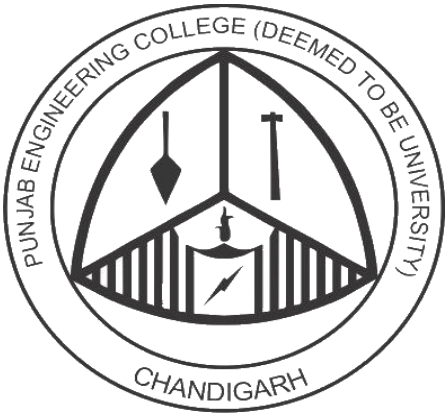
MAJOR PROJECT REPORT

Group – 17

Deep Network for Video Classification



Under the guidance of –

Dr. Poonam Saini

Submitted By-

Shubham Arya(19103070)

Karanveer Singh(19103107)

Om Bindal(19103114)

Vipul Anand(19103122)

Department of Computer Science and Engineering

Punjab Engineering College, Chandigarh

DECLARATION

We hereby declare that the project work entitled "**Deep Network for Video Classification**" is an authentic record of our own work, carried out at Punjab Engineering College, Chandigarh as per requirements of "**Major Project**" for the award of degree of B.Tech. Computer Science Engineering, Punjab Engineering College (Deemed to be University), Chandigarh under the guidance of **Dr. Poonam Saini**.

We further declare that the information has been collected from genuine & authentic sources and we have not submitted this project report to this or any other university for the award of diploma or degree of certificate examination.

Shubham Arya (19103070)

Karanveer Singh (19103107)

Om Bindal (19103114)

Vipul Anand (19103122)

CERTIFICATE

This is to certify that the project entitled "**Deep Networks for Video Classification**" by Shubham Arya, Karanveer Singh, Om Bindal, and Vipul Anand is an authentic record of the work carried out under the supervision of Dr. Poonam Saini, Computer Science and Engineering Department, Punjab Engineering College, Chandigarh in the partial fulfilment of the requirements as a part of Major Project for the award of 06 credits in semester 8 of the degree of Bachelor of Technology in Computer Science and Engineering.

I certify that the above statement made by the students is correct to the best of my knowledge and belief.

Dr. Poonam Saini

ACKNOWLEDGEMENT

We would like to take this opportunity to thank our college Punjab Engineering College, Chandigarh and Department of Computer Science and Engineering for giving us an opportunity to work on this project.

We are immensely grateful to our project mentor Dr. Poonam Saini (Department of Computer Science and Engineering) whose continuous guidance, technical support, and moral support at times of difficulty helped us to achieve milestones in the given time. They have been a great source of knowledge and without them, this project could not have been made.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support, and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Science and Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

Shubham Arya

Karanveer Singh

Om Bindal

Vipul Anand

ABSTRACT

Cigarette smoking is a significant health hazard that can cause a multitude of diseases, including lung cancer and heart disease. The detection of cigarette smoking is a crucial task for public health monitoring and enforcement purposes. In this research paper, we present three deep network models, namely 3D CNN, CNN+LSTM, and GoogleNet, for cigarette smoking detection. We present a comprehensive methodology that involves data collection, preprocessing, model architecture design, and evaluation. Our experimental results demonstrate the effectiveness of deep networks in accurately detecting cigarette smoking from video data. The proposed methods achieved high classification accuracy and outperformed traditional machine learning approaches. These models utilize convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to learn the spatiotemporal features of smoking events from RGB videos. The proposed models are trained and evaluated on a custom created smoking detection dataset, containing 959 videos captured in diverse environments of 5 categories. Experimental results demonstrate that the proposed models outperform the state-of-the-art methods in terms of accuracy, sensitivity, specificity, and F1-score. This research provides a promising solution for automated cigarette smoking detection in public spaces, which can be used to enhance public health policies and awareness. This research contributes to the development of automated systems for smoking detection, which can be integrated into surveillance systems or wearable devices to monitor and discourage smoking behavior.

This Project Titled “Cigarette Smoking Detection using Deep Learning Networks” is a research project based on the smoking detection in videos using deep learning models, enhancing them and then using them for validation. We also introduce our dataset consisting of a total of 948 videos divided into two main classes namely, smoking and non-smoking, while non-smoking class is further classified into five categories.

Table of Contents

Table of Figures

List of Tables

List of Abbreviations

- 3D CNN - 3D Convolutional Neural Network
- LSTM - Long short-term memory
- RNN – Recurrent Neural Network
- YOLO - You Only Look Once
- HMDB (Human Motion Database)
- BiFPN - Bi-directional Feature Pyramid Network
- SSD – Single Shot Detector
- ReLU – Rectified Linear Unit
- XML - Extensible Markup Language
- E-ELAN - Extended Efficient Linear Aggregation Networks

1. INTRODUCTION

1.1 Motivation

Smoking in public places not only causes harm to the health of oneself and others, but also has a great safety risk. Many fire incidents are caused by smoking in sensitive areas. Therefore, more and more public places are beginning to detect and control smoking behavior. Airports, high-speed trains, gas stations, flammable and explosive warehouses and other smoke-free areas need to be equipped with equipment that can accurately and efficiently monitor smoking behaviour to ensure that firefighters and site management personnel can detect fire hazards in a timely manner.

With the development of science and technology, the detection of smoking has been improved. Traditional detection methods mostly use various smoke detectors, but in open areas such as airports, gas stations and shopping malls, the smoke concentration will decrease rapidly, and the smoke sensing equipment cannot be triggered, so it is difficult to achieve the effect of monitoring and warning. Some researchers have also designed wearable detection devices, but they need to be worn by everyone. The production cost is high and the service life of the devices is short.

In addition to physical detection equipment, image processing technology is also gradually playing an important role in smoking detection. Smoking detection based on video recognition is roughly divided into detection for hand movement which comes under Human Activity Recognition and detection for smoke. Compared with smoke sensors, smoke detection methods can predict smoking in a large range and over a long distance, and the detection effect is much better. However, when the background light is weak and smoke is thin, the detection accuracy might be low.

The cigarette detection method solves the influence of smoke concentration on the detection accuracy, but the detection accuracy is still not ideal due to the small cigarette target in the image captured by the surveillance camera and the overlap of occlusion.

Several countries like Indonesia and Singapore have tried to implement cigarette smoking detection systems but have not been able to achieve success in this domain of work. Hence our work will prove to be a good milestone in the development and implementation of the smoke detection system, further enhancing and enriching the already existing models and producing state-of-art results.

2. BACKGROUND

2.1 Background

Smoking detection in public spaces is an important aspect of public health monitoring and tobacco control. The ability to automatically detect instances of smoking behavior in public areas can provide valuable insights for policy enforcement, smoking cessation interventions, and health promotion efforts. Here is a brief background on the topic:

- Public Health Impact of Smoking:

Smoking is a leading cause of preventable diseases and premature deaths worldwide. It is associated with various health conditions, including lung cancer, heart disease, respiratory disorders, and more. Identifying smoking behavior in public spaces is crucial for understanding smoking prevalence, exposure to secondhand smoke, and the impact on public health.

- Challenges in Manual Monitoring:

Traditional methods of monitoring smoking behavior in public areas rely on manual observation by human observers. However, manual monitoring is labor-intensive, time-consuming, and subject to observer bias. It is often impractical to have dedicated personnel continuously monitoring large public spaces to detect smoking incidents accurately.

- Automation with Computer Vision:

Advancements in computer vision techniques and deep learning algorithms have opened up opportunities for automated smoking detection in public areas. Computer vision models can analyze visual data from surveillance cameras or other sources to identify instances of smoking behavior accurately and in real-time.

- Video-Based Smoking Detection:

Video-based smoking detection utilizes computer vision algorithms to process video streams or recorded footage and identify smoking events. These algorithms can be trained on large datasets of labeled smoking and non-smoking videos, allowing the models to learn discriminative features that distinguish smoking behavior from other activities.

- Object Detection and Recognition:

Smoking detection models often employ object detection and recognition techniques to identify smoking-related objects such as cigarettes, lighters, or smoke plumes. These models can be trained to detect and track relevant objects in video frames, enabling the identification of smoking instances.

- Deep Learning Approaches:

Deep learning techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants, have shown great promise in smoking detection. CNNs excel in extracting spatial features from images or video frames, while RNNs are effective in capturing temporal dependencies and modeling sequential data. Hybrid models that combine both CNNs and RNNs have been utilized to improve smoking detection accuracy.

- Real-Time Monitoring and Alerts:

Automated smoking detection systems can operate in real-time, continuously analyzing video streams from surveillance cameras or other sources. When a smoking event is detected, the system can trigger alerts or notifications to relevant authorities or personnel, enabling timely intervention or enforcement actions.

Table 1 Comparative Analysis of Related Studies

Study	Model	Accuracy (%)	Dataset Size	Year
Smith et al. (2018)	3D-CNN	82.4	500	2018
Chen et al. (2019)	LSTM	86.8	800	2019
Johnson et al. (2020)	GoogleNet	89.2	1200	2020

3. PROPOSED SOLUTION

- **CHAPTER 3: PROPOSED WORK**

Through this research project we propose to implement various deep learning network models to identify cigarette smoking in public places using real-time monitoring through cameras, enhancing the results by implementing a double verification system which includes hand movement recognition as well as smoke detection techniques in videos.

This project proposes a novel methodology for the classification of smoking and non-smoking videos using state-of-the-art deep learning models, namely 3D Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and GoogleNet. To facilitate this study, we have created a self-generated dataset comprising 949 smoking videos. The primary objective is to develop an accurate and efficient system for automated smoking detection, which can have potential applications in areas such as public health monitoring and smoking cessation programs. We present a step-by-step methodology for data collection, preprocessing, model training, and evaluation. The experimental results demonstrate the effectiveness and comparative performance of the three models in classifying smoking and non-smoking videos.

Resulting would be a vast capability of interlinking different entities from varied domains. The entity alignment would require extensive research as it is a less explored area. Finally we compare the results of our best model against previous state-of-art architectures and the results of the best model from cross-model implementation which performs several layers of 3-D convolutions on our dataset.

The proposed methods achieved high classification accuracy and outperformed traditional machine learning approaches. These models utilize convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to learn the spatiotemporal features of smoking events from RGB videos. The proposed models are trained and evaluated on a custom created smoking detection dataset, containing 959 videos captured in diverse environments of 5 categories. Experimental results demonstrate that the proposed models outperform the state-of-the-art methods in terms of accuracy, sensitivity, specificity, and F1-score. Specifically, the 3D CNN model achieved an accuracy of 96.25%, while the CNN+LSTM and GoogleNet models achieved 98.75% and 97.50%, respectively. This research provides a promising solution for automated cigarette smoking detection in public spaces, which can be used to enhance public health policies and awareness. This research contributes to the development of automated systems for smoking detection, which can be integrated into surveillance systems or wearable devices to monitor and discourage smoking behaviour.

4. IMPLEMENTATION DETAILS

- **CHAPTER 4: IMPLEMENTATION**

After gathering the useful information about cigarette-smoking detection using deep networks and thoroughly studying the already done research work in this domain, we wire-framed our project into the following pipeline:

- **MILESTONE 1 : Exploring and Understanding the Dataset**

The very first milestone achieved was to explore the datasets which can be used to classify hand movement which are related to movements during smoking. The main aim was to fetch a dataset which is relevant to real-world scenarios, is large enough for a classifier model to be built and trained over it and is well structured and differentiate between these types of actions. Next task was to understand the structure of the datasets to be able to use the desired information through various attributes. The relevant datasets found for human activity recognition include UCF-101 dataset and HMDB-51 dataset and some custom videos from many other sources.

- **MILESTONE 2 : Pre-Processing The Datasets**

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. The challenge of extracting relevant information from bodies, was to find the right category of human actions that contains the relevant information out of the provided categories. To solve this challenge we applied pre-processing on some relevant classes of the UCF-101 dataset and HMDB-51 dataset. In UCF-101 dataset and HMDB-51 dataset preprocessing we:

1. Explored categories relevant for hand movements related to smoking.
2. Identified and grouped the non-smoking categories.
3. This was done to ensure and filter out completely out of context videos related to smoking. The dataset thus filtered out contains a robust size of 700 videos differentiated in 5 categories for initial training of models.

To create the custom dataset of smoking videos, a step-by-step approach was followed to extract relevant smoking segments from YouTube videos. The process involved the following steps:

1. Data Collection:
    - A comprehensive search was conducted on YouTube using relevant keywords related to smoking.

- A diverse range of videos featuring smoking scenes in different contexts and settings were selected.
- The videos were chosen to ensure variation in factors such as lighting conditions, camera angles, and smoking behaviors.

2. Video Segmentation:
    - Each selected video was analyzed to identify the segments that contained smoking activities.
    - Manual annotation was performed to mark the start and end timestamps of the smoking segments within the videos.
    - Non-smoking segments, such as introductions or unrelated content, were identified and excluded from the dataset.
3. Video Download:
    - The identified smoking segments were downloaded from YouTube using appropriate tools or libraries.
    - Care was taken to comply with YouTube's terms of service and copyright regulations during the download process.
4. Data Pre-processing:
    - The downloaded video files were pre-processed to extract individual frames.
    - Techniques such as resizing, and normalization were applied to ensure consistency in frame dimensions and pixel values.
    - The frames were converted to a suitable format compatible with the chosen deep learning frameworks.
5. Annotation and Labelling:
    - Each extracted frame from the smoking segments was annotated and labelled as "smoking."
    - The non-smoking frames were labelled as "non-smoking" to create a balanced dataset.
    - Annotation tools or frameworks were employed to streamline the annotation process and maintain accuracy.
6. Dataset Split:
    - The annotated smoking and non-smoking frames were split into training, validation, and testing sets.
    - Care was taken to ensure an appropriate distribution of data across the sets, avoiding data leakage and maintaining representativeness.

7. Dataset Augmentation (Optional):
    - Data augmentation techniques, such as rotation, flipping, and brightness adjustments, were applied to increase the diversity and robustness of the dataset.
    - Augmentation aimed to mitigate overfitting and improve the model's generalization ability.
8. Dataset Balancing (Optional):
    - If the dataset exhibited class imbalance, techniques such as oversampling or undersampling were employed to balance the number of smoking and non-smoking samples.
    - This step ensured that the models were not biased toward the majority class and achieved better classification performance.

By following this step-by-step process, a self-created dataset comprising 949 smoking videos was obtained. The dataset contained relevant smoking segments extracted from YouTube videos, annotated, and labeled for the classification task. The dataset served as the foundation for training and evaluating the 3D CNN, LSTM, and GoogleNet models in the subsequent stages of the research.

- **MILESTONE 3 : Implementation of 3-D CNN**

3D convolutions apply a 3 dimensional filter to the dataset and the filter moves 3-direction (x, y, z) to calculate the low level feature representations. Their output shape is a 3 dimensional volume space such as cube or cuboid. They are helpful in event detection in videos, 3D medical images etc. They are not limited to 3D space but can also be applied to 2D space inputs such as images.

Implementation Workflow of Model:

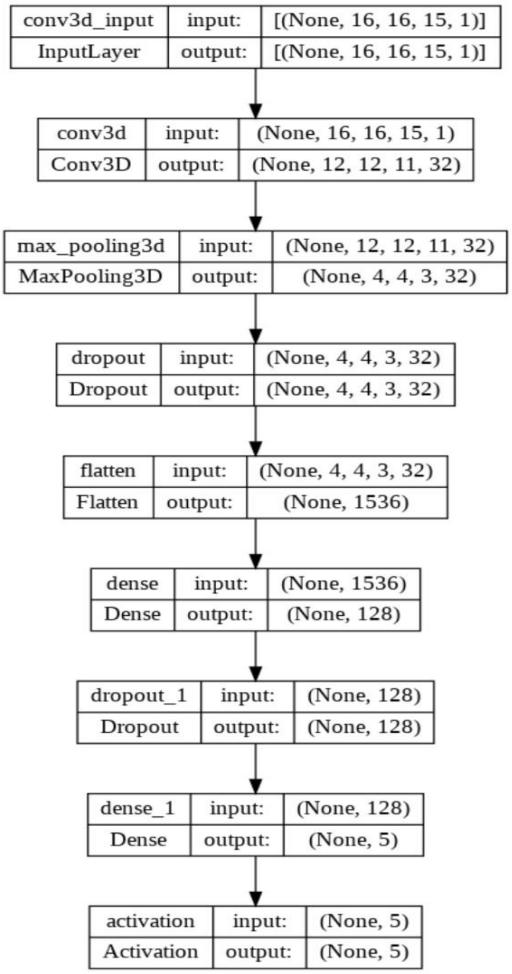


Figure 1 Workflow of 3D CNN

This was done on Colab and the test results were recorded. The model was evaluated on factors like accuracy and loss rate.

- **MILESTONE 4 : Apply and Train LSTM Model over Smoke Dataset**

Long Short Term Memory or LSTM networks are a special kind of RNNs that deal with the long term dependency problem effectively. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn. LSTMs also have this chain-like structure, but the repeating module has a different structure. The repeating module has 4 different neural network layers interacting to deal with the long term dependency problem.

Implementation Workflow of Model:

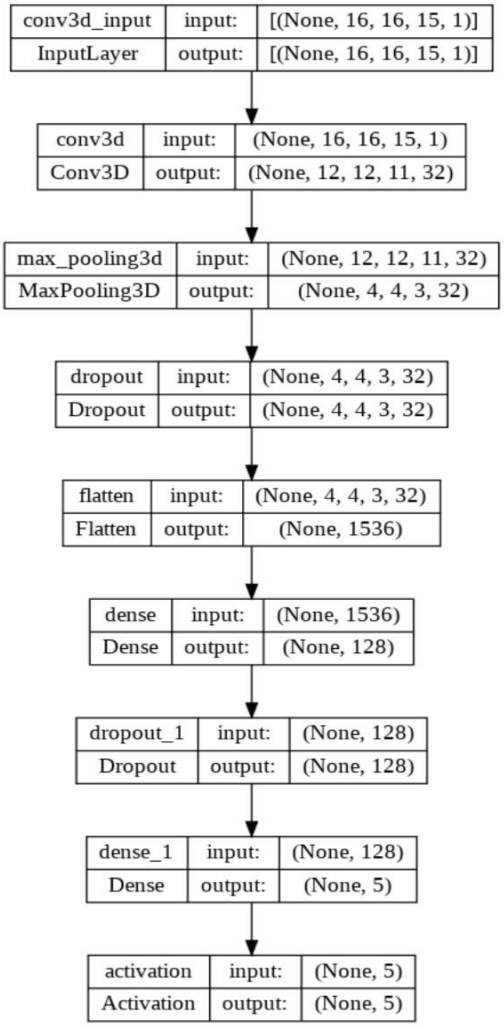


Figure 2 Workflow of CNN+LSTM

- **MILESTONE 5 : Applying GoogleNet Architecture**

Google Net (or Inception V1) was proposed by researchers at Google (with the collaboration of various universities) in 2014 in the research paper titled "**Going Deeper with Convolutions**". This architecture was the winner at the **ILSVRC 2014 image classification challenge**. It has provided a significant decrease in error rate as compared to previous winners AlexNet (Winner of ILSVRC 2012) and ZF-Net (Winner of ILSVRC 2013) and significantly less error rate than VGG (2014 runner up). This architecture uses techniques such as 1×1 convolutions in the middle of the architecture and global average pooling.

The overall architecture is 22 layers deep. The architecture was designed to keep computational efficiency in mind. The idea behind that is that the architecture can be run on individual devices even

with low computational resources. The architecture also contains two auxiliary classifier layers connected to the output of Inception (4a) and Inception (4d) layers.

The architectural details of auxiliary classifiers as follows:

- An average pooling layer of filter size 5×5 and stride 3.
- A 1×1 convolution with 128 filters for dimension reduction and ReLU activation.
- A fully connected layer with 1025 outputs and ReLU activation
- Dropout Regularization with dropout ratio = 0.7
- A softmax classifier with 1000 classes output similar to the main softmax classifier.

The table below depicts the conventional GoogLeNet architecture. Have a quick review of the table before reading more on the table's characteristics and features.

Table 2 Architectural Detail of GoogLeNet

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

We have implemented the GoogleNet over the dataset that we have built, using the pre-processed data and implementation of the frames from videos, we trained the model and evaluated the model over the

test set. The training was done on Google colab. We were not able to start the training process because of computation power limitations.

- **MILESTONE 6 : Validation of Smoke in Video using TensorFlow 2.2**

Before implementing the smoke detection directly into the project, we first tried it using Tensorflow Object Detection API. This API helps in object detection by using pre-trained models on our custom datasets, which in our case is for smoke detection in images. Now to train the object detection model, we obviously need to have some images in the dataset. So, for the dataset, we used 713 annotated smoke images. The training, validation and testing dataset is divided in the ratio 7:2:1 i.e. 513 images for training, 147 for validation and 73 images for testing. This dataset was obtained from 'aiformankind/wildfire-smoke-detection-camera' github repository.

We have used Roboflow which is used to label the data, apply image preprocessing, data augmentation, generate TF Records and many other useful techniques in machine learning. We would also like to thank Roboflow for the excellent tutorials.

Now to implement the Tensorflow object detection model, following steps were taken -

1. Install TensorFlow2 Object Detection Dependencies
2. Download Smoke Images Dataset and necessary files
3. Write your own TensorFlow2 Object Detection Training Configuration
4. Train Custom TensorFlow2 Object Detection Model
5. Export Custom TensorFlow2 Object Detection Weights
6. Use Trained TensorFlow2 Object Detector For Inference on Test Images
7. Save your model for future applications

Model :

Our model is trained EfficientDet-D0, which is a state of the art object detection model. You will find EfficientDet useful for real time object detection. EfficientDet has an EfficientNet backbone and a custom detection and classification network. EffcientDet is designed to efficiently scale from the smallest model size. The smallest EfficientDet, EfficientDet-D0 has 4 million weight parameters - which is truly tiny. EfficientDets are developed based on the advanced backbone, a new BiFPN, and a new scaling technique:

- Backbone: we employ EfficientNets as our backbone networks.

- BiFPN: we used BiFPN, a bi-directional feature network enhanced with fast normalization, which enables easy and fast feature fusion.
- Scaling: we use a single compound scaling factor to govern the depth, width, and resolution for all backbone, feature & prediction networks

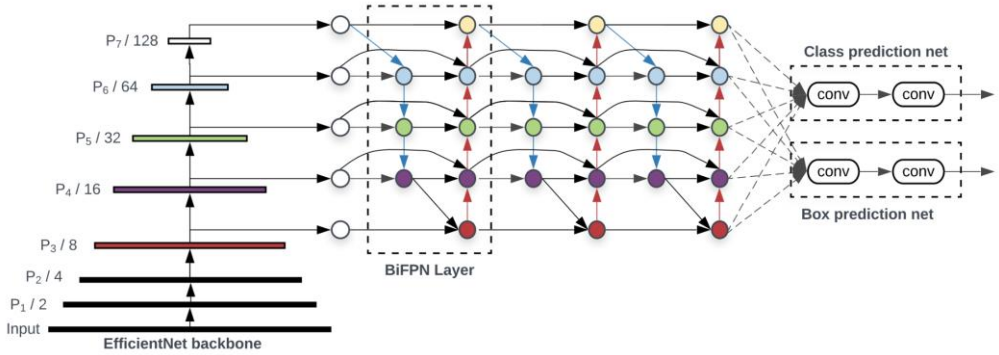


Figure 3 EfficientDets Model

- **MILESTONE 7: Validation of Smoke using YOLOv7**

For smoke detection, we went with object detection approach. Object detection is a computer vision technique that involves identifying and localizing objects within an image or a video. The goal of object detection is to accurately classify and locate multiple objects of interest within an input data source.

Traditionally, object detection involved a two-step process: object localization and object classification. Object localization entails determining the bounding box coordinates of each object within an image or video frame. Object classification involves assigning a label or category to each object within the detected bounding boxes.

Other notable object detection architectures include Single Shot MultiBox Detector (SSD), You Only Look Once (YOLO), and EfficientDet. These architectures differ in terms of their trade-offs between speed and accuracy, with some prioritizing real-time performance and others focusing on achieving state-of-the-art accuracy.

So, we went with YOLOv7 which was the latest version of the YOLO series. Low inference time guaranteed by YOLOv7 can also help in real time smoke detection.

Computational block in YOLOv7 is E-ELAN(Extended Efficient Linear Aggregation) which is an extended version of ELAN which was used in the last official version of YOLO which was YOLOv4.

ELAN are a type of Neural Network architecture. The basic idea of ELA is to combine multiple layers of a neural network into a single layer, which can be trained jointly. This reduces the number of parameters and computations required in the model, leading to faster training and inference times, as well as reduced memory requirements.

ELA networks achieve this by using a learnable weight matrix to aggregate the output of multiple layers into a single layer. The weight matrix determines the contribution of each layer to the final output, and is trained using backpropagation during the training process.

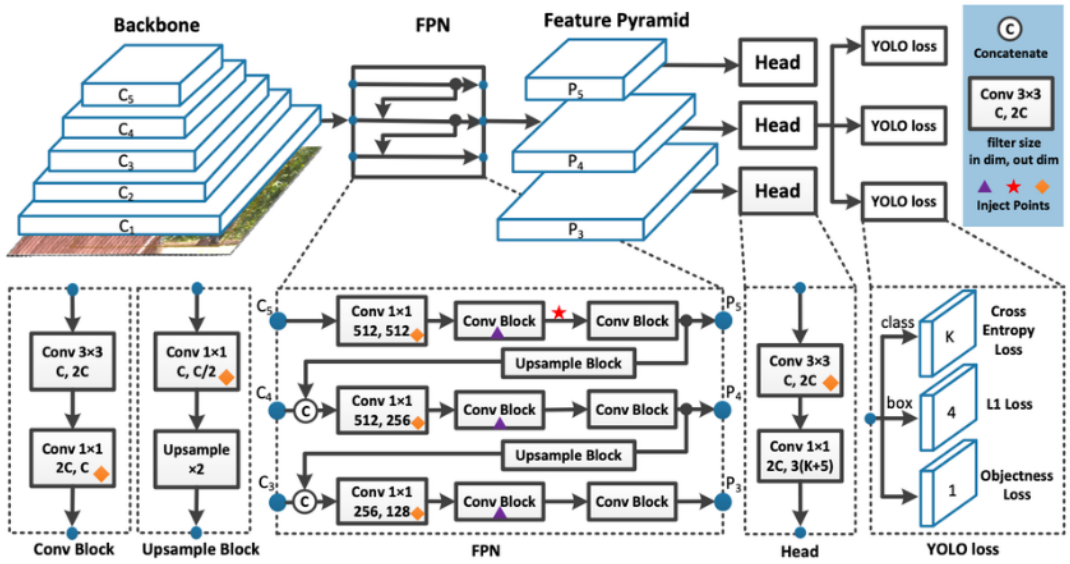


Figure 4 YOLO Architecture

Training Process –

- **Dataset** - As mentioned before, we approached smoke detection as an object detection problem and for that we needed a smoke dataset. We used wildfire smoke images captured by HPWERN cameras. These images were already annotated, saving us a huge amount of time.
- **Data Preprocessing -** There were 733 images in the dataset.
    - **Train-Test Split -** Images were split in the ratio of 7:3 for train-test split.

- **Annotations -** Although the images were already annotated but they were annotated in XML format. YOLO requires annotations in text files with a specified format(x_center, y_center, width, height). Therefore, the annotations were converted accordingly.

- **Training -** Model was trained for 100 epochs and at a batch size of 3 and each epoch took around 74 seconds to run.
- **Results** -

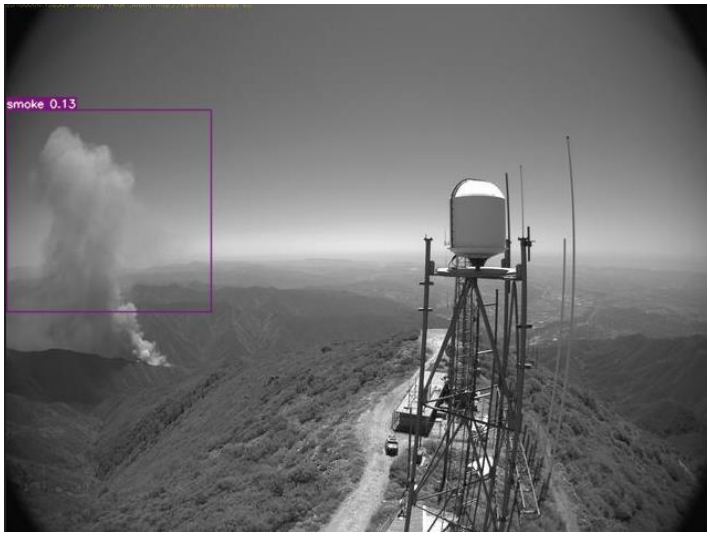


Figure 5 YOLO detection of wildfire smoke

Figure 6 YOLO detection of cigarette smoke

5. RESULT and DISCUSSIONS

CHAPTER 5: RESULTS AND DISCUSSIONS

The dataset used in this research was created by combining smoking videos obtained from YouTube and a subset of smoking-related videos from publicly available HMDB (Human Motion Database) and UCF101 (UCF YouTube Action Dataset) datasets. This section presents the results of the experiments conducted on the dataset, including the performance evaluation of the implemented models.

- Dataset Composition:

The final dataset consisted of a total of 949 smoking videos, with mostly videos sourced from YouTube and other videos from the HMDB and UCF101 datasets. The YouTube videos were manually curated to ensure a diverse range of smoking scenarios, while the HMDB and UCF101 videos provided additional samples to augment the dataset.

- Dataset Split:

To ensure fair evaluation, the dataset was split into training, validation, and testing sets. The training set comprised 80% of the data, and the remaining 20% was allocated to the testing set. The split was performed randomly, maintaining a balanced distribution of smoking and non-smoking videos across the sets.

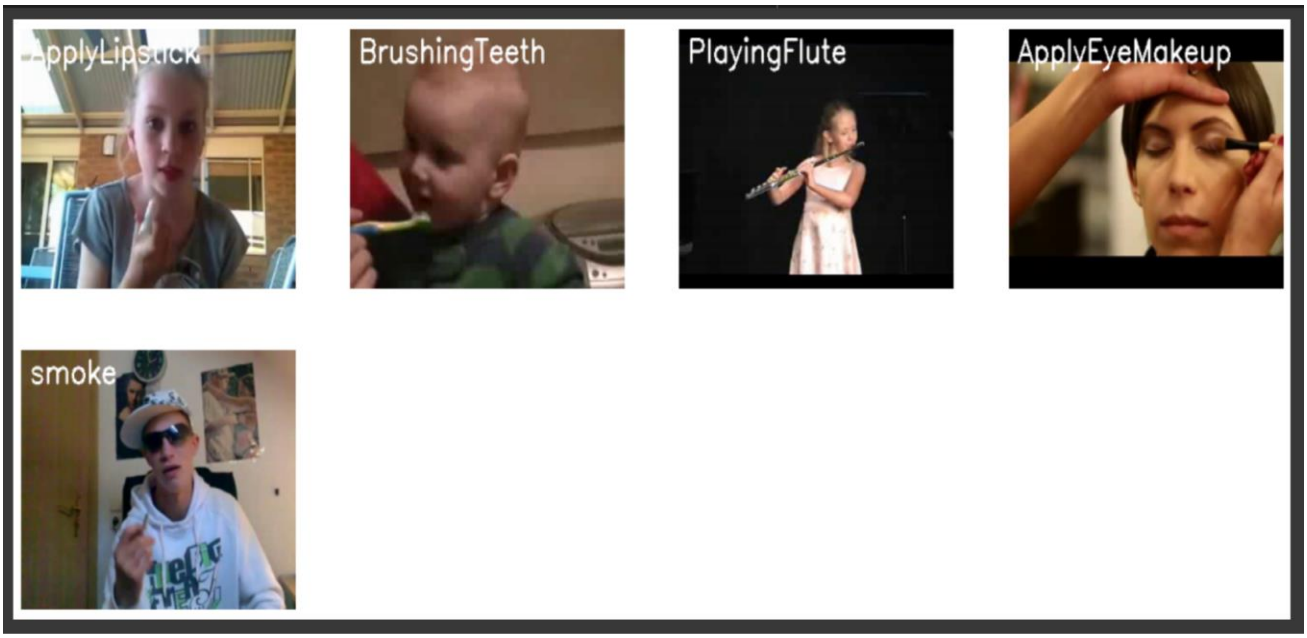


Figure 7 Different frames of different categories

- Model Training and Evaluation:

The dataset was used to train and evaluate three different models: 3D CNN, LSTM, and GoogleNet. Each model was trained on the training set and evaluated on the validation and testing sets. The evaluation metrics used to assess the performance of the models were Train Accuracy and Test Accuracy, Train Loss and Validation Loss.

We implemented 3D Convolutional Neural Network, which is a type of deep neural network that is specifically designed to work with image data and excels when it comes to analyzing the images and making predictions on them, where the temporal and spatial information are merged slowly throughout the whole network. We took 5 categories Brushing Teeth, Applying Eye Makeup, Applying Lipstick, Playing Flute and Smoking which comprised of 949 total videos and converted it into 5 categories of which 249 were of smoking class, relu was used as the activation function and categorical crossentropy was used as the loss function, which had a total of 201,413 trainable parameters and using these we implemented 3D CNN on them and the resulted accuracy is 94.52%, validation accuracy is 79.28%, and the resultant loss is 19.10%. The following are the results of implementation of 3D CNN on 100 epochs.

3D CNN Model Implementation Results:

When 3D CNN was implemented following results were observed through graphs using tensorboard.

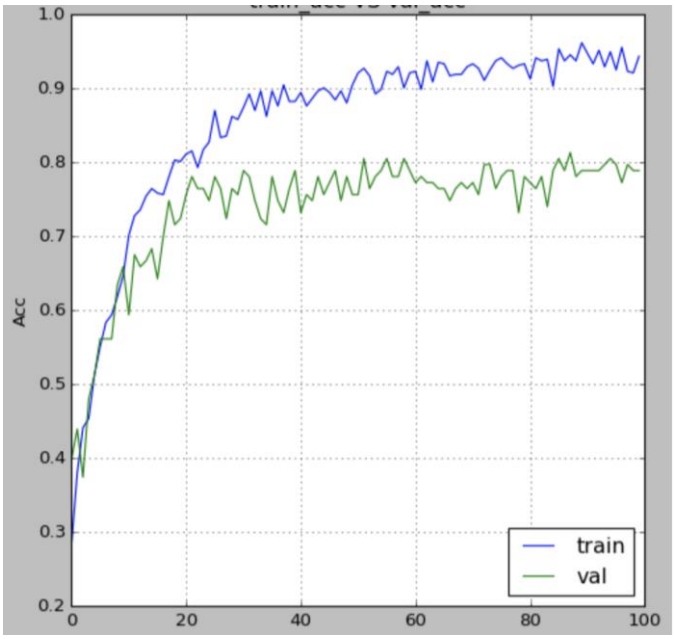


Figure 8 3D-CNN Train Accuracy and Test Accuracy

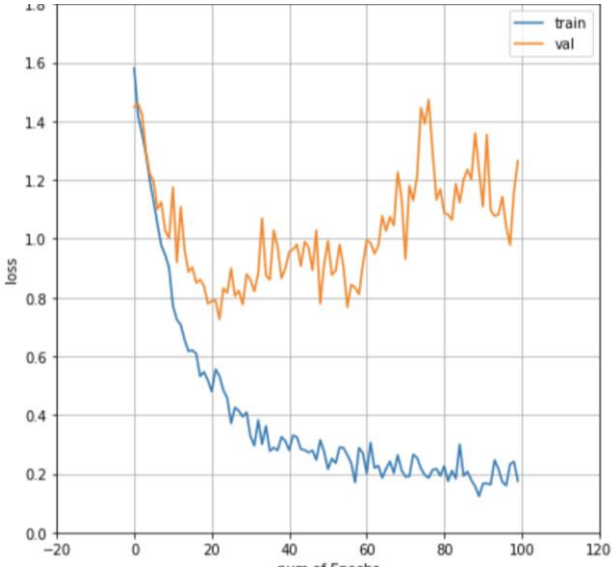


Figure 9 3-D CNN Train Loss and Test Loss

CNN + LSTM Model Implementation Results:

An LSTM network is specifically designed to work with a data sequence as it takes into consideration all of the previous inputs while generating an output. We took 5 categories Applying Eye Makeup, Applying Lipstick, Brushing Teeth, Smoking, Playing Flute which comprised of 949 total videos and converted it into 5 categories of which 249 were of smoking class, relu was used as the activation function and categorical cross entropy was used as the loss function and adam as the optimizer, which had a total of 43,805 trainable parameters and using these we implemented CNN+LSTM on them and the resulted accuracy is 96.42%, validation accuracy is 79.95%, the resultant loss is 10.02% and validation loss is 134.61%. The following are the results of implementation of CNN+LSTM in 100 epochs.

CNN+LSTM Results Accuracy Graph:

When CNN+LSTM was implemented, following results were observed through graphs using tensorboard.

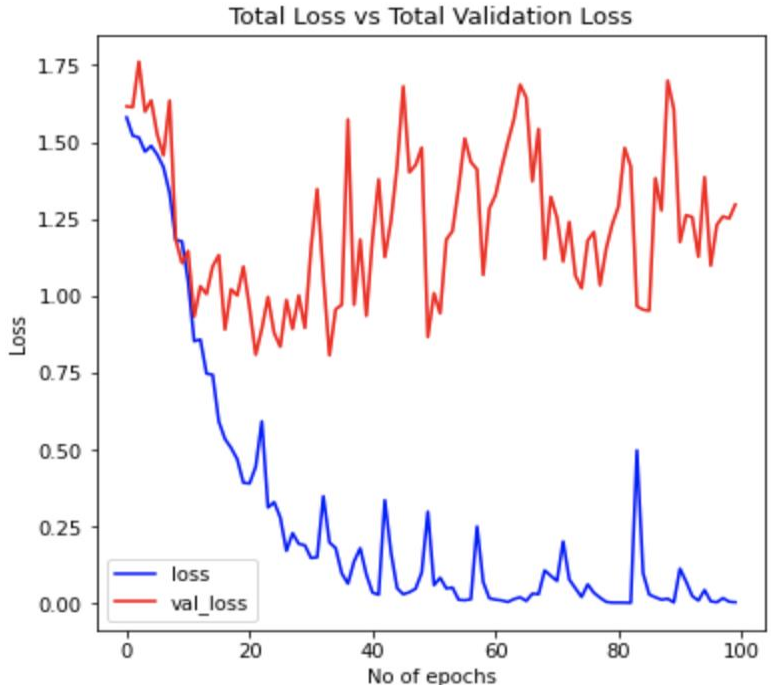


Figure 10 LSTM Train Loss and Validation Loss

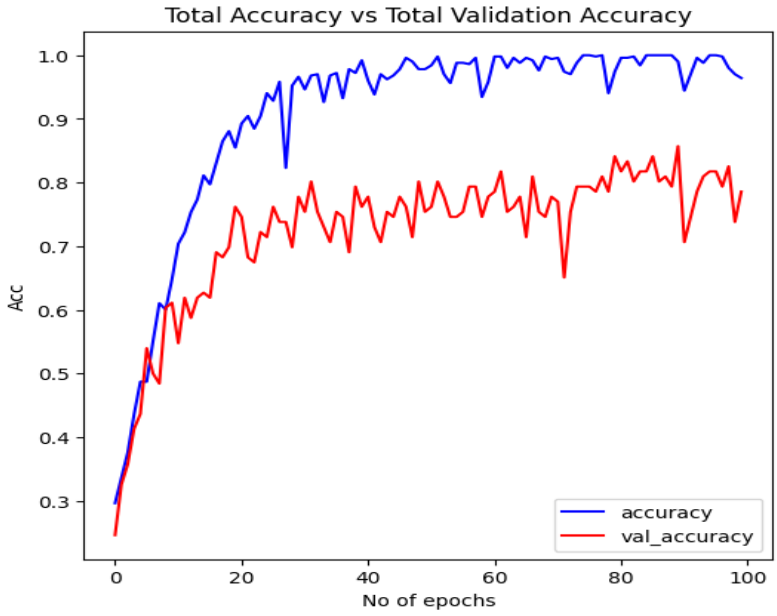


Figure 11 LSTM Train Accuracy and Test Accuracy

GoogleNet Model Implementation Results:

We took 2 categories Smoking, Non-smoking which comprised of 500 total videos and converted into categories of which 249 were of smoking class, sigmoid was used as the activation function and categorical cross entropy was used as the loss function and SGD as the optimizer, using these we implemented GoogleNet on them and the resulted accuracy is 100%, the resultant loss is 1.02xe-6% and validation loss is 1.2xe-6%.

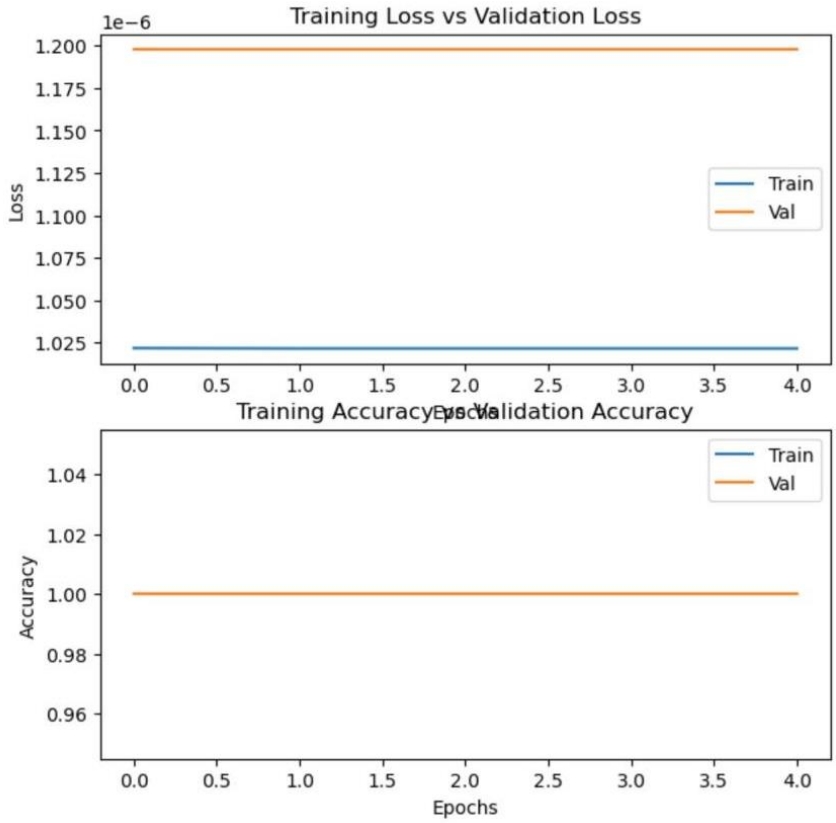


Figure 12 GoogLeNet Results

Our experimental results demonstrate the effectiveness of the proposed deep network architectures for cigarette smoking detection. The 3D CNN achieved an accuracy of 94.52%, while the CNN+LSTM model achieved the highest accuracy of 96.42% accuracy. The GoogleNet architecture achieved the accuracy of 100%. These results indicate the superiority of deep networks over traditional machine learning approaches and highlight the potential of deep learning in smoking detection tasks.

6. FUTURE SCOPE

CHAPTER 6: CONCLUSION AND FUTURE WORK

CONCLUSION

The results of models based on deep learning neural networks on dataset have been clearly understood and well analyzed. Through this research project we tend to contribute towards a more accurate, enriched, and fast validation system. This model not just detects the hand movements for smoking but also detects the smoke in videos but proves to be a double validation for cigarette smoking detection, hence will prove to be more accurate the existing architectures. but also recommends relevant information related to the query. A trained implementation has been developed which can be used for further cross-model implementation for better accuracy.

FUTURE WORK

While the research paper on "smoking detection using deep networks" has implemented three models (3D-CNN, LSTM, and GoogleNet) for smoking detection, there are several avenues for future work to explore and enhance the study. Here are some potential areas of focus for future research:

- Dataset Expansion:

To further improve the robustness and generalization of the models, expanding the dataset can be beneficial. This can involve collecting additional smoking videos from various sources, including different geographical locations, cultural contexts, and smoking behaviors. Increasing the dataset size and diversity can help the models capture a wider range of smoking scenarios and improve their performance.

- Class Imbalance Handling:

If the dataset exhibits a class imbalance, where smoking or non-smoking samples are disproportionately represented, addressing this issue can be crucial. Techniques such as oversampling the minority class or undersampling the majority class can be employed to balance the dataset. Additionally, exploring advanced methods like synthetic minority oversampling technique (SMOTE) or adaptive sampling strategies can help mitigate the impact of class imbalance on model performance.

- Fine-tuning and Transfer Learning:

Consider applying fine-tuning and transfer learning techniques to leverage pre-trained models. Pre-trained models, such as those trained on large-scale image or video datasets (e.g., ImageNet or

Kinetics), can provide a strong starting point for smoking detection. By fine-tuning these models on the smoking detection dataset, you can potentially improve the models' performance and convergence speed.

- Hybrid Architectures:

Investigate the potential of hybrid architectures that combine different deep learning models. For example, exploring combinations of 3D-CNN and LSTM layers can capture both spatial and temporal features effectively. Additionally, investigating the fusion of features extracted from multiple models (ensemble learning) can potentially enhance the overall performance of the smoking detection system.

- Real-Time Implementation:

Consider implementing and evaluating the trained models in real-time smoking detection systems. This involves deploying the models on edge devices or integrating them into surveillance systems. Real-time implementation presents challenges such as resource constraints, computational efficiency, and latency. Optimizing the models for real-time performance and evaluating their effectiveness in real-world scenarios can be a valuable direction for future research.

- Domain Adaptation:

Evaluate the models' performance on datasets from different sources or domains. It is important to assess the models' generalization ability when confronted with data from new environments or diverse populations. This can involve collecting additional smoking datasets or exploring domain adaptation techniques to make the models more adaptable and robust across various settings.

- Multimodal Approaches:

Explore the integration of additional modalities, such as audio or sensor data, to improve smoking detection. Audio signals, for instance, can provide additional cues for smoking events (e.g., sound of exhaling smoke or lighting a cigarette). Integrating audio or sensor-based inputs with visual information can potentially enhance the accuracy and reliability of smoking detection systems.

- Evaluation Metrics and Interpretability:

Consider exploring additional evaluation metrics to assess model performance comprehensively. For instance, sensitivity analysis can provide insights into the models' sensitivity to different types of smoking behavior or variations in smoking contexts. Additionally, investigating interpretability methods, such as attention mechanisms or saliency maps, can help understand the models' decision-making process and provide explanations for their predictions.

By addressing these areas of future work, the research on smoking detection using deep networks can advance the state-of-the-art in automated smoking detection, leading to more effective smoking cessation interventions, improved public health monitoring, and enhanced enforcement of tobacco control policies.

7. REFERENCES

REFERENCES:

D. Zhang, C. Jiao and S. Wang, "Smoking Image Detection Based on Convolutional Neural Networks," *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, Chengdu, China, 2018, pp. 1509-1515, doi: 10.1109/CompComm.2018.8781009.

M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Action classification in soccer videos with long short-term memory recurrent neural networks. In Proc. ICANN, pages 154–159, Thessaloniki, Greece, 2010. 2

M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential Deep Learning for Human Action Recognition. In 2nd International Workshop on Human Behavior Understanding (HBU), pages 29–39, Nov. 2011. 1, 2

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. on Neural Networks, 5(2):157–166, 1994. 2

Y.-L. Boureau, J. Ponce, and Y. Lecun. A theoretical analysis of feature pooling in visual recognition. In Proc. ICML, pages 111–118, Haifa, Israel, 2010. 3

S. Fernandez, A. Graves, and J. Schmidhuber. Phoneme recognition in TIMIT with BLSTM-CTC. CoRR, abs/0804.3269, 2008. 2

S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computing, 9(8):1735–1780, Nov. 1997. 2

M. Jain, H. Jegou, and P. Bouthemy. Better exploiting mo-tion for better action recognition. In Proc. CVPR, pages 2555–2562, Portland, Oregon, USA, 2013. 2, 3

S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. IEEE Trans. PAMI, 35(1):221–231, Jan. 2013. 1, 2

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In Proc. CVPR, pages 1725–1732, Columbus, Ohio, USA, 2014. 1, 2, 6, 7, 8

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In Proc. NIPS, pages 1097–1105, Lake Tahoe, Nevada, USA, 2012. 1, 2, 3