance videos). We studied various research papers to find solutions and approaches for similar problems. Many different techniques were explored, however by our research we decided to use CNN and BI-LSTM approach. CNN will be used for extracting features from the video frames and BI-LSTM will be used for classification.

With the above-mentioned architecture of the model, we were not able to achieve a good accuracy, so we explored further ways to get better results from our model. We used a preprocessing computer vision technique known as the optical flow to improve our results. We also experimented with frame padding. We also decided to use a transfer learning approach where we used the MobileNet model (trained by Google) in our model. The above proved to be a breakthrough for us in achieving a good accuracy.

For the fire use case the main challenge we faced was the dataset again. We trained our model on the images dataset we found on Kaggle and although we got a decent accuracy using this dataset, there is still room for much improvement in this domain. There is not much video dataset available to implement fire. We tried to look for the videos dataset for fire from different sources but could only find around 40 videos in total.

This leaves us with the line of work for the future to further improve the availability of an adequate amount of video dataset for fire detection to get better results than currently achieved. The fire case is of tremendous importance given the recent forest fire events, there is a need to develop a model for early detection of fire to avoid huge forest and countryside fires.

From the industrial and application point of view, the installation or integration of these models with the surveillance cameras is also one essential task for the future to utilize the work that has been done in this domain.

VI. REFERENCES

[1] de la Calle Robles, Luis. (2007). Fighting for Local Control: Street Violence in the Basque Country, International Studies Quarterly, 51(2), 431–455. doi:10.1111/j.1468-2478.2007.00458.x.

[2] Badger, Stephen G. (2021). NFPA's "Large-Loss Fires in the United States", National Fire Protection Association, December 2021.

[3] Sultani, W.; Chen, C.; Shah, M. (2018). Real-world Anomaly Detection in Surveillance Videos, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[4] Carneiro, S. A.; da Silva, G. P.; Guimaraes, S. J.; Pedrini, H. (2019). Fight Detection in Video Sequences Based on Multi-stream Convolutional Neural Networks, 32nd SIBGRAPI Conference on Graphics, Patterns and Images, 2019.

[5] Sudhakaran, S.; Lanz, O. (2017). Learning to Detect Violent Videos Using Convolutional Long Short-Term Memory, 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017.

[6] Mukherjee, S.; Saini, R.; Kumar, P.; Roy, P. P.; Dogra, D. P.; Kim, B.-G. (2017). Fight Detection in Hockey

Videos using Deep Network, Journal of Multimedia Information System, 4(4), 225–232, December 2017.

[7] Akti, S.; Tataroglu, G. A.; Ekenel, H. K. (2019). Vision-based Fight Detection from Surveillance Cameras, Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), 2019.

[8] Perez, M.; Kot, A. C.; Rocha, A. (2019). Detection of Real-world Fights in Surveillance Videos, ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing, 2019.

[9] Hassner, T., Itcher, Y., & Kliper-Gross, O. (2012). Violent Flows: Real-time Detection of Violent Crowd Behavior. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 1–6. doi:10.1109/CVPRW.2012.6239236.

[10] Zhang, Y., Qiao, Y., & Wang, Q. (2016). Real-time Action Recognition with Enhanced Motion Vector CNNs. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2718–2726. doi:10.1109/CVPR.2016.298.

[11] Pundir, R., Rani, R., & Yadav, J. (2020). An Efficient Approach for Violence Detection in Surveillance Video using 3D Convolutional Neural Networks. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 279–284. doi:10.1109/ICESC48915.2020.9155727.

[12] Zhou, S., Jiang, H., Cheng, Y., & Sun, W. (2017). Violence Detection in Surveillance Video Using Low-Level Descriptors. IEEE Access, 5, pp. 24268–24278. doi:10.1109/ACCESS.2017.2763739.