Implementation of Driver Vigilance System using Deep Learning and Advance Computer Vision

Harshvardhan Patil
Department of Computer Science and Engineering,
Jain college of Engineering,
Belagavi, India.

Aishwarya C Kuratti
Department of Computer Science and Engineering,
Jain college of Engineering,
Belagavi, India.

Disha Bhanushali
Department of Computer Science and Engineering,
Jain college of Engineering,
Belagavi, India.

Sandhya Belgaonkar
Department of Computer Science and Engineering,
Jain college of Engineering,
Belagavi, India.

Prof. Praveen.Y.Chitti
Department of Computer Science and Engineering,
Jain college of Engineering, Belagavi, India.

***Abstract*—Distracted driving is an established cause of motor vehicle crashes for all ages. With the rapidly growing elderly population and more adults embracing technology, distracted driving is also increasing in prevalence within that population—particularly cell phone usage behind the wheel. This research explores the behaviors and attitudes of senior drivers regarding cellphone use while driving as well as the prevalence of the mode of cell phone use behind the car such as, talking, texting, emailing, browsing the internet and navigating. It also explores possible characteristics that would predict the frequency of distracted driving. Distracted driving is an established cause of motor vehicle crashes, for all ages. Nearly 60% of crashes involving younger drivers are linked to distraction (AAAFTS, 2015). This research brief provides evidence from a recent survey that as more older adults embrace technology, distracted driving—in particular, using cell phones behind the wheel—is prevalent among them as well. According to a recent survey conducted by AAA Foundation for Traffic Safety and the University of California San Diego, the majority of drivers aged 65 and older—nearly 60%—have used their cell phone in some capacity (i.e., texting, making calls, and answering calls) while driving. More than a quarter of these older drivers have engaged in distracting behaviors while driving with a minor in the car. Among those, 32% have talked on the phone—either with hands-free or hand-held devices—with younger children (under age 11) in the car, while 42% have done so when accompanied by older children (12- to 17 year-olds). While distracted driving encompasses a wide range of risky behaviors including but not limited to eating, talking with passengers, reaching for belongings, etc., this survey focused solely on cell phone use while operating a vehicle. The findings suggest the need for interventions to reduce distracted driving behaviors among older adults, especially given the rapidly growing older adult population, with their age-associated physiologic changes, such as slower reflexes, reduced contrast sensitivity, and other driving-impairing conditions.**

I. INTRODUCTION

In recent years, the increasing number of vehicles on roads leads to an increase in traffic accidents. In 2015, the National Highway Traffic Safety Administration, part of U.S. Department of Transportation, reported that 35,092 people died in traffic accidents on the U.S. roads, a 7.2% increase in fatalities from 2014. Distracted driving was responsible for 391,000 injuries and 3477 fatalities in 2015. It is found that distracted driving was related to one-tenth of fatal crashes. Distracted driving fatalities have increased more rapidly than those caused by drunk driving, speeding and failing to wear a seatbelt. A driver is considered to be distracted when there is an activity that attracts his/her attention away from the task of driving.

There are three types of driving distractions

• Manual distraction: The driver takes his/her hands off the wheel, e.g. drinking, eating etc.

• Visual distraction: The driver looks away from the road, e.g. reading, watching the phone etc.

• Cognitive distraction: The driver's mind is not fully focused on the driving task, e.g. talking, thinking etc.

It is important to note that although driving distractions are categorized into three different types they do not always occur separately. For example, in the event of talking on the phone, two types of distractions occur at the same time: manual distraction and cognitive distraction. There are many sources that can lead to distraction. However, the most possible distractions usually come from inside the vehicle. Major motor companies such as Toyota, Nissan, Ford, and Mercedes-Benz have been introducing advanced infotainment, control panels, and display systems. Adjusting those in-vehicle devices while driving could cause a considerable distraction that may lead to traffic accidents. Another source that can influence the driving performance is phone use. Conversation on phones while driving consumes a significant amount of brain power. When doing both, the human brain activity dedicated to driving can be reduced by 37%. Text messaging while driving can cause even more distraction because it keeps not only the driver's thought but also his/her hands and eyes out of the driving task for an average time of 4.6s. A recent study pointed out that ~78% of drivers use cell phones behind the wheel which significantly increases the possibility of traffic accidents on the Indian roads. To reduce vehicle accidents and improve transportation safety, a system that can classify distracted driving is highly desirable and has attracted much research interest in recent

years. This study is motivated by developing such a distraction detection system that has the potential to be implemented in real vehicles. Therefore, the goal of this work is to develop an assisted driving system that can detect distracted driving behaviors and alert the driver to focus on the driving task. The main contributions of this study are: (i) proposing a real time distraction detection system which is developed using deep learning. (ii) implementing four types of convolutional neural networks (CNNs) for the detection system in order to determine the most suitable architecture for distraction detection; (iii) collecting our own distracted driving image dataset and (iv) developing a voice-alert system which reminds the driver to focus on the driving task when he/she gets distracted. The proposed work focuses on driver distraction activities detection via images using different kinds of machine learning techniques. The input of our model is videos of driver taken in the car. We first preprocess these videos to get input vectors, then use different classifiers (linear SVM, softmax, naive bayes, decision tree, and 2-layer neural network) to output a predicted type of distraction activity that drivers are conducting.

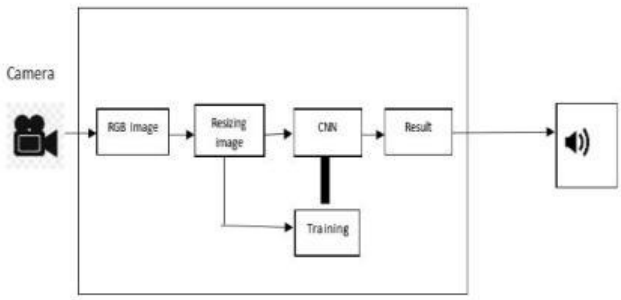


Figure 1.1: Software setup of the embedded computing system for distraction detection and alert

II. LITERATURE SURVEY

Literature survey or a literature study includes the current knowledge including substantive findings as well as theoretical and methodological contributions to a particular topic. In [1] the authors Mahbub Hussain, Jordan J. Bird and Diego R. Faria states that Image classification is one of the core problems in Computer Vision field with a large variety of practical applications. This paper proposes the study and investigation of a CNN architecture model(i.e. Inception-v3) to establish whether it would work best in terms of accuracy and efficiency with new image datasets via Transfer Learning. The retrained model is evaluated and the results are compared to some state-of-the-art approaches. Deep Learning has emerged as a new area in machine learning and is applied to a number of signal and image applications. Author used Deep Learning Algorithm namely Convolutional neural networks(CNN) in image classification [2]. Texted-based retrieval system is used to retrieve video or images from database but this is not efficient approach so to address this problems associated with Traditional system Content Based Image Retrieval and Content Based Video Retrieval were introduced. They proposed the methodology for CBIR based on image classification using Support Vector Machine [3] classifier is introduces and CBIR used C4.5 classifier. To detect an object in an image or a video the system needs to have few components in order to complete the task of detecting the object. The various techniques that are used to detect an object, localize an object, categorize an object, extract features, appearance information and many more[4]. In [5] the authors Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu and Xindong Wu states that due to object detection's close relationship with video analysis and image understanding, it has attracted much research attention in recent years. This paper provides a review on deep learning bases object detection frameworks. The Supervised Machine Learning is the search for algorithms that reason from externally supplied instances to produce general hypothesis, which then make predictions about future instances. In [6] describes various Supervised Machine Learning classification techniques, compares various supervised learning algorithms as well as determines the most efficient classification algorithm based on dataset, the number of instances and variables.

The Convolutional Neural Networks for human action recognition in videos have proposed different solutions for incorporating the appearance and motion information. In [7] a new ConvNet architecture for spatiotemporal fusion of video snippets, and evaluate its performance on standard benchmarks where this architecture achieves state-of-the-art results. In [8] the authors Fernando Moya Rueda, Rene Grzeszick, Gernot A. Fink , Sascha Feldhorst and Michael ten Hompel state that methods of HAR have been developed for classifying human movements. HAR uses as inputs signals from videos or from multichannel time-series. This paper focuses on HAR from multichannel time-series. Capturing, evaluating and analyzing signal series for recognizing human actions are critical for many applications. The video retrieval can be used for multiuser systems for video search and browsing which are useful in web applications. This paper takes the information needs and retrieval data already present in the archive, and that retrieval performance can be significantly improved when content-based image retrieval (CBIR) algorithm[9] are applied to search. With the development of multimedia data types and available bandwidth there is huge demand of video retrieval systems, as users shift from text based retrieval systems to content based retrieval systems. In [10] the authors Byeong-Ho KANG, states that Image Processing is any form of signal processing for which the input is an image, such as photographs or frames or videos .This paper presents Image and Video processing elements and current technologies related to that.

III. PROBLEM IDENTIFICATION

According to the motor vehicle safety division, one in five car accidents is caused by a distracted driver. The World Health Organization(WHO) reported 1.25 million deaths yearly due to road traffic accidents worldwide and the number is continuously increasing, Nearly fifth of these accidents are caused by distracted drivers. Therefore, our project aims to alarm the driver whenever he/she gets distracted. It mainly focuses on the driver when he/she is

texting on phone, talking on phone, drinking and operating radio.

IV. OBJECTIVES

The main objective of the proposed work is

• To reduce the number of accidents causing due to distracted drivers.
• To provide an alert system when the driver is involved in other activities apart from driving.
• To provide a system for the safety measures for drivers.

V. METHODOLOGY

6.1 Importing libraries.

Python modules can get access to code from another module by importing the file/function using import. The import statement is the most common way of invoking the import machinery. For this project, different libraries are imported for various purposes. Few are listed below.

A. NumPy

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. Operations using NumPy

 Mathematical and logical operations on arrays. Fourier transforms and routines for shape manipulation. Operations related to linear algebra. NumPy has in-built functions for linear algebra and random number generation.

B. Pandas

Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-touse data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc. Operations using pandas

 Fast and efficient DataFrame object with default and customized indexing.
 Tools for loading data into in-memory data objects from different file formats.
 Data alignment and integrated handling of missing data.
 Reshaping and pivoting of date sets.

C. Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits. Matplotlib comes with a wide variety of plots. Plots helps to understand trends, patterns, and to make correlations. They're typically instruments for reasoning about quantitative information.

D. Tensorflow

TensorFlow is a Python library for fast numerical computing created and released by Google. It is a foundation library that can be used to create Deep Learning models directly or by using wrapper libraries that simplify the process built on top of TensorFlow. TensorFlow is mainly used for: Classification, Perception, Understanding, Discovering, Prediction and Creation.

E. Keras

Keras is a powerful and easy-to-use free open source Python library for developing and evaluating deep learning models. It wraps the efficient numerical computation libraries Theano and TensorFlow and allows you to define and train neural network models in just a few lines of code. Keras doesn't handle low-level computation. Instead, it uses another library to do it, called the "Backend. So Keras is high-level API wrapper for the low-level API, capable of running on top of TensorFlow, CNTK, or Theano.

Keras High-Level API handles the way we make models, defining layers, or set up multiple input-output models. In this level, Keras also compiles our model with loss and optimizer functions, training process with fit function. Keras doesn't handle Low-Level API such as making the computational graph, making tensors or other variables because it has been handled by the "backend" engine.

F. OpenCV

OpenCV is a Python library which is designed to solve computer vision problems. It supports a wide variety of programming languages such as C++, Python, Java etc. OpenCV Python is nothing but a wrapper class for the original C++ library to be used with Python. Using this, all of the OpenCV array structures gets converted to/from NumPy arrays. This makes it easier to integrate it with other libraries which use NumPy. For example, libraries such as SciPy and Matplotlib.

6.2 Reading the dataset

We took the StateFarm dataset which contained snapshots from a video captured by a camera mounted in the car. The training set has ~22.4K labeled samples with equal distribution among the classes and 79.7 K unlabeled test samples. The provided data set has driver images, each taken in a car with a driver doing something in the car (texting, eating, talking on the phone, makeup, reaching behind, etc). This dataset is obtained from Kaggle. There are 10 classes of images:

C0: Safe Driving
C1: Texting Right
C2: Talking on phone right
C3: Texting Left
C4: Talking on phone left
C5: Operating the radio
C6: Drinking
C7: Reaching behind
C8: Talking to passengers
C9: Hair and makeup.

There are 102150 total images. Of these 17939 are training images, 4485 are validation images and 79726 are training images. All the training, validation images belong to the 10 categories shown above. The images are coloured and have 640 x 480 pixels. Now all the dataset images are read one by one. First the images of class 0 are read and

resized. After all the images in class 0 are read, images from class 1 are read and the process continues until all the 10 classes are read.

6.3 Preprocessing the images

Preprocessing of images is carried out before model is built and training process is executed. Following are the steps carried out during preprocessing. 1. Initially the images are divided into training and validation sets. 2. The images are resized to a square images i.e. 224 x 224 pixels. 3. All three channels were used during training process as these are color images. 4. The images are normalised by dividing every pixel in every image by 255. 5. To ensure the mean is zero a value of 0.5 is subtracted.

6.4 Splitting of Dataset

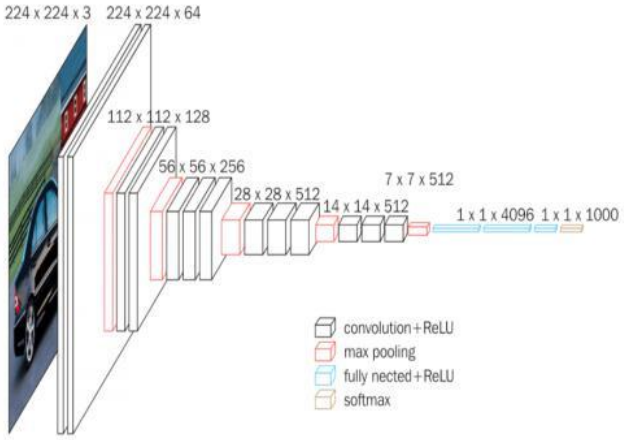
Now, after importing libraries and preprocessing all the images of the dataset. Now the images are categorized into 10 classes namely safe driving, texting on phone right, texting on phone left, operating radio etc. Now the images are further shuffled and are sent to the VGG Model.

The input to cov1 layer is of fixed size 224 x 224 RGB image.

The image is passed through a stack of convolutional (conv.)

6.5 Getting the model (Keras VGG16)

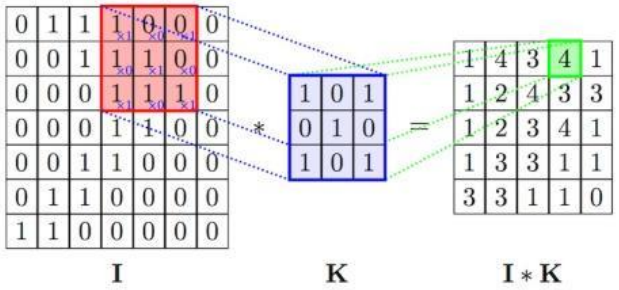
The input to cov1 layer is of fixed size 224 x 224 RGB image. The image is passed through a stack of convolutional (conv.) layers, where the filters were used with a very small receptive field: 3×3. In one of the configurations, it also utilizes 1×1 convolution filters, which can be seen as a linear transformation of the input channels. The convolution stride is fixed to 1 pixel; i.e. the padding is 1-pixel for 3×3 conv. layers. Spatial pooling is carried out by five max-pooling layers, which follow some of the conv. Layers.



Max-pooling is performed over a 2×2 pixel window, with stride 2. Fully-Connected (FC) layers follow a stack of convolutional layers (which has a different depth in different architectures): the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels. The final layer is the soft-max layer. The configuration of the fully connected layers is the same in all networks.

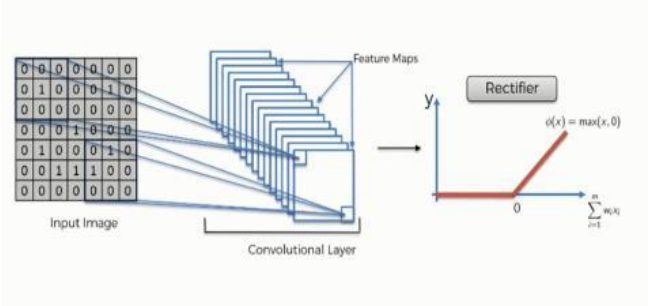
Convolutional Layer:

The CONV layer's parameters are a set of learnable filters of small dimensions (e.g. 5 × 5 × 3). The filter convolves with the input volume (across width and height in 2D) to select small areas (e.g. 5 × 5) and use these small local areas to compute dot products with weights/parameters. Each filter corresponds to a slice or a depth of one in the output volume (i.e. the depth of the output volume of a CONV layer is determined by the number of filters).



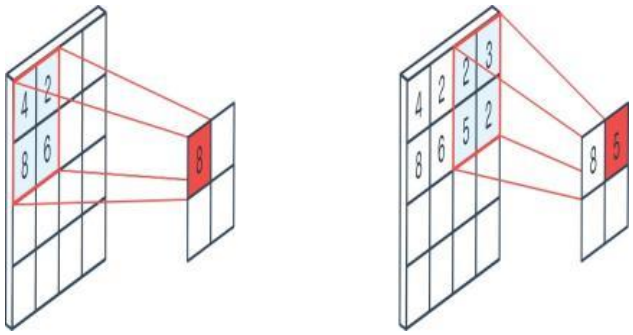
Rectifier Linear Units(ReLu) Layer:

The ReLU layer applies an element-wise non-linear activation function to increase nonlinearity in the model.



Pooling Layer:

A pooling layer reduces the 2D dimensions of the input volume (leaving the depth unchanged) to prevent the model from overfitting and getting too large (too many weights) to compute. It is done independently for each depth slice of the input volume by applying a small filter (e.g. 2× 2). The most popular pooling function is the max function (e.g. outputting the maximum values of 2× 2 regions, thus reducing the dimension by 75%).

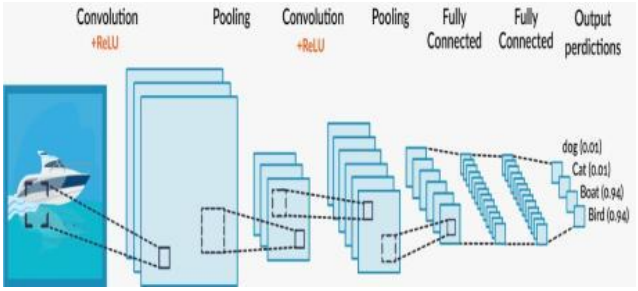


Fully-connected Layer:

Fully-connected layers, as the name suggests, is like ordinary NN where each neuron is connected to all the outputs from the previous layer. The last FC layer computes probabilities for each class. For multi-class classification, softmax is a popular choice. Softmax regression has the following log-likelihood function:

$$l(\theta) = \sum_{i=1}^{m} log \prod_{l=1}^{k} \left(\frac{exp(\theta_l^T x^{(i)})}{\sum_{j=1}^{k} exp(\theta_j^T x^{(i)})} \right)^{1\{y^{(i)}=l\}}$$

That is, Softmax trains the final layer to correctly predicts with maximal confidence for each image.



Extra layer

To maximize the value from transfer learning, we added a few extra layers to help the model adapt to our use case.

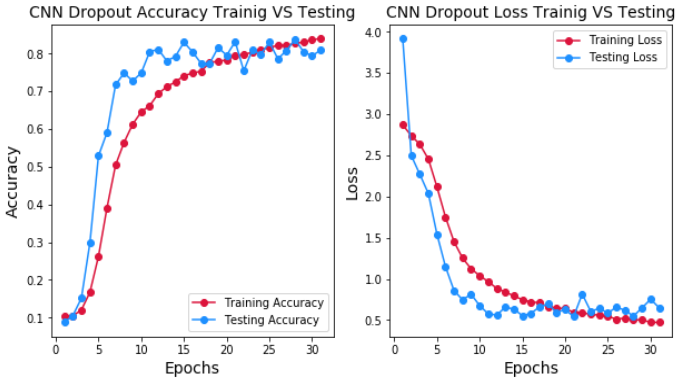
- Global average pooling: This layer retains only the average of the values in each patch.
- Dropout: This layers help in controlling for over fitting as it drops a faction of parameters.
- Batch normalization :This layer normalizes the inputs to the next layer which allows faster and more resilient training.
- Dense: This layer is the regular fully-connected layer with a specific activation function.

6.6 Train

The first question when doing transfer learning is if we should train only the extra layers added to the pre-existing architecture or if we should train all the layers. Naturally, we started by using the ImageNet weights and trained only the new layers since the number of parameters to train would be lesser and the model would train faster. We keep training the sub folder images that is the categorised train data. The 25 epochs are generated and every time the training accuracy increases. Once the accuracy rate comes to stable the training process terminates.

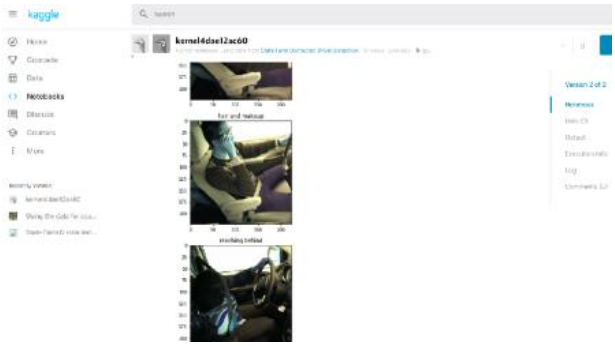
6.7 Plot graphs and compare the results

The graph is been plotted to compare the accuracy rate with the increasing epoch numbers for the training and testing data. Dropout is a regularization technique patented by google for reducing overfitting in neural networks by preventing complex co-adaptions on training data. It is an efficient way of performing model averaging with neural networks. The term “dropout” refers to dropping out units in a neural network.



6.8 Testing the test images:

The images that have been used for training and testing respectively are 9135 and 1865. The images that have been used while training are categorised data format. The testing includes the randomised dataset of images and the output is been calculated and the result is been analysed. The output will be of the format giving accuracy for each category of data used while training. The highest accuracy is been considered and given as output. The result that has been generated has an accuracy of 85%. Thus we conclude this step.



6.9 Test on real world images.

The final step is we test our model with a real world data and find the accuracy rate and the results are been recorded for different actions and people.

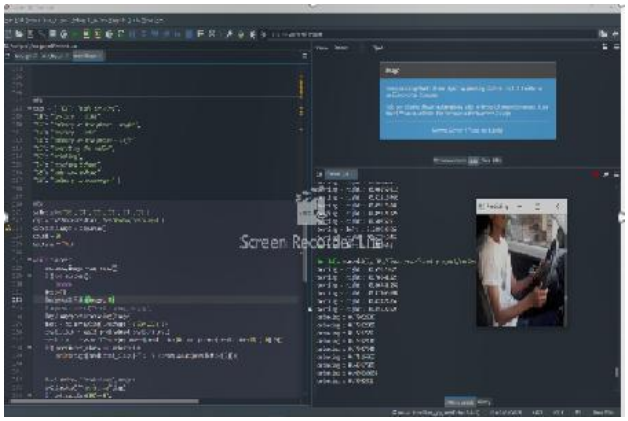


Figure 7.1:Architecture of proposed system

VI. PROPOSED WORK PLAN

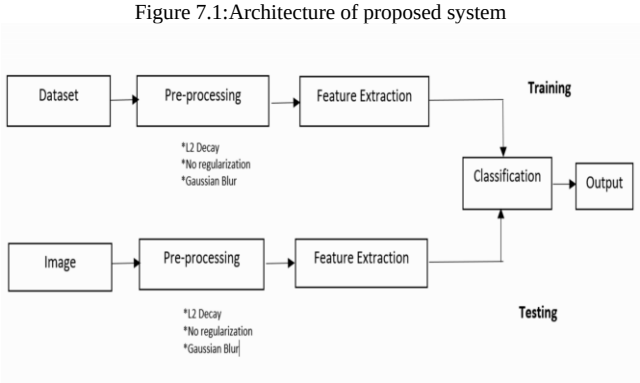


Figure 6.2:Architecture of proposed system

Training phase:
In this phase video will be converted into frames and then the victim images will be extracted through key frame extraction. Then the images will be resized and converted into the smallest pixel and the CNN filters will be applied. From these images the key features will be extracted. These features of image will be sent for classification.

Testing phase**:**
In this phase images will be extracted through key frame extraction. Then the images will be resized and converted into the smallest pixel and the CNN filters will be applied. From these images the key features will be extracted. These features of image will be tested with the training data and the desired part will be retrieved.

REFERENCES

[1] Mahbub Hussain, Jordan J. Bird and Diego R. Faria "A Study on CNN Transfer Learning for Image Classification", School of Engineering and Applied Science Aston University,UK, june 2018.
[2] Deepika Jaswal, Sowmya. V, K.P.Soman "Image Classification Using Convolutional Neural Networks",International Journal of Advancements in Research and Technology, Volume 3, June-2014.
[3] Milan R. Shetake, Sanjay. B. Waikar, "Content Based Image and Video Retrieval",International Journal of Advances in Electronics and Computer Science, Volume-2, Sept-2015.
[4] Karthik Umesh Sharma and Nileshsingh V. Thakur "A review and an approach for object detection in images", International Journal of Computational Vision and Robotics, Volume 7, Nov-2017.
[5] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu and Xindong Wu "Object Detection with Deep Learning", IEEE Transactions on neural networks and Learning Systems, 16 Apr 2019. [6] Osisanwo F.Y. ,Akinsola J.E.T. Awodele O, Hinmikaiye J.O "Supervised Machine Learning Algorithms: Classification and Comparison,International Journal of Computer Trends and Technology, Volume 48, June 2017.
[6] Christoph Feichtenhofer, Axel Pinz and Andrew Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition", 26 Sept 2016.
[7] Fernando Moya Rueda, Rene Grzeszick, Gernot A. Fink , Sascha Feldhorst and Michael ten Hompel, " Convolutional Neural Networks for Human Activity Recognition Using BodyWorn Sensors" ,25 May 2018.
[8] Vrushali A. Wankhede, Prakash S. Mohod, "A Review on Content-Based Image Retrieval from Videos using Self Learning Object Dictionary", International Journal of Science and Research, 2012.
[9] Byeong-Ho KANG, "A Review on Image and Video Processing", International Journal of Multimedia And Ubiquitos Engineering,Volume 2, April 2017.