Project Report On

Hybrid CNN-SVM model for Action Recognition from Videos

"A dissertation submitted in partial fulfillment of the requirements of 8th Semester 2022 Project-I (CS-794) examination in Computer Science and Engineering of the Maulana Abul Kalam Azad University of Technology"

Submitted by

Uttam Modi (10200119058)
Rahul Pramanik (10200119061)
Anik Mitra (10200119064)
Pratik Tamang (10200119066)

Under the guidance of

Dr. Kousik Dasgupta
Computer Science and Engineering
Kalyani Government Engineering College

Department of Computer Science and Engineering
Kalyani Government Engineering College
(Affiliated to Maulana Abul Kalam Azad University of Technology, West Bengal)
Kalyani - 741235, Nadia, WB

কল্যাণী গভঃ ইঞ্জিনিয়ারিং কলেজ
KALYANI GOVERNMENT ENGINEERING COLLEGE
(GOVERNMENT OF WEST BENGAL)
KALYANI, NADIA, PIN - 741 235, WEST BENGAL

Memo No. :

Date :

Certificate of Approval

This is to certify that this report of B. Tech 8th Sem, 2022 project, entitled "**Hybrid CNN-SVM model for Action Recognition from Videos"** is a record of bona-fide work, carried out by **Uttam Modi, Rahul Pramanik, Anik Mitra , Pratik Tamang** under my supervision and guidance.

In my opinion, the report in its present form is in partial fulfillment of all the requirements, as specified by the **Kalyani Government Engineering College** and as per regulations of the **Maulana Abul Kalam Azad University of Technology**. In fact, it has attained the standard necessary for submission. To the best of my knowledge, the results embodied in this report are original in nature and worthy of incorporation in the present version of the report for the Project-I (CS-794) 7th Sem B. Tech programme in Computer Science and Engineering in the year 2022.

Guide / Supervisor

Dr. Kousik Dasgupta
Computer Science and Engineering
Kalyani Government Engineering College

Examiner(s)

Head of the Department
Computer Science and Engineering
Kalyani Government Engineering College

Table Of Contents

Abstract:

Recent advancements in automatic analysis of static images and greater processing power have made action recognition in videos the focus of scientific research. Computer vision researchers have become increasingly interested in the task of identifying human actions in video as a consequence of its importance in various applications. We analyze the key models and methods for action recognition that include joint trajectory estimations, silhouette matching, and spatiotemporal local descriptor estimations based on human models. With the hybrid model, we combine the key features of both CNN and SVM models to solve the problem in an integrated way. Moreover, the work proposes an action recognition model from the perspective of individual applications. With the help of expert knowledge, we propose action semantic models that can deal with compound actions and activities. A domain specific database is generally used to build and test models and methods for action recognition, since the task is domain dependent. The work presented here utilizes a recent video dataset that we have created for developing action recognition methods. Additional modalities were considered to help improve our methods.

Introduction:

Action recognition is generally formulated as a task that aims to honor the primary mortal action being performed within a given videotape. Despite the constraint that similar vids substantially feature mortal conduct. It's integral to achieving general videotape understanding and benefits numerous downstream videotape understanding operations, similar as videotape reclamation, tone- driving systems, surveillance, robotics, etc. In the proposed work we've anatomized and proposed way to descry conduct for surveillance.

Literature survey:

In paper(1) explains how the author developed a deep convolutional network armature for detecting mortal conduct in vids by using the action bank features of the UCF50 database.

In paper(2) developed 3D CNN models that combine spatiotemporal features with spatial features in order to fete conduct. It's necessary for the armature to be suitable to induce multiple channels of information from conterminous input frames in order to perform complication and subsampling independently. Data from all channels is consolidated to gain the final point depiction.

In paper(3) analyzes mortal action recognition as being a grueling task due to both the spatial and temporal confines of the action data. As a result for the unpredictability of the mortal body, it's possible to insulate it into different corridor, grounded on the positions of the cadaverous joints, and perform element recognition grounded on the part- grounded descriptions.

In paper(4) In this work, videotape conception discovery ways are presented by using Support Vector machines(SVM) and Convolutional Neural Networks(CNN). A set of low- position visible features are of greatly lower size and also proposes effective union of SVM and CNNs to ameliorate conception discovery, where the being CNN toolkits can abstract frame position static descriptors. originally SVM is developed using global features likeHue_moments and HSV histogram uprooted from crucial- frame. CNN is developed using uprooted crucial-frames from vids. The two classifiers SVM and CNN are independently trained on data set and this enriches effective results. The delicacy of each classifier is collectively calculated. The emulsion of two classifiers is performed to efficiently descry the generalities in the test dataset. After the emulsion of two classifiers, the delicacy is calculated. The proposed frame using emulsion of SVM and CNN gives effective videotape conception discovery. The proposed frame is validated on standard UCF 101 dataset using delicacy as prophetic measure. The emulsion of CNN and SVM classifiers provides better results in comparison with individual

classifier

In paper(5) the author concentrated on totalities up the ongoing advancement of exertion acknowledgment from the launch. At that point dependent on the Hierarchical Filtered stir model and Nearest Neighbor classifier, do exertion acknowledgment exercising Histogram of acquainted slants(overeater) included in videotape groupings of colorful pretensions. Then they use KTH dataset for preparing and MSR exertion dataset s for testing. The examination shows that the new element birth process is feasible and has better prosecution in thecross-dataset exertion acknowledgment.

In paper(6) the author presents a videotape depiction dependent on thick directions and movement limit descriptors. Directions catch the near movement data of the videotape. A thick depiction ensures a decent addition of anterior area movement just as of the encompassing setting. A stylish in class optic sluice computation empowers a hearty and complete birth of thick directions.

In paper(7) In the paper, the problem of automatic bracket for different types of videotape is studied, the automatic bracket model of videotape content MPEG- 7 visual descriptors and support vector machine (SVM) are proposed, and gives the system frame design and the concrete process design. In this chapter, the design system of the SVM multi classifier is bettered. The new 1- 1 system is grounded on 1- 1 system, in order to ameliorate the bracket delicacy of the system. The bracket trials of the common five kinds(variety, education, machine, life and technology) of videotape are compared with the experimental results of the same type. The effectiveness and feasibility of the proposed algorithm are proved.

In paper(8) the author delved the guess that the characterization of conditioning can be supported by planning a keen element pooling procedure under the generally employed pack- of- words- put together representation. innovated with respect to programmed videotape saliency disquisition, they propose the spatial-worldly consideration aware pooling plan for highlight pooling.
CNN can learn suitable highlights by them naturally, which prompts great item acknowledgment and grouping perfection. The rest of the papers are organized as follows Section III describes the proposed system, Section IV includes the dataset medication, Section V includes the perpetration way and Section VI describes the result and discussion. compass for farther exploration in this work is detailed in the final section .

In paper(9) the author critically reviews different approaches and styles in videotape bracket with their advantage, finding, limitations, challenges, data summary, exploration gaps, and performance. From the analysis of this paper. It's concluded that a videotape- grounded approach for videotape bracket works more over textbook

and audio. The least employed process of videotape bracket becomes textbook birth. In different operations audio and video features lines are used, but as we can appreciate, also the performance of the bracket tasks can be more enhanced if the birth of both visual and audio features is taken with equal significance in the collection of videotape features. Audio- grounded results need little computing source.

In paper(10) the author a simple convolutional neural network on videotape Bracket has been made. This network gives lower computational cost. This is a veritably shallow network that gives a good delicacy rate. trials by varying literacy rate and Ages have great influence on the network's performance. The temporal features of videotape Sequences are important rather than the features in static images and this can be extended in the unborn work. In unborn our system will be added with different stride, padding, and more figures of complication layers to different optimizers and break the bracket problem in order to increase delicacy and performance.

Proposed Work:

This section details the proposed work using a hybrid CNN and SVM model. The proposed hybrid model combines the key properties of both the classifiers. While developing the novel model we had worked on a CNN with LSTM model for initial detailed understanding of the problem domain and its solution. The discussion describes the related prerequisite of our proposed work and then details both the models.

A. Dataset

UCF101 is an action recognition data set of realistic action videos, collected from YouTube, having 101 action categories. This data set is an extension of UCF50 dataset which has 50 action categories. With 13320 videos from 101 action categories, UCF101 gives the largest diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc, it is the most challenging data set to date.

Preparation of Domain specific Dataset:

The UCF101 has been considered in the work as an action recognition dataset of realistic action vids, collected from youtube, having 101 action orders. It consists of vids of 101 action orders. Split the dataset into train and test.(From UCF101 website) and maintain a CSV train for holding information to prizeframes.ThenSub-sampled to 40 frames for each videotape .

Visualize the data with it's labels:

We visualized the data along with labels to get an idea about what we will be dealing with. For visualization, we picked 20 random categories from the dataset and a random video from each selected category and will visualize the first frame of the selected videos with their associated labels written. This way were be able to visualize a subset (20 random videos) of the dataset.

Fig 1: Some sample video from the generated dataset for visualisation

Preprocessing the Dataset:

The dataset required some preprocessing. For this we read the video files from the dataset and resize the frames of the videos to a fixed width and height, to reduce the computations and normalize the data to range [0-1] by dividing the pixel values with 255, which makes convergence faster while training the network. The number of frames of a video that will be fed to the model as one sequence will be 20 so that we get the whole idea of what

is happening in the video.

Feature Extraction:

Feature extraction normally refers to the process of extracting features (informative characteristics) from a frame in a video, independently of past or future frames. It is very similar to extracting features in a static image. There exist methods that use several frames to extract more stable features or something like that. This process, for example, is usually (not exclusively) done to initialize an object tracking method to recognize the object. Feature tracking is the process of "following" or tracking some features from frame to frame. Normally different techniques are used to leverage the knowledge of the position of the features in the previous frame. A basic version of CNN as detailed below has been used for feature extraction.

Splitting Training and Testing Datasets:

The split is a technique for evaluating the performance of any Deep Learning & Neural Network Model. The procedure involves taking the dataset and dividing it into two subsets. The first subset is used to fit train the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

For the proposed work we have considered three sets of Train-Test ratios as detailed below. This has been to negate any discrepancy in results and bring variations.

- Train: 80%, Test: 20%
- Train: 70%, Test: 30%
- Train: 60%, Test: 40%

B. A proposed Convolution Neural Network with LSTM model for action recognition in video

An effort was made initially in the project to solve the problem using a version of CNN and the LSTM network, CNN has been used for feature extraction and the LSTM network is used for classification. The proposed models are described in the section below.

Convolutional Neural Network (CNN):

In the era of technology, the quickly growing demand for learnable machines for the resolution of several advanced issues, deep learning has evolved itself as the vicinity of interest to the researchers within the past few years. As researchers and engineering students tend to mimic human behavior, a serious question arises How do humans acquire knowledge? The solution to the present question is a vital ability of humans i.e. learning, that

has to be incorporated in machines, therefore the term machine learning was coined. Machine learning guarantees to scale back the efforts by creating the machines to learn themselves through past experiences victimization 3 approaches of learning specifically, learning beneath direction, while not direction and semi-supervised learning.The conventional machine learning techniques would like feature extraction because of the necessity, and this needs a site skilled. moreover, the choice of acceptable options for a given downside may be a difficult task. Deep learning techniques overcome the matter of feature choice by not requiring pre-selected options however extracting the numerous options from raw input mechanically for a retardant in hand. Deep learning model consists of a group of process layers that will learn numerous options of information through multiple levels of abstraction. Multiple levels permit the network to be told distinct options. Deep learning has emerged as an approach for achieving promising ends up in numerous applications like image recognition, speech recognition, topic classification, sentiment analysis, language translation, linguistic communication understanding, signal process, face recognition, prediction of bioactivity of tiny molecules, etc. There are unit completely different deep learning architectures like deep belief networks, continual neural networks, convolution neural networks etc.

Convolution Neural Network (CNN), typically referred to as ConvNet, has a deep feed-forward design and has astonishing ability to generalize in a very higher manner as compared to networks with totally connected layers. describes CNN because the construct of class-conscious feature detectors is in a very biologically galvanized manner. It will learn extremely abstract options and might establish objects with efficiency. The substantial reasons why CNN is taken into account higher than alternative classical models area unit as follows. First, the key interest for applying CNN lies within the plan of victimization construct of weight sharing, thanks to that the number of parameters that want coaching is well reduced, leading to improved generalization. thanks to lesser parameters, CNN is often trained swimmingly and doesn't suffer overfitting.Secondly, the classification stage is incorporated with the feature extraction stage, each uses a learning method. Thirdly, it's way more troublesome to implement giant networks victimization general models of artificial neural networks (ANN) than implementing in CNN. CNN's are units wide being employed in numerous domains thanks to their outstanding performance like image classification, object detection, face detection, speech recognition, vehicle recognition, diabetic retinopathy, facial features recognition, and plenty of a lot of. The motivation of this study is to ascertain a theoretical framework whereas adding to the data and understanding regarding CNN. the aim of this study is to gift the merger of the elementary principles of CNN and supply the main points regarding the final model, 3 commonest architectures, and learning algorithms. a replacement learning technique, ADAM projected by has conjointly been elucidated. additionally, to it, it computes the learning rate for each individual parameter. the whole structure of the sections is as follows. Section one offers the introduction and provides the aim of the study. Section two describes the final model of CNN in conjunction with all its elementary ideas. Section three introduces numerous architectures of CNN. Section four portrays the educational algorithmic rule and Section five illustrates the conclusion and future scope.

A convolutional neural network, or CNN, maybe a deep learning neural network designed to process structured arrays of data like pictures. Convolutional neural networks square measure wide utilized in pc vision and became the state of the art for several visual applications like image classification, and have conjointly found success in language process for text classification.

Convolutional neural networks square measure excellent at memorizing and feature extraction tool patterns within the input image, like lines, gradients, circles, or perhaps eyes and faces. it's this property that produces convolutional neural networks thus powerful for pc vision. in contrast to earlier pc vision algorithms, convolutional neural networks will operate directly on a raw image and don't would like any preprocessing.

A convolutional neural network may be a feed-forward neural network, typically with up to twenty or thirty layers. the facility of a convolutional neural network comes from a special reasonably layer known as the convolutional layer.

Convolutional neural networks contain several convolutional layers stacked on high of every alternative, all capable of recognizing additional subtle shapes. With 3 or four convolutional layers it's the potential to acknowledge written digits and with twenty-five layers, it's the potential to differentiate human faces.

The usage of convolutional layers in an exceedingly convolutional neural network mirrors the structure of the human cortical region, wherever a series of layers method associate degree incoming image and determine high-end features, incoming image and identify progressively and more complex features.

Convolution Neural Network General Model:

1. General Model

The simple Artificial neural network model has a single input and output layer along with multiple hidden layers. A particular neuron takes input vector X and produces output Y by performing some calculation F on it, denoted by the general equation given below.

$$F(X, W) = Y$$

Where W represents the weight vector which means the strength of interconnection between neurons of two side-by-side layers. The obtained weight vector can be now used to perform image classification. A significant amount of papers exist related to the pixel-based classification of images. However, contextual information like the shape of the image produces better results or outperforms. CNN is a model that is gaining attention because of its classification capability based on contextual information. The general model of CNN has been described

below in figure 1. A general model of CNN consists of four components namely (a) convolution layer, (b) pooling layer, (c) activation function, and (d) fully connected layer. The functionality of each component has been illustrated below.

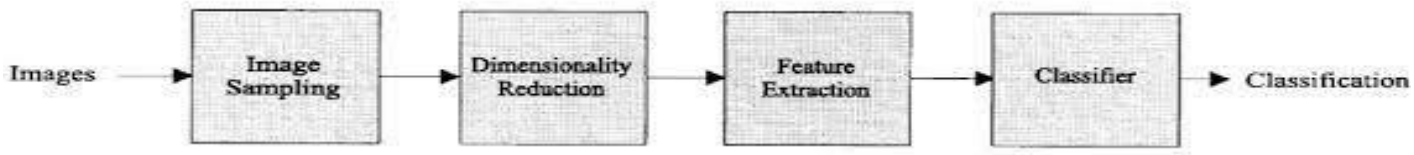


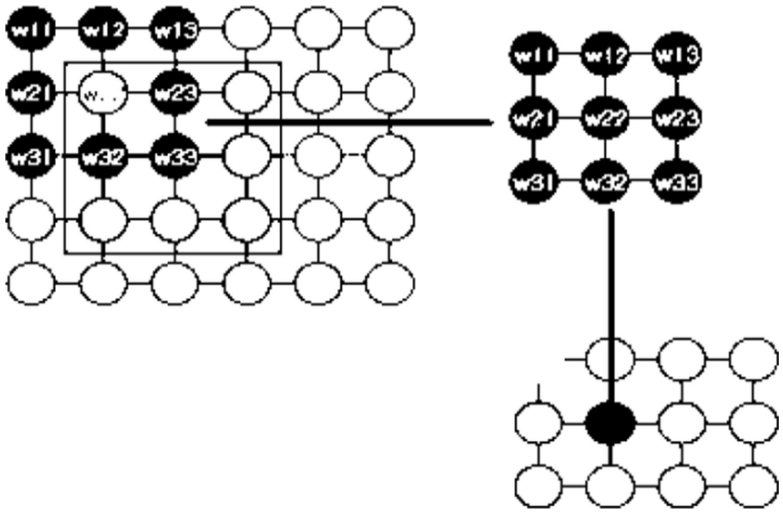
Fig 2: A basic model of CNN

2. Convolution Layer

An image to be categorized is provided to the input layer and therefore the output is that the foreseen class label computed mistreatment extracted options from the image. a private vegetative cell within the next layer is connected to some neurons within the previous layer, this native correlation is termed as receptive field. The native options from the input image are extracted by employing a receptive field. The receptive field of a vegetative cell associated with the explicit region in the previous layer forms a weight vector, that remains equal in the slightest degree points on the plane, wherever the plane refers to the neurons within the next layer. because the neurons within the plane share the same weights, so similar options occurring at totally different locations within the input file are often detected. This has been pictured within the figure shown below

The weight vector, additionally called filter or kernel, slides over the input vector to come up with the feature map. This methodology of slippy the filter horizontally furthermore as vertically is named convolution operation. This operation extracts the N range of options from the input image during a single layer representing distinct options, resulting in N filters and N feature maps. thanks to the development of native receptive fields, the quantity of trainable parameters is considerably reduced. The output within the next layer for location (i,j), is computed when applying convolution operation mistreatment the formula given as shown below:

where, X is the input provided to the layer, W is the filter or kernel that slides over the input, b is that the bias, representing the convolution operation, and σ is nonlinearity introduced within the network.



3. Pooling Layer

The exact location of a feature becomes slighter once it's been detected. Hence, the convolution layer is followed by a pooling or sub-sampling layer. the key advantage of exploiting the pooling technique is that it remarkably reduces the number of trainable parameters and introduces translation unchangeability. To perform a pooling. the operation, a window is chosen and therefore the input parts lying therein window square measure tried and true a pooling operate as shown within the figure. The exact location of a feature becomes slighter once it's been detected. Hence, the convolution layer is followed by a pooling or sub-sampling layer. the key advantage of exploiting the pooling technique is that it remarkably reduces the number of trainable parameters and introduces translation unchangeability. To perform a pooling. the operation, a window is chosen, and therefore the input parts lying therein window square measure tried and true a pooling operate as shown within the figure.

The pooling operation generates another output vector. There exist a few pooling techniques like average pooling and max-pooling, out of that max-pooling is the most typically used technique that reduces map size terribly significantly. whereas computing errors, the error isn't back-propagated to the winning unit as a result of it doesn't participate in forwarding flow.

4. Fully Connected Layer

It is the same as the totally connected network within the standard models. The output of the primary section (includes convolution and pooling repetitively) is fed into the totally connected layer, and the inner product of the weight vector and input vector is computed so as to get the final output. Gradient descent, conjointly called batch mode learning or offline formula, reduces the value performed by estimating the value over a complete coaching dataset and updates the parameters solely when one epoch, wherever associate epoch corresponds to traversing the whole dataset. It yields world minima however if the dimensions of the coaching dataset are giant, the time needed to coach the network well will increase. This approach of reducing the value performed was replaced by random gradient descent.

5. Activation Function

A vast number of research papers exist that use sigmoid and tanh etc activation functions within the typical machine and deep learning algorithms. In order to introduce non-linearity, the use of corrected long measure (ReLU) has tested itself higher than the previous, one because of 2 major factors. First, the calculation of partial derivatives of ReLU is straightforward. Second, whereas considering coaching time to be one in every the issue, the saturating nonlinearities like sigmoid portrayed by. are slower than non-saturating nonlinearities like ReLU. Third, ReLU doesn't enable gradients to disappear. however, the potency of ReLU deteriorates once an oversized gradient is flowing through the network, associate degreed an update in weight causes the vegetative cell to not get activated resulting in a Dying ReLU drawback that could be an extended issue that's typically caused.

Activation perform defines the output of input or set of inputs or in different terms defines node of the output of node that's given in inputs.

Why will we want Activation Functions?

Without activation function, weight and bias would solely have a linear transformation, or neural network is simply a statistical regression model, an equation is polynomial of 1 degree solely that is straightforward to unravel however restricted in terms of ability to unravel complicated issues or higher degree polynomials.

But opposite it, the addition of activation perform to neural network executes the non-linear transformation to input and build it capable to unravel complicated issues like language translations and image classifications.

Types of Activation Functions

We have divided all the essential neural networks into 3 major parts:

A. Binary step function
B. Linear function
C. Nonlinear activation function

A. Binary Step Neural Network Activation function

1. Binary Step function

This activation function terribly basic and it involves the mind on every occasion if we tend to attempt a certain output. it's essentially a threshold base classifier, in this, we tend to decide some threshold worth choosing output that vegetative cell ought to be activated or deactivated.

f(x) = one if x > zero else zero if x < zero

An image highlights the binary step function in neural network. Analytics Steps, analytics steps

Binary step function this, we tend to decide the edge is worth zero. it's terribly easy and helpful to classify binary issues or classifiers.

B. Linear Neural Network Activation function

2. Linear function

It is an easy line activation function wherever our function is directly proportional to the weighted total of neurons or input. Linear activation functions square measure higher in giving a good variety of activations and a line of a positive slope could increase the firing rate because the input rate will increase.

In binary, either a vegetative cell is firing or not. If you recognize gradient descent in deep learning then you'd notice that during this performance spinoff is constant.

$$Y = mZ$$

Where spin-off with regard to Z is constant m. This means the gradient is additionally constant and it's nothing to try and do with Z. In this, if the changes created in backpropagation are constant and not smitten by Z thus this can not be sensible for learning.

In this, our second layer is the output of a linear performance of previous layers' input. Wait for a second, what have we tended to learn during this that if we tend to compare all the layers and take away all the layers except the primary associate degree last then conjointly we are able to solely get an output that may be a linear function of the primary layer.

C. Non-Linear Neural Network Activation function

3. ReLU(corrected Linear unit) Activation function

Rectified linear measure or ReLU is most generally used activation function straight away that ranges from zero to time, All the negative values square measure born-again into zero, and this conversion rate is thus quick that neither it will map nor match into information properly that creates a drag, however wherever there's a drag there's an answer.

The graph outlines the variation of the corrected linear measure function in activation functions in a neural network. Rectified linear measure activation function. We use the Leaky ReLU function rather than ReLU to avoid this inappropriate, in Leaky ReLU vary is swollen which boosts the performance.

4. Leaky ReLU Activation function

The variation of the Leaky ReLU function as an associate degree activation function in a neural network is conferred within the image. Analytics Steps Leaky ReLU Activation function. We needed the Leaky ReLU activation function to unravel the 'Dying ReLU' drawback, as mentioned in ReLU, we tend to observe that every one of the negative input values is converted into zero terribly quickly, and within the case of Leaky ReLU we tend to don't build all negative inputs to zero however to a price concerning zero that solves the most important issue of ReLU activation function.

5. Sigmoid Activation function

The sigmoid activation function is employed principally because it will its task with nice potency, it essentially may be a probabilistic approach towards deciding and ranges in between zero to one, thus {when we tend to|once we|after we} have to be compelled to build a choice or to predict associate degree output we use this activation function thanks to the vary is that the minimum, therefore, the prediction would be additional correct.

Highlighting the Sigmoid activation function in the neural networks within the graphical type. Analytics Steps Sigmoid Activation function. The equation for the sigmoid function is

$$f(x) = 1/(1+e(-x))$$

The sigmoid function is primarily termed a vanishing gradient problem that happens as a result of we tend to convert massive input in between the vary of zero to one and thus their derivatives become abundant smaller that doesn't provide satisfactory output. to unravel this drawback another activation performed like ReLU is employed wherever we tend to don't have any low spinoff drawback.

6. Hyperbolic Tangent Activation Function(Tanh)

Hyperbolic Tangent(Tanh) activation function in a neural network and its variation square measure is displayed within the graph. Analytics Steps

7. Softmax Activation function

Softmax is employed primarily at the last layer i.e output layer for deciding a similar as sigmoid activation works, softmax essentially provides worth to the input variable in keeping with their weight, and therefore the total of those weights is eventually one.

Different architectures Of Convolution Neural networks:

Multiple architectures have been developed and implemented in CNN. Brief explanations of that architecture are explained below.

1.LeNet Architecture

The multi-layer networks area unit is appropriate for image recognition tasks thanks to the flexibility to be told from extremely complicated and high-dimensional information. In 1998, a projected associate degree design known as LeNet design that uses a dataset was summarized within the following paragraph. The LeNet5 design has been portrayed within the figure. it's eight layers constituting 5 convolutional layers and 3 totally connected layers. Each unit in a plane has twenty-five inputs. Units within the 1st hidden layer receive input from the five × five space, which could be a part of a whole image so an awfully little region of the input image is passed to the primary hidden layer. This native space of the input image is named the receptive field of the unit. each unit in a very plane shares a similar weight vector. The output of the unit is kept at a similar location within the feature map.

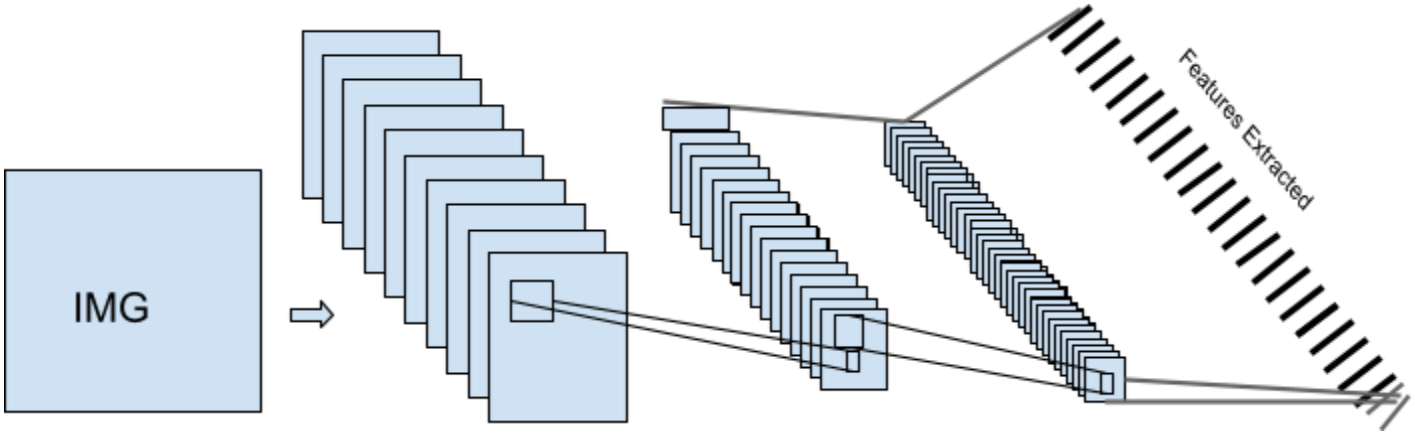


Fig 3- Detailed architecture of CNN used in the model.

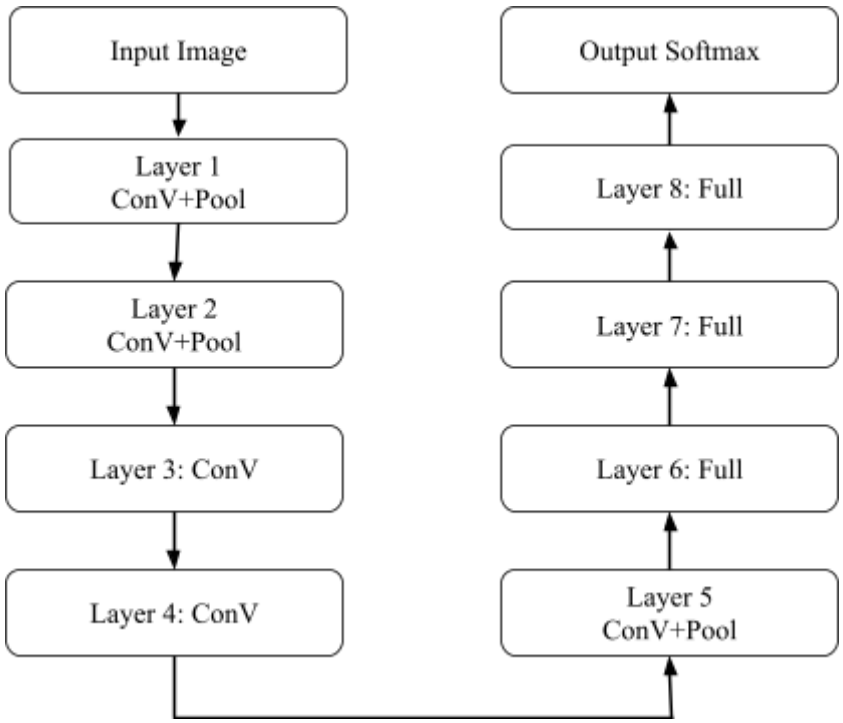
The neighboring units within the feature map square measure the results of the neighboring units within the previous layer. Thus, the result's Associate in the array overlaps contiguous receptive fields. the primary layer could be a convolution layer that consists of neurons that output sigmoid activation applied to the weighted add. As shown within the figure, whereas computing contiguous units within the feature map, if the five × five areas are chosen as array input and a horizontal shift on the realm is performed, it'll end in the overlapping of 4 rows and 5 columns. varied feature maps square measure generated that square measure the results of completely different weight vectors applied to an equivalent input image. completely different options are often extracted from the

feature maps obtained. An important property of CNN is that a small shift within the input doesn't have an effect on the feature map. The preciseness of the position of the feature in a picture isn't crucial, so to cut back the

preciseness price sub-sampling is performed. As shown in figure four, sub-sampling has been delineated within the second layer. A variety of feature maps obtained once sub-sampling square measure an equivalent as those obtained once convolution. Here, within the subsampling layer, 2×2 areas are taken as input and computed the common of the four inputs, multiply it by the trainable constant and add the trainable bias, pass it to the sigmoid. A rise within the range of feature maps is often ascertained because the spatial resolution decreases layer by layer. The training is accomplished with a back propagation technique.

2. AlexNet Architecture

Brief description of a modified variant of LeNet i.e. AlexNet architecture proposed has been summarized in this fragment. It consists of three fully connected and five convolution layers, the outputs received are passed to the 1000-way softmax in order to classify 1.2 million high-resolution images into 1000 distinct classes.



Non-saturating neurons are used along with the implementation of efficient GPUs to train the network at a faster speed. So, to make the network classify objects from millions of images, a large network is required which might ultimately lead to intense demand for training a very large number of weights, leading to the problem of overfitting. To overcome this problem, a dropout method was implemented. In this technique, the neuron that has a probability of 0.5, does not take part in forwarding and backward propagation and is damped. The neurons that are dependent on these damped neurons are compelled to learn the most robust features completely on their own, which reduces overfitting substantially. The number of iterations required to converge is doubled due to the

dropout method. The network uses two GTX 580 3GB GPUs and it takes five to six days to train the network. The major features of this architecture were the introduction of ReLU non-linearity in CNN due to which the convergence rate increased rapidly.

3.GoogleNet Architecture

A model usually called GoogleNet was projected. It absolutely was enlightened when winning the ILSVRC14 competition. the key aim was to develop a model at intervals with a lower budget, that might scale back the facility, variety of trainable parameters used, and memory consumption. The model dead reduced the number of trainable parameters employed in the network. The broad description of the design is as follows. It primarily uses twelve million fewer parameters than the model projected. The design is geared toward building a network that might acknowledge the objects in a picture with additional preciseness. This might probably be achieved by increasing the scale of the network thereby increasing the number of layers, however, a serious disadvantage of this idea was that if the network size is magnified it might increase the number of parameters to be trained therefore resulting in the matter of overfitting. Another major disadvantage is that if the quantity of filters is magnified, the computation additionally will increase resulting in an increase in overhead. The answer provided to that was to implement a distributed matrix. The units that are extremely related to mix to create a cluster within the previous layer and that they offer input to the consequent layer then are accustomed to the best configuration. The non-uniform distributed matrix also can be used however the overhead of cache misses still exists although the computations are reduced up to one hundred times. the employment of extremely tuned numerical libraries acting quicker computations additionally doesn't facilitate. Therefore, this state of art works on uniform distributed matrices.

LONG SHORT-TERM MEMORY(LSTM):

Well on a large variety of problems, and are now extensivelyused.LSTMs are explicitly designed to avoid the long- term reliance problem. Flashing back information for long ages of time is virtually their dereliction not a commodity they struggle to learn!

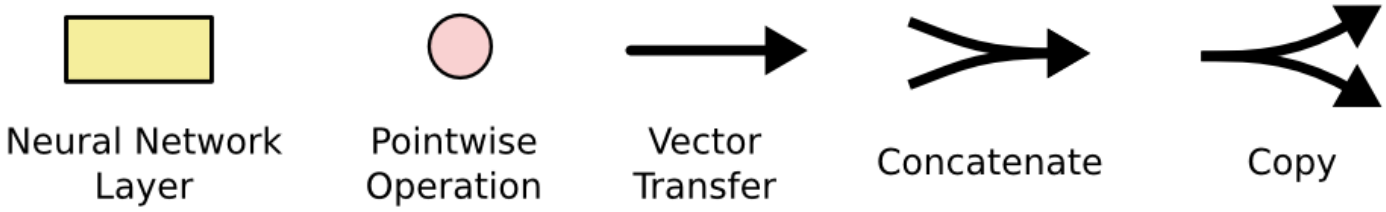
All intermittent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a veritably simple structure, similar as a single tanh subcaste.

Long Short Term Memories also have this chain- like structure, but the repeating module has a different structure rather than having a single neural network caste, there are four, interacting in a truly special way.

We ’ll walk through the LSTM illustration step by step later. For now, let’s just try to get comfortable with the memorandum we’ll be using.

In the below illustration, each line carries an entire vector, from the affair of one knot to the inputs of others. The pink circles represent pointwise operations, like vector addition, while the unheroic boxes are learned

neural network layers. Lines incorporating denote consecution, while a line separating denotes its content being copied and the clones going to different locales.



The Core Idea Behind LSTMs:

The key to LSTMs is the cell state, the horizontal line running through the highest part of the diagram. The cell state is reasonably sort of a conveyer. It runs straight down the whole chain, with just some minor linear interactions. It's terribly simple for info to simply flow on it unchanged.

The LSTM will have the flexibility to get rid of or add info to the cell state, fastidiously regulated by structures referred to as gates. Gates are the way to optionally let info through; they're composed out of a sigmoid neural internet layer and a pointwise multiplication operation.

The sigmoid layer outputs numbers between zero and one, describing what quantity of every element ought to be let through. a price of zero means that "let nothing through," whereas a price of 1 means that "let everything through!" Associate in Nursing LSTM has 3 of those gates, to guard and manage the cell state.

Understanding the working of LSTM:

The first step in our LSTM is to choose what info we're about to throw off from the cell state. This call is formed by a sigmoid layer known as the "forget gate layer." It looks at ht−1 and outputs xt-1 between zero and one for every number within the cell state Ct−1. A one represents "completely keep this" whereas a zero represents "completely get rid of this."

Let's return to our example of a language model making an attempt to predict the succeeding word supported by all the previous ones. In such a drag, the cell state may embrace the gender of this subject, so the right pronouns are often used. Once we see a brand new subject, we wish to forget the gender of the previous subject.

The next step is to choose what new info we're about to store within the cell state. This has 2 components. First, a sigmoid layer known as the "input gate layer" decides the values we'll update. Next, a tanh layer creates a vector of latest candidate values, C~t, that might be further to the state. Within the next step, we'll mix these two

to form an Associate in Array update to the state.
In the example of our language model, we'd need to feature the gender of the new subject to the cell state, to switch the previous one we're forgetting.
It's currently time to update the previous cell state, Ct−1, into the new cell state Ct. The previous steps already set what to try and do, we have a tendency to simply get to truly know.We multiply the previous state by linear unit, forgetting the items we have a tendency to forget earlier. Then we have a tendency to add it∗C~t. This can be the new candidate values, scaled by what quantity we have a tendency to set to update every state worth.
In the case of the language model, this can be wherever we'd truly drop {the info|the knowledge|the data} concerning the previous subject's gender and add the new information, as we have a tendency to set within the previous steps.

Finally, we want to choose what we're about to output. This output supports our cell state, however is a filtered version. First, we have a tendency to run a sigmoid layer that decides what components of the cell state we're about to output. Then, we have a tendency to place the cell state through tanh (to push the values to be between −1 and 1) and multiply it by the output of the sigmoid gate, so we have a tendency to solely output the components we have a tendency to set to.
For the language model example, since it simply saw a topic, it'd need to output info relevant to a verb, just in case that's what's coming back next. as an example, it'd output whether or not the topic is singular or plural, so we all know what type a verb ought to be conjugated into if that's what follows next.
.

Variants on Long Short Term Memory:

What I've delineated to this point may be a pretty traditional LSTM. however not all LSTMs area unit an equivalent because the on top of. In fact, it sounds like virtually each paper involving LSTMs uses a rather completely different version. The variations are unit minor, however its price mentions a number of them.
One common LSTM variant, introduced by Gers & Schmidhuber (2000), is adding "peephole connections." This implies that we have a tendency to let the gate layers check up on the cell state.
The top of the diagram adds peepholes to any or all the gates, however several papers can offer some peepholes and not others.
Another variation is to use coupled forget and input gates. Rather than severally deciding what to forget and what we should always add new data to, we have a tendency to create those choices alone. we have a tendency to solely forget once we're planning to input one thing in its place. we have a tendency to solely input new values to the state after we forget one thing older.

These squares measure solely many of the foremost notable LSTM variants. There square measure legion others, like Depth Gated RNNs by Yao, et al. (2015). There's additionally some utterly completely different approach to braving semipermanent dependencies, like mechanism RNNs by Koutnik, et al. (2014).
Which of those variants is best? Do the variations matter? Greff, et al. (2015) do a pleasant comparison of widespread variants, finding that they're all identical. Jozefowicz, et al. (2015) tested over 10 thousand RNN architectures, finding some that worked higher than LSTMs on certain tasks.

Proposed Hybrid CNN and LSTM model combination:

In this work, a multilayered CNN and Long Short- Term Memory model will be used. To support sequence vaticination the CNN- LSTM armature involves using Convolutional Neural Network(CNN) layers for point birth on input train combined with LSTMs. CNN LSTMs were developed for visual time series vaticination problems and thus the operation of generating textual descriptions from sequences of images. The UCF101 is an action recognition dataset of realistic action videos, collected from youtube, having 101 action orders. It consists of videos of 101 action orders. Split the dataset into train and test.(From UCF101 website) and maintain a CSV train for holding information to prizeframes.ThenSub-sampled to 40 frames for each videotape and it's passed to Inception V3 model. commencement V3 model is a pre-trained model on image- net dataset. Processing blocks are Convolution Layer, Max Pooling, ReLU- Subcaste(for thresholding), leveling and passing it to intermittent Neural Network. LSTM model handed by Keras and passed features to the LSTM model and added many further layers. Finally a subcaste with soft maximum activation gives the prognosticate class.

Implementing the LRCN Approach:

In this step, we are going to implement the primary approach by employing a combination of ConvLSTM cells. A ConvLSTM cell could be a variant of the Associate in Nursing LSTM network that contains convolutions operations within the network. It's Associate in Nursing LSTM with convolution embedded within the design, that makes it capable of distinguishing spatial options of the info whereas keeping under consideration the relation.

For video classification, this approach effectively captures the relation within the individual frames and therefore the relation across the various frames. As a results of this convolution structure, the ConvLSTM is capable of taking in third-dimensional input (width, height, num_of_channels) whereas an easy LSTM solely takes in

1-dimensional input thence Associate in Nursing LSTM is incompatible for modeling Spatio-temporal information on its own.

In this step, we are going to implement the LRCN Approach by combining Convolution and LSTM layers during a single model.we will implement the approach referred to as the semipermanent perennial Convolutional Network (LRCN), which mixes CNN and LSTM layers during a single model. The Convolutional layers are used for spatial feature extraction from the frames, and therefore the extracted spatial options are fed to LSTM layer(s) at every time-steps for temporal sequence modeling. In this fashion the network learns spatiotemporal options directly in Associate in Nursing end-to-end coaching, leading to a strong model.

.

The architecture of CNN+LSTM hybrid model:

To construct the model, we'll use Keras ConvLSTM2D repeated layers. The ConvLSTM2D layer additionally takes within the range of filters and kernel size needed for applying the convolutional operations. The output of the layers is planned within the finish and is fed to the Dense layer with softmax activation that outputs the likelihood of every action class. We will additionally use MaxPooling3D layers to cut back the scale of the frames and avoid spare computations and Dropout layers to forestall overfitting the model on the information.

Next Step is to implement our LRCN design, we've used time-distributed Conv2D layers which can be followed by MaxPooling2D and Dropout layers. The features extracted from the Conv2D layers are then planate victimization of the Flatten layer and can be fed to a LSTM layer. The Dense layer with softmax activation can then use the output from the LSTM layer to predict the action being performed.

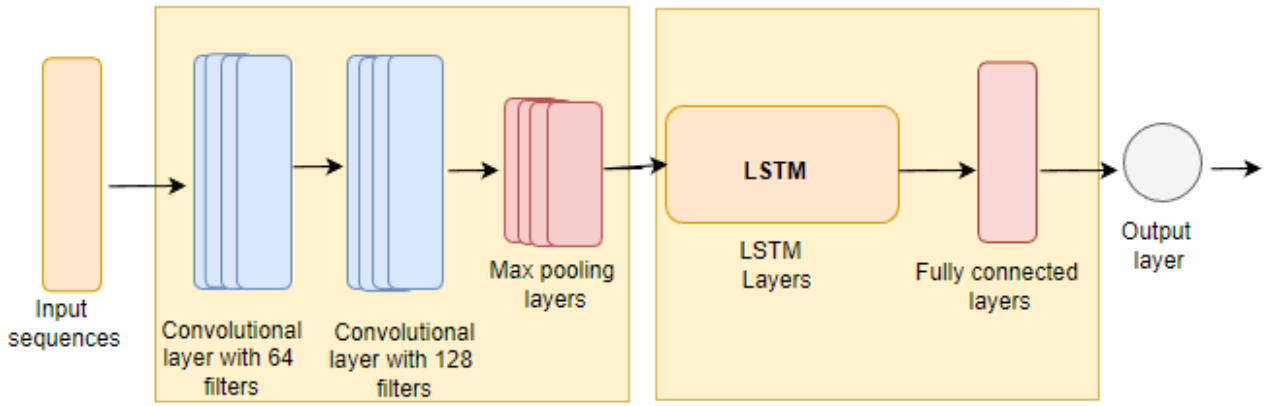


Fig 4. Proposed CNN + LSTM architecture

C. A Novel framework using CNN with Support Vector Machine(SVM) (CNN-SVM) for action recognition

In This approach we have used CNN + SVM. CNN is the same as we used previously.We change LSTM to Support Vector Machine (SVM) algorithm for measuring high accuracy of our project. The SVM aims to detect actions from multi-dimensional dataset in a space where data elements belonging to different classes are separated by a hyperplane. The SVM classifier has the ability to minimize the generalization error on unseen data. Thus, SVM thrives to overcome some of disadvantages of LSTM such as 1) LSTMs take longer to train and require more memory. 2) LSTMs are easy to overfit 3) LSTMs are sensitive to different random weight initializations

Support Vector Machine (SVM)

Support Vector Machine is used for Classification and Regression problems and is one of the most popular Supervised Learning algorithms. The aim of the SVM algorithm is to create the best line that can segregate m-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed a Support Vector Machine.SVMs were integrated within the sampling strategy to simultaneously increase feasible design space coverage, and construct feasibility constraints. It was the black-box simulator that gave rise to binary classification problems from which SVMs can be applied for the determination of optimum separating hyperplanes and is feasible for a given unit design process.The subsequent infeasible design samples were replaced by the samples of feasible designs after a few initial samples were activated by the SVMs.Specifically, SVMs with linear classification functions were used to generate linear feasibility constraints for use in optimization.

Types of SVM :-

- **Linear SVM:** Linear SVM is employed for linearly severable information, which suggests if a dataset is classified into 2 categories by employing a single line, then such information is termed as linearly severable information, and classifier is employed referred to as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is employed for non-linearly separated information, which suggests if a dataset can't be classified by employing a line, then such information is termed as non-linear information and classifier used is named as Non-linear SVM classifier.

Support Vector Machines formulation

Support Vector machines notice the ideas printed on top of. To see why, we want to specify 2 things: the hypothesis areas employed by SVM, and therefore the loss functions used. The lore read of SVM is that they notice any "optimal" hyperplane because the answer to the training downside. The best formulation of SVM is the linear one, wherever the hyperplane lies in the area of the computer file x, during this case the hypothesis area may be a set of all hyperplanes of the form:

$$f(x) = w \cdot x + b.$$

In their most general formulation, SVM finds a hyperplane during an area completely different from that of the computer file x. It's a hyperplane during a feature area induced by a kernel K (the kernel defines an inner product in this area (Wahba, 1990)). Through the kernel K the hypothesis area is outlined as a group of "hyperplanes" within the feature area induced by K. this may even be seen as a group of functions during a Reproducing Kernel metric space (RKHS) outlined by K (Wahba, 1990), (Vapnik, 1998). We tend to not discuss RKHS here and refer the reader to the literature. Thus to summarize, the hypothesis area employed by SVM may be a set of the set of hyperplanes outlined in some area - AN RKHS. This area is formally written as

$$\left\{f : \|f\|_K^2 < \infty\right\}$$

where K is the kernel that defines the RKHS, and $||f||_k^2$ is the RKHS norm of the performance (Wahba, 1990). For instance, for the linear case mentioned on top of, K is that the kernel $K(x1, x2) = x1 \cdot x2$, the functions thought of area unit of the shape $f(x) = w \cdot x + b$, and therefore the RKHS norm of those functions is just the norm of w, namely.

In fact SVM think about subsets of this area, particularly sets of the shape

$$\left\{f : \|f\|_K^2 \leq A^2\right\}$$

for some constant A. within the SLT framework mentioned on top of, the constant A is employed to outline a structure of hypothesis areas (the larger A is, the additional advanced the hypothesis area is). The goal of SVM is to seek out the answer with the "optimal" RKHS norm, that is, to seek out the best A. rather than looking out several hypothesis areas one by one by activity ERM for every alternative of A, SVM searches for AN A (or the best RKHS norm fk2) during a completely different approach, because it are going to be obvious from the SVM formulation conferred below. This "search method" for the best fk2 has been extensively mentioned within the literature (see for instance (Bartlett and Shawe-Taylor, 1998), (Burges, 1998), (Evgeniou et al., 1999)), and that

we don't discuss it here any more. The second alternative is that of the loss performed. For this we've to tell apart between SVM classifiers and SVM regressors. For classification ideally the misclassification error has to be reduced, thus a loss perform of the shape sign(-yf(x)) ought to be used (in classification y takes binary values ±1, and classification is finished by taking the sign of perform f(x)). but thanks to scaling further as process reasons (Vapnik, 1998), the particular loss perform used for SVM classification is |1-yf(x)|+ (that is, zero if 1-yf(x) < zero, and 1-yf(x) otherwise). this is often conjointly referred to as the "soft margin" loss perform thanks to its normal "margin" interpretation: the points that the loss perform is zero area unit those that have "margin"

$$yf(x)/\,||f||_k^2$$

at least $1/||f||_k^2$ (that is 1 - yf(x) ≤ 0 ⇒ yf(x)$||f||_k^2$ ≥ 1 $||f||_k^2$).The margin is a crucial geometric amount related to SVM classification. For additional info we tend to refer the reader to the literature. perform adequate toFor regression the loss perform used is that the questionable epsilon-insensitive loss perform |y-f(x)|ε that is adequate to |y-f(x)|-ε if |y-f(x)| > ε, and zero otherwise. To summarize, following the SLT ideas printed on top of for the given selections of the loss perform and therefore the hypothesis areas, SVM area unit learning machines that minimize the empirical error whereas taking under consideration the "complexity" of the hypothesis area employed by conjointly minimizing the RKHS norm of the answer a pair of K f . SVM in observation minimizes a trade off between empirical error and complexity of hypothesis area. Formally this is often done by resolution the subsequent diminution problems: SVM classification

SVM classification

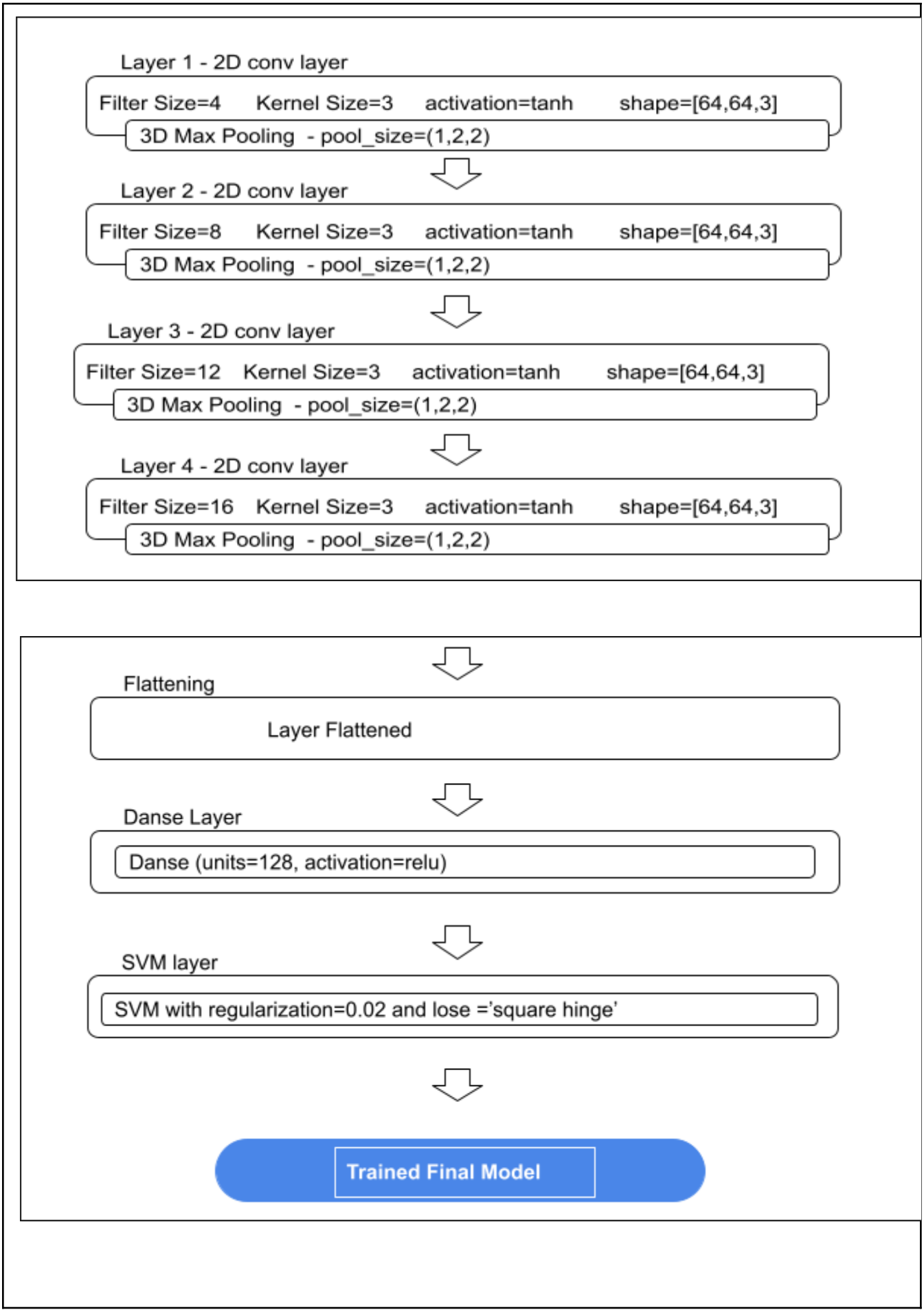
$$\min_f \|f\|_K^2 + C\sum_{i=1}^{l} |\,1 - y_i f(\mathbf{x}_i)\,|_+ \qquad (1)$$

SVM regression

$$\min_f \|f\|_K^2 + C\sum_{i=1}^{l} |y_i - f(\mathbf{x}_i)|_\varepsilon \qquad (2)$$

where C may be a questionable "regularization parameter" that controls the trade off between empirical error and complexity of the hypothesis area used. Having mentioned however SVM stems out of the idea printed on top of, we tend to currently intercommunicate their actual implementation. A consequent section in brief discusses however the diminution issues (1) and (2) is done, taking conjointly under consideration (Trafalis, 1999).

The Architecture of the proposed CNN+SVM model:


Layer 1 - 2D conv layer
Filter Size=4 Kernel Size=3 activation=tanh shape=[64,64,3]
3D Max Pooling - pool_size=(1,2,2)
Layer 2 - 2D conv layer
Filter Size=8 Kernel Size=3 activation=tanh shape=[64,64,3]
3D Max Pooling - pool_size=(1,2,2)
Layer 3 - 2D conv layer
Filter Size=12 Kernel Size=3 activation=tanh shape=[64,64,3]
3D Max Pooling - pool_size=(1,2,2)
Layer 4 - 2D conv layer
Filter Size=16 Kernel Size=3 activation=tanh shape=[64,64,3]
3D Max Pooling - pool_size=(1,2,2)
Flattening
Layer Flattened
Danse Layer
Danse (units=128, activation=relu)
SVM layer
SVM with regularization=0.02 and lose ='square hinge'
Trained Final Model

D. Model Investigation:-

Accuracy - Accuracy is a metric that typically describes how the model performs across all categories. It's helpful once all categories are a unit of equal importance. it's calculated because of the magnitude relation between range|the amount|the quantity} of correct predictions to the full number of predictions.

$$Accuracy = \frac{True_{positive} + True_{negative}}{True_{positive} + True_{negative} + False_{positive} + False_{negative}}$$

Loss - The loss operating during a neural network quantifies the distinction between the expected outcome and therefore the outcome made by the machine learning model. From the loss operation, we {are able to} derive the gradients that are accustomed to update the weights. The typical overall losses constitute the value.
For CNN + LSTM we've got used Categorical Cross-Entropy

The categorical cross-entropy is applied in multiclass classification situations. We have a tendency to add over quite 2 categories.The categorical cross-entropy is acceptable together with associate activation like the softmax which will manufacture many possibilities for the quantity of categories that add up to one.

$$\text{Loss} = -\sum_{i=1}^{\substack{\text{output} \\ \text{size}}} y_i \cdot \log \hat{y}_i$$

where y ^ i is the i-th scalar value in the model output, y i is the corresponding target value, and output size is the number of scalar values in the model output.

For CNN + SVM we have used the **Squared Hinge loss** function.
A special form of value operates that not solely penalizes misclassified samples however conjointly properly classified ones that are inside an outlined margin from the choice boundary. The hinge loss operation is most typically used to regularize soft margin support vector machines.

$$L(y, \hat{y}) = \sum_{i=0}^{N} \left(max(0, 1 - y_i \cdot \hat{y}_i)^2 \right)$$

Precision_Score - The Presicion_Score is calculated because the quantitative relation between range|the amount|the quantity} of Positive samples properly classified to the whole number of samples classified as

Positive (either properly or incorrectly). The exactness measures the model's accuracy in classifying a sample as positive.

$$Precision = \frac{True_{positive}}{True_{positive} + False_{positive}}$$

Recall_Score - The recall is calculated because the magnitude relation between variety|the amount|the quantity} of Positive samples properly classified as Positive to the full number of Positive samples. The recall measures the model's ability to notice Positive samples. the upper the recall, the additional positive samples detected.

$$Recall = \frac{True_{positive}}{True_{positive} + False_{negative}}$$

F1_Score - The F1 score will be taken as a means of exactitude and recall, wherever Associate in Nursing F1 score reaches its best worth at one and worst score at zero. The relative contribution of exactitude and recall to the F1 score are equal. It primarily wants to compare the performance of 2 classifiers.

```
F1_Score = 2 * (precision * recall) / (precision + recall)
```

Action names in the proposed dataset:-

We took the standard training and testing percentage(70 - 30) and 10 classes related to Music from the dataset.
Class_List_Names - ['PlayingCello', 'PlayingGuitar', 'PlayingDhol', 'PlayingFlute', 'PlayingPiano', 'PlayingSitar', 'PlayingTabla','PlayingViolin','PlayingDaf','Drumming']

Performance measure of hybrid CNN with LSTM model

Accuracy - 0.7879
Loss - 0.8016
Precision_Score - [0.87037037, 0.89090909, 0.73076923, 0.75609756, 0.61538462, 0.8372093 , 0.89655172, 0.56521739, 0.91111111, 0.6875]

Recall_Score - [0.81034483, 0.84482759, 0.92682927, 0.64583333, 0.88888889, 0.81818182, 0.72222222, 0.48148148, 0.82 , 0.825]

F1_Score - [0.83928571, 0.86725664, 0.8172043 , 0.69662921, 0.72727273, 0.82758621, 0.8 , 0.52 , 0.86315789, 0.75]

Confusion Matrix -

A confusion matrix contains visualized and quantised information about multiple classifiers using a reference classification system [12], [49]. Each row represents the predicted class, and each column represents instances of the ground truth classes.

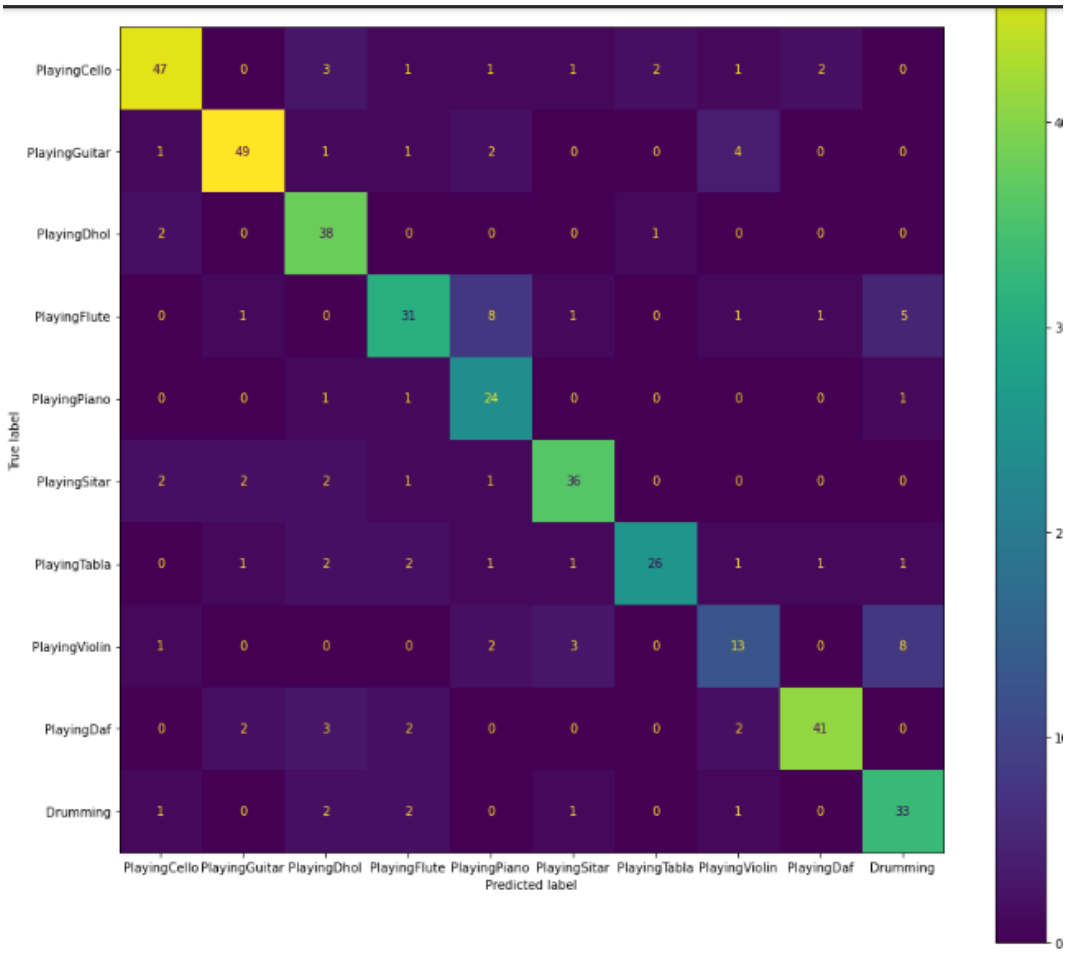


Fig 5: Confusion Matrix for UCF dataset using CNN + LSTM model for a action with 78 % accuracy

Performance measure of the proposed hybrid CNN with SVM model

Accuracy - 0.8834

Loss - 0.9492

Precision_Score - [0.83333333, 0.98181818, 0.98076923, 0.75609756, 0.82051282, 0.88372093, 0.96551724, 0.65217391, 1. , 0.83333333]

Recall_Score - [0.95744681, 0.91525424, 0.83606557, 0.83783784, 0.91428571, 0.92682927, 0.93333333, 0.75 , 0.88235294, 0.83333333]

F1_Score - [0.89108911, 0.94736842, 0.90265487, 0.79487179, 0.86486486, 0.9047619 , 0.94915254, 0.69767442, 0.9375 , 0.83333333]

Confusion Matrix generated using proposed CNN and SVM Model-

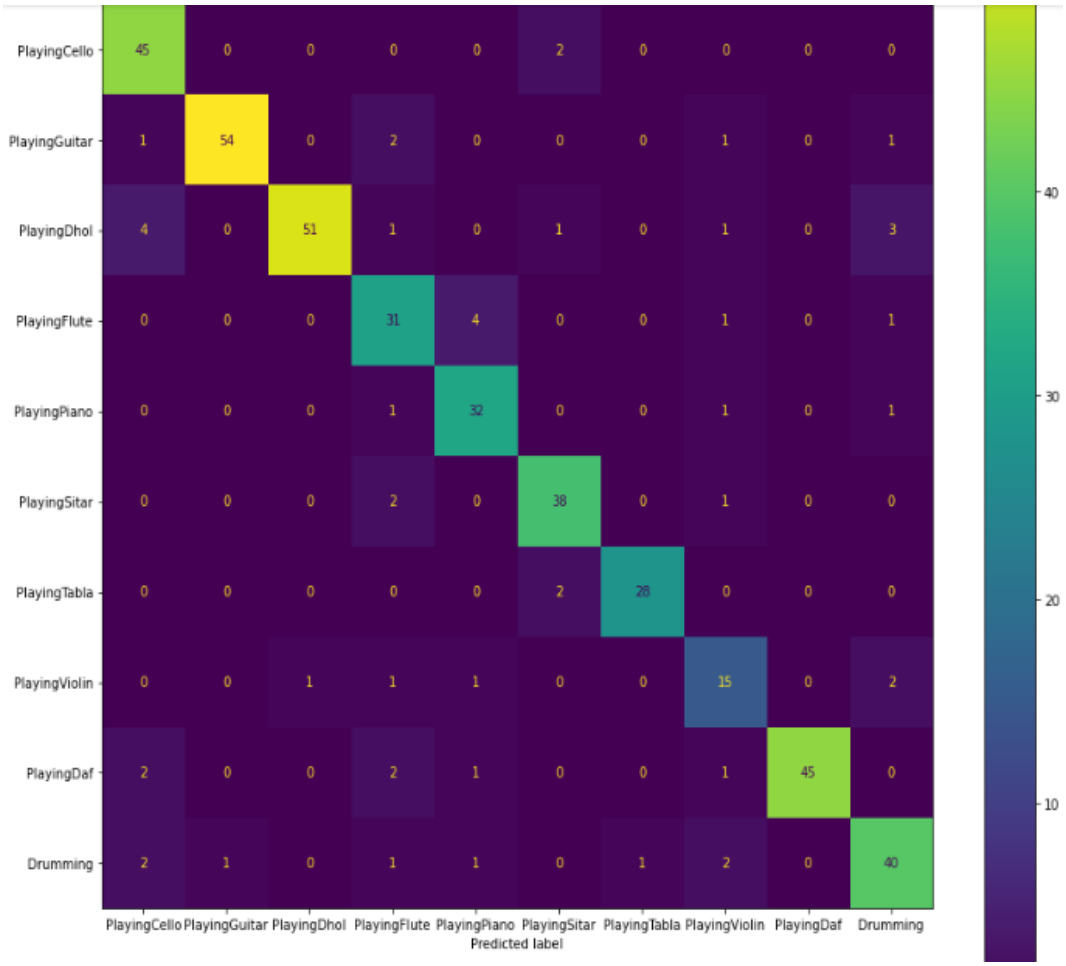


Fig 6: Confusion Matrix for UCF dataset using CNN with SVM LSTM model for an action with more

than 88 % accuracy

Comparison of Results of CNN+LSTM and CNN+SVM with different splitting percentage of training and testing dataset, the results are tabulated in Table 1

Table 1 : Performance evaluation of the proposed method with other methods based on split ratio

S.No	**Attributes:**	**80 - 20**		**70 - 30**		**60 - 40**	
		CNN +LSTM	**CNN with SVM (proposed model**	**CNN +LSTM**	**CNN with SVM (proposed model**	**CNN +LSTM**	**CNN with SVM (proposed model**
1	Accuracy	0.8497 **:** 0.8776		0.7879 : 0.8834		0.8007 : 0.8514	
2	Loss	0.6285 **:** 0.9576		0.8016 : 0.9492		0.8010 : 0.9647	
3	Recall_Score	0.8496 : 0.8776		0.7879 :0.8834		0.8006 : 0.8513	
4	F1_Score	0.8496 : 0.8776		0.7879 : 0.8834		0.8006 : 0.8513	
5	Precision_Score	0.8496 : 0.8776		0.7879 : 0.8834		0.8006 : 0.8513	

Sample Output Generated:

We compiled and trained the model then tried to predict with several fresh videos as input where we got satisfactory results. Model was able to predict what action was in that particular video. Some Samples predicted from our Model are depicted in Figure 6.

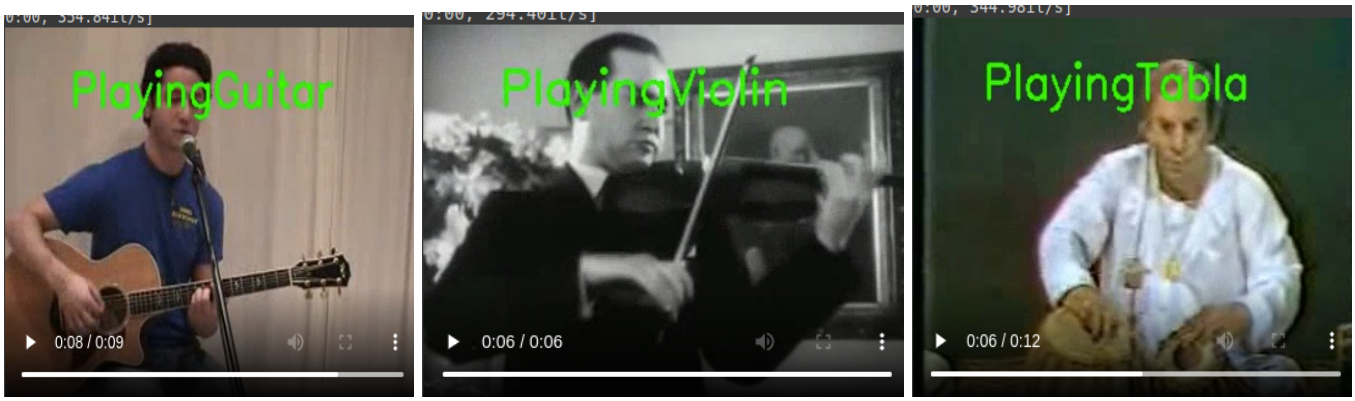


Fig 7: Some of the detected actions

Application of the Proposed work:

Automatic surveillance is nowadays gaining a lot of importance mostly after the popularity or completion of usage of online systems. Industry is very keen in designing automatic systems to catch unwanted interference. The applications range from surveillance in Industry line, online interview for job aspirants or may be Cheating on the Seat of The Examination. Automated surveillance are being extensively used by government for detecting trouble makes in crowd or terrorist. The systems are getting so sophisticated that they can predict what a terrorist can attack much before actual attack so that giving ample time to security personnel to avert the attacks.

The presented work is a part of a surveillance system for remote online based Interview systems to facilitate recruiters. A specific dataset for the particular domain is being collected in an interview system demonstrated as below in Fig 7.

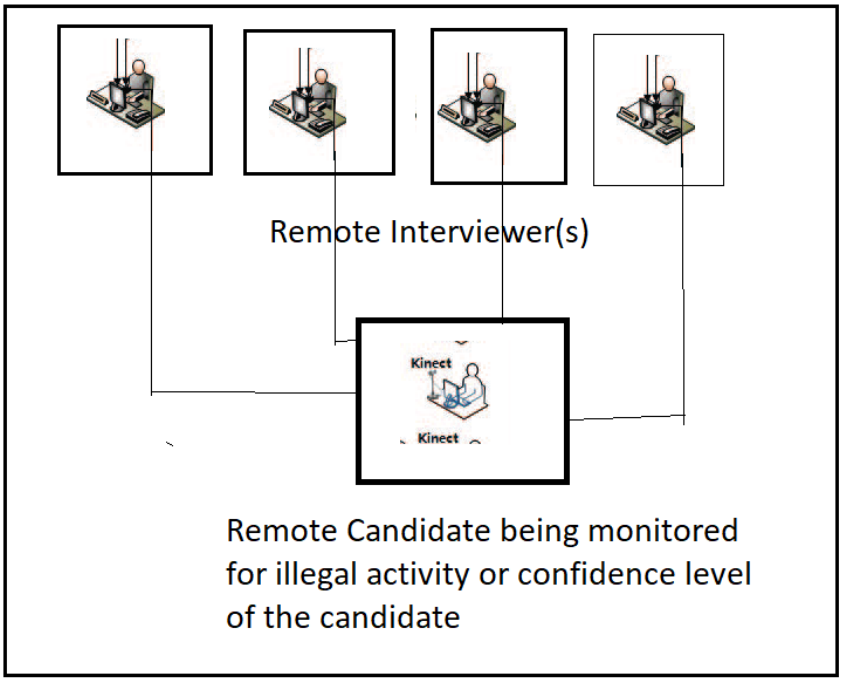


Fig 8: Remote Surveillance based Interview system

Some sample actions that is being detected is given in Fig 8:

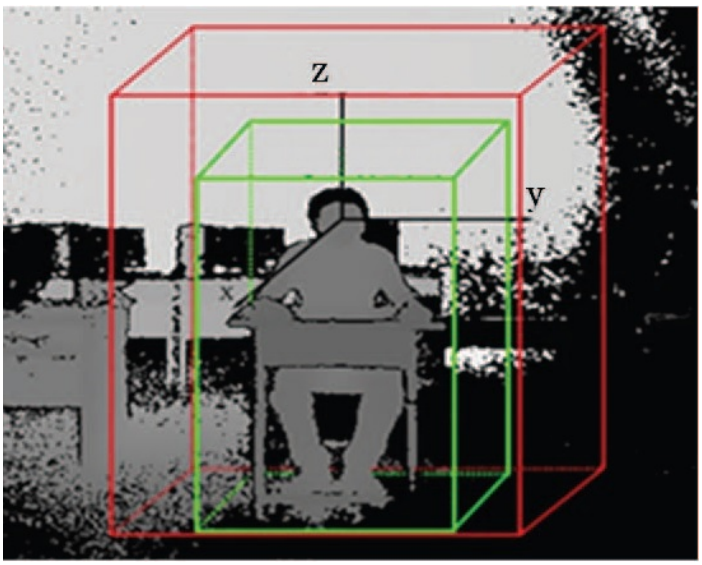


Fig 9: Window of observance for remote Interview system for automatic surveillance.

Conclusion:

In this project, we carried out a brief study of a framework for Action Recognition from videos and proposed some experiments on it. We took the UCF101 dataset and with the 101 classes/categories took some random four classes**.** Preprocessed the dataset and some of the features have been extracted as of now. We have splitted the dataset into training and testing parts and experimented with the LRCN(CNN(Convolutional Neural Network) and LSTM(Long Short-Term Memory Network)) approach which basically deals with images in

motion. We have trained the model with four classes and predicted with 0.789% accuracy till now.

Furthermore, for better classification of actions we moved to Support Vector Machine (SVM) from LSTM and gained an accuracy of 0.8834.

Future Scope:

Activity Recognition is an important problem in computer vision. AR is the basis for many applications such as video surveillance, health care, and human-computer interaction. Methodologies and technologies have made tremendous development in the past decades and have kept developing up to date. However, challenges still exist when facing realistic sceneries.

Reference:

[1] S.Sadanand and J.Corso,"Action bank: A high-level representation of activity in Video," IEEE Computer Society Conference on Computer Vision and Pattern Recognition ,pp. 1234-1241. 2012.

[2] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu ,"3D convolution neural networks for human action recognition",
IEEE Transactions On Pattern Analysis And Machine Intelligence , vol. 35, no. 1,pp. 221-231, 2013.

[3] Meng Li, Howard Leung, and Hubert P. H. Shum,"Human action recognition via skeletal and depth based feature fusion,"Proceedings of the 9th International Conference on Motion in Games,pp.123-132, 2016

[4] 1 Jeevan J.Deshmukh, 2Nita S.Patil, 3Dr.Sudhir D.Savarkar 1Student, 2 Assistant Professor, 3Professor of the IEEE, ISSN: 2321-9939.

[5] Yuanyuan Huang, Haomiao Yang,and Ping Huang, "Action recognition using hog features in different resolution video sequences," International Conference on Computer Distributed Control and Intelligent Environmental Monitoring, 2012.

[6] Heng Wang, Alexander Klaser, Cordelia Schmid,and Cheng-Lin Liu,"Dense trajectories and motion boundary descriptors for action recognition,"International journal of computer vision,pp.60-79,2013.

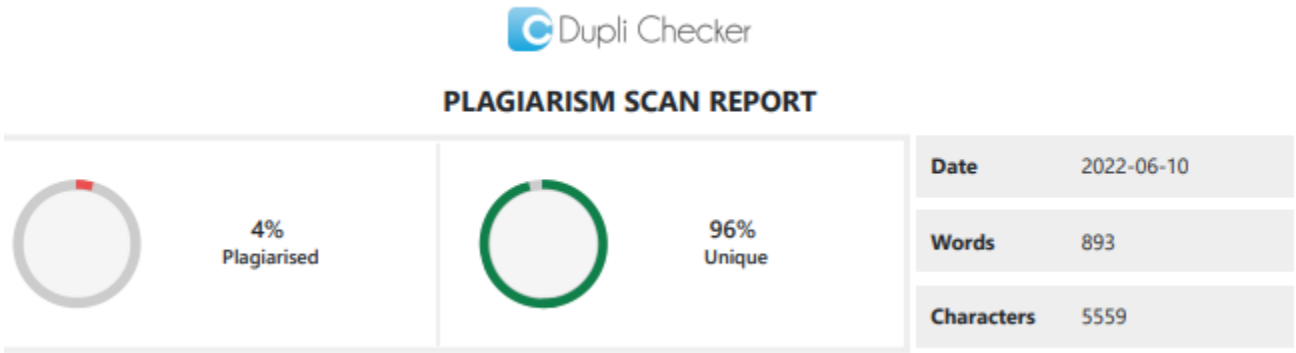
[7] Jilin Communications Polytechnic, Changchun, pp.117-126.

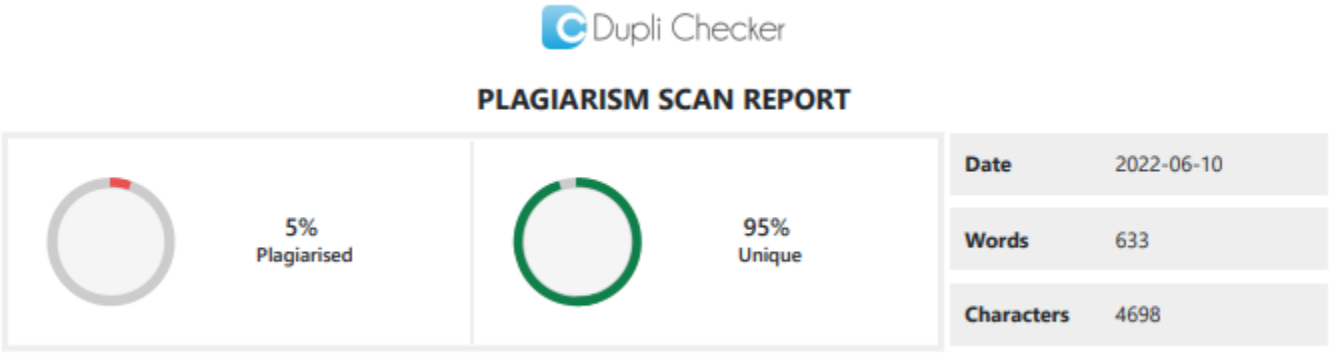
[8] TamV.Nguyen et al.,"Spatial-Temporal Attention-Aware Pooling for Action Recognition," IEEE Transactions On Circuits And Systems For Video Technology, vol. 25, no. 1,pp.77-86, 2015.

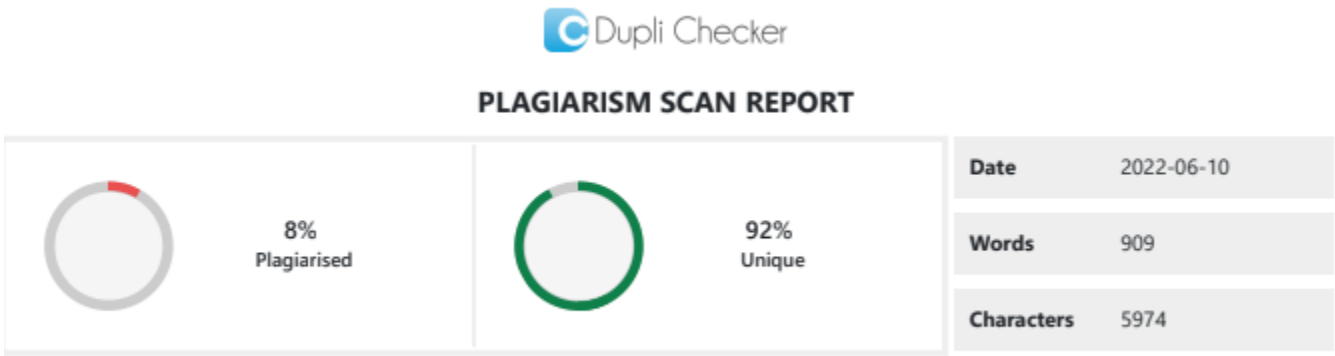
[9]Md Shofiqul Islam, Shanjida Sultana , Uttam kumar Roy , Jubayer Al Mahmud, ISSN: 2338-3070.

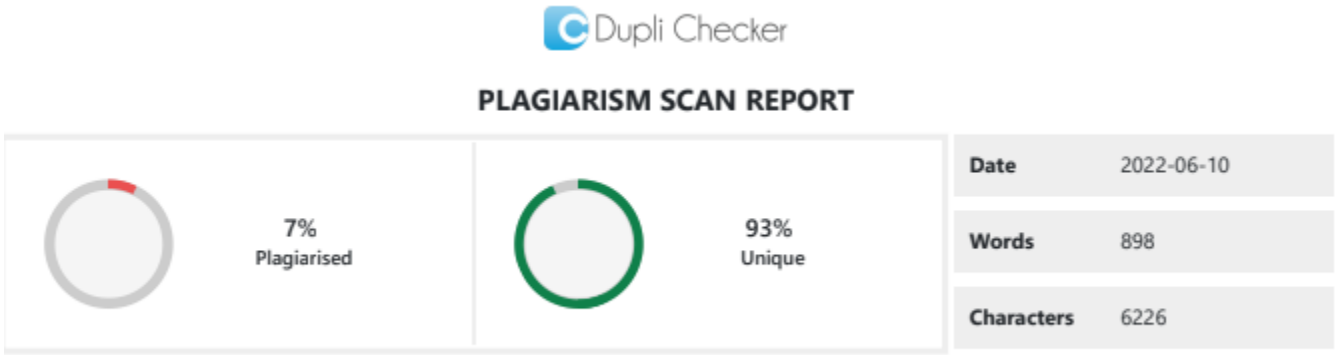
[10] M. Ramesh, K. Mahesh"Sports Video Classification with Deep Convolutional Neural Network: A test on UCF101 Dataset"ISSN: 2249 – 8958.

Plagiarism Checker









Dupli Checker

PLAGIARISM SCAN REPORT

0%
Plagiarised

100%
Unique

Date	2022-06-10
Words	984
Characters	6642

Dupli Checker

PLAGIARISM SCAN REPORT

4%
Plagiarised

96%
Unique

Date	2022-06-10
Words	979
Characters	6314

Dupli Checker

PLAGIARISM SCAN REPORT

17%
Plagiarised

83%
Unique

Date	2022-06-10
Words	974
Characters	6724

Dupli Checker

PLAGIARISM SCAN REPORT

0%
Plagiarised

100%
Unique

Date	2022-06-10
Words	1000
Characters	6476