International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 15, Issue 1 (January 2019), PP.79-86

Video Concept Detection Using Convolutional Neural Network

Ritika D Sangale[1], Nita S Patil[2], Sudhir D Sawarkar[3]
1(Department of Computer Engineering, DattaMeghe College of Engineering, Navi Mumbai, India)
2(Department of Computer Engineering, DattaMeghe College of Engineering, Navi Mumbai, India)
3(Department of Computer Engineering, DattaMeghe College of Engineering, Navi Mumbai, India)
Corresponding Author: Ritika D Sangale

ABSTRACT: *The performance of the video concept detection method depends on, the selection of the low-level visual features used to represent key-frames of a shot and CNN based Model. Video Concept detection is the task of assigning an input video one or multiple labels indicating the presence of one or multiple concepts in the video sequence. In this paper we have implemented video concept detection using a deep CNN and the low level visual features modeled using GMM supervectors. In the experiment, CNN achieves better classification accuracy on TRECVID dataset due to the capability of feature and classifier learning.*

Date of Submission: 02-04-2019 Date of acceptance: 18-04-2019

I. INTRODUCTION

Multimedia data is an inimitable source of information which presents both opportunities and challenges. Video Concept detection is defined as a process of detecting concept present in the video or selected video frame. Video is the collection of multiple key-frames. The major challenge of concept detection lies in the existence of the well-known semantic gap between the low-level visual features and the interpretation of the same data as semantics. The definition of Video Semantic Concept detection is as the task of assigning an input video labels indicating that the presence of one or more semantic concepts in the video. Such semantic concepts can be like people, activities like “fighting”, “swimming”, “running”, object which includes “sky”, “tree”, “car” and scenes which includes “outdoors”, “desert”, “beach” etc. Figure 1. shows various concepts that are present in the sample video frame.

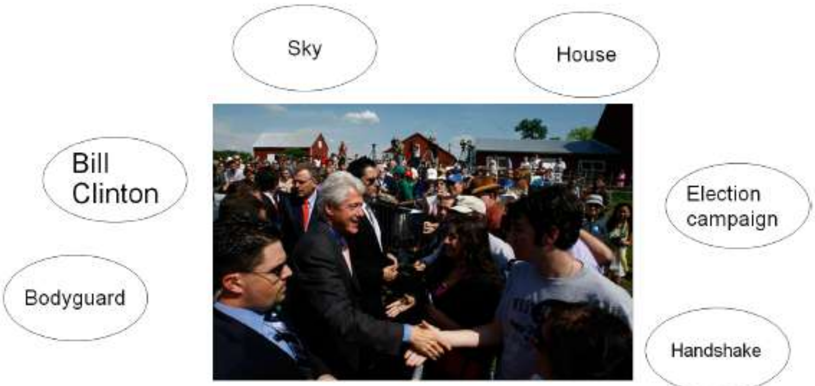


Figure 1: Various semantic concepts visible in a video frame.

Concept detection or semantic indexing are commonly used terms describing the task of recognizing concepts in video. Such semantic concepts can be anything of user’s interest that is visually observable [1]. The goal of concept detection, or high-level feature extraction, is to build mapping functions from the low-level features to the high-level concepts with machine learning techniques.

The rest of the paper is organized as follows: Section II consists of Literature Review. Section III gives a brief overview of work related to visual concept detection using low level descriptors and CNN. And also describe the various feature descriptors and methodology used for experimentation. Section IV presents experimental results Section V includes conclusion of the paper.

II. LITERATURE REVIEW

JingweiXu and LiSong, and RongXie2016 [2], proposed a novel SBD framework based on representative features extracted from CNN. This scheme is suitable for detection of both CT and GT boundaries. Authorsachieves excellent accuracy in shot Boundary Detection.

Ganesh. I. Rathod, Dipali. A. Nikam [3], proposed An Algorithm for Shot Boundary Detection and Key Frame Extraction Using Histogram Difference. This paper work a Square histogram based model is developed using frame segmentation and automatic threshold calculation. In this paper the keyframe is extracted by using a reference frame approach per shot. A total of around 40 videos of different types are tested on this model and the model is able to detect all shot boundaries and is storing the suitable frames as keyframes to represent the video summary. An efficiency of almost 95% to 98% is observed using this algorithm.

Alex Krizhevsky, IlyaSutskever, Geoffrey E. Hinton2012 [4], author trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet ILSVRC-2010 contest into the 1000 different classes.

Nakamasa Inoue and Koichi Shinoda and Zhang Xuefeng and Kazuya Ueki 2014 [7],authors proposed a high-performance semantic indexing system using a deep CNN and GMM supervectors with the six audio and visual features. The result was 28.1 % in terms of Mean InfAP, which was ranked third among participating teams in the semantic indexing task.

Samira Pouyanfar and Shu-Ching Chen [8], a novel ensemble deep classifier is proposed which fuses the results from several weak learners and different deep feature sets. In this Paper proposed framework is designed to handle the imbalanced data problem in multimedia systems.

III. SYSTEM OVERVIEW

In this section we present the methodology adopted for visual concept detection using low level features and Convolutional Neural Network (CNN)**.** Figure 2 shows the architecture of proposed work.

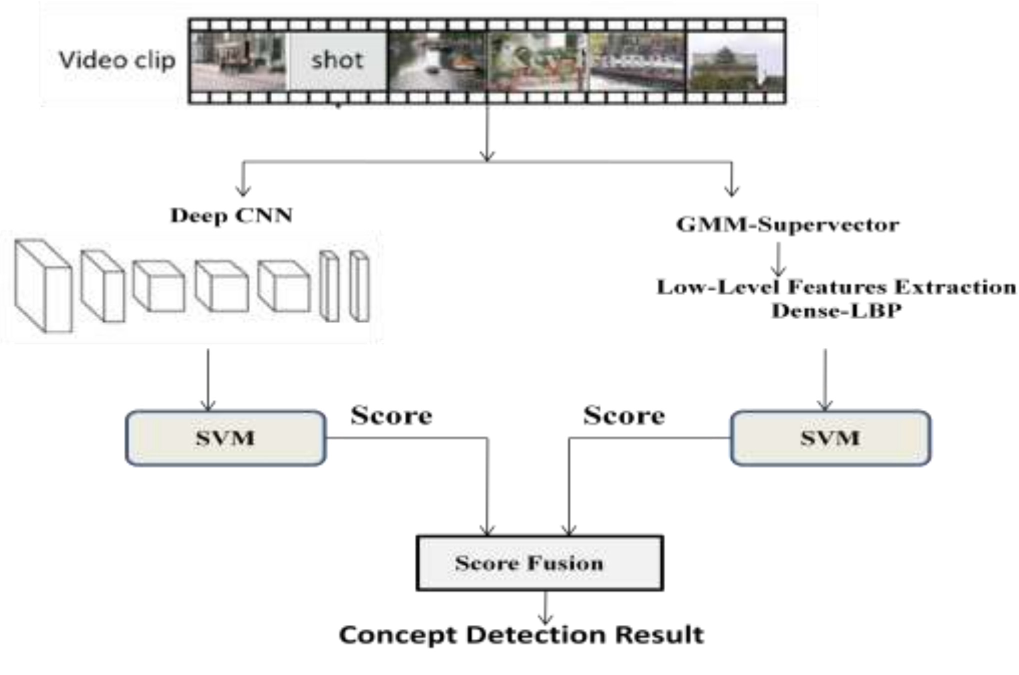


Figure 2: Proposed Architecture

Our proposed Architecture is divided into three main modules:
(1) Preprocessing stage,
(2) Deep feature extraction, and
(3) Classification stage which includes training, validation, and testing.

In our system video is taken as input, video clip is processed in three steps: shot boundaries detection, shot selection and key frame extraction from the selected shot. First the video is divided into shots, then the representative shot is determined, and finally a frame is which best represent the shot is select as key-frame.

3.1 Shot Boundary detection:

The first step in edited video analysis and characterization is shot detection. Methods for shot boundary detection usually first extract visual features from each frame, then measure similarities between frames using the extracted features, and, finally, detect shot boundaries between frames that are dissimilar [2]. Previous techniques focused on cut detection, and more current work has focused on gradual transitions detection. We detect shot boundaries of a given video by using the grid based edge comparison between consecutive frames. It also provides some general features about the video like;

- number of black frames => to capture fade effects
- number of segments in video => to keep tract of the tempo of an video. If it is an action movie trailer it is likely to be faster.
- avg distance between frames of segments => If motion in a segment is higher then this value will be higher.
- number of total segments => all the segments dectection without thresholding.

Our input video is divided into shots and all the shots are stored inside the SAVE_SHOTS Folder. Next process is we have to select any video from this folder and keyframes are extracted from the selected video shot. Figure 3 shows input video is divided into no. of shots.

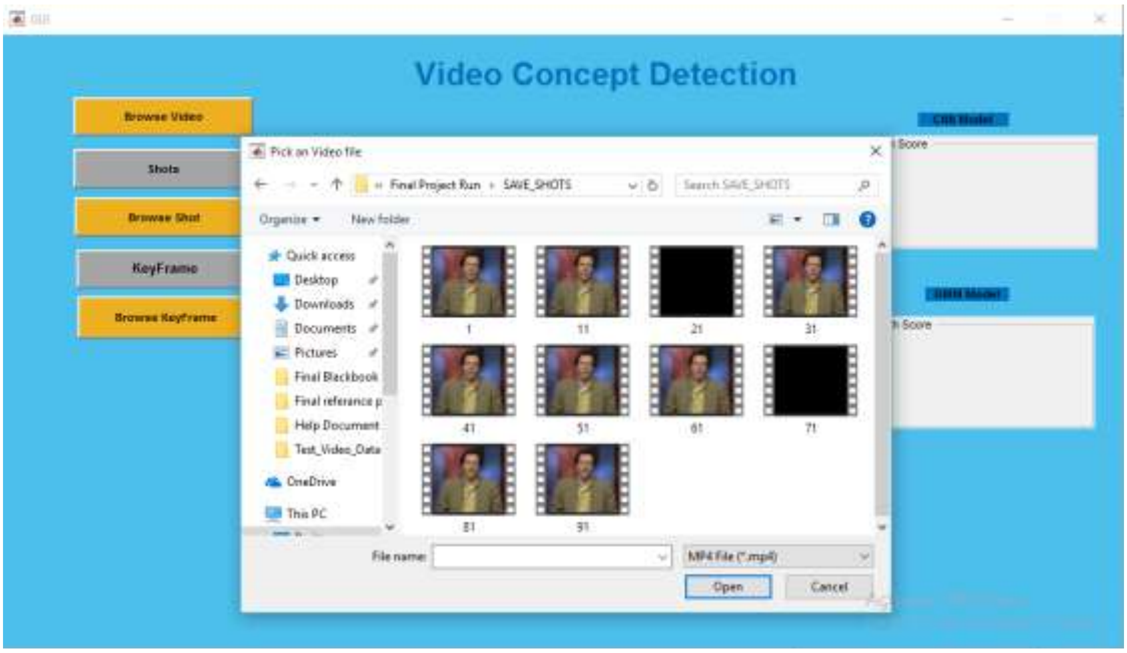


Figure 3: Video is divided into shots.

3.2 Key-frame Extraction:

We use Histogram Comparisons Method for key-frame Extraction. In Histogram Comparisons Extracts key frames from video using function VideoReader by calculating histogram difference [3].

Extracting Frames From Shot steps:

1. Calculate number of frames.
2. Read the frames of shots of video.
3. Calculate histogram of each frame and find histogram difference of the adjacent frames.
4. Calculating mean and standard deviation and extracting frames.
5. If histogram difference is greater than threshold value frame is selected as a key frame.

Extracted key-frame is given as input to Deep CNN for extracting deep features using pre-trained alexnet architecture. Figure 4 shows the shot is divided into no of key-frames.

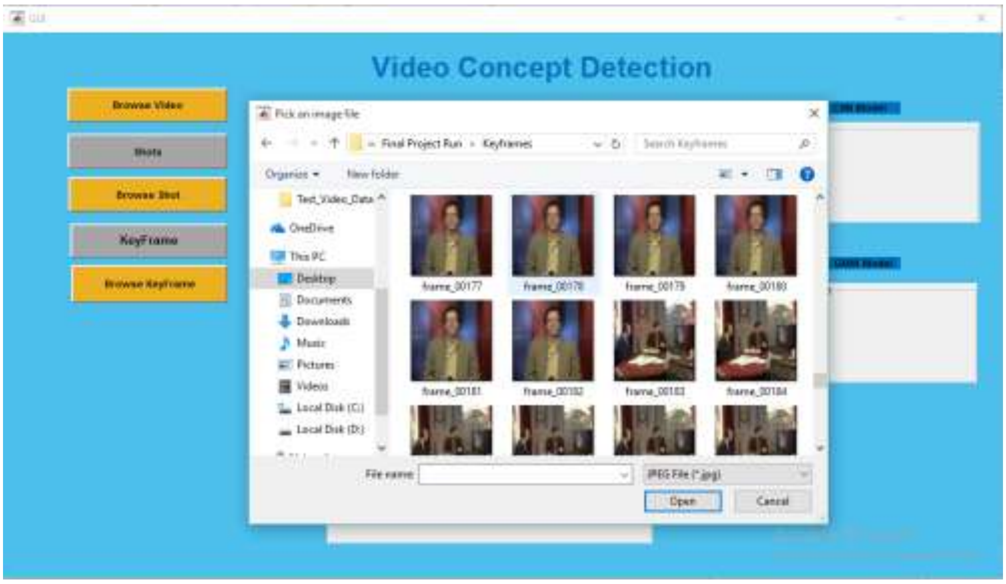


Figure 4: Key-frames extracted from shots.

Following section describes the architecture of Alexnet used in the paper.

3.3 Deep Features Using Convolutional Neural Networks (CNN):

We use deep convolutional neural network for visual feature extraction. For this purpose, instead of training an entire CNN from scratch, we take the pre-trained reference models and treat the convolutional networks as feature extractors for our datasets. Theses reference models are pre-trained on very large-scale datasets. Specifically, our selected models (Alexnet) have more impacts on the image processing field in recent years.

We used Alexnet CNN structure for deep feature extraction of key-frames. Figure 5 shows the architecture of Alexnet model trained on Imagenet dataset consisting of 1 million images.

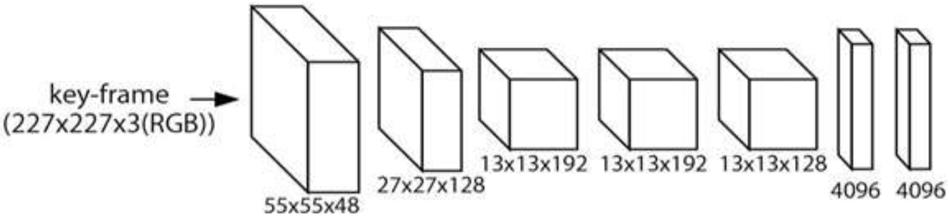


Figure 5: Architecture of Deep CNN [4]

Layer Description in Architecture of Deep CNN is as follows.
Layer 0 is Input layer where key-frame are resized to 227 x 227 x 3 and given as input.
1. Layer 1: is Max-Pooling with 48 filters, size 11×11and stride of 4, padding 0 of size: 55 x 55 x 48, (227-11)/4 + 1 = 55 is the size of the outcome and 48 depth because 1 set denotes 1 filter and there are 48 filters.
2 .Layer 2: is Max-Pooling layer with 128 filters and stride of 2 having size: 27 x 27 x 128, (55 – 3)/2 + 1 = 27 is size of outcome and depth is 128 because pooling is done independently on each layer.
3. Layer 3: is Convolution with 192 filters and stride of 2 having Size: 13 x 13 x 192, (27 – 3)/2 + 1 = 13 is size of outcome and Depth is 192 because of 192 filters.
4. Layer 4: is Convolution with 192 filters, size 3×3, stride 1, and padding 1: Size: 13 x 13 x 192, Because of padding of (3-1)/2=1, the original size is restored and 192 depth because of 192 filters.
5. Layer 5: is Max-Pooling with 128 filters and stride 2of having size 13 x 13x 128, and Depth is128 because pooling is done independently on each layer.
6. Layer 6: Fully connected with 4096 neuron: In this layer, each of the 13 x 13x 128, pixels are fed into each of the 4096 neurons and weights determined by back-propagation.
7.Layer 7: Fully Connected with 4096 neuron: Similar to layer6.finally Fully Connected with 1000 neurons this is the last layer and has 1000 neurons because ImageNET data has 1000 classes to be predicted [4].

In this paper, Deep CNN trained on the Alexnet dataset is adopted to extract features from video key-frame .A 4096-dimensional feature vector is extracted from the key-frame of each concept by using the CNN. The first to fifth layers are convolutional layers, in which the first, second, and fifth layers have max-pooling procedure. The sixth and seventh layers are fully connected. In our proposed system parameters of the CNN is trained on the dataset with 36 object categories. Finally, from each keyframe, we extract a 4096-dimensional feature at the sixth layer to train an SVM for each concept in the Video Concept Detection [4].

3.4 Low-Level Feature Extraction:

For Low-level Feature Extraction we have used Dense LBP visual feature extraction technique to extract feature from input key-frame. All the features are given to GMM Supervector. GMM supervector represents the distribution features of each key-frame. GMM supervector extracts the set of all low-level features as feature vector. This feature vector is used to train SVM for each concept.
Two models are trained on different features and are called as CNN and SVM models.

3.4.1 LBP with dense sampling (LBP-Dense): Local Binary Pattern (LBP) is a simple yet very efficient texture operator which labels the pixels of an image by thresholding the neighbourhood of each pixel and considers the result as a binary number. Due to its discriminative power and computational simplicity, LBP texture operator has become a popular approach in various applications [7].

In LBP feature first we dividing the keyframe into 8*8 block and applying LBP to each block. extractLBPFeatures () this function is used to extract the LBP features. This operator works with the eight neighbors of a pixel, using the value of this center pixel as a threshold. If a neighbor pixel has a higher gray value than the center pixel (or the the same gray value) than a one is assigned to that pixel, else it gets a zero. The LBP code for the center pixel is then produced by concatenating the eight ones or zeros to a binary code

3.4.2 GMM Supervector: This method is used to create feature vector of Local Extracted feature, there is LBP types of features are extracted from keyframe in low level feature extraction. Input Video is divided into Shot; select any shot from the divided shots. Then shot is converted into key-frames using histogram difference technique. Select keyframe is the input of the GMM Supervector. From the keyframe low level features are extracted to create feature vector. A Gaussian mixture model (GMM) is useful for modeling data that comes from one of several groups: the groups might be different from each other, but data points within the same group can be well-modeled by a Gaussian distribution. A GMM supervector consists of the parameters of a GMM for the distribution of low-level features extracted from a Keyframe [7]. SVM are used to train discriminative models for each Semantic Concept.

3.6 Dataset Design:

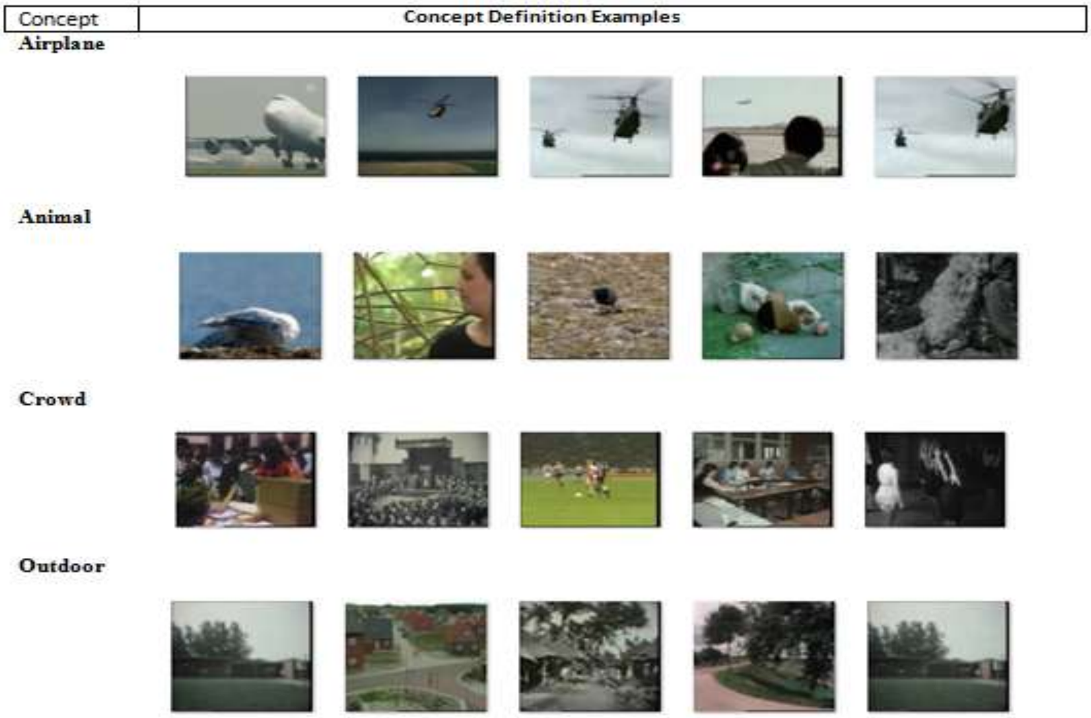
The TRECVID 2007 data set is composed of 111 video clips separated into two groups, the development set and testing set. There are 36 defined concepts in the dataset. The concept list is given in Table 1.The 36 concepts are manually annotated over these key-frames. For a concept, positive key-frame is defined as a frame containing a said concept as a visual content. Table 1 lists some of the concepts and Table 2 concept defining key-frames in the dataset.

Table 1: Concept list in TRECVID development dataset.

Sr. No.	Concept	Sr. No.	Concept
1	Airplane	19	Natural-Disaster
2	Animal	20	Office
3	Boat_ship	21	Outdoor
4	Building	22	People-Matching
5	Bus	23	Person
6	Car	24	Police-Security
7	Charts	25	Prisoner
8	Computer_Tv-Screen	26	Road
9	Court	27	Sky
10	Crowd	28	Snow
11	Desert	29	Sports
12	Explosion_Fire	30	Studio

13	Face	31	Truck
14	Flag_US	32	Urban
15	Maps	33	Vegetation
16	Meeting	34	Walking-Running
17	Military	35	Waterscape-Waterfront
18	Mountain	36	Weather

Table 2: Concept definition examples from the TRECVID development dataset.



3.7 Data preparation:

Table 3: Partition details of TRECVID development dataset

Dataset	Dataset Name	Partition	#of Key-Frames
TRECVID Development Dataset	**Partition-I**	**Training Dataset**	**180**
	Partition-II	**Testing Dataset**	**180**

CNN-SVM trained model and Low-level Feature-GMM-SVM trained model will produce Scores of the features After calculating score, we perform the score fusion . Score fusion is a feature-fusion technique where scores resulting out of classifiers are combined using some strategy and final detection scores for each concept are computed. There are three fusion strategies are Linear, Average, and Max and their details are as follows-

- **Linear**: Performs a grid search in fusion parameter space to select the optimal weights.
- **Average**: The scores resulting from each classifier are simply averaged to generate the fused score.
- **Max**: For each concept, the best performance is selected [5].

In the experimentation MAX fusion strategy is used. Scores are fused together and highest score gives the video concept detection result.

For testing a new video, video is first segmented into shots. First the video is divided into shots, and then the shot is selected by user. Key-frame extraction algorithm is run in the background and all extracted key-frames from the selected shot are displayed to the user. User can select any one keyframe from extracted keyframes. The selected key-frame is now taken as input to both the trained models which predicts score for keyframe from predefined list of concepts. Figure 6 shows the block diagram of testing process.

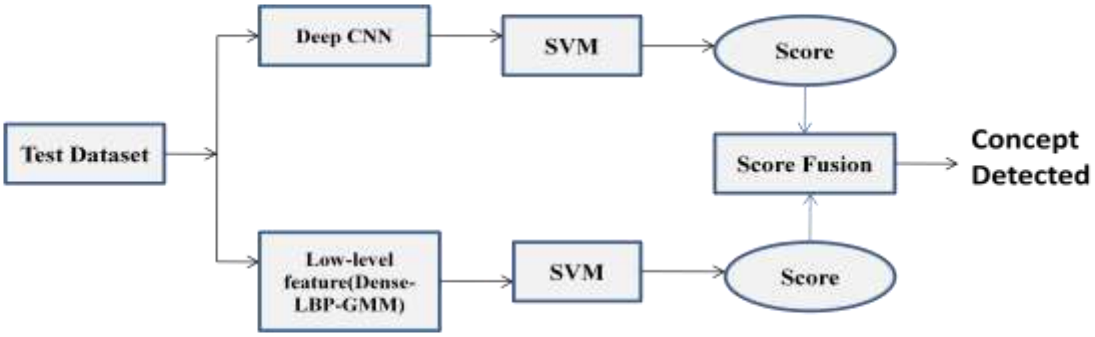


Figure 6: Block diagram of testing process.

IV. EXPERIMENTAL RESULTS

For implementation of project, we have used MATLAB R2018a as developing environment and runtime platform.

4.1 Deep CNN: To Build DCNN train model, we take the pre-trained reference model Alexnet and treat the convolutional networks as feature extractors for our dataset.In DCNN augmentedImageDatastore() function is used and this function is available on R2018a matlab edition.

An augmented image datastore generates batches of new images, with optional preprocessing such as resizing, rotation, and reflection, based on the training images.

4.2 Low-Level Feature Extraction: In low level feature we use Dense Local Binary Patterns (LBPs) ,input is taken as Gray image that should be in type Single and wndsize is the sliding window size that should be positive by default wndsize=5.Output of the dense-LBP is a matrix with LBP feature.

At the end of low level feature extraction, GMM Supervectors are used to combine all the extracted feature vector and by combining all feature vector to form normalized mean Vectors.

4.3 Classifier:

The machine learning toolbox for MATLAB R2018a contains an integrated Support Vector Machine. It is realized by function fitcecoc () which is multiclass models for support vector machines or other classifiers. fitcecoc() uses K(K – 1)/2 binary support vector machine (SVM) models using the one-versus-one coding design, where K is the number of unique class labels.

Keyframe pass through from both the models. Figure 6, will show the final detected concept of the keyframe. When the keyframe pass through our trained models it will produce scores and on the bases of this score concept are detected.

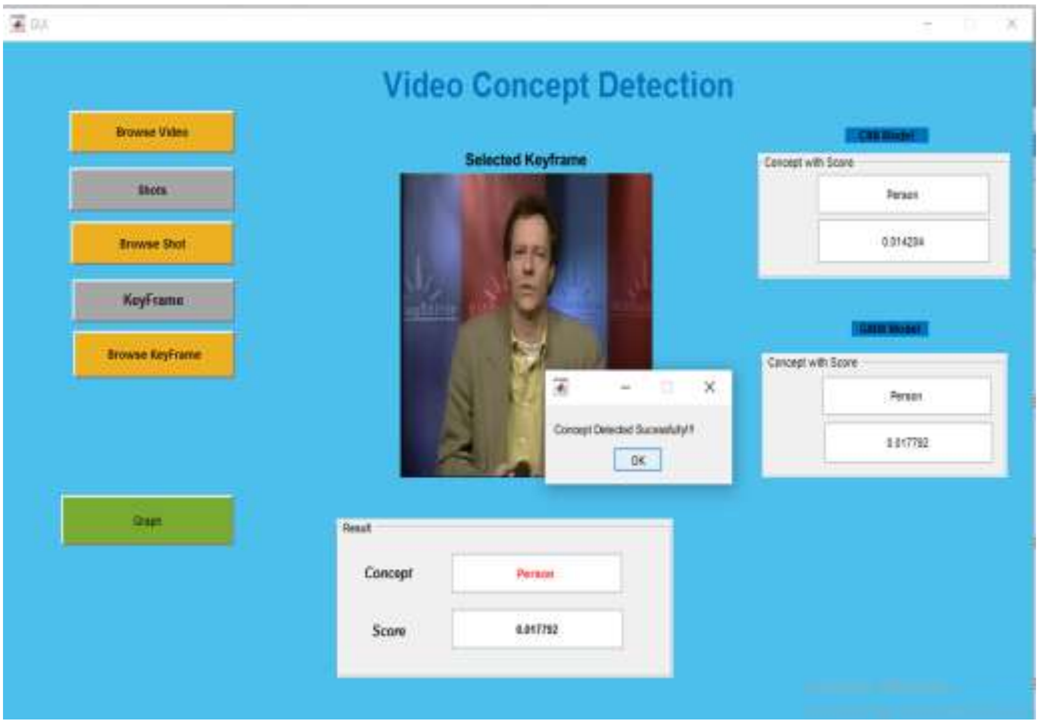
From the Figure 7, we observed that selected keyframe passes through CNN Train Model its gives 0.014201 Score and Same Keyframe passes through GMM Train model its gives 0.017792 Score.

Fused both score and Combined Both Scores:

CNN Score = 0.014201

GMM Score = 0.017792

Higher Score predict the final score and detected concept label i.e. person.



Concept with Score: Person, 0.014294

Concept with Score: Person, 0.017792

Concept Detected Sucessfully!!

Result
Concept: Person
Score: 0.017792

Figure 7: Concept Detected.

V. CONCLUSION

Looking at the all cons of the traditional approach of concept detection we proposed video concept detection using a deep CNN and GMM super vectors with the low level visual features. It is observed that CNNs achieve better classification accuracy on large scale datasets due to their capability of joint feature and classifier learning. Although our framework is simple and easy to build, experimental results show that it can achieve good performance compared to other complicated systems while requiring less resource and computational cost. This finding is important for practical applications that need to process thousands of hours of videos and hundreds of concepts. Results observed in the comparative study with other traditional methods suggest that CNN gives better accuracy and boosts the performance of the system due to unique features like shared weights and local connectivity.

REFERENCES

[1]. Gert-Jan POULISSE “Complex Semantic Concept Detection in Video”,2012.

[2]. JingweiXu, Li Song, RongXie “Shot Boundary Detection Using Convolutional Neural Networks” in IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Ghent, 2016.

[3]. Ganesh. I. Rathod, Dipali. A. Nikam, “An Algorithm for Shot Boundary Detection and Key Frame Extraction Using Histogram Difference” International Journal of Emerging Technology and Advanced Engineering,Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 8, August 2013)

[4]. Alex Krizhevsky, IlyaSutskever, Geoffrey E. Hinton “ImageNet Classification with Deep Convolutional Neural Networks” University of Toronto,2012

[5]. Nitin J. Janwe and Kishor K. Bhoyar “Semantic Video Concept Detection using Novel Mixed-Hybrid-Fusion Approach for Multi-Label Data”, Electronic Letters on Computer Vision and Image Analysis 16(3):14-29; 2017

[6]. David G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”,Computer Science Department ,University of British Columbia Vancouver, B.C., Canada ,lowe@cs.ubc.ca January 5, 2004.

[7]. Nakamasa Inoue and Koichi Shinoda and Zhang Xuefeng and Kazuya Ueki “Semantic Indexing Using Deep CNN and GMM Supervectors”, Tokyo Institute of Technology, Waseda University,2014.

[8]. Samira Pouyanfar and Shu-Ching Chen “Automatic Video Event Detection for Imbalance Data Using Enhanced Ensemble Deep Learning” International Journal of Semantic Computing World Scientific Publishing Company, 2017.

[9]. AshwinBhandare , Maithili Bhide , PranavGokhale , RohanChandavarkar Applications of Convolutional Neural Networks ” AshwinBhandare et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (5) , 2016

[10]. Jianxin Wu, “Introduction to Convolutional Neural Networks”

[11]. Ritika D Sangale,Nita S Patil,Sudhir D Sawarkar,“Convolutional Neural Network (CNN) and GMM Supervectors”,International Journal of Engineering Science Invention (IJESI)_ISSN(Online): 2319 – 6734, ISSN (Print): 2319 – 6726www.ijesi.org ||Volume 7 Issue 8 Ver IV || Aug 2018 || PP 69-75

Ritika D Sangale" Video Concept Detection Using Convolutional Neural Network" International Journal Of Engineering Research And Development , vol. 15, no. 1, 2019, pp 79-86